Research Article

Johannes Kraus*, Svetoslav Nakov and Sergey I. Repin

Reliable Numerical Solution of a Class of Nonlinear Elliptic Problems Generated by the Poisson–Boltzmann Equation

https://doi.org/10.1515/cmam-2018-0252
Received September 27, 2018; revised April 7, 2019; accepted May 3, 2019

Abstract: We consider a class of nonlinear elliptic problems associated with models in biophysics, which are described by the Poisson–Boltzmann equation (PBE). We prove mathematical correctness of the problem, study a suitable class of approximations, and deduce guaranteed and fully computable bounds of approximation errors. The latter goal is achieved by means of the approach suggested in [19] for convex variational problems. Moreover, we establish the error identity, which defines the error measure natural for the considered class of problems and show that it yields computable majorants and minorants of the global error as well as indicators of local errors that provide efficient adaptation of meshes. Theoretical results are confirmed by a collection of numerical tests that includes problems on 2D and 3D Lipschitz domains.

Keywords: Poisson–Boltzmann Equation, Semilinear Partial Differential Equations, Existence and Uniqueness of Solutions, Convergence of Finite Element Approximations, a Priori Error Estimates, Guaranteed and Efficient a Posteriori Error Bounds, Error Indicators and Adaptive Mesh Refinement

MSC 2010: 65J15, 49M29, 65N15, 65N30, 65N50, 35J20

1 Introduction

1.1 Classical Statement of the Problem

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ be a bounded domain with Lipschitz boundary $\partial\Omega$. We assume that Ω contains an interior subdomain Ω_1 with Lipschitz boundary Γ. In general, Ω_1 may consist of several disconnected parts (in this case, all of them are assumed to have Lipschitz continuous boundaries). We consider a class of nonlinear elliptic equations motivated by the Poisson–Boltzmann equation (PBE), which is widely used for computation of electrostatic interactions in a system of biomolecules in ionic solution [10, 11, 23],

$$-\nabla \cdot (\epsilon \nabla u) + k^2 \sinh(u + w) = l \quad \text{in } \Omega_1 \cup \Omega_2, \tag{1.1a}$$

$$[u]_\Gamma = 0, \tag{1.1b}$$

$$\left[\epsilon \frac{\partial u}{\partial n}\right]_\Gamma = 0, \tag{1.1c}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{1.1d}$$

***Corresponding author: Johannes Kraus,** Faculty of Mathematics, University of Duisburg-Essen, Thea-Leymann-Str. 9, 45127 Essen, Germany, e-mail: johannes.kraus@uni-due.de
Svetoslav Nakov, RICAM, Austrian Academy of Sciences, Altenbergerstraße 69, 4040 Linz, Austria, e-mail: svetoslav.nakov@oeaw.ac.at
Sergey I. Repin, University of Jyväskylä, Jyväskylän, Finland; and V. A. Steklov Institute of Mathematics in St. Petersburg, Russia, e-mail: serepin@jyu.fi, repin@pdmi.ras.ru

where $\Omega_2 := \Omega \setminus (\Omega_1 \cup \Gamma)$, the coefficients ϵ, $k \in L^\infty(\Omega)$, $\epsilon_{\max} \geq \epsilon \geq \epsilon_{\min} > 0$, w is measurable, and $l \in L^2(\Omega)$. Typically, in biophysical applications, Ω_1 is occupied by one or more macromolecules, and Ω_2 is occupied by a solution of water and moving ions. The coefficients ϵ and k represent the dielectric constant and the modified Debye–Hückel parameter, and u is the dimensionless electrostatic potential. Concerning the given functions k and w, we can identify three main cases:

(a) $k_{\max} \geq k(x) \geq k_{\min} > 0$ in Ω and $w \in L^\infty(\Omega)$,

(b) $k(x) \equiv 0$ in Ω_1, $k_{\max} \geq k(x) \geq k_{\min} > 0$ in Ω_2 and $w \in L^\infty(\Omega_2)$,

(c) $k(x) \equiv 0$ in Ω_2, $k_{\max} \geq k(x) \geq k_{\min} > 0$ in Ω_1 and $w \in L^\infty(\Omega_1)$.

Throughout the paper, the major attention is paid to case (b), which arises when solving the PBE and which is the most interesting from the practical point of view. Cases (a) and (c) can be studied analogously (with some rather obvious modifications). The case with nonhomogeneous Dirichlet boundary condition $u = g$ on $\partial\Omega$ can also be treated in this framework provided that the boundary condition is defined as the trace of a function g such that $g \in H^1(\Omega) \cap L^\infty(\Omega)$ and $\nabla g \in L^s(\Omega)$ with $s > \max\{2, d\}$.

The ability to find reliable and efficient solutions of the nonlinear Poisson–Boltzmann equation (PBE) for complex geometries of the interior domain Ω_1 (with Lipschitz boundary) and piecewise constant dielectrics is important for applications in biophysics and biochemistry, e.g., in modeling the effects of water and ion screening on the potentials in and around soluble proteins, nucleic acids, membranes, and polar molecules and ions; see [23] and the references therein. Although the solution of the linearized PBE (as in the linear Debye–Hückel theory) often yields accurate approximations [22], certain mathematical models are valid only if they are based on the nonlinear PBE.

Over the recent years, adaptive finite element methods have proved to be an adequate technique in the numerical solution of elliptic problems with sharp local features such as point sources, heterogeneous coefficients or nonsmooth boundaries or interfaces (e.g., see [4, 9] and also have been successfully used to solve the nonlinear PBE [5, 14]. Adaptivity heavily relies on reliable and efficient error indicators, which are typically developed in the framework of a posteriori error control methods. While the theory of a posteriori error estimates for linear elliptic partial differential equations is already well established and understood, it is far less developed for nonlinear problems. A posteriori error analysis based on functional estimates has already been successfully applied to variational nonlinear problems including obstacle problems in [20, 21]. The accuracy verification approach taken in this work is also based on arguments that are commonly used in duality theory and convex analysis and can be found, e.g., in [8, 17]. Fast solution methods for systems of nonlinear differential equations is another important issue that affects efficiency of computer simulation methods. Multigrid methods may provide optimal or nearly optimal algorithms (in terms of complexity) to perform this task (e.g., see [18]). However, a systematic discussion of this topic is beyond the framework of this paper.

The main questions studied in the paper are related to the well-posedness of problem (1.1) and a posteriori error estimation of its numerical solution. We use a suitable weak formulation (Definition 2.1), where the nonlinearity does not satisfy any polynomial growth condition, and consequently it does not induce a bounded mapping from $H^1_0(\Omega)$ to its dual $H^{-1}(\Omega)$. For this (more general) weak formulation, we can guarantee existence of a solution and prove its uniqueness using a result of Brezis and Browder [3]. Additionally, in Proposition 2.1, we show that the solution is bounded (here [3] is used, again together with special test functions suggested in [16, 26]). Boundedness of the solution is important and later used in the derivation of functional a posteriori error estimates. Applying the general approach from [17, 19], we derive guaranteed and computable bounds of the difference between the exact solution and any function from the respective energy class in terms of the energy and combined energy norms (equation (3.20)). Moreover, we obtain an error identity (3.19) with respect to a certain measure for the error which is the sum of the usual combined energy norm $|||\nabla(v-u)|||^2 + |||y^* - p^*|||^2_*$ and a nonlinear measure. In the case of a linear elliptic equation of the form $-\operatorname{div}(\epsilon\nabla u) + u = l$, this nonlinear measure reduces to $\|v-u\|^2_{L^2(\Omega)} + \|\operatorname{div}(y^* - p^*)\|^2_{L^2(\Omega)}$, where v and y^* are approximations to the exact solution u and the exact flux $p^* = \epsilon\nabla u$. One advantage of the presented error estimate is that it is valid for any conforming approximations of u and $\epsilon\nabla u$ and that it does not rely on Galerkin orthogonality or properties specific to the used numerical method. Another advantage is that only the mathematical structure of the problem is exploited, and therefore no mesh dependent constants are present in the estimate. Majorants of the error not only give guaranteed bounds of global (energy) error

norms but also generate efficient error indicators (cf. (1.1a), Figures 12 and 13). Also, we derive a simple, but efficient lower bound for the error in the combined energy norm. Using only the error majorant, we obtain an analog of Cea's lemma which forms a basis for the a priori convergence analysis of finite element approximations for this class of semilinear problems. Finally, we present three numerical examples that verify the accuracy of error majorants and minorants and confirm efficiency of the error indicator in mesh adaptive procedures.

The outline of the paper is as follows. In Section 2, we discuss correctness of problem (1.1) and prove an a priori $L^\infty(\Omega)$ estimate for the solution u. In Section 3, first we recall some facts from the duality theory and general a posteriori error estimation method for convex variational problems. Then we apply this abstract framework and derive explicit forms of all the respective terms. A special attention is paid to the general error identity that defines a combined error measure natural for the considered class of problems. At the end of Section 3, we prove convergence of the conforming finite element method based on P_1 Lagrange elements. In Section 4, we consider numerical examples associated with 2D and 3D problems and make a systematic comparison of numerical solutions computed by adaptive mesh refinements based on different indicators. The last section includes a summary of the presented results.

2 Variational Form of the Problem

From now on, we assume that the functions k and w fall in case (b) of Section 1.1.

Definition 2.1. A function $u \in H^1_0(\Omega)$ is called a weak solution of (1.1) if u is such that $b(x, u+w)v \in L^1(\Omega)$ for any $v \in H^1_0(\Omega) \cap L^\infty(\Omega)$ and

$$a(u, v) + \int_\Omega b(x, u+w)v\,dx = \int_\Omega lv\,dx \quad \text{for all } v \in H^1_0(\Omega) \cap L^\infty(\Omega), \tag{2.1}$$

where $a(u, v) = \int_\Omega \epsilon \nabla u \cdot \nabla v\,dx$ and $b(x, z) := k^2(x)\sinh(z)$.

Define the functional $J\colon H^1_0(\Omega) \to \mathbb{R} \cup \{+\infty\}$ by the relation

$$J(v) := \begin{cases} \int_\Omega \left[\frac{\epsilon(x)}{2}|\nabla v|^2 + k^2\cosh(v+w) - lv\right] dx & \text{if } k^2\cosh(v+w) \in L^1(\Omega), \\ +\infty & \text{if } k^2\cosh(v+w) \notin L^1(\Omega), \end{cases} \tag{2.2}$$

and consider the variational problem:

$$\text{find } u \in H^1_0(\Omega) \text{ such that} \quad J(u) = \min_{v \in H^1_0(\Omega)} J(v). \tag{2.3}$$

2.1 Existence of a Minimizer

We begin with proving that the variational problem is well-posed. First it is necessary to make some comments on specific features of the above defined variational problem associated with the term $k^2\cosh(v+w)$. Notice that, for $d \le 2$, the function $e^v \in L^2(\Omega)$ for all $v \in H^1_0(\Omega)$ (e.g., see [15, 27]) and, therefore, the set $\mathrm{dom}(J) := \{v \in H^1_0(\Omega) : J(v) < \infty\}$ is a linear subspace of $H^1_0(\Omega)$. However, if $d = 3$, then $\mathrm{dom}(J)$ is a convex set but not a linear subspace (Remark 2.1). Since $\mathrm{dom}(J)$ is convex and obviously J is convex over $\mathrm{dom}\,J$, it follows that J is convex over $H^1_0(\Omega)$ (e.g., see [8]). Next we note that J is a proper functional, i.e., J is not identically equal to $+\infty$ (e.g., $J(0) = \int_\Omega k^2\cosh(w)\,dx < \infty$) and does not take the value $-\infty$ ($J(v)$ is nonnegative). Therefore, existence of the minimizer u is guaranteed by known theorems of the calculus of variations (e.g., see [8]) if we will show that

(1) J is sequentially weakly lower semicontinuous (s.w.l.s.c.), i.e., $J(v) \le \liminf_{n\to\infty} J(v_n)$ for any sequence $\{v_n\}_{n=1}^\infty \subset H^1_0(\Omega)$ weakly converging to v in $H^1_0(\Omega)$) ($v_n \rightharpoonup v$),
(2) J is coercive, i.e., $\lim_{n\to\infty} J(v_n) = +\infty$ whenever $\|v_n\|_{H^1(\Omega)} \to \infty$.

To prove that (1) is fulfilled, notice that J is the sum of the functionals

$$\int_\Omega \Big(\frac{\epsilon}{2}|\nabla v|^2 - lv\Big)\,dx \quad \text{and} \quad \int_\Omega k^2(x)\cosh(v+w)\,dx.$$

The first functional is convex and continuous in $H_0^1(\Omega)$ and, therefore, it is s.w.l.s.c. (sequentially weakly lower semicontinuous). The second functional is convex and, for $d = 2$, it is Gateaux differentiable, which implies that it is also s.w.l.s.c. (the proof of this implication can be found in [24, Corollary 2.4]). However, for $d = 3$, the functional $\int_\Omega k^2(x)\cosh(v+w)\,dx$ is not Gateaux differentiable (Remark 2.2). Nevertheless, one can show that it is also s.w.l.s.c. For this purpose, we use Fatou's lemma and compact embedding of $H_0^1(\Omega)$ into $L^2(\Omega)$.

Let $\{v_n\}_{n=1}^\infty \subset H_0^1(\Omega)$ be a sequence weakly converging in $H_0^1(\Omega)$ to a $v \in H_0^1(\Omega)$, i.e., $v_n \rightharpoonup v$. Since the embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$ is compact, it follows that $v_n \to v$ (strongly) in $L^2(\Omega)$. Therefore, we can extract a subsequence $v_{n_m}(x) \to v(x)$, which converges almost everywhere in the pointwise sense. Recall that $k^2(x)\cosh(z(x)+w(x)) \geq 0$ for all $z \in H_0^1(\Omega)$ and that $k^2(x)\cosh(\cdot)$ is a continuous function for a.e. $x \in \Omega$. Hence, by Fatou's lemma, we obtain

$$\begin{aligned}\liminf_{m\to\infty}\int_\Omega k^2(x)\cosh(v_{n_m}(x)+w(x))\,dx &\geq \int_\Omega \liminf_{m\to\infty} k^2(x)\cosh(v_{n_m}(x)+w(x))\,dx\\ &= \int_\Omega k^2(x)\cosh(v(x)+w(x))\,dx. \qquad (2.4)\end{aligned}$$

Now it is clear that if $\{v_{n_m}\}_{m=1}^\infty$ is an arbitrary subsequence of $\{v_n\}_{n=1}^\infty$, then there exists a further subsequence $\{v_{n_{m_s}}\}_{s=1}^\infty$ for which (2.4) is satisfied. This means that (2.4) is also satisfied for the whole sequence $\{v_n\}_{n=1}^\infty$, and hence $\int_\Omega k^2\cosh(v+w)\,dx$ is s.l.w.s.c.

The coercivity of J follows from the estimate

$$\begin{aligned}J(v) &\geq \frac{1}{2}a(v,v) - \|l\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \geq \epsilon_{\min}\|\nabla v\|^2_{L^2(\Omega)} - \|l\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)}\\ &\geq \frac{\epsilon_{\min}}{1+C_{\mathrm{F}}^2}\|v\|^2_{H^1(\Omega)} - \|l\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)},\end{aligned}$$

where C_{F} is the constant in the Friedrichs inequality $\|v\|_{L^2(\Omega)} \leq C_{\mathrm{F}}\|\nabla v\|_{L^2(\Omega)}$ for all $v \in H_0^1(\Omega)$.

Thus conditions (1) and (2) are fulfilled, and the existence of a minimizer $u \in H_0^1(\Omega)$ is guaranteed. Moreover, noting that J is strictly convex, we arrive at the following result.

Theorem 2.1. *There exists a unique minimizer $u \in H_0^1(\Omega)$ of problem* (2.3).

Remark 2.1. It is worth noting that $\mathrm{dom}(J) := \{v \in H_0^1(\Omega) : k^2(x)\cosh(v+w) \in L^1(\Omega)\}$ is a linear subspace of $H_0^1(\Omega)$ for $d \leq 2$ and not a linear subspace of $H_0^1(\Omega)$ if $d \geq 3$. In dimension $d \leq 2$, from [15, 27], we know that $\mathrm{e}^v \in L^2(\Omega)$ for any $v \in H_0^1(\Omega)$, and thus $\mathrm{e}^{\lambda v_1 + \mu v_2} \in L^2(\Omega)$ for any $\lambda, \mu \in \mathbb{R}$ and any $v_1, v_2 \in H_0^1(\Omega)$.

On the other hand, if $d \geq 3$, let $\Omega = B(0,1)$, and let the inner domain Ω_2 be given by $\Omega_2 = B(0,\bar{r})$ for some $\bar{r} < 1$, where $B(0,r)$ denotes the ball in $\mathbb{R}^d$ with radius r centered at zero. Consider the function $v = \ln\frac{1}{|x|} \in H_0^1(B(0,1))$. Since $\mathrm{e}^v = \frac{1}{|x|} \in L^1(\Omega_2)$ and $\mathrm{e}^{\lambda v} = \frac{1}{|x|^\lambda} \notin L^1(\Omega_2)$ for any $\lambda \geq d$,[1] we find on the one hand that

$$\begin{aligned}\int_\Omega k^2\cosh(v+w)\,dx &= \int_{\Omega_2} k^2\frac{(\mathrm{e}^{v+w}+\mathrm{e}^{-v-w})}{2}\,dx\\ &\leq \frac{1}{2}k_{\max}^2\mathrm{e}^{\|w\|_{L^\infty(\Omega_2)}}\int_{\Omega_2}(\mathrm{e}^v+\mathrm{e}^{-v})\,dx \leq \frac{1}{2}k_{\max}^2\mathrm{e}^{\|w\|_{L^\infty(\Omega_2)}}\Big(\int_{\Omega_2}\mathrm{e}^v\,dx + |\Omega_2|\Big) < \infty.\end{aligned}$$

On the other hand,

$$\int_\Omega k^2\cosh(\lambda v+w)\,dx \geq \frac{1}{2}\int_{\Omega_2} k^2\mathrm{e}^{\lambda v+w}\,dx \geq \frac{1}{2}k_{\min}^2\mathrm{e}^{-\|w\|_{L^\infty(\Omega_2)}}\int_{\Omega_2}\mathrm{e}^{\lambda v}\,dx = \infty \quad \text{for any } \lambda > d.$$

[1] For any d, using spherical coordinates, we have $\int_{B(0,1)}\frac{1}{|x|^\lambda}\,dx \sim \int_0^1\frac{1}{\rho^\lambda}\rho^{d-1}\,d\rho = \int_0^1\frac{1}{\rho^{\lambda-d+1}}\,d\rho < \infty$ if and only if $\lambda - d + 1 < 1$, i.e., if and only if $\lambda < d$.

Hence $v \in \operatorname{dom}(J)$, but $\lambda v \notin \operatorname{dom}(J)$ for any $\lambda \geq d$ and, therefore, $\operatorname{dom}(J)$ is not a linear subspace. However, $\operatorname{dom}(J) \subset H_0^1(\Omega)$ is a convex set. Indeed, let $v_1, v_2 \in \operatorname{dom}(J)$, i.e., $k^2 \cosh(v_1 + w), k^2 \cosh(v_2 + w) \in L^1(\Omega)$. Since $k^2 \cosh(\cdot)$ is convex for almost all $x \in \Omega$ and any $\lambda \in [0, 1]$, we have

$$\int_\Omega k^2 \cosh(\lambda v_1 + (1-\lambda)v_2 + w)\,dx \leq \lambda \int_\Omega k^2 \cosh(v_1 + w)\,dx + (1-\lambda)\int_\Omega k^2 \cosh(v_2 + w)\,dx < +\infty.$$

Hence $\operatorname{dom}(J)$ is a convex set.

Remark 2.2. The functional $\int_\Omega k^2 \cosh(v + w)\,dx$ is not Gateaux differentiable at any $u \in H_0^1(\Omega) \cap L^\infty(\Omega)$ if $d = 3$ (therefore, J is also not Gateaux differentiable). In fact, $\int_\Omega k^2 \cosh(v + w)\,dx$ is discontinuous at every $u \in H_0^1(\Omega) \cap L^\infty(\Omega)$. This fact is easy to see by the following example. Let $\Omega_2 = B(0, 2) \subset \Omega$ be the ball centered at 0 with radius 2. There exists a function $v \in H_0^1(\Omega)$ such that $\int_{\Omega_2} \mathrm{e}^{\lambda v}\,dx = +\infty$ for any $\lambda > 0$. In particular, we can set $v = \phi|x|^{-1/3}$, where ϕ is a smooth function equal to 1 in $B(0, 1)$ and 0 in $\mathbb{R}^3 \setminus B(0, 2)$. Then $v \in H_0^1(\Omega)$, but $\mathrm{e}^{\lambda v} \notin L^1(\Omega_2)$ for any $\lambda > 0$ since $\mathrm{e}^{\lambda v} > |x|^{-3}$ for small enough $|x|$. In this case, for any $u \in H_0^1(\Omega) \cap L^\infty(\Omega)$ and any $\lambda > 0$, we have

$$\int_\Omega k^2 \cosh(u + \lambda v + w)\,dx \geq \frac{1}{2}\int_{\Omega_2} k^2 \mathrm{e}^{u+\lambda v+w}\,dx \geq \frac{k_{\min}^2 \mathrm{e}^{-\|u+w\|_{L^\infty(\Omega_2)}}}{2}\int_{\Omega_2} \mathrm{e}^{\lambda v}\,dx = +\infty.$$

Now our goal is to show that the minimizer u is a solution of (2.1). To prove this, we use the Lebesgue dominated convergence theorem and the fact that, at the unique minimizer u of J, we have $k^2 \cosh(u + w) \in L^1(\Omega)$. Since $J(u + \lambda v) - J(u) \geq 0$ for all $v \in H_0^1(\Omega) \cap L^\infty(\Omega)$ and any $\lambda \geq 0$, we have

$$\begin{aligned}\frac{1}{2}a(u + \lambda v, u + \lambda v) + \int_\Omega k^2 \cosh(u + \lambda v + w)\,dx - \int_\Omega l(u + \lambda v)\,dx \\ - \frac{1}{2}a(u, u) - \int_\Omega k^2 \cosh(u + w)\,dx + \int_\Omega lu\,dx \geq 0.\end{aligned} \tag{2.5}$$

Making equivalent transformations of (2.5) and dividing by $\lambda > 0$, we obtain

$$a(u, v) + \lim_{\lambda \to 0^+} \int_\Omega \frac{k^2(\cosh(u + \lambda v + w) - \cosh(u + w))}{\lambda}\,dx - \int_\Omega lv\,dx \geq 0. \tag{2.6}$$

To compute the limit in the second term of (2.6), we will apply the Lebesgue dominated convergence theorem. We have

$$\begin{aligned}f_\lambda(x) &:= \frac{k^2(x)\big(\cosh(u(x) + w(x) + \lambda v(x)) - \cosh(u(x) + w(x))\big)}{\lambda} \\ &\xrightarrow{\lambda \to 0^+} k^2(x)\sinh(u(x) + w(x))v(x) \quad \text{for a.e. } x \in \Omega.\end{aligned} \tag{2.7}$$

By the mean value theorem, we obtain

$$f_\lambda(x) = k^2(x)\sinh(\varsigma(x))v(x),$$

where $\varsigma(x) := u(x) + w(x) + \Theta(x)\lambda v(x)$ and $\Theta(x) \in (0, 1)$, a.e. $x \in \Omega$. Then

$$|f_\lambda(x)| \leq k^2(x)\Big(\frac{\mathrm{e}^{\varsigma(x)} + \mathrm{e}^{-\varsigma(x)}}{2}\Big)|v(x)|,$$

from which it follows that

$$|f_\lambda| \leq \|v\|_{L^\infty(\Omega)} k^2 \frac{\mathrm{e}^{u+w} + \mathrm{e}^{-u-w}}{2}\mathrm{e}^{|\Theta(x)|\lambda|v(x)|} \leq \|v\|_{L^\infty(\Omega)}\mathrm{e}^{\|v\|_{L^\infty(\Omega)}} \underbrace{k^2 \cosh(u + w)}_{\in L^1(\Omega)} \quad \text{for all } \lambda \leq 1. \tag{2.8}$$

From the Lebesgue dominated convergence theorem, (2.7) and (2.8), it follows that the limit in (2.6) is equal to $\int_\Omega k^2 \sinh(u + w)v\,dx$, and therefore we obtain

$$a(u, v) + \int_\Omega b(x, u + w)v\,dx - \int_\Omega lv\,dx \geq 0 \quad \text{for all } v \in H_0^1(\Omega) \cap L^\infty(\Omega). \tag{2.9}$$

Since the test functions belong to a linear manifold, (2.9) is equivalent to the weak formulation (2.1).

2.2 Uniqueness of the Solution to (2.1)

Uniqueness of the solution of (2.1) follows from the monotonicity of $b(x,\cdot)$:

$$\int_\Omega (b(x, v+w) - b(x, z+w))(v-z)\,dx \geq 0 \quad \text{for all } v, z \in H_0^1(\Omega) \cap L^\infty(\Omega). \tag{2.10}$$

If $u_1, u_2 \in H_0^1(\Omega)$ are two different solutions of (2.1), then

$$a(u_1 - u_2, v) + \int_\Omega (b(x, u_1+w) - b(x, u_2+w))v\,dx = 0 \quad \text{for all } v \in H_0^1(\Omega) \cap L^\infty(\Omega). \tag{2.11}$$

Note that the difference $u_1 - u_2$ is not necessarily in $H_0^1(\Omega) \cap L^\infty(\Omega)$. To show that we can test with $u_1 - u_2$ in (2.11), we apply a property of Sobolev spaces proved in [3].

Theorem 2.2. *Let Ω be an open set in $\mathbb{R}^d$, $T \in H^{-1}(\Omega) \cap L^1_{\mathrm{loc}}(\Omega)$, and $v \in H_0^1(\Omega)$. If there exists a function $f \in L^1(\Omega)$ such that $T(x)v(x) \geq f(x)$ a.e in Ω, then $Tv \in L^1(\Omega)$ and the duality product $\langle T, v\rangle$ in $H^{-1}(\Omega) \times H_0^1(\Omega)$ coincides with $\int_\Omega Tv\,dx$.*

We have the following situation: a locally summable function $g \in L^1_{\mathrm{loc}}(\Omega)$ defines a bounded linear functional T_g over the dense subspace $C_0^\infty(\Omega)$ of $H_0^1(\Omega)$ through the integral formula $\langle T_g, \varphi\rangle = \int_\Omega g\varphi\,dx$. It is clear that the functional T_g is uniquely extendable by continuity to a bounded linear functional $\overline{T}_g$ over the whole space $H_0^1(\Omega)$. The question is whether this extension is still representable by the same integral formula for any $v \in H_0^1(\Omega)$ (if the integral makes sense at all). If the function $v \in H_0^1(\Omega)$ is fixed, then Theorem 2.2 gives a sufficient condition for gv to be summable and for the extension $\overline{T}_g$ evaluated at v to be representable with the same integral formula as above, i.e., $\langle \overline{T}_g, v\rangle = \int_\Omega gv\,dx$.

Now, applying Theorem 2.2 to the functional T_g defined by

$$\langle T_g, v\rangle := \langle b(x, u_1+w) - b(x, u_2+w), v\rangle \quad \text{for all } v \in H_0^1(\Omega) \cap L^\infty(\Omega)$$

and the function $v = u_1 - u_2 \in H_0^1(\Omega)$, using (2.11), we conclude that

$$a(u_1 - u_2, u_1 - u_2) + \int_\Omega (b(x, u_1+w) - b(x, u_2+w))(u_1 - u_2)\,dx = 0.$$

Using the monotonicity (2.10) of $b(x,\cdot)$ and the coercivity of $a(\cdot,\cdot)$, we obtain $u_1 = u_2$.

2.3 Boundedness of the Minimizer

Next we show that the solution to problem (2.1) is essentially bounded. To prove this, we need the following lemma [16].

Lemma 2.1. *Let $\varphi(t)$ be a nonnegative function, which is nonincreasing for $s_0 \leq t < \infty$ and such that*

$$\varphi(h) \leq C\frac{\varphi(s)^\beta}{(h-s)^\alpha} \quad \text{for all } h > s > s_0,$$

where C and α are positive constants and $\beta > 1$. If $j \in \mathbb{R}$ is defined by $j^\alpha := C\varphi(s_0)^{\beta-1}2^{\frac{\alpha\beta}{\beta-1}}$, then $\varphi(s_0 + j) = 0$.

Now we present a main result of this section.

Proposition 2.1. *The unique weak solution u to problem (2.1) belongs to $L^\infty(\Omega)$. Moreover, there exists a positive constant $\bar{j} > 0$, depending only on d, Ω, $\|l\|_{L^2(\Omega)}$, $\epsilon_{\min}$, such that $\|u\|_{L^\infty(\Omega)} \leq \|w\|_{L^\infty(\Omega_2)} + \bar{j}$. If $l = 0$, then the constant $\bar{j}$ is equal to zero.*

Proof. To prove that u is bounded, we apply Theorem 2.2 once again.

The first step is to show that (2.1) holds for the test function

$$v = G_s(u) := \operatorname{sgn}(u)\max\{|u| - s, 0\}, \tag{2.12}$$

where $s \geq \|w\|_{L^\infty(\Omega_2)}$ (we notice that similar test functions G_s have been used in [16, Theorem B.2] in the context of linear elliptic problems).

It is easy to see that $G_s(0) = 0$, this function is Lipschitz continuous and, therefore, $G_s(u) \in H_0^1(\Omega)$ (e.g., see [12, 16]). Next the functional T_b defined by

$$\langle T_b, v\rangle := \int_\Omega b(x, u + w)v\,dx \quad \text{for all } v \in H_0^1(\Omega) \cap L^\infty(\Omega)$$

is bounded and linear and $b(x, u + w) \in L^1_{\mathrm{loc}}(\Omega)$. This follows from (2.1) and from the fact that the functionals $a(u, \cdot)$ and $(l, \cdot)$ belong to $H^{-1}(\Omega)$. In view of Theorem 2.2, to show that $\langle T_b, G_s(u)\rangle = \int_\Omega b(x, u + w)G_s(u)\,dx$, it suffices to verify the inequality

$$b(x, u + w)G_s(u) \geq f \quad \text{a.e. for some } f \in L^1(\Omega). \tag{2.13}$$

Choosing $s \geq \|w\|_{L^\infty(\Omega_2)}$, using the monotonicity of $b(x, \cdot)$, and the fact that $b(x, 0) = 0$, we obtain

$$b(x, u + w)G_s(u) = \begin{cases} b(x, u + w)(u - s) \geq 0 & \text{for } u > s, \\ 0 & \text{for } u \in [-s, s], \\ b(x, u + w)(u + s) \geq 0 & \text{for } u < -s, \end{cases} \tag{2.14}$$

which shows that assumption (2.13) holds for $f = 0$.

Now we are ready to prove that $u \in L^\infty(\Omega)$. First we consider the case $l = 0$. From (2.14), it follows that

$$\int_\Omega b(x, u + w)G_s(u)\,dx \geq 0. \tag{2.15}$$

Moreover, using the definition of $a(\cdot, \cdot)$ and the definition (2.12) of $G_s(u)$, we obtain

$$\begin{aligned} a(u, G_s(u)) &= \int_\Omega \epsilon\nabla u \cdot \nabla G_s(u)\,dx = \int_\Omega \epsilon\nabla G_s(u) \cdot \nabla G_s(u)\,dx \\ &\geq \epsilon_{\min}\|\nabla G_s(u)\|^2_{L^2(\Omega)} \geq \frac{\epsilon_{\min}}{C_{\mathrm{F}}^2}\|G_s(u)\|^2_{L^2(\Omega)}, \end{aligned} \tag{2.16}$$

where C_{F} is the constant in Friedrichs' inequality $\|v\|_{L^2(\Omega)} \leq C_{\mathrm{F}}\|\nabla v\|_{L^2(\Omega)}$ that holds for all $v \in H_0^1(\Omega)$. Finally, using (2.1), (2.15), and (2.16), we get $\|G_s(u)\|^2_{L^2(\Omega)} \leq 0$ for all $s \geq \|w\|_{L^\infty(\Omega_2)}$. Consequently, $|u| \leq s$ almost everywhere and for all $s \geq \|w\|_{L^\infty(\Omega_2)}$. In the case where l is not identically zero in Ω, we further estimate $a(u, G_s(u))$ from below and $\int_\Omega lG_s(u)\,dx$ from above using the Sobolev embedding $H^1(\Omega) \hookrightarrow L^q(\Omega)$, where $q = \infty$ for $d = 1$, $q < \infty$ for $d = 2$, and $q = \frac{2d}{d-2}$ for $d \geq 3$. Let q^* denote the Hölder conjugate to q. Then $q^* = 1$ for $d = 1$, $q^* = \frac{q}{q-1} > 1$ for $d = 2$, and $q^* = \frac{2d}{d+2}$ for $d > 2$. In order to treat both cases in which we are interested simultaneously, namely, $d = 2, 3$, we can take $q = 6$ and $q^* = 6/5$. By C_{E} we denote the embedding constant in the inequality $\|v\|_{L^6(\Omega)} \leq C_{\mathrm{E}}\|v\|_{H^1(\Omega)}$ for all $v \in H^1(\Omega)$, which depends only on the domain Ω and d. For $a(u, G_s(u))$, we have

$$a(u, G_s(u)) = \int_\Omega \epsilon\nabla G_s(u) \cdot \nabla G_s(u)\,dx \geq \frac{\epsilon_{\min}}{1 + C_{\mathrm{F}}^2}\|G_s(u)\|^2_{H^1(\Omega)} \tag{2.17}$$

and, for $\int_\Omega lG_s(u)\,dx$, we obtain

$$\int_\Omega lG_s(u)\,dx = \int_{A(s)} lG_s(u)\,dx \leq \|l\|_{L^{q^*}(A(s))}\|G_s(u)\|_{L^q(\Omega)} \leq C_{\mathrm{E}}\|l\|_{L^{q^*}(A(s))}\|G_s(u)\|_{H^1(\Omega)}, \tag{2.18}$$

where $A(s) := \{x \in \Omega : |u(x)| > s\}$. Combining (2.17), (2.18), (2.15), and (2.1), we arrive at the estimate

$$\frac{\epsilon_{\min}}{1 + C_{\mathrm{F}}^2}\|G_s(u)\|_{H^1(\Omega)} \leq C_{\mathrm{E}}\|l\|_{L^{q^*}(A(s))}. \tag{2.19}$$

The final step before applying Lemma 2.1 is to estimate the left-hand side of (2.19) from below in terms of $|A(h)|$ for $h > s \geq \|w\|_{L^\infty(\Omega_2)}$ and the right-hand side of (2.19) from above in terms of $|A(s)|$. Again, using the

Sobolev embedding $H^1(\Omega) \hookrightarrow L^q(\Omega)$ and Hölder's inequality yields

$$\begin{aligned}\|G_s(u)\|_{H^1(\Omega)} &\geq \frac{1}{C_{\mathrm{E}}}\Big(\int_\Omega |G_s(u)|^q\,dx\Big)^{\frac{1}{q}} = \frac{1}{C_{\mathrm{E}}}\Big(\int_{A(s)} |\,|u| - s|^q\,dx\Big)^{\frac{1}{q}}\\ &\geq \frac{1}{C_{\mathrm{E}}}\Big(\int_{A(h)} (h-s)^q\,dx\Big)^{\frac{1}{q}} = \frac{1}{C_{\mathrm{E}}}(h-s)|A(h)|^{\frac{1}{q}} \qquad (2.20)\end{aligned}$$

and

$$\|l\|_{L^{q^*}(A(s))} \leq \|l\|_{L^2(\Omega)}|A(s)|^{\frac{2-q^*}{2q^*}}. \qquad (2.21)$$

Combining (2.20), (2.21), and (2.19), we obtain the following inequality for the nonnegative and non-increasing function $\varphi(t) := |A(t)|$:

$$|A(h)| \leq \Big(\frac{C_{\mathrm{E}}^2(1+C_{\mathrm{F}}^2)}{\epsilon_{\min}}\|l\|_{L^2(\Omega)}\Big)^q \frac{|A(s)|^{\frac{q-2}{2}}}{(h-s)^q} \quad \text{for all } h > s \geq \|w\|_{L^\infty(\Omega_2)}.$$

Since $\frac{q-2}{2} = 2 > 1$, applying Lemma 2.1, we conclude that there is some $j > 0$ such that

$$\begin{aligned}0 < j^q &= \Big(\frac{C_{\mathrm{E}}^2(1+C_{\mathrm{F}}^2)}{\epsilon_{\min}}\|l\|_{L^2(\Omega)}\Big)^q |A(\|w\|_{L^\infty(\Omega_2)})|^{\frac{q-4}{2}} 2^{\frac{q(q-2)}{q-4}}\\ &\leq \Big(\frac{C_{\mathrm{E}}^2(1+C_{\mathrm{F}}^2)}{\epsilon_{\min}}\|l\|_{L^2(\Omega)}\Big)^q |\Omega|^{\frac{q-4}{2}} 2^{\frac{q(q-2)}{q-4}} =: \bar{j}^q\end{aligned}$$

and $|A(\|w\|_{L^\infty(\Omega_2)} + \bar{j})| = 0$. Hence $\|u\|_{L^\infty(\Omega)} \leq \|w\|_{L^\infty(\Omega_2)} + \bar{j}$. □

The results of this section are summarized in the following theorem.

Theorem 2.3. *Problem* (2.1) *has a unique solution* $u \in H_0^1(\Omega) \cap L^\infty(\Omega)$*, which is the unique minimizer of variational problem* (2.3).

Remark 2.3. Since $k = 0$ in Ω_1, $w \in L^\infty(\Omega_2)$ and $u \in L^\infty(\Omega)$, we conclude that (2.1) holds for all $v \in H_0^1(\Omega)$ resulting in a standard weak formulation. If k^2 is uniformly positive in the whole domain Ω and $w \in L^\infty(\Omega)$, then $\|u\|_{L^\infty(\Omega)} \leq \|w\|_{L^\infty(\Omega)} + \bar{j}$. On the other hand, if $k = 0$ in Ω_2, k^2 is uniformly positive in Ω_1, and $w \in L^\infty(\Omega_1)$, we have $\|u\|_{L^\infty(\Omega)} \leq \|w\|_{L^\infty(\Omega_1)} + \bar{j}$.

3 A Posteriori Error Estimates

3.1 Abstract Framework

First we briefly recall some results from the duality theory [8, 17]. Consider a class of variational problems having the following common form:

$$\text{find } u \in V \text{ such that} \quad J(u) = \inf_{v\in V} J(v), \quad \text{where} \quad J(v) = G(\Lambda v) + F(v). \qquad (\mathrm{P})$$

Here V, Y are reflexive Banach spaces with the norms $\|\cdot\|_V$ and $\|\cdot\|_Y$, respectively, $F\colon V \to \overline{\mathbb{R}}$, $G\colon Y \to \overline{\mathbb{R}}$ are convex and proper functionals, and $\Lambda\colon V \to Y$ is a bounded linear operator. By 0_V we denote the zero element in V. It is assumed that J is coercive and lower semicontinuous. In this case, problem (P) has a solution u, which is unique if J is strictly convex.

The spaces topologically dual to V and Y are denoted by V^* and Y^*, respectively. They are endowed with the norms $\|\cdot\|_{V^*}$ and $\|\cdot\|_{Y^*}$. Henceforth, $\langle v^*, v\rangle$ denotes the duality product of $v^* \in V^*$ and $v \in V$. Analogously, (y^*, y) is the duality product of $y^* \in Y^*$ and $y \in Y$, and $\Lambda^*\colon Y^* \to V^*$ is the operator adjoint to Λ. It is defined by the relation

$$\langle \Lambda^* y^*, v\rangle = (y^*, \Lambda v) \quad \text{for all } v \in V,\ y^* \in Y^*.$$

The functional $J^*: V^* \to \overline{\mathbb{R}}$ defined by the relation

$$J^*(v^*) := \sup_{v \in V} \{\langle v^*, v\rangle - J(v)\}$$

is called *dual* (or Fenchel conjugate) to J (see, e.g., [8]). In accordance with the general duality theory of the calculus of variations, the primal problem (P) has a dual counterpart:

$$\text{find } p^* \in Y^* \text{ such that } \quad I^*(p^*) = \sup_{y^* \in Y^*} I^*(y^*), \quad \text{where } \quad I^*(y^*) := -G^*(y^*) - F^*(-\Lambda^* y^*), \tag{P*}$$

where G^* and F^* are the functionals conjugate to G and F, respectively. Problems (P) and (P*) are generated by the Lagrangian $L: V \times Y^* \to \overline{\mathbb{R}}$ defined by the relation $L(v, y^*) = (y^*, \Lambda v) - G^*(y^*) + F(v)$. If we additionally assume that G^* is coercive and that $F(0_V)$ is finite, then it is well known that problems (P) and (P*) have unique solutions $u \in V$ and $p^* \in Y^*$ and that strong duality relations hold (see [17], or the book [8, Proposition 2.3, Remark 2.3, and Proposition 1.2 in Chapter VI]):

$$J(u) = \inf_{v \in V} J(v) = \inf_{v \in V} \sup_{y^* \in Y^*} L(v, y^*) = \sup_{y^* \in Y^*} \inf_{v \in V} L(v, y^*) = \sup_{y^* \in Y^*} I^*(y^*) = I^*(p^*).$$

Furthermore, the pair (u, p^*) is a saddle point of the Lagrangian L, i.e.,

$$L(u, y^*) \le L(u, p^*) \le L(v, p^*) \quad \text{for all } v \in V,\ y^* \in Y^*,$$

and u and p^* satisfy the relations

$$\Lambda u \in \partial G^*(p^*), \quad p^* \in \partial G(\Lambda u).$$

We have

$$J(v) - I^*(y^*) = G(\Lambda v) + F(v) + G^*(y^*) + F^*(-\Lambda^* y^*) = D_G(\Lambda v, y^*) + D_F(v, -\Lambda^* y^*) =: M^2_\oplus(v, y^*), \tag{3.1}$$

where

$$D_G(\Lambda v, y^*) := G(\Lambda v) + G^*(y^*) - (y^*, \Lambda v),$$
$$D_F(v, -\Lambda^* y^*) := F(v) + F^*(-\Lambda^* y^*) + \langle \Lambda^* y^*, v\rangle$$

are the *compound* functionals for G and F, respectively [17]. A compound functional is nonnegative by the definition. Equality (3.1) shows that D_G and D_F can vanish simultaneously if and only if $v = u$ and $y^* = p^*$. Moreover, setting $v := u$ and $y^* := p^*$ in (3.1), we obtain analogous identities for the primal and dual parts of the error:

$$J(u) - I^*(y^*) = M^2_\oplus(u, y^*) = D_G(\Lambda u, y^*) + D_F(u, -\Lambda^* y^*), \tag{3.2a}$$
$$J(v) - I^*(p^*) = M^2_\oplus(v, p^*) = D_G(\Lambda v, p^*) + D_F(v, -\Lambda^* p^*). \tag{3.2b}$$

Using the fact that $J(u) = I^*(p^*)$ and that the above equalities (3.2) hold, we obtain another important relation (see [17])

$$M^2_\oplus(v, y^*) = J(v) - I^*(y^*) = J(v) - I^*(p^*) + J(u) - I^*(y^*) = M^2_\oplus(v, p^*) + M^2_\oplus(u, y^*). \tag{3.3}$$

Notice that $M^2_\oplus(v, y^*)$ depends on the approximations v and y^* only and, therefore, is fully computable. The right-hand side of (3.3) can be viewed as a certain measure of the distance between (u, p^*) and (v, y^*), which vanishes if and only if $v = u$ and $y^* = p^*$. Hence the relation

$$D_G(\Lambda v, p^*) + D_F(v, -\Lambda^* p^*) + D_G(\Lambda u, y^*) + D_F(u, -\Lambda^* y^*) = M^2_\oplus(v, y^*) \tag{3.4}$$

establishes the equality of the computable term $M^2_\oplus(v, y^*)$ and an error measure natural for this class of variational problems.

It is worth noting that identity (3.4) can be represented in terms of norms if G and F are quadratic functionals. For example, if $V = H^1_0(\Omega)$, $V^* = H^{-1}(\Omega)$, $Y = [L^2(\Omega)]^d = Y^*$, $G(\Lambda v) = G(\nabla v) = \int_\Omega \frac{1}{2} A \nabla v \cdot \nabla v \, dx$ and $F(v) = \int_\Omega (\frac{1}{2} v^2 - lv)\, dx$ (where A is a symmetric positive definite matrix with bounded entries), then

$$D_G(\Lambda v, p^*) = \frac{1}{2} \int_\Omega A \nabla (v - u) \cdot \nabla (v - u)\, dx, \qquad D_G(\Lambda u, y^*) = \frac{1}{2} \int_\Omega A^{-1} (y^* - p^*) \cdot (y^* - p^*)\, dx,$$
$$D_F(v, -\Lambda^* p^*) = \frac{1}{2} \|v - u\|^2_{L^2(\Omega)}, \qquad D_F(u, -\Lambda^* y^*) = \frac{1}{2} \|\operatorname{div}(y^* - p^*)\|^2_{L^2(\Omega)}. \tag{3.5}$$

In this case, the minimizer of (P) solves the linear elliptic problem $-\operatorname{div}(A\nabla u) + u = l$ in Ω, and (3.4) is reduced to the error identity

$$\int_\Omega A\nabla(v-u)\cdot\nabla(v-u)\,dx + \int_\Omega A^{-1}(y^*-p^*)\cdot(y^*-p^*)\,dx + \|v-u\|^2_{L^2(\Omega)} + \|\operatorname{div}(y^*-p^*)\|^2_{L^2(\Omega)}$$
$$= \|\!|A\nabla v - y^*|\!\|^2_* + \|v - \operatorname{div} y^* - l\|^2_{L^2(\Omega)} = 2M^2_\oplus(v, y^*). \tag{3.6}$$

The sum of the first and the third term in (3.6) represents the primal, the sum of the second and fourth term the dual error.

We set $V := H^1_0(\Omega)$, $Y := [L^2(\Omega)]^d$, where $d = 2, 3$, and Λ the gradient operator $\nabla\colon H^1_0(\Omega) \to [L^2(\Omega)]^d$. We further denote $g\colon \Omega\times\mathbb{R}^3 \to \mathbb{R}$, $g(x,\xi) := \frac{\epsilon(x)}{2}|\xi|^2$, and $B\colon \Omega\times\mathbb{R}\to\mathbb{R}$, $B(x,\xi) := k^2(x)\cosh(\xi)$. With this notation, we have

$$G(\Lambda v) := \int_\Omega g(x,\nabla v(x))\,dx = \int_\Omega \frac{\epsilon}{2}|\nabla v|^2\,dx,$$
$$F(v) := \int_\Omega B(x, v(x)+w(x))\,dx - \int_\Omega lv\,dx = \int_\Omega k^2\cosh(v+w)\,dx - \int_\Omega lv\,dx,$$

and the functional J, defined by (2.2), can be written in the form $J(v) = G(\Lambda v) + F(v)$. For any $v\in V$ the functional $G(\Lambda v)$ is finite, while $F\colon V\to\mathbb{R}\cup\{+\infty\}$ may take the value $+\infty$ for some $v\in V$ if $d\ge 3$ (e.g., $v = \log\frac{1}{|x|^\alpha}$, $\alpha \ge d$ on the unit ball in $\mathbb{R}^d$). However, if $d\le 2$, then $\exp(v)\in L^1(\Omega)$ for all $v\in H^1_0(\Omega)$ and $F\colon V\to\mathbb{R}$ (see [15]). Also, $F(0_V)$ is obviously finite since $w\in L^\infty(\Omega_2)$. We set $V^* = H^{-1}(\Omega)$ and $Y^* = Y = [L^2(\Omega)]^d$. In this case, Λ^* coincides with $-\operatorname{div}$ considered as an operator from $[L^2(\Omega)]^d$ to $H^{-1}(\Omega)$. We will present the particular form of error equality (3.4) where the error is measured in a special "nonlinear norm". This measure contains the usual combined energy norm terms, i.e., the sum of the energy norms of the errors for the primal and dual problem, but also two additional nonnegative terms due to the nonlinearity $B(x,\xi)$ (or equivalently $b(x,\xi)$) which in some cases may dominate the usual energy norm terms. We start by deriving explicit expressions for G^*, F^*, and then we will use these expressions to get an explicit form of the abstract error equality (3.4).

3.2 Fenchel Conjugates of the Functionals *G* and *F*

It is easy to find that $G^*(y^*) = \int_\Omega \frac{1}{2\epsilon(x)}|y^*(x)|^2\,dx$. For $y^*\in H(\operatorname{div};\Omega)$ and an arbitrary function $z\colon\Omega_2\to\mathbb{R}$, we introduce the functional

$$I_{y^*}(z) := \int_{\Omega_2}[(\operatorname{div} y^* + l)z - B(x, z+w)]\,dx.$$

Recalling that the nonlinearity B is supported on Ω_2, we have

$$\begin{aligned}
F^*(-\Lambda^* y^*) &= \sup_{z\in H^1_0(\Omega)}[\langle -\Lambda^* y^*, z\rangle - F(z)] = \sup_{z\in H^1_0(\Omega)}[(-y^*, \Lambda z) - F(z)]\\
&= \sup_{z\in H^1_0(\Omega)}\int_\Omega [-y^*\cdot\nabla z - B(x, z+w) + lz]\,dx \quad (\text{if } y^*\in H(\operatorname{div};\Omega))\\
&= \sup_{z\in H^1_0(\Omega)}\int_\Omega [\operatorname{div} y^* z - B(x, z+w) + lz]\,dx \quad (\text{finite if } \operatorname{div} y^* + l = 0 \text{ in } \Omega_1)\\
&= \sup_{z\in H^1_0(\Omega)} I_{y^*}(z) \le \int_{\Omega_2}\sup_{\xi\in\mathbb{R}}[(\operatorname{div} y^*(x) + l(x))\xi - B(x, \xi + w(x))]\,dx\\
&= \int_{\Omega_2}[(\operatorname{div} y^*(x) + l(x))\xi_0(x) - B(x, \xi_0(x) + w(x))]\,dx = I_{y^*}(\xi_0).
\end{aligned} \tag{3.7}$$

Here $\xi_0\colon\Omega_2\to\mathbb{R}$ is computed by the condition

$$\frac{d}{d\xi}[(\operatorname{div} y^*(x) + l(x))\xi - B(x, \xi + w(x))] = 0 \quad \text{for a.e. } x\in\Omega_2, \tag{3.8}$$

which is equivalent to

$$\operatorname{div} y^*(x) + l(x) - k^2(x)\sinh(\xi + w(x)) = 0 \quad \text{for a.e. } x \in \Omega_2.$$

We notice that (3.8) is a necessary condition for a maximum which is also sufficient since $B(x, \cdot)$ is convex. The solution of the last equation exists, is unique, and is given by

$$\xi_0(x) = \operatorname{arsinh}(\rho_k(y^*)) - w(x) = \ln\big(\rho_k(y^*) + \sqrt{\rho_k^2(y^*) + 1}\big) - w(x) = \ln\big(\Theta(\rho_k(y^*))\big) - w(x),$$

where $\rho_k(y^*) := \frac{\operatorname{div} y^*(x) + l(x)}{k^2(x)}$ and $\Theta(s) := s + \sqrt{s^2 + 1}$ for $s \in \mathbb{R}$. Note that the exact solution $p^* = \epsilon\nabla u$ of dual problem (P*) also satisfies the relation $\operatorname{div}(\epsilon\nabla u) + l = 0$ because, for any $x \in \Omega_1$, it holds $k(x) = 0$. Moreover, since $u \in L^\infty(\Omega)$, $w \in L^\infty(\Omega_2)$, and $l \in L^2(\Omega)$, we see that $\operatorname{div} p^* = k^2 \sinh(u + w) + l \in L^2(\Omega)$ and thus $p^* \in H(\operatorname{div}; \Omega)$. In Proposition 3.1, we will later prove that we have not overestimated the supremum over $z \in H_0^1(\Omega)$ in (3.7) and that we actually have equalities everywhere. Denoting $S := \operatorname{arsinh}(\rho_k(y^*))$ and using the expression for $\xi_0(x)$ and the formula

$$\cosh(\operatorname{arsinh}(x)) = \sqrt{x^2 + 1} \quad \text{for all } x \in \mathbb{R},$$

for any $y^* \in H(\operatorname{div}; \Omega) \subset [L^2(\Omega)]^d = Y^*$ with $\operatorname{div} y^* + l = 0$ in Ω_1, we obtain an explicit formula for $F^*(-\Lambda^* y^*)$:

$$\begin{aligned} F^*(-\Lambda^* y^*) &= \int_{\Omega_2} \big[k^2 \rho_k(y^*)\big(\ln\big(\Theta(\rho_k(y^*))\big) - w\big) - k^2\sqrt{\rho_k^2(y^*) + 1}\big]\, dx \\ &= \int_{\Omega_2} [k^2 \sinh(S)(S - w) - k^2 \cosh(S)]\, dx. \end{aligned} \tag{3.9}$$

Remark 3.1. Since $|\ln(t + \sqrt{t^2 + 1})| \le |t|$ for all $t \in \mathbb{R}$, the function $\ln(\Theta(f(x))) - w(x)$ belongs to $L^2(\Omega_2)$ for any $f \in L^2(\Omega_2)$, and we conclude that $\xi_0(x) \in L^2(\Omega_2)$ if $y^* \in H(\operatorname{div}; \Omega)$. Therefore, the integral in (3.9) is well defined.

Now our goal is to prove that the inequality $\sup_{z \in H_0^1(\Omega)} I_{y^*}(z) \le I_{y^*}(\xi_0)$ holds as the equality. In other words, we want to prove that the error estimate remains sharp and that the computed majorant $M_\oplus^2(v, y^*)$ will be indeed zero if approximations (v, y^*) coincide with the exact solution (u, p^*).

Proposition 3.1. *For any $y^* \in H(\operatorname{div}; \Omega)$ with $\operatorname{div} y^* + l = 0$ in Ω_1, it holds*

$$\sup_{z \in H_0^1(\Omega)} I_{y^*}(z) = I_{y^*}(\xi_0) < \infty.$$

Proof. The idea is to approximate $f = \frac{\operatorname{div} y^* + l}{k^2} \in L^2(\Omega_2)$ and $w_{\restriction_{\Omega_2}} \in L^\infty(\Omega_2)$ by $C_0^\infty(\Omega_2)$ functions (in the a.e. sense) and use the Lebesgue dominated convergence theorem. Let $f_n \in C_0^\infty(\Omega_2)$ and $w_n \in C_0^\infty(\Omega_2)$ be such that $f_n(x) \to f(x)$ a.e. in Ω_2, $|f_n(x)| \le h(x) \in L^2(\Omega_2)$ (see [2, Theorem 4.9]), $w_n(x) \to w(x)$ a.e. in Ω_2, $|w_n(x)| \le m + 2$, where $m := \|w\|_{L^\infty(\Omega_2)}$. Then

$$z_n(x) := \ln(\Theta(f_n(x))) - w_n(x) \to \xi_0(x) \quad \text{a.e. in } \Omega_2$$

and $z_n \in C_0^\infty(\Omega_2) \subset H_0^1(\Omega_2) \subset H_0^1(\Omega)$ (by extending the functions by zero in Ω_1). Since $B(x, \cdot)$ is continuous, we have the pointwise a.e. in Ω_2 convergence

$$(\operatorname{div} y^*(x) + l(x))z_n(x) - B(x, z_n + w(x)) \to (\operatorname{div} y^*(x) + l(x))\xi_0(x) - B(x, \xi_0(x) + w(x))$$

Now we search for a function in $L^1(\Omega_2)$ that majorates the function $|(\operatorname{div} y^*(x) + l(x))z_n(x) - B(x, z_n + w(x))|$:

$$\begin{aligned} &|(\operatorname{div} y^*(x) + l(x))z_n(x) - k^2(x)\cosh(z_n(x) + w(x))| \\ &\qquad \le |\operatorname{div} y^*(x) + l(x)|\,|z_n(x)| + k^2(x) e^{\|w\|_{L^\infty(\Omega_2)}} e^{|z_n(x)|}. \end{aligned} \tag{3.10}$$

Our next goal is to bound $|z_n(x)|$ in (3.10). For the first summand, we have

$$|z_n(x)| = |\ln(\Theta(f_n(x))) - w_n(x)| \le |f_n(x)| + m + 2 \le h(x) + m + 2 \in L^2(\Omega_2),$$

where Remark 3.1 has been used. However, this bound cannot be used in the second term because e^h might not belong even to $L^1(\Omega_2)$. In order to find an L^1-majorant for the second summand in (3.10), we distinguish the following two cases: In the first case, $f_n(x) > 0$. Then $|\ln(\Theta(f_n(x)))| \le |\ln(\Theta(h(x)))|$.

In the second case ($f_n(x) \le 0$), we have $\Theta(f_n(x)) \le 1$. Therefore, $0 \ge f_n(x) \ge -h(x)$. Since $\Theta(s)$ is a monotonically increasing function, $\Theta(0) = 1 \ge \Theta(f_n(x)) \ge \Theta(-h(x)) > 0$. From here, we obtain

$$\ln(1) = 0 \ge \ln(\Theta(f_n(x))) \ge \ln(\Theta(-h(x))),$$

and using the relation $\Theta(-h) = \frac{1}{\Theta(h)}$, we conclude that

$$|\ln(\Theta(f_n(x)))| \le |\ln(\Theta(-h(x)))| = |\ln(\Theta(h(x)))|.$$

Finally, for almost all $x \in \Omega_2$, we have

$$\begin{aligned} |z_n(x)| &= |\ln(\Theta(f_n(x))) - w_n(x)| \\ &\le |\ln(\Theta(h(x)))| + m + 2 = \ln(\Theta(h(x))) + m + 2 \quad \text{because} \quad h(x) \ge 0 \quad \text{for a.e. } x \in \Omega_2. \end{aligned}$$

Therefore,

$$\begin{aligned} &|(\operatorname{div} y^*(x) + l(x))z_n(x) - k^2(x)\cosh(z_n(x) + w(x))| \\ &\quad \le |\operatorname{div} y^*(x) + l(x)|(h(x) + \|w\|_{L^\infty(\Omega_2)} + 2) + k^2(x)e^{2\|w\|_{L^\infty(\Omega_2)}+2}\Theta(h(x)) := H(x) \in L^2(\Omega_2), \end{aligned}$$

where, in the last line, we used the fact that $\Theta(h(x)) \in L^2(\Omega_2)$. All the conditions of the Lebesgue dominated convergence theorem are satisfied, and we see that $I_{y^*}(z_n) \to I_{y^*}(\xi_0)$ and, consequently,

$$\sup_{z \in H_0^1(\Omega)} I_{y^*}(z) = I_{y^*}(\xi_0). \qquad \square$$

3.3 Error Measures

In this section, we apply the abstract framework from Section 3.1 and derive an explicit form of relation (3.4) adapted to our problem. For any $y^* \in H(\operatorname{div}; \Omega)$ with $\operatorname{div} y^* + l = 0$ in Ω_1, the quantity $M_\oplus^2(v, y^*)$ is fully computable and is given by the relation

$$\begin{aligned} M_\oplus^2(v, y^*) &= D_G(\Lambda v, y^*) + D_F(v, -\Lambda^* y^*) \\ &= G(\Lambda v) + G^*(y^*) - (y^*, \Lambda v) + F(v) + F^*(-\Lambda^* y^*) + \langle \Lambda^* y^*, v \rangle \\ &= \int_\Omega \eta^2(x)\, dx = \frac{1}{2}|||\epsilon \nabla v - y^*|||_*^2 + D_F(v, -\Lambda^* y^*), \end{aligned} \tag{3.11}$$

where

$$\eta^2(x) = \begin{cases} \frac{1}{2\epsilon}|\epsilon \nabla v - y^*|^2, & x \in \Omega_1 \\ \frac{1}{2\epsilon}|\epsilon \nabla v - y^*|^2 + k^2\cosh(v + w) - lv & \\ \qquad + k^2\rho_k(y^*)(\ln(\Theta(\rho_k(y^*))) - w) - k^2\sqrt{\rho_k^2(y^*) + 1} - \operatorname{div} y^* v, & x \in \Omega_2 \end{cases} \tag{3.12}$$

and we have used the expressions for G^* and F^*. It is clear that $\eta^2(x) \ge 0$ since it is the sum of the compound functionals generated by $\tilde{g}_x(s) := g(x, s)$ and $\tilde{B}_x(s) - l(x)s = B(x, s + w(x)) - l(x)s$ evaluated at $(\nabla v(x), y^*(x))$ and $(v(x), \operatorname{div} y^*(x))$, respectively. It therefore qualifies as an error indicator, provided that y^* is chosen appropriately, which we demonstrate with numerical experiments in the next section. Using the expression for G^*, we obtain

$$\begin{aligned} D_G(\Lambda v, p^*) &= \frac{1}{2}\int_\Omega \epsilon|\nabla(v - u)|^2\, dx =: \frac{1}{2}|||\nabla(v - u)|||^2, \\ D_G(\Lambda u, y^*) &= \frac{1}{2}\int_\Omega \frac{1}{\epsilon}|y^* - p^*|^2\, dx =: \frac{1}{2}|||y^* - p^*|||_*^2. \end{aligned} \tag{3.13}$$

Now we find explicit expressions for the nonlinear measures $D_F(v, -\Lambda^* p^*)$ and $D_F(u, -\Lambda^* y^*)$ similar to the ones for the case of quadratic F in (3.5) for the linear elliptic equation $-\operatorname{div}(A\nabla u) + u = l$. We will need the following assertion, which is easy to prove.

Proposition 3.2. *For all $s, t \in \mathbb{R}$, it holds*

$$\frac{(t-s)^2}{2} \le A(s,t) \le \frac{(\sinh(t)-\sinh(s))^2}{2}, \tag{3.14}$$

where $A(s,t) = \cosh(t) - \cosh(s) + s\sinh(s) - t\sinh(s)$.

Since, for the exact solution u, we have $\rho_k(p^*) = \sinh(u+w)$ and $u = \operatorname{arsinh}(\rho_k(p^*)) - w$ for a.e. $x \in \Omega_2$, we find that

$$\begin{aligned} D_F(v, -\Lambda^* p^*) &= \int_{\Omega_2} \left(k^2\cosh(v+w) - lv + k^2\sinh(u+w)u - k^2\cosh(u+w) - \operatorname{div} p^* v\right) dx \\ &= \int_{\Omega_2} k^2\big(\cosh(v+w) - \cosh(u+w) + u\sinh(u+w) - v\sinh(u+w)\big)\, dx. \end{aligned}$$

Similarly, $D_F(u, -\Lambda^* y^*) = \int_{\Omega_2} k^2\big(\cosh(T) - \cosh(S) + S\sinh(S) - T\sinh(S)\big)\, dx$, where $T := \operatorname{arsinh}(\rho_k(p^*))$. The nonlinear quantities $D_F(v, -\Lambda^* p^*)$ and $D_F(u, -\Lambda^* y^*)$ measure the error in v and in $\operatorname{div} y^*$, respectively. Using inequality (3.14), we can represent these two measures in a form which resembles the corresponding estimates in the case (3.5) of a quadratic functional F, namely,

$$\int_{\Omega_2} \frac{k^2}{2}(v-u)^2\, dx \le D_F(v, -\Lambda^* p^*) \le \int_{\Omega_2} \frac{k^2}{2}(\sinh(v+w) - \sinh(u+w))^2\, dx, \tag{3.15}$$

$$\int_{\Omega_2} \frac{k^2}{2}(T-S)^2\, dx \le D_F(u, -\Lambda^* y^*) \le \int_{\Omega_2} \frac{1}{2k^2}(\operatorname{div} p^* - \operatorname{div} y^*)^2\, dx. \tag{3.16}$$

Note that, for $k \ge k_{\min} > 0$ in Ω, the following equivalences hold:

$$\int_{\Omega} \frac{k^2}{2}(v-u)^2\, dx \eqsim \|v-u\|^2_{L^2(\Omega)} \quad \text{and} \quad \int_{\Omega} \frac{1}{2k^2}(\operatorname{div} p^* - \operatorname{div} y^*)^2\, dx \eqsim \|\operatorname{div} y^* - \operatorname{div} p^*\|^2_{L^2(\Omega)}.$$

Moreover, replacing the nonlinear term $k^2\sinh(u+w)$ with u, inequalities (3.15) and (3.16) reduce to the equalities for $D_F(v, -\Lambda^* p^*)$ and $D_F(u, -\Lambda^* y^*)$ in (3.5) because, in this case, the inverse function of $f(x) = x$ is again $f(x)$. The functions on the left-hand side, in the middle, and on the right-hand side in inequality (3.14) are depicted on Figure 1.

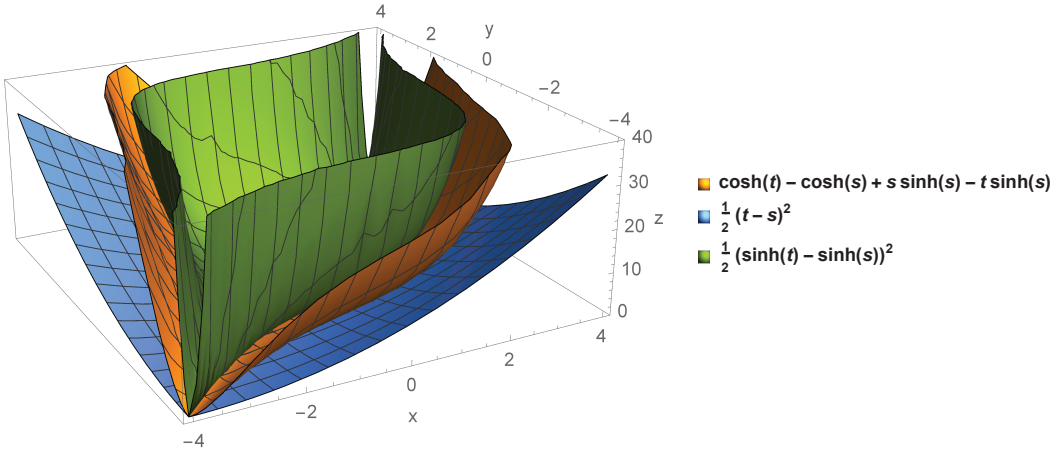

Figure 1: Functions in inequality (3.14).

Further, if v is in a δ_1-neighborhood of u in $L^\infty(\Omega)$ norm, then we can find a constant $C_1(\delta_1, \|u\|_{L^\infty(\Omega)}) > 1$ such that

$$\int_{\Omega_2} \frac{k^2}{2}(\sinh(v+w) - \sinh(u+w))^2\,dx \le C_1(\delta_1, \|u\|_{L^\infty(\Omega)}) \int_{\Omega_2} \frac{k^2}{2}(v-u)^2\,dx. \tag{3.17}$$

Analogously, if $l \in L^\infty(\Omega_2)$ and $\|\operatorname{div}(y^* - p^*)\|_{L^\infty(\Omega_2)} \le \delta_2$ (recall that when $l \in L^\infty(\Omega_2)$, $\operatorname{div} p^*$ is in $L^\infty(\Omega_2)$), then we can find a constant $C_2(\delta_2, \|\operatorname{div} p^*\|_{L^\infty(\Omega_2)}) < 1$ such that

$$C_2(\delta_2, \|\operatorname{div} p^*\|_{L^\infty(\Omega_2)}) \int_{\Omega_2} \frac{1}{2k^2}(\operatorname{div} p^* - \operatorname{div} y^*)^2\,dx \le \int_{\Omega_2} \frac{k^2}{2}(T-S)^2\,dx. \tag{3.18}$$

The constants C_1 and C_2 are just Lipschitz constants for the locally Lipschitz function sinh. Notice that if $k^2 \ge k_{\min} > 0$ in Ω, then everywhere in (3.15), (3.16), (3.17), and (3.18), the integrals are taken over the entire domain Ω. Now the abstract error identity (3.4) takes the form

$$\begin{aligned}
&\frac{1}{2}|||\nabla(u-v)|||^2 + \frac{1}{2}|||p^* - y^*|||_*^2 + \int_{\Omega_2} \frac{k^2}{2}(v-u)^2\,dx + C_2(\delta_2, \|\operatorname{div} p^*\|_{L^\infty(\Omega)}) \int_{\Omega_2} \frac{1}{2k^2}(\operatorname{div} p^* - \operatorname{div} y^*)^2\,dx \\
&\quad \le \underline{\frac{1}{2}|||\nabla(u-v)|||^2 + \frac{1}{2}|||p^* - y^*|||_*^2 + D_F(v, -\Lambda^* p^*) + D_F(u, -\Lambda^* y^*) = M_\oplus^2(v, y^*)} \\
&\quad \le \frac{1}{2}|||\nabla(u-v)|||^2 + \frac{1}{2}|||p^* - y^*|||_*^2 \\
&\qquad + C_1(\delta_1, \|u\|_{L^\infty(\Omega)}) \int_{\Omega_2} \frac{k^2}{2}(v-u)^2\,dx + \int_{\Omega_2} \frac{1}{2k^2}(\operatorname{div} y^* - \operatorname{div} p^*)^2\,dx,
\end{aligned} \tag{3.19}$$

where we have used $p^* = \epsilon\Lambda u = \epsilon\nabla u$. Relation (3.19) shows that the computable majorant $M_\oplus^2(v, y^*)$ is bounded from below and above by a multiple of one and the same error norm. Since $D_F(v, -\Lambda^* p^*) \ge 0$ and $D_F(u, -\Lambda^* y^*) \ge 0$, we also obtain a guaranteed bound on the error in the combined energy norm,

$$|||\nabla(u-v)|||^2 + |||p^* - y^*|||_*^2 \le 2M_\oplus^2(v, y^*). \tag{3.20}$$

From the pointwise equality

$$\begin{aligned}
\frac{1}{\epsilon}|\epsilon\nabla v - y^*|^2 &= \frac{1}{\epsilon}|\epsilon\nabla(v-u) - (y^* - p^*)|^2 \\
&= \epsilon|\nabla(v-u)|^2 + \frac{1}{\epsilon}|y^* - p^*|^2 - 2(y^* - p^*)\cdot\nabla(v-u),
\end{aligned} \tag{3.21}$$

after applying Young's inequality and integrating over Ω, we obtain a lower bound for the error in combined energy norm,

$$\frac{1}{2}|||\epsilon\nabla v - y^*|||_*^2 \le |||\nabla(v-u)|||^2 + |||y^* - p^*|||_*^2 \tag{3.22}$$

Remark 3.2. Integrating (3.21) over Ω, we obtain the algebraic identity

$$|||\epsilon\nabla v - y^*|||_*^2 = |||\nabla(v-u)|||^2 + |||y^* - p^*|||_*^2 - 2\int_\Omega (y^* - p^*)\cdot\nabla(v-u)\,dx, \tag{3.23}$$

from which the Prager–Synge identity is derived. Comparing the last relation with (3.19), using the fact that $M_\oplus(v, y^*)^2 = \frac{1}{2}|||\epsilon\nabla v - y^*|||_*^2 + D_F(v, -\Lambda^* y^*)$, we arrive at the relation

$$D_F(v, -\Lambda^* y^*) = D_F(v, -\Lambda^* p^*) + D_F(u, -\Lambda^* y^*) + \int_\Omega (y^* - p^*)\cdot\nabla(v-u)\,dx. \tag{3.24}$$

From here, it is seen that if the integral on the right-hand side is small compared to the other terms, then the error in v and $\operatorname{div} y^*$ measured with $D_F(v, -\Lambda^* p^*) + D_F(u, -\Lambda^* y^*)$ is controlled mainly by the computable term $D_F(v, -\Lambda^* y^*)$ in the majorant $M_\oplus^2(v, y^*)$. Moreover, (3.23) enables us to give a practical estimation of

the error in combined energy norm, which is very close to the real error in all of the experiments that we have conducted.

We conclude the section by presenting a near best approximation result. Contrary to the result in [5, Theorem 6.2], we do not make any restrictive assumptions on the meshes to ensure that the finite element approximations u_h are uniformly bounded in L^∞ norm. In our analysis, $V_h \subset L^\infty$ is a finite-dimensional subspace of H^1_0, and u_h is the minimizer of J over V_h, which is the unique solution of the Galerkin problem:

$$\text{find } u_h \in V_h \text{ such that} \quad a(u_h, v) + \int_\Omega b(x, u_h + w)v\,dx = (l, v) \quad \text{for all } v \in V_h. \tag{3.25}$$

Then, using (3.2b) and expression (3.13) for $D_G(\Lambda v, p^*)$, for any $v \in V_h$, we can write

$$\begin{aligned}|||\nabla(u_h - u)|||^2 + 2D_F(u_h, -\Lambda^* p^*) &= 2(J(u_h) - J(u)) \\ &\le 2(J(v) - J(u)) = |||\nabla(v - u)|||^2 + 2D_F(v, -\Lambda^* p^*).\end{aligned}$$

Since $2D_F(u_h, -\Lambda^* p^*) \ge 0$, using (3.15), we obtain the following generalization of Cea's lemma to the case of our nonlinear problem.

Proposition 3.3. *Let $V_h \subset L^\infty(\Omega)$ be a closed subspace of $H^1_0(\Omega)$ and $u_h \in V_h$ the Galerkin approximation of u defined by* (3.25). *Then*

$$|||\nabla(u_h - u)|||^2 \le \inf_{v \in V_h} \Big\{ |||\nabla(v - u)|||^2 + \int_{\Omega_2} k^2(\sinh(v + w) - \sinh(u + w))^2\,dx \Big\}. \tag{3.26}$$

Since we use the finite element method with P_1 Lagrange elements, let V_h be the corresponding space, where h refers to the maximum element size. By $I_h(\varphi)$, we denote the Lagrange finite element interpolant of $\varphi \in C^0(\Omega)$. Using (3.26), we can show unqualified convergence of the finite element approximations u_h to u as $h \to 0$. Let $\varepsilon > 0$, and $\bar{u} \in C_0^\infty(\Omega)$ is such that $\|\nabla(\bar{u} - u)\|_{L^2(\Omega)} \le \varepsilon$ and $\|\bar{u}\|_{L^\infty(\Omega)} \le \|u\|_{L^\infty(\Omega)} + 2$. Also, let L be the Lipschitz constant in the inequality

$$|\sinh(s) - \sinh(t)| \le L|s - t| \quad \text{for all } s, t \in [-2\|w\|_{L^\infty(\Omega_2)} - \bar{j} - 2,\ 2\|w\|_{L^\infty(\Omega_2)} + \bar{j} + 2],$$

where $\bar{j}$ is the constant from Proposition 2.1. Then, applying the triangle inequality together with Young's inequality, we obtain

$$\begin{aligned}|||\nabla(u_h - u)|||^2 \le{}& 2\big(|||\nabla(I_h(\bar{u}) - \bar{u})|||^2 + |||\nabla(\bar{u} - u)|||^2\big) \\ &+ 2\Big(\int_\Omega k^2(\sinh(I_h(\bar{u}) + w) - \sinh(\bar{u} + w))^2\,dx \\ &\qquad + \int_\Omega k^2(\sinh(\bar{u} + w) - \sinh(u + w))^2\,dx \Big).\end{aligned} \tag{3.27}$$

For the first term in (3.27), assuming mesh regularity, we have

$$|||\nabla(I_h(\bar{u}) - \bar{u})|||^2 + |||\nabla(\bar{u} - u)|||^2 \le \epsilon_{\max}(C|\bar{u}|_2^2 h^2 + \varepsilon^2),$$

where $|\bar{u}|_2$ denotes the H^2 seminorm of $\bar{u}$ and $C > 0$ is a constant depending on the mesh regularity. Using the fact that $\|I_h(\bar{u})\|_{L^\infty(\Omega)} \le \|\bar{u}\|_{L^\infty(\Omega)} \le \|u\|_{L^\infty(\Omega)} + 2$, for the second term in (3.27), we obtain the upper bound

$$\begin{aligned}2k_{\max}^2 L^2\big(\|I_h(\bar{u}) - \bar{u}\|^2_{L^2(\Omega)} + \|\bar{u} - u\|^2_{L^2(\Omega)}\big) &\le 2k_{\max}^2 L^2 C_{\mathrm{F}}^2\big(\|\nabla(I_h(\bar{u}) - \bar{u})\|^2_{L^2(\Omega)} + \|\nabla(\bar{u} - u)\|^2_{L^2(\Omega)}\big) \\ &\le 2k_{\max}^2 L^2 C_{\mathrm{F}}^2(C|\bar{u}|_2^2 h^2 + \varepsilon^2).\end{aligned}$$

This inequality shows that the right-hand side of (3.27) can be made as small as desired provided that we choose ε and h small enough, and therefore $|||\nabla(u_h - u)||| \to 0$ when $h \to 0$. Moreover, (3.26) can be also used to obtain qualified convergence of u_h in the energy norm under additional assumptions on the interface Γ, the meshes, and the regularity of u.

4 Numerical Results

In this section, we present numerical examples illustrating error identity (3.19) and performance of functional a posteriori error estimates. All numerical experiments are carried out in FreeFem++ developed and maintained by Frederich Hecht [13], and all pictures are generated in VisIt [6]. We solve adaptively the homogeneous nonlinear problem (1.1) with $w := w_{h_{\mathrm{ref}}} = g - z_{h_{\mathrm{ref}}}$, where $z_{h_{\mathrm{ref}}}$ is a good Galerkin finite element approximation of the solution z of

$$\begin{aligned} -\nabla\cdot(\epsilon\nabla z) &= -k^2\sinh(g) + l \quad \text{in } \Omega_1\cup\Omega_2,\\ [z]_\Gamma &= 0,\\ \left[\epsilon\frac{\partial z}{\partial n}\right]_\Gamma &= 0,\\ z &= 0 \qquad\qquad \text{on } \partial\Omega, \end{aligned}$$

for given functions g and l. We compare the accuracy of the adaptively computed solution u_h of (1.1) for $w = w_{h_{\mathrm{ref}}}$ to the reference solution $z_{h_{\mathrm{ref}}}$. The adaptive mesh refinement is based on the error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ on subdomains O_i, where the function η is defined in (3.12) and η^2 is the integrand of the majorant $M^2_\oplus(v, y^*)$. The factor $\sqrt{2}$ accounts for the factor 2 in (3.20). More precisely, we find approximations u_h to the exact solution $u \in H^1_0(\Omega)$ of the problem

$$\int_\Omega \epsilon\nabla u\cdot\nabla v\,dx + \int_\Omega b(x, u + w_{h_{\mathrm{ref}}})v\,dx = \int_\Omega lv\,dx = 0 \quad \text{for all } v \in H^1_0(\Omega).$$

In all examples, we use piecewise constant parameters ϵ and k, and for $y^* \in H(\mathrm{div};\Omega)$, we used a patchwise equilibrated reconstruction of the numerical flux $\epsilon\nabla u_h$ based on [1]. More precisely, we find y^* in the Raviart–Thomas space RT_0 over the same mesh such that its divergence is equal to the L^2 orthogonal projection of $k^2\sinh(u_h + w) + l$ onto the space of piecewise constants.

Recall that

$$M^2_\oplus(v, y^*) = M^2_\oplus(v, p^*) + M^2_\oplus(u, y^*),$$

where $M^2_\oplus(v, y^*) = \frac{1}{2}|||\epsilon\nabla v - y^*|||^2_* + D_F(v, -\Lambda^* y^*)$ and $M^2_\oplus(v, p^*) = J(v) - J(u) = \frac{1}{2}|||\nabla(v-u)|||^2 + D_F(v, -\Lambda^* p^*)$ is the primal error, whereas $M^2_\oplus(u, y^*) = I^*(p^*) - I^*(y^*) = \frac{1}{2}|||y^* - p^*|||^2_* + D_F(u, -\Lambda^* y^*)$ is the dual error. Further, we use v for the approximate solution u_h and u for the reference solution $z_{h_{\mathrm{ref}}}$ and define the efficiency index of the lower bound for the error in combined energy norm (3.22) by

$$I^{\mathrm{CEN,Low}}_{\mathrm{Eff}} := \frac{\frac{\sqrt{2}}{2}|||\epsilon\nabla v - y^*|||_*}{\sqrt{|||\nabla(v-u)|||^2 + |||y^* - p^*|||^2_*}}.$$

Similarly,

$$I^{\mathrm{CEN,Up}}_{\mathrm{Eff}} := \frac{\sqrt{2M^2_\oplus(v, y^*)}}{\sqrt{|||\nabla(v-u)|||^2 + |||y^* - p^*|||^2_*}}$$

defines the efficiency index of the upper bound (3.20) for the error in combined energy norm. Finally,

$$I^{\mathrm{E}}_{\mathrm{Eff}} := \frac{\sqrt{2M^2_\oplus(v, y^*)}}{|||\nabla(v-u)|||} \quad \text{and} \quad P^{\mathrm{CEN}}_{\mathrm{rel}} := \frac{|||\epsilon\nabla v - y^*|||_*}{\sqrt{|||\nabla v|||^2 + |||y^*|||^2_*}}$$

define the efficiency index of the upper bound for the error in energy norm and the practical estimate of the relative error in combined energy norm, respectively.

4.1 Example 1 (2D Problem)

In the first example, the domain Ω is a square with a side 20 with Ω_1 being a regular 15-sided polygon with a radius of its circumscribed circle equal to 2. The coefficients ϵ and k are defined by the relations

$$\epsilon(x) = \begin{cases} \epsilon_1 = 1, & x \in \Omega_1, \\ \epsilon_2 = 100, & x \in \Omega_2, \end{cases} \qquad k(x) = \begin{cases} k_1 = 0.15, & x \in \Omega_1, \\ k_2 = 0.4, & x \in \Omega_2, \end{cases}$$

and

$$g = L\Big(\exp\Big(-b_1\Big(\frac{(x_1 - c_1)^2}{\sigma_1^2} - 1\Big)\Big) - \exp\Big(-b_2\Big(\frac{(x_2 - c_2)^2}{\sigma_2^2} - 1\Big)\Big)\Big),$$

$l = 0$, where $b_1 = 2 = b_2 = 2$, $c_1 = -1$, $c_2 = 6$, $\sigma_1 = \sigma_2 = 1.5$, $L = 0.8$. The reference solution $z_{h_{\mathrm{ref}}}$ is computed on a multiply refined mesh with 50 086 142 triangles. Note that $k^2 = 0.0225$ in Ω_1 and $k^2 = 0.16$ in Ω_2. The mesh adaptation is done with the built-in function "adaptmesh" of FreeFem++. The localized error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$, computed on each vertex patch O_i of the mesh, is compared to its average value over all patches, and the local mesh size is divided by two if this average is smaller than the local value.

Table 1 illustrates the main error identity (3.3) and the convergence of its constituent parts. Further, it is seen that the dual error $2M_\oplus^2(u, y^*)$ dominates the primal error in this example. This is due to the fact that the term $2D_F(u, -\Lambda^* y^*)$, measuring the error in $\operatorname{div} y^*$ (cf. (3.16) and (3.18)), is much larger than $|||\nabla(v - u)|||^2 + D_F(v, -\Lambda^* p^*)$, where $D_F(v, -\Lambda^* p^*)$ behaves like $\|v - u\|^2_{L^2(\Omega_2)}$ (cf. (3.15) and (3.17)). As we mentioned earlier, for y^*, we use a partially equilibrated reconstruction of the numerical flux $\epsilon\nabla v$, which is the reason why the integral term in (3.23) is negligible compared to the combined energy norm of the error. This fact is confirmed by the values of the efficiency index of the lower bound (3.22).

In Table 3, we can see that $I_{\mathrm{Eff}}^{\mathrm{CEN,Low}}$ is approximately equal to $0.7071 \approx \frac{\sqrt{2}}{2}$. The value of the efficiency index with respect to the combined energy norm and the value of the ratio $D_F(v, -\Lambda^* y^*)/M_\oplus^2(v, y^*)$ are also coupled in the sense that if we have only one of these two quantities, we can estimate the other one by using the main error equality (3.19). This estimation is accurate because the integral term in (3.24) is very close to zero, and therefore $D_F(v, -\Lambda^* y^*) \approx D_F(v - \Lambda^* p^*) + D_F(u - \Lambda^* y^*)$. One more consequence of using a partially equilibrated flux is that we obtain a very accurate practical estimate of the absolute and relative error in combined energy norm as illustrated in the last two columns of Table 3.

Figure 2 depicts a mesh that is a part of a sequence of meshes obtained by mesh adaptation using the localized functional error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$. Figure 3 depicts a mesh with approximately the same number of elements but obtained by mesh adaptation using the error indicator $|||\epsilon\nabla v - y^*|||_{*(O_i)}$. The mesh in Figure 2 is refined mainly where the error in $\operatorname{div} y^*$ is the dominant part of the error $M_\oplus^2(v, -\Lambda^* p^*) + M_\oplus^2(u, -\Lambda^* y^*)$. On the other hand, the mesh in Figure 3 is refined most around the extrema of the solution. Figure 5 depicts the minimal set of elements K of a mesh T_h that contains at least 30 % of the total indicated error $\sum_{K \in T_h} |||\epsilon\nabla v - y^*|||_{*(K)}$ (greedy algorithm with a bulk factor of 0.3), where T_h is part of the same sequence as the mesh illustrated in Figure 3.

Figure 6 depicts the elements marked by the greedy algorithm using a bulk factor of 0.5 and employing the true error

$$\sqrt{2M_\oplus^2(v, p^*) + 2M_\oplus^2(u, y^*)}$$

as indicator. Figure 7 depicts elements which are marked additionally or fail to be marked by the same greedy algorithm when employing the functional error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ for the same bulk factor. The ratio of the number of these differently marked elements, that is, elements which are marked by one of the two methods but not by the other one, and the total number of elements is 0.022, and the ratio of the number of differently marked elements to the number of marked elements using the true error is 0.048 (Table 4). Comparing the indicated error and the true error elementwise, one finds that the error indicator generated by the majorant $M_\oplus^2(v, y^*)$ reproduces the local distribution of the error with a very high accuracy. This is also confirmed by Figure 8, where it can be seen that all error measures are almost identical in both cases of adaptive mesh refinement. Mesh adaptation based on the functional error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ instead of the error indicator $|||\epsilon\nabla v - y^*|||_{*(O_i)}$ (Figure 9) yields approximately twice smaller efficiency indices in energy

Example 1 (2D): $k_1 = 0.15$, $k_2 = 0.4$, $\epsilon_1 = 1$, $\epsilon_2 = 100$						
# elts	$\frac{\|v-u\|_0}{\|u\|_0}$ [%]	$\frac{\|\|\|\nabla(v-u)\|\|\|}{\|\|\|\nabla u\|\|\|}$ [%]	$\frac{\|\|\|y^*-p^*\|\|\|_*}{\|\|\|p^*\|\|\|_*}$ [%]	$2M^2_\oplus(v,y^*)$	$2M^2_\oplus(v,p^*)$	$2M^2_\oplus(u,y^*)$
196	15.0077	51.5582	86.1021	1778.14	66.5980	1711.54
347	5.69339	30.8534	41.7241	703.594	20.7780	682.816
630	4.20384	21.7715	31.4858	217.719	10.2201	207.498
1315	2.39552	15.8532	23.1244	76.8018	5.37574	71.4261
2865	1.87075	11.7353	17.1655	33.9310	2.94414	30.9869
5938	0.64611	7.93001	11.4692	16.0812	1.33874	14.7425
12006	0.36985	5.64786	8.23544	7.75232	0.67872	7.07360
24571	0.16023	3.94241	5.76054	3.85268	0.33039	3.52229
48483	0.08909	2.80265	4.09366	1.90043	0.16682	1.73361
97423	0.03961	1.97875	2.88455	0.96275	0.08304	0.87970
192905	0.02230	1.39832	2.03200	0.47524	0.04136	0.43388
386185	0.01015	0.99471	1.44616	0.24134	0.02082	0.22052

Table 1: Constituent parts of main error identity (3.3) for Example 1 (2D).

Example 1 (2D): $k_1 = 0.15$, $k_2 = 0.4$, $\epsilon_1 = 1$, $\epsilon_2 = 100$				
# elts	$\|\|\|\nabla(v-u)\|\|\|^2$	$\|\|\|y^*-p^*\|\|\|_*^2$	$2D_F(v,-\Lambda^* p^*)$	$2D_F(u,-\Lambda^* y^*)$
196	56.5057	157.588	10.0923	1553.95
347	20.2350	37.0058	0.54296	645.811
630	10.0756	21.0729	0.14450	186.425
1315	5.34235	11.3668	0.03338	60.0593
2865	2.92742	6.26338	0.01671	24.7235
5938	1.33673	2.79619	0.00200	11.9462
12006	0.67805	1.44169	0.00067	5.63191
24571	0.33038	0.70538	0.00001	2.81691
48483	0.16696	0.35622	0.00000	1.37739
97423	0.08323	0.17687	0.00000	0.70283
192905	0.04156	0.08777	0.00000	0.34611
386185	0.02103	0.04445	0.00000	0.17606

Table 2: Constituent parts of error identity (3.6) for example Example 1 (2D).

Example 1 (2D): $k_1 = 0.15$, $k_2 = 0.4$, $\epsilon_1 = 1$, $\epsilon_2 = 100$						
# elts	$\frac{D_F(v,-\Lambda^* y^*)}{M^2_\oplus(v,y^*)}$ [%]	$I^{\text{CEN,Low}}_{\text{Eff}}$	$I^{\text{CEN,Up}}_{\text{Eff}}$	$I^{\text{E,Up}}_{\text{Eff}}$	$P^{\text{CEN}}_{\text{rel}}$ [%]	True rel. error in CEN [%]
196	89.0701	0.67371	2.88191	5.60966	74.6973	70.9641
347	92.4942	0.67919	3.50597	5.89671	36.2638	36.6935
630	85.9525	0.70066	2.64380	4.64848	27.1574	27.0680
1315	78.2616	0.70681	2.14392	3.79158	19.9383	19.8250
2865	72.8992	0.70729	1.92142	3.40452	14.7523	14.7032
5938	74.3009	0.70708	1.97256	3.46846	9.87419	9.85973
12006	72.6473	0.70722	1.91238	3.38130	7.06762	7.06119
24571	73.1176	0.70708	1.92864	3.41485	4.93753	4.93591
48483	72.4826	0.70694	1.90588	3.37371	3.50789	3.50805
97423	73.0084	0.70678	1.92392	3.40108	2.47256	2.47347
192905	72.8486	0.70629	1.91692	3.38145	1.74226	1.74418
386185	72.9912	0.70546	1.91972	3.38748	1.23829	1.24114

Table 3: Efficiency indices for Example 1 (2D).

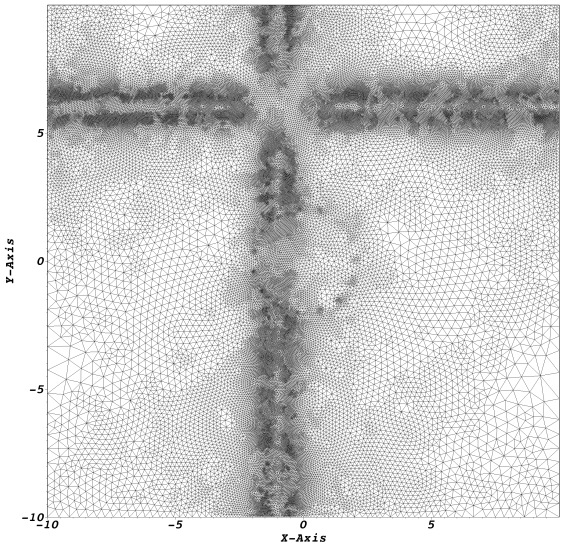

Figure 2: Mesh on the 9th level of AMR (97 423 elements) based on the error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ with flux equilibration for y^*.

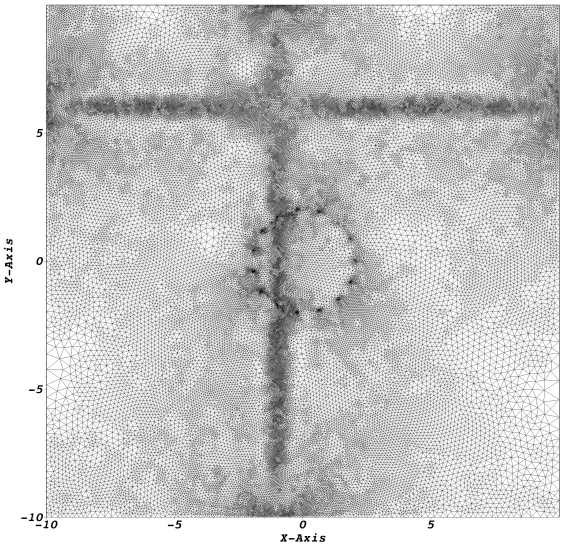

Figure 3: Mesh on the 9th level of AMR (97 353 elements) based on the error indicator $|||\epsilon\nabla v - y^*|||_{*(O_i)}$ with flux equilibration for y^*.

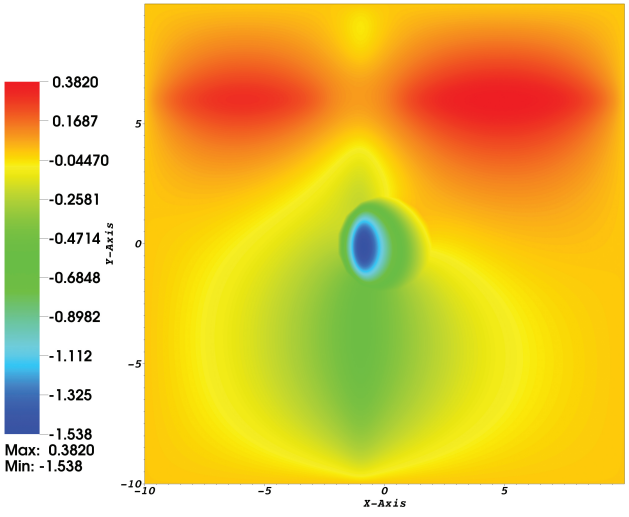

Figure 4: Reference solution for Example 1 (2D).

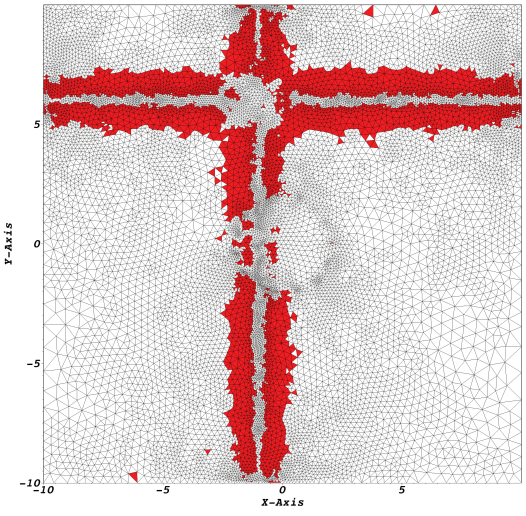

Figure 5: Mesh on the 7th level of AMR (24 122 elements) based on the error indicator $|||\epsilon\nabla v - y^*|||_{*(O_i)}$ with flux equilibration for y^*. The elements are marked by applying the error indicator $\|\sqrt{2}\eta\|_{L^2(K)}$ and using the greedy algorithm with bulk factor 0.3.

and combined energy norms and approximately twice smaller values for the full error $M_\oplus^2(v, p^*) + M_\oplus^2(u, y^*)$ on meshes with a comparable number of elements. The reason for the higher efficiency indices is that no adaptive control is applied on the nonlinear part of the error measure in (3.19), and consequently, the ratio $D_F(v, -\Lambda^* y^*)/M_\oplus^2(v, y^*)$ is increasing, reaching values close to 100 % on fine meshes. However, the error in $|||\nabla(v-u)|||$ and $|||y^* - p^*|||_*$ might be a little higher in the case of the functional error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$. For example, on the mesh from Figure 5 with 24 122 elements, $M_\oplus^2(v, p^*) + M_\oplus^2(u, y^*) = 3.8314$, $|||\nabla(v-u)||| = 0.4674$, $|||y^* - p^*|||_* = 0.6540$, whereas on a mesh with 24 571 elements from the sequence adapted with the indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$, we obtained a value of 1.9263 for $M_\oplus^2(v, y^*)$, and 0.574791 and 0.8399 for $|||\nabla(v-u)|||$ and $|||y^* - p^*|||_*$, respectively. This shows that, reducing the error in $\operatorname{div} y^*$, the functional error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ provides a better approximation for the primal and dual problem together.

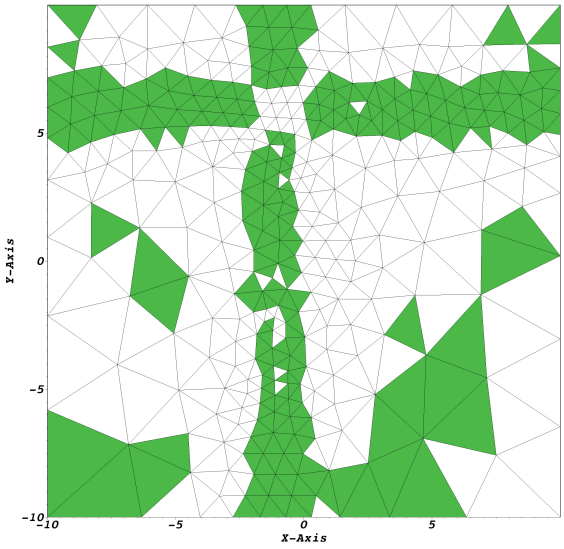

Figure 6: Mesh on the 2nd level of AMR (630 elements) based on the error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ with flux equilibration for y^*. The elements are marked by applying the true error $\sqrt{2M_\oplus^2(v,p^*)+2M_\oplus^2(u,y^*)}$ as an indicator using greedy algorithm with bulk factor 0.5.

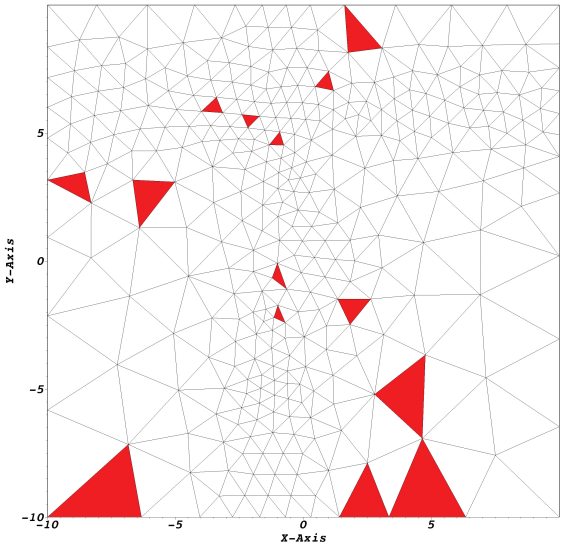

Figure 7: Mesh on the 2nd level of AMR (630 elements) based on the error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ with flux equilibration for y^*. Here we mark red those elements, which differ in the markings based on the indicator $\|\sqrt{2}\eta\|_{L^2(K)}$ and on the true error $\sqrt{2M_\oplus^2(v,p^*)+2M_\oplus^2(u,y^*)}$. Marking is done by greedy algorithm with bulk factor 0.5.

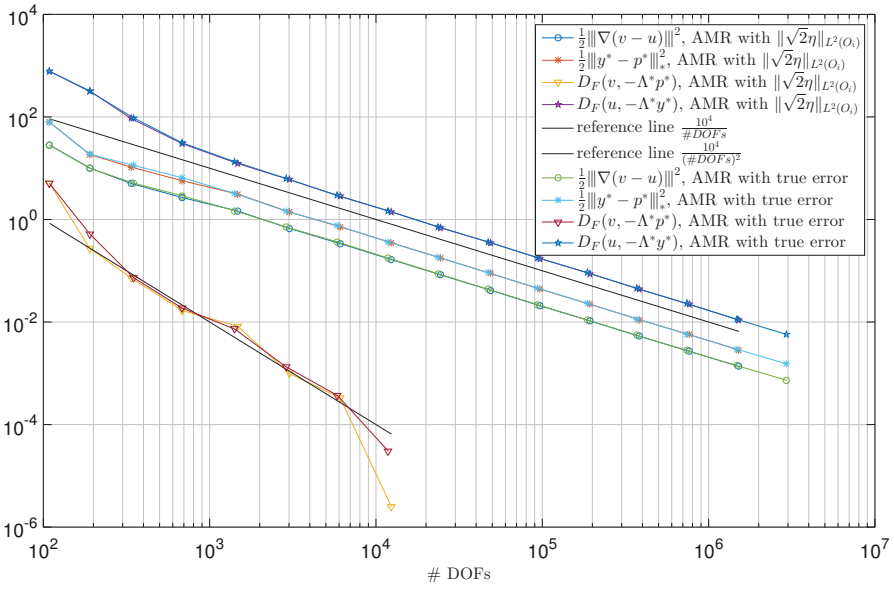

Figure 8: Comparison of errors for AMR based on the functional error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ versus AMR based on the indicator generated by the true error $\sqrt{2M_\oplus^2(v,p^*)+2M_\oplus^2(u,y^*)}$.

Now we want to demonstrate that flux equilibration is indeed an important subtask to make the proposed error bounds reliable and efficient. For this purpose, we use a simple global gradient averaging procedure, i.e., project the numerical flux $\epsilon\nabla v \in L^2(\Omega)$ onto the subspace $[V_h]^2$, where V_h is the finite element space of continuous piecewise linear functions. Then the problem in Example 1 is solved adaptively once applying the functional error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ and next applying the error indicator $|||\epsilon\nabla v - y^*|||_{*(O_i)}$. Figure 10 shows the adapted mesh with 563 965 elements, which is a part of a sequence of meshes obtained applying the functional error indicator with gradient averaging for y^*. Figure 11 shows a mesh with 444 092 elements, which is part of a sequence of meshes adapted using the second indicator with gradient averaging

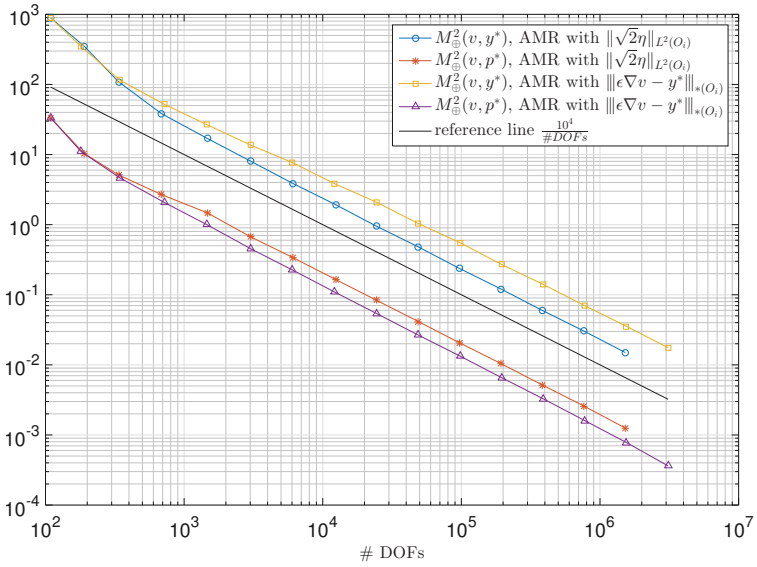

Figure 9: Comparison of errors for AMR based on the functional error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ versus AMR based on the indicator $|||\epsilon\nabla v - y^*|||_{*(O_i)}$.

Example 1 (2D): $k_1 = 0.15$, $k_2 = 0.4$, $\epsilon_1 = 1$, $\epsilon_2 = 100$			
# elts	**# marked elts with true error**	**# differently marked elts**	**differently marked elts in % of all mesh elts**
196	62	6	3.06122
347	150	10	2.88184
630	288	14	2.22222
1315	632	39	2.96578
2865	1439	113	3.94415
5938	2949	216	3.63759
12006	5981	534	4.44778
24571	12099	961	3.91111
48483	24194	2233	4.60574
97423	47784	4012	4.11812

Table 4: Marking based on true error and functional error indicator in Example 1 (2D).

for y^*. Comparing with the results based on flux equilibration for y^*, it can be seen that the mesh in Ω_2 close to the interface Γ is refined too much for both error indicators. Apart from that, the meshes on Figures 11 and 3 look quite similar, unlike the meshes on Figures 10 and 2. For meshes with a comparable number of elements, applying the indicator $|||\epsilon\nabla v - y^*|||_{*(O_i)}$ using gradient averaging instead of flux equilibration, we obtained $\sim 30\,\%$ larger values for the error $|||\nabla(v-u)|||$ and $60\,\%$ larger values for the error $|||y^* - p^*|||_*$. The difference in the errors when applying the functional indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ with gradient averaging for y^* instead of flux equilibration resulted in an even more drastic increase of the error, namely, between $40\,\%$ and $180\,\%$ for $|||\nabla(v-u)|||$ and between $64\,\%$ and $66\,\%$ for $|||y^* - p^*|||_*$, where the meshes had between 21 528 and 563 965 elements. In both cases, we obtained an increasing sequence of efficiency indices with respect to energy and combined energy norms reaching values of 133 and 107 with the functional error indicator on a mesh with 2 089 022 elements, and 570 and 269 with the error indicator $|||\epsilon\nabla v - y^*|||_{*(O_i)}$ on a mesh with 2 954 218 elements. This is due to the fact that the nonlinear term $D_F(u, -\Lambda^* y^*)$, which measures the error in $\operatorname{div} y^*$ (see (3.16) and (3.18)), dominates the other terms in the nonlinear measure $M^2_\oplus(v, p^*) + M^2_\oplus(u, y^*)$ for the error, reaching more than $99.99\,\%$ of it in both cases. In both experiments with gradient averaging for y^*, increasing values of $D_F(u, -\Lambda^* y^*)$ are in correspondence with increasing error $\|\operatorname{div} y^* - \operatorname{div} p^*\|_{L^2(\Omega)}$ and increasing efficiency indices.

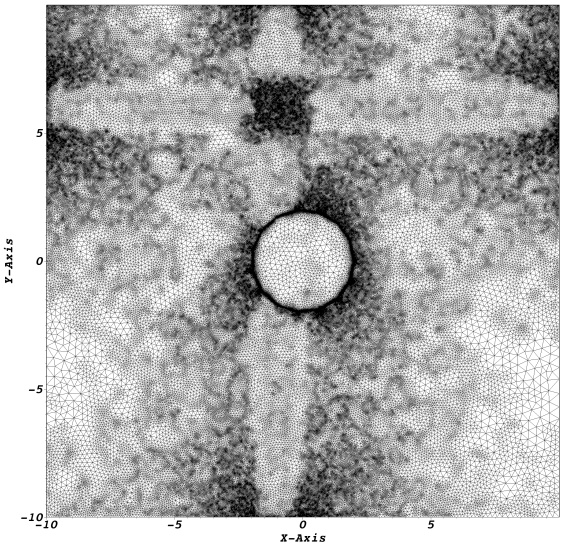

Figure 10: Mesh with 563 965 elements, adapted using the error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ with gradient averaging for y^*.

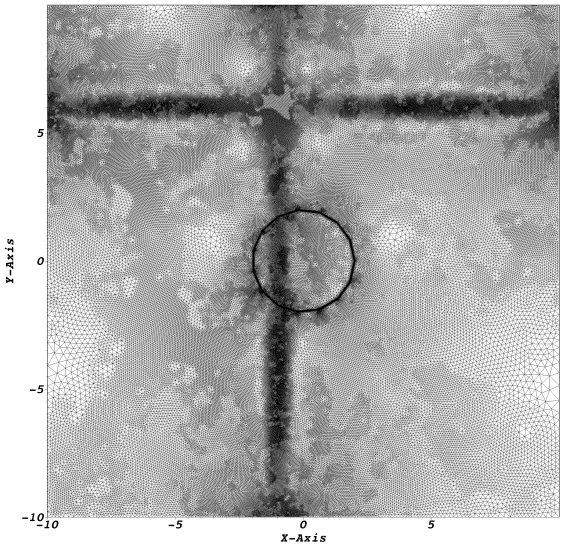

Figure 11: Mesh with 444 092 elements, adapted using the error indicator $|||\epsilon\nabla v - y^*|||_{*(O_i)}$ with gradient averaging for y^*.

4.2 Example 2 (2D Problem)

Figures 13 and 15 show how meshes depend on the indicator in another example, where $\epsilon_1 = 1\,\epsilon_2 = 100$, $k_1 = 0.2$, $k_2 = 0.3$. The functions

$$g = \exp\Big(-b_1\Big(\frac{|x-c_1|^2}{\sigma_1^2} - 1\Big)\Big) - \exp\Big(-b_2\Big(\frac{|x-c_2|^2}{\sigma_2^2} - 1\Big)\Big) \quad \text{and} \quad l = \exp\Big(-b_3\Big(\frac{|x|^2}{\sigma_3^2} - 1\Big)\Big)\sin\Big(\frac{x_1x_2}{4}\Big),$$

where $b_1 = 2.2$, $b_2 = 2.5$, $b_3 = 6$, $c_1 = (-1, 0)$, $c_2 = (5, 5)$, $\sigma_1 = \sigma_2 = 2$, $\sigma_3 = 10$. The indicator $|||\epsilon\nabla v - y^*|||_{*(O_i)}$ correctly approximates elementwise errors in the combined energy norm but does not capture the rest of the error, which results from the nonlinearity $k^2\sinh(u+w)$ and the right-hand side l in (1.1). On the other hand, the term $D_F(v, -\Lambda^* y^*)$ controls the error $D_F(v, -\Lambda^* p^*) + D_F(u, -\Lambda^* y^*)$, and this is the reason why the mesh on Figure 13 resembles the wavy features of the function $f = -k^2\sinh(u+w) + l$.

The isolines of the reference solution and of the function f are depicted on Figures 14 and 12.

4.3 Example 3 (3D Problem)

Here we consider an example close to a real physical problem. The computational domain Ω is a cube of side length 20 angstrom with a triangulated water molecule Ω_1 in it. The diameter of the water molecule, which is positioned in the center of the cube, is about 2.75 angstrom. Its shape is not changed during the mesh adaptation process. The surface mesh of the water molecule is taken from [28]. Figure 16 illustrates the initial tetrahedral mesh, which consists of 60 222 elements. It is generated using TetGen [25] and adaptively refined with the help of mmg3d [7]. Using the localized error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ computed on each vertex patch O_i of the mesh, a new local mesh size at each vertex is defined by the formula

$$h_i^{\text{new}} = h_i^{\text{old}}\Big(\max\Big\{\min\Big\{\frac{\text{AM}\{\|\sqrt{2}\eta\|_{L^2(O_j)}\}}{\|\sqrt{2}\eta\|_{L^2(O_i)}}, 1\Big\}, 0.35\Big\}\Big)$$

and supplied to mmg3d, where $\text{AM}\{\|\sqrt{2}\eta\|_{L^2(O_j)}\}$ is the arithmetic mean of $\{\|\sqrt{2}\eta\|_{L^2(O_j)}\}$ over all vertex patches O_j. The coefficients ϵ and k for this example are typical for electrostatic computations in biophysics using the PBE and are given by

$$\epsilon(x) = \begin{cases} \epsilon_1 = 2, & x \in \Omega_1, \\ \epsilon_2 = 80, & x \in \Omega_2, \end{cases} \qquad k(x) = \begin{cases} k_1 = 0, & x \in \Omega_1, \\ k_2 = 0.84, & x \in \Omega_2. \end{cases}$$

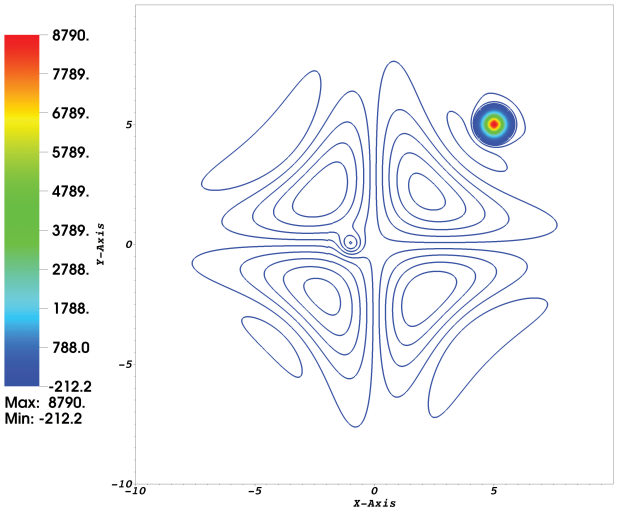

Figure 12: Function $f = -k^2 \sinh(u + w) + l$.

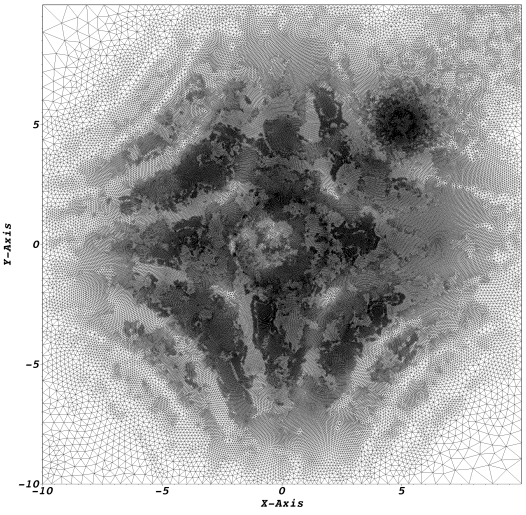

Figure 13: Mesh with 395 935 elements, obtained by AMR using the error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ with flux equilibration for y^*.

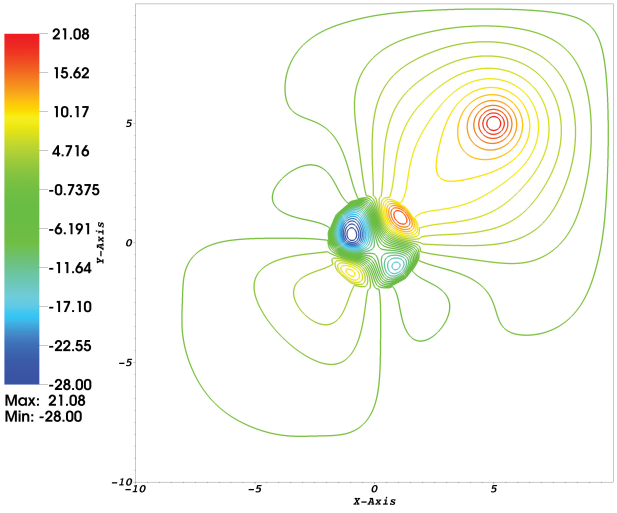

Figure 14: Reference solution.

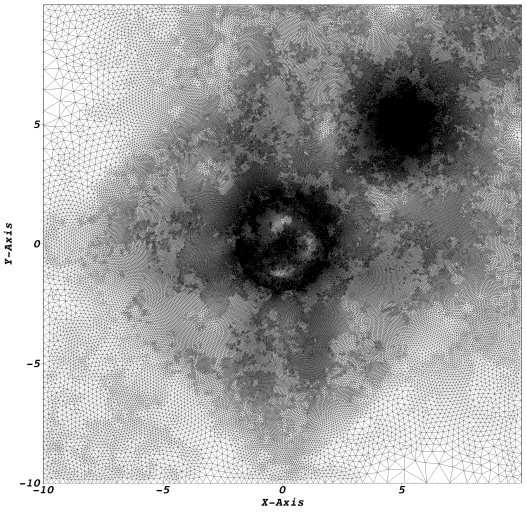

Figure 15: Mesh with 555 489 elements, obtained by AMR using the error indicator $|||\epsilon\nabla v - y^*|||_{*(O_i)}$ with flux equilibration for y^*.

We consider the homogeneous problem, i.e., $l = 0$, and

$$g = \exp\Big(-b_1\Big(\frac{|x-c_1|^2}{\sigma_1^2} - 1\Big)\Big) - \exp\Big(-b_2\Big(\frac{|x-c_2|^2}{\sigma_2^2} - 1\Big)\Big) + \exp\Big(-b_3\Big(\frac{|x-c_3|^2}{\sigma_3^2} - 1\Big)\Big) + \exp\Big(-b_4\Big(\frac{|x-c_4|^2}{\sigma_4^2} - 1\Big)\Big),$$

where $b_1 = b_2 = b_3 = b_4 = 2.3$, $c_1 = (1, 1, 0)$, $c_2 = (4, 4, 0)$, $c_2 = (0, 6, 0)$, $c_2 = (-5, 0, 0)$, $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 2$. The reference solution $z_{h_{\text{ref}}}$ is computed on a very fine mesh, obtained after a sequence of adaptive mesh refinements, that contains 79 917 007 tetrahedrons.

Since $l = 0$ in Ω_1 is a constant function, the patchwise reconstruction from [1] produces a flux y^* with zero divergence in Ω_1 and, therefore, the reliability of our majorant is guaranteed. In this example, we achieved a very tight guaranteed bound on the error in combined energy norm, as well as in energy norm. The efficiency index $I_{\text{Eff}}^{\text{CEN,Up}}$ settles at around 1.05 and the efficiency index $I_{\text{Eff}}^{\text{E,Up}}$ decreases to 1.30 (Table 6). This is

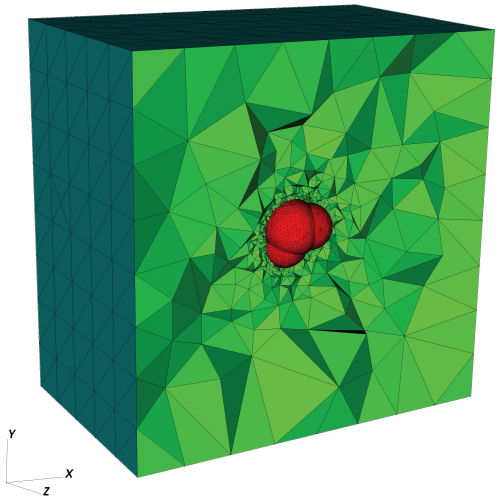

Figure 16: Initial mesh in Example 3 consisting of 60 222 tetrahedrons.

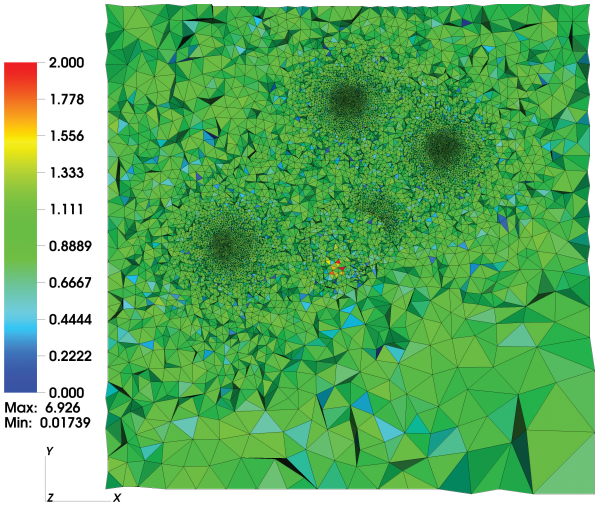

Figure 17: Ratio of the error indicator $|||\epsilon\nabla v - y^*|||_*$ and combined energy norm of the error, elementwise. Mesh on the 4th level of AMR (1.1736e+06 elements) in Example 3 using the error indicator $\|\sqrt{2}\eta\|_{L^2(O_i)}$ with flux equilibration for y^*.

Example 3 (3D): $k_1 = 0$, $k_2 = 0.84$, $\epsilon_1 = 2$, $\epsilon_2 = 80$						
# elts	$\frac{\|v-u\|_0}{\|u\|_0}$ [%]	$\frac{\|\|\nabla(v-u)\|\|}{\|\|\nabla u\|\|}$ [%]	$\frac{\|\|y^*-p^*\|\|_*}{\|\|p^*\|\|_*}$ [%]	$2M^2_\oplus(v,y^*)$	$2M^2_\oplus(v,p^*)$	$2M^2_\oplus(u,y^*)$
60222	76.8320	108.015	167.589	425569	117373	308196
103236	11.9257	46.3306	55.1210	47104.5	17845.0	29259.5
222118	1.09233	17.7353	14.9578	4484.44	2224.69	2259.75
552936	0.49820	8.67222	7.09062	965.067	513.706	451.361
1.1736 e+06	0.25609	6.58075	5.33661	539.734	295.254	244.481
2.05668e+06	0.17094	5.37625	4.18207	350.648	197.016	153.631
2.97315e+06	0.12317	4.73466	3.53852	265.167	152.783	112.385
3.90692e+06	0.10071	4.32886	3.12966	216.336	127.703	88.6336

Table 5: Constituent parts of main error identity (3.3) for Example 3 (3D).

in a good agreement with the fact that, in this example, the ratio $D_F(v, -\Lambda^* y^*)/M^2_\oplus(v, y^*)$ is well controlled and decreases to around 10 % (column 2 in Table 7). We also note that, in this example, we obtained very similar results with the error indicator $|||\epsilon\nabla v - y^*|||_{*(O_i)}$. For the efficiency index $I^{\mathrm{CEN,Low}}_{\mathrm{Eff}}$ of the lower bound on the combined energy norm of the error, we obtain values converging to approximately 0.7071, which is the approximate value of $\frac{\sqrt{2}}{2}$ (column 3 in Table 6). This means that the combined energy norm of the error

$$\sqrt{|||\nabla(v-u)|||^2 + |||y^* - p^*|||^2_*}$$

is practically equal to $|||\epsilon\nabla v - y^*|||_*$.

Another consequence of this fact is the good accuracy of the practical estimation $P^{\mathrm{CEN}}_{\mathrm{rel}}$ of the relative error in combined energy norm (columns 6 and 7 in Table 6). The tight bounds on the error also enable us to compute tight and guaranteed upper bounds on the relative error in energy norm:

$$\frac{|||\nabla(v-u)|||}{|||\nabla u|||} \le \frac{\sqrt{2M^2_\oplus(v,y^*)}}{|||\nabla v||| - \sqrt{2M^2_\oplus(v,y^*)}} =: \mathrm{RE}^{\mathrm{Up}}, \tag{4.1}$$

where (4.1) is valid if $|||\nabla v||| - \sqrt{2M^2_\oplus(v,y^*)} > 0$.

As a remark, we note that the efficiency indices with respect to the energy and combined energy norms of the error can be improved if we use a flux reconstruction in a bigger space, say, RT_1, which has better

Example 3 (3D): $k_1 = 0$, $k_2 = 0.84$, $\epsilon_1 = 2$, $\epsilon_2 = 80$					
# elts	$\|\|\|\nabla(v-u)\|\|\|^2$	$\|\|\|y^*-p^*\|\|\|_*^2$	$2D_F(v,-\Lambda^*p^*)$	$2D_F(u,-\Lambda^*y^*)$	$\mathrm{RE}^{\mathrm{Up}}$ [%]
60222	79487.0	191346	37886.0	116850	—
103236	14623.9	20699.7	3221.12	8559.78	310.049
222118	2142.92	1524.28	81.7757	735.474	33.9219
552936	512.376	342.528	1.32980	108.833	13.4714
1.1736 e+06	295.039	194.026	0.21458	50.455	9.75647
2.05668e+06	196.919	119.155	0.09743	34.4762	7.72193
2.97315e+06	152.724	85.3044	0.05857	27.0805	6.64970
3.90692e+06	127.666	66.7303	0.03663	21.9033	5.96873

Table 6: Constituent parts of error identity (3.6) for Example 3 (3D).

Example 3 (3D): $k_1 = 0$, $k_2 = 0.84$, $\epsilon_1 = 2$, $\epsilon_2 = 80$						
# elts	$\frac{D_F(v,-\Lambda^*y^*)}{M_\oplus^2(v,y^*)}$ [%]	$I_{\mathrm{Eff}}^{\mathrm{CEN,Low}}$	$I_{\mathrm{Eff}}^{\mathrm{CEN,Up}}$	$I_{\mathrm{Eff}}^{\mathrm{E,Up}}$	$P_{\mathrm{rel}}^{\mathrm{CEN}}$ [%]	True rel. error in CEN [%]
60222	40.0541	0.68627	1.25353	2.31386	92.8434	140.985
103236	20.4500	0.72828	1.15478	1.79473	47.6870	50.9159
222118	16.1172	0.71615	1.10583	1.44661	16.4040	16.4054
552936	11.2249	0.70786	1.06248	1.37241	7.90966	7.92099
1.1736 e+06	9.33477	0.70731	1.05053	1.35254	5.98505	5.99106
2.05668e+06	9.82289	0.70725	1.05327	1.33442	4.81343	4.81632
2.97315e+06	10.2057	0.70722	1.05547	1.31767	4.17784	4.17960
3.90692e+06	10.1194	0.70719	1.05492	1.30175	3.77592	3.77716

Table 7: Efficiency indices for Example 3 (3D).

approximation properties. In this way, the error in $\operatorname{div} y^*$ will decrease and, as a result, the term $D_F(v,-\Lambda^*y^*)$ and consequently the dual part of the error $2M_\oplus^2(u,y^*) = |||y^*-p^*|||_*^2 + D_F(u,-\Lambda^*y^*)$ will constitute a smaller part of the whole majorant and the error, respectively. Even better, we can minimize the majorant with respect to y^* in a subspace of $H(\operatorname{div};\Omega)$ like RT_0, possibly on another finer mesh. Note that, in the limit case, we have $\inf_{y^*\in H(\operatorname{div};\Omega)} M_\oplus^2(v,y^*) = M_\oplus^2(v,p^*) = \frac{1}{2}|||\nabla(v-u)|||^2 + D_F(v,-\Lambda^*p^*)$, and the dual error completely vanishes. In this case,

$$I_{\mathrm{Eff}}^{\mathrm{CEN,Up}} = I_{\mathrm{Eff}}^{\mathrm{E}} = \frac{\sqrt{2M_\oplus^2(v,p^*)}}{|||\nabla(v-u)|||} = \frac{\sqrt{|||\nabla(v-u)|||^2 + 2D_F(v,-\Lambda^*p^*)}}{|||\nabla(v-u)|||},$$

where the last ratio tends to 1 because, by (3.15) and (3.17), the term $D_F(v,-\Lambda^*p^*) \sim \|v-u\|_{L^2(\Omega_2)}^2$ and has a higher order of convergence than $|||\nabla(v-u)|||^2$. In practice, we can minimize the majorant with respect to y^* only once on a sufficiently big subspace of $H(\operatorname{div};\Omega)$ to find some good approximation $\overline{y}^*$ of p^* and then reuse this $\overline{y}^*$ and obtain guaranteed and tight bounds on the error in energy and combined energy norm.

A key factor that determines the efficiency index is the ratio $\frac{D_F(v,-\Lambda^*y^*)}{M_\oplus^2(v,y^*)}$. Assuming that

$$D_F(v,-\Lambda^*y^*) \approx D_F(v,-\Lambda^*p^*) + D_F(u,-\Lambda^*y^*),$$

which means that the last term in (3.24) is close to zero, we obtain from (3.19) the estimate

$$I_{\mathrm{Eff}}^{\mathrm{CEN,Up}} \approx \frac{1}{\sqrt{1-\frac{D_F(v,-\Lambda^*y^*)}{M_\oplus^2(v,y^*)}}}.$$

From what we have observed, the efficiency index $I_{\mathrm{Eff}}^{\mathrm{E,Up}}$ with respect to the energy norm usually is no more than twice bigger than $I_{\mathrm{Eff}}^{\mathrm{CEN,Up}}$ (assuming we have a good approximation y^* to p^*). Therefore, if, during the computations, we detect that this ratio is increasing, we can apply the so-called estimation with one step delay, i.e., compute the value of the majorant $M_\oplus^2(v,y^*)$ for the current mesh level with the reconstructed y^* from the next level.

5 Conclusions

We proved the existence and uniqueness of a solution u of the nonlinear elliptic problem (1.1), which appears in the context of solving the nonlinear PBE numerically by two- or three-term regularization. Also, we established the $L^\infty(\Omega)$ regularity of the solution u and an analog of Cea's lemma (cf. (3.26)). These results are used to prove convergence of P_1 FEM approximations under natural conditions on regularity of the meshes used in the constructions of Galerkin approximations.

The main result is the error identity (3.19). We deduced it by finding explicit forms of the terms in general error relations (3.1) and (3.4). The identity defines a natural error measure for the considered class of problems and forms a basis for fully computable and guaranteed tight bounds on the global errors (Table 6).

An advantage of the suggested approach is that it can be used for any conforming approximation (e.g., P^1 or P^2 finite element, IGA, or spectral approximations) and that the estimates do not contain local (mesh dependent) constants or unknown global constants.

As we have confirmed by our theoretical findings and numerical experiments, flux equilibration is an important paradigm to obtain an accurate error indicator (cf. Figures 10 and 11) as well as to guarantee that the last term in (3.11) does not dominate the majorant.

References

[1] D. Braess and J. Schöberl, Equilibrated residual error estimator for Maxwell's equations, RICAM report 2006-19, Austrian Academy of Sciences, 2006.
[2] H. Brézis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, New York, 2011.
[3] H. Brézis and F. E. Browder, Sur une propriété des espaces de Sobolev, *C. R. Acad. Sci. Paris Sér. A-B* **287** (1978), no. 3, A113–A115.
[4] C. Carstensen, M. Feischl, M. Page and D. Praetorius, Axioms of adaptivity, *Comput. Math. Appl.* **67** (2014), no. 6, 1195–1253.
[5] L. Chen, M. J. Holst and J. Xu, The finite element approximation of the nonlinear Poisson–Boltzmann equation, *SIAM J. Numer. Anal.* **45** (2007), no. 6, 2298–2320.
[6] H. Childs, E. Brugger, B. Whitlock, J. Meredith, S. Ahern, D. Pugmire, K. Biagas, M. Miller, C. Harrison, G. H. Weber, H. Krishnan, T. Fogal, A. Sanderson, C. Garth, E. Wes Bethel, D. Camp, O. Rübel, M. Durant, J. M. Favre and P. Navrátil, VisIt: An end-user tool for visualizing and analyzing very large data, in: *High Performance Visualization–Enabling Extreme-Scale Scientific Insight*, CRC Press, Boca Raton (2012), 357–372.
[7] C. Dobrzynski, MMG3D: User guide, Technical Report RT-0422, INRIA, 2012.
[8] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.
[9] M. Feischl, D. Praetorius and K. G. van der Zee, An abstract analysis of optimal goal-oriented adaptivity, *SIAM J. Numer. Anal.* **54** (2016), no. 3, 1423–1448.
[10] F. Fogolari, A. Brigo and H. Molinari, The Poisson–Boltzmann equation for biomolecular electrostatics: A tool for structural biology, *J. Mol. Recognit.* **15** (2002), 377–392.
[11] F. Fogolari, P. Zuccato, G. Esposito and P. Viglino, Biomolecular electrostatics with the linearized Poisson–Boltzmann equation, *Biophys. J.* **76** (1999), 1–16.
[12] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Classics Math., Springer, Berlin, 2001.
[13] F. Hecht, New development in freefem++, *J. Numer. Math.* **20** (2012), no. 3–4, 251–265.
[14] M. Holst, J. A. McCammon, Z. Yu, Y. C. Zhou and Y. Zhu, Adaptive finite element modeling techniques for the Poisson–Boltzmann equation, *Commun. Comput. Phys.* **11** (2012), no. 1, 179–214.
[15] B. Kawohl and M. Lucia, Best constants in some exponential Sobolev inequalities, *Indiana Univ. Math. J.* **57** (2008), no. 4, 1907–1927.
[16] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Class. Appl. Math. 31, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2000.
[17] P. Neittaanmäki and S. Repin, *Reliable Methods for Computer Simulation. Error Control and a Posteriori Estimates*, Stud. Math. Appl. 33, Elsevier Science, Amsterdam, 2004.
[18] H. Oberoi and N. Allewell, Multigrid solution of the nonlinear Poisson–Boltzmann equation and calculation of titration curves, *Biophys. J.* **65** (1993), 48–55.

[19] S. I. Repin, A posteriori error estimation for variational problems with uniformly convex functionals, *Math. Comp.* **69** (2000), no. 230, 481–500.
[20] S. I. Repin, On measures of errors for nonlinear variational problems, *Russian J. Numer. Anal. Math. Modelling* **27** (2012), no. 6, 577–584.
[21] S. Repin and J. Valdman, Error identities for variational problems with obstacles, *ZAMM Z. Angew. Math. Mech.* **98** (2018), no. 4, 635–658.
[22] I. Sakalli, J. Schöberl and E. W. Knapp, mfes: A robust molecular finite element solver for electrostatic energy computations, *J. Chem. Theory Comput.* **10** (2014), 5095–5112.
[23] K. Sharp and B. Honig, Calculating total electrostatic energies with the nonlinear Poisson–Boltzmann equation, *J. Phys. Chem* **94** (1990), 7684–7692.
[24] R. E. Showalter, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, 1977.
[25] H. Si, TetGen, a Delaunay-based quality tetrahedral mesh generator, *ACM Trans. Math. Software* **41** (2015), no. 2, Article ID 11.
[26] G. Stampacchia, Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus, *Ann. Inst. Fourier (Grenoble)* **15** (1965), no. 1, 189–258.
[27] N. S. Trudinger, On imbeddings into Orlicz spaces and some applications, *J. Math. Mech.* **17** (1967), 473–483.
[28] A collection of molecular surface meshes, https://www.rocq.inria.fr/gamma/gamma/download/affichage.php?dir=MOLECULE&name=water_mol&last_page=6, Accessed: 2017-08-18.