Guaranteed, locally efficient, and robust a posteriori estimates for nonlinear elliptic problems in iteration-dependent norms. An orthogonal decomposition result based on iterative linearization

Koondanibha Mitra, Martin Vohralík

Guaranteed, locally efficient, and robust a posteriori estimates for nonlinear elliptic problems in iteration-dependent norms. An orthogonal decomposition result based on iterative linearization

K. Mitra[1] and M. Vohralík[2,3]

[1]Hasselt University, Belgium
[2]Inria, 2 rue Simone Iff, 75589 Paris, France
[3]CERMICS, Ecole des Ponts, 77455 Marne-la-Vallée, France

July 9, 2023

Abstract

We consider numerical approximations of nonlinear, monotone, and Lipschitz-continuous elliptic problems, with gradient-dependent or gradient-independent diffusivity. For this purpose, we first employ a symmetric iterative linearization, with examples including respectively the Kačanov, Newton, Zarantonello and fixed-point (Picard), L, and M-schemes. We design an iteration-dependent norm, stemming from the concrete linearization and finite element discretization, which leads to an orthogonal decomposition of the total error into the linearization and discretization components. This norm is of reaction–diffusion type, so that available results from a posteriori analysis of linear (singularly perturbed) reaction–diffusion problems enable us to obtain guaranteed and locally efficient error bounds robust with respect to the strength of the nonlinearity. Moreover, the clear distinction and orthogonality of the error components leads to a design of adaptive stopping criteria for iterative linearizations ensuring a natural balance between discretization and linearization components. Under appropriate assumptions, we also show that the iteration-dependent norm is, up to a multiplicative factor tending to one, equivalent to the iteration-independent natural discrete energy norm. In an appendix, we address the Newton (nonsymmetric) linearization for the gradient-independent diffusivity case. Numerical experiments illustrate the theoretical developments.

Contents

1 Introduction

For $d \geq 1$, let $\Omega \subset \mathbb{R}^d$ be an open, bounded, and connected polytope with Lipschitz boundary. Let $u \in H_0^1(\Omega)$ solve the partial differential equation

$$\langle \mathcal{R}(u), \varphi \rangle = 0 \qquad \forall \varphi \in H_0^1(\Omega), \tag{1.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$ and $\mathcal{R} : H_0^1(\Omega) \to H^{-1}(\Omega)$ is a nonlinear elliptic operator which satisfies the monotonicity and Lipschitz-continuity conditions, i.e., there exist real constants $\lambda_{\mathrm{M}} \geq \lambda_{\mathrm{m}} > 0$ such that for an arbitrary $u_\ell \in H_0^1(\Omega)$,

$$\lambda_{\mathrm{m}} \operatorname{dist}(u_\ell, u) \leq \sup_{\varphi \in H_0^1(\Omega)} \frac{\langle \mathcal{R}(u_\ell) - \mathcal{R}(u), \varphi \rangle}{\|\nabla \varphi\|} \leq \lambda_{\mathrm{M}} \operatorname{dist}(u_\ell, u). \tag{1.2}$$

Above, $\operatorname{dist} : H_0^1(\Omega) \times H_0^1(\Omega) \to [0, \infty)$ represents a distance metric that is either equivalent to the $H_0^1(\Omega)$-norm, or becomes equivalent to it after a transformation of variables, and $\|\cdot\|$ is the $L^2(\Omega)$ norm. We refer to [17, 26, 32, 46, 48, 57, 68] and the references therein, as well as to what follows, for details and examples.

1.1 $H_0^1(\Omega)$-error-norm and energy difference a posteriori estimates

Let $V_\ell \subset H_0^1(\Omega)$ be a finite element space. Let a corresponding finite element approximation $u_\ell \in V_\ell$ of u from (1.1) be given by

$$\langle \mathcal{R}(u_\ell), \varphi_\ell \rangle = 0 \qquad \forall \varphi_\ell \in V_\ell. \tag{1.3}$$

It is well known from the literature [3, 4, 8, 14, 15, 18, 19, 23, 24, 27–33, 35–38, 40–46, 54, 56–59, 63] that it is possible to construct (a guaranteed) a posteriori estimator $\eta(u_\ell)$, fully computable from u_ℓ, that is also efficient in that

$$\lambda_{\mathrm{m}} \operatorname{dist}(u_\ell, u) \leq \eta(u_\ell) \leq C \lambda_{\mathrm{M}} \operatorname{dist}(u_\ell, u), \tag{1.4}$$

where $C > 0$ is a generic constant independent of u, u_ℓ, and the global Lipschitz/monotonicity ratio $\lambda_{\mathrm{M}}/\lambda_{\mathrm{m}}$. Similarly, if an associated energy J of the operator $\mathcal{R}$ exists, then one may obtain

$$J(u_\ell) - J(u) \leq \tilde{\eta}(u_\ell)^2 \leq C^2 \frac{\lambda_{\mathrm{M}}^2}{\lambda_{\mathrm{m}}^2} (J(u_\ell) - J(u)). \tag{1.5}$$

(Actually, this has recently been improved to only feature $\lambda_{\mathrm{M}}/\lambda_{\mathrm{m}}$ in [33, Theorem 4.1].) However, the ratio $\lambda_{\mathrm{M}}/\lambda_{\mathrm{m}}$ that appears in both cases above can be large, and even unbounded in the degenerate limit. Thus, the above a posteriori error estimates may become imprecise and are in particular not robust with respect to the strength of the nonlinearity represented by this ratio.

1.2 Dual norm of the residual a posteriori estimates

To circumvent the dependence on $\lambda_{\mathrm{M}}/\lambda_{\mathrm{m}}$, observe that

$$\sup_{\varphi\in H_0^1(\Omega)} \frac{\langle \mathcal{R}(u_\ell)-\mathcal{R}(u),\varphi\rangle}{\|\nabla\varphi\|} = \sup_{\varphi\in H_0^1(\Omega)} \frac{\langle \mathcal{R}(u_\ell),\varphi\rangle}{\|\nabla\varphi\|} = \|\mathcal{R}(u_\ell)\|_{H^{-1}(\Omega)}, \tag{1.6}$$

since $u \in H_0^1(\Omega)$ solves (1.1). One can then consider $\|\mathcal{R}(u_\ell)\|_{H^{-1}(\Omega)}$ directly as an error measure, since it in particular vanishes if and only if $u_\ell = u$ (it is equivalent to $\mathrm{dist}(u_\ell, u)$ by (1.2)). Moreover, $\|\mathcal{R}(u_\ell)\|_{H^{-1}(\Omega)}$, though defined as a global dual norm, is localizable by the orthogonality to the lowest-order shape functions from (1.3); this is explicitly recalled in, e.g., [6,11,16], see also the references therein. Furthermore, it can be a posteriori estimated robustly by the computable estimator $\eta(u_\ell)$, i.e.,

$$\|\mathcal{R}(u_\ell)\|_{H^{-1}(\Omega)} \leq \eta(u_\ell) \leq C\|\mathcal{R}(u_\ell)\|_{H^{-1}(\Omega)}, \tag{1.7}$$

where the constant is as in (1.4) independent of the global Lipschitz/monotonicity ratio $\lambda_{\mathrm{M}}/\lambda_{\mathrm{m}}$, see [14, 15, 23, 24]. Nevertheless, $\|\mathcal{R}(u_\ell)\|_{H^{-1}(\Omega)}$ might be too weak an error measure since it does not carry any physical units/scaling/energy information and may not correspond to an error metric either locally or globally.

1.3 Our goals

In the present contribution, we are interested in deriving a posteriori error estimates for finite element approximations (1.3) of (1.1) independent of the factor $\lambda_{\mathrm{M}}/\lambda_{\mathrm{m}}$, i.e., robust with respect to the strength of the nonlinearity, in contrast to (1.4) or (1.5). In improvement of (1.7), energy information and local norm correspondence are encoded. Based on iterative linearization, we also derive an orthogonal decomposition result of the total error into the discretization and linearization error components. This finally allows us to design an adaptive stopping criterion for iterative linearizations ensuring a natural balance between these error components.

1.4 Motivation

We start by motivating our approach.

1.4.1 Energy norm estimates in the linear case

To gain insight, let us turn our attention for a while to a linear problem. Consider the linear diffusion equation

$$\langle \mathcal{R}(u),\varphi\rangle := (\bar{\boldsymbol{K}}\nabla u, \nabla\varphi) + \langle f,\varphi\rangle = 0 \qquad \forall\varphi \in H_0^1(\Omega), \tag{1.8}$$

where $f \in H^{-1}(\Omega)$ and, for a.e. $\boldsymbol{x} \in \Omega$, $\bar{\boldsymbol{K}}(\boldsymbol{x}) \in \mathbb{R}^{d\times d}$ is symmetric and satisfies the uniform ellipticity and boundedness condition, i.e., there exist real constants $\lambda_{\mathrm{M}} \geq \lambda_{\mathrm{m}} > 0$ such that $\lambda_{\mathrm{m}}\,|\boldsymbol{y}|^2 \leq \boldsymbol{y}^{\mathrm{T}}\bar{\boldsymbol{K}}(\boldsymbol{x})\boldsymbol{y} \leq \lambda_{\mathrm{M}}\,|\boldsymbol{y}|^2$ for all $\boldsymbol{y} \in \mathbb{R}^d$. In this case, by Céa's lemma [17], one has for the finite element approximation $u_\ell \in V_\ell$ given by (1.3) the a priori characterization

$$\|\nabla(u-u_\ell)\| \leq \tfrac{\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\|\nabla(u-\varphi_\ell)\| \qquad \forall\varphi_\ell \in V_\ell. \tag{1.9}$$

The estimate above implies that u_ℓ is only a quasi-optimal approximation of u from (1.8) in V_ℓ with respect to the intrinsic $H_0^1(\Omega)$-norm, not strictly leading to error minimization and involving the ratio $\lambda_{\mathrm{M}}/\lambda_{\mathrm{m}}$. Congruently, similar to (1.2) and (1.4), one only obtains the a posteriori characterizations

$$\lambda_{\mathrm{m}}\|\nabla(u-u_\ell)\| \leq \|\mathcal{R}(u_\ell)\|_{H^{-1}(\Omega)} \leq \lambda_{\mathrm{M}}\|\nabla(u-u_\ell)\|, \tag{1.10a}$$

$$\lambda_{\mathrm{m}}\|\nabla(u-u_\ell)\| \leq \eta(u_\ell) \leq C\lambda_{\mathrm{M}}\|\nabla(u-u_\ell)\|, \tag{1.10b}$$

which involve the ratio $\lambda_{\mathrm{M}}/\lambda_{\mathrm{m}}$ and are in particular not robust with respect to the inhomogeneity and anisotropy in $\bar{\boldsymbol{K}}$.

In contrast to the above, introducing the energy norm $|||\varphi|||_{1,\bar{\boldsymbol{K}}} := (\bar{\boldsymbol{K}}\nabla\varphi, \nabla\varphi)^{\frac{1}{2}}$, which is arguably the natural norm for this system, one has by Céa's lemma that (compare with (1.9))

$$|||u - u_\ell|||_{1,\bar{\boldsymbol{K}}} = \min_{\varphi_\ell \in V_\ell} |||u - \varphi_\ell|||_{1,\bar{\boldsymbol{K}}}. \tag{1.11}$$

This entails that $u_\ell \in V_\ell$ is the optimal approximation of u from (1.8) in V_ℓ with respect to the norm $|||\cdot|||_{1,\bar{\boldsymbol{K}}}$; this equality crucially does not involve the ratio $\lambda_{\mathrm{M}}/\lambda_{\mathrm{m}}$ and is consequently robust with respect to $\lambda_{\mathrm{M}}/\lambda_{\mathrm{m}}$. Moreover, (1.11) motivates to introduce the energy dual norm

$$|||\varsigma|||_{-1,\bar{\boldsymbol{K}}} := \sup_{\varphi \in H_0^1(\Omega)} \frac{\langle \varsigma, \varphi \rangle}{|||\varphi|||_{1,\bar{\boldsymbol{K}}}}$$

for $\varsigma \in H^{-1}(\Omega)$ and consider $|||\mathcal{R}(u_\ell)|||_{-1,\bar{\boldsymbol{K}}}$ as an error measure alternative to $\|\mathcal{R}(u_\ell)\|_{H^{-1}(\Omega)}$. In fact,

$$\begin{aligned} |||\mathcal{R}(u_\ell)|||_{-1,\bar{\boldsymbol{K}}} &= \sup_{\varphi \in H_0^1(\Omega)} \frac{\langle \mathcal{R}(u_\ell), \varphi \rangle}{|||\varphi|||_{1,\bar{\boldsymbol{K}}}} = \sup_{\varphi \in H_0^1(\Omega)} \frac{(\bar{\boldsymbol{K}}\nabla u_\ell, \nabla\varphi) + \langle f, \varphi \rangle}{|||\varphi|||_{1,\bar{\boldsymbol{K}}}} \\ &= \sup_{\varphi \in H_0^1(\Omega)} \frac{(\bar{\boldsymbol{K}}\nabla(u_\ell - u), \nabla\varphi)}{|||\varphi|||_{1,\bar{\boldsymbol{K}}}} = |||u - u_\ell|||_{1,\bar{\boldsymbol{K}}}, \end{aligned}$$

so that (1.10a) is also turned into equality. Just as $|||u - u_\ell|||_{1,\bar{\boldsymbol{K}}}$ can be estimated robustly a priori using (1.11), robust a posteriori error estimates of the form (compare with (1.10b))

$$|||u - u_\ell|||_{1,\bar{\boldsymbol{K}}} \le \bar{\eta}(u_\ell) \le C|||u - u_\ell|||_{1,\bar{\boldsymbol{K}}}$$

become available at least under some conditions on $\bar{\boldsymbol{K}}$, see [5,12,65] and the references therein. The use of the energy norm and the dual energy norm of the residual also lead to robust a posteriori estimates for singularly perturbed reaction–diffusion problems, see [1,2,42,43,60,64] and the references therein.

1.4.2 A first attempt in the nonlinear case

The above approach cannot immediately be generalized to the nonlinear case. Indeed, let us consider a nonlinear extension of (1.8)

$$\langle \mathcal{R}(u), \varphi \rangle := (\bar{\boldsymbol{K}}\mathcal{D}(u)\nabla u, \nabla\varphi) + \langle f, \varphi \rangle = 0 \qquad \forall \varphi \in H_0^1(\Omega), \tag{1.12}$$

where $\mathcal{D} : \mathbb{R} \to (0, \infty)$ is a non-constant function. In this case, the norms $|||\cdot|||_{\pm 1,\bar{\boldsymbol{K}}\mathcal{D}(u)}$ can no longer be directly computed since the exact solution $u \in H_0^1(\Omega)$ is unknown, and a modification of the approach from the linear case is necessary.

1.5 The pathway on an example

We now stick to the example (1.12) and describe our approach in this model setting.

1.5.1 Iterative linearization

Iterative linearizations are commonly used to approximate the nonlinear problems (1.1) and (1.3). One of the most straightforward is the fixed point (Picard) linearization. For the specific case (1.12), the finite element discretization combined with the Picard linearization reads: let

$u_\ell^i \in V_\ell$ be an approximation of u_ℓ at some linearization iteration $i \geq 0$. We then look for $u_\ell^{i+1} \in V_\ell$ such that

$$(\bar{\boldsymbol{K}}\mathcal{D}(u_\ell^i)\nabla u_\ell^{i+1}, \nabla\varphi_\ell) + \langle f, \varphi_\ell\rangle = 0 \qquad \forall \varphi_\ell \in V_\ell. \tag{1.13}$$

We now observe that u_ℓ^{i+1} is the finite element approximation of the continuous-level solution $u_{\langle\ell\rangle}^{i+1} \in H_0^1(\Omega)$ of the linear problem

$$(\bar{\boldsymbol{K}}\mathcal{D}(u_\ell^i)\nabla u_{\langle\ell\rangle}^{i+1}, \nabla\varphi) + \langle f, \varphi\rangle = 0 \qquad \forall \varphi \in H_0^1(\Omega). \tag{1.14}$$

1.5.2 Iteration-dependent energy norm

Since $u_\ell^i \in V_\ell$ is known, at each iteration step $i \geq 0$, similarly to the linear case of Section 1.4.1, one can introduce the iteration-dependent energy norm $|\!|\!|\cdot|\!|\!|_{1,u_\ell^i}$ and the corresponding dual norm $|\!|\!|\cdot|\!|\!|_{-1,u_\ell^i}$ by

$$\begin{aligned} |\!|\!|\varphi|\!|\!|_{1,u_\ell^i} &:= (\bar{\boldsymbol{K}}\mathcal{D}(u_\ell^i)\nabla\varphi, \nabla\varphi)^{\frac{1}{2}}, && \varphi \in H_0^1(\Omega), \\ |\!|\!|\varsigma|\!|\!|_{-1,u_\ell^i} &:= \sup_{\varphi \in H_0^1(\Omega)} \frac{\langle\varsigma, \varphi\rangle}{|\!|\!|\varphi|\!|\!|_{1,u_\ell^i}}, && \varsigma \in H^{-1}(\Omega). \end{aligned}$$

1.5.3 Residuals

Let the discretization residual corresponding to the $(i+1)$ linearization iteration be given by

$$\langle \mathcal{R}_{\rm disc}^{u_\ell^i}(u_\ell^{i+1}), \varphi\rangle := (\bar{\boldsymbol{K}}\mathcal{D}(u_\ell^i)\nabla u_\ell^{i+1}, \nabla\varphi) + \langle f, \varphi\rangle, \qquad \varphi \in H_0^1(\Omega).$$

Note that from (1.14), $\langle \mathcal{R}_{\rm disc}^{u_\ell^i}(u_\ell^{i+1}), \varphi\rangle$ uniformly vanishes if and only if there is no discretization error in the sense that $u_\ell^{i+1} = u_{\langle\ell\rangle}^{i+1}$. Using (1.12) and (1.14), one observes that, for all $\varphi \in H_0^1(\Omega)$, the total residual writes

$$\langle \mathcal{R}(u_\ell^i), \varphi\rangle = (\bar{\boldsymbol{K}}\mathcal{D}(u_\ell^i)\nabla u_\ell^i, \nabla\varphi) + \langle f, \varphi\rangle = (\bar{\boldsymbol{K}}\mathcal{D}(u_\ell^i)\nabla(u_\ell^i - u_\ell^{i+1}), \nabla\varphi) + \langle \mathcal{R}_{\rm disc}^{u_\ell^i}(u_\ell^{i+1}), \varphi\rangle.$$

1.5.4 Orthogonal error decomposition

Now, using the Galerkin orthogonality from (1.13)–(1.14) reveals the orthogonal decomposition

$$\underbrace{|\!|\!|\mathcal{R}(u_\ell^i)|\!|\!|_{-1,u_\ell^i}^2}_{\substack{\text{total residual/error} \\ |\!|\!|u_\ell^i - u_{\langle\ell\rangle}^{i+1}|\!|\!|_{1,u_\ell^i}}} = \underbrace{|\!|\!|u_\ell^i - u_\ell^{i+1}|\!|\!|_{1,u_\ell^i}^2}_{\substack{\text{linearization} \\ \text{error}}} + \underbrace{\left|\!\left|\!\left|\mathcal{R}_{\rm disc}^{u_\ell^i}(u_\ell^{i+1})\right|\!\right|\!\right|_{-1,u_\ell^i}^2}_{\substack{\text{discretization residual/} \\ \text{error} \\ |\!|\!|u_\ell^{i+1} - u_{\langle\ell\rangle}^{i+1}|\!|\!|_{1,u_\ell^i}}}. \tag{1.15}$$

A detailed proof of this result is given in Section 5.1.3 for the general case. Clearly, the first term on the right-hand-side is the linearization error measured by the difference of two consecutive iterates, whereas the second one represents the discretization error on step $(i+1)$ of the linearization. Consequently, $|\!|\!|\mathcal{R}(u_\ell^i)|\!|\!|_{-1,u_\ell^i} = \left|\!\left|\!\left|u_\ell^i - u_{\langle\ell\rangle}^{i+1}\right|\!\right|\!\right|_{1,u_\ell^i}$ seems as the ideal error metric for the nonlinear problem (1.12), with orthogonal decomposition into the two error components.

1.5.5 A posteriori error estimates

The linearization error in (1.15) is directly computable. In addition, crucially, $\left|\!\left|\!\left|\mathcal{R}_{\rm disc}^{u_\ell^i}(u_\ell^{i+1})\right|\!\right|\!\right|_{-1,u_\ell^i}$ which equals $\left|\!\left|\!\left|u_\ell^{i+1} - u_{\langle\ell\rangle}^{i+1}\right|\!\right|\!\right|_{1,u_\ell^i}$ and stems from a linear (here diffusion and below possibly singularly perturbed reaction–diffusion) problem may be robustly estimated using the energy setting of the linear case of Section 1.4.1. Moreover, the total error $|\!|\!|\mathcal{R}(u_\ell^i)|\!|\!|_{-1,u_\ell^i}$ converges to

$|\!|\!|\mathcal{R}(u_\ell)|\!|\!|_{-1,\bar{\boldsymbol{K}}\mathcal{D}(u_\ell)}$ from Section 1.4.2 if the linearization is converging. This is shown for the general case in Appendix B.

1.5.6 Balancing of error components and stopping criteria

Finally, the orthogonal decomposition (1.15) allows us to naturally balance the two error components and design an adaptive stopping criterion for the linearization iterations. The linearization error component arising from our analysis has also been used in [61] to switch between linearization schemes, L-scheme and the Newton scheme to be more precise, when solving the time-discretized Richards equation.

1.6 Structure of the present contribution

In this manuscript, we consider two rather broad classes of nonlinear elliptic problems of the form (1.1); details are elaborated in Section 3 after we fix the setting in Section 2. We then in Section 4 describe the class of linearization schemes that we consider, including respectively the Kačanov, Newton, Zarantonello and the fixed-point (Picard), L, and M-schemes. Importantly, they all give rise to a symmetric linearization just as in (1.13) (of reaction–diffusion type in general). Our main results are collected in Section 5. First, an orthogonal decomposition result of the form (1.15) is shown in Section 5.1. Then, in Section 5.2, guaranteed and locally efficient a posteriori error estimates not featuring the global Lipschitz/monotonicity ratio $\lambda_{\mathrm{M}}/\lambda_{\mathrm{m}}$ are stated. Finally, an adaptive stopping criterion for iterative linearizations ensuring a natural balance between discretization and linearization components is designed in Section 5.3. In Section 6, we then present numerical results for two test problems and five different linearization schemes. They clearly illustrate the robustness as well as effectiveness of the derived a posteriori estimates. Next, in Section 7 functional analytic aspects of Section 3 are discussed in detail, and, subsequently, the details and the proof of the a posteriori estimates are presented in Section 8. Appendix A summarizes the well-posedness and consistency of the iterative linearizations introduced in Section 4 on an abstract level. The transition of our results from iteration-dependent norms to fixed norms is then discussed in Appendix B. Finally, Appendix C extends our results to the Newton linearization for the gradient-independent diffusivity case, which, in contrast to the above, gives rise to a linear but nonsymmetric (advection–reaction–diffusion) problem on each linearization step. The framework has to be extended but, under appropriate (smallness) assumptions, similar structural results as in the main body of the manuscript are obtained, with convincing numerical illustrations.

2 Basic mathematical tools and notation

Let $L^q(\Omega)$ be the space of functions that are Lebesgue-integrable with power $1 \le q \le \infty$. For $v \in L^q(\Omega)$, we denote the corresponding norm by $\|v\|_{L^q(\Omega)}$ and reduce this notation to $\|v\|$ in the specific case $q = 2$; there, the corresponding scalar product is denoted by (v, w). Similar notation is used for vector-valued functions. On a subdomain $\omega \subseteq \Omega$, we use the notation $\|v\|_\omega$ and $(v, w)_\omega$. We denote by $W^{1,q}(\Omega)$ the Sobolev space of $L^q(\Omega)$ functions having a weak derivative also in $L^q(\Omega)$, with $H^1(\Omega) := W^{1,2}(\Omega)$; $W^{1,q}_0(\Omega)$ and $H^1_0(\Omega)$ are then the trace-free subspaces.

For an open Lipschitz domain $\omega \subseteq \Omega$ with diameter $h_\omega > 0$, we will repeatedly use the Poincaré inequality: for functions $u \in H^1(\omega)$ vanishing in a trace sense in a fixed relatively open subset of $\partial\omega$ with non-zero measure, or for functions $u \in H^1(\omega)$ satisfying $\int_\omega u = 0$ (zero-mean-value property), there holds for a constant $C_\omega > 0$

$$\|u\|_\omega \le C_\omega h_\omega \|\nabla u\|_\omega. \tag{2.1}$$

If ω is convex, then $C_\omega \leq \pi^{-1}$ in the zero-mean-value case. The discussion of the other cases can be found in, e.g., [6, Section 2.1] and the references therein. We will further use the Young inequality which states that for $\rho \in \mathbb{R}$, $\rho > 0$, and a, $b \in \mathbb{R}$,

$$ab \leq \frac{1}{2\rho}a^2 + \frac{\rho}{2}b^2. \tag{2.2}$$

3 Two classes of nonlinear elliptic problems

For our analysis, we divide elliptic problems represented by (1.1) into two different classes.

3.1 Gradient-dependent diffusivity

In this case, the action of the nonlinear operator $\mathcal{R}$ applied on a function $\xi \in H_0^1(\Omega)$ is prescribed by

$$\langle \mathcal{R}(\xi), \varphi \rangle := (\boldsymbol{\sigma}(\cdot, \nabla\xi), \nabla\varphi) + \langle f(\cdot, \xi), \varphi \rangle, \qquad \varphi \in H_0^1(\Omega). \tag{3.1}$$

We suppose the following to ensure the monotonicity and Lipschitz continuity of the operator $\mathcal{R}$ in the sense of (1.2):

Assumption 3.1 (Gradient-dependent diffusivity)**.**

(A1) The flux function $\boldsymbol{\sigma} : \Omega \times \mathbb{R}^d \mapsto \mathbb{R}^d$ is bounded with respect to its first argument $\boldsymbol{x} \in \Omega$ for a fixed second argument $\boldsymbol{y} \in \mathbb{R}^d$. Moreover, it satisfies for real constants $\sigma_{\mathrm{M}} \geq \sigma_{\mathrm{m}} > 0$ the Lipschitz-continuity condition

$$|\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{z})| \leq \sigma_{\mathrm{M}} |\boldsymbol{y} - \boldsymbol{z}| \qquad \textit{for a.e. } \boldsymbol{x} \in \Omega \textit{ and all } \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^d$$

and the monotonicity condition

$$(\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{z})) \cdot (\boldsymbol{y} - \boldsymbol{z}) \geq \sigma_{\mathrm{m}} |\boldsymbol{y} - \boldsymbol{z}|^2 \qquad \textit{for a.e. } \boldsymbol{x} \in \Omega \textit{ and all } \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^d.$$

(A2) The reaction/source function $f : \Omega \times \mathbb{R} \to \mathbb{R}$ is bounded with respect to its first argument $\boldsymbol{x} \in \Omega$ for a fixed second argument $\xi \in \mathbb{R}$. It is Lipschitz continuous and increasing with respect to its second argument $\xi \in \mathbb{R}$, i.e., there exists a constant $f_{\mathrm{M}} > 0$ such that

$$0 \leq f(\boldsymbol{x}, \xi_2) - f(\boldsymbol{x}, \xi_1) \leq f_{\mathrm{M}} (\xi_2 - \xi_1) \qquad \textit{for a.e. } \boldsymbol{x} \in \Omega \textit{ and all } \xi_2 \geq \xi_1.$$

Under these conditions, there exists a unique solution $u \in H_0^1(\Omega)$ of problem (1.1) with (3.1). Moreover, the operator $\mathcal{R}$ is monotone and Lipschitz continuous, as stated in (1.2), with the simple distance metric

$$\mathrm{dist}(u_\ell, u) = \|\nabla(u_\ell - u)\|.$$

The proof of the statements above is postponed to Proposition 7.1, since it uses some well-known arguments and it is not the main focus of the paper.

Example 3.2 (Systems modeled by (3.1))**.** *Systems having the flux function $\boldsymbol{\sigma}$ of the form*

$$\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y}) = A(|\boldsymbol{y}|)\boldsymbol{y}, \tag{3.2}$$

where $A : [0, \infty) \to (0, \infty)$, satisfy (A1) if and only if $0 < \sigma_{\mathrm{m}} \leq (A(t)\, t)' \leq \sigma_{\mathrm{M}} < \infty$ for almost all $t \geq 0$, see, e.g., [33, Appendix A] and the references therein. In this case

$$|A(|\boldsymbol{y}|)\boldsymbol{y} - A(|\boldsymbol{z}|)\boldsymbol{z}| \leq \sigma_{\mathrm{M}} |\boldsymbol{y} - \boldsymbol{z}| \quad \textit{and} \quad (A(|\boldsymbol{y}|)\boldsymbol{y} - A(|\boldsymbol{z}|)\boldsymbol{z}) \cdot (\boldsymbol{y} - \boldsymbol{z}) \geq \sigma_{\mathrm{m}} |\boldsymbol{y} - \boldsymbol{z}|^2.$$

A common example of this type of system is the mean-curvature flow where

$$A(\varrho) = \sigma_{\mathrm{m}} + (\sigma_{\mathrm{M}} - \sigma_{\mathrm{m}})/\sqrt{1+\varrho^2}$$

for the two constants $0 < \sigma_{\mathrm{m}} \leq \sigma_{\mathrm{M}} < \infty$. Numerical results are presented for this system in Section 6.1. Examples of other problems close to this category are p-Laplace equations and compressive and non-Newtonian flow equations. However, these satisfy (A1) only under certain additional assumptions on boundedness.

3.2 Gradient-independent diffusivity

The second class of problems that we consider foregoes the nonlinear dependence on the gradient. In this case, we assume the following form of $\mathcal{R}$: for $\xi \in H_0^1(\Omega)$,

$$\langle \mathcal{R}(\xi), \varphi \rangle := \tau(\bar{\boldsymbol{K}}(\cdot)(\mathcal{D}(\cdot,\xi)\nabla\xi + \boldsymbol{q}(\cdot,\xi)), \nabla\varphi) + \langle f(\cdot,\xi), \varphi\rangle, \qquad \varphi \in H_0^1(\Omega). \tag{3.3}$$

We suppose that the auxiliary functions satisfy the following general conditions

Assumption 3.3 (Gradient-independent diffusivity)**.**

(B1) The diffusion coefficient $\mathcal{D} : \Omega \times \mathbb{R} \to (0,\infty)$ is bounded and Lipschitz-continuous with respect to all its arguments, i.e., there exists a constant $\mathcal{D}_{\mathrm{M}} > 0$ such that $0 < \mathcal{D} \leq \mathcal{D}_{\mathrm{M}}$, and

$$|\mathcal{D}(\boldsymbol{x}_1,\xi_1) - \mathcal{D}(\boldsymbol{x}_2,\xi_2)| \leq \mathcal{D}_{\mathrm{M}}(|\boldsymbol{x}_1 - \boldsymbol{x}_2| + |\xi_1 - \xi_2|) \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \Omega \text{ and } \xi_1, \xi_2 \in \mathbb{R}.$$

(B2) f satisfies (A2).

(B3) The permeability tensor $\bar{\boldsymbol{K}} : \Omega \to \mathbb{R}^{d\times d}$ is uniformly symmetric positive definite and bounded, i.e., there exist real constants $K_{\mathrm{M}} \geq K_{\mathrm{m}} > 0$ such that for almost all $\boldsymbol{x} \in \Omega$,

$$K_{\mathrm{m}}|\boldsymbol{y}|^2 \leq \boldsymbol{y}^{\mathrm{T}}\bar{\boldsymbol{K}}(\boldsymbol{x})\boldsymbol{y} \leq K_{\mathrm{M}}|\boldsymbol{y}|^2 \qquad \forall \boldsymbol{y} \in \mathbb{R}^d.$$

(B4) The advective flux $\boldsymbol{q} : \Omega \times \mathbb{R} \to \mathbb{R}^d$ is bounded by a constant $q_{\mathrm{M}} > 0$, i.e., $|\boldsymbol{q}| < q_{\mathrm{M}}$. Moreover, it is small with respect the diffusion coefficient $\mathcal{D}$ in that for a constant $\gamma > 0$, the following inequality is satisfied for all $\xi, \zeta \in \mathbb{R}$ and almost all $\boldsymbol{x} \in \Omega$,

$$\left|\bar{\boldsymbol{K}}^{\frac{1}{2}}(\boldsymbol{x})\left(\boldsymbol{q}(\boldsymbol{x},\zeta) - \boldsymbol{q}(\boldsymbol{x},\xi) - \sum_{j=1}^{d}\boldsymbol{e}_j\int_{\xi}^{\zeta}\partial_{x_j}\mathcal{D}(\boldsymbol{x},\varrho)\,\mathrm{d}\varrho\right)\right|^2 \leq \gamma(f(\boldsymbol{x},\zeta) - f(\boldsymbol{x},\xi))\int_{\xi}^{\zeta}\mathcal{D}(\boldsymbol{x},\varrho)\,\mathrm{d}\varrho.$$

(B5) τ is a positive real parameter.

The existence and uniqueness of a solution $u \in L^\infty(\Omega) \cap H_0^1(\Omega)$ of the model (1.1) with (3.3) under Assumption 3.3 is shown in Proposition 7.2. These assumptions further ensure that the operator $\mathcal{R}$ is monotone and Lipschitz continuous in the sense of (1.2) with a more complex distance metric

$$\mathrm{dist}(u_\ell, u) = \left\|\bar{\boldsymbol{K}}^{\frac{1}{2}}(\cdot)\nabla\int_{u}^{u_\ell}\mathcal{D}(\cdot,\varrho)\,\mathrm{d}\varrho\right\|.$$

Example 3.4 (Systems modeled by (3.3))**.** *We mention two important classes of equations that are modeled by (3.3).*

*Class I: **Semilinear equations**, corresponding to*

$$\mathcal{D} = 1, \quad \bar{\boldsymbol{K}} = \mathbb{I}, \quad \tau = 1, \text{ and } \boldsymbol{q}(\boldsymbol{x},\xi) = \boldsymbol{0}. \tag{3.4}$$

(These would of course also fit (3.1) *with $\boldsymbol{\sigma}(\boldsymbol{x}, \nabla\xi) = \nabla\xi$). They are used in the modelling of ignition of gases, gravitational influences on stars, in quantum field theory, and in many other applications (see, e.g., [49] and the references therein).*

Class II: elliptic problems arising from ***implicit time-discretization of nonlinear advection–reaction–diffusion equations****, with $\tau > 0$ being the corresponding time-step [52]. Examples are the Fischer–KPP equation, equations used to model flow through porous media (the porous medium equation, the Richards equation), and biofilms [48].*

In all cases, Assumption 3.3 is satisfied either directly (semilinear equations), or under some conditions on the auxiliary functions (time-discretized systems). A specific example of the Richards equation is discussed in more detail in Section 6.2.

Remark 3.5 (Degeneracy)**.** *Conditions (B1)–(B5) allow for degeneracy, i.e., they allow that $\mathcal{D}(\boldsymbol{x}, \xi) \searrow 0$ as $\xi \to \pm\infty$, leading to the loss of ellipticity of the problem* (3.3)*, see also Proposition 7.2. This is not covered in the gradient-dependent diffusivity case* (3.1)*.*

4 Finite element discretization and iterative linearization

To obtain a numerical approximation of $u \in H_0^1(\Omega)$ in (1.1), one usually first discretizes the system in space and then employs an iterative linearization scheme to solve the resulting system of nonlinear algebraic equations. We now describe this procedure in detail.

4.1 Finite element discretization

Let $\mathcal{T}_\ell$ denote a triangulation of Ω with matching and uniformly shape-regular simplicial meshes. Let $\mathcal{V}_\ell$ be the corresponding set of vertices. For each element $K \in \mathcal{T}_\ell$, let $h_K := \mathrm{diam}\{K\}$ denote the diameter of K. Let

$$V_\ell := H_0^1(\Omega) \cap \mathcal{P}_p(\mathcal{T}_\ell), \tag{4.1}$$

where $\mathcal{P}_p(\mathcal{T}_\ell)$ denotes the space of piecewise polynomials of degree $p \geq 1$ for the triangulation $\mathcal{T}_\ell$. We note that, for all $q \geq 2$,

$$V_\ell \subset W_0^{1,\infty}(\Omega) \subset L^\infty(\Omega) \cap W_0^{1,q}(\Omega) \subset L^\infty(\Omega) \cap H_0^1(\Omega), \tag{4.2}$$

as V_ℓ is finite dimensional. Then, $u_\ell \in V_\ell$ is the finite element approximation to the nonlinear problem (1.1) if it satisfies

$$\langle \mathcal{R}(u_\ell), \varphi_\ell \rangle = 0 \qquad \forall \varphi_\ell \in V_\ell. \tag{4.3}$$

To solve the system of nonlinear algebraic equations arising from (4.3), iterative linearization schemes are commonly used, where a sequence $\{u_\ell^i\}_{i\geq 1} \in V_\ell$ is constructed iteratively from an initial guess $u_\ell^0 \in V_\ell$. It yields u_ℓ in the limit when the linearization converges. Below, we give an abstract definition of such a linearization scheme.

4.2 Iteration-dependent scalar product

For a linearization iteration index $i \geq 0$, let $u_\ell^i \in V_\ell$ be given. We define

$$(\!(\xi, \zeta)\!)_{u_\ell^i} := (L^i \xi, \zeta) + (\mathfrak{a}^i \nabla\xi, \nabla\zeta), \qquad \xi, \zeta \in H_0^1(\Omega) \tag{4.4}$$

and assume:

Assumption 4.1 (Coefficients L^i and $\mathfrak{a}^i$)**.** *The function $L^i : \Omega \to \mathbb{R}$ and the symmetric tensor $\mathfrak{a}^i : \Omega \to \mathbb{R}^d \times \mathbb{R}^d$ are defined from the data of* (3.1) *or* (3.3) *and u_ℓ^i and satisfy for functions*

$0 \leq L_{\mathrm{m}}^i \leq L_{\mathrm{M}}^i < \infty$ *and* $0 < a_{\mathrm{m}}^i \leq a_{\mathrm{M}}^i < \infty$*, piecewise constant with respect to mesh* $\mathcal{T}_\ell$ *and sharpest possible, that*

$$L_{\mathrm{m}}^i|_K \leq L^i(\boldsymbol{x}) \leq L_{\mathrm{M}}^i|_K, \qquad \text{for a.e. } \boldsymbol{x} \in K, \forall K \in \mathcal{T}_\ell, \tag{4.5a}$$

$$a_{\mathrm{m}}^i|_K |\boldsymbol{y}|^2 \leq \boldsymbol{y}^{\mathrm{T}}\, \mathfrak{a}^i(\boldsymbol{x})\, \boldsymbol{y} \leq a_{\mathrm{M}}^i|_K |\boldsymbol{y}|^2, \qquad \forall \boldsymbol{y} \in \mathbb{R}^d, \ \text{for a.e. } \boldsymbol{x} \in K, \forall K \in \mathcal{T}_\ell. \tag{4.5b}$$

Consequently,

$$(\!(\xi, \zeta)\!)_{u_\ell^i} \text{ is a scalar product over } H_0^1(\Omega) \text{ equivalent to } (\xi,\zeta) + (\nabla\xi, \nabla\zeta) \text{ and } (\nabla\xi, \nabla\zeta). \tag{4.6}$$

Remark 4.2 (Iteration-dependent scalar product)**.** *The scalar product* $(\!(\xi, \zeta)\!)_{u_\ell^i}$ *from* (4.4) *is of reaction–diffusion type when* $L^i \neq 0$ *and of pure diffusion type when* $L^i = 0$*. The first setting in general only appears when the function* $f(\boldsymbol{x}, \xi)$ *in* (3.1) *or* (3.3) *depends (linearly or nonlinearly) on the second argument* ξ*, i.e., when it represents a (linear or nonlinear) reaction term. When it only depends on the space coordinate* $\boldsymbol{x}$*, forming a source term* $f(\boldsymbol{x})$*, typically* $L^i = 0$*, see Tables* 1 *and* 2 *below. The scalar product* $(\!(\xi, \zeta)\!)_{u_\ell^i}$ *is always of pure diffusion type for the Zarantonello linearization of Table* 1*, where* $L^i = 0$ *by definition.*

We will now use this scalar product to define the iterative linearization.

4.3 Iterative linearization

Let $u_\ell^0 \in V_\ell$ be given. For $i \geq 0$, let $u_\ell^i \in V_\ell$ denote the i^{th} iterate. Then, the $(i+1)^{\text{th}}$ iterate $u_\ell^{i+1} \in V_\ell$ is computed by solving the linear reaction–diffusion or diffusion problem (see Remark 4.2), with the difference of the iterates (increment) in the scalar product (4.4) on the left-hand side and the residual (1.1) on the right-hand side,

$$(\!(u_\ell^{i+1} - u_\ell^i, \varphi_\ell)\!)_{u_\ell^i} = -\langle \mathcal{R}(u_\ell^i), \varphi_\ell \rangle \qquad \forall \varphi_\ell \in V_\ell. \tag{4.7}$$

Since, recalling (4.2), $u_\ell^i \in H_0^1(\Omega)$, problem (4.7) is well posed by the Riesz representation theorem in view of Assumption 4.1 and (4.6). Examples of $(\!(\cdot, \cdot)\!)_{u_\ell^i}$ corresponding to the most commonly used linearization schemes are given in Section 4.6, whereas the consistency of the linearization iterations is discussed in a general functional analytic setting in Appendix A.

When expanded using (4.4), problem (4.7) writes: find $u_\ell^{i+1} \in V_\ell$ such that

$$\begin{aligned} (L^i\, u_\ell^{i+1}, \varphi_\ell) + (\mathfrak{a}^i \nabla u_\ell^{i+1}, \nabla\varphi_\ell) = &- \langle \mathcal{R}(u_\ell^i), \varphi_\ell \rangle + (L^i\, u_\ell^i, \varphi_\ell) \\ &+ (\mathfrak{a}^i \nabla u_\ell^i, \nabla\varphi_\ell) \qquad \forall \varphi_\ell \in V_\ell. \end{aligned} \tag{4.8}$$

Furthermore, we can always identify a scalar-valued source term $\mathcal{S}^i \in L^2(\Omega)$ and a vector-valued flux term $\boldsymbol{\mathcal{F}}^i \in \boldsymbol{L}^2(\Omega; \mathbb{R}^d)$, so that (4.8) further rewrites as: find $u_\ell^{i+1} \in V_\ell$ such that

$$(L^i\, u_\ell^{i+1}, \varphi_\ell) + (\mathfrak{a}^i \nabla u_\ell^{i+1}, \nabla\varphi_\ell) = -(\mathcal{S}^i, \varphi_\ell) - (\boldsymbol{\mathcal{F}}^i, \nabla\varphi_\ell) \qquad \forall \varphi_\ell \in V_\ell. \tag{4.9}$$

Indeed, there holds, for all $\varphi \in H_0^1(\Omega)$,

$$\langle \mathcal{R}(u_\ell^i), \varphi \rangle - (L^i\, u_\ell^i, \varphi) - (\mathfrak{a}^i \nabla u_\ell^i, \nabla\varphi) = (\mathcal{S}^i, \varphi) + (\boldsymbol{\mathcal{F}}^i, \nabla\varphi), \tag{4.10}$$

where, in the gradient-dependent case (3.1),

$$\mathcal{S}^i := f(\cdot, u_\ell^i) - L^i\, u_\ell^i, \tag{4.11a}$$

$$\boldsymbol{\mathcal{F}}^i := \boldsymbol{\sigma}(\cdot, \nabla u_\ell^i) - \mathfrak{a}^i \nabla u_\ell^i, \tag{4.11b}$$

whereas in the gradient-independent case (3.3),

$$\mathcal{S}^i := f(\cdot, u_\ell^i) - L^i\, u_\ell^i \tag{4.12a}$$

$$\boldsymbol{\mathcal{F}}^i := \tau \bar{\boldsymbol{K}}(\cdot)\, \boldsymbol{q}(\cdot, u_\ell^i) + (\tau \bar{\boldsymbol{K}}(\cdot) \mathcal{D}(\cdot, u_\ell^i) - \mathfrak{a}^i) \nabla u_\ell^i. \tag{4.12b}$$

4.4 Continuous-level counterpart of (4.7)

Only the discrete problems (4.7) need to be resolved in practice. For our theoretical developments below, though, the following continuous-level counterpart of (4.7) will be useful: we define $u_{\langle\ell\rangle}^{i+1} \in H_0^1(\Omega)$ to be the solution of

$$(\!(u_{\langle\ell\rangle}^{i+1} - u_\ell^i, \varphi)\!)_{u_\ell^i} = -\langle \mathcal{R}(u_\ell^i), \varphi\rangle \qquad \forall \varphi \in H_0^1(\Omega). \tag{4.13}$$

This problem is again well posed by the Riesz representation theorem. Now u_ℓ^{i+1} from (4.7) is the finite element approximation of $u_{\langle\ell\rangle}^{i+1}$ from (4.13). As for (4.9), (4.13) can be equivalently written as the linear (reaction–)diffusion problem of finding $u_{\langle\ell\rangle}^{i+1} \in H_0^1(\Omega)$ such that

$$(L^i\, u_{\langle\ell\rangle}^{i+1}, \varphi) + (\mathfrak{a}^i \nabla u_{\langle\ell\rangle}^{i+1}, \nabla\varphi) = -(\mathcal{S}^i, \varphi) - (\boldsymbol{\mathcal{F}}^i, \nabla\varphi) \qquad \forall \varphi \in H_0^1(\Omega). \tag{4.14}$$

4.5 Iteration-dependent norm

In our developments, we crucially rely on the *iteration-dependent norm* associated to the scalar product (4.4) from the iterative linearization (4.7), i.e., given as

$$|\!|\!|\xi|\!|\!|_{1,u_\ell^i} := (\!(\xi, \xi)\!)_{u_\ell^i}^{\frac{1}{2}}, \qquad \xi \in H_0^1(\Omega). \tag{4.15}$$

The Cauchy–Schwarz inequality then reads, for any $\xi, \zeta \in H_0^1(\Omega)$,

$$(\!(\xi, \zeta)\!)_{u_\ell^i} \le |\!|\!|\xi|\!|\!|_{1,u_\ell^i} |\!|\!|\zeta|\!|\!|_{1,u_\ell^i}, \tag{4.16}$$

and we have, for any $\xi \in H_0^1(\Omega)$, the duality expression

$$|\!|\!|\xi|\!|\!|_{1,u_\ell^i} = \sup_{\varphi \in H_0^1(\Omega)} \frac{(\!(\xi, \varphi)\!)_{u_\ell^i}}{|\!|\!|\varphi|\!|\!|_{1,u_\ell^i}}, \tag{4.17}$$

with the equality occurring when $\varphi = \xi / |\!|\!|\xi|\!|\!|_{1,u_\ell^i}$. We will also employ the corresponding dual norm

$$|\!|\!|\varsigma|\!|\!|_{-1,u_\ell^i} := \sup_{\varphi \in H_0^1(\Omega)} \frac{\langle \varsigma, \varphi\rangle}{|\!|\!|\varphi|\!|\!|_{1,u_\ell^i}}, \quad \varsigma \in H^{-1}(\Omega). \tag{4.18}$$

Finally, for an open Lipschitz subdomain $\omega \subseteq \Omega$, we also introduce the local counterparts of $|\!|\!|\cdot|\!|\!|_{\pm 1,u_\ell^i}$ for $\xi \in H_0^1(\Omega)$ and $\varsigma \in H^{-1}(\Omega)$ as

$$|\!|\!|\xi|\!|\!|_{1,u_\ell^i,\omega} := \left((L^i\,\xi, \xi)_\omega + (\mathfrak{a}^i \nabla\xi, \nabla\xi)_\omega\right)^{\frac{1}{2}}, \tag{4.19a}$$

$$|\!|\!|\varsigma|\!|\!|_{-1,u_\ell^i,\omega} := \sup_{\varphi \in H_0^1(\omega)} \frac{\langle \varsigma, \varphi\rangle}{|\!|\!|\varphi|\!|\!|_{1,u_\ell^i,\omega}}. \tag{4.19b}$$

4.6 Examples of common linearization schemes

We now give examples of some common linearization schemes defining the iteration-dependent scalar product (4.4) and consequently the iterates (4.7).

Scheme	$L^i(\boldsymbol{x})$	$\mathfrak{a}^i(\boldsymbol{x})$
Kačanov [68, Chapter 25]	$\partial_\xi f(\boldsymbol{x}, u_\ell^i)$	$A(\lvert\nabla u_\ell^i\rvert)$
Newton [22]	$\partial_\xi f(\boldsymbol{x}, u_\ell^i)$	$A(\lvert\nabla u_\ell^i\rvert) + \frac{A'(\lvert\nabla u_\ell^i\rvert)}{\lvert\nabla u_\ell^i\rvert} \nabla u_\ell^i \otimes \nabla u_\ell^i$
Zarantonello [66]	0	Λ (constant) > 0

Table 1: Some common linearizations to solve problem (1.1) with the gradient-dependent expression (3.1) for $\mathcal{R}$. For the Kačanov and Newton iterations, $\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y}) = A(|\boldsymbol{y}|)\boldsymbol{y}$ is assumed, as in Example 3.2.

4.6.1 Gradient-dependent diffusivity

For the gradient-dependent diffusivity case (3.1), the three most standard linearization schemes are presented in Table 1. The Kačanov (fixed point) iterations are considered in Chapter 25 of [68], and the convergence of the scheme is proven under certain restrictions on $\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y})$. The Newton [22] linearization adds a first-order correction and converges quadratically for a sufficiently precise initial guess. The Zarantonello [66] scheme converges linearly irrespective of the initial guess, for a constant $\Lambda \geq \sigma_{\mathrm{M}}^2/\sigma_{\mathrm{m}}$, where σ_{m} and σ_{M} stem from (A1). We refer to [33,36] and the references therein for additional properties.

From (4.11) and Table 1, one in particular obtains that the linearizations (4.7) read here as (4.9) with

$$\mathcal{S}^i = f(\cdot, u_\ell^i) - \partial_\xi f(\cdot, u_\ell^i)\, u_\ell^i \quad \text{and} \quad \boldsymbol{\mathcal{F}}^i = 0 \qquad \text{[Kačanov]}, \tag{4.20a}$$

$$\mathcal{S}^i = f(\cdot, u_\ell^i) - \partial_\xi f(\cdot, u_\ell^i)\, u_\ell^i \quad \text{and} \quad \boldsymbol{\mathcal{F}}^i = -A'(|\nabla u_\ell^i|)|\nabla u_\ell^i|\nabla u_\ell^i \qquad \text{[Newton]}, \tag{4.20b}$$

$$\mathcal{S}^i = f(\cdot, u_\ell^i) \quad \text{and} \quad \boldsymbol{\mathcal{F}}^i = \boldsymbol{\sigma}(\cdot, \nabla u_\ell^i) - \Lambda \nabla u_\ell^i \qquad \text{[Zarantonello]}. \tag{4.20c}$$

4.6.2 Gradient-independent diffusivity

For the gradient-independent diffusivity case (3.3), the commonly used linearization schemes are collected in Table 2. The four schemes presented are all linearly converging under some restrictions on the parameter $\tau > 0$.

Scheme	$L^i(\boldsymbol{x})$	$\mathfrak{a}^i(\boldsymbol{x})$
Picard [13]	$\partial_\xi f(\boldsymbol{x}, u_\ell^i)$	$\tau \bar{\boldsymbol{K}}(\boldsymbol{x})\, \mathcal{D}(\boldsymbol{x}, u_\ell^i)$
Jäger–Kačur [39]	$\max_{\xi\in\mathbb{R}} \left(\frac{f(\boldsymbol{x},\xi) - f(\boldsymbol{x},u_\ell^i)}{\xi - u_\ell^i}\right)$	$\tau \bar{\boldsymbol{K}}(\boldsymbol{x})\, \mathcal{D}(\boldsymbol{x}, u_\ell^i)$
L-scheme [50, 55]	L (constant) $\geq \frac{1}{2} \sup \partial_\xi f$	$\tau \bar{\boldsymbol{K}}(\boldsymbol{x})\, \mathcal{D}(\boldsymbol{x}, u_\ell^i)$
M-scheme [52]	$\partial_\xi f(\boldsymbol{x}, u_\ell^i) + M\tau$ (constant)	$\tau \bar{\boldsymbol{K}}(\boldsymbol{x})\, \mathcal{D}(\boldsymbol{x}, u_\ell^i)$

Table 2: Some common linearization techniques to solve the problem (1.1) with the gradient-independent expression (3.3) for $\mathcal{R}$.

Remark 4.3 (Newton(–Raphson) linearization)**.** *For the case* (3.3)*, the Newton(–Raphson) iterative scheme does not yield a symmetric linearization problem in general. We will discuss it separately in Appendix C.*

From (4.12) and Table 2, one in particular obtains that the linearizations (4.7) read here as (4.9) with

$$\mathcal{S}^i = f(\cdot, u_\ell^i) - L^i\, u_\ell^i \quad \text{and} \quad \boldsymbol{\mathcal{F}}^i = \tau \bar{\boldsymbol{K}}(\cdot)\boldsymbol{q}(\cdot, u_\ell^i). \tag{4.21}$$

Observe that, for all the schemes in Table 2, the second term cancels out in (4.12b).

5 Main results

In this section, we show that there exists an orthogonality relation between the total error in the finite element solution u_ℓ^i, the linearization error, and the discretization error. The iteration-dependent norm defined in Section 4.5 is crucial for this. We further state how this result allows us to estimate the total error using *guaranteed* (binding above), *efficient* (binding below, up to a constant), and *fully computable* a posteriori estimators *robust* with respect to the strength of the nonlinearities in (3.1) or (3.3). Congruently, the distinction of the error components allows us to balance them and to stop the iterative linearizations timely.

5.1 An orthogonal decomposition of the error

We start by the orthogonal decomposition of the total error in the finite element approximation u_ℓ^i.

5.1.1 Discretization residual and discretization error

Stemming from the linearized problems (4.7) and (4.13), we define the *discretization residual* $\mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(\xi) \in H^{-1}(\Omega)$, for $\xi \in H_0^1(\Omega)$, as

$$\langle \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(\xi), \varphi \rangle := (\!(\xi - u_\ell^i, \varphi)\!)_{u_\ell^i} + \langle \mathcal{R}(u_\ell^i), \varphi \rangle, \qquad \varphi \in H_0^1(\Omega). \tag{5.1}$$

In particular, using the definition (4.4) and (4.10), with the notations (4.11)–(4.12),

$$\begin{aligned} \langle \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}), \varphi \rangle &= (\!(u_\ell^{i+1} - u_\ell^i, \varphi)\!)_{u_\ell^i} + \langle \mathcal{R}(u_\ell^i), \varphi \rangle \\ &= (L^i\, u_\ell^{i+1}, \varphi) + (\mathfrak{a}^i \nabla u_\ell^{i+1} + \boldsymbol{\mathcal{F}}^i, \nabla \varphi) + (\mathcal{S}^i, \varphi). \end{aligned} \tag{5.2}$$

Note that, from (4.13), there holds

$$\langle \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_{\langle \ell \rangle}^{i+1}), \varphi \rangle = 0 \qquad \forall \varphi \in H_0^1(\Omega), \tag{5.3}$$

which justifies the name "residual": when the finite-dimensional space V_ℓ approaches $H_0^1(\Omega)$, $\mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1})$ vanishes. More precisely, since $u_\ell^{i+1} - u_{\langle \ell \rangle}^{i+1}$ is the difference between the discrete and the continuous level solutions from (4.7)–(4.13), using the following result, we find that $\left|\!\left|\!\left| \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}) \right|\!\right|\!\right|_{-1,u_\ell^i}$ represents the *discretization error* of the linearization step $i+1$:

Lemma 5.1 (Discretization residual–discretization error relation)**.** *There holds*

$$\left|\!\left|\!\left| \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}) \right|\!\right|\!\right|_{-1,u_\ell^i} = |\!|\!| u_\ell^{i+1} - u_{\langle \ell \rangle}^{i+1} |\!|\!|_{1,u_\ell^i}.$$

Proof. From (5.1) and (5.3), we have that

$$\langle \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}), \varphi \rangle \overset{(5.3)}{=} \langle \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}) - \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_{\langle \ell \rangle}^{i+1}), \varphi \rangle \overset{(5.1)}{=} (\!(u_\ell^{i+1} - u_{\langle \ell \rangle}^{i+1}, \varphi)\!)_{u_\ell^i} \qquad \forall \varphi \in H_0^1(\Omega).$$

Thus, the result follows from (4.17)–(4.18). □

5.1.2 Total residual and total error

In order to characterize the total error in $u_\ell^i \in V_\ell$, we need to consider the total residual $\mathcal{R}(u_\ell^i)$ stemming from (1.1); as motivated in Section 1, we consider $|||\mathcal{R}(u_\ell^i)|||_{-1,u_\ell^i}$ as the total error measure. Observe from (4.18) that if $u \in H_0^1(\Omega)$ solves (1.1), then

$$|||\mathcal{R}(u_\ell^i)|||_{-1,u_\ell^i} = \sup_{\varphi \in H_0^1(\Omega)} \frac{\langle \mathcal{R}(u_\ell^i), \varphi \rangle}{|||\varphi|||_{1,u_\ell^i}} = \sup_{\varphi \in H_0^1(\Omega)} \frac{\langle \mathcal{R}(u_\ell^i) - \mathcal{R}(u), \varphi \rangle}{|||\varphi|||_{1,u_\ell^i}}, \tag{5.4}$$

which is, due to (1.2) and (4.6), equivalent to the error metric $\operatorname{dist}(u_\ell^i, u)$. Additionally, using

$$\langle \mathcal{R}(u_\ell^i), \varphi \rangle \overset{(4.13)}{=} ((u_\ell^i - u_{\langle \ell \rangle}^{i+1}, \varphi))_{u_\ell^i} \qquad \forall \varphi \in H_0^1(\Omega), \tag{5.5}$$

and employing (4.17)–(4.18), one has:

Lemma 5.2 (Total residual–total error relation)**.** *There holds*

$$|||\mathcal{R}(u_\ell^i)|||_{-1,u_\ell^i} = \left|\left|\left| u_\ell^i - u_{\langle \ell \rangle}^{i+1} \right|\right|\right|_{1,u_\ell^i}.$$

5.1.3 An orthogonal decomposition of the total residual/error

The remaining component of the total residual/error is the linearization error; it comes naturally out of the following decomposition result:

Theorem 5.3 (Orthogonal decomposition of the total residual/error into linearization and discretization components)**.** *Let the nonlinear elliptic operator $\mathcal{R} : H_0^1(\Omega) \to H^{-1}(\Omega)$ be introduced in (1.1) and satisfy (1.2). Let the finite element linearization sequence $\{u_\ell^i\}_{i \geq 0} \subset V_\ell$ be given by (4.7), for the iteration-dependent scalar product (4.4) stemming from the given linearization and satisfying Assumption 4.1. Let the affine elliptic operator $\mathcal{R}_{\mathrm{disc}}^{u_\ell^i} : H_0^1(\Omega) \to H^{-1}(\Omega)$ be given by (5.1), stemming from (4.13). Let finally the norms $|||\cdot|||_{\pm 1,u_\ell^i}$ be given in (4.15), (4.18). Then, for all linearization steps $i \geq 0$,*

$$\underbrace{|||\mathcal{R}(u_\ell^i)|||_{-1,u_\ell^i}^2}_{\substack{\textit{total residual/error} \\ |||u_\ell^i - u_{\langle \ell \rangle}^{i+1}|||_{1,u_\ell^i}}} = \underbrace{|||u_\ell^i - u_\ell^{i+1}|||_{1,u_\ell^i}^2}_{\substack{\textit{linearization} \\ \textit{error}}} + \underbrace{\left|\left|\left| \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}) \right|\right|\right|_{-1,u_\ell^i}^2}_{\substack{\textit{discretization residual/} \\ \textit{error} \\ |||u_\ell^{i+1} - u_{\langle \ell \rangle}^{i+1}|||_{1,u_\ell^i}}}.$$

Proof. One has

$$\begin{aligned}
|||\mathcal{R}(u_\ell^i)|||_{-1,u_\ell^i}^2 &\overset{\text{Lemma } 5.2}{=} \left|\left|\left| u_\ell^i - u_{\langle \ell \rangle}^{i+1} \right|\right|\right|_{1,u_\ell^i}^2 = \left|\left|\left| (u_\ell^i - u_\ell^{i+1}) + (u_\ell^{i+1} - u_{\langle \ell \rangle}^{i+1}) \right|\right|\right|_{1,u_\ell^i}^2 \\
&\overset{(4.15)}{=} |||u_\ell^i - u_\ell^{i+1}|||_{1,u_\ell^i}^2 + \left|\left|\left| u_\ell^{i+1} - u_{\langle \ell \rangle}^{i+1} \right|\right|\right|_{1,u_\ell^i}^2 + 2((u_\ell^i - u_\ell^{i+1}, u_\ell^{i+1} - u_{\langle \ell \rangle}^{i+1}))_{u_\ell^i} \\
&\overset{(4.7),(4.13)}{=} |||u_\ell^i - u_\ell^{i+1}|||_{1,u_\ell^i}^2 + \left|\left|\left| u_\ell^{i+1} - u_{\langle \ell \rangle}^{i+1} \right|\right|\right|_{1,u_\ell^i}^2 \\
&\overset{\text{Lemma } 5.1}{=} |||u_\ell^i - u_\ell^{i+1}|||_{1,u_\ell^i}^2 + \left|\left|\left| \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}) \right|\right|\right|_{-1,u_\ell^i}^2,
\end{aligned}$$

where the penultimate equality follows from the Galerkin orthogonality between u_ℓ^{i+1} and $u_{\langle \ell \rangle}^{i+1}$, using that $u_\ell^i - u_\ell^{i+1}$ lies in the finite element space V_ℓ. □

5.2 Guaranteed and locally efficient a posteriori error bounds robust with respect to the strength of the nonlinearity

Based on Theorem 5.3, estimating $|\!|\!|\mathcal{R}(u_\ell^i)|\!|\!|_{-1,u_\ell^i}$ amounts to estimating $\left|\!\left|\!\left|\mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1})\right|\!\right|\!\right|_{-1,u_\ell^i}$, where the latter corresponds to the numerical approximation (4.7) of the *linear* (reaction–)diffusion problem (4.13). For this purpose, we will follow [1, 2, 42, 43, 60, 64], the pure diffusion approaches [10, 20, 21, 24], but most precisely [60, Definition 2.3].

For each vertex $\boldsymbol{a}$ in the set of vertices $\mathcal{V}_\ell$, let the patch $\mathcal{T}_{\boldsymbol{a}}$ be the set of mesh elements from $\mathcal{T}_\ell$ sharing $\boldsymbol{a}$. We denote by $\psi_{\boldsymbol{a}}$ the corresponding hat function (piecewise affine function with value 1 in $\boldsymbol{a}$ and 0 in the other vertices), with the subdomain $\omega_{\boldsymbol{a}}$ being the interior of the support of $\psi_{\boldsymbol{a}}$. Let $h_{\omega_{\boldsymbol{a}}}$ be the diameter of $\omega_{\boldsymbol{a}}$.

As in (4.1), let $\mathcal{P}_p(K)$ stand for the space of polynomials of total degree at most p on the simplex $K \in \mathcal{T}_\ell$; $\boldsymbol{\mathcal{P}}_p(K;\mathbb{R}^d)$ denotes its componentwise extension to $\mathbb{R}^d$. The broken (elementwise) spaces $\mathcal{P}_p(\mathcal{T}_{\boldsymbol{a}})$ and $\boldsymbol{\mathcal{RT}}_p(\mathcal{T}_{\boldsymbol{a}})$ on the patch $\mathcal{T}_{\boldsymbol{a}}$ are then defined by, cf. [7],

$$\begin{aligned}\mathcal{P}_p(\mathcal{T}_{\boldsymbol{a}}) &:= \{v_\ell \in L^2(\omega_{\boldsymbol{a}}), \quad v_\ell|_K \in \mathcal{P}_p(K) \quad \forall K \in \mathcal{T}_{\boldsymbol{a}}\},\\ \boldsymbol{\mathcal{RT}}_p(\mathcal{T}_{\boldsymbol{a}}) &:= \{\boldsymbol{v}_\ell \in \boldsymbol{L}^2(\omega_{\boldsymbol{a}};\mathbb{R}^d), \quad \boldsymbol{v}_\ell|_K \in \boldsymbol{\mathcal{RT}}_p(K) := \boldsymbol{\mathcal{P}}_p(K;\mathbb{R}^d) + \boldsymbol{x}\mathcal{P}_p(K) \quad \forall K \in \mathcal{T}_{\boldsymbol{a}}\}.\end{aligned}$$

We remark that $\boldsymbol{\mathcal{P}}_p(\mathcal{T}_{\boldsymbol{a}};\mathbb{R}^d) \subsetneq \boldsymbol{\mathcal{RT}}_p(\mathcal{T}_{\boldsymbol{a}}) \subsetneq \boldsymbol{\mathcal{P}}_{p+1}(\mathcal{T}_{\boldsymbol{a}};\mathbb{R}^d)$. Decompose $\mathcal{V}_\ell$ into interior vertices $\mathcal{V}_\ell^{\mathrm{int}}$ and boundary vertices $\mathcal{V}_\ell^{\mathrm{ext}}$. The local mixed finite element spaces are then defined by

$$Q_{\boldsymbol{a}} := \begin{cases} \mathcal{P}_p(\mathcal{T}_{\boldsymbol{a}}) & \text{if } \boldsymbol{a} \in \mathcal{V}_\ell^{\mathrm{ext}} \text{ or } L_{\mathrm{M}}^i|_{\omega_{\boldsymbol{a}}} > 0,\\ \{v_\ell \in \mathcal{P}_p(\mathcal{T}_{\boldsymbol{a}}), \quad (v_\ell, 1)_{\omega_{\boldsymbol{a}}} = 0\} & \text{if } \boldsymbol{a} \in \mathcal{V}_\ell^{\mathrm{int}} \text{ and } L_{\mathrm{M}}^i|_{\omega_{\boldsymbol{a}}} = 0,\end{cases} \tag{5.6a}$$

$$\boldsymbol{V}_{\boldsymbol{a}} := \begin{cases} \{\boldsymbol{v}_\ell \in \boldsymbol{\mathcal{RT}}_p(\mathcal{T}_{\boldsymbol{a}}) \cap \boldsymbol{H}(\mathrm{div},\omega_{\boldsymbol{a}}),\ \boldsymbol{v}_\ell{\cdot}\boldsymbol{n} = 0 \text{ on } \partial\omega_{\boldsymbol{a}}\} \text{ if } \boldsymbol{a} \in \mathcal{V}_\ell^{\mathrm{int}},\\ \{\boldsymbol{v}_\ell \in \boldsymbol{\mathcal{RT}}_p(\mathcal{T}_{\boldsymbol{a}}) \cap \boldsymbol{H}(\mathrm{div},\omega_{\boldsymbol{a}}),\ \boldsymbol{v}_\ell{\cdot}\boldsymbol{n} = 0 \text{ on } \partial\omega_{\boldsymbol{a}} \setminus \{\psi_{\boldsymbol{a}} > 0\}\} \text{ if } \boldsymbol{a} \in \mathcal{V}_\ell^{\mathrm{ext}}.\end{cases} \tag{5.6b}$$

In order to define our estimators, we solve independent small discrete local linear reaction–diffusion problems in a dual formulation on the vertex patches $\mathcal{T}_{\boldsymbol{a}}$. They ultimately yield the equilibrated flux–potential pair $\boldsymbol{\sigma}_\ell^{i+1}$, ϕ_ℓ^{i+1} and read (details of the notation are postponed to Section 8.1 below):

Definition 5.4 (Equilibrated flux $\boldsymbol{\sigma}_\ell^{i+1}$ and potential ϕ_ℓ^{i+1})**.** *Let $u_\ell^{i+1} \in V_\ell$ be defined by (4.7) or, equivalently, (4.9). For each vertex $\boldsymbol{a} \in \mathcal{V}_\ell$, let $(\boldsymbol{\sigma}_{\boldsymbol{a}}^{i+1}, \phi_{\boldsymbol{a}}^{i+1}) \in \boldsymbol{V}_{\boldsymbol{a}} \times Q_{\boldsymbol{a}}$ be defined by the constrained minimization problem*

$$(\boldsymbol{\sigma}_{\boldsymbol{a}}^{i+1}, \phi_{\boldsymbol{a}}^{i+1}) := \mathop{\arg\min}_{\substack{(\boldsymbol{v}_\ell, q_\ell) \in \boldsymbol{V}_{\boldsymbol{a}} \times Q_{\boldsymbol{a}},\\ \nabla\cdot\boldsymbol{v}_\ell + \Pi_{\ell,p}(L^i\, q_\ell) =\\ -\nabla\psi_{\boldsymbol{a}}\cdot\boldsymbol{\Pi}_{\ell,p-1}^{\mathrm{RT}}(\mathfrak{a}^i\nabla u_\ell^{i+1} + \boldsymbol{\mathcal{F}}^i)\\ -\Pi_{\ell,p}(\psi_{\boldsymbol{a}}\mathcal{S}^i)}} \begin{bmatrix} (\|w^i(a_{\mathrm{m}}^i)^{-\frac{1}{2}}(\psi_a \boldsymbol{\Pi}_{\ell,p-1}^{\mathrm{RT}}(\mathfrak{a}^i\nabla u_\ell^{i+1} + \boldsymbol{\mathcal{F}}^i) + \boldsymbol{v}_\ell)\|_{\omega_{\boldsymbol{a}}}^2 \\ + \|(L^i)^{\frac{1}{2}}(\Pi_{\ell,p}(\psi_{\boldsymbol{a}}\, u_\ell^{i+1}) - q_\ell)\|_{\omega_{\boldsymbol{a}}}^2 \end{bmatrix}, \tag{5.7}$$

with, respectively, the coefficients $0 \le L_{\mathrm{M}}^i$, $0 < a_{\mathrm{m}}^i$ defined in (4.5), the scalar- and vector-valued source terms $\mathcal{S}^i$ and $\boldsymbol{\mathcal{F}}^i$ in (4.11) or (4.12), the elementwise constant weight function $w^i > 0$ in (8.3a), and the projection operators $\Pi_{\ell,p} : L^2(\Omega) \to \mathcal{P}_p(\mathcal{T}_\ell)$, $\boldsymbol{\Pi}_{\ell,p-1}^{\mathrm{RT}} : \boldsymbol{L}^2(\Omega;\mathbb{R}^d) \to \boldsymbol{\mathcal{RT}}_{p-1}(\mathcal{T}_\ell;\mathbb{R}^d)$ in (8.1) below. Then, after extending $\boldsymbol{\sigma}_{\boldsymbol{a}}^{i+1}$, $\phi_{\boldsymbol{a}}^{i+1}$ by zero from each patch subdomain $\omega_{\boldsymbol{a}}$ to Ω, we define

$$\boldsymbol{\sigma}_\ell^{i+1} := \sum_{\boldsymbol{a} \in \mathcal{V}_\ell} \boldsymbol{\sigma}_{\boldsymbol{a}}^{i+1}, \quad \phi_\ell^{i+1} := \sum_{\boldsymbol{a} \in \mathcal{V}_\ell} \phi_{\boldsymbol{a}}^{i+1}. \tag{5.8}$$

Remark 5.5 (Flux equilibration)**.** *The flux equilibration of Definition 5.4 is in general of reaction–diffusion type, following [60, Definition 2.3]. If, however, $L^i = 0$, as in Remark 4.2, Definition 5.4 boils down to the usual pure diffusion equilibration of [10, 20, 21, 24] where the weight function w^i equals one and the potentials $\phi_{\boldsymbol{a}}^{i+1}$ and ϕ_ℓ^{i+1} are not needed.*

Definition 5.4 in particular implies:

Lemma 5.6 (Properties of $\boldsymbol{\sigma}_\ell^{i+1}$ and ϕ_ℓ^{i+1})**.** *The equilibrated flux and potential from Definition 5.4 are well defined and satisfy*

$$\boldsymbol{\sigma}_\ell^{i+1} \in \boldsymbol{\mathcal{RT}}_p(\mathcal{T}_\ell) \cap \boldsymbol{H}(\mathrm{div}, \Omega), \qquad \phi_\ell^{i+1} \in \mathcal{P}_p(\mathcal{T}_\ell). \tag{5.9}$$

The potential contributions satisfy the mean value property

$$(L^i\, \phi_{\boldsymbol{a}}^{i+1}, 1)_{\omega_{\boldsymbol{a}}} = (L^i\, u_\ell^{i+1}, \psi_{\boldsymbol{a}})_{\omega_{\boldsymbol{a}}} \qquad \forall \boldsymbol{a} \in \mathcal{V}_\ell^{\mathrm{int}} \tag{5.10}$$

and the flux fulfils the mass balance

$$\Pi_{\ell,p}(L^i \phi_\ell^{i+1}) + \nabla{\cdot}\boldsymbol{\sigma}_\ell^{i+1} = -\Pi_{\ell,p}\mathcal{S}^i. \tag{5.11}$$

Let a subdomain $\omega \subseteq \Omega$ correspond to some mesh elements from $\mathcal{T}_\ell$ and let $\widetilde{\omega}$ correspond to ω and the mesh elements neighboring by vertex. With the equilibrated flux $\boldsymbol{\sigma}_\ell^{i+1}$ and potential ϕ_ℓ^{i+1}, we can now define our (discretization error) estimators:

$$\eta^i_{\mathrm{F},\omega} := \|\mathsf{w}^i(a^i_{\mathrm{m}})^{-\frac{1}{2}}(\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}(\mathfrak{a}^i\nabla u_\ell^{i+1} + \boldsymbol{\mathcal{F}}^i) + \boldsymbol{\sigma}_\ell^{i+1})\|_\omega, \tag{5.12a}$$

$$\eta^i_{\mathrm{S},\omega} := \|(L^i)^{\frac{1}{2}}(u_\ell^{i+1} - \phi_\ell^{i+1})\|_\omega, \tag{5.12b}$$

$$\eta^i_{\mathrm{osc},\omega} := \left(\sum_{K\in\mathcal{T}_\ell,\, K\subseteq\omega} (\widetilde{\mathsf{w}}^i_K\|(I - \Pi_{\ell,p})(\mathcal{S}^i + L^i\phi_\ell^{i+1})\|_K)^2\right)^{\frac{1}{2}}, \tag{5.12c}$$

$$\eta^i_{\mathrm{quad,F},\omega} := \|(\mathfrak{a}^i)^{-\frac{1}{2}}((\mathbb{I} - \boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1})(\mathfrak{a}^i\nabla u_\ell^{i+1} + \boldsymbol{\mathcal{F}}^i))\|_\omega, \tag{5.12d}$$

$$\eta^i_{\mathrm{quad,S},\omega} := \left(\sum_{\boldsymbol{a}\in\mathcal{V}_\ell,\, \omega_{\boldsymbol{a}}\subseteq\widetilde{\omega}} \|(L^i)^{\frac{1}{2}}(\Pi_{\ell,p}(\psi_{\boldsymbol{a}}u_\ell^{i+1}) - \Pi_{\ell,p,L^i}(\psi_{\boldsymbol{a}}u_\ell^{i+1}))\|^2_{\omega_{\boldsymbol{a}}}\right)^{\frac{1}{2}}, \tag{5.12e}$$

$$\eta^i_{\mathrm{quad,S,osc},\omega} := \left(\sum_{\boldsymbol{a}\in\mathcal{V}_\ell,\, \omega_{\boldsymbol{a}}\subseteq\widetilde{\omega}} \frac{h^2_{\omega_{\boldsymbol{a}}}}{\min_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{m}}}\|(I - \Pi_{\ell,p})(\psi_{\boldsymbol{a}}(\mathcal{S}^i + L^i\, u_\ell^{i+1}))\|^2_{\omega_{\boldsymbol{a}}}\right)^{\frac{1}{2}}. \tag{5.12f}$$

They respectively express the $\boldsymbol{H}(\mathrm{div}, \Omega)$ non-conformity of the direct finite element fluxes $\mathfrak{a}^i\nabla u_\ell^{i+1} + \boldsymbol{\mathcal{F}}^i$, a possible weighted discrepancy between u_ℓ^{i+1} and ϕ_ℓ^{i+1} important for large values of L^i, the possible oscillations in the source term $\mathcal{S}^i$ and $L^i\phi_\ell^{i+1}$, and quadrature errors for the flux and source terms. Note that the quadrature estimator $\eta^i_{\mathrm{quad,S},\omega}$ vanishes if L^i is elementwise constant; similarly, if $\mathfrak{a}^i$ is elementwise constant and $\boldsymbol{\mathcal{F}}^i \in \boldsymbol{\mathcal{P}}_p(\mathcal{T}_\ell, \mathbb{R}^d)$ (often satisfied in the lowest polynomial degree setting $p = 1$), then $\eta^i_{\mathrm{quad,F},\omega} = 0$.

Introducing

$$\eta^i_{\mathrm{disc},\omega} := \left(\sum_{\substack{K\in\mathcal{T}_\ell\\ K\subseteq\omega}} (\eta^i_{\mathrm{F},K} + \eta^i_{\mathrm{S},K} + \eta^i_{\mathrm{osc},K} + \eta^i_{\mathrm{quad,F},K})^2\right)^{\frac{1}{2}}, \qquad \text{(discretization estimator)} \tag{5.13a}$$

$$\eta^i_{\mathrm{lin},\omega} := |\!|\!| u_\ell^{i+1} - u_\ell^i |\!|\!|_{1,u_\ell^i,\omega}, \qquad \text{(linearization estimator)} \tag{5.13b}$$

$$\eta^i_\omega := ([\eta^i_{\mathrm{lin},\omega}]^2 + [\eta^i_{\mathrm{disc},\omega}]^2)^{\frac{1}{2}}, \qquad \text{(total estimator)} \tag{5.13c}$$

our second main result reads:

Theorem 5.7 (Guaranteed, locally efficient, and robust a posteriori estimates)**.** *Let the assumptions of Theorem 5.3 be satisfied. For a subdomain $\omega \subseteq \Omega$ corresponding to some mesh elements from $\mathcal{T}_\ell$, let $\widetilde{\omega}$ correspond to ω and the mesh elements neighboring by vertex. For any linearization step $i \geq 0$, there holds*

$$|||\mathcal{R}(u^i_\ell)|||^2_{-1,u^i_\ell} \leq [\eta^i_\Omega]^2, \tag{5.14a}$$

(guaranteed reliability)

$$[\eta^i_\omega]^2 \lesssim \vartheta^2_{\widetilde{\omega},\ell,i} \left|\!\left|\!\left| \mathcal{R}^{u^i_\ell}_{\mathrm{disc}}(u^{i+1}_\ell) \right|\!\right|\!\right|^2_{-1,u^i_\ell,\widetilde{\omega}} + [\eta^i_{\mathrm{lin},\omega}]^2 + [\eta^i_{\mathrm{osc},\widetilde{\omega}}]^2 + \vartheta^2_{\widetilde{\omega},\ell,i}[\eta^i_{\mathrm{quad,F},\widetilde{\omega}}]^2 + [\eta^i_{\mathrm{quad,S},\widetilde{\omega}}]^2 + [\eta^i_{\mathrm{quad,S,osc},\widetilde{\omega}}]^2, \tag{5.14b}$$

(local efficiency)

$$[\eta^i_\Omega]^2 \lesssim \vartheta^2_{\Omega,\ell,i} |||\mathcal{R}(u^i_\ell)|||^2_{-1,u^i_\ell} + [\eta^i_{\mathrm{osc},\Omega}]^2 + \vartheta^2_{\Omega,\ell,i}[\eta^i_{\mathrm{quad,F},\Omega}]^2 + [\eta^i_{\mathrm{quad,S},\Omega}]^2 + [\eta^i_{\mathrm{quad,S,osc},\Omega}]^2, \tag{5.14c}$$

(global efficiency)

where the local variability constant $\vartheta_{\omega,\ell,i} \geq 1$ is defined as $\vartheta_{\omega,\ell,i} := \max_{\boldsymbol{a}\in\mathcal{V}_\ell,\, \omega_{\boldsymbol{a}}\subseteq\omega} \vartheta_{\boldsymbol{a},\ell,i}$ with

$$\vartheta^2_{\boldsymbol{a},\ell,i} := \frac{\max_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{M}}}{\min_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{m}}} \qquad \textit{if } \frac{h^2_{\omega_{\boldsymbol{a}}} \max_{\omega_{\boldsymbol{a}}} L^i_{\mathrm{M}}}{\max_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{M}}} \leq 1, \tag{5.15a}$$

(diffusion-dominated case)

$$\vartheta^2_{\boldsymbol{a},\ell,i} \textit{only depends on } \frac{\max_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{M}}}{\min_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{m}}}, \frac{\max_{\omega_{\boldsymbol{a}}} L^i_{\mathrm{M}}}{\min_{\omega_{\boldsymbol{a}}} L^i_{\mathrm{m}}} \qquad \textit{if } \frac{h^2_{\omega_{\boldsymbol{a}}} \max_{\omega_{\boldsymbol{a}}} L^i_{\mathrm{M}}}{\max_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{M}}} > 1 \tag{5.15b}$$

(reaction-dominated case)

and measures the maximum local patch-wise variation of coefficients L^i and $\mathfrak{a}^i$ of (4.4) in ω, and the constant in $\lesssim$ only depends on the space dimension d, the shape-regularity constant of $\mathcal{T}_\ell$, and the polynomial degree p.

The proof of Theorem 5.3 is postponed to Section 8. We conclude this section by three remarks.

Remark 5.8 (Theorem 5.7)**.** *The estimate (5.14a) gives a guaranteed and fully computable upper bound on the error. The error lower bounds (5.14b)–(5.14c) feature a generic constant hidden in $\lesssim$ (only depending on p, d, and the regularity of $\mathcal{T}_\ell$), so that their quality (robustness) solely hinges on $\vartheta_{\boldsymbol{a},\ell,i}$ from (5.15).*

Remark 5.9 (Robustness in the Zarantonello case)**.** *The Zarantonello linearization of Table 1 gives $L^i = 0$ and $\mathfrak{a}^i = \Lambda =$ constant, so that case (5.15a) always applies and the variability constants $\vartheta_{\boldsymbol{a},\ell,i} = 1$. In this case, the estimates of Theorem 5.3 are always robust with respect to the strength of the nonlinearity.*

Remark 5.10 (Cases (5.15a) and (5.15b))**.** *Case (5.15a) corresponds to a situation where the reaction coefficients L^i from the linearization scalar product (4.4) do not dominate over the diffusion coefficients $\mathfrak{a}^i$ from (4.4); note the presence of the mesh size $h_{\omega_{\boldsymbol{a}}}$ implying that this case always prevails under sufficient local mesh refinement. Also recall from Remark 4.2 and Tables 1 and 2 that typically $L^i = 0$ when $f(\boldsymbol{x}, \xi)$ in (3.1) or (3.3) only depends on the space coordinate $\boldsymbol{x}$, i.e., when $f(\boldsymbol{x})$ is a source term.*

Remark 5.11 (The variability constants $\vartheta_{\boldsymbol{a},\ell,i}$ from (5.15))**.** *The constants $\vartheta_{\boldsymbol{a},\ell,i}$ from (5.15) describe a local variation of the linearization coefficient $\mathfrak{a}^i$ via the piecewise constants $a^i_{\mathrm{m}}, a^i_{\mathrm{M}}$ from Assumption 4.1 in (5.15a), and additionally on the local variation of L^i via $L^i_{\mathrm{m}}, L^i_{\mathrm{M}}$ in (5.15b). This is expected to be (much) smaller than the global variation of $\mathfrak{a}^i$ and L^i over* *the whole computational domain Ω due to the interior regularity of solutions of elliptic equations [26, Section 6.3], and definitely much smaller than the ratio $\lambda_{\mathrm{M}}/\lambda_{\mathrm{m}}$ from* (1.4) *or* (1.5)*; $\vartheta_{\boldsymbol{a},\ell,i}$ can be small and/or tend to one with local mesh refinement even for strong nonlinearities. Moreover, even in the worst case, at least for the gradient-dependent diffusivity* (3.1) *in the case of Example 3.2 and with f only depending on the space coordinate $\boldsymbol{x}$, $\vartheta_{\boldsymbol{a},\ell,i} \leq (\sigma_{\mathrm{M}}/\sigma_{\mathrm{m}})^{1/2}$ for the Kačanov and Newton linearization schemes of Table 1 as discussed in [33, Sections 2.3.2 and 3.7], which still outperforms* (1.4) *and* (1.5) *by the factor $(\sigma_{\mathrm{M}}/\sigma_{\mathrm{m}})^{1/2}$.*

Remark 5.12 (A posteriori assessment of robustness)**.** *Whenever $\vartheta_{\Omega,\ell,i}$ is of order of unity, the estimates of Theorem 5.7 are robust with respect to the strength of the nonlinearity. Crucially, $\vartheta_{\Omega,\ell,i}$ can be easily computed at any mesh $\mathcal{T}_\ell$ and linearization iteration i in any setting (problem, linearization scheme) in* (5.15a)*, so that robustness is assessed a posteriori.*

5.3 Balancing of the linearization and discretization components and stopping criteria for iterative linearizations

The orthogonal error decomposition of Theorem 5.3 together with the computable estimators of Theorem 5.7 yield the following adaptive stopping criterion for iterative linearizations that ensures a natural balance between discretization and linearization components, in that it requests that the linearization component only forms a user-given μ-fraction of the total error.

Algorithm 5.13 (Adaptive linearization)**.** *Let $\mu \in (0,1)$ and $u_\ell^0 \in V_\ell$ be fixed. Pursue the iterative linearization* (4.7) *until, for some $\bar{i} \geq 0$,*

$$\eta^{\bar{i}}_{\mathrm{lin},\Omega} \leq \mu\, \eta^{\bar{i}}_{\Omega}, \tag{5.16}$$

where $\eta^{\bar{i}}_{\mathrm{lin},\Omega}$ and $\eta^{\bar{i}}_{\Omega}$ are computed following (5.13)*.*

Moreover, the linearization estimator $\eta^{i}_{\mathrm{lin},\Omega}$ can also be used to switch between linearization schemes in an efficient way, see [61] for an example. However, this is not pursued in this work.

6 Numerical experiments

In this section, we test the numerical performance of the estimators of Theorem 5.7 and the stopping criteria of Algorithm 5.13. In what follows, a fixed parameter $\mu = 0.05$ from (5.16) has been used. Three levels of discretization are considered with an initial mesh $\mathcal{T}_1$ fixed and its uniform refinements $\mathcal{T}_\ell$, where we choose $\ell \in \{2,4\}$. Following Section 4.1, the corresponding finite element spaces V_ℓ are given as $\mathcal{P}_1(\mathcal{T}_\ell) \cap H^1(\Omega)$, i.e., $p = 1$. For linearization step $i \geq 0$ and $\ell \in \{1,2,4\}$, let $u^{i+1}_{\mathrm{ref}} \in V_{10}$ (i.e. $\ell = 10$) be a fine-mesh approximation to $u^{i+1}_{\langle\ell\rangle}$ from (4.13) that we use for approximate evaluation of the error; indeed, by Lemma 5.2, one has that

$$|||\mathcal{R}(u^i_\ell)|||_{-1,u^i_\ell} = \left|\left|\left| u^i_\ell - u^{i+1}_{\langle\ell\rangle} \right|\right|\right|_{1,u^i_\ell} \approx |||u^i_\ell - u^{i+1}_{\mathrm{ref}}|||_{1,u^i_\ell}. \tag{6.1}$$

With these definitions, we express the quality of the estimates by introducing the global and local effectivity indices:

$$\text{Eff. Ind.} := \eta^i_\Omega / |||u^i_\ell - u^{i+1}_{\mathrm{ref}}|||_{1,u^i_\ell}, \tag{6.2a}$$

$$(\text{Eff. Ind.})_K := \eta^i_K / |||u^i_\ell - u^{i+1}_{\mathrm{ref}}|||_{1,u^i_\ell,K}, \qquad K \in \mathcal{T}_\ell, \tag{6.2b}$$

where η^i_Ω and η^i_K are defined in (5.13c). The numerical solutions are computed using the FreeFem++ software [34] and the codes are available online in a Git repository. We consider the following two test cases:

- Section 6.1. Gradient-dependent diffusivity case (3.1): mean curvature flow of Example 3.2, a singular reaction–source term, inhomogeneous Dirichlet boundary conditions, and known exact solution.

- Section 6.2. Gradient-independent diffusivity case (3.3): one time step of a backward Euler time discretization of the Richards equation with nonlinear degenerate advection–reaction–diffusion terms, anisotropic permeability tensor, and mixed boundary conditions.

6.1 Gradient-dependent diffusivity

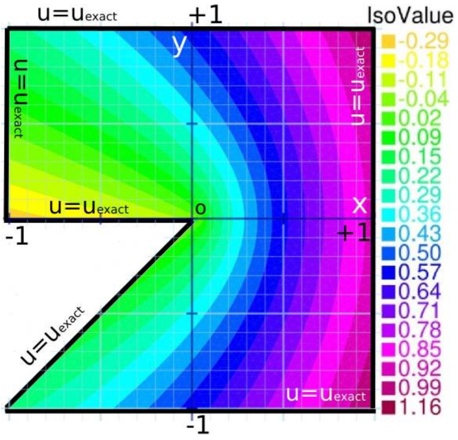

Figure 1: [Section 6.1] The computational domain and the exact solution for the gradient-dependent diffusivity test case.

The computational domain Ω is the square $(-1,1)^2$ with a triangular section removed, see Figure 1. In Ω, we consider, for (3.1),

$$\boldsymbol{\sigma}(\boldsymbol{x},\nabla\xi) := A(|\nabla\xi|)\nabla\xi \text{ with } A(\varrho) := 2\tau + 1/\sqrt{1+\varrho^2}, \tag{6.3a}$$

following Example 3.2, where τ is a parameter. Thus $\sigma_{\mathrm{m}} = 2\tau$, $\sigma_{\mathrm{M}} - \sigma_{\mathrm{m}} = 1$, and Assumption (A1) is satisfied. We define

$$f(\boldsymbol{x},\xi) := \nu\,\xi - g(\boldsymbol{x}), \tag{6.3b}$$

where $\nu = 10^{-2}$ is a fixed parameter. The function g follows from our choice of the exact solution: Let $\boldsymbol{x} \in \Omega$ be represented in polar coordinates by the radius-azimuth pair (r,θ) and let $\lambda := 4/7$ be a fixed parameter. Then,

$$u_{\mathrm{exact}}(\boldsymbol{x}) := r^{\lambda}\cos(\lambda\theta), \quad \text{yielding} \quad g(\boldsymbol{x}) := \left[r^{-1}\frac{(1-\lambda)(\lambda\, r^{\lambda-1})^3}{(1+(\lambda\, r^{\lambda-1})^2)^{\frac{3}{2}}} + \nu r^{\lambda}\right]\cos(\lambda\theta). \tag{6.4}$$

Observe that $g \in H^{-1}(\Omega)$ but $g \notin L^2(\Omega)$ due to the singularity at $r = 0$. Thus, Assumption (A2) is not satisfied, but our numerical results seem to carry over even to this case. We set the Dirichlet condition $u = u_{\mathrm{exact}}$ on $\partial\Omega$.

We consider two linearization schemes from Table 1, the Kačanov and the Zarantonello ones. In the Zarantonello case, we choose $\Lambda = 1$; this does not satisfy the theoretical requirement $\Lambda \geq \sigma_{\mathrm{M}}^2/\sigma_{\mathrm{m}}$ from the a priori analysis [66] but numerically leads to a fast convergence (no requirement on Λ appears in our a posteriori theory). In fact, for the theoretical choice of $\Lambda = \sigma_{\mathrm{M}}^2/\sigma_{\mathrm{m}}$, the Zarantonello scheme barely converges for $\tau < 10^{-1}$. The initial guess is taken to be $u_\ell^0 = \frac{11}{10}u_{\mathrm{exact}}$.

For $\tau = 0.01$, Figure 2 (left) plots the global effectivity indices for these schemes against the linearization iteration index i. The indices are between 1–2 in all cases. Figure 2 (center) shows the distribution of the local errors. As expected, the error is concentrated around the

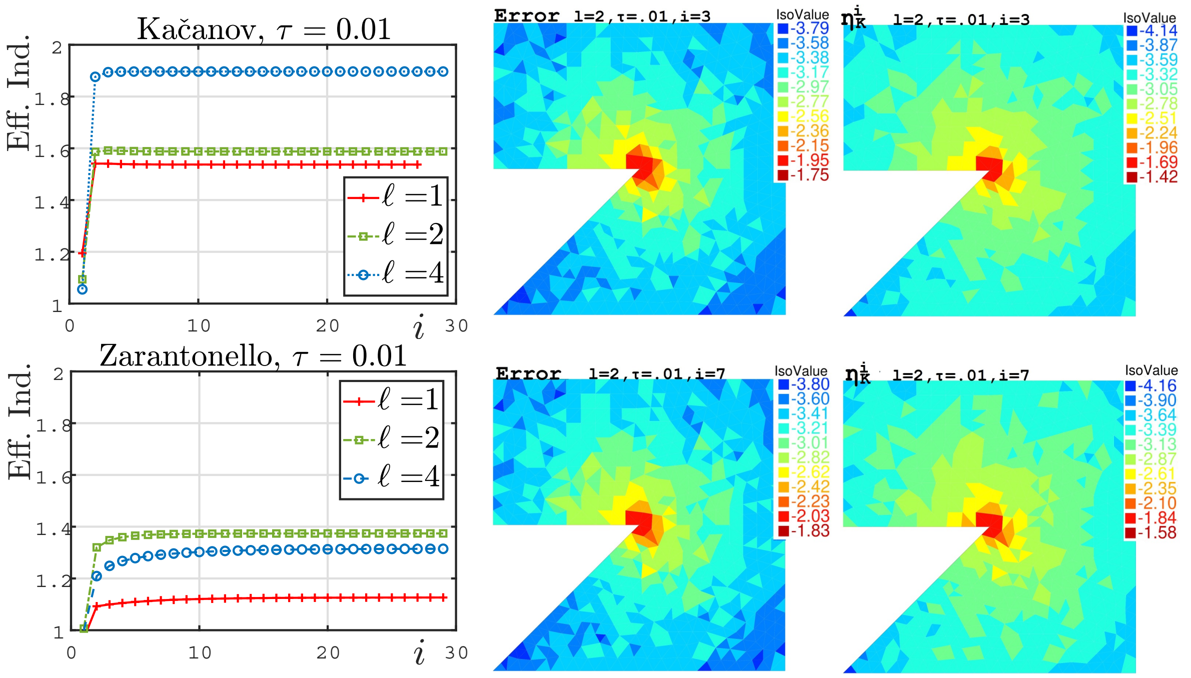

Figure 2: [Section 6.1, $\tau = 0.01$] Comparison between errors and estimators. The Kačanov linearization (top) and the Zarantonello linearization (bottom). Global effectivity indices (6.2a) as functions of the refinement level ℓ and the linearization iteration i (left). Elementwise errors $|\!|\!| u_\ell^i - u_{\text{ref}}^{i+1} |\!|\!|_{1,u_\ell^i,K}$ (center) and estimators η_K^i (right) for $\ell = 2$ and at the linearization iteration $i = \bar{i}$ satisfying the stopping criterion (5.16) (both in logarithmic scale).

singular re-entrant corner at the origin. It decreases radially, and its minima are situated at the other corners. The distribution of the estimators η_K^i, shown in Figure 2 (right), reproduces the same pattern. Figure 3 shows the corresponding local effectivity indices (6.2b). In almost every mesh element, these indices are between 0.75–1.5, and in all the cases they are in the order of unity. The effectivity indices around the singularity at the origin are about 1.5–2.0 in all the simulations, which we find to be particularly impressive. Note that an effectivity index less than 1 does not contradict the local efficiency estimate of Theorem 5.7 since, instead of the neighborhood $\widetilde{K}$, the estimates are computed over the individual mesh elements $K \in \mathcal{T}_\ell$.

The robustness of the estimates is next exhibited in Figure 4. For a fixed mesh, it plots the global effectivity indices (6.2a) of the linearization schemes against $1/\tau$, where we consider $\tau \in [10^{-3}, 1]$. Though the Lipschitz continuity/monotonicity ratio $\sigma_{\text{M}}/\sigma_{\text{m}} = (1 + 2\tau)/2\tau$, expressing the strength of the nonlinearity, varies over three orders of magnitude, the effectivity indices stay essentially intact. This illustrates our robustness result from Theorem 5.7. More precisely, for the Zarantonello linearization, we know from Remark 5.9 that the variability constant $\vartheta_{\Omega,\ell,i}$ equals one, so that our estimates are robust. For the Kačanov scheme, as explained in Remark 5.12, we can first compute the variability constants $\vartheta_{\Omega,\ell,i}$ to a posteriori assess the robustness also in this case, see Figure 4 (left) for the values of $\vartheta_{\Omega,\ell,i}$. We recall that $\vartheta_{\Omega,\ell,i}$ is from (5.15) computed easily and locally, and without the knowledge of u_{ref}^{i+1} or u_{exact} needed to compute the (global) effectivity indices (6.2a).

Figure 5 further illustrates how the errors, the linearization estimator $\eta_{\text{lin},\Omega}^i$, and the total estimator η_Ω^i vary with the linearization iterations. Linearization error dominates the total error at the start of the iterations but becomes negligible as the iterations converge, where the discretization component becomes dominant. The appropriateness of (5.16) is evident from the plots, since the adaptive criterion is met only after 3 iterations for the Kačanov scheme, and after 7–9 iterations for the Zarantonello scheme, and there is no significant change in the total

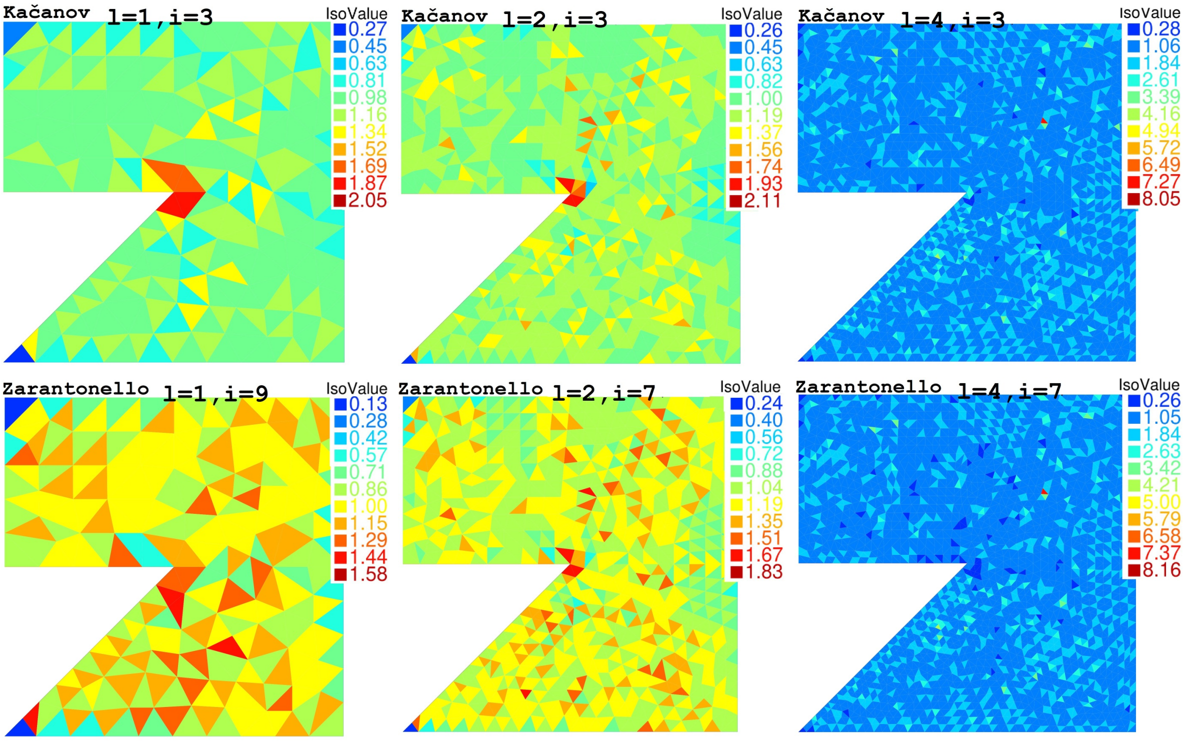

Figure 3: [Section 6.1, $\tau = 0.01$] Local effectivity indices (6.2b) for the Kačanov (top) and the Zarantonello (bottom) linearizations at different discretization levels ℓ and for $i = \bar{i}$ from (5.16).

errors beyond this point. Furthermore, from the simulations, the Kačanov scheme is found to be much faster than the Zarantonello scheme. The Zarantonello scheme, though, leads to slightly better effectivity indices, probably due to the local variability constants $\vartheta_{\boldsymbol{a},\ell,i}$ from (5.15) being 1 for this scheme (see Remark 5.9).

Finally, Figure 6 shows the relative magnitude of the components of the discretization estimator $\eta^i_{\text{disc},\Omega}$ defined in (5.13a). The flux estimator $\eta^i_{\text{F},\Omega}$ is the largest contributor to $\eta^i_{\text{disc},\Omega}$ followed by the data-oscillation estimator $\eta^i_{\text{osc},\Omega}$. Since, in Table 1, $L^i = \nu$ for the Kačanov scheme and $L^i = 0$ for the Zarantonello scheme, we have that $\eta^i_{\text{quad,S},\Omega} = 0$ in both cases, whereas $\eta^i_{\text{quad,F},\Omega} < 10^{-5}$. With $L^i = 0$, the Zarantonello scheme is a pure diffusion linearization, see Remarks 4.2 and 5.5, so that $\eta^i_{\text{S},\Omega} = 0$.

6.2 Gradient-independent diffusivity: the Richards equation

We choose here Ω as the unit square, $\Omega = (0,1) \times (0,1)$. The backward Euler time-discretized Richards equation is characterized here by the following choices of the functions in (3.3):

$$\begin{gathered} f(\boldsymbol{x},\xi) = S(\xi) - S(\bar{u}(\boldsymbol{x})), \quad \mathcal{D}(\boldsymbol{x},\xi) = \kappa(S(\xi)), \quad \boldsymbol{q}(\boldsymbol{x},\xi) = -\kappa(S(\xi))\,\boldsymbol{g}, \\ \bar{\boldsymbol{K}} = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}, \quad \boldsymbol{g} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \tau \in [10^{-3}, 1]. \end{gathered} \tag{6.5}$$

Here $\boldsymbol{g}$ is the gravity vector, and the functions $S : \mathbb{R} \to (0,1]$ and $\kappa : [0,1] \to [0,1]$, known as the saturation and relative permeability functions, respectively, are modeled using the van Genuchten [62] parametrization, expressed in a nondimensional setting for $\lambda \in (0,1)$ as

$$S(\xi) := \left(1 + (2-\xi)^{\frac{1}{1-\lambda}}\right)^{-\lambda}, \quad \kappa(s) := \sqrt{s}\left(1 - (1 - s^{\frac{1}{\lambda}})^{\lambda}\right)^2. \tag{6.6}$$

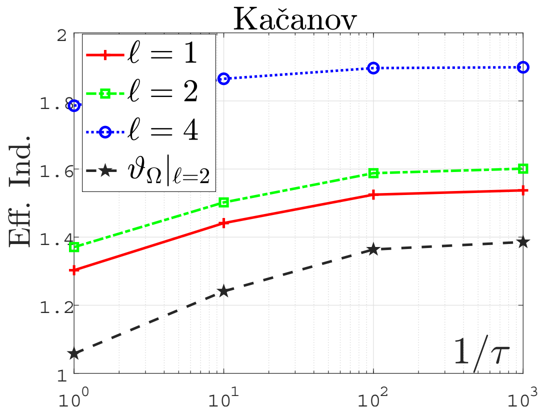

Figure 4: [Section 6.1, τ varies] Global effectivity indices (6.2a) vs. $1/\tau$ for the Kačanov (left) and Zarantonello (right) linearizations at different discretization levels ℓ for $i = \bar{i}$ from (5.16). For the Kačanov scheme, the variability constant $\vartheta_{\Omega,\ell,i}$ from (5.15) is also plotted for $\ell = 2$. For Zarantonello linearization, this constant is always 1.

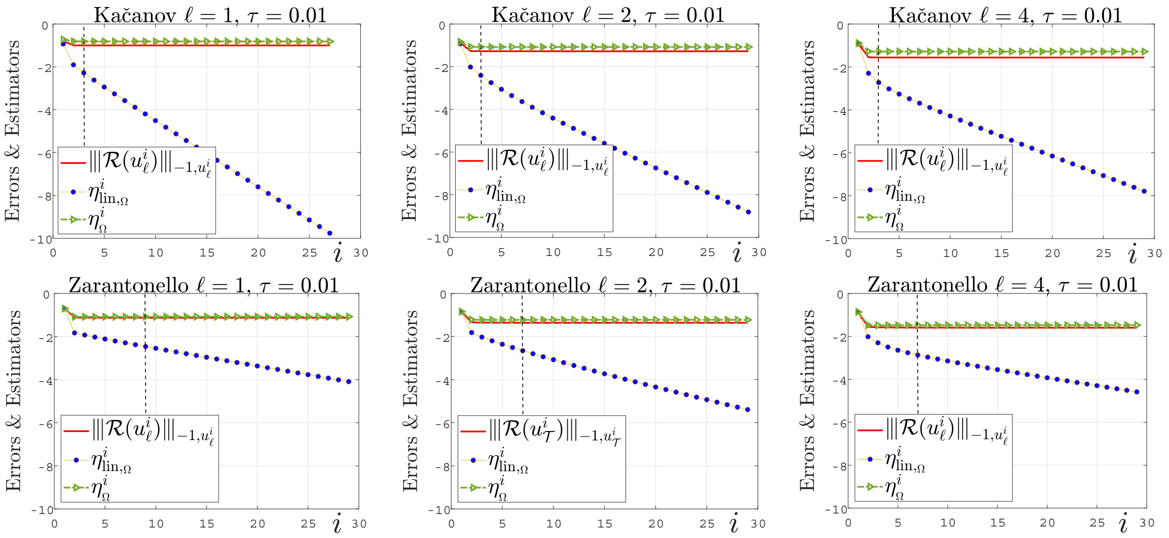

Figure 5: [Section 6.1, $\tau = 0.01$] Total error $|||\mathcal{R}(u_\ell^i)|||_{-1,u_\ell^i}$ approximated using (6.1), linearization error $\eta_{\mathrm{lin},\Omega}^i$, and the total estimator η_Ω^i against iteration index i in a $\log_{10}$-scale: the Kačanov scheme (top), the Zarantonello scheme (bottom). The vertical black dashed line on each plot indicates the stopping criterion $i = \bar{i}$ from (5.16).

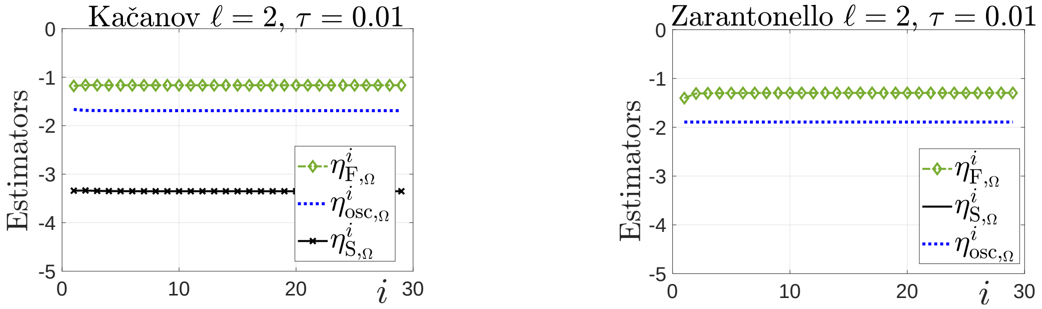

Figure 6: [Section 6.1, $\tau = 0.01$] The components (5.12) of the discretization estimator $\eta_{\mathrm{disc},\Omega}^i$ vs. iteration i for $\tau = 0.01$. For both schemes, $\eta_{\mathrm{quad,S},\Omega}^i = 0$, $\eta_{\mathrm{quad,F},\Omega}^i < 10^{-5}$, and for the Zarantonello scheme, $\eta_{\mathrm{S},\Omega}^i = 0$.

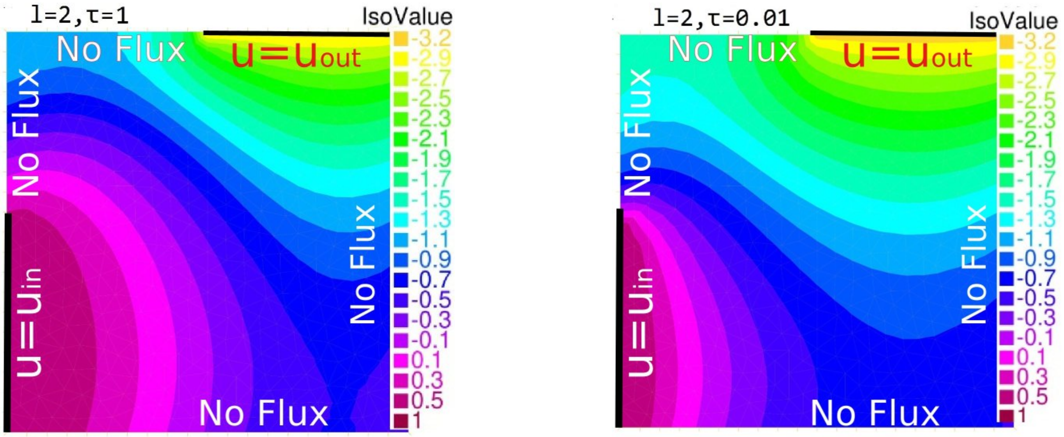

Figure 7: [Section 6.2] The domain $\Omega = (0,1)^2$ with the boundary conditions and reference solutions of (1.1) for $\tau = 1$ (left) and $\tau = 0.01$ (right).

More details on the Richards equation can be found in [53] and the references therein. For the rest of this paper, we will use $\lambda = 0.5$. The mixed (Dirichlet and no-flux) boundary condition used for the problem is shown in Figure 7 with $u_{\text{in}} = 1$ and $u_{\text{out}} = -3$. Finally, in the context of the time-dependent Richards equation, the quantity $\bar{u}$ represents the solution of the previous time step. Here, it is arbitrarily taken to be the solution of the Darcy problem $\nabla\cdot[\bar{\boldsymbol{K}}(\nabla\bar{u}-\boldsymbol{g})] = 0$ in Ω, completed with the mixed boundary condition mentioned above. It is straightforward to verify that the Richards equation satisfies the conditions stated in (B1)–(B3). Furthermore, it satisfies (B4) since in order for the constant $\gamma > 0$ to exist, one requires by Cauchy's mean value theorem that $(\kappa'(S(\xi))^2 S'(\xi)/\kappa(S(\xi))$ is bounded for all $\xi \in \mathbb{R}$, a requirement that is satisfied for $\lambda = 0.5$ ($\gamma \approx 1.2$ in this case).

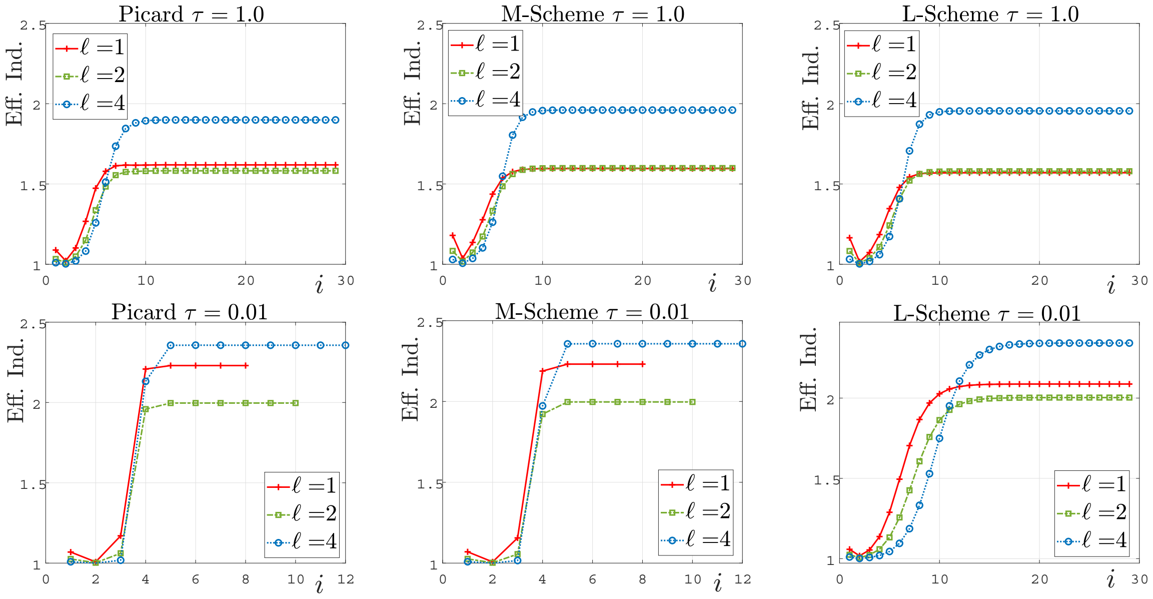

Figure 8: [Section 6.2, $\tau = 1$ and $\tau = 0.01$] Global effectivity indices (6.2) for different linearization schemes, τ, ℓ, and i values. Results for $\tau = 1$ (top) and $\tau = 0.01$ (bottom). The Picard scheme (left), the M-scheme (center), and the L-scheme (right).

We take the initial guess for the iterations as $u_\ell^0 = 0$ and show results for numerical schemes mentioned in Table 2. We fix the parameters for L- and M-schemes as $L = 1/6$ and $M = 0.1$ respectively. The M-scheme is chosen to depict the base numerical results since it is a hybrid between the L-scheme and the Picard (fixed point) iterations [52]. Global effectivity indices, computed using (6.2a), are plotted in Figure 8 for the different schemes, at all discretization levels, different time-step lengths τ, and iterations. It is seen that the effectivity indices stay above 1 in all cases, thus validating the guaranteed upper bound (5.14a). As the linearization iterations converge, these indices reach a stable value as expected.

To inspect the robustness of our estimators with respect to the strength of the nonlinearities, we in Figure 9 plot the effectivity indices against the Lipschitz/monotonicity constant of the operator $\mathcal{R}$, represented in this case by $1/\tau$. Despite the parameter τ varying between 1 and 10^{-3}, thus approaching the singularly perturbed regime, the effectivity indices stay bounded between 1.5 and 2.6. From Table 2 and (5.15), the variability constant $\vartheta_{\Omega,\ell,i}$ essentially depends on the patchwise variations of $\mathcal{D}(\boldsymbol{x}, u_\ell^i) = \kappa(S(u_\ell^i))$. Scaled by half, we plot $\vartheta_{\Omega,\ell,i}$ for all the schemes on the second mesh level $\ell = 2$. We observe uniformly moderate values, which immediately ensure the observed robustness, see Remark 5.12 (recall that $\vartheta_{\Omega,\ell,i}$ is easily and locally computable, whereas we can only compute the (global) effectivity indices (6.2a) knowing/approximating the exact solution).

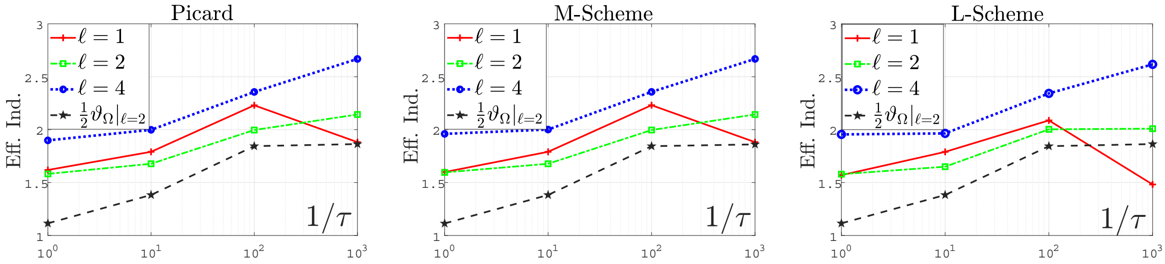

Figure 9: [Section 6.2, τ varies] Global effectivity indices (6.2a) vs. $1/\tau$ for the Picard (left), M- (center), and L-schemes (right) at different discretization levels ℓ for $i = \bar{i}$ from (5.16). The variability constant $\vartheta_{\Omega,\ell,i}$ from (5.15) is also plotted (scaled by half and for $\ell = 2$).

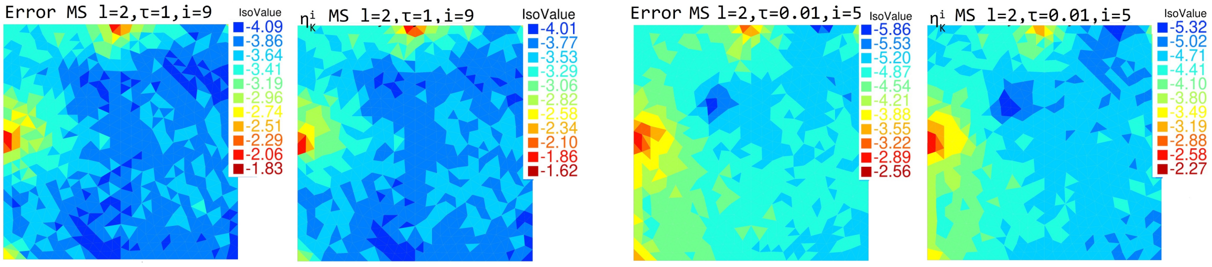

Figure 10: [Section 6.2, $\tau = 1$ and $\tau = 0.01$] The elementwise error approximated by $|\!|\!|u_\ell^i - u_{\text{ref}}^{i+1}|\!|\!|_{1,u_\ell^i,K}$ vs. the elementwise estimators η_K^i for $\ell = 2$ for the M-scheme at $i = \bar{i}$ from (5.16); $\tau = 1$ (left) and $\tau = 0.01$ (right) (all in logarithmic scale).

Figure 10 presents the distribution of the error for the M-scheme at $\ell = 2$ compared to the distribution of η_K^i. Using the maximum principle, we obtain from (6.6) that $\min_\Omega \kappa(S(u)) = \kappa(S(u_{\text{out}})) \approx 5 \times 10^{-4}$, whereas $\max_\Omega \kappa(S(u)) = \kappa(S(u_{\text{in}})) = 1$. As a result, the elementwise errors and estimators are seen to vary 2–4 orders of magnitude. Nevertheless, the plots suggest that the estimator η_K^i captures both the magnitude and the distribution of the error accurately. This is quantified by the local effectivity indices (6.2b) plotted in Figure 11. In most of the domain, the effectivity index is between 1 and 2.5, and in all cases it is in the order of unity. For the other linearization schemes, the local effectivity indices also stay within this range, despite the estimator η_Ω^i varying by about 2 to 4 orders of magnitude, τ varying 3 orders of magnitude, and the diffusion coefficient $\kappa \circ S$ in (6.5)–(6.6) varying by 3 to 4 orders of magnitude. Hence, we consider the match to be excellent.

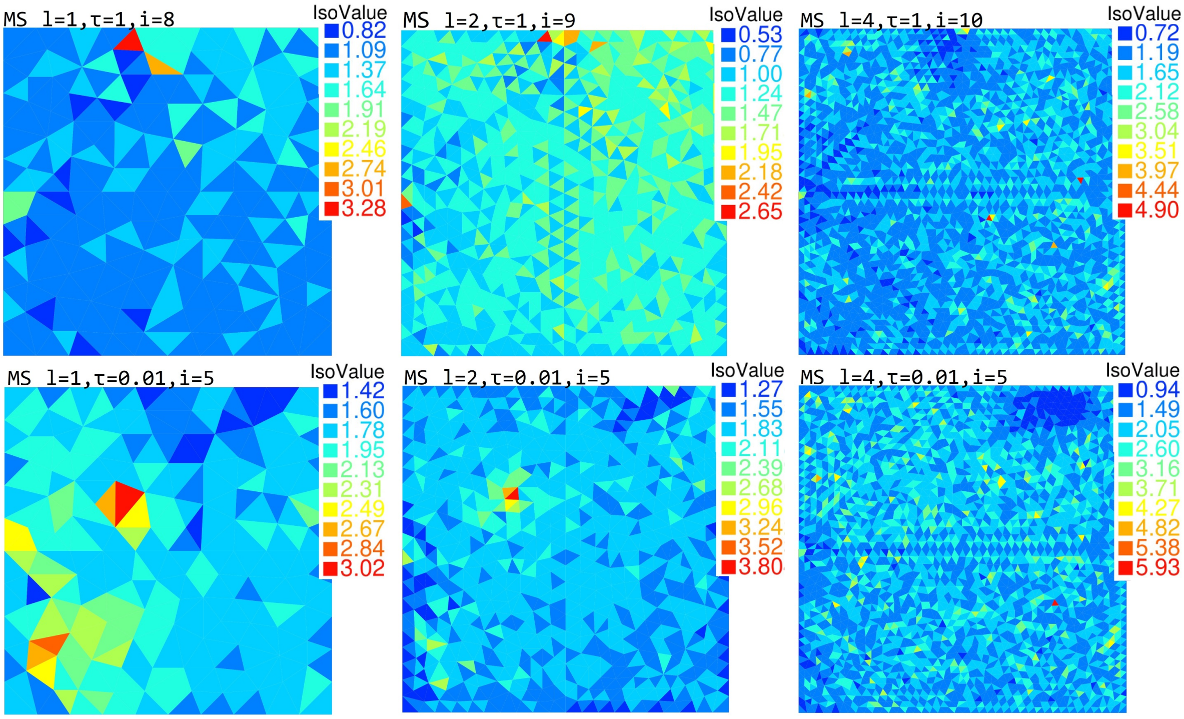

Figure 11: [Section 6.2, $\tau = 1$ and $\tau = 0.01$] Local effectivity indices (6.2b) for the M-scheme at discretization levels $\ell = 1, 2, 4$ for $i = \bar{i}$ from (5.16); $\tau = 1$ (top) and $\tau = 0.01$ (bottom).

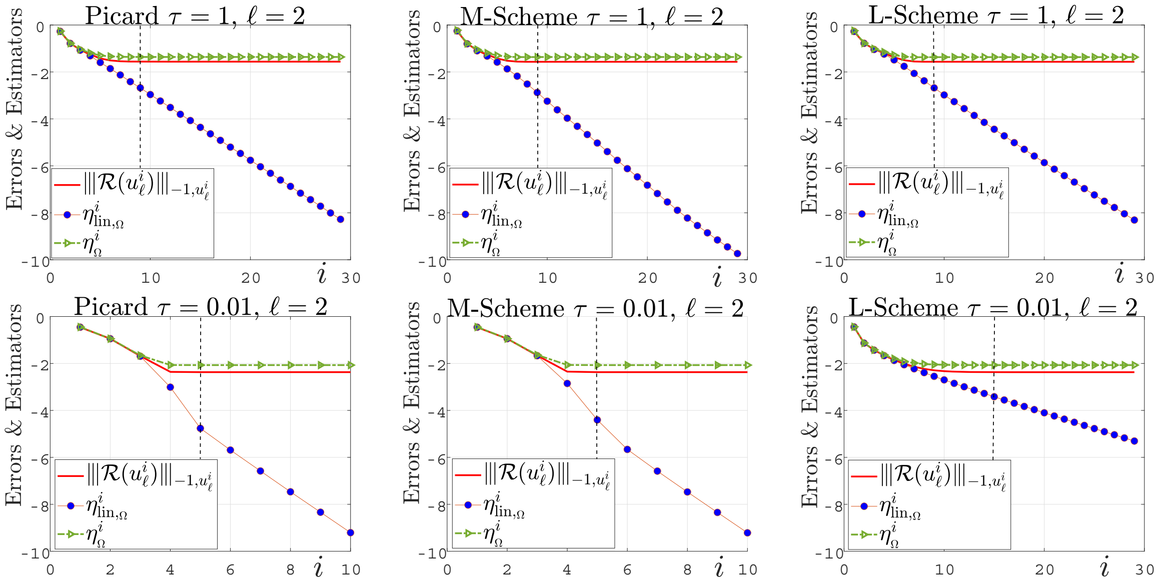

Figure 12: [Section 6.2, $\tau = 1$ and $\tau = 0.01$] Total error $|||\mathcal{R}(u_\ell^i)|||_{-1,u_\ell^i}$ approximated using (6.1), linearization error $\eta^i_{\mathrm{lin},\Omega}$, and total estimator η^i_Ω against iteration index i in a $\log_{10}$-scale. Fixed $\ell = 2$. The vertical black dashed line indicates the stopping criterion $i = \bar{i}$ from (5.16). Results are for $\tau = 1$ (top) and $\tau = 0.01$ (bottom). The Picard scheme (left), the M-scheme (center), and the L-scheme (right).

Next, we investigate how the error, the linearization estimator $\eta^i_{\mathrm{lin},\Omega}$, and the total estimator η^i_Ω vary with the iteration i for different linearization schemes. This is shown in Figure 12. At the beginning, the dominant error component is the linearization error, which is roughly equal to the total error. However, as the iterations advance, the total error is first dominated by the discretization error, and then dictated by it. The results also clearly portray the known trends of the linearization schemes [50, 52], i.e., the L-scheme converges more slowly as τ is decreased, whereas the Picard and the M-schemes speed up, and the M-scheme is faster than the Picard scheme particularly for larger values of τ. Figure 12 further illustrates the efficiency of (5.16), since the iterations are stopped after they cease to have any significant effect on the quality of the solutions.

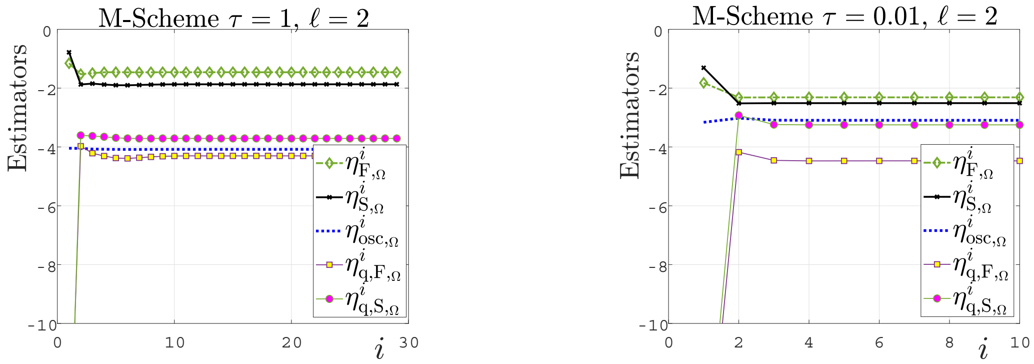

Figure 13: [Section 6.2, $\tau = 1$ and $\tau = 0.01$] The relative magnitudes of components of the discretization estimator $\eta^i_{\mathrm{disc},\Omega}$ from Section 5.2 plotted against iteration i for the M-scheme at discretization level $\ell = 2$; $\tau = 1$ (left) and $\tau = 0.01$ (right).

Finally, Figure 13 shows the relative magnitude of the components of the discretization estimator $\eta^i_{\mathrm{disc},\Omega}$, see (5.13a). The flux estimator $\eta^i_{\mathrm{F},\Omega}$ is the dominant component in all cases, followed in order by the potential estimator $\eta^i_{\mathrm{S},\Omega}$, the data-oscillation estimator $\eta^i_{\mathrm{osc},\Omega}$, and the quadrature estimator $\eta^i_{\mathrm{quad,F},\Omega}$ (we also plot $\eta^i_{\mathrm{quad,S},\Omega}$).

7 Well-posedness of the models (3.1) and (3.3)

In this section, we collect the existence and uniqueness proofs for the models (3.1) and (3.3).

7.1 Well-posedness of the gradient-dependent diffusivity model (3.1)

Proposition 7.1 (Monotonicity and Lipschitz continuity of $\mathcal{R}$ when defined by (3.1), existence and uniqueness of the weak solution of (1.1))**.** *Assume the expression* (3.1) *for* $\mathcal{R} : H^1_0(\Omega) \to H^{-1}(\Omega)$ *and (A1)–(A2). Then, for all* ξ, $\zeta \in H^1_0(\Omega)$*,*

$$\sigma_{\mathrm{m}} \|\nabla(\xi - \zeta)\| \leq \sup_{\varphi \in H^1_0(\Omega)} \frac{\langle \mathcal{R}(\xi) - \mathcal{R}(\zeta), \varphi \rangle}{\|\nabla \varphi\|} \leq (\sigma_{\mathrm{M}} + (C_\Omega\, h_\Omega)^2 f_{\mathrm{M}})\, \|\nabla(\xi - \zeta)\|, \tag{7.1}$$

and there exists a unique solution $u \in H^1_0(\Omega)$ *of* (1.1)*.*

Proof. The operator $\mathcal{R}$ from (3.1) is monotone and continuous since

$$\begin{aligned}
\langle \mathcal{R}(\xi) - \mathcal{R}(\zeta), \xi - \zeta \rangle &\overset{(3.1)}{=} (\boldsymbol{\sigma}(\cdot, \nabla\xi) - \boldsymbol{\sigma}(\cdot, \nabla\zeta), \nabla(\xi - \zeta)) + \langle f(\cdot, \xi) - f(\cdot, \zeta), \xi - \zeta \rangle \\
&\overset{(A2)}{\geq} (\boldsymbol{\sigma}(\cdot, \nabla\xi) - \boldsymbol{\sigma}(\cdot, \nabla\zeta), \nabla(\xi - \zeta)) \overset{(A1)}{\geq} \sigma_{\mathrm{m}} \|\nabla(\xi - \zeta)\|^2, \\
\sup_{\varphi \in H_0^1(\Omega)} \frac{\langle \mathcal{R}(\xi) - \mathcal{R}(\zeta), \varphi \rangle}{\|\nabla\varphi\|} &\overset{(A1)}{\leq} \sigma_{\mathrm{M}} \|\nabla(\xi - \zeta)\| + \sup_{\varphi \in H_0^1(\Omega)} \frac{\|f(\cdot, \xi) - f(\cdot, \zeta)\| \|\varphi\|}{\|\nabla\varphi\|} \\
&\overset{(2.1)}{\leq} \sigma_{\mathrm{M}} \|\nabla(\xi - \zeta)\| + C_\Omega\, h_\Omega\, \|f(\cdot, \xi) - f(\cdot, \zeta)\| \\
&\overset{(A2)}{\leq} \sigma_{\mathrm{M}} \|\nabla(\xi - \zeta)\| + C_\Omega\, h_\Omega\, f_{\mathrm{M}} \|\xi - \zeta\| \\
&\overset{(2.1)}{\leq} (\sigma_{\mathrm{M}} + (C_\Omega\, h_\Omega)^2 f_{\mathrm{M}}) \|\nabla(\xi - \zeta)\|.
\end{aligned}$$

The existence of $u \in H_0^1(\Omega)$ solving (1.1) then follows from the nonlinear Lax–Milgram theorem [67, Chapter 2] due to the monotonicity and Lipschitz continuity of $\mathcal{R}$. □

7.2 Well-posedness of the gradient-independent diffusivity model (3.3)

Proposition 7.2 (Monotonicity and Lipschitz continuity of $\mathcal{R}$ when defined by (3.3), existence and uniqueness of the weak solution of (1.1))**.** *Assume the expression* (3.3) *for* $\mathcal{R} : H_0^1(\Omega) \to H^{-1}(\Omega)$ *and* (B1)–(B5) *with* $0 < \tau < (4/\gamma)$*. Moreover, let the Kirchhoff transform function* $\mathcal{K} : \Omega \times \mathbb{R} \to \mathbb{R}$*, defined as*

$$\mathcal{K}(\boldsymbol{x}, \xi) := \int_0^\xi \mathcal{D}(\boldsymbol{x}, \varrho)\, \mathrm{d}\varrho, \tag{7.2}$$

satisfy $\lim_{\xi \to \pm\infty} \mathcal{K}(\boldsymbol{x}, \xi) = \pm\infty$ *for all* $\boldsymbol{x} \in \Omega$*. Then, for all* $\xi, \zeta \in H_0^1(\Omega)$*,*

$$\|\nabla(\mathcal{K}(\cdot, \xi) - \mathcal{K}(\cdot, \zeta))\| \lesssim \sup_{\varphi \in H_0^1(\Omega)} \frac{\langle \mathcal{R}(\xi) - \mathcal{R}(\zeta), \varphi \rangle}{\|\nabla\varphi\|} \lesssim \|\nabla(\mathcal{K}(\cdot, \xi) - \mathcal{K}(\cdot, \zeta))\|, \tag{7.3}$$

where the hidden constants depend only on Ω *and the constants* τ*,* γ*,* $\mathcal{D}_{\mathrm{M}}$*,* f_{M}*,* K_{m}*,* K_{M}*, and* q_{M}*. Moreover, there exists a unique solution* $u \in L^\infty(\Omega) \cap H_0^1(\Omega)$ *of* (1.1)*.*

Proof. To prove monotonicity of $\mathcal{R}$, we introduce $\boldsymbol{v}(\boldsymbol{x}, \xi) := \boldsymbol{q}(\boldsymbol{x}, \xi) - \sum_{j=1}^d \boldsymbol{e}_j \int_0^\xi \partial_{x_j} \mathcal{D}(\boldsymbol{x}, \varrho)\, \mathrm{d}\varrho$ to rewrite (3.3) using (7.2) and $\nabla \mathcal{K}(\boldsymbol{x}, \xi) = \mathcal{D}(\boldsymbol{x}, \xi) \nabla\xi + \sum_{j=1}^d \boldsymbol{e}_j \int_0^\xi \partial_{x_j} \mathcal{D}(\boldsymbol{x}, \varrho)\, \mathrm{d}\varrho$ as

$$\langle \mathcal{R}(\xi), \varphi \rangle = \tau(\bar{\boldsymbol{K}}(\cdot)(\nabla \mathcal{K}(\cdot, \xi) + \boldsymbol{v}(\cdot, \xi)), \nabla\varphi) + \langle f(\cdot, \xi), \varphi \rangle. \tag{7.4}$$

Dropping the $\boldsymbol{x}$-dependence from the functions f, $\bar{\boldsymbol{K}}$, $\mathcal{K}$, and $\boldsymbol{v}$ for simplicity, we have for any $\xi, \zeta \in H_0^1(\Omega)$,

$$\langle \mathcal{R}(\xi) - \mathcal{R}(\zeta), \varphi \rangle = \langle f(\xi) - f(\zeta), \varphi \rangle + \tau(\bar{\boldsymbol{K}}(\nabla(\mathcal{K}(\xi) - \mathcal{K}(\zeta)) + \boldsymbol{v}(\xi) - \boldsymbol{v}(\zeta)), \nabla\varphi). \tag{7.5}$$

For obtaining the monotonicity bound, we test (7.5) with $\varphi = \mathcal{K}(\xi) - \mathcal{K}(\zeta) = \int_\zeta^\xi \mathcal{D} \overset{(B1)}{\in} H_0^1(\Omega)$ and use the Young inequality (2.2) with $\rho = 2/(\tau\gamma)$, where $\gamma > 0$ is defined in (B4):

$$\begin{aligned}
&\langle \mathcal{R}(\xi) - \mathcal{R}(\zeta),\, \mathcal{K}(\xi) - \mathcal{K}(\zeta) \rangle \\
&\overset{(7.5)}{=} \langle f(\xi) - f(\zeta), \overset{\xi}{\underset{\zeta}{\int}} \mathcal{D} \rangle + \tau \|\bar{\boldsymbol{K}}^{\frac{1}{2}} \nabla(\mathcal{K}(\xi) - \mathcal{K}(\zeta))\|^2 + \tau(\bar{\boldsymbol{K}}(\boldsymbol{v}(\xi) - \boldsymbol{v}(\zeta)), \nabla(\mathcal{K}(\xi) - \mathcal{K}(\zeta))) \\
&\overset{(2.2)}{\geq} \langle f(\xi) - f(\zeta), \overset{\xi}{\underset{\zeta}{\int}} \mathcal{D} \rangle + \left(1 - \frac{\tau\gamma}{4}\right) \tau \|\bar{\boldsymbol{K}}^{\frac{1}{2}} \nabla(\mathcal{K}(\xi) - \mathcal{K}(\zeta))\|^2 - \gamma^{-1} \|\bar{\boldsymbol{K}}^{\frac{1}{2}} (\boldsymbol{v}(\xi) - \boldsymbol{v}(\zeta))\|^2 \\
&\overset{(B4)}{\geq} \left(1 - \frac{\tau\gamma}{4}\right) \tau \|\bar{\boldsymbol{K}}^{\frac{1}{2}} \nabla(\mathcal{K}(\xi) - \mathcal{K}(\zeta))\|^2 \overset{(B3)}{\geq} \left(1 - \frac{\tau\gamma}{4}\right) \tau K_{\mathrm{m}} \|\nabla(\mathcal{K}(\xi) - \mathcal{K}(\zeta))\|^2.
\end{aligned} \tag{7.6}$$

Hence, if $\tau < 4/\gamma$, then we have the lower bound by using $\varphi = \mathcal{K}(\xi) - \mathcal{K}(\zeta)$ in (7.3).

For obtaining the Lipschitz-continuity result, we first define for some $\varepsilon \in (0, \mathcal{D}_{\mathrm{M}})$ (with $\mathcal{D}_{\mathrm{M}} > 0$ from (B1)) and $\xi \in \mathbb{R}$,

$$\mathcal{D}_\varepsilon(\xi) := \max(\varepsilon, \mathcal{D}(\xi)) \text{ and } \mathcal{K}_\varepsilon(\xi) := \int_0^\xi \mathcal{D}_\varepsilon(\varrho)\mathrm{d}\varrho. \tag{7.7a}$$

Since, for $\zeta \in \mathbb{R}$,

$$|\mathcal{K}(\xi) - \mathcal{K}(\zeta)| = \Big|\int_\zeta^\xi \mathcal{D}\Big| \leq \Big|\int_\zeta^\xi \mathcal{D}_\varepsilon\Big| = |\mathcal{K}_\varepsilon(\xi) - \mathcal{K}_\varepsilon(\zeta)| \leq \|\mathcal{D}_\varepsilon\|_{L^\infty}|\xi - \zeta|, \tag{7.7b}$$

we have the following properties of the regularized function $\mathcal{K}_\varepsilon$:

$$\max(\varepsilon|\xi - \zeta|, |\mathcal{K}(\xi) - \mathcal{K}(\zeta)|) \leq |\mathcal{K}_\varepsilon(\xi) - \mathcal{K}_\varepsilon(\zeta)| \overset{(B1)}{\leq} \mathcal{D}_{\mathrm{M}}|\xi - \zeta|. \tag{7.7c}$$

Then, substituting $\mathcal{K}$ by $\mathcal{K}_\varepsilon$ in (7.5), the operator $\mathcal{R}$ still remains monotone following the steps of (7.6). Moreover, repeatedly using the Poincaré inequality (2.1), we obtain

$$\begin{aligned}
&\sup_{\varphi \in H_0^1(\Omega),\, \|\nabla\varphi\|=1} \langle \mathcal{R}(\xi) - \mathcal{R}(\zeta), \varphi\rangle \\
&\overset{(B3)}{\leq} \sup_{\varphi \in H_0^1(\Omega),\, \|\nabla\varphi\|=1} \|f(\xi) - f(\zeta)\|\|\varphi\| + \tau K_{\mathrm{M}}\|\nabla(\mathcal{K}_\varepsilon(\xi) - \mathcal{K}_\varepsilon(\zeta))\| \\
&\qquad + \tau {K_{\mathrm{M}}}^{\frac{1}{2}}\|\bar{\boldsymbol{K}}^{\frac{1}{2}}(\boldsymbol{v}(\xi) - \boldsymbol{v}(\zeta))\| \\
&\overset{(2.1),(B4)}{\leq} C_\Omega\, h_\Omega\|f(\xi) - f(\zeta)\| + \tau K_{\mathrm{M}}\|\nabla(\mathcal{K}_\varepsilon(\xi) - \mathcal{K}_\varepsilon(\zeta))\| \\
&\qquad + \tau(\gamma K_{\mathrm{M}})^{\frac{1}{2}}\|f(\xi) - f(\zeta)\|^{\frac{1}{2}}\|\mathcal{K}(\xi) - \mathcal{K}(\zeta)\|^{\frac{1}{2}} \\
&\overset{(B1),(B2),(7.7b),(2.2)}{\lesssim} \|\xi - \zeta\| + \tau K_{\mathrm{M}}\|\nabla(\mathcal{K}_\varepsilon(\xi) - \mathcal{K}_\varepsilon(\zeta))\| \\
&\overset{(7.7c)}{\leq} \tfrac{1}{\varepsilon}\|\mathcal{K}_\varepsilon(\xi) - \mathcal{K}_\varepsilon(\zeta)\| + \tau K_{\mathrm{M}}\|\nabla(\mathcal{K}_\varepsilon(\xi) - \mathcal{K}_\varepsilon(\zeta))\| \\
&\overset{(2.1)}{\leq} (\tfrac{C_\Omega h_\Omega}{\varepsilon} + \tau K_{\mathrm{M}})\|\nabla(\mathcal{K}_\varepsilon(\xi) - \mathcal{K}_\varepsilon(\zeta))\|.
\end{aligned}$$

Then, the existence of u_ε with $\mathcal{K}_\varepsilon(u_\varepsilon) \in H_0^1(\Omega)$ solving (1.1) follows from the monotonicity and continuity of $\mathcal{R}$ using the nonlinear Lax–Milgram theorem [67, Chapter 2] (see also [68, Chapter 25]). Moreover, $\mathcal{K}_\varepsilon(u_\varepsilon) \in L^\infty(\Omega)$, which is evident from Theorem 13.1 of [47, Chapter 3], since $\boldsymbol{q}$, $\partial_{x_j}\mathcal{D}$, f are all bounded in (7.4). In particular, since the L^∞ bounds on these functions do not depend on ε, the bound on $\|\mathcal{K}_\varepsilon(u_\varepsilon)\|_{L^\infty(\Omega)}$ is also independent of ε. Then, from $|\mathcal{K}_\varepsilon(\xi)| \geq |\mathcal{K}(\xi)|$ (as evident from putting $\zeta = 0$ in (7.7c)), we have that $\|\mathcal{K}(u_\varepsilon)\|_{L^\infty(\Omega)}$ is bounded irrespective of ε. The assumption $\mathcal{K}(\xi) \to \pm\infty$ as $\xi \to \pm\infty$ further implies that there exists a constant $R \in (0, \infty)$ independent of $\varepsilon > 0$ such that $\|u_\varepsilon\|_{L^\infty(\Omega)} \leq R$. Define $\min_{|v| \leq R} \mathcal{D}(v) =: \varepsilon_0 > 0$, the positivity following due to (B1) since $\{|v| \leq R\}$ is compact. Then, for any $0 < \varepsilon < \varepsilon_0$, we have that u_ε also solves the unregularized problem as $\mathcal{D}(u_\varepsilon) \geq \varepsilon_0 > \varepsilon$, and thus, $\mathcal{K}(u_\varepsilon) = \mathcal{K}_\varepsilon(u_\varepsilon)$ from (7.7a). Moreover, $u = u_\varepsilon \in H_0^1(\Omega)$ (as $(\mathcal{K}_\varepsilon)^{-1}$ is Lipschitz from (7.7c)) and it is the unique solution due to the monotonicity of the operator $\mathcal{R}$, see (7.6). □

8 A posteriori error estimates of Theorem 5.7

We show in this section how Theorem 5.3 together with the results on a posteriori error analysis of linear reaction–diffusion problems lead to the a posteriori error estimates of the nonlinear problem (1.1) given in Theorem 5.7.

8.1 Equilibrated flux and potential of Definition 5.4, proof of Lemma 5.6

Let us first give all the details for the construction of Definition 5.4.

8.1.1 Orthogonal projections

Let the standard L^2-orthogonal projections $\Pi_{\ell,p} : L^2(\Omega) \to \mathcal{P}_p(\mathcal{T}_\ell)$ and $\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1} : \boldsymbol{L}^2(\Omega;\mathbb{R}^d) \to \boldsymbol{\mathcal{RT}}_{p-1}(\mathcal{T}_\ell;\mathbb{R}^d)$ be defined as

$$\text{for } \zeta \in L^2(\Omega), \qquad (\Pi_{\ell,p}\zeta, v_\ell) = (\zeta, v_\ell) \qquad \forall v_\ell \in \mathcal{P}_p(\mathcal{T}_\ell), \tag{8.1a}$$

$$\text{for } \boldsymbol{\zeta} \in \boldsymbol{L}^2(\Omega;\mathbb{R}^d), \quad (\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\zeta}, \boldsymbol{v}_\ell) = (\boldsymbol{\zeta}, \boldsymbol{v}_{\boldsymbol{\ell}}) \qquad \forall \boldsymbol{v}_\ell \in \boldsymbol{\mathcal{RT}}_{p-1}(\mathcal{T}_\ell;\mathbb{R}^d). \tag{8.1b}$$

Note that (8.1) can be equivalently rewritten elementwise. We will also need the L^i-weighted projection $\Pi_{\ell,p,L^i} : \mathcal{P}_{p+1}(\mathcal{T}_\ell) \to \mathcal{P}_p(\mathcal{T}_\ell)$

$$\text{for } \zeta_\ell \in \mathcal{P}_{p+1}(\mathcal{T}_\ell), \quad (L^i \Pi_{\ell,p,L^i}\zeta_\ell, v_\ell) = (L^i\zeta_\ell, v_\ell) \qquad \forall v_\ell \in \mathcal{P}_p(\mathcal{T}_\ell), \tag{8.1c}$$

defined again elementwise and meaning, for all $K \in \mathcal{T}_\ell$,

$$\begin{aligned} \Pi_{\ell,p,L^i}\zeta_\ell|_K &= \zeta_\ell|_K \qquad \text{when } L^i \text{ is zero on } K,\\ (L^i\Pi_{\ell,p,L^i}\zeta_\ell, v_\ell)_K &= (L^i\zeta_\ell, v_\ell)_K \qquad \forall v_\ell \in \mathcal{P}_p(K) \text{ otherwise.} \end{aligned}$$

This last projection only serves to describe data oscillations. When L^i is elementwise constant, then $\Pi_{\ell,p,L^i} = \Pi_{\ell,p}$.

8.1.2 Trace and inverse inequalities

For all $K \in \mathcal{T}_\ell$ there exists a constant $C_{\mathrm{Tr}} > 0$ depending only on the space dimension $d \geq 1$ and shape-regularity of the mesh $\mathcal{T}_\ell$ such that the there holds

$$\textbf{Trace inequality:} \ \|v\|_{\partial K} \leq C_{\mathrm{Tr}} \|\nabla v\|_K^{\frac{1}{2}} \|v\|_K^{\frac{1}{2}} \quad \forall v \in H^1(K) \text{ with } \int_K v = 0. \tag{8.2a}$$

Similarly, for any $\boldsymbol{v} \in \boldsymbol{\mathcal{RT}}_p(K)$, there exist constants $C_{\mathrm{inv},p}$, $C_{\mathrm{inv},p,\partial} > 0$ which depend only on d, the shape-regularity of the mesh, and the polynomial degree p such that

$$\textbf{Inverse inequalities:} \ \begin{cases} h_K \|\nabla{\cdot}\boldsymbol{v}\|_K \leq C_{\mathrm{inv},p}\|\boldsymbol{v}\|_K,\\ h_K^{\frac{1}{2}}\|\boldsymbol{v}{\cdot}\boldsymbol{n}\|_{\partial K} \leq C_{\mathrm{inv},p,\partial}\|\boldsymbol{v}\|_K. \end{cases} \tag{8.2b}$$

Possible expressions of the constants in (8.2) are provided in Section 2.2 of [60], along with the proofs of these statements.

8.1.3 Weights

As shown in [60, 64], in order to obtain robust reliability and efficiency estimates for singularly perturbed linear reaction–diffusion problems (for large values of L^i), weights need to be introduced. Recalling (4.5), we define a piecewise constant function w^i

$$\mathsf{w}^i|_K := \min\left(1, \frac{C_\star}{h_K^{\frac{1}{2}}}\left(\frac{a^i_{\mathrm{m}}|_K}{L^i_{\mathrm{m}}|_K}\right)^{\frac{1}{4}}\right), \qquad K \in \mathcal{T}_\ell, \tag{8.3a}$$

where, with reference to constants introduced in (8.2),

$$C_\star := \frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{\pi}} C_{\mathrm{inv},p} + C_{\mathrm{Tr}}\, C_{\mathrm{inv},p,\partial}\right).$$

To treat data oscillation terms, the usual cut-off factor [64] for $K \in \mathcal{T}_\ell$ is

$$\widetilde{\mathsf{w}}^i_K := \min\left(\frac{1}{(L^i_{\mathrm{m}}|_K)^{\frac{1}{2}}}, \frac{h_K}{\pi(a^i_{\mathrm{m}}|_K)^{\frac{1}{2}}}\right). \tag{8.3b}$$

8.1.4 Shorhand notations

It will be useful to use below the shortened notation

$$\boldsymbol{\tau}^{i+1} := \mathfrak{a}^i \nabla u_\ell^{i+1} + \boldsymbol{\mathcal{F}}^i. \tag{8.4}$$

We will also employ, for $K \in \mathcal{T}_\ell$, as in (4.19a),

$$|\!|\!|\varphi|\!|\!|_{i,K} := \left((L^i \varphi, \varphi)_K + (\mathfrak{a}^i \nabla\varphi, \nabla\varphi)_K\right)^{\frac{1}{2}}. \tag{8.5}$$

Finally, for a mesh element $K \in \mathcal{T}_\ell$, we denote its set of vertices by $\mathcal{V}_K \subset \mathcal{V}_\ell$.

8.1.5 Definition 5.4

The existence and uniqueness of $\boldsymbol{\sigma}_{\boldsymbol{a}}^{i+1}$ and $\phi_{\boldsymbol{a}}^{i+1}$ from Definition 5.4 are standard following [7], see also the discussion after [60, Definition 2.3]. Problem (5.7) is of reaction–diffusion type when $L^i|_K > 0$ for at least one $K \subset \omega_{\boldsymbol{a}}$. When $L^i|_{\omega_{\boldsymbol{a}}} = 0$, it reduces to a pure diffusion equilibration, see Remarks 4.2 and 5.5. Note that, recalling the notation (8.4), the source term in (5.7) satisfies

$$-(\Pi_{\ell,p}(\psi_{\boldsymbol{a}}\mathcal{S}^i) + \nabla\psi_{\boldsymbol{a}}{\cdot}\boldsymbol{\Pi}_{\ell,p-1}^{\mathrm{RT}}\boldsymbol{\tau}^{i+1}, 1)_{\omega_{\boldsymbol{a}}} \overset{(8.1)}{=} -(\mathcal{S}^i, \psi_{\boldsymbol{a}})_{\omega_{\boldsymbol{a}}} - (\boldsymbol{\tau}^{i+1}, \nabla\psi_{\boldsymbol{a}})_{\omega_{\boldsymbol{a}}} \overset{(4.9)}{=} (L^i\, u_\ell^{i+1}, \psi_{\boldsymbol{a}})_{\omega_{\boldsymbol{a}}}. \tag{8.6}$$

Let $L^i|_{\omega_{\boldsymbol{a}}} = 0$. Then, when $\boldsymbol{a} \in \mathcal{V}_\ell^{\mathrm{int}}$ is an interior vertex, (5.6b) leads to

$$(\nabla{\cdot}\boldsymbol{\sigma}_{\boldsymbol{a}}^{i+1}, 1)_{\omega_{\boldsymbol{a}}} = (\boldsymbol{\sigma}_{\boldsymbol{a}}^{i+1}{\cdot}\boldsymbol{n}, 1)_{\partial\omega_{\boldsymbol{a}}} = 0, \tag{8.7}$$

and we have the necessary Neumann compatibility condition, the zero mean value of the source term in (5.7), from (8.6). Here, actually, $\phi_{\boldsymbol{a}}^{i+1}$ becomes undefined, and we tacitly replace it by $\Pi_{\ell,p}(\psi_{\boldsymbol{a}}\, u_\ell^{i+1})$, so that $\phi_\ell^{i+1} = u_\ell^{i+1}$. Anyhow, shall L^i be zero everywhere, this is consistent since $\eta_{\mathrm{S},\omega}^i$ of (5.12b) features $(L^i)^{\frac{1}{2}}(u_\ell^{i+1} - \phi_\ell^{i+1}) = 0$.

Practically, $\boldsymbol{\sigma}_{\boldsymbol{a}}^{i+1}$ and $\phi_{\boldsymbol{a}}^{i+1}$ are computed by solving the following Euler–Lagrange equations: find $(\boldsymbol{\sigma}_{\boldsymbol{a}}^{i+1}, \phi_{\boldsymbol{a}}^{i+1}) \in \boldsymbol{V}_{\boldsymbol{a}} \times Q_{\boldsymbol{a}}$ such that for all $(\boldsymbol{v}_\ell, q_\ell) \in \boldsymbol{V}_{\boldsymbol{a}} \times Q_{\boldsymbol{a}}$,

$$\begin{aligned}
(\boldsymbol{\sigma}_{\boldsymbol{a}}^{i+1}, (\mathsf{w}^i)^2(a_{\mathrm{m}}^i)^{-1}\boldsymbol{v}_\ell)_{\omega_{\boldsymbol{a}}} - (\phi_{\boldsymbol{a}}^{i+1}, \nabla{\cdot}\boldsymbol{v}_\ell)_{\omega_{\boldsymbol{a}}} &= -(\psi_{\boldsymbol{a}}\boldsymbol{\Pi}_{\ell,p-1}^{\mathrm{RT}}\boldsymbol{\tau}^{i+1}, (\mathsf{w}^i)^2(a_{\mathrm{m}}^i)^{-1}\boldsymbol{v}_\ell)_{\omega_{\boldsymbol{a}}} \\
&\quad - (\psi_{\boldsymbol{a}} u_\ell^{i+1}, \nabla{\cdot}\boldsymbol{v}_\ell)_{\omega_{\boldsymbol{a}}}, \\
(\nabla{\cdot}\boldsymbol{\sigma}_{\boldsymbol{a}}^{i+1}, q_\ell)_{\omega_{\boldsymbol{a}}} + (L^i\phi_{\boldsymbol{a}}^{i+1}, q_\ell)_{\omega_{\boldsymbol{a}}} &= -(\psi_{\boldsymbol{a}}\mathcal{S}^i + \nabla\psi_{\boldsymbol{a}}{\cdot}\boldsymbol{\Pi}_{\ell,p-1}^{\mathrm{RT}}\boldsymbol{\tau}^{i+1}, q_\ell)_{\omega_{\boldsymbol{a}}}.
\end{aligned}$$

8.1.6 Proof of Lemma 5.6

We have discussed the existence and uniqueness of $\boldsymbol{\sigma}_{\boldsymbol{a}}^{i+1}$ and $\phi_{\boldsymbol{a}}^{i+1}$ in Section 8.1.5 above. The inclusions (5.9) then follow immediately from (5.8) together with the no-flow boundary conditions imposed in (5.6b).

For any vertex $\boldsymbol{a} \in \mathcal{V}$, the minimization constraint in (5.7) and property (8.6) yield

$$(\nabla{\cdot}\boldsymbol{\sigma}_{\boldsymbol{a}}^{i+1}, 1)_{\omega_{\boldsymbol{a}}} + (L^i\, \phi_{\boldsymbol{a}}^{i+1}, 1)_{\omega_{\boldsymbol{a}}} \overset{(5.7)}{=} -(\Pi_{\ell,p}(\psi_{\boldsymbol{a}}\mathcal{S}^i) + \nabla\psi_{\boldsymbol{a}}{\cdot}\boldsymbol{\Pi}_{\ell,p-1}^{\mathrm{RT}}\boldsymbol{\tau}^{i+1}, 1)_{\omega_{\boldsymbol{a}}} \overset{(8.6)}{=} (L^i\, u_\ell^{i+1}, \psi_{\boldsymbol{a}})_{\omega_{\boldsymbol{a}}}. \tag{8.8}$$

Thus, using (8.7) for an interior vertex $\boldsymbol{a} \in \mathcal{V}_\ell^{\mathrm{int}}$, we conclude the mean value property (5.10).

Finally, the minimization constraint in (5.7) together with the partition of unity

$$\sum_{\boldsymbol{a}\in\mathcal{V}_\ell} \psi_{\boldsymbol{a}} = 1 \tag{8.9}$$

and (5.8) imply (5.11).

8.2 Proof of global reliability estimate in Theorem 5.7

Recalling Theorem 5.3, we need to show that

$$\left|\!\left|\!\left| \mathcal{R}^{u_\ell^i}_{\mathrm{disc}}(u_\ell^{i+1}) \right|\!\right|\!\right|^2_{-1,u_\ell^i} \le \sum_{K\in\mathcal{T}_\ell} \left(\eta^i_{\mathrm{F},K} + \eta^i_{\mathrm{S},K} + \eta^i_{\mathrm{osc},K} + \eta^i_{\mathrm{quad,F},K}\right)^2. \tag{8.10}$$

Recall the definition of $\mathcal{R}^{u_\ell^i}_{\mathrm{disc}}$ from (5.1) and the norms $|||\cdot|||_{\pm 1,u_\ell^i}$ from (4.15), (4.18), and let $\varphi \in H^1_0(\Omega)$ be fixed. Using (5.2), as well as (5.9), the Green theorem, and (5.11), we have

$$\begin{aligned}
\langle \mathcal{R}^{u_\ell^i}_{\mathrm{disc}}(u_\ell^{i+1}), \varphi\rangle &\overset{(5.2)}{=} (L^i\, u_\ell^{i+1} - L^i\, \phi_\ell^{i+1}, \varphi) + (\boldsymbol{\tau}^{i+1} + \boldsymbol{\sigma}_\ell^{i+1}, \nabla\varphi) + (\mathcal{S}^i + L^i\, \phi_\ell^{i+1} + \nabla{\cdot}\boldsymbol{\sigma}_\ell^{i+1}, \varphi)\\
&\overset{(5.11)}{=} (L^i(u_\ell^{i+1} - \phi_\ell^{i+1}), \varphi) + (\mathbf{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{\sigma}_\ell^{i+1}, \nabla\varphi)\\
&\quad + (\boldsymbol{\tau}^{i+1} - \mathbf{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1}, \nabla\varphi) + (\mathcal{S}^i - \Pi_{\ell,p}\mathcal{S}^i, \varphi) + (L^i\, \phi_\ell^{i+1} - \Pi_{\ell,p}(L^i\, \phi_\ell^{i+1}), \varphi)\\
&= \sum_{K\in\mathcal{T}_\ell} [(L^i(u_\ell^{i+1} - \phi_\ell^{i+1}), \varphi)_K + (\mathbf{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{\sigma}_\ell^{i+1}, \nabla\varphi)_K\\
&\quad + ((\mathbb{I} - \mathbf{\Pi}^{\mathrm{RT}}_{\ell,p-1})\boldsymbol{\tau}^{i+1}, \nabla\varphi)_K + ((I - \Pi_{\ell,p})(\mathcal{S}^i + L^i\phi_\ell^{i+1}), \varphi)_K].
\end{aligned} \tag{8.11}$$

We now follow the proof of [60, Theorem 3.1]. Let $K \in \mathcal{T}_\ell$ be fixed. For the third term of (8.11), one directly obtains, using the Cauchy–Schwarz inequality,

$$((\mathbb{I} - \mathbf{\Pi}^{\mathrm{RT}}_{\ell,p-1})\boldsymbol{\tau}^{i+1}, \nabla\varphi)_K \overset{(5.12d)}{\le} \eta^i_{\mathrm{quad,F},K}\|(\mathfrak{a}^i)^{\frac12}\nabla\varphi\|_K \overset{(8.5)}{\le} \eta^i_{\mathrm{quad,F},K}|||\varphi|||_{i,K}. \tag{8.12}$$

Denote $t^i := (I - \Pi_{\ell,p})(\mathcal{S}^i + L^i\phi_\ell^{i+1})$. To estimate the fourth term of (8.11), there are two possibilities. First, using the Cauchy–Schwarz inequality, (4.5a), and (8.5),

$$(t^i, \varphi)_K \le \|(L^i_{\mathrm{m}}|_K)^{-\frac12} t^i\|_K \|(L^i_{\mathrm{m}}|_K)^{\frac12}\varphi\|_K \le (L^i_{\mathrm{m}}|_K)^{-\frac12}\|t^i\|_K|||\varphi|||_{i,K}. \tag{8.13a}$$

Second, since the definition of $\Pi_{\ell,p}$ from (8.1a) allows to subtract a constant on each $K \in \mathcal{T}_\ell$, using again the Cauchy–Schwarz inequality, (4.5b), and (8.5)

$$\begin{aligned}
(t^i, \varphi)_K = (t^i, \varphi - \tfrac{1}{|K|}\textstyle\int_K \varphi)_K &\overset{(2.1)}{\le} \frac{h_K}{\pi}\|t^i\|_K\|\nabla\varphi\|_K\\
&\le \frac{h_K}{\pi(a^i_{\mathrm{m}}|_K)^{\frac12}}\|t^i\|_K\|(a^i_{\mathrm{m}}|_K)^{\frac12}\nabla\varphi\|_K \le \frac{h_K}{\pi(a^i_{\mathrm{m}}|_K)^{\frac12}}\|t^i\|_K|||\varphi|||_{i,K}.
\end{aligned} \tag{8.13b}$$

Overall, using the notation $\widetilde{\mathrm{w}}^i_K$ from (8.3b),

$$(t^i, \varphi)_K \le \widetilde{\mathrm{w}}^i_K\, \|t^i\|_K\, |||\varphi|||_{i,K} \overset{(5.12c)}{=} \eta^i_{\mathrm{osc},K}|||\varphi|||_{i,K}. \tag{8.14}$$

The first term of (8.11) is simply estimated as (note that if L^i is zero, this term vanishes)

$$(L^i(u_\ell^{i+1} - \phi_\ell^{i+1}), \varphi)_K \le \|(L^i)^{\frac12}(u_\ell^{i+1} - \phi_\ell^{i+1})\|_K\|(L^i)^{\frac12}\varphi\|_K \overset{(5.12b),\,(8.5)}{\le} \eta^i_{\mathrm{S},K}|||\varphi|||_{i,K}. \tag{8.15}$$

For the second term of (8.11), one can estimate as above

$$(\mathbf{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{\sigma}_\ell^{i+1}, \nabla\varphi)_K \le \|(a^i_{\mathrm{m}})^{-\frac12}(\mathbf{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{\sigma}_\ell^{i+1})\|_K\|(a^i_{\mathrm{m}})^{\frac12}\nabla\varphi\|_K. \tag{8.16a}$$

However, this is not an optimal bound if reaction is the dominant effect (in a singularly perturbed regime which occurs, e.g., when $\tau \ll 1$ in Table 2). Nevertheless, since $(\mathbf{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{\sigma}_\ell^{i+1})|_K$

is a polynomial lying in $\boldsymbol{\mathcal{RT}}_p(K)$ for all mesh elements K, it is possible to get using the trace and inverse inequalities (8.2) the bound

$$(\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{\sigma}^{i+1}_{\ell}, \nabla\varphi)_K \le \eta^i_{\mathrm{F},K} |||\varphi|||_{i,K}. \tag{8.16b}$$

The proof of this estimate is given in [60, proof of Theorem 3.1], with $(a^i_{\mathrm{m}})^{\frac{1}{2}}$ replaced by ε and $(L^i_{\mathrm{m}})^{\frac{1}{2}}$ by κ, yielding the cut-off $\mathsf{w}^i|_K$ given by (8.3a).

Combining (8.12)–(8.16) in (8.11) yields

$$\langle \mathcal{R}^{u^i_\ell}_{\mathrm{disc}}(u^{i+1}_\ell), \varphi\rangle \le \sum_{K\in\mathcal{T}_\ell} (\eta^i_{\mathrm{F},K} + \eta^i_{\mathrm{S},K} + \eta^i_{\mathrm{osc},K} + \eta^i_{\mathrm{quad,F},K})\, |||\varphi|||_{i,K}. \tag{8.17}$$

Upon using the Cauchy–Schwarz inequality one obtains (8.10) after noting that $|||\varphi|||_{1,u^i_\ell} = (\sum_{K\in\mathcal{T}_\ell} |||\varphi|||^2_{i,K})^{\frac{1}{2}}$. □

8.3 Local efficiency bounds

The efficiency proofs follow the methodology of [60]. We first prove the following lemma:

Lemma 8.1 (Local efficiency bounds)**.** *Under the assumptions of Theorem 5.7, one has*

$$[\eta^i_{\mathrm{disc},\omega}]^2 \lesssim \sum_{\substack{\boldsymbol{a}\in\mathcal{V}_\ell,\\ \omega_{\boldsymbol{a}}\subseteq\widetilde{\omega}}} \left[\vartheta^2_{\boldsymbol{a},\ell,i}\left(\sup_{\varphi\in H^1_0(\omega_{\boldsymbol{a}})} \frac{\langle \mathcal{R}^{u^i_\ell}_{\mathrm{disc}}(u^{i+1}_\ell), \varphi\rangle}{|||\varphi|||_{1,u^i_\ell,\omega_{\boldsymbol{a}}}}\right)^2 \right.$$
$$\left. + [\eta^i_{\mathrm{osc},\omega_{\boldsymbol{a}}}]^2 + \vartheta^2_{\boldsymbol{a},\ell,i}[\eta^i_{\mathrm{quad,F},\omega_{\boldsymbol{a}}}]^2 + [\eta^i_{\mathrm{quad,S},\omega_{\boldsymbol{a}}}]^2 + [\eta^i_{\mathrm{quad,S,osc},\omega_{\boldsymbol{a}}}]^2\right],$$

where the constant in $\lesssim$ depends only on the space dimension d, the shape-regularity constant of $\mathcal{T}_\ell$, and the polynomial degree p.

Proof. Let $\mathcal{V}_{\widetilde{\omega}} \subseteq \mathcal{V}_\ell$ be the collection of vertices of $\boldsymbol{a}\in\mathcal{V}_\ell$ such that $\omega_{\boldsymbol{a}}\subseteq\widetilde{\omega}$. Then, one has

$$\begin{aligned}
[\eta^i_{\mathrm{disc},\omega}]^2 &\overset{(5.13\mathrm{a})}{\lesssim} \sum_{K\in\mathcal{T}_\ell,\, K\subseteq\omega} ([\eta^i_{\mathrm{osc},K}]^2 + [\eta^i_{\mathrm{quad,F},K}]^2) + \sum_{K\in\mathcal{T}_\ell,\, K\subseteq\omega} ([\eta^i_{\mathrm{S},K}]^2 + [\eta^i_{\mathrm{F},K}]^2) \\
&\overset{(5.12)}{\lesssim} \sum_{\boldsymbol{a}\in\mathcal{V}_{\widetilde{\omega}}} ([\eta^i_{\mathrm{osc},\omega_{\boldsymbol{a}}}]^2 + [\eta^i_{\mathrm{quad,F},\omega_{\boldsymbol{a}}}]^2) + \sum_{K\in\mathcal{T}_\ell,\, K\subseteq\omega} \Big[\|(L^i)^{\frac{1}{2}}(u^{i+1}_\ell - \phi^{i+1}_\ell)\|^2_K \\
&\qquad + \|\mathsf{w}^i(a^i_{\mathrm{m}})^{-\frac{1}{2}}(\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{\sigma}^{i+1}_\ell)\|^2_K\Big].
\end{aligned} \tag{8.18a}$$

Then, calling the last summation above T_1, using the linearity of $\Pi_{\ell,p}$ from (8.1a), we have

$$\begin{aligned}
T_1 &\overset{(5.8),(8.9)}{=} \sum_{K\in\mathcal{T}_\ell,\, K\subseteq\omega} \left[\left\|\mathsf{w}^i(a^i_{\mathrm{m}})^{-\frac{1}{2}} \sum_{\boldsymbol{a}\in\mathcal{V}_K} (\psi_{\boldsymbol{a}}\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{\sigma}^{i+1}_{\boldsymbol{a}})\right\|^2_K \right. \\
&\qquad \left. + \left\|(L^i)^{\frac{1}{2}} \sum_{\boldsymbol{a}\in\mathcal{V}_K} (\Pi_{\ell,p}(\psi_{\boldsymbol{a}}u^{i+1}_\ell) - \phi^{i+1}_{\boldsymbol{a}})\right\|^2_K\right] \\
&\lesssim \sum_{\boldsymbol{a}\in\mathcal{V}_{\widetilde{\omega}}} \Big[\|\mathsf{w}^i(a^i_{\mathrm{m}})^{-\frac{1}{2}}(\psi_{\boldsymbol{a}}\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{\sigma}^{i+1}_{\boldsymbol{a}})\|^2_{\omega_{\boldsymbol{a}}} \\
&\qquad + \|(L^i)^{\frac{1}{2}}(\Pi_{\ell,p}(\psi_{\boldsymbol{a}}u^{i+1}_\ell) - \phi^{i+1}_{\boldsymbol{a}})\|^2_{\omega_{\boldsymbol{a}}}\Big].
\end{aligned} \tag{8.18b}$$

The terms inside the summation now correspond to the minimization (5.7) and can be estimated as in the proof of [60, Theorem 4.4] by considering two cases. Let $\boldsymbol{a}\in\mathcal{V}_{\widetilde{\omega}}$ be fixed. We henceforth suppose that $L^i \neq 0$ on $\omega_{\boldsymbol{a}}$, so that $L^i_{\mathrm{M}}|_{\omega_{\boldsymbol{a}}} > 0$ and $Q_{\boldsymbol{a}} = \mathcal{P}_p(\mathcal{T}_{\boldsymbol{a}})$ from (5.6a); the pure diffusion case $L^i = 0$ on $\omega_{\boldsymbol{a}}$ is standard (easier) and treated as in [24, 25] and the references therein.

Diffusion-dominated case ($h^2_{\omega_{\boldsymbol{a}}} \max_{\omega_{\boldsymbol{a}}} L^i_{\mathrm{M}} / \max_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{M}} \leq 1$). We adjust the methodology from [60], proof of Lemma 4.2, Case 2. Let $(\boldsymbol{v}_{\boldsymbol{a}}, q_{\boldsymbol{a}}) \in \boldsymbol{V}_{\boldsymbol{a}} \times Q_{\boldsymbol{a}}$ be given by

$$
\begin{aligned}
q_{\boldsymbol{a}} &:= (\Pi_{\ell,p,L^i}(\psi_{\boldsymbol{a}} u_\ell^{i+1}))|_{\omega_{\boldsymbol{a}}},\\
\boldsymbol{v}_{\boldsymbol{a}} &:= \mathop{\arg\min}_{\substack{\boldsymbol{w}_\ell \in \boldsymbol{V}_{\boldsymbol{a}},\\ \nabla\cdot\boldsymbol{w}_\ell = \underbrace{-\Pi_{\ell,p}(\psi_{\boldsymbol{a}}(\mathcal{S}^i + L^i\, u_\ell^{i+1})) - \nabla\psi_{\boldsymbol{a}}\cdot\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1}}_{=:g_{\boldsymbol{a}}}}} \|\psi_{\boldsymbol{a}}\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{w}_\ell\|_{\omega_{\boldsymbol{a}}}.
\end{aligned}
$$

The above minimization corresponds to a pure diffusion problem and has piecewise polynomial data $\psi_{\boldsymbol{a}}\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} \in \boldsymbol{\mathcal{RT}}_p(\mathcal{T}_{\boldsymbol{a}})$ and $g_{\boldsymbol{a}} \in \mathcal{P}_p(\mathcal{T}_{\boldsymbol{a}})$. Moreover, for any interior vertex $\boldsymbol{a} \in \mathcal{V}^{\mathrm{int}}_\ell$, we see from the second and third equalities in (8.8) that the mean value of the above divergence constraint is zero, $(g_{\boldsymbol{a}}, 1)_{\omega_{\boldsymbol{a}}} = 0$ so that the Neumann compatibility condition for the above $\boldsymbol{v}_{\boldsymbol{a}} \in \boldsymbol{V}_{\boldsymbol{a}}$ is satisfied (recall the zero normal trace constraint from (5.6b)). Using

$$
\Pi_{\ell,p}(L^i q_{\boldsymbol{a}}) = \Pi_{\ell,p}(L^i \psi_{\boldsymbol{a}} u_\ell^{i+1}), \tag{8.19}
$$

the pair $(\boldsymbol{v}_{\boldsymbol{a}}, q_{\boldsymbol{a}})$ crucially satisfies the constraint in (5.7) in that

$$
\nabla\cdot\boldsymbol{v}_{\boldsymbol{a}} + \Pi_{\ell,p}(L^i q_{\boldsymbol{a}}) = -\Pi_{\ell,p}(\psi_{\boldsymbol{a}}\mathcal{S}^i) - \nabla\psi_{\boldsymbol{a}}\cdot\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1}. \tag{8.20}
$$

To see (8.19), note that from (8.1c), $q_{\boldsymbol{a}} \in Q_{\boldsymbol{a}}$ can be equivalently defined by

$$
(L^i q_{\boldsymbol{a}}, q_\ell)_{\omega_{\boldsymbol{a}}} = (L^i \psi_{\boldsymbol{a}} u_\ell^{i+1}, q_\ell)_{\omega_{\boldsymbol{a}}} \qquad \forall q_\ell \in Q_{\boldsymbol{a}}.
$$

In particular

$$
q_{\boldsymbol{a}} = \Pi_{\ell,p}(\psi_{\boldsymbol{a}} u_\ell^{i+1})
$$

when L^i is elementwise constant, cf. the discussion below (8.1c).

Define the quadrature contribution

$$
\eta^i_{\mathrm{quad,S},\boldsymbol{a}} := \|(L^i)^{\frac{1}{2}}(\Pi_{\ell,p}(\psi_{\boldsymbol{a}} u_\ell^{i+1}) - \Pi_{\ell,p,L^i}(\psi_{\boldsymbol{a}} u_\ell^{i+1}))\|_{\omega_{\boldsymbol{a}}}
$$

and note that it vanishes when L^i is elementwise constant. Let

$$
H^1_*(\omega_{\boldsymbol{a}}) := \begin{cases} \{v \in H^1(\omega_{\boldsymbol{a}}),\, (v,1)_{\omega_{\boldsymbol{a}}} = 0\} & \text{if } \boldsymbol{a} \in \mathcal{V}^{\mathrm{int}}_\ell,\\ \{v \in H^1(\omega_{\boldsymbol{a}}),\, v = 0 \text{ on } \partial\omega_{\boldsymbol{a}} \cap \{\psi_{\boldsymbol{a}} > 0\}\} & \text{if } \boldsymbol{a} \in \mathcal{V}^{\mathrm{ext}}_\ell. \end{cases}
$$

Bounding from above the patchwise contribution of T_1 in (8.18b), since $(\boldsymbol{\sigma}^{i+1}_{\boldsymbol{a}}, \phi^{i+1}_{\boldsymbol{a}})$ from (5.7) are the minimizers, we have, also relying on $\mathsf{w}^i \leq 1$ from (8.3a) ($\mathsf{w}^i = 1$ when diffusion dominates),

$$
\begin{aligned}
&\|\mathsf{w}^i(a^i_{\mathrm{m}})^{-\frac{1}{2}}(\psi_{\boldsymbol{a}}\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{\sigma}^{i+1}_{\boldsymbol{a}})\|^2_{\omega_{\boldsymbol{a}}} + \|(L^i)^{\frac{1}{2}}(\Pi_{\ell,p}(\psi_{\boldsymbol{a}} u_\ell^{i+1}) - \phi^{i+1}_{\boldsymbol{a}})\|^2_{\omega_{\boldsymbol{a}}}\\
&\leq \|\mathsf{w}^i(a^i_{\mathrm{m}})^{-\frac{1}{2}}(\psi_{\boldsymbol{a}}\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{v}_{\boldsymbol{a}})\|^2_{\omega_{\boldsymbol{a}}} + [\eta^i_{\mathrm{quad,S},\boldsymbol{a}}]^2\\
&\leq \frac{1}{\min_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{m}}}\|\psi_{\boldsymbol{a}}\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{v}_{\boldsymbol{a}}\|^2_{\omega_{\boldsymbol{a}}} + [\eta^i_{\mathrm{quad,S},\boldsymbol{a}}]^2\\
&\lesssim \frac{1}{\min_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{m}}}\Bigg[\sup_{\varphi \in H^1_*(\omega_{\boldsymbol{a}})} \frac{(g_{\boldsymbol{a}}, \varphi)_{\omega_{\boldsymbol{a}}} - (\psi_{\boldsymbol{a}}\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1}, \nabla\varphi)_{\omega_{\boldsymbol{a}}}}{\|\nabla\varphi\|_{\omega_{\boldsymbol{a}}}}\Bigg]^2 + [\eta^i_{\mathrm{quad,S},\boldsymbol{a}}]^2\\
&= \frac{1}{\min_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{m}}}\Bigg[\sup_{\varphi \in H^1_*(\omega_{\boldsymbol{a}})} \frac{(\Pi_{\ell,p}(\psi_{\boldsymbol{a}}(\mathcal{S}^i + L^i\, u_\ell^{i+1})), \varphi)_{\omega_{\boldsymbol{a}}} + (\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1}, \nabla(\psi_{\boldsymbol{a}}\varphi))_{\omega_{\boldsymbol{a}}}}{\|\nabla\varphi\|_{\omega_{\boldsymbol{a}}}}\Bigg]^2 + [\eta^i_{\mathrm{quad,S},\boldsymbol{a}}]^2\\
&\leq \frac{1}{\min_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{m}}}\Bigg[\sup_{\varphi \in H^1_*(\omega_{\boldsymbol{a}})} \Bigg(\frac{(\mathcal{S}^i + L^i\, u_\ell^{i+1}, \psi_{\boldsymbol{a}}\varphi)_{\omega_{\boldsymbol{a}}} + (\boldsymbol{\tau}^{i+1}, \nabla(\psi_{\boldsymbol{a}}\varphi))_{\omega_{\boldsymbol{a}}}}{|\!|\!|\psi_{\boldsymbol{a}}\varphi|\!|\!|_{1,u^i_\ell,\omega_{\boldsymbol{a}}}} \frac{|\!|\!|\psi_{\boldsymbol{a}}\varphi|\!|\!|_{1,u^i_\ell,\omega_{\boldsymbol{a}}}}{\|\nabla\varphi\|_{\omega_{\boldsymbol{a}}}}\Bigg)\\
&\quad + \sup_{\varphi \in H^1_*(\omega_{\boldsymbol{a}})} \frac{((\Pi_{\ell,p} - I)(\psi_{\boldsymbol{a}}(\mathcal{S}^i + L^i\, u_\ell^{i+1})), \varphi)_{\omega_{\boldsymbol{a}}} + ((\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1} - \mathbb{I})\boldsymbol{\tau}^{i+1}, \nabla(\psi_{\boldsymbol{a}}\varphi))_{\omega_{\boldsymbol{a}}}}{\|\nabla\varphi\|_{\omega_{\boldsymbol{a}}}}\Bigg]^2 + [\eta^i_{\mathrm{quad,S},\boldsymbol{a}}]^2.
\end{aligned}
\tag{8.21a}
$$

The third (crucial) inequality follows from [60, Lemma 4.1], itself evoking [9, Theorem 7] and [25, Corollaries 3.3 and 3.6]. In the last expression in (8.21a), recalling (5.2) and (8.4), the first term, say T_2, features the residual $\langle \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}), \psi_{\boldsymbol{a}}\varphi\rangle$; note also that $\psi_{\boldsymbol{a}}\varphi \in H_0^1(\Omega)$ from the definition of the space $H^1_*(\omega_{\boldsymbol{a}})$. The second term T_3 is then related to quadrature and oscillation errors. We now bound them separately.

We first bound T_2. For this purpose, note that, cf., e.g., [9, estimate below (11)],

$$\|\nabla(\psi_{\boldsymbol{a}}\varphi)\|_{\omega_{\boldsymbol{a}}} \lesssim \|\nabla\varphi\|_{\omega_{\boldsymbol{a}}} \qquad \forall \varphi \in H^1_*(\omega_{\boldsymbol{a}}).$$

Moreover, the Poincaré inequality (2.1) and (4.5) give

$$\begin{aligned} |\!|\!| \psi_{\boldsymbol{a}}\varphi |\!|\!|^2_{1,u_\ell^i,\omega_{\boldsymbol{a}}} &= \|(L^i)^{\frac{1}{2}}(\psi_{\boldsymbol{a}}\varphi)\|^2_{\omega_{\boldsymbol{a}}} + \|(\mathfrak{a}^i)^{\frac{1}{2}}\nabla(\psi_{\boldsymbol{a}}\varphi)\|^2_{\omega_{\boldsymbol{a}}} \\ &\lesssim (h^2_{\omega_{\boldsymbol{a}}} \max_{\omega_{\boldsymbol{a}}} L^i_{\mathrm{M}} + \max_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{M}})\|\nabla(\psi_{\boldsymbol{a}}\varphi)\|^2_{\omega_{\boldsymbol{a}}}. \end{aligned}$$

Now, straightforwardly, under the assumption $h^2_{\omega_{\boldsymbol{a}}} \max_{\omega_{\boldsymbol{a}}} L^i_{\mathrm{M}} / \max_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{M}} \le 1$ of (5.15a),

$$\frac{1}{\min_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{m}}}(h^2_{\omega_{\boldsymbol{a}}} \max_{\omega_{\boldsymbol{a}}} L^i_{\mathrm{M}} + \max_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{M}}) \le \left(h^2_{\omega_{\boldsymbol{a}}} \frac{\max_{\omega_{\boldsymbol{a}}} L^i_{\mathrm{M}}}{\max_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{M}}} + 1 \right) \frac{\max_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{M}}}{\min_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{m}}} \le 2 \frac{\max_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{M}}}{\min_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{m}}}.$$

Consequently,

$$\begin{aligned} T_2 &= \frac{1}{\min_{\omega_{\boldsymbol{a}}}(a^i_{\mathrm{m}})^{\frac{1}{2}}} \sup_{\varphi \in H^1_*(\omega_{\boldsymbol{a}})} \left(\frac{\langle \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}), \psi_{\boldsymbol{a}}\varphi\rangle}{|\!|\!| \psi_{\boldsymbol{a}}\varphi |\!|\!|_{1,u_\ell^i,\omega_{\boldsymbol{a}}}} \frac{|\!|\!| \psi_{\boldsymbol{a}}\varphi |\!|\!|_{1,u_\ell^i,\omega_{\boldsymbol{a}}}}{\|\nabla\varphi\|_{\omega_{\boldsymbol{a}}}} \right) \\ &\lesssim \vartheta_{\boldsymbol{a},\ell,i} \sup_{\varphi \in H^1_0(\omega_{\boldsymbol{a}})} \frac{\langle \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}), \varphi\rangle}{|\!|\!| \varphi |\!|\!|_{1,u_\ell^i,\omega_{\boldsymbol{a}}}}. \end{aligned} \tag{8.21b}$$

As for T_3, note that, using the above bound $\|\nabla(\psi_{\boldsymbol{a}}\varphi)\|_{\omega_{\boldsymbol{a}}} \lesssim \|\nabla\varphi\|_{\omega_{\boldsymbol{a}}}$,

$$\begin{aligned} &\frac{1}{\min_{\omega_{\boldsymbol{a}}}(a^i_{\mathrm{m}})^{\frac{1}{2}}}((\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1} - \mathbb{I})\boldsymbol{\tau}^{i+1}, \nabla(\psi_{\boldsymbol{a}}\varphi))_{\omega_{\boldsymbol{a}}} \\ &\overset{(5.15a)}{\le} \vartheta_{\boldsymbol{a},\ell,i}\|(\mathfrak{a}^i)^{-\frac{1}{2}}(\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1} - \mathbb{I})\boldsymbol{\tau}^{i+1}\|_{\omega_{\boldsymbol{a}}}\|\nabla(\psi_{\boldsymbol{a}}\varphi)\|_{\omega_{\boldsymbol{a}}} \overset{(5.12d)}{\lesssim} \vartheta_{\boldsymbol{a},\ell,i}\eta^i_{\mathrm{quad,F},\omega_{\boldsymbol{a}}}\|\nabla\varphi\|_{\omega_{\boldsymbol{a}}}. \end{aligned}$$

Let

$$\eta^i_{\mathrm{quad,S,osc},\boldsymbol{a}} := \frac{h_{\omega_{\boldsymbol{a}}}}{\min_{\omega_{\boldsymbol{a}}}(a^i_{\mathrm{m}})^{\frac{1}{2}}}\|(\Pi_{\ell,p} - I)(\psi_{\boldsymbol{a}}(\mathcal{S}^i + L^i\, u_\ell^{i+1}))\|_{\omega_{\boldsymbol{a}}}.$$

Since $((\Pi_{\ell,p} - I)\zeta, \varphi)_{\omega_{\boldsymbol{a}}} = ((\Pi_{\ell,p} - I)\zeta, \varphi - \Pi_{\ell,p}\varphi)_{\omega_{\boldsymbol{a}}}$ for $\zeta \in L^2(\omega_{\boldsymbol{a}})$, the Poincaré inequality (2.1) gives

$$\frac{1}{\min_{\omega_{\boldsymbol{a}}}(a^i_{\mathrm{m}})^{\frac{1}{2}}}((\Pi_{\ell,p} - I)(\psi_{\boldsymbol{a}}(\mathcal{S}^i + L^i\, u_\ell^{i+1})), \varphi)_{\omega_{\boldsymbol{a}}} \overset{(2.1)}{\lesssim} \eta^i_{\mathrm{quad,S,osc},\boldsymbol{a}}\|\nabla\varphi\|_{\omega_{\boldsymbol{a}}}.$$

Thus, overall,

$$T_3 \lesssim \vartheta_{\boldsymbol{a},\ell,i}\eta^i_{\mathrm{quad,F},\omega_{\boldsymbol{a}}} + \eta^i_{\mathrm{quad,S,osc},\boldsymbol{a}}. \tag{8.21c}$$

The claim of Lemma 8.1 in this case follows by combining the inequalities in (8.18) and (8.21).

Reaction-dominated case ($h^2_{\omega_{\boldsymbol{a}}} \max_{\omega_{\boldsymbol{a}}} L^i_{\mathrm{M}} / \max_{\omega_{\boldsymbol{a}}} a^i_{\mathrm{M}} > 1$). We adjust the methodology from [60], proof of Lemma 4.2, Case 1. Hence, one chooses $(\boldsymbol{v}_{\boldsymbol{a}}, q_{\boldsymbol{a}}) \in \boldsymbol{V}_{\boldsymbol{a}} \times Q_{\boldsymbol{a}}$ as

$$(L^i q_{\boldsymbol{a}}, q_\ell)_{\omega_{\boldsymbol{a}}} = (-\Pi_{\ell,p}(\psi_{\boldsymbol{a}}\mathcal{S}^i) + \psi_{\boldsymbol{a}}\nabla\cdot_\ell \boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} - \rho_{\boldsymbol{a}}, q_\ell)_{\omega_{\boldsymbol{a}}} \qquad \forall q_\ell \in Q_{\boldsymbol{a}}, \tag{8.22a}$$

$$\boldsymbol{v}_{\boldsymbol{a}} := \arg\min_{\substack{\boldsymbol{w}_\ell \in \boldsymbol{V}_{\boldsymbol{a}}, \\ \nabla\cdot\boldsymbol{w}_\ell = g_{\boldsymbol{a}}}} \|(\psi_{\boldsymbol{a}}\boldsymbol{\Pi}^{\mathrm{RT}}_{\ell,p-1}\boldsymbol{\tau}^{i+1} + \boldsymbol{w}_\ell)\|_{\omega_{\boldsymbol{a}}}, \tag{8.22b}$$

where $\nabla\cdot_\ell$ is the elementwise divergence and

$$\rho_{\boldsymbol{a}} := \frac{1}{|\omega_{\boldsymbol{a}}|}(-\mathcal{S}^i - L^i u_\ell^{i+1} + \nabla\cdot_\ell \boldsymbol{\Pi}_{\ell,p-1}^{\mathrm{RT}}\boldsymbol{\tau}^{i+1}, \psi_{\boldsymbol{a}})_{\omega_{\boldsymbol{a}}},$$
$$g_{\boldsymbol{a}} := -\nabla\cdot_\ell(\psi_{\boldsymbol{a}}\boldsymbol{\Pi}_{\ell,p-1}^{\mathrm{RT}}\boldsymbol{\tau}^{i+1}) + \rho_a.$$

Observe that, $(g_{\boldsymbol{a}}, 1)_{\omega_{\boldsymbol{a}}} = 0$ using (4.9); we also need (8.1b) and $\nabla\psi_{\boldsymbol{a}} \in \boldsymbol{\mathcal{RT}}_0(\mathcal{T}_\ell;\mathbb{R}^d)$. As in (8.20), we also have

$$\nabla\cdot\boldsymbol{v}_{\boldsymbol{a}} + \Pi_{\ell,p}(L^i q_{\boldsymbol{a}}) = -\Pi_\ell(\psi_{\boldsymbol{a}}\mathcal{S}^i) - \nabla\psi_{\boldsymbol{a}}\cdot\boldsymbol{\Pi}_{\ell,p-1}^{\mathrm{RT}}\boldsymbol{\tau}^{i+1}.$$

Hence, an upper bound of the patch-wise terms in (8.18b) can be obtained from Definition 5.4 with this particular choice of $(\boldsymbol{v}_{\boldsymbol{a}}, q_{\boldsymbol{a}})$. The proof follows [60, proof of Lemma 4.2, Case 1], itself relying on [64]. □

8.4 Proof of efficiency estimates in Theorem 5.7

We can now accomplish the proof of the efficiency bounds in Theorem 5.7. Observe that $\vartheta_{\boldsymbol{a},\ell,i}^2 \le \vartheta_{\widetilde{\omega},\ell,i}^2$ by definition and

$$\sum_{\substack{\boldsymbol{a}\in\mathcal{V}_\ell,\\ \omega_{\boldsymbol{a}}\subseteq\widetilde{\omega}}} ([\eta_{\mathrm{osc},\omega_{\boldsymbol{a}}}^i]^2 + \vartheta_{\boldsymbol{a},\ell,i}^2[\eta_{\mathrm{quad,F},\omega_{\boldsymbol{a}}}^i]^2 + [\eta_{\mathrm{quad,S},\omega_{\boldsymbol{a}}}^i]^2 + [\eta_{\mathrm{quad,S,osc},\omega_{\boldsymbol{a}}}^i]^2)$$
$$\lesssim [\eta_{\mathrm{osc},\widetilde{\omega}}^i]^2 + \vartheta_{\widetilde{\omega},\ell,i}^2[\eta_{\mathrm{quad,F},\widetilde{\omega}}^i]^2 + [\eta_{\mathrm{quad,S},\widetilde{\omega}}^i]^2 + [\eta_{\mathrm{quad,S,osc},\widetilde{\omega}}^i]^2.$$

Hence, the proof of (5.14b) is completed from Lemma 8.1 by showing that

$$\sum_{\substack{\boldsymbol{a}\in\mathcal{V}_\ell\\ \omega_{\boldsymbol{a}}\subseteq\widetilde{\omega}}} \left(\sup_{\varphi\in H_0^1(\omega_{\boldsymbol{a}})} \frac{\langle\mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}), \varphi\rangle}{|\!|\!|\varphi|\!|\!|_{1,u_\ell^i,\omega_{\boldsymbol{a}}}}\right)^2 \lesssim \left|\!\left|\!\left|\mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1})\right|\!\right|\!\right|_{-1,u_\ell^i,\widetilde{\omega}}^2, \tag{8.23}$$

followed by adding $[\eta_{\mathrm{lin},\omega}^i]^2$ to both sides of the inequality (see the definition of η_ω^i in (5.13c)).

To show (8.23), we proceed as in [16, Theorem 3.3] and the references therein. For each $\boldsymbol{a} \in \mathcal{V}_\ell$ such that $\omega_{\boldsymbol{a}} \subseteq \widetilde{\omega}$, let us consider the Riesz representer $r_{\boldsymbol{a}} \in H_0^1(\omega_{\boldsymbol{a}}) \subseteq H_0^1(\widetilde{\omega})$ satisfying for all $\varphi \in H_0^1(\omega_{\boldsymbol{a}})$,

$$(\!(r_{\boldsymbol{a}}, \varphi)\!)_{u_\ell^i} = \langle\mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}), \varphi\rangle, \qquad \text{so that} \quad \sup_{\varphi\in H_0^1(\omega_{\boldsymbol{a}})} \frac{\langle\mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}), \varphi\rangle}{|\!|\!|\varphi|\!|\!|_{1,u_\ell^i,\omega_{\boldsymbol{a}}}} = |\!|\!|r_{\boldsymbol{a}}|\!|\!|_{1,u_\ell^i,\omega_{\boldsymbol{a}}}. \tag{8.24}$$

Then, defining $r := \sum_{\boldsymbol{a}\in\mathcal{V}_\ell,\,\omega_{\boldsymbol{a}}\subseteq\widetilde{\omega}} r_{\boldsymbol{a}} \in H_0^1(\widetilde{\omega})$, one has

$$\sum_{\substack{\boldsymbol{a}\in\mathcal{V}_\ell\\ \omega_{\boldsymbol{a}}\subseteq\widetilde{\omega}}} |\!|\!|r_{\boldsymbol{a}}|\!|\!|_{1,u_\ell^i,\omega_{\boldsymbol{a}}}^2 = \sum_{\substack{\boldsymbol{a}\in\mathcal{V}_\ell\\ \omega_{\boldsymbol{a}}\subseteq\widetilde{\omega}}} \langle\mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}), r_{\boldsymbol{a}}\rangle = \langle\mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}), r\rangle$$
$$\le \left|\!\left|\!\left|\mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1})\right|\!\right|\!\right|_{-1,u_\ell^i,\widetilde{\omega}} |\!|\!|r|\!|\!|_{1,u_\ell^i,\widetilde{\omega}} \le (d+1)^{\frac{1}{2}} \left|\!\left|\!\left|\mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1})\right|\!\right|\!\right|_{-1,u_\ell^i,\widetilde{\omega}} \left(\sum_{\substack{\boldsymbol{a}\in\mathcal{V}_\ell\\ \omega_{\boldsymbol{a}}\subseteq\widetilde{\omega}}} |\!|\!|r_{\boldsymbol{a}}|\!|\!|_{1,u_\ell^i,\omega_{\boldsymbol{a}}}^2\right)^{\frac{1}{2}},$$

Finally, the global efficiency estimate (5.14c) follows from choosing $\omega = \Omega$ in (5.14b) together with $\left|\!\left|\!\left|\mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1})\right|\!\right|\!\right|_{-1,u_\ell^i}^2 + [\eta_{\mathrm{lin},\Omega}^i]^2 = |\!|\!|\mathcal{R}(u_\ell^{i+1})|\!|\!|_{-1,u_\ell^i}^2$ from Theorem 5.3 and $\vartheta_{\Omega,\ell,i} \ge 1$ from (5.15).

Appendices

Appendix A Iterative linearizations in abstract spaces: well-posedness and consistency

Here, we generalize the iterative linearization defined in Sections 4.3 and 4.4 to an abstract setting, and show well-posedness and consistency of the iterations. With reference to the classes of problems discussed in Section 3, let us introduce the space

$$\mathcal{W}_q := \begin{cases} W_0^{1,q}(\Omega) & \text{for gradient-dependent diffusivity case (3.1)}, \\ L^\infty(\Omega) \cap W_0^{1,q}(\Omega) & \text{for gradient-independent diffusivity case (3.3)}, \end{cases} \tag{A.1}$$

for $q \in [2, \infty]$. Observe that the finite element space $\mathcal{V}_\ell \subset \mathcal{W}_q$ due to (4.2). We equip $\mathcal{W}_q$ with the norm $\|v\|_{\mathcal{W}_q} := \|\nabla v\|_{L^q(\Omega)}$ for (3.1), and $\|v\|_{\mathcal{W}_q} := \|\nabla v\|_{L^q(\Omega)} + \|v\|_{L^\infty(\Omega)}$ for (3.3).

In extension of (4.4), for a given parameter function $v \in \mathcal{W}_2$, let

$$(\!(\xi,\, \zeta)\!)_v := (L(\cdot, v)\, \xi, \zeta) + (\mathfrak{a}(\cdot, v)\nabla\xi, \nabla\zeta), \qquad \xi, \zeta \in H_0^1(\Omega). \tag{A.2}$$

Similarly, extending (4.15) and (4.18), let

$$|\!|\!|\xi|\!|\!|_{1,v} := (\!(\xi,\, \xi)\!)_v^{\frac{1}{2}}, \qquad \xi \in H_0^1(\Omega), \tag{A.3a}$$

$$|\!|\!|\varsigma|\!|\!|_{-1,v} := \sup_{\varphi \in H_0^1(\Omega)} \frac{\langle \varsigma, \varphi\rangle}{|\!|\!|\varphi|\!|\!|_{1,v}}, \quad \varsigma \in H^{-1}(\Omega). \tag{A.3b}$$

Generalizing Assumption 4.1, here we assume

Assumption A.1 (Coefficients L and $\mathfrak{a}$). *The function $L : \Omega \times \mathbb{R} \to \mathbb{R}$ and the symmetric tensor $\mathfrak{a} : \Omega \times \mathbb{R} \to \mathbb{R}^d \times \mathbb{R}^d$ are measurable and satisfy for continuous functions $L_{\rm m}$, $L_{\rm M} : [0, \infty) \to [0, \infty)$ and $a_{\rm m}$, $a_{\rm M} : [0, \infty) \to (0, \infty)$, for almost every $\boldsymbol{x} \in \Omega$, and for all $v \in \mathcal{W}_2$,*

$$0 \le L_{\rm m}(\|v\|_{\mathcal{W}_2}) \le L(\boldsymbol{x}, v(\boldsymbol{x})) \le L_{\rm M}(\|v\|_{\mathcal{W}_2}), \tag{A.4a}$$

$$0 < a_{\rm m}(\|v\|_{\mathcal{W}_2})|\boldsymbol{y}|^2 \le \boldsymbol{y}^{\rm T}\mathfrak{a}(\boldsymbol{x}, v(\boldsymbol{x}))\, \boldsymbol{y} \le a_{\rm M}(\|v\|_{\mathcal{W}_2})|\boldsymbol{y}|^2 \qquad \forall \boldsymbol{y} \in \mathbb{R}^d \setminus \{0\}. \tag{A.4b}$$

The functions $L_{\rm m/M}$ and $a_{\rm m/M}$ can be completely independent of the argument, for instance for the Zarantonello scheme of Table 1. Here, we adhere to a more broad setting of (A.4).

We show that the linearization scheme (4.7) is a special case of a more general iterative linearization strategy involving the inner product $(\!(\cdot,\, \cdot)\!)_v$ from (A.2), and we also show that it is a well-posed and consistent strategy. For this purpose, let us introduce

Definition A.2 (Linearization operator $\mathcal{I}$). *Let $(\!(\cdot,\, \cdot)\!)_v$ be defined in (A.2), with the coefficients L and $\mathfrak{a}$ satisfying Assumption A.1. Let $q \ge 2$ be fixed. Then, we define the linearization operator $\mathcal{I} : \mathcal{W}_q \to \mathcal{W}_q$ as: for a given $v \in \mathcal{W}_q$, $\mathcal{I}v \in \mathcal{W}_q$ solves*

$$(\!(\mathcal{I}v - v,\, \varphi)\!)_v = -\langle \mathcal{R}(v), \varphi\rangle \qquad \forall \varphi \in H_0^1(\Omega). \tag{A.5}$$

Here $\mathcal{I}v - v$ is the increment in v towards the weak solution $u \in H_0^1(\Omega)$ of (1.1), and $\langle \mathcal{R}(v), \varphi\rangle$ defines the residual relation with respect to v. With reference to (4.7) and (4.13), $u_{\langle\ell\rangle}^{i+1} = \mathcal{I}u_\ell^i$. The operator $\mathcal{I}$ extends the linear iteration in a continuous level for some initial guess $u_0 \in \mathcal{W}_q$, i.e. $u^i = (\mathcal{I})^i u_0$ for $i \ge 1$. We need to additionally assume here that the space dimension is limited to $1 \le d \le 3$ to ensure its well-posedness:

Proposition A.3 (Well-posedness of the operator $\mathcal{I}$). *Assume (A1)–(A2) if $\mathcal{R} : H_0^1(\Omega) \to H^{-1}(\Omega)$ is defined by (3.1), and (B1)–(B5) if $\mathcal{R}$ is defined by (3.3). For spatial dimension $d \in \{1, 2, 3\}$, let a finite $q \geq 2$ satisfy $d < q \leq 2\max(2, d)/(\max(2, d) - 2)$. Let the space $\mathcal{W}_q$ be defined in (A.1), and let $(\!(\cdot, \cdot)\!)_v$ be defined in (A.2), with the coefficients L and $\mathfrak{a}$ satisfying Assumption A.1. Then, the linearization operator $\mathcal{I} : \mathcal{W}_q \to \mathcal{W}_q$ of Definition A.2 is well-defined.*

Proof. Let $v \in \mathcal{W}_q$ be fixed. Observe that this implies $v \in \mathcal{W}_2$ since $\mathcal{W}_q \subseteq \mathcal{W}_2$ as $q \geq 2$, and in particular $v \in H_0^1(\Omega)$. Consequently, in view of Assumption A.1, $(\!(\cdot, \cdot)\!)_v$ is a $H_0^1(\Omega)$-inner product. Then, existence and uniqueness of $\mathcal{I}v \in H_0^1(\Omega)$ follow from the Riesz representation theorem, and we are left to show that $\mathcal{I}v \in \mathcal{W}_q$.

First, let $q = 2$. By the assumption $d < q$, this only arises when $d = 1$. In this case, cf., e.g., [26, Chapter 5.6], $H_0^1(\Omega) \subset C(\bar{\Omega}) \subset L^\infty(\Omega)$, so that indeed $\mathcal{I}v \in \mathcal{W}_q$.

Second, let $q > \max\{2, d\}$ be finite and consider the expression (3.1) of $\mathcal{R}$. Since $\mathcal{I}v \in H_0^1(\Omega)$ and $q \leq 2\max(2, d)/(\max(2, d) - 2)$, Sobolev's embedding theorem [26, Chapter 5.6] further gives $\mathcal{I}v \in L^q(\Omega)$. Using (A.4a), we then have $L(\boldsymbol{x}, v)\mathcal{I}v \in L^q(\Omega)$. Moreover, one has from Assumption 3.1 that

$$|\boldsymbol{\sigma}(\boldsymbol{x}, \nabla v)| \leq |\boldsymbol{\sigma}(\boldsymbol{x}, \nabla v) - \boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{0})| + |\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{0})| \overset{(A1)}{\leq} \sigma_{\mathrm{M}}|\nabla v| + |\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{0})| \overset{(A1)}{\in} L^q(\Omega),$$

$$|f(\boldsymbol{x}, v)| \leq |f(\boldsymbol{x}, v) - f(\boldsymbol{x}, 0)| + |f(\boldsymbol{x}, 0)| \overset{(A2)}{\leq} f_{\mathrm{M}}|v| + |f(\boldsymbol{x}, 0)| \overset{(A2)}{\in} L^q(\Omega).$$

Thus, rearranging (A.5) gives

$$(\mathfrak{a}(\cdot, v)\nabla\mathcal{I}v, \nabla\varphi) = (\mathcal{S}^\star, \varphi) + (\boldsymbol{F}^\star, \nabla\varphi) \quad \forall\, \varphi \in H_0^1(\Omega), \text{ with } \mathcal{S}^\star \in L^q(\Omega) \text{ and } \boldsymbol{F}^\star \in \boldsymbol{L}^q(\Omega; \mathbb{R}^d).$$

Now, as $\mathfrak{a}(\boldsymbol{x}, v)$ satisfies the ellipticity condition due to (A.4b), applying Meyers' theorem [51, Theorem 1], we see that $\mathcal{I}v \in W_0^{1,q}(\Omega) = \mathcal{W}_q$.

Third, let $q > \max\{2, d\}$ be finite and consider the expression (3.3) of $\mathcal{R}$. Using Assumption 3.3, this yields

$$f(\cdot, v) \in L^q(\Omega) \text{ and } \bar{\boldsymbol{K}}(\cdot)(\mathcal{D}(\cdot, v)\nabla v + \boldsymbol{q}(\cdot, v)) \in \boldsymbol{L}^q(\Omega; \mathbb{R}^d).$$

The last inclusion holds since $\bar{\boldsymbol{K}}\boldsymbol{q} \in \boldsymbol{L}^\infty(\Omega; \mathbb{R}^d)$, and $\mathcal{D}\bar{\boldsymbol{K}}\nabla v \in \boldsymbol{L}^q(\Omega; \mathbb{R}^d)$ as $v \in W_0^{1,q}(\Omega)$. Then the conclusion $\mathcal{I}v \in W_0^{1,q}(\Omega)$ follows as above. Finally $\mathcal{I}v \in W^{1,q}(\Omega)$ and $q > d$ implies $\mathcal{I}v \in C(\bar{\Omega}) \subset L^\infty(\Omega)$ from the Morrey's inequality [26, Chapter 5.6], so that indeed $\mathcal{I}v \in \mathcal{W}_q$. □

Proposition A.4 (Consistency of the operator $\mathcal{I}$). *Under the assumptions of Proposition A.3, let $u \in \mathcal{W}_2$ be the unique weak solution of (1.1) for a continuous operator $\mathcal{R} : H_0^1(\Omega) \to H^{-1}(\Omega)$. Then, $\mathcal{I}v = v$ if and only if $v = u$. Moreover, for a given $u_0 \in \mathcal{W}_q$, if the sequence $\{(\mathcal{I})^i u_0 \in \mathcal{W}_q\}$ is bounded uniformly in $\mathcal{W}_2$ with respect to $i \geq 0$ and Cauchy in $H_0^1(\Omega)$, then $\operatorname{dist}((\mathcal{I})^i u_0, u) \to 0$ as $i \to \infty$ with the metric $\operatorname{dist} : H_0^1(\Omega) \times H_0^1(\Omega) \to [0, \infty)$ introduced in (1.2).*

Proof. If $u \in \mathcal{W}_2$ is the weak solution of (1.1), then $(\!(\mathcal{I}u - u, \varphi)\!)_u = -\langle\mathcal{R}(u), \varphi\rangle = 0$ for all $\varphi \in H_0^1(\Omega)$, yielding $\mathcal{I}u - u = 0$. On the other hand, if $\mathcal{I}v = v$ then also $\langle\mathcal{R}(v), \varphi\rangle = 0$ for all $\varphi \in H_0^1(\Omega)$ which gives $v = u$ due to the uniqueness of u.

If $(\mathcal{I})^i u_0$ is bounded in $\mathcal{W}_2$, then there exists a constant $C > 0$ such that for all $\varphi \in H_0^1(\Omega)$,

$$\begin{aligned} |\langle\mathcal{R}((\mathcal{I})^i u_0) - \mathcal{R}(u), \varphi\rangle| = |\langle\mathcal{R}((\mathcal{I})^i u_0), \varphi\rangle| &= |(\!((\mathcal{I})^{i+1} u_0 - (\mathcal{I})^i u_0, \varphi)\!)_{(\mathcal{I}^i)u_0}| \\ &\leq C\|\nabla((\mathcal{I})^{i+1} u_0 - (\mathcal{I})^i u_0)\|\|\nabla\varphi\|. \end{aligned}$$

The right-hand side goes to 0 as $i \to \infty$ as the sequence $\{(\mathcal{I})^i u_0\}_{i\geq 0}$ is Cauchy. Consequently $\operatorname{dist}((\mathcal{I}^i)u_0, u) \to 0$ due to (1.2).

□

Appendix B Error estimates in a fixed norm

The norm $|||\cdot|||_{-1,u_\ell^i}$, introduced in (4.18), is not a fixed norm since its definition changes with the iteration index i. In this section, we show how the bounds for $|||\mathcal{R}(u_\ell^i)|||_{-1,u_\ell^i}$, presented in Section 5, can be converted to give estimates of $\mathcal{R}(u_\ell^i)$ in the fixed norm $|||\cdot|||_{-1,u_\ell}$, where $u_\ell \in V_\ell$ is the finite element solution of the nonlinear problem introduced in (4.3). For this purpose, we adopt the setting of Appendix A, in particular (A.2), (recall $V_\ell \subset \mathcal{W}_q \subseteq \mathcal{W}_2$ from (4.2) and since $q \geq 2$), and assume that:

Assumption B.1 (Dependence of $(\!(\cdot,\cdot)\!)_v$ on $v \in V_\ell$)**.** *For $v, w \in V_\ell$ and $\xi \in H_0^1(\Omega)$, there exists a positive function $F \in C(\Omega \times \mathbb{R})$, Lipschitz with respect to the second argument $\xi \in \mathbb{R}$, i.e.,*

$$|F(\boldsymbol{x},\xi_1) - F(\boldsymbol{x},\xi_2)| \leq F_{\mathrm{M}}\, |\xi_1 - \xi_2| \quad \forall \boldsymbol{x} \in \Omega \text{ and } \xi_1, \xi_2 \in \mathbb{R}, \tag{B.1}$$

such that

$$|||\xi|||_{1,w}^2 \leq \max\left(\sup_{\boldsymbol{x}\in\Omega}\left\{\tfrac{F(\boldsymbol{x},w)}{F(\boldsymbol{x},v)}\right\}, 1\right) |||\xi|||_{1,v}^2. \tag{B.2}$$

Remark B.2 (Applicability of Assumption B.1)**.** *Assumption B.1 is satisfied for most of the schemes presented in Tables 1 and 2 with the corresponding F-functions given in Table 3. For example, in the case of the L-scheme, one may choose $F = \mathcal{D}$ since*

$$\begin{aligned}
|||\xi|||_{1,w}^2 =& \|L^{\frac12}\xi\|^2 + \|\tau^{\frac12}\bar{\boldsymbol{K}}^{\frac12}(\cdot)\mathcal{D}(\cdot,w)^{\frac12}\nabla\xi\|^2\\
\leq& \max\left(\sup_{\boldsymbol{x}\in\Omega}\left\{\tfrac{\mathcal{D}(\boldsymbol{x},w)}{\mathcal{D}(\boldsymbol{x},v)}\right\}, 1\right)\left[\|L^{\frac12}\xi\|^2 + \|\tau^{\frac12}\bar{\boldsymbol{K}}^{\frac12}(\cdot)\mathcal{D}(\cdot,v)^{\frac12}\nabla\xi\|^2\right]\\
=& \max\left(\sup_{\boldsymbol{x}\in\Omega}\left\{\tfrac{\mathcal{D}(\boldsymbol{x},w)}{\mathcal{D}(\boldsymbol{x},v)}\right\}, 1\right)|||\xi|||_{1,v}^2.
\end{aligned}$$

However, the Newton scheme and the Kačanov scheme do not satisfy Assumption B.1 directly. In these cases, the function F also needs to depend on derivatives, i.e., $F = F(\boldsymbol{x}, v, \nabla v)$. We have not considered this option here for simplicity.

Scheme	$F(\boldsymbol{x},\xi)$
Zarantonello	Λ
Picard	$\partial_\xi f(\boldsymbol{x},\xi)$ or $\mathcal{D}(\boldsymbol{x},\xi)$
Jäger-Kačur	$\max_{\zeta\in\mathbb{R}}\left(\frac{f(\boldsymbol{x},\zeta)-f(\boldsymbol{x},\xi)}{\zeta-\xi}\right)$ or $\mathcal{D}(\boldsymbol{x},\xi)$
L-scheme	$\mathcal{D}(\boldsymbol{x},\xi)$
M-scheme	$\partial_\xi f(\boldsymbol{x},\xi) + M\tau$ or $\mathcal{D}(\boldsymbol{x},\xi)$

Table 3: The choices of F in Assumption B.1 for different linearization schemes given in Tables 1 and 2. For Picard, Jäger–Kačur, and M-scheme, the choice yielding higher $\sup_{\boldsymbol{x}\in\Omega}\{F(\boldsymbol{x},w)/F(\boldsymbol{x},v)\}$ value needs to be selected.

If stable iterative schemes converge with a linear rate in $H_0^1(\Omega)$, and more importantly in $L^2(\Omega)$, they are also expected to converge linearly in $L^\infty(\Omega)$. This is proven for the L- and the M-schemes in Proposition 3.2 of [52] for the continuous solutions. Here we assume it for the finite element solutions:

Assumption B.3 (Contraction in L^∞ of the linear iterations)**.** *There exists $\alpha \in (0,1)$ such that*

$$\|u_\ell^{i+1} - u_\ell^i\|_{L^\infty(\Omega)} \leq \alpha \|u_\ell^i - u_\ell^{i-1}\|_{L^\infty(\Omega)} \quad \forall\, i \geq 1. \tag{B.3}$$

Remark B.4 (Estimating $\alpha \in (0,1)$). *Note that the theoretical estimates for the contraction rates always depend on the Lipschitz/monotonicity ratio $\lambda_{\mathrm{M}}/\lambda_{\mathrm{m}}$ from (1.2), and thus give a pessimistic theoretical value of $\alpha \approx 1$ whenever this ratio is large. However, in Assumption B.3, the parameter α can be directly estimated from the numerical solutions as*

$$\alpha \geq \alpha_i := \|u_\ell^{i+1} - u_\ell^i\|_{L^\infty(\Omega)} / \|u_\ell^i - u_\ell^{i-1}\|_{L^\infty(\Omega)}.$$

Thus, for the guaranteed linearly converging schemes, like the Zarantonello, Jäger–Kačur, L- and M-schemes, one expects to get a tight practical bound on $\alpha \in (0,1)$ away from 1 with only a couple of iterations.

Theorem B.5 (Error estimates in the fixed norm $|\!|\!|\cdot|\!|\!|_{-1,u_\ell}$). *Let $u \in H_0^1(\Omega)$ solve (1.1) and $u_\ell \in V_\ell$ solve (4.3). Let the sequence $\{u_\ell^i\}_{i\geq 0} \subset V_\ell$ be defined by (4.7), using the scalar product (4.4) and the norms (4.15) and (4.18). Let Assumptions B.1 and B.3 hold. Then, for any $\tilde{u} \in H_0^1(\Omega)$,*

$$\frac{|\!|\!|\mathcal{R}(\tilde{u})|\!|\!|_{-1,u_\ell^i}}{(1 + \mathfrak{C}_{\mathrm{m}}^i \|u_\ell^{i+1} - u_\ell^i\|_{L^\infty(\Omega)})^{\frac{1}{2}}} \leq |\!|\!|\mathcal{R}(\tilde{u})|\!|\!|_{-1,u_\ell} \leq (1 + \mathfrak{C}_{\mathrm{M}}^i \|u_\ell^{i+1} - u_\ell^i\|_{L^\infty(\Omega)})^{\frac{1}{2}} \, |\!|\!|\mathcal{R}(\tilde{u})|\!|\!|_{-1,u_\ell^i}, \tag{B.4}$$

where the constants $\mathfrak{C}_{\mathrm{m}}^i$, $\mathfrak{C}_{\mathrm{M}}^i \geq 0$, given by

$$\mathfrak{C}_{\mathrm{m}}^i := \frac{1}{1-\alpha} \sup_{\boldsymbol{x} \in \Omega} \left\{ \frac{1}{F(\boldsymbol{x}, u_\ell^i)} \sup_{v \in I_\ell^i(\boldsymbol{x})} \left| \frac{F(\boldsymbol{x}, v) - F(\boldsymbol{x}, u_\ell^i)}{v - u_\ell^i} \right| \right\}, \tag{B.5a}$$

$$\mathfrak{C}_{\mathrm{M}}^i := \frac{1}{1-\alpha} \sup_{\boldsymbol{x} \in \Omega} \left\{ F(\boldsymbol{x}, u_\ell^i) \sup_{v \in I_\ell^i(\boldsymbol{x})} \left| \frac{\frac{1}{F(\boldsymbol{x},v)} - \frac{1}{F(\boldsymbol{x},u_\ell^i)}}{v - u_\ell^i} \right| \right\}, \tag{B.5b}$$

are uniformly bounded with respect to iteration i, where the interval $I_\ell^i(\boldsymbol{x})$ is defined for all $\boldsymbol{x} \in \Omega$ as

$$I_\ell^i(\boldsymbol{x}) := \left[u_\ell^i(\boldsymbol{x}) - \tfrac{\|u_\ell^{i+1} - u_\ell^i\|_{L^\infty(\Omega)}}{1-\alpha}, u_\ell^i(\boldsymbol{x}) + \tfrac{\|u_\ell^{i+1} - u_\ell^i\|_{L^\infty(\Omega)}}{1-\alpha} \right].$$

Proof. Observe that Assumption B.3 implies that $u_\ell^i \to u_\ell$ as $i \to \infty$ and

$$\|u_\ell^{n+1} - u_\ell^n\|_{L^\infty(\Omega)} \leq \alpha^{n-i} \|u_\ell^{i+1} - u_\ell^i\|_{L^\infty(\Omega)}$$

for $n \geq i$. This gives

$$\begin{aligned} \|u_\ell - u_\ell^i\|_{L^\infty(\Omega)} &\leq \sum_{n=i}^{\infty} \|u_\ell^{n+1} - u_\ell^n\|_{L^\infty(\Omega)} \leq \|u_\ell^{i+1} - u_\ell^i\|_{L^\infty(\Omega)} \sum_{n=0}^{\infty} \alpha^n \\ &= \frac{\|u_\ell^{i+1} - u_\ell^i\|_{L^\infty(\Omega)}}{1-\alpha} \leq \frac{\alpha^i}{1-\alpha} \|u_\ell^1 - u_\ell^0\|_{L^\infty(\Omega)}, \end{aligned} \tag{B.6a}$$

which also yields for all $\boldsymbol{x} \in \Omega$ that

$$u_\ell^i(\boldsymbol{x}) - \tfrac{\|u_\ell^{i+1} - u_\ell^i\|_{L^\infty(\Omega)}}{1-\alpha} \leq u_\ell(\boldsymbol{x}) \leq u_\ell^i(\boldsymbol{x}) + \tfrac{\|u_\ell^{i+1} - u_\ell^i\|_{L^\infty(\Omega)}}{1-\alpha}, \quad \text{or} \quad u_\ell(\boldsymbol{x}) \in I_\ell^i(\boldsymbol{x}). \tag{B.6b}$$

Hence, if $\tilde{\varphi} = \arg\max_{\varphi \in H_0^1(\Omega)} (|\langle \mathcal{R}(\tilde{u}), \varphi \rangle| / |\!|\!|\varphi|\!|\!|_{1,u_\ell})$, then Assumption B.1 yields

$$|\!|\!|\mathcal{R}(\tilde{u})|\!|\!|_{-1,u_\ell} = \frac{|\langle \mathcal{R}(\tilde{u}), \tilde{\varphi} \rangle|}{|\!|\!|\tilde{\varphi}|\!|\!|_{1,u_\ell}} = \left(\frac{|\langle \mathcal{R}(\tilde{u}), \tilde{\varphi} \rangle|}{|\!|\!|\tilde{\varphi}|\!|\!|_{1,u_\ell^i}} \right) \left(\frac{|\!|\!|\tilde{\varphi}|\!|\!|_{1,u_\ell^i}}{|\!|\!|\tilde{\varphi}|\!|\!|_{1,u_\ell}} \right) \leq |\!|\!|\mathcal{R}(\tilde{u})|\!|\!|_{-1,u_\ell^i} \sqrt{\sup_\Omega \frac{F(u_\ell^i)}{F(u_\ell)}}. \tag{B.7}$$

Here, we have dropped the $\boldsymbol{x}$-dependence of F for simplifying the notation and we have assumed that $\sup_\Omega F(u_\ell^i)/F(u_\ell) > 1$, since the $\sup_\Omega F(u_\ell^i)/F(u_\ell) \leq 1$ case is trivial in that it gives

$|\!|\!|\mathcal{R}(\tilde{u})|\!|\!|_{-1,u_\ell} \leq |\!|\!|\mathcal{R}(\tilde{u})|\!|\!|_{-1,u_\ell^i}$ directly. Then, we get the upper estimate of $|\!|\!|\mathcal{R}(\tilde{u})|\!|\!|_{-1,u_\ell}$ by observing

$$\begin{aligned}\sup_\Omega \tfrac{F(u_\ell^i)}{F(u_\ell)} &= 1+\sup_\Omega \tfrac{F(u_\ell^i)-F(u_\ell)}{F(u_\ell)} = 1+\sup_\Omega\left\{F(u_\ell^i)\left(\tfrac{1}{F(u_\ell)}-\tfrac{1}{F(u_\ell^i)}\right)\right\}\\ &\overset{\text{(B.6b)}}{\leq} 1+\sup_\Omega\left\{F(u_\ell^i)\left|\frac{\frac{1}{F(u_\ell)}-\frac{1}{F(u_\ell^i)}}{u_\ell-u_\ell^i}\right|\right\}\|u_\ell-u_\ell^i\|_{L^\infty(\Omega)}\\ &\overset{\text{(B.5b),(B.6a)}}{\leq} 1+\mathfrak{C}_{\mathrm{M}}^i\|u_\ell^{i+1}-u_\ell^i\|_{L^\infty(\Omega)}.\end{aligned}\tag{B.8}$$

The proof of the lower estimate is similar and uses that

$$|\!|\!|\mathcal{R}(\tilde{u})|\!|\!|_{-1,u_\ell^i} \leq (1+\mathfrak{C}_{\mathrm{m}}^i\|u_\ell^{i+1}-u_\ell^i\|_{L^\infty(\Omega)})^{\frac{1}{2}}|\!|\!|\mathcal{R}(\tilde{u})|\!|\!|_{-1,u_\ell}.$$

Finally, to see that the constants $\mathfrak{C}_{\mathrm{m}}^i$ are uniformly bounded, note that $\|u_\ell^i\|_{L^\infty(\Omega)}$ are bounded uniformly due to (B.6a), and consequently, $I_\ell^i(\boldsymbol{x})$ are subsets of a compact set $I_\ell \subset \mathbb{R}$ which does not depend on the iteration index i. The positivity and continuity of F then implies that $\inf_{v\in I_\ell} F(v) =: F_{\mathrm{m}} > 0$. Hence, we get

$$\mathfrak{C}_{\mathrm{m}}^i \overset{\text{(B.5a)}}{\leq} \frac{1}{1-\alpha}\sup_{\boldsymbol{x}\in\Omega}\left\{\sup_{v\in I_\ell^i(\boldsymbol{x})}\left|\tfrac{F(\boldsymbol{x},v)-F(\boldsymbol{x},u_\ell^i)}{v-u_\ell^i}\right|\frac{1}{\inf_{v\in I_\ell^i(\boldsymbol{x})}F(\boldsymbol{x},v)}\right\} \overset{\text{(B.1)}}{\leq} \frac{F_{\mathrm{M}}}{(1-\alpha)F_{\mathrm{m}}}.\tag{B.9a}$$

The same argument shows that also

$$\mathfrak{C}_{\mathrm{M}}^i \leq \frac{F_{\mathrm{M}}}{(1-\alpha)F_{\mathrm{m}}}.\tag{B.9b}$$

□

The following result shows that if the iterations $\{u_\ell^i\}_{i\geq 0}$ are converging, then the iteration-dependent dual norm $|\!|\!|\mathcal{R}(u_\ell^i)|\!|\!|_{-1,u_\ell^i}$ converges to the error of the finite element solution measured in the iteration-independent natural discrete dual norm $|\!|\!|\mathcal{R}(u_\ell)|\!|\!|_{-1,u_\ell}$.

Corollary B.6 (Convergence of $|\!|\!|\mathcal{R}(u_\ell^i)|\!|\!|_{-1,u_\ell^i}$ to $|\!|\!|\mathcal{R}(u_\ell)|\!|\!|_{-1,u_\ell}$)**.** *Let $\mathcal{R} : H_0^1(\Omega) \to H^{-1}(\Omega)$ from (1.1) be a continuous operator, and let the assumptions of Theorem B.5 hold. Then*

$$|\!|\!|\mathcal{R}(u_\ell^i)|\!|\!|_{-1,u_\ell^i} \to |\!|\!|\mathcal{R}(u_\ell)|\!|\!|_{-1,u_\ell} \quad \textit{as } i\to\infty.$$

Proof. Using the triangle inequality,

$$\Big||\!|\!|\mathcal{R}(u_\ell^i)|\!|\!|_{-1,u_\ell^i} - |\!|\!|\mathcal{R}(u_\ell)|\!|\!|_{-1,u_\ell}\Big| \leq \Big||\!|\!|\mathcal{R}(u_\ell)|\!|\!|_{-1,u_\ell^i} - |\!|\!|\mathcal{R}(u_\ell)|\!|\!|_{-1,u_\ell}\Big| + \Big||\!|\!|\mathcal{R}(u_\ell^i)|\!|\!|_{-1,u_\ell^i} - |\!|\!|\mathcal{R}(u_\ell)|\!|\!|_{-1,u_\ell^i}\Big|.$$

The first term vanishes due to Theorem B.5, in particular the uniform bounds on $\mathfrak{C}_{\mathrm{m}}^i$ and $\mathfrak{C}_{\mathrm{M}}^i$ from (B.9). The second term on the right-hand-side vanishes due to the continuity of the operator $\mathcal{R}$. More precisely, as $u_\ell^i \to u_\ell$,

$$\begin{aligned}\Big||\!|\!|\mathcal{R}(u_\ell^i)|\!|\!|_{-1,u_\ell^i} - |\!|\!|\mathcal{R}(u_\ell)|\!|\!|_{-1,u_\ell^i}\Big| &\leq |\!|\!|\mathcal{R}(u_\ell^i)-\mathcal{R}(u_\ell)|\!|\!|_{-1,u_\ell^i} = \sup_{\varphi\in H_0^1(\Omega)}\frac{\langle\mathcal{R}(u_\ell^i)-\mathcal{R}(u_\ell),\varphi\rangle}{|\!|\!|\varphi|\!|\!|_{1,u_\ell^i}}\\ &\overset{\text{(B.2)}}{\leq} \max\left(\sup_\Omega\sqrt{\frac{F(u_\ell)}{F(u_\ell^i)}},1\right)\sup_{\varphi\in H_0^1(\Omega)}\frac{\langle\mathcal{R}(u_\ell^i)-\mathcal{R}(u_\ell),\varphi\rangle}{|\!|\!|\varphi|\!|\!|_{1,u_\ell}} \leq C|\!|\!|\mathcal{R}(u_\ell^i)-\mathcal{R}(u_\ell)|\!|\!|_{-1,u_\ell} \longrightarrow 0.\end{aligned}$$

Here, the uniform boundedness of $F(u_\ell)/F(u_\ell^i)$ in Ω with respect to the iteration index i has been used which can be proven following the arguments from Theorem B.5, see (B.8)–(B.9). □

Remark B.7 (Convergence in the iteration-dependent norm $|\!|\!|\cdot|\!|\!|_{1,u_\ell^i}$)**.** *Following Lemma 5.2, the statement of Corollary B.6 can be rewritten as*

$$\Big|\!\Big|\!\Big|u_\ell^i - u_{\langle\ell\rangle}^{i+1}\Big|\!\Big|\!\Big|_{1,u_\ell^i} \to |\!|\!|u_\ell - u_{\langle\ell\rangle}|\!|\!|_{1,u_\ell} \quad \textit{as } i\to\infty.$$

Appendix C Error estimates for the Newton linearization for the gradient-independent diffusivity

In this section, we consider the Newton scheme for the gradient-independent diffusivity case (3.3). This leads to a nonsymmetric linearization problem which does not fit the symmetric framework of Sections 4 and 5.

C.1 Newton iterations

The Newton(–Raphson) scheme for iterative linearization of (4.3) in the gradient-independent diffusivity case (3.3), see [22], is defined in the spirit of (4.4), see also (A.2), but with an advection–reaction–diffusion linearization bilinear form $\mathfrak{L}_{\mathrm{N}}$: for $u_\ell^i \in V_\ell$ and ξ, $\zeta \in H_0^1(\Omega)$,

$$\mathfrak{L}_{\mathrm{N}}(u_\ell^i; \xi, \zeta) := \langle L_{\mathrm{N}}^i\, \xi, \zeta\rangle + (\mathfrak{a}_{\mathrm{N}}^i \nabla\xi + \boldsymbol{w}_{\mathrm{N}}^i \xi, \nabla\zeta), \tag{C.1}$$

where L_{N}^i and $\mathfrak{a}_{\mathrm{N}}^i$ satisfy Assumption 4.1 taken for L^i and $\mathfrak{a}^i$. In particular, with reference to (3.3) and Table 2,

$$\begin{aligned} &L_{\mathrm{N}}^i(\boldsymbol{x}) := \partial_\xi f(\boldsymbol{x}, u_\ell^i),\ \mathfrak{a}_{\mathrm{N}}^i(\boldsymbol{x}) := \tau\, \bar{\boldsymbol{K}}(\boldsymbol{x})\, \mathcal{D}(\boldsymbol{x}, u_\ell^i), \\ &\boldsymbol{w}_{\mathrm{N}}^i(\boldsymbol{x}) := \tau \bar{\boldsymbol{K}}(\boldsymbol{x})\, [\partial_\xi \mathcal{D}(\boldsymbol{x}, u_\ell^i)\, \nabla u_\ell^i + \partial_\xi \boldsymbol{q}(\boldsymbol{x}, u_\ell^i)]\,. \end{aligned} \tag{C.2}$$

Starting from an initial guess $u_\ell^0 \in V_\ell$, the Newton iterates $\{u_\ell^i\}_{i\geq 1} \subset V_\ell$ are then obtained by solving, similar to (4.7), the problems

$$\mathfrak{L}_{\mathrm{N}}(u_\ell^i; u_\ell^{i+1} - u_\ell^i, \varphi_\ell) = -\langle \mathcal{R}(u_\ell^i), \varphi_\ell\rangle \qquad \forall \varphi_\ell \in V_\ell. \tag{C.3}$$

Although $\mathfrak{L}_{\mathrm{N}}$ is bilinear, it is not symmetric, and thus it does not correspond to an inner product, in contrast to (4.4). Consequently, the orthogonality relation obtained in Theorem 5.3 does not hold in this case. However, using the techniques of Sections 5 and 8, it is still possible to provide an upper and a lower bound on an iteration-dependent dual norm of $\mathcal{R}(u_\ell^i)$ which converge to an orthogonal relation when the asymmetry (advection) vanishes. The following restrictions on $\boldsymbol{w}_{\mathrm{N}}^i$ are assumed for this purpose.

Assumption C.1 (Properties of the advection vector $\boldsymbol{w}_{\mathrm{N}}^i$)**.** *For $\{u_\ell^i\}_{i\geq 0} \subset V_\ell$ stemming from the iterative Newton linearization (C.3), we assume that $\boldsymbol{w}_{\mathrm{N}}^i \in \boldsymbol{L}^\infty(\Omega;\mathbb{R}^d)$ and there exists a constant $C_N \in [0,2)$ such that*

$$\boldsymbol{w}_{\mathrm{N}}^i(\boldsymbol{x})\, (\mathfrak{a}_{\mathrm{N}}^i(\boldsymbol{x}))^{-1}\, \boldsymbol{w}_{\mathrm{N}}^i(\boldsymbol{x}) \leq C_N^2\, L_{\mathrm{N}}^i(\boldsymbol{x}) \qquad \textit{for a.e. } \boldsymbol{x} \in \Omega,\ \forall i \geq 0. \tag{C.4}$$

Remark C.2 (Validity of Assumption C.1)**.** *For problems having linear diffusion coefficients, i.e. $\partial_\xi \mathcal{D} = 0$ in (3.3), Assumption C.1 holds if $L_{\mathrm{N}}^i > 0$ and τ (which corresponds to the time-step size for time-discretized unsteady problems, cf. Class II in Example 3.4) is small enough. For problems having nonlinear diffusion coefficients, such as the Richards equation (see Section 6.2), the condition is still satisfied if u_ℓ^i is bounded in $W^{1,\infty}(\Omega)$ uniformly with respect to i and the time-step size $\tau > 0$ is small. The time-step and non-degeneracy restrictions are well-known constraints in proving the convergence of the Newton scheme in applications, see [52] and the references therein.*

Similarly to (4.15) and (4.18) we define, for any $i \geq 0$, the energy norms

$$|||\xi|||_{1,u_\ell^i} := (\|(L_{\mathrm{N}}^i)^{\frac{1}{2}}\xi\|^2 + \|(\mathfrak{a}_{\mathrm{N}}^i)^{\frac{1}{2}}\nabla\xi\|^2)^{\frac{1}{2}}, \quad |||\varsigma|||_{-1,u_\ell^i} := \sup_{\varphi\in H_0^1(\Omega)} \frac{\langle\varsigma,\varphi\rangle}{|||\varphi|||_{1,u_\ell^i}}. \tag{C.5}$$

Contrary to (4.15), however, we do not have $|||\xi|||_{1,u_\ell^i}^2 = \mathfrak{L}_{\mathrm{N}}(u_\ell^i;\xi,\xi)$ for $\xi \in H_0^1(\Omega)$. Instead, by the Assumption C.1, there holds

Lemma C.3 (Inequalities for $\mathfrak{L}_{\mathrm{N}}$ and the $|\!|\!|\cdot|\!|\!|_{\pm 1, u_\ell^i}$ norms)**.** *For all ξ, $\zeta \in H_0^1(\Omega)$,*

$$|\mathfrak{L}_{\mathrm{N}}(u_\ell^i; \xi, \zeta)| \leq (1 + C_{\mathrm{N}}) |\!|\!|\xi|\!|\!|_{1,u_\ell^i} |\!|\!|\zeta|\!|\!|_{1,u_\ell^i}, \tag{C.6a}$$

$$(1 - \tfrac{C_{\mathrm{N}}}{2}) |\!|\!|\xi|\!|\!|_{1,u_\ell^i}^2 \leq \mathfrak{L}_{\mathrm{N}}(u_\ell^i; \xi, \xi) \leq \left(1 + \tfrac{C_{\mathrm{N}}}{2}\right) |\!|\!|\xi|\!|\!|_{1,u_\ell^i}^2. \tag{C.6b}$$

Proof. Observe from (C.1) and (C.5) that

$$|\mathfrak{L}_{\mathrm{N}}(u_\ell^i; \xi, \zeta)| \leq |\!|\!|\xi|\!|\!|_{1,u_\ell^i} |\!|\!|\zeta|\!|\!|_{1,u_\ell^i} + |(\boldsymbol{w}_{\mathrm{N}}^i \xi, \nabla \zeta)|.$$

For (C.6a), we estimate the last term using the Cauchy–Schwarz inequality:

$$\begin{aligned} |(\boldsymbol{w}_{\mathrm{N}}^i \xi, \nabla \zeta)| &\leq \|(\mathfrak{a}_{\mathrm{N}}^i)^{-\frac{1}{2}} \boldsymbol{w}_{\mathrm{N}}^i \xi\| \|(\mathfrak{a}_{\mathrm{N}}^i)^{\frac{1}{2}} \nabla \zeta\| \\ &\overset{(C.4)}{\leq} C_N \|(L_{\mathrm{N}}^i)^{\frac{1}{2}} \xi\| \|(\mathfrak{a}_{\mathrm{N}}^i)^{\frac{1}{2}} \nabla \zeta\| \overset{(C.5)}{\leq} C_{\mathrm{N}} |\!|\!|\xi|\!|\!|_{1,u_\ell^i} |\!|\!|\zeta|\!|\!|_{1,u_\ell^i}. \end{aligned} \tag{C.7}$$

For proving (C.6b), we note that

$$|(\boldsymbol{w}_{\mathrm{N}}^i \xi, \nabla \xi)| \overset{(C.4)}{\leq} C_N \|(L_{\mathrm{N}}^i)^{\frac{1}{2}} \xi\| \|(\mathfrak{a}_{\mathrm{N}}^i)^{\frac{1}{2}} \nabla \xi\| \overset{(2.2)}{\leq} \frac{C_N}{2} (\|(L_{\mathrm{N}}^i)^{\frac{1}{2}} \xi\|^2 + \|(\mathfrak{a}_{\mathrm{N}}^i)^{\frac{1}{2}} \nabla \xi\|^2) = \frac{C_N}{2} |\!|\!|\xi|\!|\!|_{1,u_\ell^i}^2.$$

□

C.2 Bounds on the dual norm of the residual by linearization and discretization components

Extending the definition of $\mathcal{R}_{\mathrm{disc}}^{u_\ell^i} : H_0^1(\Omega) \to H^{-1}(\Omega)$ from (5.1) to the Newton case, one has

$$\langle \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(\xi), \varphi \rangle := \mathfrak{L}_{\mathrm{N}}(u_\ell^i; \xi - u_\ell^i, \varphi) + \langle \mathcal{R}(u_\ell^i), \varphi \rangle, \qquad \varphi \in H_0^1(\Omega), \tag{C.8}$$

so that using (C.3), $\langle \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}), \varphi_\ell \rangle = 0$ for all $\varphi_\ell \in V_\ell$. Then, in extension of Theorem 5.3, we have:

Theorem C.4 (Upper and lower bounds on error for non-symmetric $\mathfrak{L}_{\mathrm{N}}$ (Newton))**.** *In the context of the gradient-independent diffusivity case (3.3), let the bilinear form $\mathfrak{L}_{\mathrm{N}}$ be defined by (C.1), the sequence $\{u_\ell^i\}_{i \geq 0} \subset V_\ell$ by (C.3), the norms $|\!|\!|\cdot|\!|\!|_{\pm 1, u_\ell^i}$ by (C.5), and $\mathcal{R}_{\mathrm{disc}}^{u_\ell^i} : H_0^1(\Omega) \to H^{-1}(\Omega)$ by (C.8). Let Assumption C.1 hold for some $C_{\mathrm{N}} \in [0, 2)$. Then, one has*

$$\begin{aligned} &\left(1 - \tfrac{C_{\mathrm{N}}}{2}\right)^2 \Bigg[\tfrac{1}{(1+C_{\mathrm{N}})^2} \left|\!\left|\!\left| \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}) \right|\!\right|\!\right|_{-1,u_\ell^i}^2 + |\!|\!|u_\ell^i - u_\ell^{i+1}|\!|\!|_{1,u_\ell^i}^2 \\ &\qquad\qquad - \tfrac{4C_{\mathrm{N}}}{2-C_{\mathrm{N}}} \left|\!\left|\!\left| \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}) \right|\!\right|\!\right|_{-1,u_\ell^i} |\!|\!|u_\ell^i - u_\ell^{i+1}|\!|\!|_{1,u_\ell^i} \Bigg] \\ &\leq |\!|\!|\mathcal{R}(u_\ell^i)|\!|\!|_{-1,u_\ell^i}^2 \leq (1 + C_{\mathrm{N}})^2 \Bigg[\tfrac{4}{(2-C_{\mathrm{N}})^2} \left|\!\left|\!\left| \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}) \right|\!\right|\!\right|_{-1,u_\ell^i}^2 + |\!|\!|u_\ell^i - u_\ell^{i+1}|\!|\!|_{1,u_\ell^i}^2 \\ &\qquad\qquad + \tfrac{4C_{\mathrm{N}}}{2-C_{\mathrm{N}}} \left|\!\left|\!\left| \mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1}) \right|\!\right|\!\right|_{-1,u_\ell^i} |\!|\!|u_\ell^i - u_\ell^{i+1}|\!|\!|_{1,u_\ell^i} \Bigg]. \end{aligned}$$

Remark C.5 (Relation to Theorem 5.3)**.** *If the asymmetry (the advection term $\boldsymbol{w}_{\mathrm{N}}^i$) in $\mathfrak{L}_{\mathrm{N}}$ disappears, then the constant C_{N} in (C.4) can be taken as 0. In this case, both the upper and the lower estimates of $|\!|\!|\mathcal{R}(u_\ell^i)|\!|\!|_{-1,u_\ell^i}$ in Theorem C.4 change to the orthogonal relation of Theorem 5.3.*

Proof. Similarly to (4.13), let $u^{i+1}_{\langle\ell\rangle} \in H^1_0(\Omega)$ solve

$$\mathfrak{L}_{\rm N}(u^i_\ell; u^{i+1}_{\langle\ell\rangle} - u^i_\ell, \varphi) = -\langle \mathcal{R}(u^i_\ell), \varphi\rangle \qquad \forall \varphi \in H^1_0(\Omega). \tag{C.9}$$

Then, following (C.3)–(C.9) and Lemma C.3, we have that

$$|||\mathcal{R}(u^i_\ell)|||_{-1,u^i_\ell} \overset{(C.9)}{=} \sup_{\varphi\in H^1_0(\Omega)} \frac{\mathfrak{L}_{\rm N}(u^i_\ell; u^{i+1}_{\langle\ell\rangle} - u^i_\ell, \varphi)}{|||\varphi|||_{1,u^i_\ell}} \begin{cases} \overset{(C.6a)}{\leq} (1 + C_{\rm N})\Big|\Big|\Big|u^{i+1}_{\langle\ell\rangle} - u^i_\ell\Big|\Big|\Big|_{1,u^i_\ell}, \\ \overset{(C.6b)}{\geq} \left(1 - \frac{C_{\rm N}}{2}\right) \Big|\Big|\Big|u^{i+1}_{\langle\ell\rangle} - u^i_\ell\Big|\Big|\Big|_{1,u^i_\ell}. \end{cases} \tag{C.10}$$

Expanding $\Big|\Big|\Big|u^{i+1}_{\langle\ell\rangle} - u^i_\ell\Big|\Big|\Big|^2_{1,u^i_\ell}$, one has

$$\begin{aligned} \Big|\Big|\Big|u^{i+1}_{\langle\ell\rangle} - u^i_\ell\Big|\Big|\Big|^2_{1,u^i_\ell} &\overset{(C.5)}{=} \Big|\Big|\Big|u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell\Big|\Big|\Big|^2_{1,u^i_\ell} + |||u^{i+1}_\ell - u^i_\ell|||^2_{1,u^i_\ell} \\ &\quad + 2[(L^i_{\rm N}(u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell), u^{i+1}_\ell - u^i_\ell) + (\mathfrak{a}^i_{\rm N}\nabla(u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell), \nabla(u^{i+1}_\ell - u^i_\ell))]. \end{aligned} \tag{C.11a}$$

Observe that since $u^{i+1}_\ell - u^i_\ell \in V_\ell$, we have by the definition of finite element solutions (C.3) that

$$\mathfrak{L}_{\rm N}(u^i_\ell; u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell, u^{i+1}_\ell - u^i_\ell) = 0. \tag{C.11b}$$

Hence, we have following the proof of Lemma C.3,

$$\begin{aligned} &|(L^i_{\rm N}(u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell), u^{i+1}_\ell - u^i_\ell) + (a^i_{\rm N}\nabla(u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell), \nabla(u^{i+1}_\ell - u^i_\ell))| \\ &\overset{(C.1)}{=} |\mathfrak{L}_{\rm N}(u^i_\ell; u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell, u^{i+1}_\ell - u^i_\ell) - (\boldsymbol{w}^i_{\rm N}(u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell), \nabla(u^{i+1}_\ell - u^i_\ell))| \\ &\overset{(C.11b)}{=} |(\boldsymbol{w}^i_{\rm N}(u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell), \nabla(u^{i+1}_\ell - u^i_\ell))| \overset{(C.7)}{\leq} C_N\Big|\Big|\Big|u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell\Big|\Big|\Big|_{1,u^i_\ell} |||u^{i+1}_\ell - u^i_\ell|||_{1,u^i_\ell}. \end{aligned} \tag{C.11c}$$

Finally, noting that

$$\begin{aligned} \Big|\Big|\Big|u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell\Big|\Big|\Big|_{1,u^i_\ell} &\overset{(C.6b)}{\leq} \frac{2}{2 - C_{\rm N}} \frac{\mathfrak{L}_{\rm N}(u^i_\ell; u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell, u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell)}{\Big|\Big|\Big|u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell\Big|\Big|\Big|_{1,u^i_\ell}} \\ &\leq \frac{2}{2 - C_{\rm N}} \sup_{\varphi\in H^1_0(\Omega)} \frac{\mathfrak{L}_{\rm N}(u^i_\ell; u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell, \varphi)}{|||\varphi|||_{1,u^i_\ell}} \overset{(C.8),(C.9)}{=} \frac{2}{2 - C_{\rm N}} \Big|\Big|\Big|\mathcal{R}^{u^i_\ell}_{\rm disc}(u^{i+1}_\ell)\Big|\Big|\Big|_{-1,u^i_\ell}, \end{aligned} \tag{C.11d}$$

we have the upper bound. The proof of the lower bound is similar, but requires the estimation

$$\Big|\Big|\Big|\mathcal{R}^{u^i_\ell}_{\rm disc}(u^{i+1}_\ell)\Big|\Big|\Big|_{-1,u^i_\ell} \leq (1 + C_{\rm N})\Big|\Big|\Big|u^{i+1}_{\langle\ell\rangle} - u^{i+1}_\ell\Big|\Big|\Big|_{1,u^i_\ell},$$

which follows from the last equality in (C.11d) together with (C.6a). □

C.3 A posteriori error bounds

With the above developments, the a posteriori analysis of Sections 5 and 8 carries over to the Newton scheme. In particular, the counterpart of Theorem 5.7 for the Newton scheme becomes:

Theorem C.6 (Guaranteed, efficient, and robust a posteriori estimates for non-symmetric $\mathfrak{L}_{\rm N}$ (Newton))**.** *Let the assumptions of Theorem C.4 be satisfied. Rearranging (C.3), write it in the form (4.9) with $\mathcal{S}^i$ given by (4.12a) and $\boldsymbol{\mathcal{F}}^i := \tau\bar{\boldsymbol{K}}(\cdot)\,\boldsymbol{q}(\cdot, u^i_\ell) + (\tau\bar{\boldsymbol{K}}(\cdot)\mathcal{D}(\cdot, u^i_\ell) - \mathfrak{a}^i_N)\nabla u^i_\ell +$*

$\boldsymbol{w}_N^i(u_\ell^{i+1}-u_\ell^i) \overset{(C.2)}{=} \tau\bar{\boldsymbol{K}}(\cdot)\,\boldsymbol{q}(\cdot,u_\ell^i)+\boldsymbol{w}_N^i(u_\ell^{i+1}-u_\ell^i)$, *which only has the reaction–diffusion operator given by* L_N^i *and* $\mathfrak{a}_N^i$ *on the left-hand side and the computed* u_ℓ^{i+1} *on the right-hand side. Define a posteriori estimators as in* (5.12)*, following the steps of Section* 5.2 *with in particular the equilibration given by Definition* 5.4*. Let* $\vartheta_{\Omega,\ell,i} \geq 1$ *be as in Theorem* 5.7*. Finally, let the constants* $\theta_{N,\mathrm{m}}$ *and* $\theta_{N,\mathrm{M}}$*, which converge to* 1 *as* $C_N \to 0$*, be defined as*

$$\theta_{N,\mathrm{m}} := \inf_{t\geq 0} \frac{t^2 - C_{\mathrm{N}}(2-C_{\mathrm{N}})t + \frac{1}{(1+C_{\mathrm{N}})^2}}{(1-C_{\mathrm{N}}/2)^{-2}(t^2+1)} > -\infty, \qquad \theta_{N,\mathrm{M}} := \sup_{t\geq 0} \frac{t^2 + \frac{4C_{\mathrm{N}}}{2-C_{\mathrm{N}}}t + \frac{4}{(2-C_{\mathrm{N}})^2}}{(1+C_{\mathrm{N}})^{-2}(t^2+1)} \in (0,\infty). \tag{C.12}$$

Then, there holds

$$|||\mathcal{R}(u_\ell^i)|||_{-1,u_\ell^i}^2 \leq \theta_{N,\mathrm{M}}\,[\eta_\Omega^i]^2, \tag{guaranteed reliability}$$

and, assuming $\theta_{N,\mathrm{m}} > 0$*,*

$$\theta_{N,\mathrm{m}}[\eta_\Omega^i]^2 \lesssim \vartheta_{\Omega,\ell,i}^2|||\mathcal{R}(u_\ell^i)|||_{-1,u_\ell^i}^2 + \theta_{N,\mathrm{m}}\left[[\eta_{\mathrm{osc},\Omega}^i]^2 + \vartheta_{\Omega,\ell,i}^2[\eta_{\mathrm{quad,F},\Omega}^i]^2 + [\eta_{\mathrm{quad,S},\Omega}^i]^2 + [\eta_{\mathrm{quad,S,osc},\Omega}^i]^2\right], \tag{global efficiency}$$

where the hidden constant in $\lesssim$ *has the same dependence as in Theorem* 5.7*.*

Remark C.7 (The asymmetry factors from (C.12) $\theta_{N,\mathrm{m}}$ and $\theta_{N,\mathrm{M}}$)**.** *The asymmetry factors* $\theta_{N,\mathrm{m}}$ *and* $\theta_{N,\mathrm{M}}$ *are introduced to simplify in Theorem* C.6 *the bounds introduced in Theorem* C.4 *and to make them similar to those of Theorem* 5.7*. Observe that* $\theta_{N,\mathrm{m}}$ *and* $\theta_{N,\mathrm{M}}$ *are computable by calculating the extremas of the functions in* (C.12)*. Thus, similarly to the variability constant* $\vartheta_{\Omega,\ell,i}$*, they can be estimated a posteriori.*

Proof. As in Section 8, the dual norm of the discretization residual $\mathcal{R}_{\mathrm{disc}}^{u_\ell^i}(u_\ell^{i+1})$ is estimated above and below by (8.10) and (5.14b) respectively, since the steps to obtain the estimates remain exactly the same after redefining $\boldsymbol{\mathcal{F}}^i$. Then the estimates are proven by using (C.12) in Theorem C.4. We omit the details for the sake of conciseness. □

C.4 Numerical experiments

We repeat here the numerical experiments of Section 6.2 for the Newton linearization (C.3). The effectivity indices are defined as in (6.2), where $u_{\mathrm{ref}}^{i+1} \in V_{10}$ now solves the problem

$$(L_{\mathrm{N}}^i(u_{\mathrm{ref}}^{i+1}-u_\ell^i),\varphi_{\mathrm{ref}}) + (\mathfrak{a}_{\mathrm{N}}^i\nabla(u_{\mathrm{ref}}^{i+1}-u_\ell^i),\nabla\varphi_{\mathrm{ref}}) = -\langle\mathcal{R}(u_\ell^i),\varphi_{\mathrm{ref}}\rangle \quad \forall\varphi_{\mathrm{ref}} \in V_{10}, \tag{C.13}$$

thus, yielding $|||u_\ell^i - u_{\mathrm{ref}}^{i+1}|||_{1,u_\ell^i} \approx |||\mathcal{R}(u_\ell^i)|||_{-1,u_\ell^i}$. Figure 14 shows the results. The global effectivity indices defined in (6.2) can go below 1 in this case, since the factor $\theta_{N,\mathrm{M}}$ is not included. This is observed for $\tau = 1$ at iteration $i = 3$. However, the global effectivity indices stay between 1 and 2.5 in all other cases. Local effectivity indices also remain close to 1 in most of the domain. We find the agreement to be excellent. However, it is to be noted that the estimators only indicate the actual error if Assumption C.1 holds, which cannot always be guaranteed a priori. The quadratic convergence property of the Newton scheme is exhibited in the right plot of Figure 14.

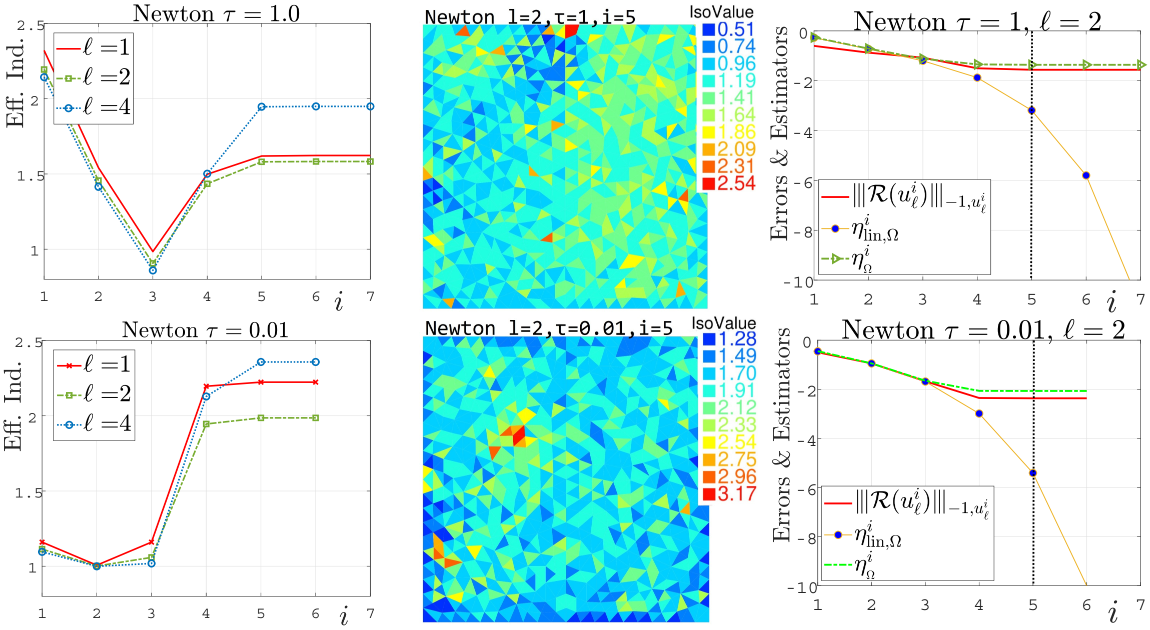

Figure 14: [Appendix C, setting of Section 6.2, τ varies] Numerical results for the Newton scheme, $\tau = 1$ (top) and $\tau = 0.01$ (bottom). Global effectivity indices (left); local effectivity indices for $\ell = 2$ at the iteration $i = \bar{i}$ from (5.16) (center); the total error $|||\mathcal{R}(u_\ell^i)|||_{-1,u_\ell^i}$, the linearization estimator $\eta^i_{\mathrm{lin},\Omega}$, and the total estimator η^i_Ω against the iteration index i (right, logarithmic scale). The vertical black-dotted lines indicate the stopping criterion $i = \bar{i}$ from (5.16) with $\mu = 0.05$. The quadratic convergence property of the Newton scheme can clearly be observed from the ever-steepening slope of $\eta^i_{\mathrm{lin},\Omega}$.

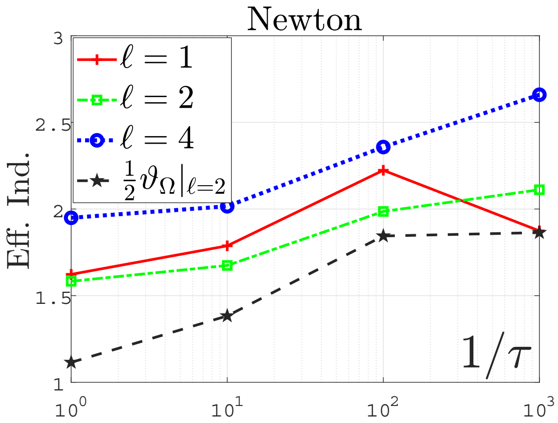

Figure 15: [Appendix C, setting of Section 6.2, τ varies] Global effectivity indices (6.2a) vs. $1/\tau$ for the Newton linearization at different discretization levels ℓ for $i = \bar{i}$ from (5.16). The variability constant $\vartheta_{\Omega,\ell,i}$ from (5.15) is also plotted (scaled by half and for $\ell = 2$).

Figure 15 plots the global effectivity indices (6.2a) against the strength of the nonlinearity represented by $1/\tau$. These indices remain between 1.5 and 2.6, numerically confirming the robustness of our a posteriori estimators. In fact, the obtained effectivity indices for the Newton scheme are almost identical to those of the Picard scheme shown in Figure 9. This results from the fact that the advection term $(\boldsymbol{w}^i_{\mathrm{N}}(u_\ell^{i+1} - u_\ell^i), \nabla\varphi_\ell)$ in (C.3) is negligible at $i = \bar{i}$ compared to the symmetric part of $\mathfrak{L}_{\mathrm{N}}$, which makes the two schemes almost coincide.

Acknowledgements

K. Mitra was supported by Fonds voor Wetenschappelijk Onderzoek (FWO) through the Junior Postdoctoral Fellowship during the completion of this work.

References

[1] Ainsworth, M., and Babuška, I. Reliable and robust a posteriori error estimation for singularly perturbed reaction-diffusion problems. *SIAM J. Numer. Anal.* **36** (1999), 331–353.

[2] Ainsworth, M., and Vejchodský, T. Robust error bounds for finite element approximation of reaction-diffusion problems with non-constant reaction coefficient in arbitrary space dimension. *Comput. Methods Appl. Mech. Engrg.* **281** (2014), 184–199. http://dx.doi.org/10.1016/j.cma.2014.08.005.

[3] Bartels, S., and Milicevic, M. Primal-dual gap estimators for *a posteriori* error analysis of nonsmooth minimization problems. *ESAIM Math. Model. Numer. Anal.* **54** (2020), 1635–1660. https://doi.org/10.1051/m2an/2019074.

[4] Bernardi, C., Dakroub, J., Mansour, G., and Sayah, T. A posteriori analysis of iterative algorithms for a nonlinear problem. *J. Sci. Comput.* **65** (2015), 672–697. https://doi.org/10.1007/s10915-014-9980-4.

[5] Bernardi, C., and Verfürth, R. Adaptive finite element methods for elliptic equations with non-smooth coefficients. *Numer. Math.* **85** (2000), 579–608. https://doi.org/10.1007/PL00005393.

[6] Blechta, J., Málek, J., and Vohralík, M. Localization of the $W^{-1,q}$ norm for local a posteriori efficiency. *IMA J. Numer. Anal.* **40** (2020), 914–950. https://doi.org/10.1093/imanum/drz002.

[7] Boffi, D., Brezzi, F., and Fortin, M. *Mixed finite element methods and applications*, vol. **44** of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2013. https://doi.org/10.1007/978-3-642-36519-5.

[8] Botti, M., and Riedlbeck, R. Equilibrated stress tensor reconstruction and a posteriori error estimation for nonlinear elasticity. *Comput. Methods Appl. Math.* **20** (2020), 39–59. https://doi.org/10.1515/cmam-2018-0012.

[9] Braess, D., Pillwein, V., and Schöberl, J. Equilibrated residual error estimates are p-robust. *Comput. Methods Appl. Mech. Engrg.* **198** (2009), 1189–1197. http://dx.doi.org/10.1016/j.cma.2008.12.010.

[10] Braess, D., and Schöberl, J. Equilibrated residual error estimator for edge elements. *Math. Comp.* **77** (2008), 651–672. http://dx.doi.org/10.1090/S0025-5718-07-02080-7.

[11] Buhr, A., Engwer, C., Ohlberger, M., and Rave, S. ArbiLoMod, a simulation technique designed for arbitrary local modifications. *SIAM J. Sci. Comput.* **39** (2017), A1435–A1465. https://doi.org/10.1137/15M1054213.

[12] Cai, D., Cai, Z., and Zhang, S. Robust equilibrated a posteriori error estimator for higher order finite element approximations to diffusion problems. *Numer. Math.* **144** (2020), 1–21. https://doi.org/10.1007/s00211-019-01075-1.

[13] Celia, M., Bouloutas, E., and Zarba, R. General mass-conservative numerical solution for the unsaturated flow equation. *Water Resources Research* **26** (1990), 1483–1496.

[14] Chaillou, A., and Suri, M. A posteriori estimation of the linearization error for strongly monotone nonlinear operators. *J. Comput. Appl. Math.* **205** (2007), 72–87. http://dx.doi.org/10.1016/j.cam.2006.04.041.

[15] Chaillou, A. L., and Suri, M. Computable error estimators for the approximation of nonlinear problems by linearized models. *Comput. Methods Appl. Mech. Engrg.* **196** (2006), 210–224. http://dx.doi.org/10.1016/j.cma.2006.03.008.

[16] Ciarlet, Jr., P., and Vohralík, M. Localization of global norms and robust a posteriori error control for transmission problems with sign-changing coefficients. *ESAIM Math. Model. Numer. Anal.* **52** (2018), 2037–2064. https://doi.org/10.1051/m2an/2018034.

[17] Ciarlet, P. G. *The Finite Element Method for Elliptic Problems*, vol. **4** of *Studies in Mathematics and its Applications*. North-Holland, Amsterdam, 1978.

[18] Cohen, A., Dahmen, W., and DeVore, R. Adaptive wavelet schemes for nonlinear variational problems. *SIAM J. Numer. Anal.* **41** (2003), 1785–1823. http://dx.doi.org/10.1137/S0036142902412269.

[19] Cottereau, R., Chamoin, L., and Díez, P. Strict error bounds for linear and nonlinear solid mechanics problems using a patch-based flux-free method. *Mechanics & Industry* **11** (2010), 249–254. https://doi.org/10.1051/meca/2010049.

[20] Cottereau, R., Díez, P., and Huerta, A. Strict error bounds for linear solid mechanics problems using a subdomain-based flux-free method. *Comput. Mech.* **44** (2009), 533–547. https://doi.org/10.1007/s00466-009-0388-1.

[21] Destuynder, P., and Métivet, B. Explicit error bounds in a conforming finite element method. *Math. Comp.* **68** (1999), 1379–1396. http://dx.doi.org/10.1090/S0025-5718-99-01093-5.

[22] Deuflhard, P. *Newton methods for nonlinear problems: affine invariance and adaptive algorithms*, vol. **35**. Springer Science & Business Media, Berlin, 2005.

[23] El Alaoui, L., Ern, A., and Vohralík, M. Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems. *Comput. Methods Appl. Mech. Engrg.* **200** (2011), 2782–2795. http://dx.doi.org/10.1016/j.cma.2010.03.024.

[24] Ern, A., and Vohralík, M. Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs. *SIAM J. Sci. Comput.* **35** (2013), A1761–A1791. http://dx.doi.org/10.1137/120896918.

[25] Ern, A., and Vohralík, M. Stable broken H^1 and $\boldsymbol{H}(\mathrm{div})$ polynomial extensions for polynomial-degree-robust potential and flux reconstruction in three space dimensions. *Math. Comp.* **89** (2020), 551–594. http://dx.doi.org/10.1090/mcom/3482.

[26] Evans, L. *Partial differential equations*, vol. **19**. American Mathematical Society, 2022.

[27] Gantner, G., Haberl, A., Praetorius, D., and Schimanko, S. Rate optimality of adaptive finite element methods with respect to overall computational costs. *Math. Comp.* **90** (2021), 2011–2040. https://doi.org/10.1090/mcom/3654.

[28] Garau, E. M., Morin, P., and Zuppa, C. Convergence of an adaptive Kačanov FEM for quasi-linear problems. *Appl. Numer. Math.* **61** (2011), 512–529. https://doi.org/10.1016/j.apnum.2010.12.001.

[29] Guo, M., Han, W., and Zhong, H. Legendre-Fenchel duality and a generalized constitutive relation error. arXiv preprint 1611.05589, https://arxiv.org/abs/1611.05589, 2016.

[30] Haberl, A., Praetorius, D., Schimanko, S., and Vohralík, M. Convergence and quasi-optimal cost of adaptive algorithms for nonlinear operators including iterative linearization and algebraic solver. *Numer. Math.* **147** (2021), 679–725. https://doi.org/10.1007/s00211-021-01176-w.

[31] Han, W. A posteriori error analysis for linearization of nonlinear elliptic problems and their discretizations. *Math. Methods Appl. Sci.* **17** (1994), 487–508. http://dx.doi.org/10.1002/mma.1670170702.

[32] Han, W. *A posteriori error analysis via duality theory*, vol. **8** of *Advances in Mechanics and Mathematics*. Springer-Verlag, New York, 2005.

[33] Harnist, A., Mitra, K., Rappaport, A., and Vohralík, M. Robust energy a posteriori estimates for nonlinear elliptic problems. HAL Preprint 04033438, https://hal.inria.fr/hal-04033438, 2023.

[34] Hecht, F. New development in FreeFem++. *J. Numer. Math.* **20** (2012), 251–265. https://doi.org/10.1515/jnum-2012-0013.

[35] Heid, P., Praetorius, D., and Wihler, T. P. Energy contraction and optimal convergence of adaptive iterative linearized finite element methods. *Comput. Methods Appl. Math.* **21** (2021), 407–422. https://doi.org/10.1515/cmam-2021-0025.

[36] Heid, P., and Wihler, T. Adaptive iterative linearization Galerkin methods for nonlinear problems. *Mathematics of Computation* **89** (2020), 2707–2734. https://doi.org/10.1090/mcom/3545.

[37] Holst, M., Szypowski, R., and Zhu, Y. Adaptive finite element methods with inexact solvers for the nonlinear Poisson-Boltzmann equation. In *Domain decomposition methods in science and engineering XX*, vol. **91** of *Lect. Notes Comput. Sci. Eng.* Springer, Heidelberg, 2013, pp. 167–174. https://doi.org/10.1007/978-3-642-35275-1.

[38] Houston, P., Süli, E., and Wihler, T. A posteriori error analysis of *hp*-version discontinuous Galerkin finite-element methods for second-order quasi-linear elliptic PDEs. *IMA Journal of Numerical Analysis* **28** (2008), 245–273. https://doi.org/10.1093/imanum/drm009.

[39] Jäger, W., and Kačur, J. Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes. *ESAIM: Mathematical Modelling and Numerical Analysis* **29** (1995), 605–627.

[40] Karátson, J., and Korotov, S. Sharp upper global a posteriori error estimates for nonlinear elliptic variational problems. *Appl. Math.* **54** (2009), 297–336. http://dx.doi.org/10.1007/s10492-009-0020-x.

[41] Kim, K.-Y. A posteriori error estimators for locally conservative methods of nonlinear elliptic problems. *Appl. Numer. Math.* **57** (2007), 1065–1080. http://dx.doi.org/10.1016/j.apnum.2006.09.010.

[42] Kopteva, N. Energy-norm a posteriori error estimates for singularly perturbed reaction-diffusion problems on anisotropic meshes. *Numer. Math.* **137** (2017), 607–642. https://doi.org/10.1007/s00211-017-0889-3.

[43] Kopteva, N. Fully computable a posteriori error estimator using anisotropic flux equilibration on anisotropic meshes. ArXiv Preprint 1704.04404, https://arxiv.org/abs/1704.04404, 2017.

[44] Kovács, B. On the numerical performance of a sharp a posteriori error estimator for some nonlinear elliptic problems. *Appl. Math.* **59** (2014), 489–508. https://doi.org/10.1007/s10492-014-0068-0.

[45] Ladevèze, P., and Moës, N. A posteriori constitutive relation error estimators for nonlinear finite element analysis and adaptive control. In *Advances in adaptive computational methods in mechanics (Cachan, 1997)*, vol. **47** of *Stud. Appl. Mech.* Elsevier Sci. B. V., Amsterdam, 1998, pp. 231–256. https://doi.org/10.1016/S0922-5382(98)80013-5.

[46] Ladevèze, P., and Pelle, J.-P. *Mastering calculations in linear and nonlinear mechanics.* Springer-Verlag, New York, 2005.

[47] Ladyzhenskaya, O., Uraltseva, N., and Ehrenpreis, L. *Linear and quasilinear elliptic equations.* Academic Press, 1968.

[48] Li, H., Wu, Z., Yin, J., and Zhao, J. *Nonlinear diffusion equations.* World Scientific, Singapore, 2001.

[49] Lions, P. On the existence of positive solutions of semilinear elliptic equations. *SIAM review* **24** (1982), 441–467.

[50] List, F., and Radu, F. A study on iterative methods for solving Richards' equation. *Computational Geosciences* **20** (2016), 341–353. https://doi.org/10.1007/s10596-016-9566-3.

[51] Meyers, N. G. An L^p-estimate for the gradient of solutions of second order elliptic divergence equations. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3)* **17** (1963), 189–206.

[52] Mitra, K., and Pop, I. A modified L-scheme to solve nonlinear diffusion problems. *Computers & Mathematics with Applications* **77** (2019), 1722–1738. 7th International Conference on Advanced Computational Methods in Engineering (ACOMEN 2017), https://doi.org/10.1016/j.camwa.2018.09.042.

[53] Mitra, K., and Vohralík, M. A posteriori error estimates for the Richards equation. HAL Preprint 03328944, submitted for publication, https://hal.inria.fr/hal-03328944, 2022.

[54] Neittaanmaki, P., and Repin, S. A posteriori error identities for nonlinear variational problems. *Ann. Acad. Rom. Sci. Ser. Math. Appl.* **7** (2015), 157–172.

[55] Pop, I., Radu, F., and Knabner, P. Mixed finite elements for the Richards equation: linearization procedure. *Journal of Computational and Applied Mathematics* **168** (2004), 365–373. https://doi.org/10.1016/j.cam.2003.04.008.

[56] Pousin, J., and Rappaz, J. Consistency, stability, a priori and a posteriori errors for Petrov-Galerkin methods applied to nonlinear problems. *Numer. Math.* **69** (1994), 213–231. http://dx.doi.org/10.1007/s002110050088.

[57] Repin, S. *A posteriori estimates for partial differential equations*, vol. **4** of *Radon Series on Computational and Applied Mathematics*. Walter de Gruyter GmbH & Co. KG, Berlin, 2008. http://dx.doi.org/10.1515/9783110203042.

[58] Repin, S. I. A posteriori error estimation for nonlinear variational problems by duality theory. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **243** (1997), 201–214, 342.

[59] Repin, S. I. A posteriori error estimation for variational problems with uniformly convex functionals. *Math. Comp.* **69** (2000), 481–500. http://dx.doi.org/10.1090/S0025-5718-99-01190-4.

[60] Smears, I., and Vohralík, M. Simple and robust equilibrated flux a posteriori estimates for singularly perturbed reaction–diffusion problems. *ESAIM Math. Model. Numer. Anal.* **54** (2020), 1951–1973. https://doi.org/10.1051/m2an/2020034.

[61] Stokke, J. S., Mitra, K., Storvik, E., Both, J. W., and Radu, F. A. An adaptive solution strategy for Richards' equation. arXiv 2301.02055, https://doi.org/10.48550/arXiv.2301.02055, 2023.

[62] van Genuchten, M. A closed form for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc. Amer. J.* **44** (1980), 892–898.

[63] Verfürth, R. A posteriori error estimates for nonlinear problems. Finite element discretizations of elliptic equations. *Math. Comp.* **62** (1994), 445–475. http://www.jstor.org/stable/2153518.

[64] Verfürth, R. Robust a posteriori error estimators for a singularly perturbed reaction-diffusion equation. *Numer. Math.* **78** (1998), 479–493. https://doi.org/10.1007/s002110050322.

[65] Vohralík, M. Guaranteed and fully robust a posteriori error estimates for conforming discretizations of diffusion problems with discontinuous coefficients. *J. Sci. Comput.* **46** (2011), 397–438. http://dx.doi.org/10.1007/s10915-010-9410-1.

[66] Zarantonello, E. *Solving functional equations by contractive averaging.* Mathematics Research Center, United States Army, University of Wisconsin, 1960.

[67] Zeidler, E. *Applied Functional Analysis: Applications to Mathematical Physics (Applied Mathematical Sciences)(v. 108).* Springer, 1995.

[68] Zeidler, E. *Nonlinear Functional Analysis and Its Applications: II/B: Nonlinear Monotone Operators.* Springer Science & Business Media, 2013.