Freie Universität Berlin
Department of Mathematics and Computer Science

Adaptive finite element methods for Stokes–Darcy

Master Thesis

submitted by: Moritz Hoffmann

supervised by: Prof. Dr. Volker John
Dr. Alfonso Caiazzo

Berlin, February 10, 2016

Eidesstaatliche Erklärung

Ich, Moritz Hoffmann, erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Masterarbeit selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe.
Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht.
Diese Arbeit wurde in gleicher oder ähnlicher Form noch bei keiner anderen Universität als Prüfungsleistung eingereicht und ist auch noch nicht veröffentlicht.

Berlin, 10.02.2016,

Contents

1 Introduction 4

2 Models for flow problems 5
2.1 The Darcy equations 5
2.1.1 Weak formulation 6
2.1.2 Finite element discretization 9
2.2 The Stokes equations 12
2.2.1 Weak formulation 13
2.2.2 Finite element discretization 16
2.3 The Stokes–Darcy equations 18
2.3.1 Weak formulation 20
2.3.2 Finite element discretization 21

3 A posteriori error estimates 24
3.1 Residual based a posteriori error estimates 27
3.1.1 Application to the Darcy equations 33
3.1.2 Application to the Stokes equations 43
3.2 Dual weighted residual method 50
3.2.1 Application to the Darcy equations 57
3.2.2 Application to the Stokes equations 59
3.2.3 Practical evaluation of the error estimators 61

4 Cell marking strategies 63

5 Cell refinement 64

6 Numerical studies 67
6.1 Scalar convection–diffusion equations 67
6.1.1 The Hemker problem 69
6.1.2 Example with boundary layers 73
6.2 Stokes equations 75
6.2.1 Disc with a crack 75
6.2.2 Flow around a cylinder 81
6.3 Stokes–Darcy problem 87

7 Conclusion and outlook 89

1 Introduction

A popular tool to describe phenomena and problems in physics and engineering is the usage of partial differential equations. They pose an abstract model to the actual situation and allow theoretical as well as application based investigations. Since an analytical solution is often unknown, this approach leads to numerical simulations, making use of computers in order to find a discrete approximation. In the following, the finite element method will be considered, which is a popular numerical technique to find these discrete approximations.

The general form of the finite element method involves choosing a grid for the domain of the considered partial differential equation. However, if the grid is too coarse, one might not resolve the solution sufficiently well. On the other hand, it is usually a priori not known where the grid needs to be finer. A globally finer grid is not always within the realm of possibility because it implies a massive increase in the limited resources of required memory and computational power in order to obtain an approximation.

One way of circumventing this problem is the following: Instead of calculating just one approximated solution on a fixed grid, one tries to derive a posteriori error estimators, i.e., error estimators which give information about the magnitude of the error of a given approximated solution to the actual solution on certain portions of the grid. The hope is that refining the areas of the grid in which the error is particularly large will increase the overall accuracy significantly. Iterating this process results in a so-called adaptive algorithm.

A general adaptive algorithm for stationary problems can look like the following.

1.1 Algorithm [A general adaptive algorithm]**.** This algorithm is a modified version of [Ver13, Algorithm 1.1]. Let the data of a partial differential equation and a tolerance ε be given, then one looks for a numerical solution of the problem with an error less than ε.

(i) Construct an initial coarse mesh $\mathcal{T}_0$ representing sufficiently well the geometry and data of the problem; set $k = 0$.

(ii) *Solve*: Solve the discrete problem associated with $\mathcal{T}_k$.

(iii) *Estimate*: For every element K in $\mathcal{T}_k$ compute a local a posteriori error indicator.

(iv) If the global error estimate is less than ε, stop, otherwise:

  (a) *Mark*: Decide by the local error indicators which elements have to be refined.

  (b) *Refine*: Construct the next mesh $\mathcal{T}_{k+1}$. Increase k by 1 and return to step (ii).

In this thesis, three different partial differential equations are being considered: The incompressible and stationary Stokes equations which model free viscous flow, the Darcy equations which model flow through porous media and convection–diffusion equations which describe the transportation and diffusion of scalar quantities like temperature or concentration, as well as a coupled Stokes–Darcy system which can describe for example the free flow of a river and the resulting flow in its riverbed in the field of geosciences.

The goal is to investigate the adaptive algorithm for these problem types theoretically as well as practically for different kinds of error indicators and different parameters in the marking strategy of step (iv).(a). In particular, the theoretical part of the thesis includes the Stokes, Darcy, and Stokes–Darcy equations, numerical simulations were performed for the convection–diffusion, Stokes, and Stokes–Darcy equations. The convection–diffusion equations were chosen for the numerical simulations because under certain assumptions they can be seen as a generalization of the Darcy equations and it is more likely that the solution of a convection–diffusion problem possesses features that require the application of an adaptive algorithm, making it easier to find suitable examples.

In the following sections, steps (ii)-(iv) of the algorithm are described successively.

Section 2 deals with the Darcy, Stokes, and Stokes–Darcy equations. Existence and uniqueness of a weak solution, as well as the corresponding finite element discretization are discussed for the Darcy and Stokes equations, for the Stokes–Darcy equations, the coupling conditions as well as possible finite element discretizations are introduced. Therefore, this section can be identified with step (ii) of the adaptive algorithm.

In virtue of step (iii), in Section 3 error estimators are being derived first in an abstract setting and then applied to the Stokes, and Darcy equations. The considered error estimators are a residual based a posteriori error estimator in the energy norm and a dual weighted residual error estimator, which indicates the error of the approximated solution with respect to a functional of interest.

The next step of the algorithm is step (iv), being subject of Section 4 and Section 5, which are the counterparts of steps (iv).(a) and (iv).(b), respectively.

Section 6 deals with numerical studies for the convection–diffusion, Stokes and Stokes–Darcy equations. The differences between the residual based a posteriori error estimators and the dual weighted residual error estimators as well as the performance of the adaptive algorithm under certain configurations of step (iv).(a) are discussed.

Section 7 gives conclusions and an outlook.

2 Models for flow problems

This section deals with models for flow problems and their respective discretizations, i.e., it considers point (ii) of Algorithm 1.1. First a model for flow through porous media is introduced, followed by the Stokes equations. Finally, the coupled problem of the two previously introduced models is considered.

2.1 The Darcy equations

The Darcy equations describe the behavior of fluids in porous media like sand. First they were obtained as results from experiments by Henry Darcy in 1856 [Dar56], later it was found that they can be deduced from the Navier–Stokes equations and therefore also pose a theoretical result. The dimensionless problem associated to the equations usually reads as follows:

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary and $f \in L^2(\Omega)$ a source term. Then one wants to find the fluid velocity $\mathbf{u} : \Omega \to \mathbb{R}^d$ and the piezometric head $\varphi : \Omega \to \mathbb{R}$ such that

$$\begin{cases} \mathbf{u} + \mathbb{K}\,\nabla\varphi = 0 & \text{in } \Omega, \\ \nabla \cdot \mathbf{u} = f & \text{in } \Omega, \end{cases} \tag{1}$$

where $\mathbb{K}$ is the hydraulic conductivity tensor, describing the characteristics of the porous medium. The piezometric head gives information on the liquid pressure at a specific place. This formulation of the problem is known as the “mixed form”.

2.1 Remark [On the velocity term]**.** The velocity $\mathbf{u}$ of the Darcy equations is not a pointwise velocity like expected in a free flow, it is rather a specific discharge that happens to have the unit of a velocity: Let $Q[m^3/s]$ be the total discharge (volume per time) and $A[m^2]$ be the cross sectional area to flow, then the not yet nondimensionalized averaged velocity is given by

$$\mathbf{v} := \frac{Q}{A} \left[\frac{m^3}{s \cdot m^2} = \frac{m}{s}\right].$$

By taking the divergence of the first equation of (1) and substituting the second equation, one obtains the simpler "primal form" of the Darcy equations:

$$-\nabla \cdot \mathbb{K}\, \nabla\varphi = f \text{ in } \Omega. \tag{2}$$

Note that the two forms are equivalent if the functions are smooth enough. However in the setting of a finite element discretization as introduced later the two forms are likely to yield different solutions.

To obtain a well-posed problem, one needs boundary conditions. Therefore, decompose the boundary $\partial\Omega$ into two relatively open, disjoint parts $\Gamma_{\text{nat}}, \Gamma_{\text{ess}} \subset \partial\Omega$, such that

$$\overline{\Gamma}_{\text{nat}} \cup \overline{\Gamma}_{\text{ess}} = \partial\Omega \text{ and } \Gamma_{\text{nat}} \cap \Gamma_{\text{ess}} = \emptyset.$$

Here, Γ_{nat} denotes the part of Γ for the natural boundary conditions, i.e., the boundary conditions that are being incorporated into the weak formulation by substitution and Γ_{ess} denotes the part of Γ for the essential boundary conditions, i.e., the boundary conditions that are being incorporated into the ansatz and test spaces. The solution of the Darcy equations should then fulfill

$$\begin{cases} (-\,\mathbb{K}\,\nabla\varphi)\cdot\mathbf{n} = u_{\text{nat}} & \text{on } \Gamma_{\text{nat}}, \\ \varphi = \varphi_{\text{ess}} & \text{on } \Gamma_{\text{ess}}, \end{cases} \tag{3}$$

where $\mathbf{n}$ denotes the outer unit normal vector.

2.1.1 Weak formulation

For simplicity, in the following only the primal form (2) with $\mathbb{K} = K \cdot \mathbb{I}$ is being considered where K denotes a positive scalar value and $\mathbb{I}$ the identity, whereas in general it can only be assumed that $\mathbb{K}$ is symmetric and positive definite. Further it is assumed that the essential boundary has positive measure, i.e., $\text{meas}(\Gamma_{\text{ess}}) > 0$.

To derive a weak formulation, consider a test function ψ which vanishes close to the essential boundary, namely

$$\psi \in C^\infty_{\Gamma_{\text{ess}}}(\Omega) := \left\{ v \in C^\infty(\Omega) \cap H^1(\Omega) : {}^{\exists U \subset \mathbb{R}^d \text{ open neighborhood of } \Gamma_{\text{ess}}}_{\text{s.t. } v(x)=0\ \forall x \in U\cap\Omega} \right\}.$$

Multiplication of (2) with ψ, integration and integration by parts yields

$$(\mathbb{K}\,\nabla\varphi, \nabla\psi)_0 = (f, \psi)_0 - \langle u_{\text{nat}}, \psi\rangle_{\Gamma_{\text{nat}}}.$$

In order to apply Hilbert space theory, a test space is needed that is complete in the norm of $H^1(\Omega)$, which is not the case for $C^\infty_{\Gamma_{\text{ess}}}(\Omega)$. The norm of $H^1(\Omega)$ is given by

$$\|u\|_{H^1(\Omega)} = \|u\|_1 := (\|\nabla u\|_0^2 + \|u\|_0^2)^{\frac{1}{2}}.$$

Consequently, one considers the completion with respect to that norm, i.e.,

$$V := H^1_{\Gamma_{\text{ess}}}(\Omega) = \overline{C^\infty_{\Gamma_{\text{ess}}}(\Omega)}^{H^1(\Omega)}, \tag{4}$$

as a new test space, which now contains functions that vanish on the essential boundary in the sense of traces. With respect to inclusion, the space V is therefore in between $H^1_0(\Omega)$ and $H^1(\Omega)$.

2.2 Remark [Homogeneous essential boundary conditions for the primal form]**.** For the analysis, homogeneous essential boundary conditions can be assumed: Due to the surjectivity of the trace operator, the boundary condition $\varphi_{\text{ess}} \in H^{1/2}(\Gamma_{\text{ess}})$ can be extended into the interior of Ω. Thus with $\tilde{\varphi} := \varphi - \varphi_{\text{ess}}$ being the sought solution of (2) with homogeneous essential boundary conditions, one automatically obtains the solution $\varphi = \tilde{\varphi} + \varphi_{\text{ess}}$ of the original problem.

Even though V is the completion with respect to the $H^1(\Omega)$-norm, one can apply a different norm simplifying the analysis with the following result.

2.3 Theorem [The Poincaré-Friedrichs inequality]**.** Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary and $\operatorname{meas}(\Gamma_{\text{ess}}) > 0$. Then, for all $u \in H^1_{\Gamma_{\text{ess}}}(\Omega)$ one has

$$\|u\|_0 \leq C\|\nabla u\|_0. \tag{5}$$

Proof. The proof relies on the "Norm Equivalence Theorem of Sobolev" which can be found in [Ste08, Theorem 2.6]. It states: Let $f : H^1(\Omega) \to \mathbb{R}$ be a given linear bounded functional satisfying

$$0 \leq |f(v)| \leq c_f \|v\|_1 \quad \forall v \in H^1(\Omega)$$

and $f(c) = 0 \Rightarrow c \equiv 0$ for $c \in P_0(\Omega)$, where $P_0(\Omega)$ denotes the ring of polynomials mapping from Ω to $\mathbb{R}$ of degree 0, i.e., constant polynomials. Then

$$\|v\|_{1,f} := \left(|f(v)|^2 + \|\nabla v\|_0^2\right)^{1/2}$$

defines an equivalent norm in $H^1(\Omega)$.

Now let

$$f(v) := \left(\int_{\Gamma_{\text{ess}}} v^2\right)^{1/2}.$$

It is

$$0 \leq |f(v)| = \left(\int_{\Gamma_{\text{ess}}} v^2\right)^{1/2} \leq \left(\int_{\partial\Omega} v^2\right)^{1/2} = \|v\|_{L^2(\partial\Omega)} \leq c_f \|v\|_{H^1(\Omega)}$$

where the last inequality is due to the boundedness of the trace operator. Further one has for $c \in P_0(\Omega)$ that

$$0 = f(c) = \left(\int_{\Gamma_{\text{ess}}} c^2\right)^{1/2} = |c| \left(\operatorname{meas}(\Gamma_{\text{ess}})\right)^{1/2}$$
$$\Rightarrow c = 0.$$

Therefore all the requirements for the equivalence theorem are satisfied and $\|\cdot\|_{1,f}$ defines an equivalent norm on $H^1(\Omega)$.

Let $u \in H^1_{\Gamma_{\text{ess}}}(\Omega)$ be arbitrary. Since u vanishes on Γ_{ess} in the sense of traces, one has $f(u) = 0$. Due to the norm equivalence it is

$$\begin{aligned} \|u\|_1^2 = \|u\|_0^2 + \|\nabla u\|_0^2 &\leq C_1 \left(f(u) + \|\nabla u\|_0^2\right) \\ \Rightarrow \|u\|_0^2 &\leq C_2 \left(f(u) + \|\nabla u\|_0^2\right) = C_2 \|\nabla u\|_0^2 \end{aligned}$$

for some constants C_1, C_2, C_3. □

Now V can be equipped with the $H^1(\Omega)$-semi-norm $|u|_1 := \|\nabla u\|_0$ which in this case is even a norm since $H^1(\Omega) \subset L^2(\Omega)$ and therefore

$$0 = |u|_1 = \|\nabla u\|_0 \overset{(5)}{\geq} \frac{1}{C}\|u\|_0 \Rightarrow u = 0.$$

The weak formulation then reads as follows: Find $\widetilde{\varphi} := \varphi - \varphi_{\text{ess}} \in V$ with $\varphi \in H^1(\Omega)$ such that

$$a(\widetilde{\varphi}, \psi) = \langle f, \psi\rangle - \langle u_{\text{nat}}, \psi\rangle_{\Gamma_{\text{nat}}} - (\mathbb{K}\,\nabla\varphi_{\text{ess}}, \nabla\psi)_0 =: \langle \tilde{f}, \psi\rangle \;\; \forall\psi \in V, \tag{6}$$

with

$$a : V \times V \to \mathbb{R},\; (\widetilde{\varphi}, \psi) \mapsto (\mathbb{K}\,\nabla\widetilde{\varphi}, \nabla\psi)_0. \tag{7}$$

Here φ_{ess} denotes an extension of the essential boundary data $\varphi_{\text{ess}} \in H^{1/2}(\Gamma_{\text{ess}})$ into the interior of Ω.

2.4 Theorem [Existence and uniqueness of a weak solution]**.** Problem (6) with $f \in L^2(\Omega)$, $u_{\text{nat}} \in H^{-1/2}(\Gamma_{\text{nat}})$ and homogeneous essential boundary data has a unique solution.

Proof. This theorem's proof is based on an application of the theorem of Lax-Milgram, which states that there exists exactly one solution to

$$a(\varphi, \psi) = \langle \tilde{f}, \psi\rangle$$

if $a : V \times V \to \mathbb{R}$ is bounded and coercive and $\tilde{f}$ is linear and bounded.

- Boundedness of $a(\cdot,\cdot)$: By applying $\mathbb{K} = K\,\mathbb{I}$ for some scalar K and the Cauchy–Schwarz inequality one obtains

$$|a(\varphi, \psi)| = K|(\nabla\varphi, \nabla\psi)_0| \leq K|\varphi|_1|\psi|_1.$$

- Coercivity of $a(\cdot,\cdot)$ follows directly from the definition of the V-norm: $a(\varphi, \varphi) = K|\varphi|_1^2$.
- Boundedness of the right-hand side is obtained by the triangle inequality, the Cauchy–Schwarz inequality and the boundedness of the trace operator:

$$\begin{aligned}|\langle \tilde{f}, \psi\rangle| &= |(f, \psi)_0 - \langle u_{\text{nat}}, \psi\rangle_{\Gamma_{\text{nat}}} - (\mathbb{K}\,\nabla\varphi_{\text{ess}}, \nabla\psi)_0| \\ &\leq \|f\|_0\|\psi\|_0 + C\|u_{\text{nat}}\|_{-\frac{1}{2}}|\psi|_1 + K|\varphi_{\text{ess}}|_1|\psi|_1 < \infty.\end{aligned}$$

□

2.5 Remark [Existence and uniqueness for inhomogeneous essential boundary data]**.** When considering the problem with inhomogeneous essential boundary data $\varphi_{\text{ess}} \not\equiv 0$, it can be reduced to a problem with homogeneous essential boundary data by subtracting an extension of it from the sought solution, see Remark 2.2. Since then the sought solution is in V, existence and uniqueness can be provided with the above theorem and therefore holds for the original problem by readding the extension as well.

2.6 Remark [Recovering the Darcy velocity in the primal form]**.** One can recover the Darcy velocity in the primal form by

$$\mathbf{u} = -\,\mathbb{K}\,\nabla\varphi.$$

However in a discrete setting by taking the gradient of the pressure, one loses accuracy and

$$\nabla \cdot \mathbf{u} = f$$

might no longer be satisfied.

2.1.2 Finite element discretization

In order to solve the Darcy equations numerically, one can use a finite element approach. This section corresponds to points (i) and (ii) of Algorithm 1.1. Concerning the first point, let $\mathcal{T}$ be a partition of the domain Ω which satisfies the following conditions:

(i) It is
$$\overline{\Omega} = \bigcup_{T \in \mathcal{T}} T,$$
i.e., there is no additional approximation of the boundary necessary.

(ii) The essential boundary Γ_{ess} is the union of $(d-1)$-dimensional faces of elements in $\mathcal{T}$.

(iii) Affine equivalence: $\mathcal{T}$ consists of simplices and parallelepipeds. Therefore every element in $\mathcal{T}$ is the image under an affine map of either the reference simplex
$$\widehat{K}_d = \left\{ \mathbf{x} \in \mathbb{R}^d : \mathbf{x} \geq 0, \sum_{i=1}^{d} x_i \leq 1 \right\}$$
or the reference cube
$$\widehat{K}_d = [0,1]^d.$$

(iv) Admissibility: Any two elements in $\mathcal{T}$ are either disjoint or share a complete lower dimensional face of their boundaries. This property prevents hanging nodes, see Definition 5.1.

(v) Shape regularity: For any element $K \in \mathcal{T}$, the ratio of its diameter h_K to the diameter ρ_K of the largest ball completely contained in K is bounded from above independently of K, as sketched in Figure 1.

Regarding the last point, to every partition $\mathcal{T}$ the *shape parameter*
$$C_{\mathcal{T}} := \max_{K \in \mathcal{T}} \frac{h_K}{\rho_K} \tag{8}$$
is associated. When considering families of partitions which are for instance obtained by global or local refinement, it must be uniformly bounded with respect to all partitions. In two dimensions, this property ensures that the element angles are all bounded away from zero and therefore do not become too small.

Denote for every partition $\mathcal{T}$ the sets $\mathcal{N}$ containing all 0-dimensional faces, i.e., all vertices, and $\mathcal{E}$ containing all $(d-1)$-dimensional faces, i.e., all facets, of all elements of $\mathcal{T}$. The skeleton Σ of $\mathcal{T}$ is given by the union of all facets in $\mathcal{E}$. Whenever there is a subscript character denoting a set like K, Ω, Γ_{ess} on $\mathcal{N}$ or $\mathcal{E}$, the restriction to the faces of $\mathcal{N}$ or $\mathcal{E}$ respectively contained in the subscript set is being considered. For every d-face and $(d-1)$-face $F \in \mathcal{T} \cup \mathcal{E}$, denote its diameter by $h_F := \operatorname{diam}(F)$. Further, let $\omega_K, \widetilde{\omega}_K, \omega_E, \widetilde{\omega}_E, \omega_z$ be subsets of $\mathcal{T}$ for $K \in \mathcal{T}$, $E \in \mathcal{E}$, and $z \in \mathcal{N}$ as shown in Figure 2 defined by
$$\omega_K = \bigcup_{\substack{K' \in \mathcal{T} \\ \mathcal{E}_K \cap \mathcal{E}_{K'} \neq \emptyset}} K', \quad \widetilde{\omega}_K = \bigcup_{\substack{K' \in \mathcal{T} \\ \mathcal{N}_K \cap \mathcal{N}_{K'} \neq \emptyset}} K', \quad \omega_E = \bigcup_{\substack{K' \in \mathcal{T} \\ E \in \mathcal{E}_{K'}}} K', \quad \widetilde{\omega}_E = \bigcup_{\substack{K' \in \mathcal{T} \\ \mathcal{N}_E \cap \mathcal{N}_{K'} \neq \emptyset}} K', \quad \omega_z = \bigcup_{\substack{K' \in \mathcal{T} \\ z \in \mathcal{N}_{K'}}} K'.$$

The second point of Algorithm 1.1 is based on the Galerkin method, i.e., the infinite-dimensional function space V is being approximated by a finite-dimensional subspace $V^h \subset V$

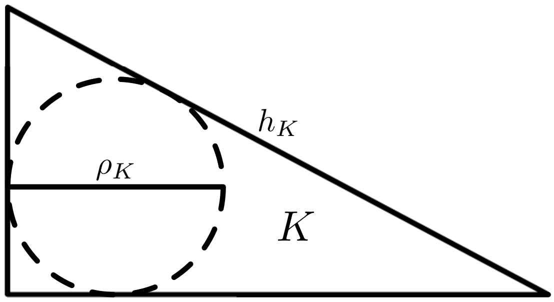


Figure 1: Sketch of the shape parameter of a single simplex in two dimensions.

with a basis $\{\psi_i\}_{i=1}^N$ where h represents the refinement level and N is the dimension and therefore the number of degrees of freedom of V^h.

It however remains to specify how to choose the space V^h. In finite element methods, one can either define the space directly on the elements $K \in \mathcal{T}$ or on reference elements which then get mapped onto the respective elements in $\mathcal{T}$. Here, the latter approach is outlined. To this end, let $P(K) \subset C^s(K)$, $s \in \mathbb{N}$, be a finite-dimensional space on a mesh cell $K \in \mathcal{T}$. Usually, $P(K)$ consists of polynomials. Due to the affine equivalence property of $\mathcal{T}$ these local spaces can also be defined on the reference element by

$$P(K) := \{\varphi \circ F_K^{-1} : \varphi \in P(\widehat{K}_d)\}, \tag{9}$$

where F_K is an affine transform mapping $\widehat{K}_d$ onto $K \in \mathcal{T}$. Let $\Phi_{K,1}, \dots, \Phi_{K,N_K} : C^s(K) \to \mathbb{R}$ be linear functionals on the mesh cells and assume that the local spaces' bases can be transformed into $\{\phi_{K,i}\}_{i=1}^{N_K}$ such that

$$\Phi_{K,i}(\phi_{K,j}) = \delta_{ij}$$

for all $i, j = 1, 2, \dots, N_K$. Further let $\Phi_1, \dots, \Phi_N$ be linear functionals whose restriction on $K \in \mathcal{T}$ yields the previously defined $\Phi_{K,i}$.

The subdomain ω_i denotes the union of all mesh cells K such that there is a $\psi \in P(K)$ with $\Phi_i(\psi) \neq 0$. A function φ defined on Ω with $\varphi\big|_K \in P(K)$ for each $K \in \mathcal{T}$ is called continuous with respect to the functional Φ_i if

$$\Phi_i(\varphi\big|_{K_1}) = \Phi_i(\varphi\big|_{K_2}) \quad \forall K_1, K_2 \in \omega_i.$$

The global finite element space is then defined by

$$V^h = \{\psi : \Omega \to \mathbb{R} : \psi\big|_K \in P(K),\, \psi \text{ is continuous with respect to } \Phi_1, \dots, \Phi_N\} \cap V. \tag{10}$$

The weak problem (6) discretized by the finite element method is then to find $\varphi^h - \varphi^h_{\Gamma_{\text{ess}}} \in V^h$ such that

$$a^h(\varphi^h - \varphi^h_{\Gamma_{\text{ess}}}, \psi^h) := (\mathbb{K}\,\nabla(\varphi^h - \varphi^h_{\Gamma_{\text{ess}}}), \nabla\psi^h)_0 = \langle \widetilde{f}, \psi^h \rangle \tag{11}$$

for all $\psi^h \in V^h$, where $\varphi^h_{\Gamma_{\text{ess}}}$ denotes an extension of the essential boundary data into the domain.

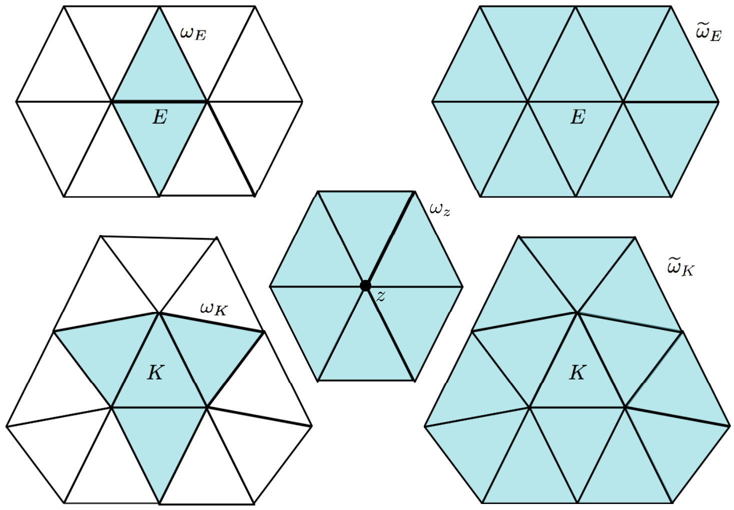


Figure 2: Subsets ω_E (top left), $\widetilde{\omega}_E$ (top right), ω_K (lower left), $\widetilde{\omega}_K$ (lower right) and ω_z (middle) for $z \in \mathcal{N}$, $E \in \mathcal{E}$ and $K \in \mathcal{T}$, i.e., all elements, which have in case of ω_E or ω_K at least one edge and in case of $\widetilde{\omega}_E$ and $\widetilde{\omega}_K$ at least one vertex in common; in case of ω_z all elements which share z.

2.7 Remark [On the test functions]**.** The formulation of the weak problem (6), (7) suggests that one tests with all elements of V^h. However it suffices to test with the basis functions, i.e.,

$$a^h(\varphi^h - \varphi^h_{\Gamma_{\text{ess}}}, \psi^h) = \langle \tilde{f}, \psi^h \rangle \; \forall \psi^h \in V^h \Leftrightarrow a(\varphi^h - \varphi^h_{\Gamma_{\text{ess}}}, \psi^h_i) = \langle \tilde{f}, \psi^h_i \rangle \; \forall i \in \{1, 2, \ldots, N\}.$$

Indeed, one can represent every $\psi^h \in V^h$ as linear combination of the basis functions

$$\psi^h = \sum_{i=1}^{N} \alpha_i \psi^h_i$$

for some $\alpha_i \in \mathbb{R}$ and insert this into the problem's equation. Applying that $a^h(\cdot, \cdot)$ and the dual pairing are bilinear yields

$$a^h(\varphi^h - \varphi^h_{\Gamma_{\text{ess}}}, \psi^h) = \sum_{i=1}^{N} \alpha_i a^h(\varphi^h - \varphi^h_{\Gamma_{\text{ess}}}, \psi^h_i) = \sum_{i=1}^{N} \alpha_i \langle \tilde{f}, \psi^h_i \rangle = \langle \tilde{f}, \psi^h \rangle.$$

This equation holds for all basis functions. On the other hand it also holds for all functions from V^h, so for the basis functions in particular. Altogether it is equivalent if one tests with all functions from V^h or just with its basis functions.

Since one now is in the setting of a finite-dimensional space V^h and a bilinear operator $a^h(\cdot, \cdot)$, one can express the discretized problem (11) in terms of a system of linear equations. To this end, let

$$\varphi^h = \sum_{i=1}^{N} \alpha_i \psi_i$$

be the sought solution. Inserting it into (11) yields

$$a^h(\varphi^h, \psi_j) = \sum_{i=1}^{N} \alpha_i a^h(\psi_i, \psi_j) = \langle \tilde{f}, \psi_j \rangle + a^h(\varphi^h_{\Gamma_{\text{ess}}}, \psi_j)$$

for all basis functions ψ_j of V^h, $j = 1, 2, \ldots, N$, which is equivalent to the matrix-vector form equation

$$A\underline{\boldsymbol{\phi}} = \mathbf{f},$$

where A is a $N \times N$ matrix with

$$(A)_{ij} = a^h(\psi_j, \psi_i),$$

$\mathbf{f} = (\langle \tilde{f}, \psi_1 \rangle + a^h(\varphi^h_{\Gamma_{\text{ess}}}, \psi_1), \ldots, \langle \tilde{f}, \psi_N \rangle + a^h(\varphi^h_{\Gamma_{\text{ess}}}, \psi_N))^T$ is a vector containing the right-hand side entries and $\underline{\boldsymbol{\phi}}$ is the solution vector containing the weights α_i for each row i.

2.2 The Stokes equations

The Stokes equations model flows of fluids with high viscosity and can be derived from the Navier–Stokes equations. In direct comparison, they are identical except for the additional convective term in the Navier–Stokes equations. Here they are not discussed in their full generality but in the incompressible and stationary case. They read in the dimensionless Cauchy stress form as follows:

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded domain with Lipschitz continuous boundary $\partial\Omega$, its outer unit normal $\mathbf{n}$ and a source term $\mathbf{f}$. Then find $(\mathbf{u}, p) : \Omega \to \mathbb{R}^d \times \mathbb{R}$, such that

$$\begin{cases} -\nabla \cdot \mathbb{T}(\mathbf{u}, p) = \mathbf{f} & \text{in } \Omega, \\ \nabla \cdot \mathbf{u} = 0 & \text{in } \Omega, \end{cases} \tag{12}$$

where $\mathbf{u}$ and p denote the fluid's velocity and pressure, respectively. The map

$$\mathbb{T}(\mathbf{u}, p) := 2\nu\, \mathbb{D}(\mathbf{u}) - p\,\mathbb{I}$$

is known as the Cauchy stress tensor giving information about the stress inside the fluid, $\nu := \mathrm{Re}^{-1}$ the kinematic viscosity and

$$\mathbb{D}(\mathbf{u}) := \frac{1}{2}(\nabla\mathbf{u} + \nabla\mathbf{u}^T)$$

the so-called deformation tensor, which is the symmetric part of the gradient $\nabla\mathbf{u}$. Similarly to the Darcy equations, in the smooth case an equivalent formulation can be found, which however yields a different discretization and therefore other results in simulations. This alternative form is known as the Laplace form and is given by (12) with

$$\mathbb{T}(\mathbf{u}, p) := \nu\nabla\mathbf{u} - p\,\mathbb{I}\,. \tag{13}$$

2.8 Remark [On the equivalence of the Cauchy stress and the Laplacian form]**.** The equivalence holds if $\nabla \cdot \mathbf{u} = 0$ is given. In a discrete setting however, this condition usually is only fulfilled approximatively by $\mathbf{u}^h$.

The problem can be completed by setting boundary conditions. Here, the focus is on Dirichlet and Neumann-type boundary conditions. Similar as for the Darcy equations, split $\partial\Omega$ into two relatively open disjoint sets Γ_{ess} denoting the essential and Γ_{nat} denoting the natural boundary conditions which correspond to Dirichlet and Neumann boundary conditions respectively, such that

$$\overline{\Gamma}_{\text{ess}} \cup \overline{\Gamma}_{\text{nat}} = \partial\Omega \text{ and } \Gamma_{\text{ess}} \cap \Gamma_{\text{nat}} = \emptyset.$$

Further, the essential boundary conditions are assumed to be homogeneous. The conditions are then given by

$$\begin{cases} \mathbf{u} = \mathbf{0} & \text{on } \Gamma_{\text{ess}}, \\ \mathbb{T}(\mathbf{u}, p) \cdot \mathbf{n} = T_N & \text{on } \Gamma_{\text{nat}}, \end{cases} \tag{14}$$

where $T_N \in H^{-1/2}(\Gamma_{\text{nat}})$ and $\mathbb{T}$ is defined as in (12) or (13).

2.9 Remark [Homogeneous essential boundary conditions]**.** If the essential boundary conditions are not homogeneous, the problem can still be transformed into an equivalent problem with homogeneous essential boundary conditions by searching for $\tilde{\mathbf{u}} := \mathbf{u} - \mathbf{u}_{\text{ess}}$ instead of $\mathbf{u}$. Here, $\mathbf{u}_{\text{ess}}$ denotes an extension of the essential boundary conditions into the interior of Ω. The existence of such an extension is provided by the surjectivity of the trace operator.

2.2.1 Weak formulation

Let $\mathbf{u} \in (C^2(\Omega) \cap C^1(\overline{\Omega}))^d$ and $p \in C^1(\Omega) \cap C(\overline{\Omega})$ be classical solutions of (12) or (13). Let $\mathbf{v}$ and q be corresponding test functions for velocity and pressure, where

$$(C^\infty_{\Gamma_{\text{ess}}}(\Omega))^d \text{ with } C^\infty_{\Gamma_{\text{ess}}}(\Omega) := \left\{ v \in C^\infty(\Omega) \cap H^1(\Omega) : \substack{\exists U \subset \mathbb{R}^d \text{ open neighborhood of } \Gamma_{\text{ess}} \\ \text{s.t. } v(x)=0 \ \forall x \in U \cap \Omega} \right\}$$

and $C^\infty(\Omega)$ are first candidates for the respective spaces. In the test space for $\mathbf{v}$ is not enough structure to make Hilbert or Banach space theory applicable, thus its completion with respect to the $H^1(\Omega)$-norm is being considered. This also explains the intersection with $H^1(\Omega)$ in the definition, as $C^\infty(\Omega) \not\subset H^1(\Omega)$, but $C^\infty(\Omega) \cap H^1(\Omega)$ is dense in $H^1(\Omega)$, see, e.g., [MS64]. For the same reason, $C^\infty(\Omega)$ is being completed with respect to the $L^2(\Omega)$-norm, yielding as test spaces

$$\mathbf{v} \in \mathbf{V} := \left(\overline{C^\infty_{\Gamma_{\text{ess}}}(\Omega)}^{H^1(\Omega)} \right)^d \text{ and } q \in Q := \overline{C^\infty(\Omega)}^{L^2(\Omega)}. \tag{15}$$

2.10 Theorem [On the pressure test space]**.** In fact, one obtains a pressure space

$$Q = L^2(\Omega).$$

Proof. The space $L^2(\Omega)$ is usually defined as the set of measurable functions f on Ω satisfying

$$\int\limits_\Omega |f|^2 < \infty,$$

under the equivalence relation $\sim$ defined by $f \sim g :\Leftrightarrow f = g$ almost everywhere together with the $L^2(\Omega)$ scalar product. In Lemma 4.2.1 and Corollary 4.2.2 of [Bog07] it is shown that $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$, in particular this holds for $C^\infty(\Omega)$. Thus by definition, the completion of $C^\infty(\Omega)$ with respect to the $L^2(\Omega)$-norm yields the $L^2(\Omega)$ Hilbert space. □

Multiplication of (12) with the test functions, integration, and integration by parts results in equations of the form

$$\left\{\begin{aligned} (2\nu\,\mathbb{D}(\mathbf{u}),\mathbb{D}(\mathbf{v}))_0 - (\nabla\cdot\mathbf{v},p)_0 &= \langle\tilde{\mathbf{f}},\mathbf{v}\rangle \text{ , for (12),}\\ (\nu\nabla\mathbf{u},\nabla\mathbf{v})_0 - (\nabla\cdot\mathbf{v},p)_0 &= \langle\tilde{\mathbf{f}},\mathbf{v}\rangle \text{ , for (13),}\\ -(\nabla\cdot\mathbf{u},q)_0 &= 0, \end{aligned}\right. \tag{16}$$

where $\langle\tilde{\mathbf{f}},\mathbf{v}\rangle := (\mathbf{f},\mathbf{v})_0 + \langle T_N,\mathbf{v}\rangle_{\Gamma_{\text{nat}}}$ is the source term including the natural boundary conditions. These equations can be expressed in terms of bilinear forms $a : V \times V \to \mathbb{R}$ and $b : V \times Q \to \mathbb{R}$ with

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) &:= (2\nu\,\mathbb{D}(\mathbf{u}),\mathbb{D}(\mathbf{v}))_0 \text{ , for (12),}\\ a(\mathbf{u},\mathbf{v}) &:= (\nu\nabla\mathbf{u},\nabla\mathbf{v})_0 \qquad \text{ , for (13),}\\ b(\mathbf{v},p) &:= -(\nabla\cdot\mathbf{v},p)_0, \end{aligned}$$

satisfying

$$\left\{\begin{aligned} a(\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) &= \langle\tilde{\mathbf{f}},\mathbf{v}\rangle,\\ b(\mathbf{u},q) &= 0. \end{aligned}\right. \tag{17}$$

The analysis can be simplified by applying the $H^1(\Omega)$-semi-norm instead of the $H^1(\Omega)$-norm for the velocity space. The semi-norm is in fact a norm on $\mathbf{V}$ due to the Poincaré-Friedrichs inequality (5) and $\mathbf{V} \subset (H^1(\Omega))^d \subset (L^2(\Omega))^d$:

$$0 = |\mathbf{u}|_1 = \|\mathbf{u}\|_0 \geq \frac{1}{C}\|\mathbf{u}\|_0 \Rightarrow \mathbf{u} = \mathbf{0}.$$

2.11 Remark [On the pressure space]**.** If $\Gamma_{\text{nat}} = \emptyset$, there are no boundary conditions for p. This has the consequence that p is only fixed up to a constant: Let $c \in \mathbb{R}\setminus\{0\}$ be a scalar value and $(\mathbf{u},p)$ be a solution of the Stokes problem without natural boundary conditions, then $(\mathbf{u},p+c)$ is also a solution, as

$$b(\mathbf{v},c) = c\int\limits_{\Omega}\mathbf{v} = c\int\limits_{\Gamma_{\text{ess}}}\mathbf{v}\cdot\mathbf{n} = 0 \Rightarrow b(\mathbf{v},p+c) = b(\mathbf{v},p) + b(\mathbf{v},c) = b(\mathbf{v},p).$$

A possible remedy for this issue is changing the pressure space into

$$Q = L_0^2(\Omega) := \left\{ f \in L^2(\Omega) : \int\limits_{\Omega} f = 0 \right\}.$$

In practice one can use the original pressure space and simply fix p on a single node and then modify the approximated solution by $p - \overline{p}_\Omega$, where $\overline{p}_\Omega$ denotes the space average of p.

2.12 Lemma [Estimate divergence by gradient]**.** For all $\mathbf{v} \in \mathbf{V}$ it is $\|\nabla\cdot\mathbf{v}\|_0 \leq \sqrt{d}\|\nabla\mathbf{v}\|_0$.

Proof. Let $\mathbf{v} = (v_1,\ldots,v_d) \in \mathbf{V}$ and denote

$$\mathbf{d} := \left(\frac{\partial v_1}{\partial x_1},\ldots,\frac{\partial v_d}{\partial x_d}\right)^T.$$

Then one can estimate

$$\begin{aligned}\|\nabla\cdot\mathbf{v}\|_0^2 &= \int\limits_\Omega \left(\sum_{i=1}^d \frac{\partial v_i}{\partial x_i}\right)^2 = \int\limits_\Omega |(\mathbf{1},\mathbf{d})_{\ell^2}|^2 \leq \int\limits_\Omega \|\mathbf{1}\|_{\ell^2}^2\|\mathbf{d}\|_{\ell^2}^2 &&\text{, using Cauchy–Schwarz,}\\ &= d\int\limits_\Omega \sum_{i=1}^d \left(\frac{\partial v_i}{\partial x_i}\right)^2 \leq d\int\limits_\Omega \sum_{i,j=1}^d \left(\frac{\partial v_i}{\partial x_j}\right)^2 \\ &= d\|\nabla\mathbf{v}\|_0^2.\end{aligned}$$

□

2.13 Theorem [Existence and uniqueness of a weak solution]**.** There exists exactly one solution of the Stokes problem in its weak formulation (17).

Proof. This problem can be applied to the framework of saddle-point problems introduced in Chapter I of [GR81]. According to Theorem 4.1 of § 4 of Chapter I thereof, this type of problem has a unique solution if $a(\cdot,\cdot)$ is continuous and coercive on the space of weakly divergence free functions

$$\mathbf{V}_{\text{div}} := \{\mathbf{v}\in\mathbf{V} : (\nabla\cdot\mathbf{v},q)_0 = 0\ \forall q\in Q\}$$

and that $b(\cdot,\cdot)$ is continuous and satisfies the so-called inf-sup condition. The inf-sup condition is fulfilled if there is a constant $\beta_{\text{is}} > 0$ such that

$$\inf_{\substack{q\in Q\\ q\neq 0}} \sup_{\substack{\mathbf{v}\in\mathbf{V}\\ \mathbf{v}\neq\mathbf{0}}} \frac{b(\mathbf{v},q)}{\|\mathbf{v}\|_{\mathbf{V}}\|q\|_Q} \geq \beta_{\text{is}}. \tag{18}$$

The continuity and coercivity of $a(\cdot,\cdot)$ has to be shown separately for the Cauchy stress form and the Laplacian form.

(i) Continuity of $a(\cdot,\cdot)$:

- For (12) it is

$$\begin{aligned}|a(\mathbf{v},\mathbf{w})| &= |(2\nu\,\mathbb{D}(\mathbf{v}),\mathbb{D}(\mathbf{w}))| \leq 2\nu\|\,\mathbb{D}(\mathbf{v})\|_0\|\,\mathbb{D}(\mathbf{w})\|_0 &&\text{, using Cauchy–Schwarz,}\\ &\leq 2\nu\left(\frac{\|\nabla\mathbf{v}\|_0+\|\nabla\mathbf{v}^T\|_0}{2}\cdot\frac{\|\nabla\mathbf{w}\|_0+\|\nabla\mathbf{w}^T\|_0}{2}\right) &&\text{, using triangle inequality,}\\ &= 2\nu\|\mathbf{v}\|_{\mathbf{V}}\|\mathbf{w}\|_{\mathbf{V}} &&\text{, using } \|\nabla\mathbf{v}\|_0 = \|\nabla\mathbf{v}^T\|_0.\end{aligned}$$

- For (13) it is

$$|a(\mathbf{v},\mathbf{w})| = |(\nu\nabla\mathbf{v},\nabla\mathbf{w})_0| \leq \nu\|\mathbf{v}\|_{\mathbf{V}}\|\mathbf{w}\|_{\mathbf{V}}, \text{ using Cauchy–Schwarz.}$$

(ii) Coercivity of $a(\cdot,\cdot)$:

- For (12) coercivity can be shown by using Korn's first inequality. It states that if $\Omega\subset\mathbb{R}^d$ is a bounded domain with Lipschitz boundary, $\Gamma\subset\partial\Omega$ open with $\operatorname{meas}(\Gamma) > 0$, then there is a constant $\kappa > 0$ such that

$$\|\,\mathbb{D}(\mathbf{v})\|_0^2 \geq \kappa\|\nabla\mathbf{v}\|_0^2$$

for all $\mathbf{v} \in H^1_\Gamma(\Omega)$, see, e.g., [Nef02].

With this inequality one obtains

$$(2\nu\,\mathbb{D}(\mathbf{u}), \mathbb{D}(\mathbf{u}))_0 = 2\nu\|\,\mathbb{D}(\mathbf{u})\|^2 \geq 2\nu\kappa\|\mathbf{u}\|^2_{\mathbf{V}}$$

and therefore coercivity of $a(\cdot, \cdot)$.

- For (13) it is $a(\mathbf{v}, \mathbf{v}) = (\nu\nabla\mathbf{v}, \nabla\mathbf{v})_0 = \nu\|\mathbf{v}\|^2_{\mathbf{V}}$, thus coercive.

(iii) Continuity of $b(\cdot, \cdot)$: One has for all $\mathbf{v} \in \mathbf{V}$, $q \in Q$ that

$$\begin{aligned} |b(\mathbf{v}, q)| = |(\nabla \cdot \mathbf{v}, q)_0| &\leq \|\nabla \cdot \mathbf{v}\|_0\|q\|_Q && \text{, using Cauchy–Schwarz,} \\ &\leq \sqrt{d}\|\nabla\mathbf{v}\|_0\|q\|_Q && \text{, using Lemma 2.12,} \\ &= \sqrt{d}\|\mathbf{v}\|_{\mathbf{V}}\|q\|_Q. \end{aligned}$$

(iv) Inf-sup condition: Let $q \in Q$ be arbitrary. Due to [GR81, Lemma 3.2], there is a unique $\mathbf{v} \in \mathbf{V}^\perp_{\text{div}}$ such that

$$q = \nabla \cdot \mathbf{v} \text{ and } \|\mathbf{v}\|_{\mathbf{V}} \leq C\|q\|_Q,$$

where the space $\mathbf{V}^\perp_{\text{div}}$ is the orthogonal complement of $\mathbf{V}_{\text{div}}$ with respect to the scalar product of $\mathbf{V}$

$$(\mathbf{v}, \mathbf{w})_{\mathbf{V}} = \int\limits_\Omega (\nabla\mathbf{v}, \nabla\mathbf{w}).$$

It follows with the above result, that

$$\frac{b(\mathbf{v}, q)}{\|\mathbf{v}\|_{\mathbf{V}}} = \frac{(\nabla \cdot \mathbf{v}, q)_0}{\|\mathbf{v}\|_{\mathbf{V}}} = \frac{(q, q)_0}{\|\mathbf{v}\|_{\mathbf{V}}} = \frac{\|q\|^2_Q}{\|\mathbf{v}\|_{\mathbf{V}}} \geq \frac{1}{C}\|q\|_Q.$$

Since the right-hand side does not depend on $\mathbf{v}$ anymore and $b(\tilde{\mathbf{v}}, q) = 0$ for all $\tilde{\mathbf{v}} \in \mathbf{V} \setminus \mathbf{V}^\perp_{\text{div}} = \mathbf{V}_{\text{div}}$, it is

$$\sup_{\substack{\mathbf{v}\in\mathbf{V} \\ \mathbf{v}\neq\mathbf{0}}} \frac{b(\mathbf{v}, q)}{\|\mathbf{v}\|_{\mathbf{V}}} \geq \frac{1}{C}\|q\|_Q.$$

Because q was arbitrary, one obtains

$$\inf_{\substack{q\in Q \\ q\neq 0}} \sup_{\substack{\mathbf{v}\in\mathbf{V} \\ \mathbf{v}\neq\mathbf{0}}} \frac{b(\mathbf{v}, q)}{\|\mathbf{v}\|_{\mathbf{V}}\|q\|_Q} \geq \frac{1}{C} =: \beta_{\text{is}} > 0.$$

□

2.2.2 Finite element discretization

Similarly to the discretization of the Darcy equations, one considers finite-dimensional subspaces $\mathbf{V}^h \subset \mathbf{V}$ and $Q^h \subset Q$ denoted by $\mathbf{V}^h/Q^h$. With regard to the existence and uniqueness of a solution in the discrete setting, these spaces should be connected by a discrete variant of the inf-sup condition (18), i.e., there is a constant $\beta^h_{\text{is}} > 0$ such that

$$\inf_{\substack{q^h\in Q^h \\ q^h\neq 0}} \sup_{\substack{\mathbf{v}^h\in\mathbf{V}^h \\ \mathbf{v}^h\neq\mathbf{0}}} \frac{b^h(\mathbf{v}^h, q^h)}{\|\mathbf{v}^h\|_{\mathbf{V}^h}\|q^h\|_{Q^h}} \geq \beta^h_{\text{is}}, \tag{19}$$

where $b^h : \mathbf{V}^h \times Q^h \to \mathbb{R}$ is the discrete variant of $b(\cdot,\cdot)$. Since it is assumed here that $\mathbf{V}^h$ and Q^h are actual subspaces of $\mathbf{V}$ and Q, i.e., conforming finite element discretizations are considered, it is

$$a(\mathbf{u}^h, \mathbf{v}^h) = a^h(\mathbf{u}^h, \mathbf{v}^h) \text{ and } b(\mathbf{u}^h, q^h) = b^h(\mathbf{u}^h, q^h) \quad \forall \mathbf{u}^h, \mathbf{v}^h \in \mathbf{V}^h, q^h \in Q^h,$$

where $a^h : \mathbf{V}^h \times \mathbf{V}^h \to \mathbb{R}$ denotes the discrete variant of $a(\cdot,\cdot)$. The discretization of the weak problem (17) then reads: Find $(\mathbf{u}^h, p^h) \in \mathbf{V}^h \times Q^h$ such that

$$\begin{cases} a^h(\mathbf{u}^h, \mathbf{v}^h) + b^h(\mathbf{v}^h, p^h) = \langle \widetilde{\mathbf{f}}, \mathbf{v}^h \rangle, \\ \qquad\qquad\qquad b^h(\mathbf{u}^h, q^h) = 0, \end{cases} \tag{20}$$

for all $\mathbf{v}^h \in \mathbf{V}^h$ and $q^h \in Q^h$.

In order to obtain a matrix-vector form of the problem, let

$$(\mathbf{u}^h, p^h) = \left(\sum_{i=1}^{dN_u} u_i^h \mathbf{w}_i, \ \sum_{i=1}^{N_p} p_i^h q_i \right)$$

be the sought solution with $\{\mathbf{w}_i\}_{i=1}^{dN_u}$ and $\{q_i\}_{i=1}^{N_p}$ being the respective bases for $\mathbf{V}^h$ and Q^h as well as $\underline{\mathbf{u}} = (u_i^h)_{i=1}^{dN_u}$ and $\underline{\mathbf{p}} = (p_i^h)_{i=1}^{N_p}$ being the unknown coefficients that need to be determined. Inserting this expression into (20) yields

$$\begin{cases} \sum_{i=1}^{dN_u} u_i^h a^h(\mathbf{w}_i, \mathbf{w}_j) + \sum_{i=1}^{N_p} p_i^h b^h(\mathbf{w}_j, q_i) = \langle \widetilde{\mathbf{f}}, \mathbf{w}_j \rangle, \\ \qquad\qquad\qquad\qquad \sum_{i=1}^{dN_u} u_i^h b^h(\mathbf{w}_i, q_k) = 0, \end{cases}$$

for all $j \in \{1, 2, \ldots, dN_u\}$ and $k \in \{1, 2, \ldots, N_p\}$, which is equivalent to the block-matrix-vector form

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{\mathbf{u}} \\ \underline{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \underline{\mathbf{f}} \\ \underline{\mathbf{0}} \end{pmatrix}$$

with

$$\begin{aligned} (A)_{ij} &= a^h(\mathbf{w}_j, \mathbf{w}_i) && , \forall i, j \in \{1, 2, \ldots, dN_u\}, \\ (B)_{ij} &= b^h(\mathbf{w}_j, q_i) && , \forall i \in \{1, 2, \ldots, N_p\} \forall j \in \{1, 2, \ldots, dN_u\}, \\ (\underline{\mathbf{f}})_i &= \langle \widetilde{\mathbf{f}}, \mathbf{w}_i \rangle && , \forall i \in \{1, 2, \ldots, dN_u\}. \end{aligned}$$

2.14 Remark [On the structure of the block A]**.** When it comes to the implementation and therefore memory requirements, one can investigate if there is some inherent structure that allows to reduce these requirements. A possible approach to construct finite element functions for the vector-valued space $\mathbf{V}^h$ is to choose its basis like

$$\mathbf{V}^h = \operatorname{span}\{\mathbf{w}_i : i = 1, \ldots, dN_u\} = \operatorname{span}\{w_i \mathbf{e}_j : j = 1, \ldots, d, \ i = 1, \ldots, N_u\},$$

see, e.g., Chapter 1, § 4 of [Tem01]. Here, N_u denotes the number of unknowns for one component of the velocity space. By this specific way of constructing the basis functions, each of them does not vanish in one component only. In the following, assume that $d = 3$, the case $d = 2$ is analogous. Let $\mathbf{w}_i = w_i\mathbf{e}_k$ and $\mathbf{w}_j = w_j\mathbf{e}_l$ be basis functions. In the Cauchy stress case (12), it is

$$\begin{aligned} A_{ij} &= a^h(\mathbf{w}_j, \mathbf{w}_i) = (2\nu\,\mathbb{D}(\mathbf{w}_j), \mathbb{D}(\mathbf{w}_i))_0 = 2\left(\frac{\nabla\mathbf{w}_j + \nabla\mathbf{w}_j^T}{2}, \frac{\nabla\mathbf{w}_i + \nabla\mathbf{w}_i^T}{2}\right)_0 \\ &= \frac{1}{2}\big((\nabla\mathbf{w}_j, \nabla\mathbf{w}_i)_0 + (\nabla\mathbf{w}_j, \nabla\mathbf{w}_i^T)_0 + (\nabla\mathbf{w}_j^T, \nabla\mathbf{w}_i)_0 + (\nabla\mathbf{w}_j^T, \nabla\mathbf{w}_i^T)_0\big) \\ &= (\nabla\mathbf{w}_j, \nabla\mathbf{w}_i)_0 + (\nabla\mathbf{w}_j^T, \nabla\mathbf{w}_i)_0. \end{aligned}$$

The first term is the term of the Laplace case (13) and vanishes if $k \neq l$, for example

$$(\nabla(w_i\mathbf{e}_1), \nabla(w_j\mathbf{e}_2))_0 = \int_\Omega \begin{pmatrix} \partial_x w_i & \partial_y w_i & \partial_z w_i \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} : \begin{pmatrix} 0 & 0 & 0 \\ \partial_x w_j & \partial_y w_j & \partial_z w_j \\ 0 & 0 & 0 \end{pmatrix} d\mathbf{x} = 0.$$

Further, if $k = l$, the result of the first term's scalar product is independent of the chosen component, yielding

$$A = \begin{pmatrix} A_{11} & & \\ & A_{11} & \\ & & A_{11} \end{pmatrix} + \widetilde{A},$$

where $\widetilde{A} = 0$ in the Laplace case. In the Cauchy stress case the terms of $\widetilde{A}$ do not vanish anymore if $k \neq l$ but it holds that $\widetilde{A}_{kl} = \widetilde{A}_{lk}^T$, since

$$(\nabla(w_j\mathbf{e}_k)^T, \nabla(w_i\mathbf{e}_l))_0 = (\partial_{x_k} w_j, \partial_{x_l} w_i)_0 = (\nabla(w_i\mathbf{e}_l)^T, w_j\mathbf{e}_k)_0.$$

This gives

$$A = \begin{pmatrix} A_{11} + \widetilde{A}_{11} & \widetilde{A}_{12} & \widetilde{A}_{13} \\ \widetilde{A}_{12}^T & A_{22} + \widetilde{A}_{22} & \widetilde{A}_{23} \\ \widetilde{A}_{13}^T & \widetilde{A}_{23}^T & A_{33} + \widetilde{A}_{33} \end{pmatrix}$$

and therefore one needs to store six matrix blocks instead of just one in comparison to the Laplace case.

2.3 The Stokes–Darcy equations

This section deals with the coupled Stokes–Darcy system, i.e., free flow of high viscosity coupled with flow through porous media. The coupling is realized by splitting the domain Ω into two parts Ω_f and Ω_p for the Stokes and Darcy system respectively, such that

$$\overline{\Omega} = \overline{\Omega}_f \cup \overline{\Omega}_p, \quad \Omega_f \cap \Omega_p = \emptyset, \quad \overline{\Omega}_f \cap \overline{\Omega}_p = \Gamma,$$

where Γ is the so-called interface between Ω_f and Ω_p, being responsible for the information exchange between the two systems. Such a domain could look like the illustration in Figure 3. One obtains

$$\begin{cases} -\nabla\cdot\mathbb{T}(\mathbf{u}_f, p_f) = \mathbf{f}_f \ , \text{ in } \Omega_f, \\ \nabla\cdot\mathbf{u}_f = 0 \ , \text{ in } \Omega_f, \\ -\nabla\cdot\mathbb{K}\,\nabla\varphi_p = \tilde{f}_p \ , \text{ in } \Omega_p, \end{cases} \tag{21}$$

where $\mathbf{u}_f : \Omega_f \to \mathbb{R}^d$ denotes the Stokes velocity, $p_f : \Omega_f \to \mathbb{R}$ the Stokes pressure and $\varphi_p : \Omega_p \to \mathbb{R}$ the piezometric head or Darcy pressure. To complete the problem one has

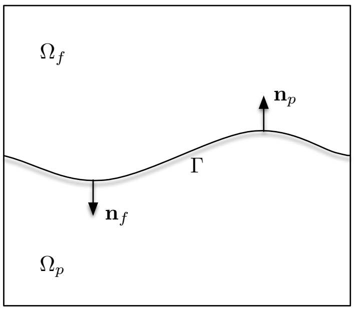


Figure 3: A sketch of the domain used in the Stokes–Darcy problem.

to assign boundary conditions. To this end, split $\partial\Omega$ into relatively open, disjoint parts $\Gamma_{f,n}, \Gamma_{f,e} \subset \partial\Omega_f \setminus \Gamma$ and $\Gamma_{p,n}, \Gamma_{p,e} \subset \partial\Omega_p \setminus \Gamma$ denoting natural and essential parts of the boundary of the Stokes and Darcy domains respectively, such that

$$\begin{aligned} \overline{\Gamma}_{f,n} \cup \overline{\Gamma}_{f,e} &= \partial\Omega \setminus \Gamma, \quad \Gamma_{f,n} \cap \Gamma_{f,e} = \emptyset, \\ \overline{\Gamma}_{p,n} \cup \overline{\Gamma}_{p,e} &= \partial\Omega \setminus \Gamma, \quad \Gamma_{p,n} \cap \Gamma_{p,e} = \emptyset. \end{aligned}$$

On these parts one can now impose the conditions

$$\begin{cases} \mathbf{u}_f = \mathbf{u}_{f,\mathrm{ess}} & , \text{ on } \Gamma_{f,e}, \\ \mathbb{T}(\mathbf{u}_f, p_f) \cdot \mathbf{n} = T_{f,\mathrm{nat}} & , \text{ on } \Gamma_{f,n}, \\ \varphi_p = \varphi_{p,\mathrm{ess}} & , \text{ on } \Gamma_{p,e}, \\ (-\,\mathbb{K}\,\nabla\varphi) \cdot \mathbf{n} = u_{p,\mathrm{nat}} & , \text{ on } \Gamma_{p,n}. \end{cases} \tag{22}$$

To obtain a well-posed problem and actual information exchange between the two systems, further conditions on the interface are imposed. To this end, let $\mathbf{n}_f$ be the outer unit normal vector of Ω_f and $\mathbf{n}_p$ be the outer unit normal vector of Ω_p with $\mathbf{n}_f = -\mathbf{n}_p$ on Γ. Then,

- to obtain a continuous normal velocity, impose

$$\mathbf{u}_f \cdot \mathbf{n}_f = \mathbf{u}_p \cdot \mathbf{n}_f = -(\mathbb{K}\,\nabla\varphi_p) \cdot \mathbf{n}_f, \tag{23}$$

- to preserve normal stress, impose

$$-\mathbf{n}_f \cdot \mathbb{T}(\mathbf{u}_f, p_f) \cdot \mathbf{n}_f = g\varphi_p, \tag{24}$$

where g denotes the gravitational acceleration,

- and to exchange information about the tangential velocity, impose the so-called Beavers–Joseph–Saffman condition. For this last condition, let $\boldsymbol{\tau}_i$, $i \in \{1, \ldots, d-1\}$, be pairwise orthogonal tangential vectors on Γ and let

$$\alpha_i = \alpha_{\mathrm{BJ}} \sqrt{\boldsymbol{\tau}_i^T\,\mathbb{K}\,\boldsymbol{\tau}_i},$$

be a constant where α_{BJ} is dimensionless and depends only on the structure of the porous medium. It then reads

$$\mathbf{u}_f \cdot \boldsymbol{\tau}_i + \alpha_i \boldsymbol{\tau}_i \cdot \mathbb{T}(\mathbf{u}_f, p_f) \cdot \mathbf{n}_f = 0, \tag{25}$$

see, e.g., [GKR13, Section 2.1].

2.15 Remark [On the interface conditions]**.** The Beavers–Joseph–Saffman condition is strictly speaking no coupling condition, as it does not relate quantities from Ω_f and Ω_p, but rather a boundary condition on Γ for the Stokes problem. It was developed out of the Beavers–Joseph condition which included the Darcy velocity $\mathbf{u}_p$ but it was found out that it can be neglected compared to the other quantities in the condition, see, e.g., [DQ09, Section 3].

Further, in the first condition the Stokes and Darcy velocities are coupled directly in normal direction, even though the Darcy velocity is an averaged quantity whereas the Stokes velocity is of a pointwise nature.

2.3.1 Weak formulation

To derive a weak formulation, consider test functions from spaces similar to the ones used in the Stokes and Darcy problems, namely

$$\mathbf{V}_f = \left\{\mathbf{v} \in \left(H^1(\Omega_f)\right)^d : \mathbf{v}\big|_{\Gamma_{f,e}} = \mathbf{0}\right\},\ Q_f = L^2(\Omega_f),\ V_p = \left\{v \in H^1(\Omega_p) : v\big|_{\Gamma_{p,e}} = 0\right\},$$

corresponding to Stokes velocity, Stokes pressure and Darcy pressure, respectively.

Multiplication of the equations of (21) with respective test functions $\mathbf{v} \in \mathbf{V}_f$, $q \in Q_f$, $\psi \in V_p$, integration and integration by parts yields the weak formulation of the Darcy equations (6) and of the Stokes equations (17) with different right-hand sides. Decomposition of the Stokes velocity test function $\mathbf{v}$ into its normal and tangential components

$$\mathbf{v} = (\mathbf{v}\cdot\mathbf{n}_f)\cdot\mathbf{n}_f + \sum_{i=1}^{d-1}(\mathbf{v}\cdot\boldsymbol{\tau}_i)\cdot\boldsymbol{\tau}_i$$

allows substitution of the Beavers–Joseph–Saffman condition (25), yielding, after rearranging terms, a system of the form

$$\begin{cases} a_f(\mathbf{u}_f,\mathbf{v}) + b_f(\mathbf{v},p_f) - \langle \mathbf{n}_f\cdot\mathcal{T}(\mathbf{u}_f,p_f)\cdot\mathbf{n}_f, \mathbf{v}\cdot\mathbf{n}_f\rangle_\Gamma = \langle \mathbf{f}_f^1,\mathbf{v}\rangle, \\ b_f(\mathbf{u}_f,q) = \langle f_f^2, q\rangle, \\ a_p(\varphi_p,\psi) + \langle \mathbb{K}\,\nabla\varphi_p\cdot\mathbf{n}_f,\psi\rangle_\Gamma = \langle f_p,\psi\rangle, \end{cases} \tag{26}$$

with bilinear forms

$$\begin{aligned} &a_f : \mathbf{V}_f\times\mathbf{V}_f\to\mathbb{R},\ (\mathbf{u}_f,\mathbf{v})\mapsto \sum_{i=1}^{d-1}\frac{1}{\alpha_i}\langle \mathbf{u}_f\cdot\boldsymbol{\tau}_i,\mathbf{v}\cdot\boldsymbol{\tau}_i\rangle_\Gamma + \begin{cases}(2\nu\,\mathbb{D}(\mathbf{u}_f),\mathbb{D}(\mathbf{v}))_{0,\Omega_f} & \text{for (12)}\\ (\nu\nabla\mathbf{u}_f,\nabla\mathbf{v})_{0,\Omega_f} & \text{for (13)}\end{cases}, \\ &b_f : \mathbf{V}_f\times Q_f\to\mathbb{R},\ (\mathbf{v},p_f)\mapsto -(\nabla\cdot\mathbf{v},p_f)_{0,\Omega_f}, \\ &a_p : Q_p\times Q_p\to\mathbb{R},\ (\varphi_p,\psi)\mapsto(\mathbb{K}\,\nabla\varphi,\nabla\psi)_{0,\Omega_p}, \end{aligned}$$

and right-hand sides

$$\begin{aligned} \mathbf{f}_f^1 \in \mathbf{V}_f{}',\ \langle\mathbf{f}_f^1,\mathbf{v}\rangle &= (\mathbf{f}_f,\mathbf{v})_{0,\Omega_f} + \langle T_{\text{f,nat}},\mathbf{v}\rangle_{\Gamma_{f,n}}, \\ f_f^2 \in Q_f',\ \langle f_f^2,q\rangle &= (\nabla\cdot\mathbf{u}_{f,\text{ess}},q)_{0,\Omega_f}, \\ f_p \in V_p',\ \langle f_p,\psi\rangle &= (\mathbb{K}\,\nabla\varphi_{p,\text{ess}},\nabla\psi)_{0,\Omega_p} - \langle u_{p,\text{nat}},\psi\rangle_\Gamma. \end{aligned}$$

There is more than one way to include the remaining two interface conditions (23) and (24), in particular they can either be included such that they pose Neumann conditions or as a weighted linear combination, i.e., as Robin conditions.

(i) *The Neumann–Neumann coupling.* In the Neumann–Neumann coupling, the interface conditions are included as Neumann conditions into the systems, yielding the problem to find $(\mathbf{u}_f, p_f, \varphi_p) \in \mathbf{V}_f \times Q_f \times Q_p$ such that

$$\begin{cases} a_f(\mathbf{u}_f, \mathbf{v}) + b_f(\mathbf{v}, p_f) + \langle g\varphi_p, \mathbf{v} \cdot \mathbf{n}_f \rangle_\Gamma = \langle \mathbf{f}_f^1, \mathbf{v} \rangle, \\ b_f(\mathbf{u}_f, q) = \langle f_f^2, q \rangle, \\ a_p(\varphi_p, \psi) - \langle \mathbf{u}_f \cdot \mathbf{n}_f, \psi \rangle_\Gamma = \langle f_p, \psi \rangle, \end{cases} \tag{27}$$

is satisfied for all $(\mathbf{v}, q, \psi) \in \mathbf{V}_f \times Q_f \times Q_p$.

If $\text{meas}(\Gamma_{f,n}) = \emptyset$, the Stokes pressure is, as in Remark 2.11, only fixed up to a constant and one can change the pressure space to $Q_f = L_0^2(\Omega_f)$.

(ii) *The Robin–Robin coupling.* In the Robin–Robin coupling, a weighted linear combination out of the conditions is inserted into the systems. To this end, let $\gamma_f \geq 0$ and $\gamma_p > 0$ be weights, such that

$$\begin{aligned} \gamma_f \mathbf{u}_f \cdot \mathbf{n}_f + \mathbf{n}_f \cdot \mathbb{T}(\mathbf{u}_f, p_f) \cdot \mathbf{n}_f &= -\gamma_f(\mathbb{K}\,\nabla\varphi_p) \cdot \mathbf{n}_f - g\varphi_p \qquad &&\text{, on } \Gamma, \\ -\gamma_p \mathbf{u}_f \cdot \mathbf{n}_f + \mathbf{n}_f \cdot \mathbb{T}(\mathbf{u}_f, p_f) \cdot \mathbf{n}_f &= \gamma_p(\mathbb{K}\,\nabla\varphi_p) \cdot \mathbf{n}_f - g\varphi_p \qquad &&\text{, on } \Gamma. \end{aligned}$$

These equations can now be inserted as Robin boundary conditions into the weak Stokes equations and weak Darcy equations, respectively. The problem then is to find $(\mathbf{u}_f, p_f, \varphi_p) \in \mathbf{V}_f \times Q_f \times Q_p$ such that

$$\begin{cases} a_f(\mathbf{u}_f, \mathbf{v}) + \langle \gamma_f \mathbf{u}_f \cdot \mathbf{n}_f, \mathbf{v} \cdot \mathbf{n}_f \rangle_\Gamma + b(\mathbf{v}, p_f) + \langle \eta_f, \mathbf{v} \cdot \mathbf{n}_f \rangle_\Gamma = \langle \mathbf{f}_f^1, \mathbf{v} \rangle, \\ b_f(\mathbf{u}_f, q) = \langle f_f^2, q \rangle, \\ a_p(\varphi_p, \psi) + \langle \gamma_p^{-1} \varphi_p, \psi \rangle_\Gamma + \langle \eta_p, \psi \rangle_\Gamma = \langle f_p, \psi \rangle, \end{cases} \tag{28}$$

for all $(\mathbf{v}, q, \psi) \in \mathbf{V}_f \times Q_f \times Q_p$, where

$$\eta_f := -\gamma_f(\mathbb{K}\,\nabla\varphi_p) \cdot \mathbf{n}_f - g\varphi_p, \qquad \eta_p := -\mathbf{u}_f \cdot \mathbf{n}_f + \gamma_p^{-1} \mathbf{n}_f \cdot \mathbb{T}(\mathbf{u}_f, p_f) \cdot \mathbf{n}_f$$

are interface variables which correspond to the information exchange between the Stokes and the Darcy system. The restrictions on γ_f and γ_p are such that

$$a_f(\mathbf{u}_f, \mathbf{v}) + \langle \gamma_f \mathbf{u}_f \cdot \mathbf{n}_f, \mathbf{v} \cdot \mathbf{n}_f \rangle_\Gamma \text{ and } a_p(\varphi_p, \psi) + \langle \gamma_p^{-1} \varphi_p, \psi \rangle_\Gamma$$

are coercive.

2.16 Remark. Considering the Robin–Robin coupling, one still can approximate the Neumann–Neumann coupling by $\gamma_f = 0$ and $\gamma_p \to \infty$.

2.3.2 Finite element discretization

Combining the results of the finite element discretization of the Darcy problem in Section 2.1.2 and of the Stokes problem in Section 2.2.2, one obtains finite-dimensional subspaces $\mathbf{V}_f^h \subset \mathbf{V}_f$, $Q_f^h \subset Q_f$ and $Q_p^h \subset Q_p$. Since a conforming finite element discretization is considered, the discrete bilinear forms are restrictions of the continuous ones, i.e.,

$$a_f^h(\mathbf{u}_f^h, \mathbf{v}^h) = a_f(\mathbf{u}_f^h, \mathbf{v}^h), \quad b_f^h(\mathbf{v}^h, p^h) = b_f(\mathbf{v}^h, p^h), \quad a_p^h(\varphi^h, \psi^h) = a_p(\varphi^h, \psi^h),$$

for all $\mathbf{u}_f^h, \mathbf{v}^h \in \mathbf{V}_f^h$, $p^h \in Q_f^h$, $\varphi_p^h, \psi^h \in Q_p^h$. The discretization of the Robin–Robin coupling (28) then reads: Find $(\mathbf{u}_f^h, p_f^h, \varphi_p^h) \in \mathbf{V}_f^h \times Q_f^h \times Q_p^h$ such that

$$\left\{ \begin{aligned} a_f(\mathbf{u}_f^h,\mathbf{v}^h) + \langle \gamma_f \mathbf{u}_f^h\cdot\mathbf{n}_f, \mathbf{v}^h\cdot \mathbf{n}_f\rangle_\Gamma + b(\mathbf{v}^h,p_f^h) + \langle \eta_f^h, \mathbf{v}^h\cdot\mathbf{n}_f\rangle_\Gamma &= \langle \mathbf{f}_f^1,\mathbf{v}^h\rangle,\\ b_f(\mathbf{u}_f^h,q^h) &= \langle f_f^2, q^h\rangle,\\ a_p(\varphi_p^h,\psi^h)+\langle\gamma_p^{-1}\varphi_p^h,\psi^h\rangle_\Gamma + \langle\eta_p^h,\psi^h\rangle_\Gamma &= \langle f_p,\psi^h\rangle, \end{aligned}\right. \tag{29}$$

for all $(\mathbf{v}^h, q^h, \psi^h) \in \mathbf{V}_f^h \times Q_f^h \times Q_p^h$, where

$$\eta_f^h := -\gamma_f(\mathbb{K}\,\nabla\varphi_p^h)\cdot\mathbf{n}_f - g\varphi_p^h, \qquad \eta_p^h := -\mathbf{u}_f^h\cdot\mathbf{n}_f + \gamma_p^{-1}\mathbf{n}_f\cdot\mathbb{T}(\mathbf{u}_f^h,p_f^h)\cdot\mathbf{n}_f$$

are the discrete versions of η_f and η_p, respectively. As before, with $\gamma_f = 0$ and $\gamma_p \to \infty$, the discretization of the Neumann–Neumann coupling can be derived.

For obtaining a matrix-vector form of the problem, let

$$\left(\mathbf{u}_f^h, p_f^h, \varphi_p^h\right) = \left(\sum_{i=1}^{N_u} u_i^h \mathbf{w}_i, \sum_{i=1}^{N_p} p_i^h q_i, \sum_{i=1}^{N_\varphi} \varphi_i^h \psi_i\right)$$

be the sought solution with

$$\mathbf{V}_f^h = \operatorname{span}\{\mathbf{w}_i\}_{i=1}^{N_u},\ Q_f^h = \operatorname{span}\{q_i\}_{i=1}^{N_p},\ Q_p^h = \operatorname{span}\{\psi_i\}_{i=1}^{N_\varphi},$$

as well as

$$\underline{\mathbf{u}}_f = (u_i^h)_{i=1}^{N_u},\ \underline{\mathbf{p}}_f = (p_i^h)_{i=1}^{N_p},\ \underline{\boldsymbol{\varphi}}_p = (\varphi_i^h)_{i=1}^{N_\varphi}$$

being the coefficients that need to be determined. Further, define

$$\underline{\mathbf{f}}_1 = (\langle \mathbf{f}_f^1, \mathbf{w}_i\rangle)_{i=1}^{N_u},\ \underline{\mathbf{f}}_2 = (\langle f_f^2, q_i\rangle)_{i=1}^{N_p},\ \underline{\mathbf{f}}_3 = (\langle f_p, \psi_i\rangle)_{i=1}^{N_\varphi}.$$

(i) In case of the Neumann–Neumann coupling, this yields the problem to find $\mathbf{u}_f, \mathbf{p}_f, \boldsymbol{\varphi}_p$, such that

$$\left\{ \begin{aligned} \textstyle\sum_{i=1}^{N_u} u_i^h a_f^h(\mathbf{w}_i,\mathbf{w}_j) + \sum_{i=1}^{N_p} p_i^h b^h(\mathbf{w}_j,q_i) + \sum_{i=1}^{N_\varphi}\varphi_i^h\langle g\psi_i, \mathbf{w}_j\cdot\mathbf{n}_f\rangle_\Gamma &= \langle \mathbf{f}_f^1,\mathbf{w}_j\rangle,\\ \textstyle\sum_{i=1}^{N_u} u_i^h b^h(\mathbf{w}_i,q_k) &= \langle f_f^2, q_k\rangle,\\ \textstyle\sum_{i=1}^{N_\varphi}\varphi_i^h a_p^h(\psi_i,\psi_l) - \sum_{i=1}^{N_u} u_i^h\langle \mathbf{w}_i\cdot\mathbf{n}_f, \psi_l\rangle_\Gamma &= \langle f_p,\psi_l\rangle, \end{aligned}\right.$$

for all $j \in \{1,2,\dots,N_u\}$, $k \in \{1,2,\dots,N_p\}$, $l \in \{1,2,\dots,N_\varphi\}$. This is equivalent to the block-matrix-vector form

$$\begin{pmatrix} A & B & C_\Gamma^S \\ B^T & 0 & 0 \\ C_\Gamma^D & 0 & D \end{pmatrix} \begin{pmatrix} \underline{\mathbf{u}}_f \\ \underline{\mathbf{p}}_f \\ \underline{\boldsymbol{\varphi}}_p \end{pmatrix} = \begin{pmatrix} \underline{\mathbf{f}}_1 \\ \underline{\mathbf{f}}_2 \\ \underline{\mathbf{f}}_3 \end{pmatrix}$$

with

$$\begin{aligned} (A)_{ij} &= a_f(\mathbf{w}_j,\mathbf{w}_i) \in \mathbb{R}^{N_u\times N_u}, & (B)_{ij} &= b_f(\mathbf{w}_i,q_j) \in \mathbb{R}^{N_u\times N_p},\\ (C_\Gamma^S)_{ij} &= \langle g\psi_j, \mathbf{w}_i\cdot\mathbf{n}_f\rangle_\Gamma \in \mathbb{R}^{N_u\times N_\varphi}, & (C_\Gamma^D)_{ij} &= \langle \mathbf{w}_j\cdot\mathbf{n}_f, \psi_i\rangle_\Gamma \in \mathbb{R}^{N_\varphi\times N_u},\\ (D)_{ij} &= a_p(\psi_j,\psi_i) \in \mathbb{R}^{N_\varphi\times N_\varphi}\,. \end{aligned}$$

(ii) In case of the Robin–Robin coupling, one has two possibilities. Either one re-substitutes η_f and η_p into the equations or one treats them as separate variables, reducing the direct coupling between the two systems to the interface. The former approach yields the block-matrix-vector equation

$$\begin{pmatrix} A_{\text{rob}} & B & C_\varphi^S \\ B^T & 0 & 0 \\ C_u^D & C_p^D & D_{\text{rob}} \end{pmatrix} \begin{pmatrix} \underline{\mathbf{u}}_f \\ \underline{\mathbf{p}}_f \\ \underline{\boldsymbol{\varphi}}_p \end{pmatrix} = \begin{pmatrix} \underline{\mathbf{f}}_1 \\ \underline{\mathbf{f}}_2 \\ \underline{\mathbf{f}}_3 \end{pmatrix}$$

with

$$\begin{aligned} (A_{\text{rob}})_{ij} &= a_f(\mathbf{w}_j, \mathbf{w}_i) + \langle \gamma_f \mathbf{w}_j \cdot \mathbf{n}_f, \mathbf{w}_i \cdot \mathbf{n}_f \rangle_\Gamma \in \mathbb{R}^{N_u \times N_u}, \\ (B)_{ij} &= b_f(\mathbf{w}_i, q_j) \in \mathbb{R}^{N_u \times N_p}, \\ (D_{\text{rob}})_{ij} &= a_p(\psi_j, \psi_i) + \langle \gamma_p^{-1} \psi_j, \psi_i \rangle_\Gamma \in \mathbb{R}^{N_\varphi \times N_\varphi}, \\ (C_\varphi^S)_{ij} &= \langle g\psi_j, \mathbf{w}_i \cdot \mathbf{n}_f \rangle_\Gamma + \langle \gamma_f (\mathbb{K}\, \nabla \psi_j) \cdot \mathbf{n}_f, \mathbf{w}_i \cdot \mathbf{n}_f \rangle_\Gamma \in \mathbb{R}^{N_u \times N_\varphi}, \\ (C_u^D)_{ij} &= \langle 2\nu\gamma_p^{-1} \mathbf{n}_f \cdot \mathbb{D}(\mathbf{w}_j) \cdot \mathbf{n}_f, \psi_i \rangle_\Gamma - \langle \mathbf{w}_j \cdot \mathbf{n}_f, \psi_i \rangle_\Gamma \in \mathbb{R}^{N_\varphi \times N_u}, \\ (C_p^D)_{ij} &= \langle -\gamma_p^{-1} q_j, \psi_i \rangle_\Gamma \in \mathbb{R}^{N_\varphi \times N_p}\,. \end{aligned}$$

The latter approach yields a more complicated but less coupled block-matrix-vector equation. For its derivation, let $\eta_f = \sum_{i=1}^{N_{\eta,f}} \eta_f^i \Lambda_i$ and $\eta_p = \sum_{i=1}^{N_{\eta,p}} \eta_p^i \Lambda_i$ be sought as well, where Λ_i are basis functions restricted to the interface Γ. Defining $\underline{\boldsymbol{\eta}}_f := (\eta_f^i)_{i=1}^{N_{\eta,f}}$ and $\underline{\boldsymbol{\eta}}_p := (\eta_p^i)_{i=1}^{N_{\eta,p}}$, one obtains a system of the form

$$\begin{pmatrix} D_{\gamma_p} & 0 & 0 & 0 & E_p \\ R_p & -\mathbb{I} & 0 & 0 & 0 \\ 0 & E_f & A_{\gamma_f} & B & 0 \\ 0 & 0 & B^T & 0 & 0 \\ 0 & 0 & R_f^1 & R_f^2 & -\mathbb{I} \end{pmatrix} \begin{pmatrix} \underline{\boldsymbol{\varphi}}_p \\ \underline{\boldsymbol{\eta}}_f \\ \underline{\mathbf{u}}_f \\ \underline{\mathbf{p}}_f \\ \underline{\boldsymbol{\eta}}_p \end{pmatrix} = \begin{pmatrix} \underline{\mathbf{f}}_3 \\ \mathbf{0} \\ \underline{\mathbf{f}}_1 \\ \underline{\mathbf{f}}_2 \\ \mathbf{0} \end{pmatrix}$$

with

$$\begin{aligned} (A_{\gamma_f})_{ij} &= a_f(\mathbf{w}_j, \mathbf{w}_i) + \langle \gamma_f \mathbf{w}_j \cdot \mathbf{n}_f, \mathbf{w}_i \cdot \mathbf{n}_f \rangle_\Gamma \in \mathbb{R}^{N_u \times N_u}, \\ (R_f^1)_{ii} &= 2\nu\gamma_p^{-1} \mathbf{n}_f \cdot \mathbb{D}(\mathbf{w}_i) \cdot \mathbf{n}_f - \mathbf{w}_i \cdot \mathbf{n}_f \in \mathbb{R}^{N_u \times N_u}, \\ (B)_{ij} &= b_f(\mathbf{w}_i, q_j) \in \mathbb{R}^{N_u \times N_p}, & (D_{\gamma_p})_{ij} &= a_p(\psi_j, \psi_i) + \langle \gamma_p^{-1} \psi_j, \psi_i \rangle_\Gamma \in \mathbb{R}^{N_\varphi \times N_\varphi}, \\ (E_f)_{ij} &= \langle \Lambda_j, \mathbf{w}_i \cdot \mathbf{n}_f \rangle_\Gamma \in \mathbb{R}^{N_u \times N_{\eta,f}}, & (E_p)_{ij} &= \langle \Lambda_j, \psi_i \rangle_\Gamma \in \mathbb{R}^{N_\varphi \times N_{\eta,p}} \\ (R_f^2)_{ii} &= \gamma_p^{-1} q_i \in \mathbb{R}^{N_p \times N_p}, & (R_p)_{ii} &= -\gamma_f (\mathbb{K}\, \nabla \psi_i) \cdot \mathbf{n}_f - g\psi_i \in \mathbb{R}^{N_\varphi \times N_\varphi} \end{aligned}$$

and $\mathbb{I}$ being the identity.

Instead of solving the whole system at once, one can consider to solve the system iteratively by, e.g., a block Gauss–Seidel method, as it otherwise might be too large. For the formally decoupled Robin–Robin Stokes–Darcy problem, one obtains a method as presented in the following algorithm.

2.17 Algorithm [Block–Gauss–Seidel for the Robin–Robin coupling]**.** This algorithm solves the Robin–Robin coupling of the Stokes–Darcy equations in a Block–Gauss–Seidel manner.

(i) *Initialization.* Let

$$\begin{pmatrix} \underline{\boldsymbol{\varphi}}_p^{(0)} & \underline{\boldsymbol{\eta}}_f^{(0)} & \underline{\mathbf{u}}_f^{(0)} & \underline{\mathbf{p}}_f^{(0)} & \underline{\boldsymbol{\eta}}_p^{(0)} \end{pmatrix}^T \in \mathbb{R}^{N_\varphi + N_{\eta,f} + N_u + N_p + N_{\eta,p}}$$

be a randomly chosen start vector.

(ii) *Forward substitution in step $k \to k+1$.* For iteration step k, update the solution vector by

$$\begin{pmatrix} \underline{\boldsymbol{\varphi}}_p^{(k+1)} \\ \underline{\boldsymbol{\eta}}_f^{(k+1)} \\ \begin{pmatrix} \underline{\mathbf{u}}_f^{(k+1)} \\ \underline{\mathbf{p}}_f^{(k+1)} \end{pmatrix} \\ \underline{\boldsymbol{\eta}}_p^{(k+1)} \end{pmatrix} = \begin{pmatrix} D_{\gamma_p}^{-1} \left(\underline{\mathbf{f}}_3 - E_p \underline{\boldsymbol{\eta}}_p^{(k)} \right) \\ R_p \underline{\boldsymbol{\varphi}}_p^{(k+1)} \\ \begin{pmatrix} A_{\gamma_f} & B \\ B^T & 0 \end{pmatrix}^{-1} \left(\begin{pmatrix} \underline{\mathbf{f}}_1 \\ \underline{\mathbf{f}}_2 \end{pmatrix} - \begin{pmatrix} E_f \underline{\boldsymbol{\eta}}_f^{(k+1)} \\ 0 \end{pmatrix} \right) \\ R_f^1 \underline{\mathbf{u}}_f^{(k+1)} + R_f^2 \underline{\mathbf{p}}_f^{(k+1)} \end{pmatrix}.$$

(iii) If not converged, go back to step (ii), otherwise stop.

3 A posteriori error estimates

Motivation. Sometimes one faces problems that have singularities in their solution or quantities of interest which simply cannot be resolved by global grid refinement. This issue can be addressed by locally refining the grid appropriately.

One kind of error estimates are so called “a priori” error estimates. A priori means from the earlier which already indicates, that these estimates can be performed before the actual simulation. However, they usually depend on the continuous solution, only yield information on the asymptotic error behavior and therefore are not of great use concerning adaptive grid refinement.

Thus, one needs estimates that somehow integrate an already approximated solution and do not depend on the continuous solution. These estimates are called “a posteriori”, i.e., from the later. Generally, they are needed for two tasks:

(i) Control adaptive grid refinement by local error estimates.

(ii) Estimate the global error to be used as stopping criterion.

To meet the above mentioned requirements one considers two types of inequalities:

Given the data of a partial differential equation on a domain Ω with solution u and approximated solution u^h the estimate dealing with the overall accuracy is of the form

$$\|u - u^h\|_\Omega \leq C\eta, \tag{30}$$

where C is a positive constant which is independent of Ω, the refinement level h and u. Concerning the size of C at least the order of magnitude should be known and η is an error indicator computable using u^h and not u.

For identifying the regions at which refinement is needed a local estimate of the form

$$\eta_K \leq C\|u - u^h\|_{\omega(K)} \tag{31}$$

is considered, where $\omega(K)$ denotes a small neighborhood of a mesh cell K and η_K is computable using u^h and not u. Such a small neighborhood can, e.g., look like one of the neighborhoods that are shown in Figure 2. Typically, η and η_K are related by

$$\eta = \left(\sum_K \eta_K^2 \right)^{\frac{1}{2}}.$$

The idea is that large values of η_K indicate large local errors, so one needs to prove that the constant C in equation (31) can be bounded from above and below independently of K. On the other hand one hopes that refining regions with large local errors will have a great impact on the overall error. In Figure 4 an example of a local grid refinement is displayed. One can see that, in contrast to global and uniform refinement, the refined cells concentrate around the midpoint, i.e., where the values of η_K are largest.

In this section and the contained subsections, a few a posteriori error estimates will be derived, therefore they correspond to step (iii) of Algorithm 1.1.

3.1 Definition [Reliable error indicators]**.** If $\eta \leq$ "tolerance" holds and this implies that the true error is also smaller than the tolerance up to a multiplicative constant, the error indicator is called *reliable.* In particular an error indicator satisfying the upper bound (30) is reliable because

$$\eta \leq \mathrm{TOL} \Rightarrow \|u - u^h\|_\Omega \overset{(30)}{\leq} C\eta \leq C\,\mathrm{TOL}.$$

3.2 Definition [Locally efficient error indicators]**.** If $\eta_K \geq$ "tolerance" implies that the true error is also greater than the tolerance up to a multiplicative constant, the error indicator is called *locally efficient.* In particular an error indicator satisfying the local lower bound (31) is locally efficient because

$$\eta_K \geq \mathrm{TOL} \Rightarrow C\|u - u^h\|_{\omega(K)} \overset{(31)}{\geq} \eta_K \geq \mathrm{TOL}.$$

Sometimes this property is also referred to just as "efficient".

In order to classify a posteriori error indicators by quality, one can consider the so-called efficiency index.

3.3 Definition [Efficiency index]**.** The *efficiency index* of an a posteriori error indicator is given by the ratio of the estimated error to the actual error, i.e.,

$$I_{\text{eff}} := \frac{\eta}{\|u - u^h\|}. \tag{32}$$

If I_{eff} and I_{eff}^{-1} are bounded for all meshes, the overall process, i.e., Algorithm 1.1, is called *efficient.* It is called *asymptotically exact* if its efficiency index tends to one when the mesh-size tends to zero.

If the reliability of the error indicator is not given, it might happen that even if it is smaller or equal than the tolerance, the true error might still be large. On the other hand, if it is reliable but not efficient, a large local estimate does not necessarily indicate large local errors and one therefore might refine more than actually necessary or in the wrong subregions.

Thus, reliability is a crucial property of an a posteriori error indicator, if one wants to resolve the features of the solution that can be controlled with the considered norm up to a certain tolerance; efficiency is a good property with respect to run time and usage of resources, but it is not necessary.

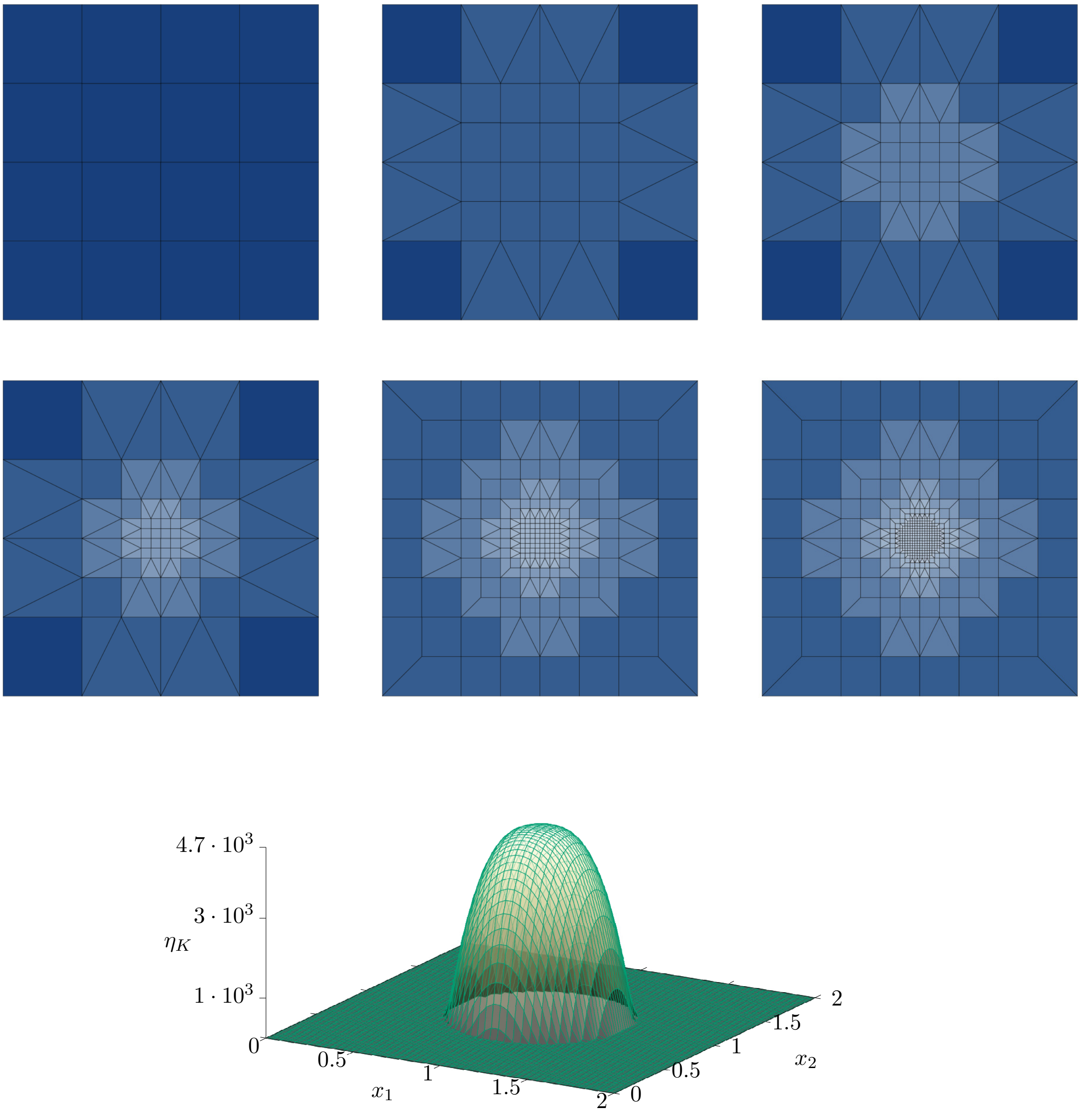


Figure 4: Local grid refinement. The upper six images display from left to right the adaptively refined grids $\mathcal{T}_0, \dots, \mathcal{T}_5$ for values of η_K as in the lower picture, i.e., a peak of the estimated error in the middle. The background color of each cell is tinted correspondingly to which $\mathcal{T}_k$ of the grid hierarchy it belongs.

3.1 Residual based a posteriori error estimates

One possibility to obtain a posteriori error estimates is estimating the error by the strong form residual of the problem in a suitable norm.

Here, following [Ver13, Section 4.1], first a common framework for abstract linear elliptic problems is set up. To this end, let $(X, \|\cdot\|_X), (Y, \|\cdot\|_Y)$ be Banach spaces. Further, let

$$\mathcal{L}(X,Y) := \{F : X \to Y : F \text{ continuous and linear}\}$$

be the space of continous linear mappings from X to Y equipped with the operator norm

$$\|F\|_{\mathcal{L}(X,Y)} := \sup_{\varphi \in X, \varphi \neq 0} \frac{\|F\varphi\|_Y}{\|\varphi\|_X}, \tag{33}$$

let $Y^* := \mathcal{L}(Y, \mathbb{R})$ and $\langle \ell, \varphi \rangle_Y$ be the dual pairing of Y^* and Y. Moreover, let $\mathcal{L}^2(X, Y, \mathbb{R})$ be the space of continuous bilinear mappings from $X \times Y$ to $\mathbb{R}$ equipped with the norm

$$\|B\|_{\mathcal{L}^2(X,Y,\mathbb{R})} := \sup_{\varphi \in X \setminus \{0\}} \sup_{\psi \in Y \setminus \{0\}} \frac{|B(\varphi, \psi)|}{\|\varphi\|_X \|\psi\|_Y}.$$

3.4 Remark. The spaces $\mathcal{L}^2(X, Y, \mathbb{R})$ and $\mathcal{L}(X, Y^*)$ are isomorphic under $B(\varphi, \psi) = \langle F\varphi, \psi \rangle_Y$ for all $\varphi \in X$, $\psi \in Y$ for $B \in \mathcal{L}^2(X, Y, \mathbb{R})$ and $F \in \mathcal{L}(X, Y^*)$.

Given a bilinear map $B \in \mathcal{L}^2(X, Y, \mathbb{R})$ and a linear functional $\ell \in Y^*$, consider the problem to find $\varphi \in X$ such that

$$B(\varphi, \psi) = \langle \ell, \psi \rangle_Y \quad \forall \psi \in Y, \tag{34}$$

or equivalently for $L \in \mathcal{L}(X, Y^*)$,

$$L\varphi = \ell. \tag{35}$$

3.5 Theorem [Hahn–Banach, [EMT04, Theorem 3.1.2]]**.** Let E be a subspace of a Banach space X, let E^* be its dual space and let $f_0 \in E^*$. Then, there exists an extension $f \in X^*$ such that $f\big|_E = f_0$ (i.e., $f(x) = f_0(x)$ for $x \in E$) and $\|f\|_{X^*} = \|f_0\|_{E^*}$. That is,

$$\sup_{\substack{x \in X \\ x \neq 0}} \frac{|f(x)|}{\|x\|_X} = \sup_{\substack{x \in E \\ x \neq 0}} \frac{|f_0(x)|}{\|x\|}.$$

Proof. See [EMT04, Chapter 9]. □

The Hahn–Banach Theorem has one corollary which can be used to show existence and uniqueness of a solution.

3.6 Corollary [Corollary of the Hahn–Banach Theorem, [EMT04, Corollary 3.1.7]]**.** Let $L \subset X$ be a subspace of a normed space X and let $x \in X$ such that $\operatorname{dist}(x, L) = d > 0$. Then, there exists $f \in X^*$ such that $\|f\|_{X^*} = 1$, $f(L) = 0$ and $f(x) = d$.

Proof. First consider $L_1 = \operatorname{span}\{x, L\}$; that is,

$$L_1 = \{\lambda x + y : \lambda \in \mathbb{R}, y \in L\}.$$

(i) Let $z = \lambda x + y$ and define the function $f_0 : L_1 \to \mathbb{R},\ z = \lambda x + y \mapsto \lambda \cdot d$. This function is well defined because $\lambda \in \mathbb{R}$ and $y \in L$ are uniquely defined by z. Indeed, assume z had two representations

$$z = \lambda_1 x + y_1 = \lambda_2 x + y_2 \Rightarrow y_1 - y_2 = \lambda_2 x - \lambda_1 x = (\lambda_2 - \lambda_1)x.$$

If $\lambda_2 - \lambda_1 = 0$, one obtains $y_1 - y_2 = 0 \Rightarrow y_1 = y_2$ and therefore uniqueness. Otherwise one obtains that the left-hand side is in L since L is a space. Assume the right-hand side would be in L as well, then

$$\frac{1}{\lambda_2 - \lambda_1}((\lambda_2 - \lambda_1)x) = x \in L,$$

which is a contradiction to the initial assumption that $\operatorname{dist}(x, L) = d > 0$ and therefore $(\lambda_2 - \lambda_1)x \notin L$, which means that z uniquely defines $\lambda \in \mathbb{R}$ and $y \in L$.

(ii) The function f_0 is linear, it is $f_0(L) = 0$ and $f_0(x) = d$. Indeed, let $z_1 = \lambda_1 x + y_1, z_2 = \lambda_2 x + y_2 \in L_1$, $\alpha \in \mathbb{R}$, then

$$f_0(z_1 + \alpha z_2) = d \cdot (\lambda_1 + \alpha\lambda_2) = f_0(z_1) + \alpha f_0(z_2).$$

Since the pair (λ, y) is uniquely defined by z, it is $z \in L \Rightarrow \lambda_1 = \lambda_2 = 0$ and thus $f_0(L) = 0$. Analogously for $f_0(x) = d$.

(iii) It is $\|f_0\|_{L_1^*} = 1$. Indeed, let $z \in L_1$ and assume that the corresponding $\lambda \neq 0$. Then one obtains

$$\|z\|_{L_1} = \|\lambda x + y\|_{L_1} = |\lambda| \cdot \left\| x + \frac{y}{\lambda} \right\|_{L_1} = |\lambda| \cdot \left\| x - \left(-\frac{y}{\lambda} \right) \right\|_{L_1} \geq |\lambda| d = |f_0(z)|,$$

since $-\frac{y}{\lambda} \in L$ and $\operatorname{dist}(x, L) = d > 0$. If $\lambda = 0$, one has the same estimate

$$\|z\|_{L_1} \geq 0 = |f_0(z)|.$$

This means that

$$\|f_0\|_{L_1^*} = \sup_{\substack{z \in L_1 \\ z \neq 0}} \frac{|f_0(z)|}{\|z\|_{L_1}} \leq \sup_{\substack{z \in L_1 \\ z \neq 0}} \frac{\|z\|_{L_1}}{\|z\|_{L_1}} = 1.$$

On the other hand there is due to $\operatorname{dist}(x, L) = d$ and L being a subspace a sequence $(y_n)_{n \in \mathbb{N}}$ in L such that $\|x + y_n\|_{L_1} = \|x - (-y_n)\|_{L_1} \to d$ for $n \to \infty$. This yields with the definition of the operator norm

$$d = \lim_{n \to \infty} |f_0(x + y_n)| \leq \lim_{n \to \infty} \|f_0\|_{L_1^*} \|x + y_n\|_{L_1} = \|f_0\|_{L_1^*} d$$

and therefore the other direction $\|f_0\|_{L_1^*} \geq 1$.

(iv) Now the Hahn–Banach Theorem can be applied providing the existence of an extension $f \in X^*$ of f_0 such that $\|f\|_{X^*} = \|f_0\|_{L_1^*} = 1$ and $f\big|_{L_1} = f_0$. This means that $f(L) = 0$ and $f(x) = d$.

□

3.7 Theorem [Existence and uniqueness of a solution, see [Ver13, Proposition 4.1]]**.** Assume the space Y is reflexive, that

$$\sup_{\varphi \in X} B(\varphi, \psi) > 0 \tag{36}$$

for all $\psi \in Y \setminus \{0\}$, and that B fulfills the inf-sup condition, i.e., that there is a constant $\beta > 0$, such that

$$\inf_{\varphi \in X \setminus \{0\}} \sup_{\psi \in Y \setminus \{0\}} \frac{B(\varphi, \psi)}{\|\varphi\|_X \|\psi\|_Y} = \beta. \tag{37}$$

Then there is a unique solution of (34) or equivalently (35) for every right-hand side $\ell \in Y^*$ and the solution depends continuously on the right-hand side.

Proof. Denote by $L \in \mathcal{L}(X, Y^*)$ the linear mapping which corresponds to B via the isomorphism given in Remark 3.4. Condition (37) then implies that L is injective. Indeed, assuming it is not injective, there are $\varphi_1, \varphi_2 \in X$ with $\varphi_1 \neq \varphi_2$ and $L\varphi_1 = L\varphi_2$. Equivalently, this means

$$\langle L(\varphi_1 - \varphi_2), \psi \rangle_Y = B(\varphi_1 - \varphi_2, \psi) = 0 \quad \forall \psi \in Y$$

and $\|\varphi_1 - \varphi_2\|_X > 0$. Thus there is $\varphi := \varphi_1 - \varphi_2 \in X \setminus \{0\}$ for which (37) cannot hold, which is a contradiction.

Since L is per definition continuous and surjective on the range of L, it is a bijection on range(L) and the range of its inverse operator is again X and thus closed. Therefore, the Closed Range Theorem of Banach (see [Yos80, p. 205]) can be applied. It yields that the range of L is a closed subspace of Y^*.

Furthermore, L is surjective. Indeed, assume it would not be surjective, then there exists $\psi_0^* \in Y^*$ with $\psi_0^* \notin$ range(L). Since the range of L is closed and therefore a subspace of Y^*, one can apply the above Corollary 3.6 of the Hahn–Banach Theorem, yielding that there exists a map $F \in Y^{**}$ such that $F(\psi^*) = 0$ for all $\psi^* \in$ range(L), $F(\psi_0^*) = 1$ and $\|F\|_{Y^{**}} = \text{dist}(\psi_0^*, \text{range}(L)) > 0$. By the reflexivity of the Banach space and therefore the isometry of the canonical inclusion or evaluation map $i : Y^{**} \to Y$, the element F can be identified with $\psi_0 \in Y$, such that

$$\langle L\varphi, \psi_0 \rangle_Y = 0$$

for all $\varphi \in X$ and $\|\psi_0\|_Y = \|F\|_{Y^{**}} > 0$. Therefore, it is $\psi_0 \in Y \setminus \{0\}$ which contradicts assumption (36). Hence, L is surjective and thus bijective.

By the bijectivity one obtains

$$\text{range}(L) = Y^*.$$

This is the case if and only if L has a continuous inverse, see [Yos80, Corollary VII.1]. Thus L is an isomorphism and since problems (34) and (35) are equivalent, this proves the assertion. □

For discretization purposes, let $X^h \subset X$ and $Y^h \subset Y$ be finite-dimensional subspaces on which $B^h \in \mathcal{L}^2(X^h, Y^h, \mathbb{R})$ and $\ell^h \in \mathcal{L}(Y^h, \mathbb{R})$ are defined. Then $\varphi^h \in X^h$ is sought such that

$$B^h(\varphi^h, \psi^h) = \langle \ell^h, \psi^h \rangle_{Y^h} \quad \forall \psi \in Y. \tag{38}$$

Using the isomorphism given in Remark 3.4, one can equivalently express the above problem as

$$L^h \varphi^h = \ell^h$$

for the map $L^h \in \mathcal{L}(X^h, \mathcal{L}(Y^h, \mathbb{R}))$ associated to B^h.

3.8 Theorem [Existence and uniqueness of a discrete solution]**.** Assume that

$$\sup_{\varphi^h \in X^h} B^h(\varphi^h, \psi^h) > 0 \tag{39}$$

for all $\psi^h \in Y^h \setminus \{0\}$, and that B^h fulfills a discrete variant of the inf-sup condition, i.e., that there is a constant $\beta^h > 0$, such that

$$\inf_{\varphi^h \in X^h \setminus \{0\}} \sup_{\psi^h \in Y^h \setminus \{0\}} \frac{B^h(\varphi^h, \psi^h)}{\|\varphi^h\|_{X^h} \|\psi^h\|_{Y^h}} = \beta^h. \tag{40}$$

Then there is a unique solution of (38) for every right-hand side $\ell^h \in \mathcal{L}(Y^h, \mathbb{R})$ and the solution depends continuously on the right-hand side.

Proof. The finite-dimensionality of X^h and Y^h gives continuity of B^h and ℓ^h, as well as reflexivity of X^h and Y^h. Therefore the statement follows immediately from Theorem 3.7. □

After having set up the abstract framework and given conditions for existence and uniqueness of the continuous problem and the finite-dimensional one, one can start deriving a posteriori error indicators. To this end, assume that the conditions of Theorems 3.7 and 3.8 are fulfilled and let $\varphi \in X$ be the unique solution of (34) and $\varphi \in X^h$ be the unique solution of (38).

It holds due to the linearity of L that

$$R := L(\varphi - \varphi^h) = \ell - L\varphi^h. \tag{41}$$

Using the inequality (33), one obtains

$$\|L(\varphi - \varphi^h)\|_{Y^*} \leq \|L\|_{\mathcal{L}(X,Y^*)} \|\varphi - \varphi^h\|_X \Rightarrow \|\varphi - \varphi^h\|_X \geq \|L\|^{-1}_{\mathcal{L}(X,Y^*)} \|\ell - L\varphi^h\|_{Y^*}$$

and on the other hand

$$\|L^{-1}L(\varphi - \varphi^h)\|_X = \|\varphi - \varphi^h\|_X = \|L^{-1}(\ell - L\varphi^h)\|_X \leq \|L^{-1}\|_{\mathcal{L}(Y^*,X)} \|\ell - L\varphi^h\|_{Y^*}$$

yielding altogether

$$\|L\|^{-1}_{\mathcal{L}(X,Y^*)} \|\ell - L\varphi^h\|_{Y^*} \leq \|\varphi - \varphi^h\|_X \leq \|L^{-1}\|_{\mathcal{L}(Y^*,X)} \|\ell - L\varphi^h\|_{Y^*}. \tag{42}$$

The above inequalities establish an equivalence between the norm of the residual $\|\ell - L\varphi^h\|_{Y^*}$ and the error $\|\varphi - \varphi\|_X$. However calculating $\|\ell - L\varphi^h\|_{Y^*}$ is not really practical for a posteriori indicators since it involves solving an infinite-dimensional problem.

Further, one should be able to localize the estimates to a single element $K \in \mathcal{T}$. Therefore, it is advantageous if one can write the residual $R = \ell - L\varphi^h$ in form of an integral over the domain. Finally, to get the a posteriori aspect into the estimates, Galerkin orthogonality is desired, since then it holds by linearity of the integral representation, that

$$\langle R, v^h \rangle = 0 \Rightarrow \langle R, v \rangle = \langle R, v - v^h \rangle$$

for some $v \in X$ and $v^h \in X^h$.

In particular, in the following subsections of this section, a posteriori error estimates will be derived making use of the below assumptions.

(i) Y is a subspace of $H^p(\Omega)$ for some p. This condition obviously holds for the considered problems.

(ii) *L^p representation*: The residual $R = \ell - L\varphi^h$ can be represented in terms of two functions $r \in L^p(\Omega)^*$ and $j \in L^p(\Sigma)^*$ such that

$$\langle R, v\rangle_Y = \int\limits_\Omega rv + \int\limits_\Sigma jv \tag{43}$$

for all $v \in H^p_\Gamma(\Omega)$, where Σ denotes the skeleton of the grid which was used for discretization.

(iii) *Galerkin orthogonality*: The residual satisfies $\langle R, \psi^h\rangle_Y = 0$ for all $\psi^h \in Y^h$.

3.9 Remark [On the Galerkin orthogonality]**.** The Galerkin orthogonality is no actual requirement but merely a technical assumption. In the case that B^h and ℓ^h are restrictions of B and ℓ respectively, i.e.,

$$B^h(\varphi^h, \psi^h) = B(\varphi^h, \psi^h),\ \langle \ell^h, \psi^h\rangle_{Y^h} = \langle \ell, \psi^h\rangle_Y$$

for all $\varphi^h \in X^h$, $\psi^h \in Y^h$, it is

$$\begin{aligned}\langle R, \psi^h\rangle_Y &= \langle \ell - L\varphi^h, \psi^h\rangle_Y = \langle \ell, \psi^h\rangle_Y - B(\varphi^h, \psi^h) = \langle \ell^h, \psi^h\rangle_{Y^h} - B(\varphi^h, \psi^h)\\ &= B^h(\varphi^h, \psi^h) - B(\varphi^h, \psi^h) = 0\end{aligned}$$

and thus Galerkin orthogonality is given. If B^h is not a restriction of B, it is suggested in [Ver13, Section 4.1.4] to introduce a restriction or projection operator $Q^h : Y \to Y^h$ and split the residual into the form

$$\ell - L\varphi^h = (\mathbb{I}_Y - Q^h)^*(\ell - L\varphi^h) + Q^{h,*}(\ell - L\varphi^h)$$

where $\mathbb{I}_Y$ denotes the identity on Y. One then obtains with the triangle inequality

$$\|\ell - L\varphi^h\|_{Y^*} \leq \|(\mathbb{I}_Y - Q^h)^*(\ell - L\varphi^h)\|_{Y^*} + \|Q^{h,*}(\ell - L\varphi^h)\|_{Y^*}$$

and therefore one obtains with (42)

$$\|\varphi - \varphi^h\|_X \leq \|L^{-1}\|_{\mathcal{L}(Y^*,X)} \Big(\|(\mathbb{I}_Y - Q^h)^*(\ell - L\varphi^h)\|_{Y^*} + \|Q^{h,*}(\ell - L\varphi^h)\|_{Y^*}\Big). \tag{44}$$

Note that if Galerkin orthogonality is given, the terms involving Q^h vanish and the above inequality reduces to the second part of (42).

Upper bounds for the error indicator now can in principal be derived by combining the conditions with inequalities of the Poincaré and Friedrichs type.

3.10 Remark [On deriving a lower bound]**.** For lower bounds one can choose a finite-dimensional subspace $\widetilde{Y}^h$ of Y such that

$$Y^h \subset \widetilde{Y}^h \subset Y$$

and replace $\|\ell - L\varphi^h\|_{Y^*}$ by $\|\ell - L\varphi^h\|_{\widetilde{Y}^{h,*}}$. By the inclusion it is $\|\ell - L\varphi^h\|_{\widetilde{Y}^{h,*}} \leq \|\ell - L\varphi^h\|_{Y^*}$ and therefore with (42):

$$\|\ell - L\varphi^h\|_{\widetilde{Y}^{h,*}} \leq \|L\|_{\mathcal{L}(X,Y^*)}\|\varphi - \varphi^h\|_X. \tag{45}$$

Here the enriched space $\widetilde{Y}^h$ is being considered instead of Y^h since the lower bound is of a local nature and one therefore can use cut-off functions to restrict the considered domain, which could, e.g., yield spaces of the form

$$\widetilde{Y}^h = \operatorname{span}\{\psi_K v\big|_K, \psi_E v\big|_E : K \in \mathcal{T}, E \in \mathcal{E}, v \in Y^h\} \supset Y^h,$$

where ψ_K and ψ_E are so-called cut-off functions, which basically restrict the multiplied function's value to a small neighborhood of K and E, respectively.

In order to derive an error indicator a desirable property is

$$\|(\mathbb{I}_Y - Q^h)^*(\ell - L\varphi^h)\|_{Y^*} + \|Q^{h,*}(\ell - L\varphi^h)\|_{Y^*} \leq C\|\ell - L\varphi^h\|_{\widetilde{Y}^{h,*}} \tag{46}$$

for a known constant C. It basically says that the error in the space Y^*, belonging to the continuous problem, can be bounded by some constant times the error in the enriched space $\widetilde{Y}^{h,*}$, which belongs to the discretized problem and thus is easier to compute. This property might not be easy to obtain, since the left-hand side involves a supremum over an infinite-dimensional space, whereas the right-hand side involves a supremum over a finite-dimensional space. However, if it can be established, finding an error indicator $\eta_\mathcal{T}$ depending on φ^h and the data of the problem is only a matter of showing that there exists a lower bound for $\|\ell - L\varphi^h\|_{\widetilde{Y}^{h,*}}$ and an upper bound for both $\|(\mathbb{I}_Y - Q^h)^*(\ell - L\varphi^h)\|_{Y^*}$ and $\|Q^{h,*}(\ell - L\varphi^h)\|_{Y^*}$.

If the data of the problem needs to be approximated in the solution process, it introduces an additional data error $\theta_\mathcal{T}$. Altogether, one obtains the following theorem:

3.11 Theorem [A posteriori error estimation for an abstract linear elliptic equation [Ver13, Theorem 4.7]]**.** Assume that the conditions of Theorem 3.7 and Theorem 3.8 are satisfied and denote by φ and φ^h the unique solutions of problems (34) and (38), respectively. Assume that there are a restriction operator $Q^h \in \mathcal{L}(Y, Y^h)$, a finite-dimensional subspace $\widetilde{Y}^h$ of Y with $Y^h \subset \widetilde{Y}^h \subset Y$, an error indicator $\eta_\mathcal{T}$, which only depends on the discrete solution φ^h and the given data of the variational problem (34), and a data error $\theta_\mathcal{T}$, which only depends on the data of the variational problem, such that the estimates

$$\|(\mathbb{I}_y - Q^h)^*(\ell - L\varphi^h)\|_{Y^*} \leq c_A(\eta_\mathcal{T} + \theta_\mathcal{T}), \quad \|Q^{h,*}(\ell - L\varphi^h)\|_{Y^*} \leq c_C(\eta_\mathcal{T} + \theta_\mathcal{T}) \tag{47}$$

and

$$\eta_\mathcal{T} \leq c_I\left(\|\ell - L\varphi^h\|_{\widetilde{Y}^{h,*}} + \theta_\mathcal{T}\right) \tag{48}$$

are fulfilled. Then the error $\varphi - \varphi^h$ can be estimated from above by

$$\|\varphi - \varphi^h\|_X \leq \|L^{-1}\|_{\mathcal{L}(Y^*,X)}(c_A + c_C)(\eta_\mathcal{T} + \theta_\mathcal{T})$$

and from below by

$$\eta_\mathcal{T} \leq c_I\left(\|L\|_{\mathcal{L}(X,Y^*)}\|\varphi - \varphi^h\|_X + \theta_\mathcal{T}\right).$$

Proof. For the upper bound it is

$$\begin{aligned}\|\varphi - \varphi^h\|_X &\leq \|L^{-1}\|_{\mathcal{L}(Y^*,X)}\left(\|(\mathbb{I}_Y - Q^h)^*(\ell - L\varphi^h)\|_{Y^*} + \|Q^{h,*}(\ell - L\varphi^h)\|_{Y^*}\right) &&\text{, with (44),}\\ &\leq \|L^{-1}\|_{\mathcal{L}(Y^*,X)}(c_A + c_C)(\eta_\mathcal{T} + \theta_\mathcal{T}) &&\text{, with (47).}\end{aligned}$$

For the lower bound it is

$$\begin{aligned}
\eta_{\mathcal{T}} &\le c_I \left(\|\ell - L\varphi^h\|_{\widetilde{Y}^{h,*}} + \theta_{\mathcal{T}} \right) && \text{, with (48),} \\
&\le c_I \left(\|L\|_{\mathcal{L}(X,Y^*)} \|\varphi - \varphi^h\|_X + \theta_{\mathcal{T}} \right) && \text{, with (45).}
\end{aligned}$$

□

3.12 Remark [On the Galerkin orthogonality, continuing Remark 3.9]**.** If the Galerkin orthogonality is given in terms of a bilinear map B^h that is simply the restriction of B, the constant c_C in the upper bound vanishes and so does the second inequality of (47), the first inequality simplifies to

$$\|\ell - L\varphi^h\|_{Y^*} \le c_A(\eta_{\mathcal{T}} + \theta_{\mathcal{T}}).$$

3.13 Remark [Quality of the error indicator]**.** The quantity

$$\|L\|_{\mathcal{L}(X,Y^*)} \|L^{-1}\|_{\mathcal{L}(Y^*,X)} (c_A + c_C) c_I$$

measures the quality of the error indicator $\eta_{\mathcal{T}}$. It corresponds to the condition number of the linear operator, except that it also takes the constants c_A, c_C and c_I into account. If they become large, so does the condition number. It should be uniformly bounded with respect to parameters of the differential equation and its discretization. This uniformity is often referred to as *robustness*, cf. [Ver13, Remark 4.8].

The quantity $\|L^{-1}\|_{\mathcal{L}(Y^*,X)}$ is actually the inverse inf-sup constant β^{-1}. Indeed,

$$\begin{aligned}
\|L^{-1}\|_{\mathcal{L}(Y^*,X)} &= \sup_{\psi^* \in Y^* \setminus \{0\}} \frac{\|L^{-1}\psi^*\|_X}{\|\psi^*\|_{Y^*}} && \text{, per definition of the operator norm,} \\
&= \sup_{L\varphi \in Y^* \setminus \{0\}} \frac{\|\varphi\|_X}{\|L\varphi\|_{Y^*}} && \text{, for } L\varphi = \psi^* \text{ since } L \text{ is an isomorphism,} \\
&= \left(\inf_{\varphi \in X \setminus \{0\}} \frac{\|L\varphi\|_{Y^*}}{\|\varphi\|_X} \right)^{-1} \\
&= \left(\inf_{\varphi \in X \setminus \{0\}} \sup_{\psi \in Y \setminus \{0\}} \frac{\langle L\varphi, \psi \rangle}{\|\varphi\|_X \|\psi\|_Y} \right)^{-1} && \text{, per definition of the operator norm,} \\
&= \beta^{-1}.
\end{aligned}$$

3.1.1 Application to the Darcy equations

This section applies the results of Section 2.1 to the abstract framework presented in the previous section. The corresponding parameters are

$$\begin{aligned}
X = Y &= H^1_{\Gamma_{\text{ess}}}(\Omega), \\
B(\varphi, \psi) &= a(\varphi, \psi) = (\mathbb{K}\, \nabla\varphi, \nabla\psi)_0 && \text{, see (7),} \\
\langle \ell, \psi \rangle_Y &= \langle \tilde{f}, \psi \rangle && \text{, see (6),} \\
X^h = Y^h &= V^h && \text{, see (10),} \\
B^h(\varphi^h, \psi^h) &= B(\varphi^h, \psi^h) = a^h(\varphi^h, \psi^h) && \text{, see (11),} \\
\langle \ell^h, \psi^h \rangle_{Y^h} &= \langle \ell, \psi^h \rangle_Y = \langle f, \psi^h \rangle - \langle u_{\text{nat}}, \psi^h \rangle_{\Gamma_{\text{nat}}} && \text{, see (11).}
\end{aligned}$$

Since the discrete forms B^h and ℓ^h are restrictions of B and ℓ respectively, one has Galerkin orthogonality and therefore needs no restriction operator Q^h, see Remark 3.9.

As in (41), the residual R of φ^h is the case of $\varphi_{\text{ess}} = 0$ is given by

$$\langle R, \psi\rangle := (f, \psi)_0 + \langle u_{\text{nat}}, \psi\rangle_{\Gamma_{\text{nat}}} - a(\varphi^h, \psi) \tag{49}$$

for all $\psi \in V$. To derive the L^p representation (43), the residual can be transformed as follows. One obtains for all $\psi \in V$

$$\begin{aligned}\langle R, \psi\rangle &= \int\limits_{\Omega} f\psi + \int\limits_{\Gamma_{\text{nat}}} u_{\text{nat}}\psi - \int\limits_{\Omega} \mathbb{K}\,\nabla\varphi^h \nabla\psi \\ &= \int\limits_{\Omega} f\psi + \int\limits_{\Gamma_{\text{nat}}} u_{\text{nat}}\psi + \sum_{K\in\mathcal{T}} \left(\int\limits_{K} K\Delta\varphi^h\psi - \int\limits_{\partial K} \mathbb{K}\,\nabla\varphi^h\psi\cdot\mathbf{n}_K \right) \\ &= \sum_{K\in\mathcal{T}} \int\limits_{K} (f + K\Delta\varphi^h)\psi + \sum_{E\in\mathcal{E}_{\Gamma_{\text{nat}}}} \int\limits_{E} (u_{\text{nat}} - \mathbb{K}\,\nabla\varphi^h\cdot\mathbf{n}_E)\psi - \sum_{E\in\mathcal{E}_{\Omega}} \int\limits_{E} [\![\mathbb{K}\,\nabla\varphi^h\cdot\mathbf{n}_E]\!]_E\psi\end{aligned}$$

by element-wise integration by parts and the assumption that $\mathbb{K} = K\,\mathbb{I}$. Here, $[\![\psi]\!]_E$ denotes the jump of ψ across the facet $E \in \mathcal{E}$ in direction of $\mathbf{n}_E$. To define the jump, let $K_1, K_2 \in \mathcal{T}$ be cells with the common edge $E \in \mathcal{E}$. The vector $\mathbf{n}_E$ is fixed as the outer unit normal vector of K_1 at E. Then, the jump of a function ψ across the edge E in the point $\mathbf{s} \in E$ is defined by

$$[\![\psi]\!]_E(\mathbf{s}) := \lim_{\substack{\mathbf{y}\to\mathbf{s}\\ \mathbf{y}\in K_1}} \psi(\mathbf{y}) - \lim_{\substack{\mathbf{y}\to\mathbf{s}\\ \mathbf{y}\in K_2}} \psi(\mathbf{y}).$$

It should be noted that this definition depends on the enumeration of the cells; if the order is reversed, the direction of $\mathbf{n}_E$ changes and therefore also the sign of the jump.

3.14 Remark [On the essential boundary data]**.** If one is in the situation that $\operatorname{meas}(\Gamma_{\text{ess}}) > 0$ and $\varphi_{\text{ess}} \neq 0$, one obtains in (49) the additional right-hand side term $-(\mathbb{K}\,\nabla\varphi_{\text{ess}}, \nabla\psi)_0$. However, this term can be included in $\langle f, \psi\rangle$, allowing to treat homogeneous and inhomogeneous essential boundary conditions in the same way.

It then is

$$\langle R, \psi\rangle = \int\limits_{\Omega} r\psi + \int\limits_{\Sigma} j\psi$$

with

$$r\big|_K := f + K\Delta\varphi^h$$

and

$$j\big|_E := \begin{cases} -[\![\mathbb{K}\,\nabla\varphi^h\cdot\mathbf{n}_E]\!]_E & \text{, for } E \in \mathcal{E}_\Omega, \\ u_{\text{nat}} - \mathbb{K}\,\nabla\varphi^h\cdot\mathbf{n}_E & \text{, for } E \in \mathcal{E}_{\Gamma_{\text{nat}}}, \\ 0 & \text{, for } E \in \mathcal{E}_{\Gamma_{\text{ess}}}, \end{cases}$$

for all $K \in \mathcal{T}$, $E \in \mathcal{E} = \mathcal{E}_{\Gamma_{\text{ess}}} \cup \mathcal{E}_{\Gamma_{\text{nat}}} \cup \mathcal{E}_\Omega$.

In order to meet the conditions of Theorem 3.11, an upper bound for the residual has to be provided. To this end, one can use nodal shape functions. These are specific functionals of the finite element discretization as considered in Section 2.1.2. They are uniquely defined as, in the case of simplices linear and in the case of quadrilaterals bilinear, continuous functions λ_z on each element for all $z \in \mathcal{N}$ with

$$\lambda_z(z) = 1 \text{ and } \lambda_z(w) = 0 \text{ for all } w \in \mathcal{N}\setminus\{z\}.$$

3.15 Lemma. It is

$$0 \le \lambda_z \le 1, \;\; \operatorname{supp} \lambda_z = \omega_z, \quad \sum_{z \in \mathcal{N}_K} \lambda_z = 1, \quad \sum_{z \in \mathcal{N}_E} \lambda_z = 1.$$

Proof. The second assertion $\operatorname{supp} \lambda_z = \omega_z$ is given by the definition of the λ_z. For the other assertions, the convexity of $K \in \mathcal{T}$ is needed. Since $K = \operatorname{conv}(\mathcal{N}_K)$, i.e., the convex hull of its vertices, one obtains

$$K = \left\{ \mathbf{x} \in \mathbb{R}^d : \mathbf{x} = \sum_{i=1}^{|\mathcal{N}_K|} \alpha_i \mathbf{v}_i, \; \mathbf{v}_i \in \mathcal{N}_K, \; \alpha_i \ge 0, \; \sum_{i=1}^{|\mathcal{N}_K|} \alpha_i = 1 \right\},$$

see, e.g., [Gru07, p. 42]. The weights α_i summing up to one and being non-negative implies, that in particular $0 \le \alpha_i \le 1$.

Let $\mathbf{x} = \sum_{i=1}^{|\mathcal{N}_K|} \alpha_i \mathbf{v}_i \in K$ be fixed. Using the linearity of the λ_z and $\lambda_{\mathbf{v}_i}(\mathbf{v}_j) = \delta_{ij}$, one obtains

$$\sum_{j=1}^{|\mathcal{N}_K|} \lambda_{\mathbf{v}_j}(\mathbf{x}) = \sum_{j=1}^{|\mathcal{N}_K|} \lambda_{\mathbf{v}_j} \left(\sum_{i=1}^{|\mathcal{N}_K|} \alpha_i \mathbf{v}_i \right) = \sum_{j=1}^{|\mathcal{N}_K|} \alpha_j = 1,$$

i.e., the third assertion.

The fourth assertion follows by the same arguments, since faces of convex polytopes are convex and $E \in \mathcal{E}$ is a face of some $K \in \mathcal{T}$, cf. [Gru07, p. 245]. □

3.16 Remark [On the uniqueness of the nodal shape functions]**.** Taking into account the mapping between reference cell and grid cell $F_K : \widehat{K}_d \to K$ as introduced in Section 7, one can write

$$\lambda_z = \widehat{\lambda}_z \circ F_K^{-1},$$

i.e., λ_z takes values on the reference cell $\widehat{K}_d$. The condition on the nodal shape functions is for vertices $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and basis functions $\mathcal{B} = \{\beta_1, \ldots, \beta_n\}$ that

$$\delta_{ij} = \lambda_i(\mathbf{v}_j) = \sum_{k=1}^{n} \alpha_{ik} \beta_k(\mathbf{v}_j),$$

where $\alpha_{i1}, \ldots, \alpha_{in}$ are the coefficients for the i-th nodal shape function $\lambda_i = \lambda_{\mathbf{v}_i}$. This condition can also be expressed as the matrix product $\mathbb{I} = AB$ with $(A)_{ij} = \alpha_{ij}$ and $B_{ij} = \beta_i(\mathbf{v}_j)$. Thus, for existence and uniqueness of the nodal shape functions, B has to be invertible.

In the case of simplices, the basis of linear functions is given by

$$\mathcal{B} = \{1, x_1, \ldots, x_d\}.$$

Therefore, B is given by

$$B = \begin{pmatrix} 1 & \cdots & 1 \\ v_1^{(1)} & \cdots & v_n^{(1)} \\ \vdots & & \vdots \\ v_1^{(d)} & \cdots & v_n^{(d)} \end{pmatrix}.$$

The determinant of the matrix is $\det B = d!\text{meas}(\widehat{K}_d)$ and therefore positive, assuming that the simplex is not degenerated, which is given in case of the reference simplex. Therefore, the nodal shape functions are uniquely defined on simplices.

In the case of quadrilaterals, the basis is given by

$$\mathcal{B} = \left\{ \prod_{i=1}^{d} x_i^{\alpha_i} = \mathbf{x}^{\boldsymbol{\alpha}} : \alpha_i \in \{0,1\} \right\} = \bigcup_{i=0}^{d} \mathcal{B}_i$$

with $\mathcal{B}_i = \{\mathbf{x}^{\boldsymbol{\alpha}} \in \mathcal{B} : |\boldsymbol{\alpha}| = i\}$. Without loss of generality it can be assumed that $\widehat{K}_d = [0,1]^d$. Further it can be assumed that the basis functions are ordered with respect to the $\mathcal{B}_i$, i.e., if $\beta_i \in \mathcal{B}_{k_1}$ and $\beta_j \in \mathcal{B}_{k_2}$ with $k_1 < k_2$, then $i < j$. Now the vertices $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}$ can be ordered such that they correspond directly to the multi-indices of the definition of the β_i, i.e., for $\mathbf{v}^{(i)}$ it should hold that $v_j^{(i)} = \alpha_j$ where α_j is the j-th entry of $\boldsymbol{\alpha}$ in $\beta_i = \mathbf{x}^{\boldsymbol{\alpha}}$.

In the case of $d = 3$ this means that

$$\mathcal{B} = \{1\} \cup \{x_1, x_2, x_3\} \cup \{x_1x_2, x_1x_3, x_2x_3\} \cup \{x_1x_2x_3\}$$

with corresponding vertices

$$\mathcal{N}_{\widehat{K}_d} = \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right\} \cup \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\} \cup \left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right\} \cup \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}.$$

Grouping the basis functions with $|\boldsymbol{\alpha}| = k$ for $k = 0, \ldots, d$ and therefore also grouping the vertices with k entries that are 1 yields a block-wise representation

$$B = \begin{pmatrix} D_0 & * & \cdots & * \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & D_d \end{pmatrix},$$

where the D_i are identity matrices up to interchanged columns. Indeed, the rows in B belonging to D_i correspond to basis functions with $|\boldsymbol{\alpha}| = i$ and the columns correspond to vertices $\mathbf{v} \in \mathcal{N}_{\widehat{K}_d}$ with

$$h(\mathbf{v}) := |\{j : v_j = 1\}| = i,$$

i.e., the number of entries in $\mathbf{v}$ that are 1 should be i. Therefore, for $\beta = \mathbf{x}^{\boldsymbol{\alpha}} \in \mathcal{B}_{|\boldsymbol{\alpha}|}$ and $\mathbf{v}$ with $h(\mathbf{v}) = k$, one obtains $\beta(\mathbf{v}) = 0$ if $k < |\boldsymbol{\alpha}|$, since β is a product out of more than k factors. If $k = |\boldsymbol{\alpha}|$, the there is exactly one $\beta \in \mathcal{B}_{|\boldsymbol{\alpha}|}$ such that $\beta(\mathbf{v}) = 1$.

Again in the case of $d = 3$, one obtains

$$B = \left(\begin{array}{c|ccc|ccc|c} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right).$$

Altogether this yields

$$\det B = \prod_{i=0}^{d} \det D_i = 1,$$

which proves that the nodal shape functions exist and are unique.

These nodal shape functions allow the construction of the quasi-interpolation operator of Clément (cf. [Ver13, Section 3.5.1])

$$I^h : L^1(\Omega) \to V^h,\ v \mapsto \begin{cases} 0 & \text{, if } z \in \mathcal{N}_{\Gamma_{\text{ess}}}, \\ \sum_{z\in\mathcal{N}} \frac{\int_{\omega_z} \lambda_z v}{\int_{\omega_z} \lambda_z} \lambda_z & \text{, otherwise,} \end{cases} \tag{50}$$

with the approximation properties

$$\begin{aligned} \|v - I^h v\|_{L^p(K)} &\leq C_K \|v\|_{L^p(\widetilde{\omega}_K)}, \\ \|v - I^h v\|_{L^p(K)} &\leq C_K h_K \|\nabla v\|_{L^p(\widetilde{\omega}_K)}, \\ \|\nabla(v - I^h v)\|_{L^p(K)} &\leq C_K \|\nabla v\|_{L^p(\widetilde{\omega}_K)}, \end{aligned} \tag{51}$$

$$\|v - I^h v\|_{L^p(E)} \leq C_E h_E^{1-\frac{1}{p}} \|\nabla v\|_{L^p(\widetilde{\omega}_E)}, \tag{52}$$

for $1 \leq p < \infty$.

3.17 Theorem [Upper bound for the residual]**.** The residual in the dual norm of V is bounded from above by

$$\|R\|_{V^*} \leq C(\eta_{\mathcal{T}} + \theta_{\mathcal{T}})$$

with

$$\begin{aligned} \eta_{\mathcal{T}} = \Bigg(&\sum_{K\in\mathcal{T}} h_K^2 \|f^h + K\Delta\varphi^h\|_{0,K}^2 + \sum_{E\in\mathcal{E}_\Omega} h_E \left\| [\![\, \mathbb{K} \nabla\varphi^h \cdot \mathbf{n}_E]\!] \right\|_{0,E}^2 \\ &+ \sum_{E\in\mathcal{E}_{\Gamma_{\text{nat}}}} h_E \|u_{\text{nat}}^h - \mathbb{K} \nabla\varphi^h \cdot \mathbf{n}_E\|_{0,E} \Bigg)^{\frac{1}{2}}, \end{aligned} \tag{53}$$

$$\theta_{\mathcal{T}} = \left(\sum_{K\in\mathcal{T}} h_K \|f - f^h\|_{0,K}^2 + \sum_{E\in\mathcal{E}_{\Gamma_{\text{nat}}}} \|u_{\text{nat}} - u_{\text{nat}}^h\|_{0,E}^2 \right)^{\frac{1}{2}}, \tag{54}$$

where f^h and u_{nat}^h are piecewise polynomial approximations of f and u_{nat} respectively.

Proof. Let $v \in V$ be arbitrary and $v^h = I^h v \in V^h$. By the Galerkin orthogonality it is

$$\langle R, v^h \rangle = 0 \Rightarrow \langle R, v \rangle = \int_\Omega r(v - v^h) + \int_\Sigma j(v - v^h).$$

The residual can be estimated by

$$\begin{aligned}
|\langle R, v\rangle| &= \left|\sum_{K\in\mathcal{T}}\int_K r(v-I^h v) + \sum_{E\in\mathcal{E}}\int_E j(v-I^h v)\right| \\
&\leq \sum_{K\in\mathcal{T}} \|r\|_{0,K}\|v-I^h v\|_{0,K} + \sum_{E\in\mathcal{E}} \|j\|_{0,E}\|v-I^h v\|_{0,E} && \text{, using Cauchy–Schwarz,} \\
&\leq C\left(\sum_{K\in\mathcal{T}} h_K\|r\|_{0,K}\|v\|_{V(\widetilde{\omega}_K)} + \sum_{E\in\mathcal{E}} h_E^{1/2}\|j\|_{0,E}\|v\|_{V(\widetilde{\omega}_E)}\right) && \text{, using (51) and (52),} \\
&\leq C\left(\sum_{K\in\mathcal{T}} h_K^2\|r\|_{0,K}^2 + \sum_{E\in\mathcal{E}} h_E\|j\|_{0,E}^2\right)^{\frac{1}{2}} \\
&\qquad \times \left(\sum_{K\in\mathcal{T}} \|v\|_{V(\widetilde{\omega}_K)}^2 + \sum_{E\in\mathcal{E}} \|v\|_{V(\widetilde{\omega}_E)}^2\right)^{\frac{1}{2}},
\end{aligned}$$

where the last step was obtained by applying the Cauchy–Schwarz inequality for sums. Further it is

$$\left(\sum_{K\in\mathcal{T}} \|v\|_{V(\widetilde{\omega}_K)}^2 + \sum_{E\in\mathcal{E}} \|v\|_{V(\widetilde{\omega}_E)}^2\right)^{\frac{1}{2}} \leq C\|v\|_V,$$

with the constant taking into account that the faces are being counted multiple times. This estimate shows that the norm of the residual is bounded from above by

$$\|R\|_{V^*} \leq C\left(\sum_{K\in\mathcal{T}} h_K^2\|r\|_{0,K}^2 + \sum_{E\in\mathcal{E}} h_E\|j\|_{0,E}^2\right)^{\frac{1}{2}}.$$

Substitution of $f = f^h + f - f^h$, $u_{\text{nat}} = u_{\text{nat}}^h + u_{\text{nat}} - u_{\text{nat}}^h$ and an application of the triangle inequality plus the inequality

$$\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$$

for $a, b \geq 0$ yields the statement of the theorem. □

The above theorem already gives an error indicator which is reliable. The next condition of the Theorem 3.11 is a bound from below, i.e., efficiency. The efficiency estimate should be applicable to a local portion of the grid, which is why one could use cut-off functions, often due to their shape also referred to as 'bubble functions'. These are defined on the elements $K \in \mathcal{T}$ and their facets $E \in \mathcal{E}$ by

$$\psi_K = \beta_K \prod_{z\in\mathcal{N}_K} \lambda_z, \qquad \psi_E = \beta_E \prod_{z\in\mathcal{N}_E} \lambda_z$$

with factors β_K and β_E such that the functions attain the value 1 at the barycenters of K and E respectively, as indicated for β_K in Figure 5. Further it holds that $\psi_K\big|_{\partial K} = 0$ and $\psi_E\big|_{\partial\omega_E} = 0$.

3.18 Lemma [Properties of the cut-off functions]**.** It is $\operatorname{supp}(\psi_K) = K$, $\operatorname{supp}(\psi_E) = \omega_E$, $0 \leq \psi_K \leq 1$, $0 \leq \psi_E \leq 1$ and $\max_{x\in K}\psi_K(x) = \max_{x\in E}\psi_E(x) = 1$. Further, the following

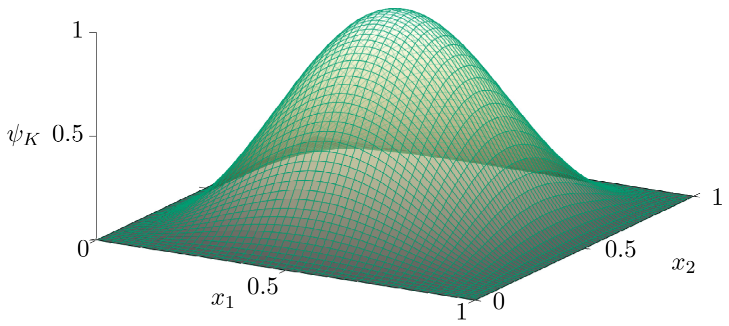


Figure 5: Graph of the cut-off function ψ_K for $K = [0,1]^2$.

inverse estimates hold for all polynomials $v \in P(K)$ of degree k and $w \in P(E)$ of degree k:

$$\|v\|_{0,K} \le C_{K,k} \|\psi_K^{1/2} v\|_{0,K}, \tag{55}$$

$$\|\nabla(\psi_K v)\|_{0,K} \le C_{K,k} h_K^{-1} \|v\|_{0,K}, \tag{56}$$

$$\|w\|_{0,E} \le C_{E,k} \|\psi_E^{1/2} w\|_{0,E}, \tag{57}$$

$$\|\nabla(\psi_E w)\|_{0,K} \le C_{E,K,k} h_E^{-1/2} \|w\|_{0,E}, \tag{58}$$

$$\|\psi_E w\|_{0,K} \le C_{E,K,k} h_E^{1/2} \|w\|_{0,E}. \tag{59}$$

Proof. The estimates can be found in [Ver13, Proposition 3.37]. Since $\operatorname{supp} \lambda_z = \omega_z$, the product in the definition of the cut-off functions restricts their support to K and ω_E respectively, i.e.,

$$\operatorname{supp} \psi_K = \bigcap_{z \in \mathcal{N}_K} \omega_z = K$$

and

$$\operatorname{supp} \psi_E = \bigcap_{z \in \mathcal{N}_E} \omega_z = \omega_E.$$

The support regarding the edge yields ω_E since ω_E is contained in all ω_z. □

3.19 Theorem [Lower bound for the residual]**.** The residual can be bounded from below by

$$\eta_{R,K} \le c_* \left(\| \mathbb{K} \nabla(\varphi - \varphi^h)\|_{\omega_K}^2 + \theta^h \right)^{\frac{1}{2}}$$

with

$$\eta_{R,K} = \Bigg(h_K^2 \|f_K^h + K\Delta\varphi^h\|_{0,K}^2 + \frac{1}{2} \sum_{E \in \mathcal{E}_{K,\Omega}} h_E \left\| [\![\nabla\varphi^h \cdot \mathbf{n}_E]\!]_E \right\|_{0,E}^2 \\ + \sum_{E \in \mathcal{E}_{K,\Gamma_{\text{nat}}}} h_E \|u_{\text{nat},E}^h - \nabla\varphi^h \cdot \mathbf{n}_E\|_{0,E}^2 \Bigg)^{\frac{1}{2}} \tag{60}$$

where f_K^h and u_{nat}^h are piecewise polynomial approximations as defined below and $\theta_{\mathcal{T}}$ is defined as in (54).

Proof. First, f and u_{nat} are being replaced on every element $K \in \mathcal{T}$ and edge $E \in \mathcal{E}_{\Gamma_{\mathrm{nat}}}$ by their respective piecewise polynomial approximations f^h and u_{nat}^h. Set

$$f_{\mathcal{T}}^h = \sum_{K \in \mathcal{T}} f_K^h \chi_K, \qquad u_{\mathrm{nat},\mathcal{T}}^h = \sum_{E \in \mathcal{E}_{\Gamma_{\mathrm{nat}}}} u_{\mathrm{nat},E}^h \chi_E,$$

where χ_A denotes the characteristic function with respect to the set A.

The local bound is now going to be derived by successively inserting appropriate test functions into the L^p representation of the residual.

(i) Let $K \in \mathcal{T}$ be arbitrary and insert

$$w_K = (f_K^h + K\Delta\varphi^h)\psi_K$$

into the L^p representation of the residual, yielding

$$\begin{aligned} \langle R, w_K \rangle &= \int\limits_\Omega r w_K + \int\limits_\Sigma j w_K = \int\limits_K r w_K \qquad &&\text{, since } \operatorname{supp} w_K \subset K, \\ &= (\mathbb{K}\,\nabla(\varphi - \varphi^h), \nabla w_K)_K. \end{aligned}$$

It then is

$$(r, w_K)_{0,K} + (f_K^h - f, w_K)_{0,K} = (\mathbb{K}\,\nabla(\varphi - \varphi^h), \nabla w_K)_{0,K} + (f_K^h - f, w_K)_{0,K}$$

and on the other hand

$$(r, w_K)_{0,K} + (f_K^h - f, w_K)_{0,K} = (f_K^h + K\Delta\varphi^h, w_K)_{0,K} = ((f_K^h + K\Delta\varphi^h)^2, \psi_K)_{0,K},$$

by the definition of r. This yields

$$(\mathbb{K}\,\nabla(\varphi - \varphi^h), \nabla w_K)_{0,K} + (f_K^h - f, w_K)_{0,K} = ((f_K^h + K\Delta\varphi^h)^2, \psi_K)_{0,K}.$$

The right-hand side can be estimated with the inverse estimate (55), resulting in

$$((f_K^h + K\Delta\varphi^h)^2, \psi_K)_{0,K} \geq C_{K,k}^{-2} \|f_K^h + K\Delta\varphi^h\|_{0,K}^2.$$

The left-hand side's terms can be estimated by

$$\begin{aligned} |(\mathbb{K}\,\nabla(\varphi - \varphi^h), \nabla w_K)_{0,K}| &\leq \|\,\mathbb{K}\,\nabla(\varphi - \varphi^h)\|_{0,K} \|\nabla w_K\|_{0,K} &&\text{, using Cauchy–Schwarz,} \\ &\leq \|\,\mathbb{K}\,\nabla(\varphi - \varphi^h)\|_{L^2(K)} \\ &\quad \times C_{K,k} h_K^{-1} \|f_K^h + K\Delta\varphi^h\|_{0,K} &&\text{, using (56),} \\ |((f_K^h - f), w_K)_{0,K}| &\leq \|f_K^h - f\|_{0,K} \|w_K\|_{0,K} &&\text{, using Cauchy–Schwarz,} \\ &\leq \|f_K^h - f\|_{0,K} \|f_K^h + K\Delta\varphi^h\|_{0,K} &&\text{, using } 0 \leq \psi_K \leq 1. \end{aligned}$$

Thus, combining the above estimates and rearranging terms yields

$$h_K \|f_K^h + K\Delta\varphi^h\|_{0,K} \leq C^3 \|\,\mathbb{K}\,\nabla(\varphi - \varphi^h)\|_{0,K} + C^2 h_K \|f - f_K^h\|_{0,K}, \tag{61}$$

where the constant C only depends on the shape parameter of K and the polynomial degree k.

(ii) Let $E \in \mathcal{E}_\Omega$ be arbitrary and insert

$$w_E = -\llbracket \mathbb{K}\,\nabla\varphi^h \cdot \mathbf{n}_E \rrbracket_E \psi_E$$

into the L^p representation of the residual, yielding

$$\begin{aligned}
\langle R, w_E\rangle &= (r, w_E)_0 + (j, w_E)_{0,\Sigma} = (r, w_E)_{0,\omega_E} + (j, w_E)_{0,E} && \text{, since } \operatorname{supp} w_E \subset \omega_E,\\
\langle R, w_E\rangle &= \langle L(\varphi - \varphi^h), w_E\rangle && \text{, using (41),}\\
&= (\mathbb{K}\,\nabla(\varphi - \varphi^h), \nabla w_E)_{0,\omega_E}.
\end{aligned}$$

This gives together with

$$\begin{aligned}
(j, w_E)_{0,E} &= (\mathbb{K}\,\nabla(\varphi - \varphi^h), \nabla w_E)_{0,\omega_E} - (r, w_E)_{0,\omega_E} && \text{, and,}\\
(j, w_E)_{0,E} &= (-\llbracket \mathbb{K}\,\nabla\varphi^h \cdot \mathbf{n}_E \rrbracket_E, w_E)_{0,E} && \text{, by definition of } j,\\
&= (\llbracket \mathbb{K}\,\nabla\varphi^h \cdot \mathbf{n}_E \rrbracket_E^2, \psi_E)_{0,E} && \text{, by definition of } w_E,
\end{aligned}$$

that

$$\begin{aligned}
&(\llbracket \mathbb{K}\,\nabla\varphi^h \cdot \mathbf{n}_E \rrbracket_E^2, \psi_E)_{0,E}\\
&\quad= (\mathbb{K}\,\nabla(\varphi - \varphi^h), \nabla w_E)_{0,\omega_E} - (r, w_E)_{0,\omega_E}\\
&\quad= (\mathbb{K}\,\nabla(\varphi - \varphi^h), \nabla w_E)_{0,\omega_E} - (f + K\Delta\varphi^h + f^h_{\omega_E} - f^h_{\omega_E}, w_E)_{0,\omega_E}\\
&\quad= (\mathbb{K}\,\nabla(\varphi - \varphi^h), \nabla w_E)_{0,\omega_E} - \sum_{K\subset\omega_E} (f^h_K + K\Delta\varphi^h, w_E)_{0,K} - \sum_{K\subset\omega_E} (f - f^h_E, w_E)_{0,K}.
\end{aligned}$$

The terms are again being estimated separately. The left-hand side can be estimated with the inverse estimate (57)

$$(\llbracket \mathbb{K}\,\nabla\varphi^h \cdot \mathbf{n}_E \rrbracket_E^2, \psi_E)_{0,E} \geq C_{E,k}^{-2} \left\| \llbracket \mathbb{K}\,\nabla\varphi^h \cdot \mathbf{n}_E \rrbracket_E \right\|_{0,E}^2.$$

The terms on the right-hand side can be estimated by

$$\begin{aligned}
&|(\mathbb{K}\,\nabla(\varphi - \varphi^h), \nabla w_E)_{0,K}|\\
&\quad\leq \|\mathbb{K}\,\nabla(\varphi - \varphi^h)\|_{0,K} \|\nabla w_E\|_{0,K} && \text{, using Cauchy–Schwarz,}\\
&\quad\leq \|\mathbb{K}\,\nabla(\varphi - \varphi^h)\|_{0,K} C_{E,k,k} h_E^{-\frac{1}{2}} \left\| \llbracket \mathbb{K}\,\nabla\varphi^h \cdot \mathbf{n}_E \rrbracket_E \right\|_{0,E} && \text{, using (58),}\\
&|(f^h_K + K\Delta\varphi^h, w_E)_{0,K}|\\
&\quad\leq \|f^h_K + K\Delta\varphi^h\|_{0,K} \|w_E\|_{0,K} && \text{, using Cauchy–Schwarz,}\\
&\quad\leq \|f^h_K + K\Delta\varphi^h\|_{0,K} C_{E,K,k} h_E^{\frac{1}{2}} \left\| \llbracket \mathbb{K}\,\nabla\varphi^h \cdot \mathbf{n}_E \rrbracket_E \right\|_{0,E} && \text{, using (59),}\\
&|(f - f^h_E, w_E)_{0,K}|\\
&\quad\leq \|f - f^h_E\|_{0,K} \|w_E\|_{0,K} && \text{, using Cauchy–Schwarz,}\\
&\quad\leq \|f - f^h_E\|_{0,K} C_{E,K,k} h_E^{\frac{1}{2}} \left\| \llbracket \mathbb{K}\,\nabla\varphi^h \cdot \mathbf{n}_E \rrbracket_E \right\|_{0,E} && \text{, using (59).}
\end{aligned}$$

yielding

$$\begin{aligned}
C_{E,k}^{-2} \left\| \llbracket \mathbb{K}\,\nabla\varphi^h \cdot \mathbf{n}_E \rrbracket_E \right\|_{0,E} &\leq C_{E,k,k} h_E^{-\frac{1}{2}} \|\mathbb{K}\,\nabla(\varphi - \varphi^h)\|_{0,\omega_E}\\
&\quad+ \sum_{K\subset\omega_E} C_{E,K,k} h_E^{\frac{1}{2}} \|f^h_K + K\Delta\varphi^h\|_{0,K} + \sum_{K\subset\omega_E} C_{E,K,k} h_E^{\frac{1}{2}} \|f - f^h_E\|_{0,K}.
\end{aligned}$$

Combining this estimate with (61) gives

$$C_{E,k}^{-2} \left\| [\![\, \mathbb{K}\, \nabla \varphi^h \cdot \mathbf{n}_E]\!]_E \right\|_{0,E} \leq C \| \, \mathbb{K}\, \nabla(\varphi - \varphi^h) \|_{0,\omega_E} + C \sum_{K \subset \omega_E} h_K \| f - f_K^h \|_{0,K}, \tag{62}$$

where the constants only depend on the shape parameter of $\mathcal{T}$ and the polynomial degree used in the inverse estimates.

(iii) Let $E \in \mathcal{E}_{\Gamma_{\text{nat}}}$ be an arbitrary edge on the boundary with natural boundary conditions and let

$$w_E = (u_{\text{nat}}^h - \nabla \varphi^h \cdot \mathbf{n}_E)\psi_E$$

be a test function. It then is

$$\begin{aligned}
(j, w_E)_{0,E} &= (\mathbb{K}\, \nabla(\varphi - \varphi^h), \nabla w_E)_K - (r, w_E)_{0,K} && \text{, same as in (ii),} \\
(j, w_E)_{0,E} &= (u_{\text{nat}} - \nabla \varphi^h \cdot \mathbf{n}_E, w_E)_{0,E} && \text{, by definition of } j, \\
&= (u_{\text{nat}}^h - \nabla \varphi^h \cdot \mathbf{n}_E, w_E)_{0,E} + (u_{\text{nat}} - u_{\text{nat}}^h, w_E)_{0,E} \\
&= ((u_{\text{nat}}^h - \nabla \varphi^h \cdot \mathbf{n}_E)^2, \psi_E)_{0,E} + (u_{\text{nat}} - u_{\text{nat}}^h, w_E)_{0,E} && \text{, by definition of } w_E,
\end{aligned}$$

which implies that

$$\begin{aligned}
&((u_{\text{nat}}^h - \nabla \varphi^h \cdot \mathbf{n}_E)^2, \psi_E)_{0,E} \\
&= (\mathbb{K}\, \nabla(\varphi - \varphi^h), \nabla w_E)_{0,K} - (r, w_E)_{0,E} + (u_{\text{nat}}^h - u_{\text{nat}}, w_E)_{0,E} \\
&= (\mathbb{K}\, \nabla(\varphi - \varphi^h), \nabla w_E)_{0,K} - (f_K^h + K\Delta\varphi^h, w_E)_{0,K} \\
&\quad - (f - f_K^h, w_E)_{0,K} + (u_{\text{nat}}^h - u_{\text{nat}}, w_E)_{0,E}.
\end{aligned}$$

The L^2 scalar product is over K instead of ω_E since it is a boundary facet. Same as in (i) and (ii) the terms are now being estimated separately. The left-hand side can be estimated with (57) by

$$((u_{\text{nat}}^h - \nabla \varphi^h \cdot \mathbf{n}_E)^2, \psi_E)_{0,E} \geq C_{E,k}^{-2} \left\| u_{\text{nat}}^h - \nabla \varphi^h \cdot \mathbf{n}_E \right\|_{0,E}^2 .$$

The terms on the right-hand side can be estimated by

$$\begin{aligned}
|(\mathbb{K}\, \nabla(\varphi - \varphi^h), \nabla w_E)_{0,K}| &\leq \| \, \mathbb{K}\, \nabla(\varphi - \varphi^h) \|_{0,K} C_{E,K,k} h_E^{-\frac{1}{2}} \| u_{\text{nat}}^h - \nabla \varphi^h \cdot \mathbf{n}_E \|_{0,E}, \\
|(f_K^h + K\Delta\varphi^h, w_E)_{0,K}| &\leq \| f_K^h + K\Delta\varphi^h \|_{0,E} C_{E,K,k} h_E^{\frac{1}{2}} \| u_{\text{nat}}^h - \nabla \varphi^h \cdot \mathbf{n}_E \|_{0,E}, \\
|(f - f_K^h, w_E)_{0,K}| &\leq \| f - f_K^h \|_{0,K} C_{E,K,k} h_E^{\frac{1}{2}} \| u_{\text{nat}}^h - \nabla \varphi^h \cdot \mathbf{n}_E \|_{0,E}, \\
|(u_{\text{nat},E}^h - u_{\text{nat}}, w_E)_{0,E}| &\leq \| u_{\text{nat},E}^h - u_{\text{nat}} \|_{0,E} \| u_{\text{nat}}^h - \nabla \varphi^h \cdot \mathbf{n}_E \|_{0,E},
\end{aligned}$$

with application of the Cauchy–Schwarz inequality and estimates (58), (59), (59) and that $0 \leq \psi_E \leq 1$ respectively. Altogether this gives

$$\begin{aligned}
h_E^{\frac{1}{2}} \| u_{\text{nat}}^h - \nabla \varphi^h \cdot \mathbf{n}_E \|_{0,E} \leq C \| \, \mathbb{K}\, \nabla(\varphi - \varphi^h) \|_{0,K} &+ C h_E \| f_K^h + K\Delta\varphi^h \|_{0,E} \\
&+ C h_E \| f - f_K^h \|_{0,K} + C h_E^{\frac{1}{2}} \| u_{\text{nat}} - u_{\text{nat},E}^h \|_{0,E},
\end{aligned}$$

where the constants only depend on the shape parameter of $\mathcal{T}$ and the polynomial degree used in the inverse estimates. Using (61) this can be further estimated by

$$\begin{aligned}
&h_E^{\frac{1}{2}} \| u_{\text{nat}}^h - \nabla \varphi^h \cdot \mathbf{n}_E \|_{0,E} \\
&\quad \leq C \| \, \mathbb{K}\, \nabla(\varphi - \varphi^h) \|_{0,K} + C(h_K + h_E) \| f - f_K^h \|_{0,K} + C h_E^{\frac{1}{2}} \| u_{\text{nat}} - u_{\text{nat},E}^h \|_{0,E}.
\end{aligned} \tag{63}$$

The above derived estimates (61), (62) and (63) can now be applied to $\eta_{R,K}$ together with a variant of Young's inequality

$$(a+b)^2 \leq 2a^2 + 2b^2$$

for $a, b \in \mathbb{R}$, yielding the theorem's statement, where the factor of $\frac{1}{2}$ in $\eta_{R,K}$ takes into account that every edge is being counted twice while summing up the jumps. □

Theorems 3.17 and 3.19 show that all assumptions of Theorem 3.11 of the abstract setting are fulfilled with

$$Y^h \subset \widetilde{Y}^h = \operatorname{span}\{\psi_K v\big|_K,\ \psi_E v\big|_E : K \in \mathcal{T}, E \in \mathcal{E}, v \in V^h\}.$$

Therefore,

$$\eta_{\mathcal{T}} = \left(\sum_{K\in\mathcal{T}} \eta_{R,K}^2\right)^{\frac{1}{2}}$$

as defined in (53) and (60) indeed bounds the error in the energy norm $|\cdot|_1$ from above and below and thus is reliable and efficient.

3.1.2 Application to the Stokes equations

This section applies the results of Section 2.2 to the abstract framework for linear problems presented in Section 3. The parameters are

$$\begin{aligned}
X = Y &= \mathbf{V} \times Q && \text{, see (15),}\\
\|(\mathbf{u}, p)\|_X &= \left(\|\mathbf{u}\|_{\mathbf{V}}^2 + \|p\|_Q^2\right)^{\frac{1}{2}} = \left(\|\nabla \mathbf{u}\|_0^2 + \|p\|_0^2\right)^{\frac{1}{2}}, &&\\
B((\mathbf{u}, p), (\mathbf{v}, q)) &= a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) - b(\mathbf{u}, q) && \text{, see (17),}\\
\langle \ell, (\mathbf{v}, q)\rangle_Y &= \langle \widetilde{\mathbf{f}}, \mathbf{v}\rangle = (\mathbf{f}, \mathbf{v})_0 + \langle T_N, \mathbf{v}\rangle_{\Gamma_{\text{nat}}} && \text{, see (17),}\\
B^h((\mathbf{u}^h, p^h), (\mathbf{v}^h, q^h)) &= B((\mathbf{u}^h, p^h), (\mathbf{v}^h, q^h)) && \text{, see (20),}\\
\langle \ell^h, (\mathbf{v}^h, q^h)\rangle_{Y^h} &= \langle \tilde{\mathbf{f}}, \mathbf{v}^h\rangle && \text{, see (20).}
\end{aligned}$$

The equation for $B(\cdot,\cdot)$ is obtained by subtracting the two equations in (17). Since a conforming finite element discretization is assumed, Galerkin orthogonality is given. In order to derive the L^p representation (43), one can apply element-wise integration by parts as in the following. Let $(\mathbf{v}, q) \in \mathbf{V} \times Q$ be a test function. Inserting this test function into the residual equation of the Laplace form yields

$$\begin{aligned}
\langle R, (\mathbf{v}, q)\rangle &= (\mathbf{f}, \mathbf{v})_0 + \langle T_N, \mathbf{v}\rangle_{\Gamma_{\text{nat}}} - a(\mathbf{u}^h, \mathbf{v}) - b(\mathbf{v}, p^h) + b(\mathbf{u}^h, q) \qquad \text{, see (17),}\\
&= (\mathbf{f}, \mathbf{v})_0 + \langle T_N, \mathbf{v}\rangle_{\Gamma_{\text{nat}}} + \sum_{K\in\mathcal{T}} \Big((\nu \Delta \mathbf{u}^h, \mathbf{v})_{0,K} - (\nu \nabla \mathbf{u}^h \cdot \mathbf{n}_K, \mathbf{v})_{0,\partial K} \Big)\\
&\quad - \sum_{K\in\mathcal{T}} \Big((\mathbf{v}, \nabla p^h)_{0,K} - (\mathbf{v}\cdot \mathbf{n}_K, p^h)_{0,\partial K} \Big) - \sum_{K\in\mathcal{T}} (\nabla \cdot \mathbf{u}^h, q)_{0,K}\\
&= \sum_{K\in\mathcal{T}} (\mathbf{f} + \nu\Delta \mathbf{u}^h - \nabla p^h, \mathbf{v})_{0,K} - \sum_{E\in\mathcal{E}_\Omega} \left(\llbracket (\nu \nabla \mathbf{u}^h - p^h\, \mathbb{I}) \cdot \mathbf{n}_E \rrbracket_E, \mathbf{v}\right)_{0,E}\\
&\quad + \sum_{E\in\mathcal{E}_{\Gamma_{\text{nat}}}} (T_N - (\nu\nabla \mathbf{u}^h + p^h\, \mathbb{I})\cdot \mathbf{n}_E, \mathbf{v})_{0,E} - \sum_{K\in\mathcal{T}} (\nabla\cdot \mathbf{u}^h, q)_{0,K}.
\end{aligned}$$

Similarly, inserting $(\mathbf{v},q)$ into the Cauchy stress form yields

$$
\begin{aligned}
\langle R,(\mathbf{v},q)\rangle &= (\mathbf{f},\mathbf{v})_0 + \langle T_N,\mathbf{v}\rangle_{\Gamma_{\text{nat}}} + \sum_{K\in\mathcal{T}} \Big((2\nu\nabla\cdot\mathbb{D}(\mathbf{u}^h),\mathbf{v})_{0,K} - (2\nu\,\mathbb{D}(\mathbf{u}^h)\cdot\mathbf{n}_K,\mathbf{v})_{0,\partial K}\Big) \\
&\quad - \sum_{K\in\mathcal{T}} \Big((\nabla p^h,\mathbf{v})_{0,K} - (p^h,\mathbf{v}\cdot\mathbf{n}_K)_{0,\partial K}\Big) - \sum_{K\in\mathcal{T}} (\nabla\cdot\mathbf{u}^h,q)_{0,K} \\
&= \sum_{K\in\mathcal{T}} (\mathbf{f}+\nabla\cdot(2\nu\,\mathbb{D}(\mathbf{u}^h)-p^h\,\mathbb{I}),\mathbf{v})_{0,K} - \sum_{E\in\mathcal{E}_\Omega} \Big([\![(2\nu\,\mathbb{D}(\mathbf{u}^h)-p^h\,\mathbb{I})\cdot\mathbf{n}_E]\!]_E,\mathbf{v}\Big)_{0,E} \\
&\quad + \sum_{E\in\mathcal{E}_{\Gamma_{\text{nat}}}} (T_N - (2\nu\,\mathbb{D}(\mathbf{u}^h)-p^h\,\mathbb{I})\cdot\mathbf{n}_E,\mathbf{v})_{0,E} - \sum_{K\in\mathcal{T}} (\nabla\cdot\mathbf{u}^h,q)_{0,K},
\end{aligned}
$$

giving altogether

$$
\langle R,(\mathbf{v},q)\rangle = \int\limits_\Omega r(\mathbf{v},q)^T + \int\limits_\Sigma j(\mathbf{v},q)^T \tag{64}
$$

with

$$
\begin{aligned}
r\big|_K &= (\mathbf{f}+\nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h),\nabla\cdot\mathbf{u}^h)^T, \\
j\big|_E &= \begin{cases} (-[\![\,\mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E]\!]_E,0)^T & \text{, on } \mathcal{E}_\Omega, \\ (T_N - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E,0)^T & \text{, on } \mathcal{E}_{\Gamma_{\text{nat}}}, \\ (\mathbf{0},0)^T & \text{, on } \mathcal{E}_{\Gamma_{\text{ess}}}, \end{cases}
\end{aligned}
$$

for all $K\in\mathcal{T}$ and $E\in\mathcal{E}$.

3.20 Remark [On the essential boundary data]**.** If one has inhomogeneous essential boundary data, one obtains in (64) the additional right-hand side term

$$
\begin{cases} -(2\nu\,\mathbb{D}(\mathbf{u}_{\text{ess}}),\mathbb{D}(\mathbf{v}))_0 & \text{, in case of the Cauchy stress form,} \\ -(\nu\nabla\mathbf{u}_{\text{ess}},\nabla\mathbf{v}) & \text{, in case of the Laplacian form.} \end{cases}
$$

This term however can be included into the source term $(\mathbf{f},\mathbf{v})_0$, allowing to treat inhomogeneous and homogeneous essential boundary data in the same way.

In order to apply Theorem 3.11, upper and lower bounds need to be provided.

3.21 Theorem [Upper bound for the residual]**.** The residual in the dual norm of X is bounded from above by

$$
\|R\|_{X^*} \le C(\eta_\mathcal{T}+\theta_\mathcal{T})
$$

with

$$
\begin{aligned}
\eta_\mathcal{T} = \Bigg(&\sum_{K\in\mathcal{T}} h_K^2\|\mathbf{f}^h - \nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h)\|_{0,K}^2 + \sum_{K\in\mathcal{T}} \|\nabla\cdot\mathbf{u}^h\|_{0,K}^2 \\
&+ \sum_{E\in\mathcal{E}_\Omega} h_E\|\,[\![\,\mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E]\!]_E\|_{0,E}^2 + \sum_{E\in\mathcal{E}_{\Gamma_{\text{nat}}}} h_E\|T_N^h - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\|_{0,E}^2\Bigg)^{\frac{1}{2}},
\end{aligned} \tag{65}
$$

$$
\theta_\mathcal{T} = \Bigg(\sum_{K\in\mathcal{T}} h_K^2\|\mathbf{f}-\mathbf{f}^h\|_{0,K}^2 + \sum_{E\in\mathcal{E}_{\Gamma_{\text{nat}}}} h_E\|T_N - T_N^h\|_{0,E}^2\Bigg)^{\frac{1}{2}}, \tag{66}
$$

where $\mathbf{f}^h$ denotes a piecewise polynomial approximation of $\mathbf{f}$ and T_N^h denotes a piecewise polynomial approximation of T_N.

Proof. Let $(\mathbf{v}, q) \in \mathbf{V} \times Q$ be arbitrary and $\mathbf{v}^h = I^h \mathbf{v} \in \mathbf{V}^h$, where I^h is the quasi-interpolation operator (50) applied to the components of $\mathbf{v}$. With the Galerkin orthogonality one obtains

$$\langle R, (\mathbf{v}^h, q^h) \rangle = 0 \Rightarrow \langle R, (\mathbf{v}, q) \rangle = \int\limits_{\Omega} r(\mathbf{v} - I^h \mathbf{v}, q - I^h q)^T + \int\limits_{\Sigma} j(\mathbf{v} - I^h \mathbf{v}, q - I^h q)^T.$$

With this identity the residual can be estimated by

$$\begin{aligned}
|\langle R, (\mathbf{v}^h, q^h) \rangle| \leq & \sum_{K \in \mathcal{T}} \|\mathbf{f} + \nabla \cdot \mathbb{T}(\mathbf{u}^h, p^h)\|_{0,K} \|\mathbf{v} - I^h \mathbf{v}\|_{0,K} \\
& + \sum_{E \in \mathcal{E}_\Omega} \| [\![\, \mathbb{T}(\mathbf{u}^h, p^h) \cdot \mathbf{n}_E]\!]_E \|_{0,E} \|\mathbf{v} - I^h \mathbf{v}\|_{0,E} \\
& + \sum_{E \in \mathcal{E}_{\Gamma_{\text{nat}}}} \|T_N - \mathbb{T}(\mathbf{u}^h, p^h) \cdot \mathbf{n}_E\|_{0,E} \|\mathbf{v} - I^h \mathbf{v}\|_{0,E} \\
& + \sum_{K \in \mathcal{T}} \|\nabla \cdot \mathbf{u}^h\|_{\mathbf{V},K} \|q\|_{0,K} && \text{, using Cauchy–Schwarz,} \\
\leq & \sum_{K \in \mathcal{T}} C_E h_K \|\mathbf{f} + \nabla \cdot \mathbb{T}(\mathbf{u}^h, p^h)\|_{0,K} \|\mathbf{v}\|_{\mathbf{V}(\tilde{\omega}_K)} && \text{, using (51),} \\
& + \sum_{E \in \mathcal{E}_\Omega} C_E h_E^{\frac{1}{2}} \| [\![\, \mathbb{T}(\mathbf{u}^h, p^h) \cdot \mathbf{n}_E]\!]_E \|_{0,E} \|\mathbf{v}\|_{\mathbf{V}(\tilde{\omega}_E)} && \text{, using (52),} \\
& + \sum_{E \in \mathcal{E}_{\Gamma_{\text{nat}}}} C_E h_E^{\frac{1}{2}} \|T_N - \mathbb{T}(\mathbf{u}^h, p^h) \cdot \mathbf{n}_E\|_{0,E} \|\mathbf{v}\|_{\mathbf{V}(\tilde{\omega}_E)} && \text{, using (52),} \\
& + \sum_{K \in \mathcal{T}} \|\nabla \cdot \mathbf{u}^h\|_{\mathbf{V},K} \|q\|_{0,K} \\
\leq & \, C \left(\sum_{K \in \mathcal{T}} h_K^2 \|\mathbf{f} + \nabla \cdot \mathbb{T}(\mathbf{u}^h, p^h)\|_{0,K}^2 \right)^{\frac{1}{2}} \|\mathbf{v}\|_{\mathbf{V}} \\
& + C \left(\sum_{E \in \mathcal{E}_\Omega} h_E \| [\![\, \mathbb{T}(\mathbf{u}^h, p^h) \cdot \mathbf{n}_E]\!]_E \|_{0,E}^2 \right)^{\frac{1}{2}} \|\mathbf{v}\|_{\mathbf{V}} \\
& + C \left(\sum_{E \in \mathcal{E}_{\Gamma_{\text{nat}}}} h_E \|T_N - \mathbb{T}(\mathbf{u}^h, p^h) \cdot \mathbf{n}_E\|_{0,E}^2 \right)^{\frac{1}{2}} \|\mathbf{v}\|_{\mathbf{V}} \\
& + C \left(\sum_{K \in \mathcal{T}} \|\nabla \cdot \mathbf{u}^h\|_{0,K}^2 \right)^{\frac{1}{2}} \|q\|_0 && \text{, using Cauchy–Schwarz,} \\
\leq & \, C \|(\mathbf{v}, q)\|_X \Bigg(\sum_{K \in \mathcal{T}} h_K^2 \|\mathbf{f} + \nabla \cdot \mathbb{T}(\mathbf{u}^h, p^h)\|_{0,K}^2 \\
& + \sum_{K \in \mathcal{T}} \|\nabla \cdot \mathbf{u}^h\|_{0,K}^2 + \sum_{E \in \mathcal{E}_\Omega} h_E \| [\![\, \mathbb{T}(\mathbf{u}^h, p^h) \cdot \mathbf{n}_E]\!]_E \|_{0,E}^2 \\
& + \sum_{E \in \mathcal{E}_{\Gamma_{\text{nat}}}} h_E \|T_N - \mathbb{T}(\mathbf{u}^h, p^h) \cdot \mathbf{n}_E\|_{0,E}^2 \Bigg)^{\frac{1}{2}},
\end{aligned}$$

where the last inequality is due to an implication of Young's inequality, namely

$$(\sqrt{a}\|\mathbf{v}\|_{\mathbf{V}}+\sqrt{b}\|q\|_Q)^2 \leq (\|\mathbf{v}\|^2+\|q\|^2)(a+b) \Leftrightarrow 2\sqrt{ab}\|\mathbf{v}\|_{\mathbf{V}}\|q\|_Q \leq \|\mathbf{v}\|^2_{\mathbf{V}}b+\|q\|^2_Q a,$$

for $a, b \geq 0$. Since $\mathbf{v}$ and q were arbitrary, replacement of $\mathbf{f}$ by $\mathbf{f}^h+\mathbf{f}-\mathbf{f}^h$, T_N by $T_N^h+T_N-T_N^h$ and an application of the triangle inequality yields the theorem's statement. □

3.22 Theorem [Lower bound for the residual]**.** It holds that

$$\eta_{R,K} \leq C\left(\|\nu\nabla(\mathbf{u}-\mathbf{u}^h)\|^2_{\omega_K}+\|p-p^h\|^2_{\omega_K}+\theta_{\mathcal{T}}\right)$$

with

$$\eta_{R,K} = \left(h_K^2\|\mathbf{f}^h+\nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h)\|^2_{0,K}+\|\nabla\cdot\mathbf{u}^h\|^2_{0,K}+\frac{1}{2}\sum_{E\in\mathcal{E}_{K,\Omega}} h_E\left\|\llbracket\,\mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\rrbracket_E\right\|^2_E\right.$$
$$\left.+\sum_{E\in\mathcal{E}_{K,\Gamma_{\mathrm{nat}}}} h_E\|T_N^h-\mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\|^2_{0,E}\right) \tag{67}$$

and $\theta_{\mathcal{T}}$ as in Theorem 3.21.

Proof. Similarly to the lower bound of the Darcy equation's residual a posteriori estimator, this lower bound is going to be provided by successively inserting appropriate test functions into the L^2 representation of the residual. To this end, let

$$\mathbf{f}^h_{\mathcal{T}} := \sum_{K\in\mathcal{T}}\mathbf{f}^h\chi_K, \quad T^h_{N,\mathcal{T}} := \sum_{E\in\mathcal{E}_{\mathrm{nat}}} T_N^h\chi_E$$

be cell-restricted versions of $\mathbf{f}^h$ and T_N^h respectively.

(i) Let $K\in\mathcal{T}$ be arbitrary and insert

$$(\mathbf{w}_K, 0) = ((\mathbf{f}^h+\nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h))\psi_K, 0)$$

as test function into the residual equation, yielding

$$\begin{aligned}\langle R,(\mathbf{w}_K,0)\rangle &= (r,(\mathbf{w}_K,0))_K && \text{, since } \operatorname{supp}\mathbf{w}_K\subset K,\\ \langle R,(\mathbf{w}_K,0)\rangle &= B((\mathbf{u}-\mathbf{u}^h,p-p^h),(\mathbf{w}_K,0)) && \text{, using (41),}\\ &= a(\mathbf{u}-\mathbf{u}^h,\mathbf{w}_K)+b(\mathbf{w}_K,p-p^h).\end{aligned}$$

Inserting the definition of $r\big|_K$ and $\mathbf{w}_K$ into the first equation gives

$$\begin{aligned}\langle R,(\mathbf{w}_K,0)\rangle = (r,(\mathbf{w}_K,0)) &= (\mathbf{f}+\nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h),\mathbf{w}_K)_K\\ &= ((\mathbf{f}^h+\nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h))^2,\psi_K)_K+(\mathbf{f}-\mathbf{f}^h_K,\mathbf{w}_K)_K,\end{aligned}$$

yielding

$$((\mathbf{f}^h+\nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h))^2,\psi_K)_K = a(\mathbf{u}-\mathbf{u}^h,\mathbf{w}_K)+b(\mathbf{w}_K,p-p^h)+(\mathbf{f}-\mathbf{f}^h_K,\mathbf{w}_K)_K.$$

The left-hand side can be estimated by applying the inverse estimate (55) which results in

$$((\mathbf{f}^h+\nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h))^2,\psi_K)_K \geq C_{K,k}^{-2}\|\mathbf{f}^h_K+\nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h)\|^2_{0,K}.$$

The terms on the right-hand side are being estimated individually. To this end, note that one obtains by application of the triangle inequality that

$$\| \mathbb{D}(\mathbf{v})\|_0 = \frac{1}{2}\|\nabla\mathbf{v} + \nabla\mathbf{v}^T\|_0 \leq \|\nabla\mathbf{v}\|_0.$$

Thus, when deriving upper bounds it can be assumed that the laplace version of $a(\cdot,\cdot)$ is being used, as the Cauchy–Stress version yields the same results up to a factor of 2 which can be absorbed into the constants. The terms can be estimated by

$$\begin{aligned}
&|a(\mathbf{u}-\mathbf{u}^h, \mathbf{w}_K)| \\
&\quad \leq C_a\|\nu\nabla(\mathbf{u}-\mathbf{u}^h)\|_{0,K}\|\nabla\mathbf{w}_K\|_{0,K} && \text{, using Cauchy–Schwarz,} \\
&\quad \leq C_a C_{K,k} h_K^{-1}\|\nu\nabla(\mathbf{u}-\mathbf{u}^h)\|_{0,K}\|\mathbf{f}^h + \nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h)\|_{0,K} && \text{, using (56),} \\
&|b(\mathbf{w}_K, p-p^h)| \\
&\quad \leq \|\nabla\cdot\mathbf{w}_K\|_{0,K}\|p-p^h\|_{0,K} && \text{, using Cauchy–Schwarz,} \\
&\quad \leq \sqrt{d}\|\nabla\mathbf{w}_K\|_{0,K}\|p-p^h\|_{0,K} && \text{, using Lemma 2.12,} \\
&\quad \leq \sqrt{d}C_{K,k}h_K^{-1}\|p-p^h\|_{0,K}\|\mathbf{f}^h + \nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h)\|_{0,K} && \text{, using (56),} \\
&|(\mathbf{f}-\mathbf{f}_K^h, \mathbf{w}_K)_{0,K}| \\
&\quad \leq \|\mathbf{f}-\mathbf{f}_K^h\|_{0,K}\|\mathbf{w}_K\|_{0,K} && \text{, using Cauchy–Schwarz,} \\
&\quad \leq \|\mathbf{f}-\mathbf{f}_K^h\|_{0,K}\|\mathbf{f}^h + \nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h)\|_{0,K} && \text{, using } 0 \leq \psi_K \leq 1,
\end{aligned}$$

where

$$C_a = \begin{cases} 1 & \text{, if } a(\mathbf{u},\mathbf{v}) = (\nu\nabla\mathbf{u}, \nabla\mathbf{v})_0, \\ 2 & \text{, if } a(\mathbf{u},\mathbf{v}) = (2\nu\,\mathbb{D}(\mathbf{u}), \mathbb{D}(\mathbf{v}))_0. \end{cases}$$

Combining the above estimates and dividing by $\|\mathbf{f}^h - \nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h)\|_{0,K}$ yields

$$\begin{aligned}
&h_K\|\mathbf{f}^h + \nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h)\|_{0,K} \\
&\qquad \leq C\left(\|\nu\nabla(\mathbf{u}-\mathbf{u}^h)\|_{0,K} + \|p-p^h\|_{0,K} + h_K\|\mathbf{f}-\mathbf{f}^h\|_{0,K}\right),
\end{aligned} \tag{68}$$

where the constant only depends on the spatial dimension of the problem, the shape parameter of K and the polynomial degree k.

(ii) Similarly to step (i), one can assume the laplace version of $a(\cdot,\cdot)$. Let $E \in \mathcal{E}_\Omega$ be arbitrary and

$$(w_E, 0) = (-[\![\, \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E]\!]_E\psi_E, 0)$$

a test function which inserted into the residual equation yields

$$\begin{aligned}
\langle R, (\mathbf{w}_E, 0)\rangle &= (r, (\mathbf{w}_E, 0))_{0,\omega_E} + (j, (\mathbf{w}_E, 0))_{0,E} && \text{, since} \operatorname{supp}\mathbf{w}_E \subset \omega_E, \\
\langle R, (\mathbf{w}_E, 0)\rangle &= a(\mathbf{u}-\mathbf{u}^h, \mathbf{w}_E) + b(\mathbf{w}_E, p-p^h) && \text{, using (41).}
\end{aligned}$$

Thus on the one hand one obtains by using the definition of j, that

$$(j, (\mathbf{w}_E, 0))_{0,E} = (-[\![\, \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E]\!]_E, \mathbf{w}_E)_{0,E} = ([\![\, \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E]\!]_E^2, \psi_E)_{0,E},$$

and on the other hand using

$$\begin{aligned}
(r, (\mathbf{w}_E, 0))_{0,\omega_E} &= (\mathbf{f} + \nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h) + \mathbf{f}_{\omega_E}^h - \mathbf{f}_{\omega_E}^h)_{0,\omega_E} \\
&= \sum_{K\subset\omega_E}(\mathbf{f}_K^h + \nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h), \mathbf{w}_E) - \sum_{K\subset\omega_E}(\mathbf{f}-\mathbf{f}_K^h, \mathbf{w}_E)_{0,\omega_E}
\end{aligned}$$

together with the definition of the residual as above, that

$$\begin{aligned}
&(j,(\mathbf{w}_E,0))_{0,E} = a(\mathbf{u}-\mathbf{u}^h,\mathbf{w}_E) + b(\mathbf{w}_E,p-p^h) - (r,(\mathbf{w}_E,0))_{0,\omega_E}\\
&\quad = \sum_{K\subset\omega_E} (\nu\nabla(\mathbf{u}-\mathbf{u}^h),\nabla\mathbf{w}_E)_{0,K} + \sum_{K\subset\omega_E} (\nabla\cdot\mathbf{w}_E,p-p^h)_{0,K}\\
&\qquad - \sum_{K\subset\omega_E} (\mathbf{f}_K^h + \nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h),\mathbf{w}_E) - \sum_{K\subset\omega_E} (\mathbf{f}-\mathbf{f}_K^h,\mathbf{w}_E)_{0,\omega_E}.
\end{aligned}$$

By applying the inverse estimate (57) to the first equation for j one obtains

$$\left(\llbracket\,\mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\rrbracket_E^2,\psi_E\right)_{0,E} \geq C_{E,k}^{-2}\left\|\llbracket\,\mathbb{T}(\mathbf{u}^h,p^h)\rrbracket_E\right\|_{0,E}^2.$$

The terms of the second equation for j can be estimated separately by

$$\begin{aligned}
|(\nu\nabla(\mathbf{u}-\mathbf{u}^h),\nabla\mathbf{w}_E)_{0,K}| &\leq C_{E,K,k}h_E^{-\frac{1}{2}}\|\nu\nabla(\mathbf{u}-\mathbf{u}^h)\|_{0,K}\left\|\llbracket\,\mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\rrbracket\right\|_{0,E},\\
|(\nabla\cdot\mathbf{w}_E,p-p^h)_{0,K}| &\leq C_{E,K,k}h_E^{-\frac{1}{2}}\sqrt{d}\|p-p^h\|_{0,K}\left\|\llbracket\,\mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\rrbracket\right\|_{0,E},\\
|(\mathbf{f}_K^h+\nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h),\mathbf{w}_E)_{0,K}| &\leq C_{E,K,k}h_E^{\frac{1}{2}}\|\mathbf{f}_K^h+\nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h)\|_{0,K}\left\|\llbracket\,\mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\rrbracket\right\|_{0,E},\\
|(\mathbf{f}-\mathbf{f}_K^h,\mathbf{w}_E)_{0,K}| &\leq C_{E,K,k}h_E^{\frac{1}{2}}\|\mathbf{f}-\mathbf{f}_K^h\|_{0,K}\left\|\llbracket\,\mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\rrbracket\right\|_{0,E},
\end{aligned}$$

with application of the Cauchy–Schwarz inequality and estimates (58), Lemma 2.12 and (58), (59) and (59) respectively. Applying these estimates to the equations for j yields

$$\begin{aligned}
&C_{E,k}^{-2}h_E^{\frac{1}{2}}\left\|\llbracket\,\mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\rrbracket\right\|_{0,E} \leq \sum_{K\subset\omega_E} C_{E,K,k}\|\nu\nabla(\mathbf{u}-\mathbf{u}^h)\|_{0,K}\\
&\quad + \sqrt{d}\sum_{K\subset\omega_E} C_{E,K,k}\|p-p^h\|_{0,K} + h_E\sum_{K\subset\omega_E} C_{E,K,k}\|\mathbf{f}_K^h+\nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h)\|_{0,K}\\
&\quad + h_E\sum_{K\subset\omega_E} C_{E,K,k}\|\mathbf{f}-\mathbf{f}_K^h\|_{0,K}.
\end{aligned}$$

By applying the previous lower bound estimate (68) one obtains

$$\begin{aligned}
h_E^{\frac{1}{2}}\left\|\llbracket\,\mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\rrbracket_E\right\|_{0,E} \leq C\Bigg(&\|\nu\nabla(\mathbf{u}-\mathbf{u}^h)\|_{0,\omega_E} + \|p-p^h\|_{0,\omega_E}\\
&+ \sum_{K\subset\omega_E} h_K\|\mathbf{f}-\mathbf{f}_K^h\|_{0,K}\Bigg),
\end{aligned} \tag{69}$$

where in the last term h_E has been estimated by a of the triangulation, concrete mesh cells and edges independent constant times h_K, which is possible due to the shape regularity.

(iii) Similarly to (i) and (ii), one can assume the laplace version of $a(\cdot,\cdot)$. Let $E\subset\mathcal{E}_{\Gamma_{\mathrm{nat}}}$ be an arbitrary edge on the natural boundary and

$$(\mathbf{w}_E,0) = ((T_N^h - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E)\psi_E,0)$$

a test function. By taking into account that in this case it is $\operatorname{supp}\mathbf{w}_E \subset \omega_E = K$, one obtains similarly as in (ii)

$$\begin{aligned}(j,(\mathbf{w}_E,0))_{0,E} &= (T_N - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E, \mathbf{w}_E)\\ &= ((T_N^h - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E)^2, \psi_E) + (T_N - T_N^h, \mathbf{w}_E)_{0,K}, \text{ and,}\\ (j,(\mathbf{w}_E,0))_{0,E} &= a(\mathbf{u}-\mathbf{u}^h,\mathbf{w}_E) + b(\mathbf{w}_E, p-p^h) - (r,(\mathbf{w}_E,0))_{0,K}\\ &= (\nu\nabla(\mathbf{u}-\mathbf{u}^h), \nabla\mathbf{w}_E)_{0,K} + (\nabla\cdot\mathbf{w}_E, p-p^h)_{0,K}\\ &\quad + (\mathbf{f}_K^h + \nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h), \mathbf{w}_E)_{0,K} - (\mathbf{f}-\mathbf{f}_K^h)_{0,K},\end{aligned}$$

by definition of j and its relation to the cell part of the residual, respectively. As in (i) and (ii), estimating the terms individually yields

$$\begin{aligned}((T_N^h - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E)^2, \psi_E) &\geq C_{E,k}^{-2}\|T_N^h - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\|_{0,E}^2,\\ |(\nu\nabla(\mathbf{u}-\mathbf{u}^h), \mathbf{w}_E)_{0,K}| &\leq C_{E,K,k}h_E^{-\frac{1}{2}}\|\nu\nabla(\mathbf{u}-\mathbf{u}^h)\|_{0,K}\|T_N^h - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\|_{0,E},\\ |(\nabla\cdot\mathbf{w}_E, p-p^h)_{0,K}| &\leq C_{E,K,k}h_E^{-\frac{1}{2}}\sqrt{d}\|p-p^h\|_{0,K}\|T_N^h - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\|_{0,E},\\ |(\mathbf{f}_K^h + \nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h), \mathbf{w}_E)_{0,K}| &\leq C_{E,K,k}h_E^{\frac{1}{2}}\|\mathbf{f}_K^h + \nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h)\|_{0,K}\|T_N^h - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\|_{0,E},\\ |(\mathbf{f}-\mathbf{f}_K^h, \mathbf{w}_E)_{0,K}| &\leq C_{E,K,k}h_E^{\frac{1}{2}}\|\mathbf{f}-\mathbf{f}_K^h\|_{0,K}\|T_N^h - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\|_{0,E},\\ |(T_N - T_N^h, \mathbf{w}_E)_{0,K}| &\leq \|T_N - T_N^h\|_{0,K}\|T_N^h - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\|_{0,E},\end{aligned}$$

by applying the Cauchy–Schwarz inequality and (57), (58), Lemma 2.12 and (58), (59), (59) and that $0 \leq \psi_E \leq 1$, respectively. Combining these estimates, dividing by $\|T_N^h - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\|_{0,E}$ and multiplication with $h_E^{\frac{1}{2}}$ yields

$$\begin{aligned}C_{E,k}^{-2}h_E^{\frac{1}{2}}\|T_N^h - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\|_{0,E} \leq C\Bigg(&\|\nu(\mathbf{u}-\mathbf{u}^h)\|_{0,K} + \|p-p^h\|_{0,K}\\ &+ h_E\|\mathbf{f}_K^h + \nabla\cdot\mathbb{T}(\mathbf{u}^h,p^h)\|_{0,K} + h_E\|\mathbf{f}-\mathbf{f}_K^h\|_{0,K} + h_E^{\frac{1}{2}}\|T_N - T_N^h\|_{0,K}\Bigg)\end{aligned}$$

Further, one can estimate h_E by h_K like in (ii) and apply estimate (68), yielding

$$\begin{aligned}h_E^{\frac{1}{2}}\|T_N^h - \mathbb{T}(\mathbf{u}^h,p^h)\cdot\mathbf{n}_E\|_{0,E} \leq C\Bigg(&\|\nu(\mathbf{u}-\mathbf{u}^h)\|_{0,K} + \|p-p^h\|_{0,K}\\ &+ h_K\|\mathbf{f}-\mathbf{f}_K^h\|_{0,K} + h_E^{\frac{1}{2}}\|T_N - T_N^h\|_{0,E}\Bigg). \qquad (70)\end{aligned}$$

With use of Lemma 2.12 and $\nabla\cdot\mathbf{u} = 0$ one can obtain

$$\|\nabla\cdot\mathbf{u}^h\|_{0,K} = \|\nabla\cdot(\mathbf{u}-\mathbf{u}^h)\|_{0,K} \leq \sqrt{d}\|\nabla(\mathbf{u}-\mathbf{u}^h)\|_{0,K}.$$

Inserting this estimate together with estimates (68), (69) and (70) into the definition of $\eta_{R,K}$ (67) and taking into account that the inner edges are being counted twice gives the statement of the theorem. □

Theorems 3.21 and 3.22 show that all assumptions of Theorem 3.11 of the abstract setting are fulfilled with

$$Y^h \subset \widetilde{Y}^h = \operatorname{span}\{\psi_K\mathbf{v}\big|_K, \psi_E\mathbf{v}\big|_E : K \in \mathcal{T}, E \in \mathcal{E}, \mathbf{v} \in X = \mathbf{V}\times Q\}.$$

Thus,

$$\eta_{\mathcal{T}} = \left(\sum_{K \in \mathcal{T}} \eta_{R,K}^2 \right)^{\frac{1}{2}}$$

as defined in (65) and (67) bounds the error in the energy norm $\|(\mathbf{v}, q)\|_X = \left(|\mathbf{v}|_1^2 + \|q\|_0^2\right)^{\frac{1}{2}}$ for $(\mathbf{v}, q) \in V \times Q$ from below and above and is therefore reliable and efficient.

3.2 Dual weighted residual method

The previously considered a posteriori error estimates measure the error in a norm which is natural to the problem itself. This however might not be desirable in practice when one is interested in specific quantities like, e.g., drag and lift coefficients of a flow around an obstacle. Further, one usually has unknown constants C which might be far from 1 on the right-hand side of (30) and (31), cf. [Joh00, Section 5].

The dual weighted residual method (also known as DWR method for short) deals with both of these issues. It is a goal-oriented approach since it gives information on the error with respect to a functional of interest. In particular, local residuals are being multiplied with weights which indicate the dependence of the error on them. These weights are obtained by the solution of a linear dual problem. Thus, it is comparably cheap to calculate the weights in the setting of a nonlinear problem and more expensive in the setting of a linear problem.

In the following, the DWR method is going to be derived in an abstract setting based on the analysis presented in [BR03a] and [BR03b]. To this end, let $(X, \|\cdot\|_X)$ be a Banach space and $\mathcal{L} : X \to \mathbb{R}$ be a differential functional. Then, a stationary point $x \in X$ of $\mathcal{L}(\cdot)$ is being sought, i.e.,

$$\mathcal{L}'(x)(y) = 0 \quad \forall y \in X, \tag{71}$$

where the derivative refers to the first parenthesis and L is linear with respect to all arguments in the second parenthesis. This stationary point corresponds to a solution of an optimization problem with the goal to solve a variational problem subject to minimizing the error with respect to a given functional. The discretization is performed by a conforming Galerkin method using a finite-dimensional subspace $X^h \subset X$. The corresponding discrete problem is then to find $x^h \in X^h$ such that

$$\mathcal{L}'(x^h)(y^h) = 0 \quad \forall y^h \in X^h. \tag{72}$$

The goal of the next lemma and proposition is to find a suitable representation of the approximation error in $\mathcal{L}$, i.e., $\mathcal{L}(x) - \mathcal{L}(x^h)$, which is later going to be used in the derivation of an error representation in the setting of a variational problem.

3.23 Lemma [Truncation error in the trapezoidal rule]**.** Let $f : [a, b] \to \mathbb{R}$ be a twice continuously differentiable function, then the trapezoidal rule is given by

$$\int_a^b f \approx (b-a)\frac{f(a) + f(b)}{2}.$$

Its truncation error fulfills the identity

$$\int_a^b f(\tau)d\tau - \frac{f(b) + f(a)}{2}(b-a) = \frac{1}{2}\int_a^b \left(\left(\tau - \frac{a+b}{2}\right)^2 - \left(\frac{b-a}{2}\right)^2 \right) f''(\tau)d\tau.$$

Proof. Integration by parts yields

$$\frac{1}{2}\int\limits_a^b \left(\left(\tau - \frac{a+b}{2}\right)^2 - \left(\frac{b-a}{2}\right)^2\right) f''(\tau) d\tau = \int\limits_a^b \left(\frac{a+b}{2} - \tau\right) f'(\tau) d\tau$$
$$= \int\limits_a^b f(\tau) d\tau - \frac{f(a)+f(b)}{2}(b-a).$$

□

3.24 Proposition [Error representation of the abstract problem]**.** For the Galerkin approximation of the variational problem (71) it is

$$\mathcal{L}(x) - \mathcal{L}(x^h) = \frac{1}{2}\mathcal{L}'(x^h)(x - I^h x) + R, \tag{73}$$

where $e := x - x^h$ is the error, $I^h x \in X^h$ is arbitrary and

$$R := \frac{1}{2}\int\limits_0^1 \tau(\tau-1)\mathcal{L}'''(x^h + \tau e)(e,e,e) d\tau$$

is the remainder and vanishes if $\mathcal{L}(\cdot)$ is quadratic.

Proof. With

$$\mathcal{L}'(\overline{xx^h})(y) := \int\limits_0^1 \mathcal{L}'(x^h + \tau e)(y) d\tau$$

it is by the fundamental theorem of calculus that

$$\mathcal{L}'(\overline{xx^h})(e) = \mathcal{L}(x^h + \tau e)\Big|_{\tau=0}^{\tau=1} = \mathcal{L}(x^h + x - x^h) - \mathcal{L}(x^h) = \mathcal{L}(x) - \mathcal{L}(x^h).$$

Application of the trapezoidal rule to $\mathcal{L}'(\overline{xx^h})(e)$ with $a = 0$ and $b = 1$ and the fact that $e \in X$ and therefore $\mathcal{L}'(x)(e) = 0$ yields

$$\mathcal{L}'(\overline{xx^h})(e) = \frac{\mathcal{L}'(x)(e) + \mathcal{L}'(x^h)(e)}{2} + \frac{1}{2}\int\limits_0^1 \tau(\tau-1)\mathcal{L}'''(x^h + \tau e)(e,e,e) d\tau$$
$$= \frac{1}{2}\mathcal{L}'(x^h)(e) + R.$$

Let $I^h x \in X^h$ be arbitrary. By linearity it then is

$$\mathcal{L}'(x^h)(e) = \mathcal{L}'(x^h)(x - I^h x) + \mathcal{L}'(x^h)(I^h x - x^h) = \mathcal{L}'(x^h)(x - I^h x),$$

since $I^h x - x^h \in X^h$ and therefore $\mathcal{L}'(x^h)(I^h x - x^h) = 0$. □

The above derived abstract problem can now be applied to a standard Galerkin approximation of variational forms. Let

$$a : V \times V \to \mathbb{R},\ a(u)(v) = f(v) \quad \forall v \in V$$

be a semilinear form on V and f a linear functional. Further, let $J : V \to \mathbb{R}$ be a differentiable function which represents the quantity of interest and whose error $J(u) - J(u^h)$ should be minimized. The task of computing $J(u)$ can ultimately also be formulated as the following optimization problem: Find $u \in V$ such that

$$J(u) \to \min, \quad a(u)(v) = f(v) \quad \forall v \in V. \tag{74}$$

If there is a unique solution of the problem to find $u \in V$ such that

$$a(u)(v) = f(v) \quad \forall v \in V,$$

the minimization problem (74) is trivial since there is only one solution that can be inserted into $J(\cdot)$.

For deriving a solution of (74), the Euler–Lagrange approach is being applied, i.e., one considers the Lagrangian functional

$$L(u, z) = J(u) + f(z) - a(u)(z)$$

where $z \in V$ is the so-called adjoint variable. With respect to the abstract setting, the functional L is the functional $\mathcal{L}$. Minima of the optimization problem correspond to stationary points of $L(\cdot, \cdot)$, i.e.,

$$\begin{aligned} 0 = \partial_u L(u, z) &= J'(u)(w) - a'(u)(w, z) \ \forall w \in V, \\ 0 = \partial_z L(u, z) &= f(v) - a(u)(v) \qquad\quad\ \ \forall v \in V, \end{aligned} \tag{75}$$

where the first equation is the adjoint equation.

When applying the Galerkin method in order to obtain a discretized problem, one considers a finite-dimensional subspace $V^h \subset V$ and seeks $(u^h, z^h) \in V^h \times V^h$ such that

$$\begin{aligned} a'(u^h)(w^h, z^h) &= J'(u^h)(w^h) \quad \forall w^h \in V^h, \\ a(u^h)(v^h) &= f(v^h) \qquad\quad\ \forall v^h \in V^h. \end{aligned} \tag{76}$$

With the solution of the discrete problem above, one can define the the so-called primal residual

$$\rho : V \to V^*, \quad \rho(u^h)(\cdot) = f(\cdot) - a(u^h)(\cdot), \tag{77}$$

and associated to z^h the adjoint or dual residual

$$\rho^* : V \to V^*, \quad \rho^*(z^h)(\cdot) = J'(u^h)(\cdot) - a'(u^h)(\cdot, z^h), \tag{78}$$

together with the adjoint error $e^* := z - z^h$.

3.25 Theorem [Error representation]**.** Let $a(\cdot)(\cdot)$ and $J(\cdot)$ be three times differentiable, (u, z) a solution of (74) and (u^h, z^h) a solution of (76). Then there is the a posteriori error representation

$$J(u) - J(u^h) = \frac{1}{2}\rho(u^h)(z - I^h z) + \frac{1}{2}\rho^*(z^h)(u - I^h u) + R, \tag{79}$$

where $I^h z \in V^h$ and $I^h u \in V^h$ are arbitrary. The remainder term is given by

$$R := \frac{1}{2} \int_0^1 \tau(\tau - 1)\big(J'''(u^h + \tau e)(e, e, e) - a'''(u^h + \tau e)(e, e, e, z^h + \tau e^*) \\ - 3a''(u^h + \tau e)(e, e, e^*)\big) d\tau.$$

Proof. This proof is based on the application of Proposition 3.24 in the general framework. To this end, set $X = V \times V$, $X^h = V^h \times V^h$, $x = (u, z)$, $x^h = (u^h, z^h)$ and $\mathcal{L}(x) = L(u, z)$.

Let $(u, z) \in V \times V$ and $(u^h, z^h) \in V^h \times V^h$ be the solutions of (75) and (76), respectively. Since these are stationary points, the requirements for Proposition 3.24 are fulfilled. Further, note that

$$\begin{aligned} \mathcal{L}(x) - \mathcal{L}(x^h) &= J(u) + f(z) - a(u)(z) - (J(u^h) + f(z^h) - a(u^h)(z^h)) \\ &= J(u) - J(u^h). \end{aligned}$$

Application of the proposition yields then that

$$\begin{aligned} J(u) - J(u^h) &= \frac{1}{2}\mathcal{L}'(x^h)(x - I^h x) + R \\ &= \frac{1}{2}(J'(u^h)(u - I^h u) - a'(u^h)(u - I^h u, z^h) + f(z - I^h z) - a(u^h)(z - I^h z)) + R \\ &= \frac{1}{2}\rho(z - I^h z) + \frac{1}{2}\rho^*(u - I^h u) + R, \end{aligned}$$

where $I^h x = (I^h u, I^h z) \in X^h$ is arbitrary. In order to compute the remainder

$$R = \frac{1}{2} \int_0^1 \tau(\tau - 1)\mathcal{L}'''(x^h + \tau e)(e, e, e) d\tau,$$

one needs to calculate the third derivative of $\mathcal{L}$. Since the dependence of $\mathcal{L}(x) = L(u, z)$ on z is linear, second order and higher derivatives with respect to u vanish and one obtains

$$\mathcal{L}''' = \partial_{uuu}\mathcal{L} + 3\partial_{uuz}\mathcal{L} + \partial_{uzz}\mathcal{L} + \partial_{zzz}\mathcal{L}.$$

This yields

$$\begin{aligned} &\mathcal{L}'''(x^h + \tau(e, e^*))((e, e^*), (e, e^*), (e, e^*)) \\ &\quad = J'''(u^h + \tau e) - a'''(u^h + \tau e)(e, e, e, z^h + \tau e^*) - 3a''(u^h + \tau e)(e, e, e^*) \end{aligned}$$

and therefore the statement. □

The remainder R in (79) is cubic in the errors e and e^* and therefore usually can be neglected. Neglecting it results in the error indicator

$$\eta(u^h, z^h) := \frac{1}{2}\rho(u^h)(z - I^h z) + \frac{1}{2}\rho^*(z^h)(u - I^h u). \tag{80}$$

Its evaluation however requires the exact primal and dual solutions u and z.

In the following a simplified error representation is developed which involves only one residual. To this end, the next proposition considers a term which reflects the difference between the primal and dual residual. This difference term is then used to absorb the dual residual into a remainder term with one order less.

3.26 Proposition. Let the primal and dual residuals be given as in (77) and (78) respectively, then it is

$$\rho^*(z^h)(u - I^h u) = \rho(u^h)(z - I^h z) + \Delta\rho, \tag{81}$$

where $I^h u$, $I^h z \in V^h$ are arbitrary and

$$\Delta\rho = \int\limits_0^1 \Big(a''(u^h + \tau e)(e, e, z^h + \tau e^*) - J''(u^h + \tau e)(e, e)\Big)\, d\tau.$$

Proof. First, define

$$g(\tau) := J'(u^h + \tau e)(u - I^h u) - a'(u^h + \tau e)(u - I^h u, z^h + \tau e^*).$$

Using the definitions $e = u - u^h$ and $e^* = z - z^h$ as well as the first equation of the optimization problem (75), one obtains

$$g(1) = J'(u)(u - I^h u) - a'(u)(u - I^h u, z) = 0.$$

On the other hand it is

$$g(0) = J'(u^h)(u - I^h u) - a'(u^h)(u - I^h u, z^h) \overset{(78)}{=} \rho^*(z^h)(u - I^h u).$$

By using the linearity of the latter arguments of $a'(\cdot)(\cdot, \cdot)$ and the definition of functional derivatives, the derivative of $g(\tau)$ can be obtained by

$$\begin{aligned}
g'(\tau) &= \lim_{h\to 0} \frac{g(\tau + h) - g(\tau)}{h} \\
&= J''(u^h + \tau e)(e, u - I^h u) \\
&\quad - \lim_{h\to 0} \frac{a'(u^h + (\tau + h)e)(u - I^h u, z^h + (\tau + h)e^*) - a'(u^h + \tau e)(u - I^h u, z^h + \tau e^*)}{h} \\
&= J''(u^h + \tau e)(e, u - I^h u) \\
&\quad - \lim_{h\to 0} \frac{a'(u^h + (\tau + h)e)(u - I^h u, z^h + \tau e^*) - a'(u^h + \tau e)(u - I^h u, z^h + \tau e^*)}{h} \\
&\quad - \lim_{h\to 0} \frac{a'(u^h + (\tau + h)e)(u - I^h u, h e^*)}{h} \\
&= J''(u^h + \tau e)(e, u - I^h u) - a''(u^h + \tau e)(e, u - I^h u, z^h + \tau e^*) \\
&\quad - a'(u^h + \tau e)(u - I^h u, e^*).
\end{aligned}$$

Applying the fundamental theorem of calculus yields

$$\begin{aligned}
\rho^*(z^h)(u - I^h u) &= \int\limits_1^0 g'(\tau) d\tau \\
&= \int\limits_0^1 \Big(a''(u^h + \tau e)(e, u - I^h u, z^h + \tau e^*) - J''(u^h + \tau e)(e, u - I^h u)\Big)\, d\tau \\
&\quad + \int\limits_0^1 a'(u^h + \tau e)(u - I^h u, e^*) d\tau,
\end{aligned}$$

where the last term is the primal residual, since

$$\int_0^1 a'(u^h+\tau e)(u-I^hu, e^*)d\tau = a(u^h+e)(e^*) - a(u^h)(e^*) = a(u)(e^*) - a(u^h)(e^*)$$

$$= f(e^*) - a(u^h)(e^*) = \rho(u^h)(z-z^h).$$

Since it was initially assumed that $f(\cdot)$ is linear and $\rho(u^h)(v^h) = 0$ for all $v^h \in V^h$, one obtains

$$\rho(u^h)(z-z^h) = f(z) - a(u^h)(z) - \underbrace{(f(z^h) - a(u^h)(z^h))}_{=0} = \rho(u^h)(z - I^hz)$$

with an arbitrary $I^hz \in V^h$ and thus altogether

$$\begin{aligned}\rho^*(z^h)(u-I^hu) &= \rho(u^h)(z-I^hz)\\ &\quad + \int_0^1 \Big(a''(u^h+\tau e)(e, u-I^hu, z^h+\tau e^*) - J''(u^h+\tau e)(e, u-I^hu)\Big)\, d\tau.\end{aligned}$$

It remains to show that the last term is the remainder $\Delta\rho$. By applying the first equation of the optimization problem (75), one obtains

$$\begin{aligned}&\int_0^1 \Big(a''(u^h+\tau e)(e, u-I^hu, z^h+\tau e^*) - J''(u^h+\tau e)(e, u-I^hu)\Big)\, d\tau\\ &= \Delta\rho + \int_0^1 \Big(a''(u^h+\tau e)(e, u^h-I^hu, z^h+\tau e^*) - J''(u^h+\tau e)(e, u^h-I^hu)\Big)\, d\tau\\ &= \Delta\rho + \Big(a'(u)(u^h-I^hu, z) - J'(u)(u^h-I^hu)\Big)\\ &\quad - \Big(a'(u^h)(u^h-I^hu, z^h) - J'(u^h)(u^h-I^hu)\Big)\\ &\overset{(75)}{=} \Delta\rho + 0 - 0 = \Delta\rho.\end{aligned}$$

□

3.27 Theorem [Simplified error representation]**.** Let the primal and dual residuals be given as in (77) and (78) respectively, then there is the a posteriori error representation

$$J(u) - J(u^h) = \rho(u^h)(z - I^hz) + R \tag{82}$$

with $I^hz \in V^h$ arbitrary and a quadratic remainder

$$R = \int_0^1 \Big(a''(u^h+\tau e)(e,e,z) - J''(u^h+\tau e)(e,e)\Big)\, \tau d\tau.$$

Proof. Using the in Theorem 3.25 derived error representation together with the in Proposition 3.26 derived connection between the residuals yields

$$J(u) - J(u^h) = \frac{1}{2}\rho(u^h)(z-I^hz) + \frac{1}{2}\rho^*(z^h)(u-I^hu) + \widetilde{R} = \rho(u^h)(z-I^hz) + \frac{1}{2}\Delta\rho + \widetilde{R},$$

where $\widetilde{R}$ denotes the remainder of the error representation of Theorem 3.25. It remains to show that $R = \widetilde{R} + \frac{1}{2}\Delta\rho$. To this end, consider the terms separately. Integration by parts yields

$$\begin{aligned}
\Delta\rho &= \Big[\tau\left(a''(u^h+\tau e)(e,e,z^h+\tau e^*) - J''(u^h+\tau e)(e,e)\right)\Big]_{\tau=0}^{\tau=1} \\
&\quad - \int\limits_0^1 \Big(a'''(u^h+\tau e)(e,e,e,z^h+\tau e^*) - J'''(u^h+\tau e)(e,e,e) + a''(u^h+\tau e)(e,e,e^*)\Big)\tau d\tau \\
R &= \left[\frac{1}{2}\tau^2\left(a''(u^h+\tau e)(e,e,z) - J''(u^h+\tau e)(e,e)\right)\right]_{\tau=0}^{\tau=1} \\
&\quad - \frac{1}{2}\int\limits_0^1 \left(a'''(u^h+\tau e)(e,e,z) - J'''(u^h+\tau e)(e,e,e)\right)\tau^2 d\tau.
\end{aligned}$$

When considering $R - \widetilde{R} - \frac{1}{2}\Delta\rho$, the first term of R cancels with the first term of $\frac{1}{2}\Delta\rho$. Simplifying the expression by collecting like terms, using the linearity of the latter arguments of $a'''(\cdot)(\cdot,\cdot,\cdot,\cdot)$ and the definition of e^* yields

$$\begin{aligned}
R - \widetilde{R} - \frac{1}{2}\Delta\rho = \frac{1}{2}\Bigg(&\int\limits_0^1 \left(a'''(u^h+\tau e)(e,e,e,\tau e^* - e^*) + 3a''(u^h+\tau e)(e,e,e^*)\right)\tau^2 d\tau \\
&- 2\int\limits_0^1 a''(u^h+\tau e)(e,e,e^*)\tau d\tau\Bigg).
\end{aligned}$$

Applying integration by parts once more and using linearity finally gives

$$\begin{aligned}
&R - \widetilde{R} - \frac{1}{2}\Delta\rho \\
&= \frac{1}{2}\Bigg(- \Big[\tau^2 a''(u^h+\tau e)(e,e,e^*)\Big]_{\tau=0}^{\tau=1} + \int\limits_0^1 a'''(u^h+\tau e)(e,e,e,e^*)\tau^2 d\tau \\
&\quad + \Big[\tau^3 a''(u^h+\tau e)(e,e,e^*)\Big]_{\tau=0}^{\tau=1} - \int\limits_0^1 a'''(u^h+\tau e)(e,e,e,e^*)\tau^3 d\tau \\
&\quad + \int\limits_0^1 a'''(u^h+\tau e)(e,e,e,\tau e^* - e^*)\tau^2 d\tau\Bigg) \\
&= \frac{1}{2}\Bigg(- \int\limits_0^1 a'''(u^h+\tau e)(e,e,e,e^*)\tau^3 d\tau + \int\limits_0^1 a'''(u^h+\tau e)(e,e,e,\tau e^*)\tau^2 d\tau\Bigg) \\
&= 0
\end{aligned}$$

and therefore the statement of the theorem. □

3.28 Remark [On the remainder]**.** When considering a linear functional $J(\cdot)$ and a linear variational equation, the remainder term of the simplified error representation and $\Delta\rho$ vanish.

If this is not the case, the remainder is still quadratic in e and thus can be assumed to be relatively small. Also, the two residuals do not coincide anymore and their difference $\Delta\rho$ indicates the influence of the nonlinearity on the error.

Neglecting the remainder gives the simplified dual weighted residual error estimator

$$\widehat{\eta}(u^h) = \rho(u^h)(z - I^h z). \tag{83}$$

3.29 Remark [Application to a linear problem with a linear functional]**.** In case of a linear problem $a(\cdot)(\cdot) = a(\cdot, \cdot)$ it is

$$a'(u^h)(\cdot, z^h) = a(\cdot, z^h).$$

If the functional is linear as well, it is

$$J'(u^h)(\cdot) = J(\cdot).$$

Considering Proposition 3.26, one can see that the remainder term Δp vanishes if both $a(\cdot)(\cdot)$ and $J(\cdot)$ are linear and therefore the primal residual and dual residual coincide. Applying this equality of residuals to the error representation (79) yields

$$J(u) - J(u^h) = J(e) = \rho(u^h)(z - I^h z),$$

where the remainder vanishes, since all higher order derivatives of the variational form and the functional vanish.

3.2.1 Application to the Darcy equations

This section applies the results of Section 3.2 to the Darcy equations. Since the problem is linear, the dual problem and its conforming Galerkin discretization are to find $z \in V$ and $z \in V^h$ respectively, such that

$$a(w, z) = J'(u)(w) \qquad \forall w \in V, \tag{84}$$

$$a(w^h, z^h) = J'(u^h)(w^h) \qquad \forall w^h \in V^h, \tag{85}$$

where $a(\cdot, \cdot)$ is given as in (7).

3.30 Theorem [DWR error indicators for the Darcy equations]**.** For the simplified error indicator (83) one obtains in the setting of the Darcy equations that

$$|J(u) - J(u^h)| \le \eta_{\mathcal{T}} := \sum_{K \in \mathcal{T}} \eta_{\mathrm{DWR},K} = \sum_{K \in \mathcal{T}} \rho_K \omega_K$$

with

$$\rho_K := \left(\|r\|_K^2 + \frac{1}{4h_K} \|j\|_{\partial K}^2 \right)^{\frac{1}{2}},$$

$$\omega_K := \left(\|z - I^h z\|_K^2 + h_K \|z - I^h z\|_{\partial K}^2 \right)^{\frac{1}{2}}.$$

Assuming that the functional $J'(\cdot)(\cdot)$ can be written in the form of an integral over the domain, i.e.,

$$J'(\cdot)(\cdot) = \sum_{K \in \mathcal{T}} \int_K j_K(\cdot)(\cdot),$$

one obtains for the error indicator (80) that

$$|J(u) - J(u^h)| \leq \widetilde{\eta}_{\mathcal{T}} := \sum_{K \in \mathcal{T}} \left(\frac{1}{2} \eta_{\mathrm{DWR},K} + \frac{1}{2} \eta^*_{\mathrm{DWR},K} \right) = \sum_{K \in \mathcal{T}} \left(\frac{1}{2} \rho_K \omega_K + \frac{1}{2} \rho^*_K \omega^*_K \right)$$

with $\eta_{\mathrm{DWR},K}$ as in the case of the simplified error indicator and

$$\rho^*_K = \left(\left\| j_K(u^h)(u - I^h u) + K \Delta z^h \right\|^2_K + \frac{1}{4 h_K} \left\| \sum_{E \in \mathcal{E}_{\partial K}} [\![\, \mathbb{K} \nabla z^h \cdot \mathbf{n}_E]\!]_E \right\|^2_{\partial K} \right)^{\frac{1}{2}},$$

$$\omega^*_K = \left(\|u - I^h u\|^2_K + \|u - I^h u\|^2_{\partial K} \right)^{\frac{1}{2}}.$$

Proof. Using the L^p representation of the residual which was derived in Section 3.1.1 and the triangle inequality, one obtains for the error indicator (83) of the abstract framework that

$$|J(u) - J(u^h)| = |\rho(u^h)(z - I^h z)| \leq \sum_{K \in \mathcal{T}} \left| \left(r\big|_K, z - I^h z \right)_K + \frac{1}{2} \left(j\big|_{\partial K}, z - I^h z \right)_{\partial K} \right|.$$

The factor of $\frac{1}{2}$ takes into account, that every facet of K is being considered twice in the sum. Application of the Cauchy–Schwarz inequality and the Cauchy–Schwarz inequality for sums yields the statement for the simplified error indicator.

The second part of the theorem requires to develop a L^p representation of the dual residual. To this end, let $w^h \in V^h$ be a test function and apply cell-wise integration by parts to the bilinear form, i.e.,

$$\begin{aligned} a(w^h, z^h) &= (\mathbb{K} \nabla w^h, \nabla z^h)_0 = \int_\Omega (\nabla z^h)^T \, \mathbb{K} \nabla w^h \\ &= \int_\Omega (\mathbb{K}^T \nabla z^h)^T \nabla w^h \\ &= \sum_{K \in \mathcal{T}} \left(\int_{\partial K} \mathbb{K} \nabla z^h w^h \cdot \mathbf{n}_K - \int_K \nabla \cdot (\mathbb{K}^T \nabla z^h) w^h \right) \qquad (86) \\ &= \sum_{K \in \mathcal{T}} \left(\int_{\partial \Omega} K \nabla z^h w^h \cdot \mathbf{n}_K - \int_K K \Delta z^h w^h \right) \qquad \text{, with } \mathbb{K} = K \, \mathbb{I}, \end{aligned}$$

which yields for the dual residual that

$$\begin{aligned} \rho^*(z^h)(w^h) &= J'(u^h)(w^h) - a'(u^h)(\cdot, z^h) \\ &= J'(u^h)(w^h) - a(w^h, z^h) \qquad \text{, by linearity,} \\ &= J'(u^h)(w^h) + \sum_{K \in \mathcal{T}} \left(\int_K K \Delta z^h w^h - \int_{\partial K} \mathbb{K} \nabla z^h w^h \cdot \mathbf{n}_K \right). \end{aligned}$$

Using the assumed representation of $J'(\cdot)(\cdot)$, one can restrict $\rho^*(z^h)(w^h)$ to a single cell $K \in \mathcal{T}$ by

$$\rho^*\big|_K(z^h)(w^h) = (j_K(u^h)(w^h) + K \Delta z^h, w^h)_K + \sum_{E \in \partial K} ([\![\, \mathbb{K} \nabla z^h \cdot \mathbf{n}_E]\!]_E, w^h)_E.$$

Applying the arguments of the first part of the theorem yields the statement. □

3.31 Remark. As one can see, the structure of the error indicator is the norm of the residual times some weighting factor depending on the solution of the dual problem. However, for practical use it is suggested in [BR03b, Remark 3.2] not to evaluate the norms but to evaluate the scalar products directly, e.g.,

$$\eta_{\mathcal{T}} = \sum_{K \in \mathcal{T}} \left| \left(r|_K, z - I^h z \right)_K + \frac{1}{2} \left(j|_{\partial K}, z - I^h z \right)_{\partial K} \right|,$$

since further estimation leads to less sharp estimates.

Further, equation (86) was simplified by using the assumption that $\mathbb{K} = K\,\mathbb{I}$. In general, it can only be assumed that $\mathbb{K}$ is positive definite and one obtains

$$\rho_K^* = \left(\left\| j_K(u^h)(u - I^h u) + \nabla \cdot (\mathbb{K}^T \nabla z^h) \right\|_K^2 + \frac{1}{4h_K} \left\| \sum_{E \in \mathcal{E}_{\partial K}} [\![\, \mathbb{K} \nabla z^h \cdot \mathbf{n}_E]\!]_E \right\|_{\partial K} \right)^{\frac{1}{2}}.$$

Also, it was assumed that the functional $J'(\cdot)(\cdot)$ can be written in the form of an integral over the domain. This assumption is especially fulfilled if $J(\cdot)$ is linear and continuous, since then $J'(\cdot)(\cdot) = J(\cdot)$ and the Riesz representation theorem gives the existence of a $j \in V$ such that $J(w) = (j, w)_V = (j, w)_0$ for all $w \in V$.

3.2.2 Application to the Stokes equations

This section applies the results of Section 3.2 to the Stokes equations. Let $\boldsymbol{\varphi} = (\mathbf{u}, p) \in \mathbf{V} \times Q$, $\boldsymbol{\varphi}^h = (\mathbf{u}^h, p^h) \in \mathbf{V}^h \times Q^h$ be the solution and the solution of the Galerkin discretization respectively. As the Darcy equations, they yield a linear problem and thus the continuous and discretized dual problems are to find $\boldsymbol{\psi} = (\mathbf{z}, r) \in \mathbf{V} \times Q$ and $\boldsymbol{\psi}^h = (\mathbf{z}^h, r^h) \in \mathbf{V}^h \times Q^h$ respectively, such that

$$\begin{aligned} a(\mathbf{w}, \mathbf{z}) + b(\mathbf{z}, s) - b(\mathbf{w}, r) &= J'(\boldsymbol{\varphi})((\mathbf{w}, s)) \qquad && \forall (\mathbf{w}, s) \in \mathbf{V} \times Q, \\ a(\mathbf{w}^h, \mathbf{z}^h) + b(\mathbf{z}^h, s^h) - b(\mathbf{w}^h, r^h) &= J'(\boldsymbol{\varphi}^h)((\mathbf{w}^h, s^h)) \qquad && \forall (\mathbf{w}^h, w^h) \in \mathbf{V}^h \times Q^h, \end{aligned}$$

where the bilinear forms $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are given as in (16). Since it is $a(\cdot, \cdot)$ is symmetric, one obtains

$$\begin{aligned} a(\mathbf{z}, \mathbf{w}) - b(-\mathbf{z}, s) + b(\mathbf{w}, -r) &= J'(\boldsymbol{\varphi})((\mathbf{w}, s)) \qquad && \forall (\mathbf{w}, s) \in \mathbf{V} \times Q, \\ a(\mathbf{z}^h, \mathbf{w}^h) - b(-\mathbf{z}^h, s^h) + b(\mathbf{w}^h, -r^h) &= J'(\boldsymbol{\varphi}^h)((\mathbf{w}^h, s^h)) \qquad && \forall (\mathbf{w}^h, s^h) \in \mathbf{V}^h \times Q^h, \end{aligned}$$

which is the weak formulation of

$$\begin{cases} -\nabla \cdot \mathbb{T}(\mathbf{z}, -r) = J_v'(\boldsymbol{\varphi}) \\ \nabla \cdot (-\mathbf{z}) = \nabla \cdot \mathbf{z} = J_p'(\boldsymbol{\varphi}) \end{cases}, \quad \begin{cases} -\nabla \cdot \mathbb{T}(\mathbf{z}^h, -r^h) = J_v'(\boldsymbol{\varphi}^h) \\ \nabla \cdot (-\mathbf{z}^h) = \nabla \cdot \mathbf{z}^h = J_p'(\boldsymbol{\varphi}^h) \end{cases},$$

with an L^2 representation $\{J_v', J_p'\}$ of the functional $J'(\cdot)$, i.e., again a (possibly compressible) Stokes problem but with negated pressure field. Since flow is always directed from areas with high pressure to areas with low pressure, the dual problem reverses the direction of the flow field.

3.32 Theorem [DWR error indicators for the Stokes equations]**.** For the simplified error indicator (83) one obtains in the setting of the Stokes equations that

$$|J(\boldsymbol{\varphi}) - J(\boldsymbol{\varphi}^h)] \leq \eta_{\mathcal{T}} := \sum_{K\in\mathcal{T}} \eta_{\mathrm{DWR},K} = \sum_{K\in\mathcal{T}} \rho_K\omega_K$$

with

$$\rho_K := \left(\|r\|_K^2 + \frac{1}{4h_K}\|j\|_{\partial K}^2\right)^{\frac{1}{2}},$$

$$\omega_K := \left(\|\boldsymbol{\psi} - I^h\boldsymbol{\psi}\|_K^2 + h_K\|\boldsymbol{\psi} - I^h\boldsymbol{\psi}\|_{\partial K}^2\right)^{\frac{1}{2}},$$

where r and j are given as in the L^p representation of the residual (64).

Assuming that the functional $J'(\cdot)(\cdot)$ can be written in the form of an integral over the domain, i.e.,

$$J'(\cdot)(\cdot) = \sum_{K\in\mathcal{T}}\int_K j_K(\cdot)(\cdot),$$

one obtains for the error indicator (80) that

$$|J(\boldsymbol{\varphi}) - J(\boldsymbol{\varphi}^h)| \leq \widetilde{\eta}_{\mathcal{T}} := \sum_{K\in\mathcal{T}}\left(\frac{1}{2}\eta_{\mathrm{DWR},K} + \frac{1}{2}\eta^*_{\mathrm{DWR},K}\right) = \sum_{K\in\mathcal{T}}\left(\frac{1}{2}\rho_K\omega_K + \frac{1}{2}\rho^*_K\omega^*_K\right)$$

with $\eta_{\mathrm{DWR},K}$ given as in the case of the simplified error indicator and

$$\rho^*_K = \Bigg(\left\|\left(j_K(\boldsymbol{\varphi}^h)(\boldsymbol{\varphi} - I^h\boldsymbol{\varphi}) + \nabla\cdot\mathbb{T}(\mathbf{z}^h, -r^h), \nabla\cdot\mathbf{z}^h\right)\right\|_K^2$$

$$+ \frac{1}{4h_K}\sum_{E\in\mathcal{E}_{\partial K}}\left\|-[\![\, \mathbb{T}(\mathbf{z}^h, -r^h)\cdot\mathbf{n}_E]\!]_E\right\|_{\partial E}^2\Bigg)^{\frac{1}{2}},$$

$$\omega^*_K = \left(\|\boldsymbol{\varphi} - I^h\boldsymbol{\varphi}\|_K^2 + h_K\|\boldsymbol{\varphi} - I^h\boldsymbol{\varphi}\|_{\partial K}^2\right)^{\frac{1}{2}}.$$

Proof. The proof can be performed analogously to the proof of Theorem 3.30. To this end, use the L^p representation of the residual which was derived in Section 3.1.2 and apply the triangle inequality to the error indicator (83) of the abstract framework, yielding

$$|J(\boldsymbol{\varphi}) - J(\boldsymbol{\varphi}^h)| = |\rho(\boldsymbol{\varphi}^h)(\boldsymbol{\psi} - I^h\boldsymbol{\psi})| \leq \sum_{K\in\mathcal{T}}\left|\left(r\big|_K, \boldsymbol{\psi} - I^h\boldsymbol{\psi}\right)_K + \frac{1}{2}\left(j\big|_{\partial K}, \boldsymbol{\psi} - I^h\boldsymbol{\psi}\right)_{\partial K}\right|.$$

The factor of $\frac{1}{2}$ takes into account, that every facet of K is being considered twice in the sum. Application of the Cauchy–Schwarz inequality and the Cauchy–Schwarz inequality for sums yields the statement for the simplified error indicator.

Since the dual problem is again a problem of the Stokes kind, one can replace the solution $(\mathbf{u}, p)$ by $(\mathbf{z}, -r)$, the solution of the discretized problem $(\mathbf{u}^h, p^h)$ by $(\mathbf{z}^h, -r^h)$ and then repeat the first part of the proof. □

3.33 Remark. As in Remark 3.31, it might be of advantage to evaluate the scalar products directly since further estimation leads to less sharp estimates. In the case of a linear and continuous functional $J(\cdot)$, the form assumed in the theorem of an integral over the domain exists.

3.2.3 Practical evaluation of the error estimators

Evaluating the error estimators possesses the problem that either the solution of the continuous dual problem z or even both, z and the solution of the continuous primal problem u have to be known, which in general is not the case. In the following, the practical evaluation of the simplified error estimator $\eta_{\mathcal{T}}$ will be discussed, therefore only z is replaced by some suitable discrete approximation $\widetilde{z}$. There are three main approaches to this task:

(i) *Approximation by a higher-order method.* In this approach one replaces z by a solution $\widetilde{z}$ to the dual problem on a finer grid or with a higher order finite element method. This however means, that the computation of the dual, auxiliary problem in order to apply the estimator in fact dominates the actual problem with respect to computational costs. An approximation can be achieved by, e.g., solving the dual problem on a coarser grid with higher order finite elements, i.e., $\widetilde{z} = I_{2h}^{h} z_{(2)}^{h}$, where $z_{(2)}^{h}$ denotes the solution of the discretized problem.

(ii) *Approximation by higher-order interpolation.* Another way of obtaining an approximation of z that does not require modifying the finite element spaces or changing the grid is based on interpolation. If the order of the finite elements is p, one tries to get an improved approximation by patch-wise interpolation of order $p' > p$ with the interpolation nodes being the vertices on a patch which usually consists of four elements in the case of quadrilaterals. If the grid contains hanging nodes or nonconforming finite elements have been used, special care is required to preserve the higher-order accuracy of the interpolation. In this case the function z^h is discontinuous on the vertices and it is suggested in [GHT04, Section 3] to apply the higher-order interpolation once to get continuity and another time for actual interpolation purposes, i.e., $\widetilde{z} = I_{\text{interp}}^{2} z^h$ where I_{interp} denotes the interpolation.

(iii) *Approximation by difference quotients.* Using the representation $|J(u) - J(u^h)| = \sum_{K \in \mathcal{T}} \rho_K \omega_K$ and the nodal linear interpolation fulfilling $I^h z(x) = z(x)$ for all $x \in \mathcal{N}$, one can apply the following interpolation estimates to the weights ω_K.

3.34 Proposition [Interpolation estimates]**.** Assuming that $z \in H^2(\Omega)$, the interpolation estimates

$$\begin{aligned} \|z - I^h z\|_{0,K}^2 &\leq C h_K^4 \|\Delta z\|_{0,K}^2, \\ h_K \|z - I^h z\|_{0,\partial K}^2 &\leq C h_K^4 \|\Delta z\|_{0,K}^2 \end{aligned}$$

for $K \in \mathcal{T}$ hold.

Proof. In [Cia91, Theorem 16.2] it is stated that for s being the highest order partial derivative in the definition of the degrees of freedom and $m, k \in \mathbb{N} \cup \{0\}$, $p, q \in [1, \infty]$ such that the inclusions

$$\begin{aligned} W^{k+1,p}(\widehat{K}) &\hookrightarrow C^s(\widehat{K}), \\ W^{k+1,p}(\widehat{K}) &\hookrightarrow W^{m,q}(\widehat{K}), \\ P_k(\widehat{K}) \subset P(K) &\subset W^{m,q}(\widehat{K}) \end{aligned}$$

hold, one obtains the interpolation estimate

$$\|v - I^h v\|_{W^{m,q}(K)} \leq C h_K^{k+1-m} |v|_{W^{k+1,p}(K)}$$

for all $v \in W^{k+1,p}(K)$. Here, $\widehat{K}$ is the reference element corresponding to $K \in \mathcal{T}$, $W^{k,p}(\widehat{K})$ denotes the Sobolev space containing functions whose weak derivatives up to order k are contained in $L^p(\widehat{K})$, $P_k(\widehat{K})$ is the space of polynomials of degree $\leq k$ and $P(K)$ is defined as in (9). For $m = 0$, $k = 1$ and $p = q = 2$ the first part of the statement is yielded directly.

For the second part of the statement the inverse estimate [Cia91, Equation (17.23)] is needed, which states that for $h = \max_{K\in\mathcal{T}} h_K$ one obtains

$$\|v - I^h v\|_{0,\partial K} \leq Ch^{-\frac{1}{2}}\|v - I^h v\|_{0,K}.$$

Applying the first part of the proof yields

$$h\|v - I^h v\|_{0,\partial K}^2 \leq Ch_K^4\|\Delta v\|_{0,K}^2$$

and with $h \geq h_K$ for all $K \in \mathcal{T}$ the second part of the statement. □

Application of the above proposition to ω_K then yields $\omega_K \leq Ch_K^2\|\Delta z\|_K$. Now the second derivatives Δz can be replaced by a second order difference quotient $\Delta^h z^h$. This yields

$$\rho_K\omega_K \leq C\|\Delta^h z^h\|_K \left(h_K^4\|r\|_K^2 + \frac{1}{4}h_K^3\|j\|_{\partial K}^2\right)^{\frac{1}{2}}$$

and therefore with the Cauchy–Schwarz inequality the error estimator

$$\eta_{\mathcal{T}} = \left(\sum_{K\in\mathcal{T}} \left\|\Delta^h z^h\right\|_K^2 \left(h_K^4\|r\|_K^2 + \frac{1}{4}h_K^3\|j\|_{\partial K}^2\right)\right)^{\frac{1}{2}}.$$

In [Ver13, Section 1.11] it is also noted that if the estimate's domain K is convex, the regularity estimate $\|\Delta z\|_{0,K} \leq C$ holds for some known constant C, yielding with the Cauchy–Schwarz inequality the error indicator

$$\eta_{\mathcal{T}} = \left(\sum_{K\in\mathcal{T}} h_K^4\rho_K^2\right)^{\frac{1}{2}}.$$

This means that, compared to the residual based error estimators (53) and (65), which estimate the error in the energy norm, one gains an additional factor of h_K cell-wise.

3.35 Remark [On the quality of the evaluation]**.** The quality of the error estimator can be measured with a modified version of the efficiency index (32), which reads

$$I_{\text{eff}}^{\text{DWR}} := \left|\frac{\eta_{\mathcal{T}}}{J(u) - J(u^h)}\right|. \qquad (87)$$

It represents the degree over-estimation and should be close to one. In [BR03a, Section 4.1] it is reported that the approximation by a higher-order method as well as the approximation by a higher-order interpolation yield an efficiency index close to one with decreasing tolerance. The approximation by difference quotients is reported to cause over-estimation, i.e., $I_{\text{eff}}^{\text{DWR}} \gg 1$.

4 Cell marking strategies

In the previous sections models for flow problems and corresponding a posteriori error estimates have been derived. In virtue of the adaptive process as described by Algorithm 1.1, this sections deals with part (iv).(a), i.e., strategies determining, which cells are to be refined based on the previously derived error estimates. The two most popular approaches, cf. [Ver13, Section 2.1.1], are the so-called "Maximum strategy" and the "Equilibration strategy" or "Dörfler marking".

The maximum strategy takes the maximal value of η_K, $K \in \mathcal{T}$, and marks every cell with an error that is higher than a prescribed percentage of that maximal value.

4.1 Algorithm [Maximum strategy, see [Ver13, Algorithm 2.1]]**.** *Given*: A partition $\mathcal{T}$, error indicators η_K for the elements $K \in \mathcal{T}$ and a threshold $\theta \in (0,1)$.

Sought: A subset $\widetilde{\mathcal{T}} \subset \mathcal{T}$ of marked elements that should be refined.

(i) Compute $\eta_{\mathcal{T},\max} = \max_{K\in\mathcal{T}} \eta_K$.

(ii) For each $K \in \mathcal{T}$: If $\eta_K \geq \theta\eta_{\mathcal{T},\max}$, mark K for refinement and put it into the set $\widetilde{\mathcal{T}}$.

Instead of taking the maximal η_K as reference value, one could also take the total error $\eta_{\mathcal{T}} = \left(\sum_{K\in\mathcal{T}} \eta_K^2\right)^{\frac{1}{2}}$ as reference value and mark the cells $K \in \mathcal{T}$ with $\eta_K \geq \theta\eta_{\mathcal{T}}$ for some $\theta \in (0,1)$. For $\theta \approx 0$, one obtains global refinement and for $\theta \approx 1$ one obtains very few marked cells which represent most of the error. In the latter case, one is very efficient in terms of degrees of freedom but needs a lot of iterations of the whole adaptive process. Since a single iteration is rather costly, one could take a smaller θ, which then also potentially results in more degrees of freedom. A strategy with the intention mark as few triangles as possible while sustaining an uniform convergence of the whole process is the above mentioned equilibration strategy which was first proposed in [Dör96, Section 4.2].

4.2 Algorithm [Equilibration strategy, adapted from [Ver13, Algorithm 2.2]]**.** *Given*: A partition $\mathcal{T}$, error indicators η_K for the elements $K \in \mathcal{T}$ and a threshold $\theta \in (0,1)$.

Sought: A subset $\widetilde{\mathcal{T}} \subset \mathcal{T}$ of marked elements that should be refined.

(i) Compute $\Theta_{\mathcal{T}} = \sum_{K\in\mathcal{T}} \eta_K^2$, set $\Sigma_{\mathcal{T}} = 0$ and $\widetilde{\mathcal{T}} = \emptyset$.

(ii) If $\Sigma_{\mathcal{T}} \geq (1-\theta)\Theta_{\mathcal{T}}$ return $\widetilde{\mathcal{T}}$ and stop, otherwise continue with step (iii).

(iii) Compute $\widetilde{\eta}_{\mathcal{T},\max} = \max_{K\in\mathcal{T}\setminus\widetilde{\mathcal{T}}} \eta_K$.

(iv) For all elements $K \in \mathcal{T}\setminus\widetilde{\mathcal{T}}$: Check whether $\eta_K = \widetilde{\eta}_{\mathcal{T},\max}$. If this is the case, put K in $\widetilde{\mathcal{T}}$ and add η_K^2 to $\Sigma_{\mathcal{T}}$, otherwise skip K.

Once all elements have been checked, return to step (ii).

For $\theta \to 0$ and $\theta \to 1$ the same behaviors as in the maximum strategy can be observed.

When dealing with problems that possess a layer in their solution, it might happen that the majority of the cells has only little error and most of it concentrates on very few cells. The cells that have only little error might split up into cells which have almost no error and cells which have significant error but still less than the ones that hold the majority of the overall error. If this is the case, the cells with the "medium" error might never be refined, which is disadvantageous for the convergence of the adaptive process. Being disadvantageous for the convergence means that one potentially needs more iterations to arrive at a sufficiently

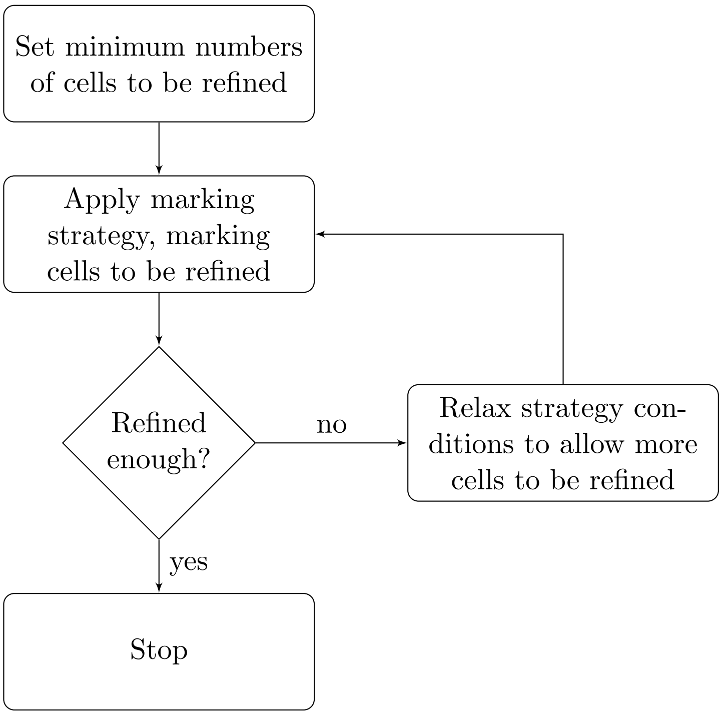

Set minimum numbers of cells to be refined
Apply marking strategy, marking cells to be refined
Refined enough?
no
Relax strategy conditions to allow more cells to be refined
yes
Stop

Figure 6: Algorithm that marks at least a minimal number of cells for refinement per iteration of the adaptive process, see [Joh00, Section 4].

accurate solution, which is more expensive than just refining more cells hoping to refine at least some of the medium error ones. In total a cell marking algorithm could look like in Figure 6.

Instead of relaxing the marking conditions like depicted in Figure 6, one could also just mark the first

$$\varepsilon_M \approx 1\% \tag{88}$$

cells holding the largest error and apply the marking strategy to the remaining cells, like proposed in [Ver13, Section 2.1.1].

5 Cell refinement

Following the steps of Algorithm 1.1, this section corresponds to (iv).(b), the actual refinement of marked cells and therefore construction of $\mathcal{T}_{k+1}$ out of $\mathcal{T}_k$. Here, only the two-dimensional case will be considered.

Unlike uniform or global refinement where all cells in the current mesh $\mathcal{T}_k$ are refined in order to obtain $\mathcal{T}_{k+1}$, the kind of refinement relevant to the adaptive algorithm is of a local nature. Often (cf., [Ver13, Section 2.1.2]), refinement of cells is achieved by introducing vertices at the midpoint of the edges and then connecting the edge midpoints within each triangle or quadrilateral, as shown in Figure 7. This scheme is known as the dyadic split or the 1-to-4 split since one cell is split into four and the resulting cells are called *red*. Its application yields cells whose shape parameter $\frac{h_K}{\rho_K}$ does not change but also hierarchy of meshes in which it is possible that cells of a higher refinement-level meet cells of a lower refinement-level, introducing hanging nodes as described in the following definition.

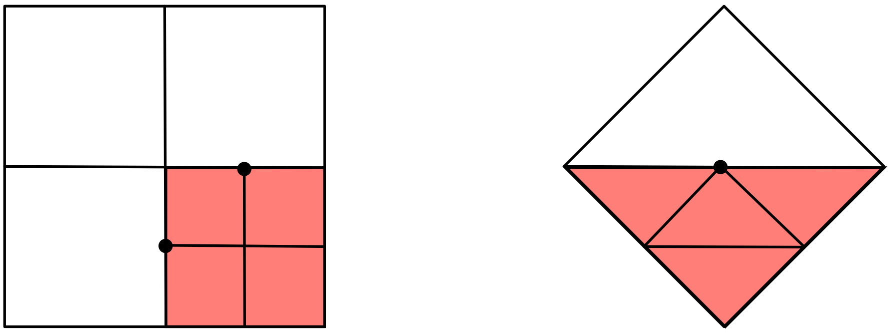


Figure 7: Refinement of triangles and quadrilaterals by the 1-to-4 split. The refined triangles and quadrilaterals are red, the resulting hanging nodes are denoted by $\bullet$.

5.1 Definition [Hanging nodes, conforming and nonconforming grid]**.** Hanging nodes are vertices that arise when two elements of $\mathcal{T}$ are not disjoint and do not share a complete lower-dimensional face of their boundaries, i.e., violate the admissibility condition on the grid described in Section 2.1.2 like shown in Figure 7. If a grid has no hanging nodes, it is called conforming, otherwise it is called nonconforming.

When a conforming grid is desired, one needs to resolve these hanging nodes by refining adjacent unrefined cells. There are several ways of resolving, the here presented way can be found in [Ver13, Section 2.1.3] and is the so called red-green-blue or red-green-blue-purple refinement for triangles or quadrilaterals, respectively. The elements obtained are called green, blue and purple and constructed by

- bisecting exactly one edge of an element for green elements,
- bisecting exactly two edges of an element for blue elements,
- bisecting exactly three edges of a quadrilateral for purple elements.

Figure 8 illustrates the different green, blue and purple refinements. In constrast to the red refinement, the green, blue and purple refinements do create elements that are less shape regular, which, if not prevented, could lead to a violation to the shape regularity condition. This can be done by imposing conditions on the green and blue refinement:

- In a blue refinement, the longest edge of the refinement edges is bisected first, i.e., the refinement as in the upper picture of Figure 9 is forbidden.
- When a green refinement can be performed, it is first checked if the edge that is to be bisected was refined in the last n_{gen} refinements. In that case, a blue refinement is performed instead. In [Ver13] a value of $n_{\text{gen}} = 1$ is suggested, as larger values can cause a massive increase of the refinement area. The lower part of Figure 9 shows a situation in which a green refinement can be performed but is forbidden.

As it was already mentioned in the condition on the green refinement, conforming closures of the grid have the consequence that the marking of cells which are to be refined is less local. On the other hand, one has to take special care in a purely nonconforming approach as well. Here, the continuity across edges and faces gets violated complicating the matrix assembly and the situation may arise that neighboring elements vastly differ in size. To prevent this,

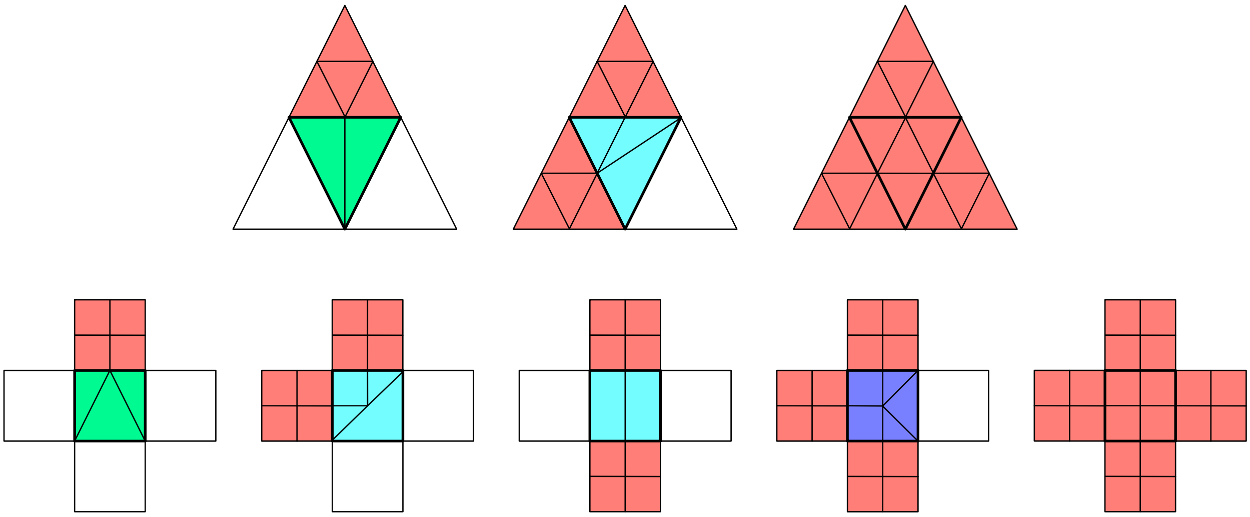


Figure 8: Resolving of hanging nodes in triangles (top row) and quadrilaterals (bottom row). The element that is being subdivided has a bold boundary. In the top row from left to right: Green, blue and red refinement. In the bottom row from left to right: Green, blue, blue, purple and red refinement.

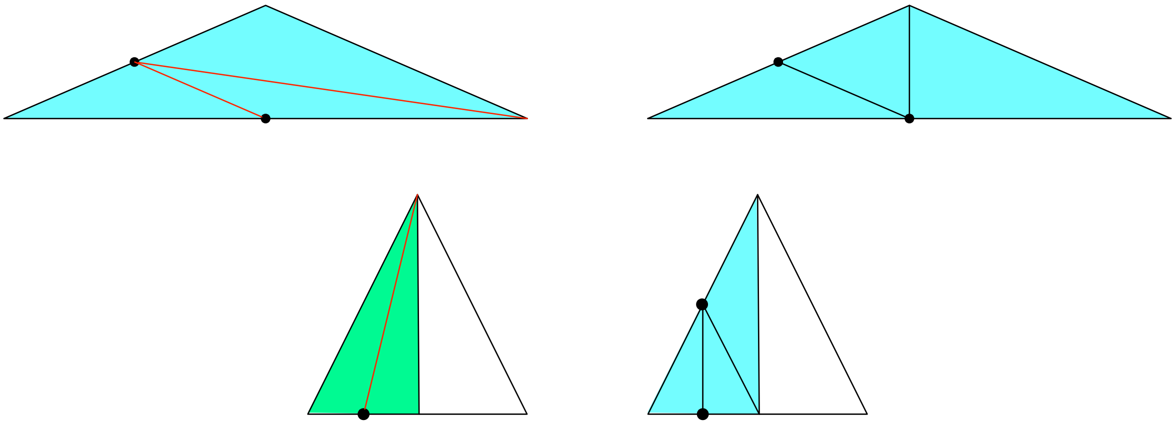


Figure 9: Forbidden blue and green refinements, hanging nodes are denoted with a •. In the upper left the forbidden blue refinement, in the upper right the correct blue refinement. In the lower left the for $n_{\mathrm{gen}} = 1$ forbidden green refinement, in the lower right the corresponding blue refinement which introduces a new hanging node.

refinement is often restricted to “balanced” or “1-regular” meshes (cf., e.g., [BSW83, Section 2]), i.e., meshes where the refinement level of adjacent elements is not allowed to differ more than one. One possibility to achieve 1-regular grids is to remove green, blue and purple refinements before the next refinement. This however yields grid hierarchies which are no subsets of each other, as depicted in Figure 4.

In the numerical studies the software ParMooN was used, which is the parallel version of the finite element package MooNMD [JM04]. Here, the refinement is restricted to 1-regular grids.

This section concludes by construction of $\mathcal{T}_{k+1}$ out of $\mathcal{T}_k$ the last step of the adaptive process as in Algorithm 1.1. In terms of sections, one now would go back to Section 2 and solve the considered problem on the newly obtained mesh.

6 Numerical studies

This chapter deals with numerical examples and studies for Algorithm 1.1 in the residual based case and the dual weighted residual case. The examples' solutions should possess layers or singularities such that an application of a posteriori estimates makes sense or is even necessary. Further, instead of considering the Darcy equations like in the previous sections, here the scalar convection–diffusion equations will be considered. In the case of $\mathbb{K} = K\,\mathbb{I}$, they can be seen as a generalization of the Darcy equations.

The first two examples deal with convection–diffusion equations, followed by another two examples for the Stokes equations and one example with the coupled Stokes–Darcy system. For both the convection–diffusion and Stokes equations there is one example for the residual based error estimator in the energy norm and one example concerning the dual weighted residual method, as introduced in Section 3.1 and Section 3.2, respectively.

All computations were performed using the parallel finite element package ParMooN, which is a fork of MooNMD [JM04]. The arising linear systems of equations were solved using the direct solver UMFPACK. The adaptive algorithm was applied using the maximum strategy of Algorithm 4.1.

6.1 Scalar convection–diffusion equations

Scalar convection–diffusion equations model the transport of scalar quantities like temperature or concentration. In the steady-state, dimensionless case, the associated problem reads as follows:

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary and $f \in L^2(\Omega)$ a source term. Then one wants to find $u : \Omega \to \mathbb{R}$ such that

$$\begin{cases} -\varepsilon\Delta u + \mathbf{b}\cdot\nabla u + cu = f & \text{, in } \Omega, \\ u = g & \text{, on } \Gamma_D, \\ \varepsilon\nabla u \cdot \mathbf{n} = g_N & \text{, on } \Gamma_N, \end{cases} \tag{89}$$

where g and g_N denote Dirichlet and Neumann boundary conditions on relatively open boundary components $\Gamma_D \subset \partial\Omega$ and $\Gamma_N \subset \partial\Omega$, respectively, such that

$$\overline{\Gamma}_D \cup \overline{\Gamma}_N = \partial\Omega \text{ and } \Gamma_D \cap \Gamma_N = \emptyset,$$

and $\varepsilon > 0$ is the diffusion, $\mathbf{b} : \Omega \to \mathbb{R}^d$ the convection, and $c : \Omega \to \mathbb{R}$ the reaction parameter. The weak formulation of problem (89) reads as follows:

Find $u \in H^1(\Omega)$ such that

$$\begin{cases} a(u,v) + b(u,v) + c(u,v) = (f,v) + (g_N, v)_{\Gamma_N} \ , \\ \qquad\qquad\qquad\qquad\quad u = g \qquad\qquad\qquad\quad , \text{ on } \Gamma_D, \end{cases} \tag{90}$$

for all

$$v \in V_0 := \left\{ v \in H^1(\Omega) : v\big|_{\Gamma_D} = 0 \right\},$$

where

$$a(u,v) := (\varepsilon \nabla u, \nabla v)_0, \quad b(u,v) := (\mathbf{b} \cdot \nabla u, v), \quad c(u,v) := (cu, v).$$

Similarly as in Section 2.1.2 and Section 2.2.2, one can discretize the problem using a conforming standard Galerkin discretization with a finite-dimensional subspace $V^h \subset V_0$. For simplicity of the presentation homogeneous Dirichlet boundary conditions are assumed. Then, a solution $u^h \in V^h$ is sought such that

$$a(u^h, v^h) + b(u^h, v^h) + c(u^h, v^h) = (f, v^h) + (g_N, v^h)_{\Gamma_N} \quad \forall v^h \in V^h. \tag{91}$$

However, on coarse or moderately fine meshes the above discretization sometimes cannot resolve all important features of the solution like in the case of the Hemker problem, see Section 6.1.1. One then can equip the discretization with a stabilizing component. Here, the SUPG stabilization was used, introducing an additional term on the left-hand side of (91), namely,

$$\sum_{K \in \mathcal{T}} (-\varepsilon \Delta u^h + \mathbf{b} \cdot \nabla u^h + cu^h - f, \delta_K \mathbf{b} \cdot \nabla v^h)_{0,K} \quad \forall v \in V^h.$$

The stabilization parameter δ_K is chosen for $d = 2$ as

$$\delta_K(x,y) = \frac{\widetilde{h}_K}{2p\|\mathbf{b}(x,y)\|} \zeta(\mathrm{Pe}_K(x,y))$$

with

$$\mathrm{Pe}_K(x,y) = \frac{\|\mathbf{b}(x,y)\| \widetilde{h}_K}{2p\varepsilon}, \quad \zeta(\alpha) = \coth \alpha - \alpha^{-1},$$

where $\widetilde{h}_K$ is the length of a mesh cell $K \in \mathcal{T}$ in the direction of $\mathbf{b}$, and p is the degree of the used finite element space, see, e.g., [JS14].

Generally, residual based a posteriori error indicators for the scalar convection–diffusion equations are of the form

$$\eta_{*,K}^2 = \alpha_K \|f + \varepsilon \Delta u^h - \mathbf{b} \cdot \nabla u^h - cu^h\|_{0,K}^2 + \sum_{E \in \mathcal{E}_K \setminus \mathcal{E}_{\Gamma_N}} \frac{\beta_E}{2} \left\| [\![\varepsilon \nabla u^h \cdot \mathbf{n}_E]\!]_E \right\|_{0,E}^2 \tag{92}$$

$$+ \sum_{E \in \mathcal{E}_{K,\Gamma_N}} \beta_E \|\varepsilon \nabla u^h \cdot \mathbf{n}_E - g_N\|_{0,E}^2, \tag{93}$$

see [Joh00, Section 3.4]. For

$$\alpha_K = h_K^4, \quad \beta_E = h_E^3 \tag{94}$$

one obtains an residual based a posteriori error indicator in the L^2-norm, see [Joh00, Section 3.5]. Another a posteriori error indicator $\eta_{K,\mathrm{eng}}$ for the Galerkin discretization with SUPG stabilization in the energy norm was derived in [Ver98] with

$$\alpha_K = \min\{h_K^2 \varepsilon^{-1}, 1\}, \quad \beta_E = \min\{h_E \varepsilon^{-1}, \varepsilon^{-\frac{1}{2}}\}. \tag{95}$$

In the following, two numerical examples of the scalar convection–diffusion equations will be studied. The first example considers a problem of which the solution is unknown, using the error indicator given by (95). The quality of the grids was measured in terms of overshoots and undershoots, i.e., unphysical values in the computed solution and it is going to be studied how the parameteres ε_M and θ in the marking strategy influence the results in terms of required computational effort.

The second example compares the numerical solution of the scalar convection–diffusion equations to a given analytical solution which possesses layers. To obtain the numerical solution, Algorithm 1.1 was applied using the dual weighted residual method with the mean value of the solution in an area away from the layers as functional of interest. Since the analytical solution is known, efficiency indices can be computed and it is going to be analyzed how they are influenced by choosing different values of ε_M (88) in the marking strategy.

6.1.1 The Hemker problem

The Hemker problem was introduced in [Hem96] and describes in the case of $\varepsilon \ll \|\mathbf{b}\|_{L^\infty(\Omega)} = 1$ heat transfer from a hot circle. The solution has two interior layers in direction of the heat transfer and a boundary layer at the hot circle. Its domain is given by

$$\Omega = \{(x, y) \in \mathbb{R}^2 : x \in [-3, 9],\ y \in [-3, 3],\ x^2 + y^2 > 1\}$$

with relatively open boundary components Γ_D and Γ_N. The problem's parameters are $\varepsilon = 10^{-4}$, $\mathbf{b} = (1, 0)^T$, $f = 0$, $g_N = 0$, $g = u_D$ with

$$u_D(x, y) = \begin{cases} 0 & , \text{if } (x, y) \in \{-3\} \times [-3, 3], \\ 1 & , \text{if } x^2 + y^2 = 1. \end{cases}$$

The initial triangulation of the domain with marked boundary components can be seen in Figure 10. A numerical solution on a fine grid with no visible unphysical values is presented in Figure 11. One can see the heat transfer and the resulting boundary and inner layers.

The computations were carried out for the energy-norm error indicator (95) and different values of $\theta \in (0, 1)$ and different values of the minimum number of cells to be refined $\varepsilon_M \in (0, 1]$ per iteration of the maximum strategy of Section 4 using P_1 finite elements. If not sufficiently many mesh cells were marked for refinement by θ, the parameter was updated by $\theta \to 0.7\theta$. One should, if possible, also apply uniform refinement of the initial grid before applying the adaptive algorithm until the most important features of the solution become recognizable, e.g., the position of the layers. This step can be seen as a kind of pre-processing of the initial grid until it becomes sufficiently fine. The number of uniform refinement steps highly depends on the problem and on the initial grid and thus has to be found by numerical tests. In this example, no uniform refinement was performed before applying the adaptive algorithm. Since the analytical solution is unknown, the quality of the numerical solutions is here measured in terms of overshoots c_o and undershoots c_u. These are values of the solution that are not physical and can be caused by spurious oscillations. In this case, the values for the undershoots are defined by the minimal value of the discrete solution and the values for the overshoots are defined by the maximal value of the discrete solution subtracted by one, i.e.,

$$c_u := \min u^h, \quad c_o := \max u^h - 1,$$

see [ACF+11, Section 3.1].

When applying Algorithm 1.1 to the problem with cell marking strategy parameters $\theta = \frac{1}{2}$ and $\varepsilon_M = 25\%$, cf. Section 4, a sequence of grids is generated of which the visually most

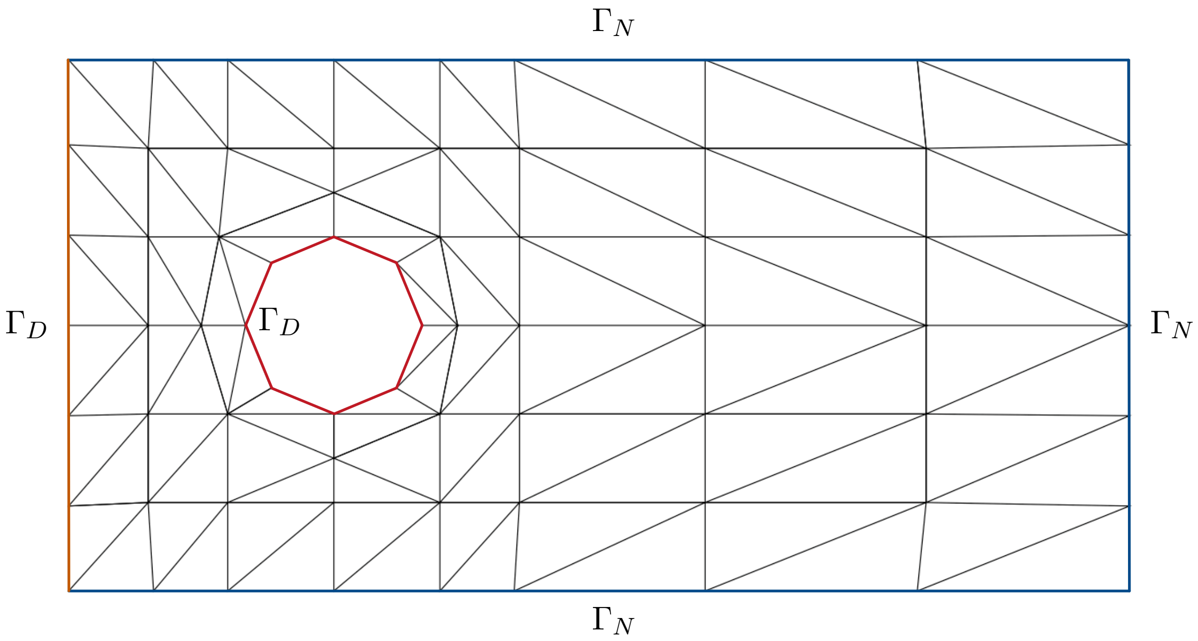


Figure 10: Initial triangulation with 104 cells of the domain of the Hemker example, see Section 6.1.1. The red part of Γ_D corresponds to $u_D = 1$, the orange part of Γ_D corresponds to $u_D = 0$. On Γ_N homogeneous boundary conditions are prescribed.

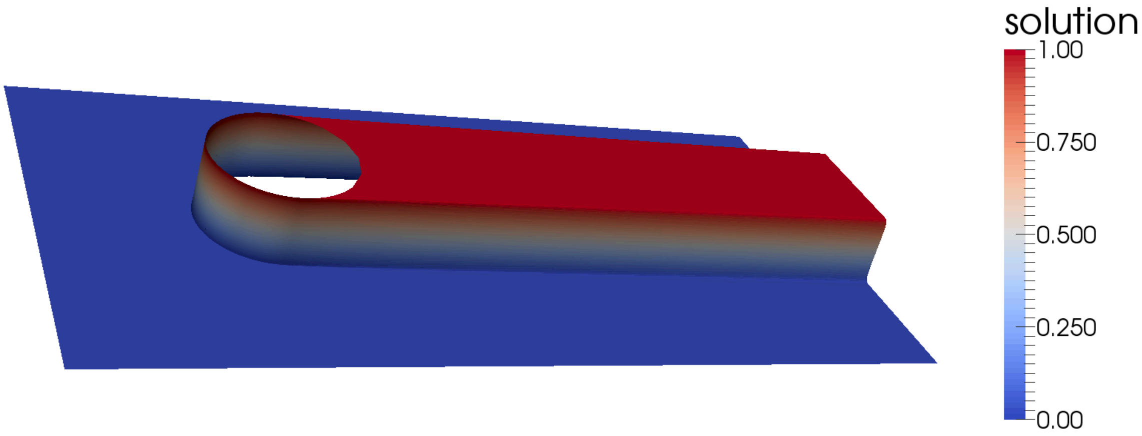


Figure 11: Solution of the Hemker problem.

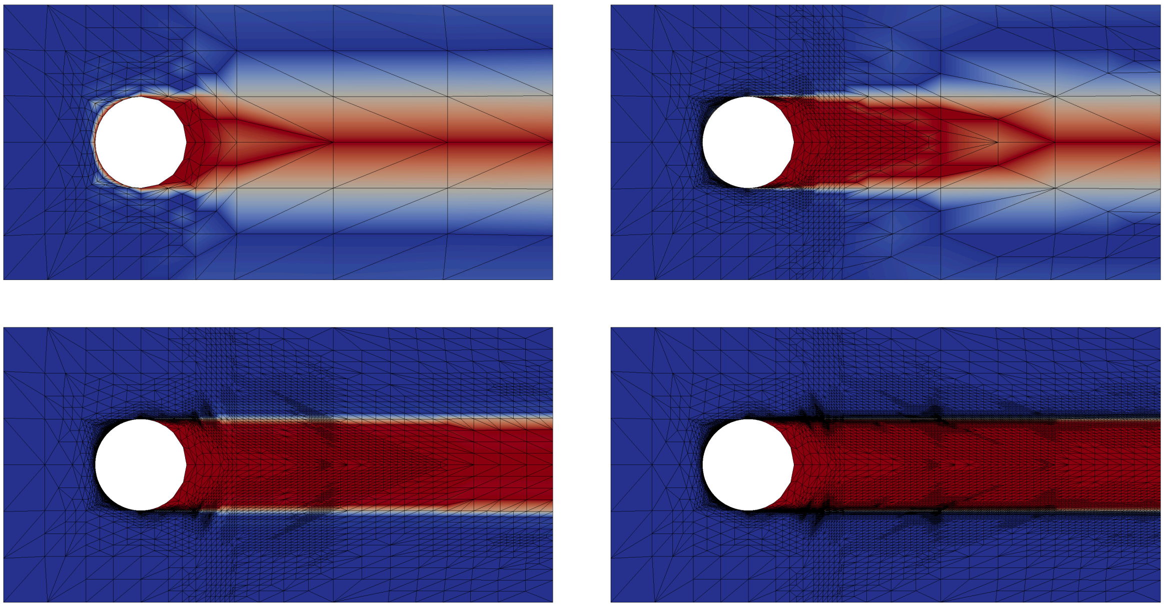


Figure 12: The evolution of grids for $\theta = \frac{1}{2}$ and a minimum of $\varepsilon_M = 25\%$ cells to be refined per iteration of Algorithm 1.1. From top-left to bottom-right the grids belong to iterations 2, 5, 8 and 11 of Algorithm 1.1, the grids consisted out of 484, 4206, 29987 and 217891 cells, respectively.

significant elements are depicted in Figure 12. One can see that the refinement concentrates mostly around the boundary layer at the hot circle and the two interior layers in direction of the heat transfer. This behavior is expected, since these are the regions in which the solution changes most and thus requires a better resolution than elsewhere.

The goal is to compare the efficiency of the algorithm for different values of θ and different values of ε_M and ultimately compare these results to uniform refinement. In Figure 13, tests were run for $\varepsilon_M = 5\%$ and $\theta \in \{0.1, 0.3, 0.7, 0.9\}$.

One can observe that for all tested values of θ, the algorithm reduces c_o and c_u after a certain number of iterations. This number of iterations is higher for higher values of θ and lower for lower values of θ. Nevertheless, the number of iterations is no good measure for the efficiency. Therefore, the number of degrees of freedom, i.e., efficiency in terms of memory usage, as well as the cumulative execution time on a single-core 3.5 GHz CPU are being compared in Table 1 for the level that yields values of c_o and c_u which are closest to 10^{-2} and -10^{-2}, respectively. The algorithm terminated after 35 levels or once the system to be solved became larger than 2 GB which is an internal memory restriction of the direct solver that was used.

The data indicates that the ε_M parameter dominates the θ parameter as soon as it is ≥ 0.25. Indeed, θ had to be decreased several times in the numerical tests. Also, it seems to be easier to reduce the overshoots c_o by uniform refinement than the undershoots c_u.

In the case of $\varepsilon_M = 0.75$, the undershoots were even worsened for all values of θ by applying the adaptive algorithm and for $\varepsilon_M = 0.5$, the target quantity of -10^{-2} was never reached before the memory requirements exceeded 2 GB. Further, the overshoots c_o require about twice as much time (about 20 seconds) and more degrees of freedom for $\varepsilon_M = 0.05$ than for

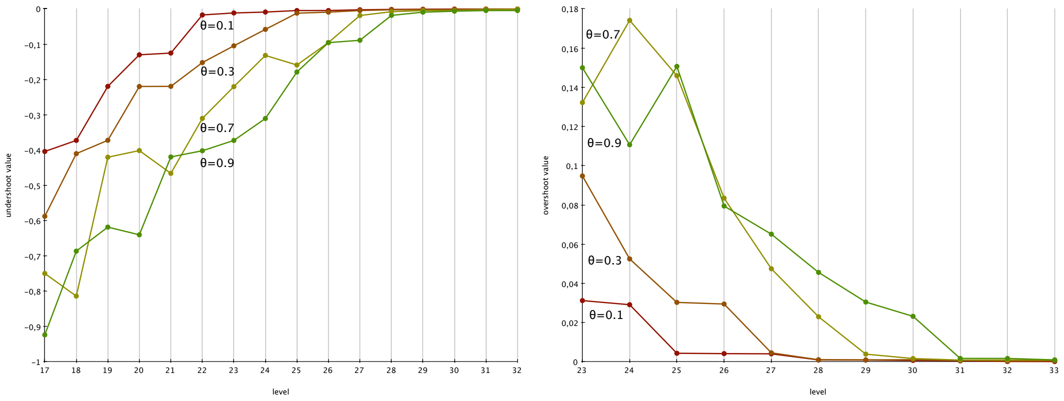
Table 1: Tables comparing required level, degrees of freedom (d.o.f.) and computation time in seconds to get closest to $c_u = -10^{-2}$ (a) and $c_o = 10^{-2}$ (b) for different values of ε_M and θ.

(a) Table showing required level, degrees of freedom (d.o.f.) and needed time to get closest to $c_u = -10^{-2}$ for different values of ε_M and θ.

		c_u	Level	d.o.f.	time in seconds
$\varepsilon_M = 0.05$	$\theta = 0.1$	-0.00918006	24	106237	15.797729
	$\theta = 0.3$	-0.00945207	26	107181	16.295698
	$\theta = 0.7$	-0.00819104	28	110582	16.210087
	$\theta = 0.9$	-0.00949332	29	87307	14.07319
$\varepsilon_M = 0.25$	$\theta = 0.1$	-0.00199572	12	168352	13.473212
	$\theta = 0.3$	-0.00074229	12	177456	14.074377
	$\theta = 0.7$	-0.00201526	12	158205	12.687869
	$\theta = 0.9$	-0.000766904	12	175689	14.104466
$\varepsilon_M = 0.50$	$\theta = 0.1$	-0.373561	10	857838	66.255399
	$\theta = 0.3$	-0.373561	10	833742	64.993061
	$\theta = 0.7$	-0.373561	10	810890	81.338537
	$\theta = 0.9$	-0.373561	10	821472	64.373674
$\varepsilon_M = 0.75$	$\theta = 0.1$	-0.625846	0	70	0.029041
	$\theta = 0.3$	-0.625846	0	70	0.036025
	$\theta = 0.7$	-0.625846	0	70	0.021676
	$\theta = 0.9$	-0.625846	0	70	0.036853

(b) Table showing required level, degrees of freedom (d.o.f.) and computation time to get closest to $c_o = 10^{-2}$ for different values of ε_M and θ.

		c_o	Level	d.o.f.	time in seconds
$\varepsilon_M = 0.05$	$\theta = 0.1$	0.00438	25	141218	20.921739
	$\theta = 0.3$	0.00462	27	149824	21.804098
	$\theta = 0.7$	0.00394	29	153826	21.849847
	$\theta = 0.9$	0.00176	31	156241	24.17644
$\varepsilon_M = 0.25$	$\theta = 0.1$	0.00696	11	87472	7.139732
	$\theta = 0.3$	0.00689	11	91034	7.340307
	$\theta = 0.7$	0.00696	11	82659	6.719419
	$\theta = 0.9$	0.00691	11	90370	7.393496
$\varepsilon_M = 0.50$	$\theta = 0.1$	0.00445	8	133147	9.602799
	$\theta = 0.3$	0.00445	8	129870	9.486261
	$\theta = 0.7$	0.00445	8	126574	12.430737
	$\theta = 0.9$	0.00445	8	127758	9.412674
$\varepsilon_M = 0.75$	$\theta = 0.1$	0.00026	7	121941	12.647923
	$\theta = 0.3$	0.00026	7	114115	11.822957
	$\theta = 0.7$	0.00026	7	108481	8.313629
	$\theta = 0.9$	0.00026	7	103287	10.95962



(a) Undershoots of the solution u of the Hemker problem for levels $17, \dots, 32$.

(b) Overshoots of the solution u of the Hemker problem for levels $23, \dots, 33$.

Figure 13: Undershoots (a) and overshoots (b) of the Hemker problem for a minimum number of cells to refine of $\varepsilon_M = 5\%$ per level and refinement tolerances $\theta \in \{0.1, 0.3, 0.7, 0.9\}$ in red, orange, yellow and green, respectively. The x-axis represents the current level, the y-axis the undershoot or overshoot.

the other values of ε_M.

On the other hand, the undershoots were resolved by $\varepsilon_M \in \{0.05, 0.25\}$ in almost the same time, despite that $\varepsilon_M = 0.05$ requires about twice as many iterations as $\varepsilon_M = 0.25$. In the case of $\varepsilon_M = 0.05$, one can observe that for increasing θ, one becomes more and more memory efficient. For $\theta = 0.9$, one even has about half the degrees of freedom than for $\varepsilon_M = 0.25$ and any value of θ used.

When comparing the results to uniform refinement, the trend continues, i.e., the undershoots were not resolved up to -10^{-2} until the memory requirements exceeded 2 GB and the overshoots were resolved up to 10^{-2} at level 6 with 214144 degrees of freedom, which is approximately between 1.3 and 2.4 times the number of degrees of freedom that were required for the adaptively refined grids. Thus, as one would expect, uniform refinement is much more memory inefficient than adaptive refinement. On the other hand, this result was computed in 6.79 seconds, which is faster than the adaptive refinement except for the configuration with parameters $\varepsilon_M = 0.25$ and $\theta = 0.7$.

Altogether, if one is interested in reducing c_o as well as $|c_u|$, the parameter ε_M should either be around 0.25 regardless of θ or one should uniformly refine the initial grid a few times before applying the adaptive algorithm with a value of ε_M around 0.05 and a value of θ around 0.9 to be as efficient as possible with respect to computation time and memory requirements.

6.1.2 Example with boundary layers

This example is Example 5.2 of [Joh00]. The solution is prescribed by

$$u(x,y) = xy^2 - y^2 \exp\left(\frac{2(x-1)}{\varepsilon}\right) - x \exp\left(\frac{3(y-1)}{\varepsilon}\right) + \exp\left(\frac{2(x-1)+3(y-1)}{\varepsilon}\right)$$

on the unit-square $\Omega = (0,1)^2$ with $\varepsilon \in \{10^{-2}, 10^{-4}, 10^{-6}\}$, $\mathbf{b} = (2,3)^T$, $c = 1$, and $\partial\Omega = \Gamma_D$, see Figure 14. The right-hand side and boundary conditions are chosen according to the

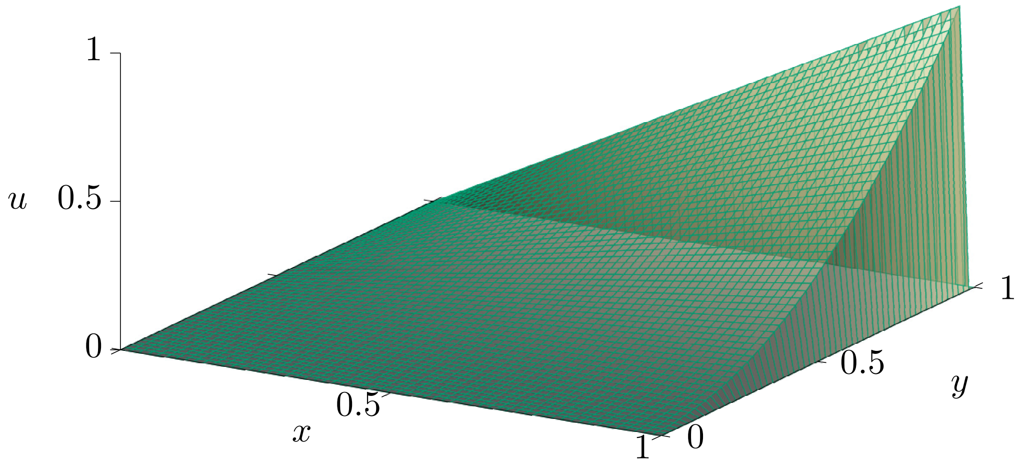


Figure 14: Solution for the example given in Section 6.1.2 for $\varepsilon = 10^{-6}$.

solution. The solution possesses boundary layers at $x = 1$ and $y = 1$. Goal of this section is to apply the dual weighted residual method to this example for a functional of interest that considers an area away from the boundary layers, compare the resulting mesh to a mesh that was constructed by application of a residual based a posteriori error estimator and then analyze its modified efficiency index (87) for different values of ε.

The computations were carried out for the dual weighted residual method in the L^2-norm by reweighting the error indicator (94), the weights were evaluated by difference quotients, cf. Section 3.2.3. The used finite element space was P_2. Concerning the cell marking strategy, the minimum number of cells to be refined was chosen as $\varepsilon_M = 25\%$ together with $\theta = 0.5$ and an update of $\theta \to 0.7\theta$ if not sufficiently many cells were marked for refinement. The functional of interest is the function average value in $B = \{(x,y)^T : (x-0.7)^2 + (y-0.3)^2 \leq 0.01^2\}$, i.e.,

$$J(u) = \frac{1}{|B|} \int\limits_B u. \tag{96}$$

The initial grid and a selection of adaptively refined grids can be seen in Figure 15. One can observe that the dual weighted residual method with η_{dwr,L^2} indeed concentrates the refinement around B and leaves the layer area untouched, whereas the error indicator η_{L^2} yields a grid that is very fine exclusively in the area of the layer. Aside from the region $B \subset \Omega$ that gets refined, there is also a stripe of refined cells from B into direction of the negative gradient of the solution. The images of the dual solution in Figure 16 possess this stripe as well with a slight decay into the direction of the boundary and are 0 elsewhere, which coincides with the sequence of grids one can observe in Figure 15.

The degrees of freedom and efficiency indices for $\varepsilon \in \{10^{-2}, 10^{-4}, 10^{-6}\}$ are shown in Table 3. Generally the values indicate a strong over-estimation of the error which could be because

Table 3: Efficiency indices and degrees of freedom for the example of Section 6.1.2 for $\varepsilon \in \{10^{-2}, 10^{-4}, 10^{-6}\}$. The "-" indicates that the memory requirements exceeded 2 GB and therefore the direct solver ran out of memory.

	$\varepsilon_M = 10^{-2}$		$\varepsilon_M = 10^{-4}$		$\varepsilon_M = 10^{-6}$	
Level	d.o.f.	$I_{\text{eff}}^{\text{DWR}}$	d.o.f.	$I_{\text{eff}}^{\text{DWR}}$	d.o.f.	$I_{\text{eff}}^{\text{DWR}}$
0	1089	232.646	1089	520.165	1089	536.665
1	1968	1234.49	2018	1065.5	2018	1075.18
2	3559	2957.45	3842	1056.27	3974	1092.1
3	7166	3044.73	7346	310.602	7490	338.037
4	14275	1560.94	14083	237.817	14103	264.182
5	29320	485.532	26414	157.408	26855	168.199
6	60608	337.526	61657	125.988	61383	100.6
7	115883	633.54	131491	120.078	133910	56.3008
8	224739	249.064	261908	151.483	269900	35.9318
9	414182	442.575	711988	546.933	624412	22.5777
10	812489	476.89	1394826	1038.08	1252220	8.01729
11	1517876	36.4234	-	-	-	-

of the approximation of the continuous dual solution by difference quotients, as reported in [BR03a, Section 4.1]. At least in the cases of $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-6}$ one can see decreasing over-estimation, in the case of $\varepsilon = 10^{-4}$ one might need more iterations.

6.2 Stokes equations

This section considers two numerical examples for the Stokes equations. First an example with a known solution is being compared with results from the literature, as well as analyzed with respect to the parameters ε_M and θ, which are part of the chosen cell marking strategy, see Section 4.

In the second example, the goal is to compute a concrete physical quantity. The computations were carried out for the dual weighted residual method with a functional of interest that represents the physical quantity, the unweighted error estimator (65) and for stepwise uniformly refined grids. The different methods' results are compared with respect to their memory consumption that is required to calculate the quantity up to a certain tolerance.

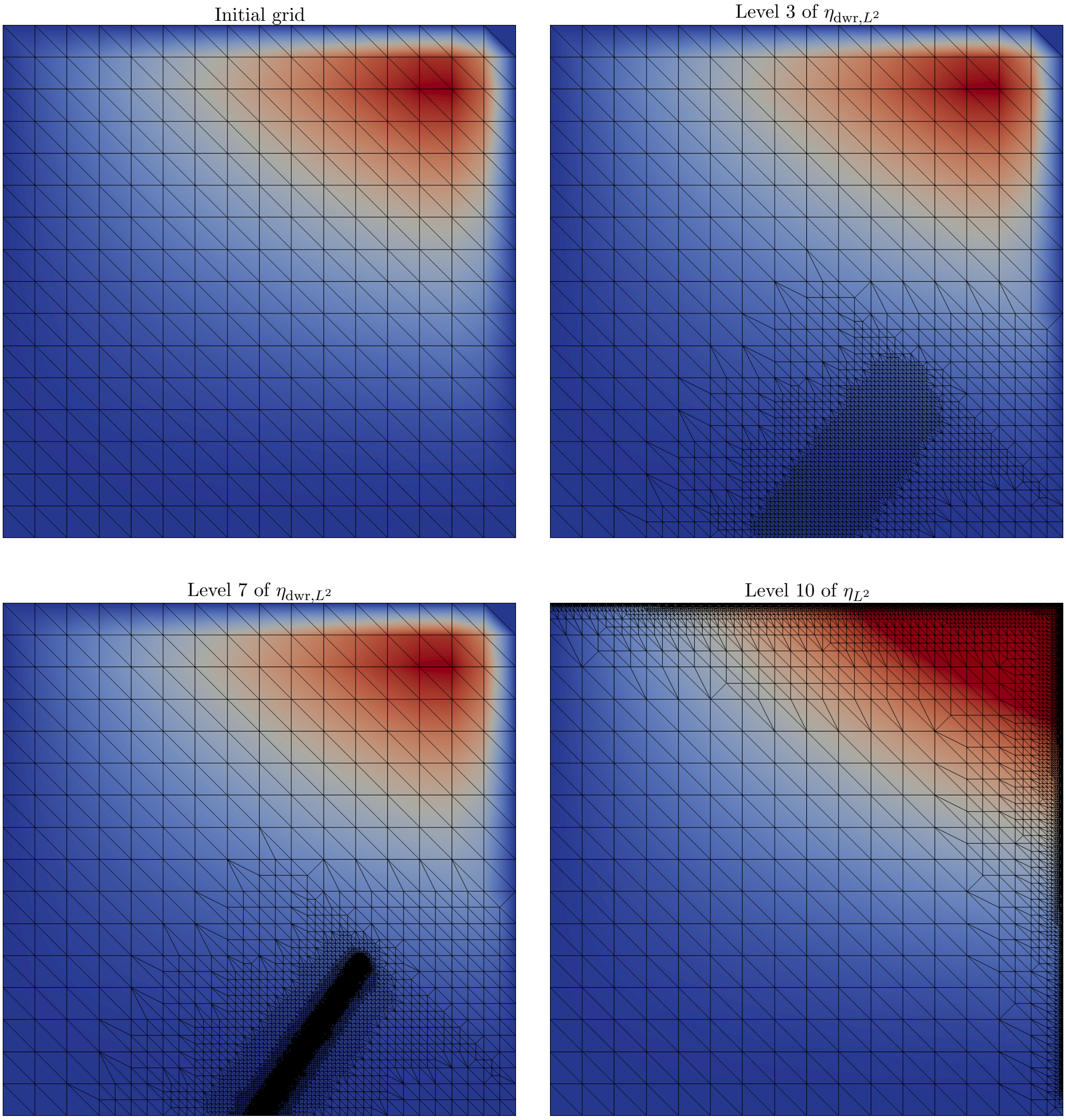
6.2.1 Disc with a crack

This example deals with a Stokes problem for which the analytical solution is known. It is given in polar coordinates $(r, \theta) \in [0, \infty) \times [0, 2\pi)$ by

$$\begin{aligned} \mathbf{u} &= \sqrt{r}\mathbf{u}_b = \sqrt{r}\left(\cos\frac{\theta}{2} - \cos\frac{3\theta}{2}, 3\sin\frac{\theta}{2} - \sin\frac{3\theta}{2}\right), \\ p &= -\frac{4}{\sqrt{r}}\cos\frac{\theta}{2}, \end{aligned}$$

with

$$x = r\cos\theta, \quad y = r\sin\theta,$$



Initial grid

Level 3 of η_{dwr,L^2}

Level 7 of η_{dwr,L^2}

Level 10 of η_{L^2}

Figure 15: The initial grid (top left) with 512 cells and 1089 degrees of freedom, the grid after 3 (top right) and 7 (bottom left) iterations of the dual weighted residual method using the functional of interest given by (96) with 3691 cells and 7490 degrees of freedom and 66855 cells and 133910 degrees of freedom, respectively. Further, the grid after 10 (bottom right) iterations of the residual estimator η_{L^2} given by (94) with 324045 cells and 676698 degrees of freedom.

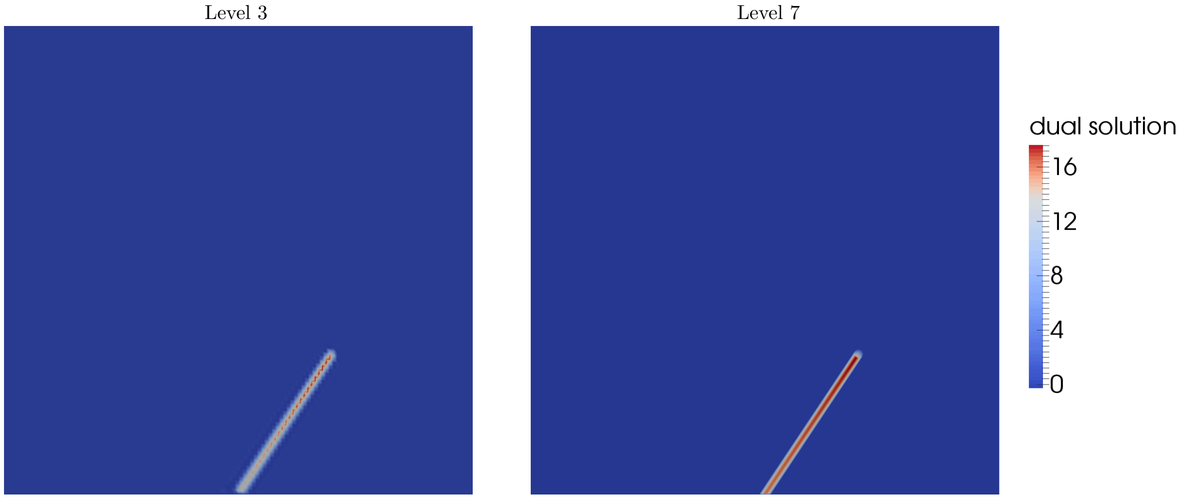


Figure 16: Dual solution of the problem given in Section 6.1.2 for levels 3 and 7 with 3691 and 66855 cells, respectively.

where $(x, y) \in \mathbb{R}^2$, see [Joh98, Example 11] and [BW90, Section 5.1].

The domain is a disc of radius 1 with the center $(0, 0) \in \mathbb{R}^2$ and a crack along the x-axis between the points $(0, 0) \in \mathbb{R}^2$ and $(1, 0) \in \mathbb{R}^2$. The domain and its initial triangulation are schematically depicted in Figure 17. The example's boundary conditions are Dirichlet boundary conditions which are given on the circle by $\mathbf{u}_b$ and on the crack by homogeneous boundary conditions, the right-hand side is according to the prescribed solution set to be homogeneous. The viscosity is set to $\nu = 1$.

The solution possesses in the pressure component a singularity in the origin, as can be observed in Figure 18.

The computations were carried out using the a posteriori error indicator in the energy norm (65) and P_2/P_1 finite elements. The cell marking strategy was chosen to be the maximum strategy of Algorithm 4.1 for different parameters of ε_M and θ. If not sufficiently many cells were marked for refinement, θ was updated by $\theta \to 0.7\theta$. As in [Joh98], the quality of the computed solution $(\mathbf{u}^h, p^h)$ was measured in terms of $\|\mathbf{u} - \mathbf{u}^h\|_0$ and the resolution of the singularity of the pressure. The resolution is measured with the minimal value $p^h_{\min}$ and the maximal value $p^h_{\max}$.

In Table 4 one can see for a selection of levels the cumulative needed time on a single-core 2.3 GHz CPU, degrees of freedom, the L^2 error of the velocity as well as the resolution of the singularity for uniform refinement and the adaptive algorithm configured with $\varepsilon_M \in \{0.05, 0.25, 0.50\}$ and $\theta \in \{0.1, 0.5, 0.9\}$. For each configuration (ε_M, θ) there are two levels, the first one being the one with an L^2 error in the velocity closest to 0.0067, which is the error of the uniform refinement for level 5 and the second selected level being the last level that could be computed before the direct solver ran out of memory.

One can see that in terms of degrees of freedom, smaller values of ε_M with larger values of θ are of advantage. While the using the adaptive algorithm with $\varepsilon_M = 0.5$ and $\theta = 0.1$ at level 5 results in roughly 26% of the degrees of freedom of the uniform refinement at level 5, the same value of ε_M with $\theta = 0.9$ already results in about 21%. The data in Table 4 shows that this trend continues, the best configuration seems to be $\varepsilon_M = 0.05$ with $\theta = 0.9$, where only about 1% of the degrees of freedom are used for comparable results. The corollary that adaptive refinement needs significantly fewer degrees of freedom to achieve results that

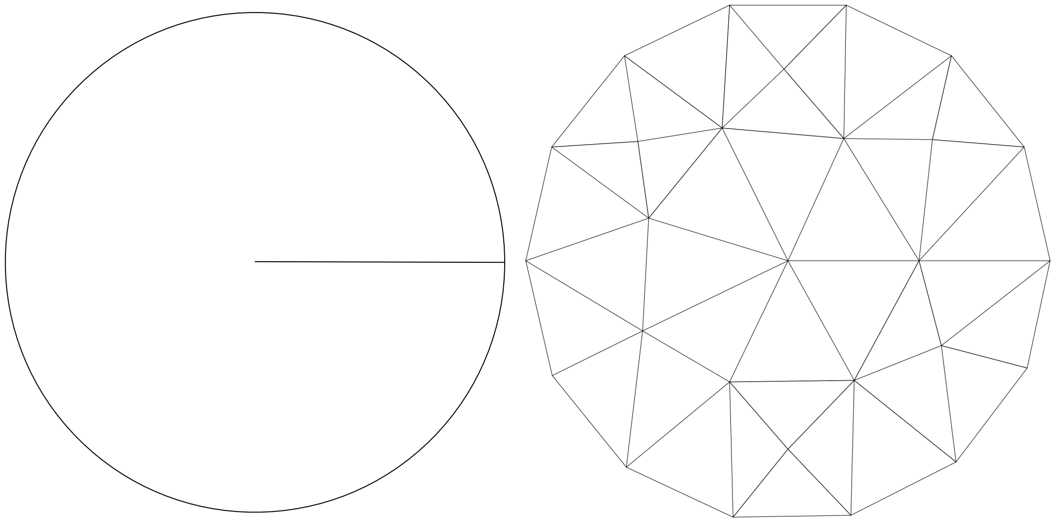


Figure 17: Sketch of the domain and its initial triangulation for the example presented in Section 6.2.1.

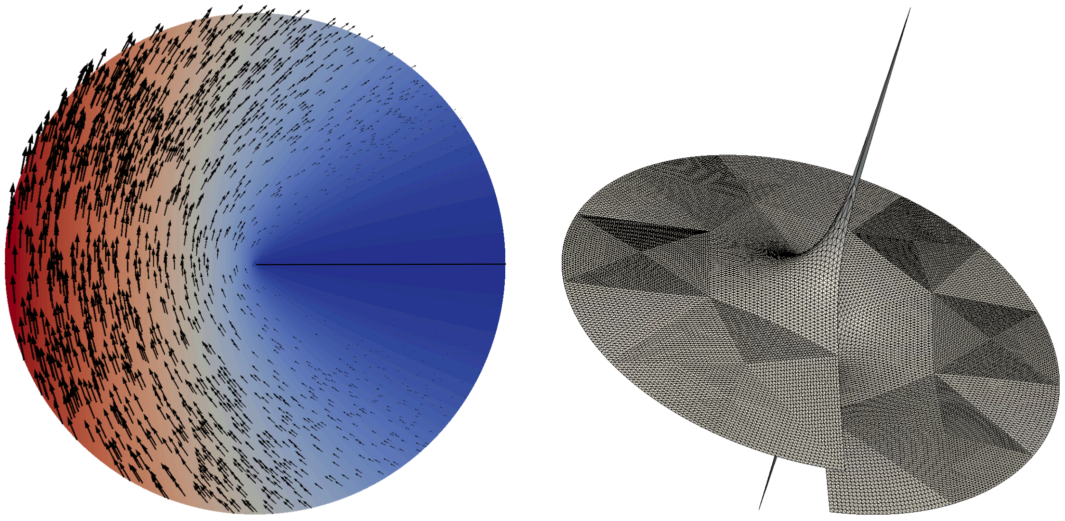
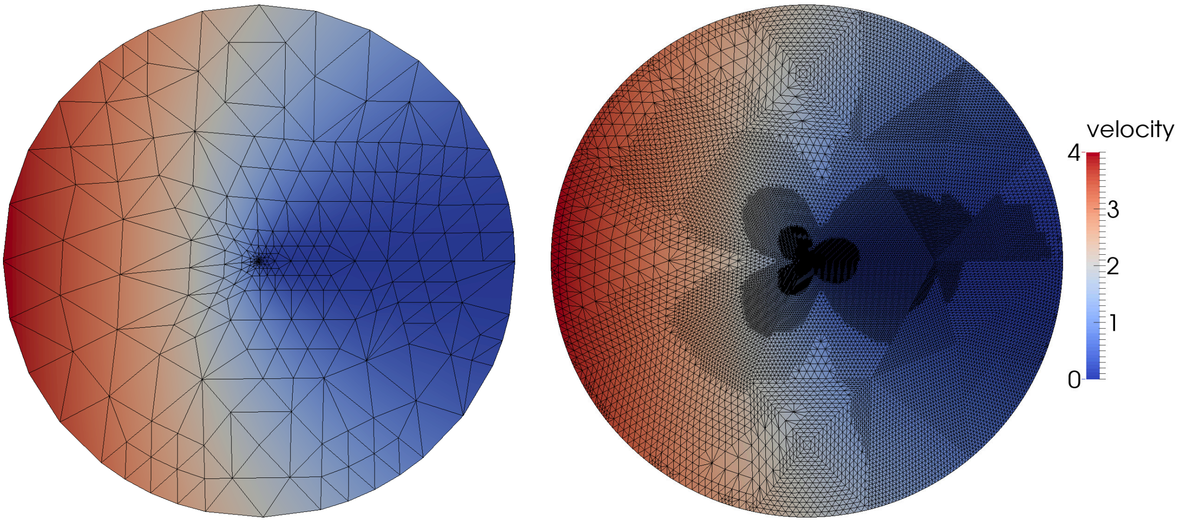


Figure 18: Solution of the example presented in Section 6.2.1. The velocity component of the solution is left, the pressure component is right.

Table 4: Table displaying degrees of freedom, L^2 error in velocity, $p^h_{\min}$, $p^h_{\max}$ and cumulative needed time on a single-core 2.3 GHz CPU for a selection of levels of uniform refinement and the adaptive algorithm configured with $\varepsilon_M \in \{0.05, 0.25, 0.50\}$ and $\theta \in \{0.1, 0.5, 0.9\}$ for the numerical example of Section 6.2.1.

		Level	total d.o.f.	$\|\mathbf{u} - \mathbf{u}^h\|_0$	$p^h_{\min}$	$p^h_{\max}$	time
uniform refinement		4	43347	0.014	-34.6964	34.4638	$0.44s$
		5	174243	0.0067	-49.4769	49.1456	$2s$
$\varepsilon_M = 0.50$	$\theta = 0.1$	5	45879	0.0067	-49.4769	49.1456	$1.1s$
		7	350406	0.0017	-99.5474	98.8813	$10s$
	$\theta = 0.5$	5	38934	0.0067	-49.4769	49.1456	$1.2s$
		7	303060	0.0017	-99.5474	98.8813	$8.8s$
	$\theta = 0.9$	5	37263	0.0067	-49.4769	49.1456	$1.1s$
		7	287913	0.0017	-99.5474	98.8813	$8.9s$
$\varepsilon_M = 0.25$	$\theta = 0.1$	5	11847	0.0067	-49.4775	49.1462	$0.57s$
		10	355230	0.00021	-282.043	280.156	$25s$
	$\theta = 0.5$	5	8031	0.0067	-49.4787	49.1473	$0.41s$
		11	500334	0.0001	-398.918	396.249	$23s$
	$\theta = 0.9$	5	9048	0.0067	-49.4777	49.1471	$0.24s$
		10	274305	0.00021	-282.043	280.156	$16s$
$\varepsilon_M = 0.05$	$\theta = 0.1$	5	5028	0.0067	-49.4839	49.1524	$0.4s$
		19	543381	$2.4 \cdot 10^{-6}$	-6383.93	6341.26	$1.3 \cdot 10^2 s$
	$\theta = 0.5$	6	1803	0.005	-44.528	44.1117	$0.13s$
		23	537501	$2.3 \cdot 10^{-6}$	-16210.4	16045.9	$2.1 \cdot 10^2 s$
	$\theta = 0.9$	7	1713	0.0051	-44.5286	44.1115	$0.25s$
		26	512400	$2.4 \cdot 10^{-6}$	-22924.1	22709.4	$2.8 \cdot 10^2 s$



Level 10 with $\varepsilon_M = 0.05,\ \theta = 0.9$

Level 10 with $\varepsilon_M = 0.25,\ \theta = 0.9$

Figure 19: Grids for the problem of Section 6.2.1 for $\varepsilon_M \in \{0.05, 0.25\}$ and $\theta = 0.9$ at level 10 with 2356 and 230786 degrees of freedom, respectively.

are comparable to uniform refinement agrees with what was found in [Joh98, Example 11]. Besides the degrees of freedom, the computational time decreases as well with decreasing ε_M and increasing θ. In particular, the shortest execution time was yielded by $\varepsilon = 0.05$ and $\theta = 0.5$ and needed only 6.5% of the time that uniform refinement required.

The figures of the levels that were the last levels before the memory barrier was hit indicate that again smaller values of ε_M with values of θ close to 1.0 are advantageous to achieve better results in terms of the quality of the computed solution. On the other hand, one can observe that for $\varepsilon_M = 0.05$ the cumulative time of the maximal levels is about 10 times as high as the time for the respective levels of $\varepsilon_M = 0.25$. This is the case, since for small values of ε_M and large values of θ, fewer cells are refined on each level and therefore more levels can be solved in total, where each subsequent level needs more time than the former ones and the number of cells increases. This can also be seen in Figure 19, where the grid of level 10 for $\varepsilon_M = 0.25$ possesses about 100 times as many degrees of freedom as the grid of level 10 for $\varepsilon = 0.05$. Further, on each level, a discrete needs to be solved which means that more levels require more solves and therefore more time. In particular, the number of solved levels for $\varepsilon = 0.05$ is about twice as high as the number of solved levels for $\varepsilon = 0.25$.

Further, one can observe the the differences for different values of θ and a fixed value of ε_M decrease with increasing ε_M. This behavior could be observed for the Hemker problem of Section 6.1.1 as well. But in contrast to the Hemker problem with an advantageous configuration of $\varepsilon_M = 0.25$, here larger values of θ together with smaller values of ε_M seem to work better. This might be the case, since the solution of the Hemker problem possesses layers which require more extensive refinement, whereas this problem possesses a point singularity and therefore requires very concentrated refinement. That larger values of θ with smaller ε_M yield sequences of grids whose refinement is more concentrated around the singularity can be observed in Figure 19.

In Figure 20 one can see the efficiency index (32) for $\varepsilon_M \in \{0.05, 0.25, 0.5\}$ and $\theta = 0.5$. The figure shows a strong over-estimation of the error and boundedness from below for $\varepsilon_M = 0.05$. Therefore, in this case the error estimator is reliable. On the other hand, for $\varepsilon_M \in \{0.25, 0.5\}$

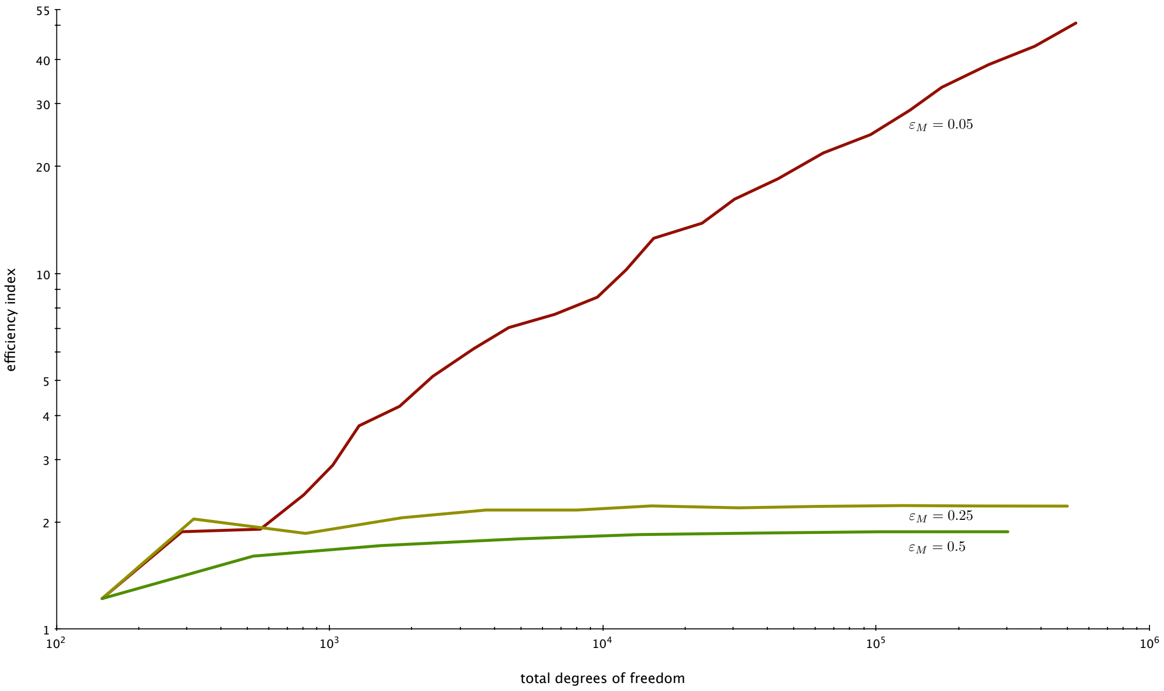


Figure 20: Efficiency index for the example of Section 6.2.1 for values of $\varepsilon_M \in \{0.05, 0.25, 0.5\}$ and $\theta = 0.5$.

the efficiency index is roughly 2 and is bounded from above and below, which means that the error estimator is not only reliable but also efficient.

6.2.2 Flow around a cylinder

In this example, a flow around a cylinder is considered. Goal is to compare the performance of uniform refinement, adaptive refinement, and the dual weighted residual method in terms of computing the drag coefficient. The drag coefficient is a concrete physical quantity describing the resistance of an object in a fluid environment. Although this example was introduced in [ST96] for the Navier–Stokes equations, here the Stokes equations are being considered.

The example's domain is given by

$$\Omega = [0\text{m}, 2.2\text{m}] \times [0\text{m}, H\,\text{m}] \cap S \text{ with } S = \{\mathbf{x}\text{m} : \mathbf{x} \in \mathbb{R}^2,\ \|\mathbf{x}\|_{\ell_2}^2 \leq 0.05\} + (0.155\text{m}, 0.155\text{m})^T,$$

where $H = 0.41$, as depicted in Figure 21. The boundary $\partial\Omega$ consists of relatively open components $\Gamma_{D,\text{in}}$, Γ_D, Γ_N and ∂S, where on Γ_D and ∂S homogeneous Dirichlet boundary conditions, on Γ_N homogeneous Neumann boundary conditions (outflow), and on $\Gamma_{D,\text{in}}$ Dirichlet boundary conditions (inflow) with

$$\mathbf{u}|_{\Gamma_{D,\text{in}}}(x, y) = 1.2\frac{y(H - y)}{H^2}\text{m}\,\text{s}^{-1}$$

are prescribed. The right-hand side $\mathbf{f}$ is set to be homogeneous and the viscosity is given by $\nu = 10^{-3}\text{m}^2\,\text{s}^{-1}$.

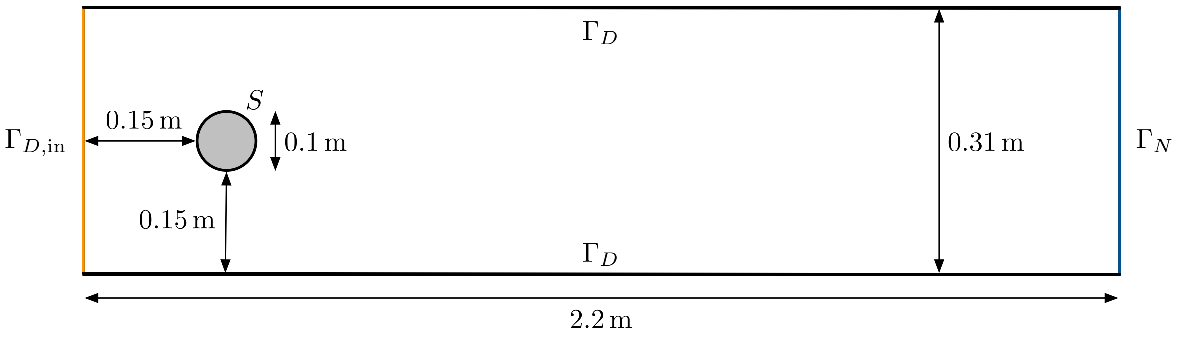


Figure 21: Sketch of the domain used in the example presented in Section 6.2.2.

The drag coefficient is a dimensionless quantity defined by

$$c_d := \frac{2}{\rho \overline{U}^2 D} \int_{\partial S} \left(\rho\nu \frac{\partial u_{\mathbf{t}_S}}{\partial \mathbf{n}} n_y - p n_x \right), \tag{97}$$

where $D = 0.1$m is the cylinder's diameter, $\rho = 1\mathrm{kg}\,\mathrm{m}^{-3}$ is the fluid's density,

$$\overline{U} = \frac{2}{3}\mathbf{u}(0, H/2)\mathrm{m}\,\mathrm{s}^{-1}$$

is the mean inflow velocity, $\mathbf{n} = (n_x, n_y)^T$ is the normal vector on ∂S directing into Ω and $u_{\mathbf{t}_S}$ is the tangential velocity for the tangential vector $\mathbf{t}_S = (n_y, -n_x)^T$.

Since the drag coefficient is defined by a path integral and for numerical simulations, ∂S needs to be approximated by a discrete boundary, the error in the quadrature might become large. To prevent this, one can transform the line integral into an integral over Ω by applying integration by parts.

To this end, one can see that

$$\begin{aligned} \frac{\partial u_1}{\partial \mathbf{t}} &= \nabla u_1 \cdot \mathbf{t} = 0 \Leftrightarrow \frac{\partial u_1}{\partial x} n_y = \frac{\partial u_1}{\partial y} n_x, \\ \frac{\partial u_2}{\partial \mathbf{t}} &= \nabla u_2 \cdot \mathbf{t} = 0 \Leftrightarrow \frac{\partial u_2}{\partial x} n_y = \frac{\partial u_2}{\partial y} n_x, \end{aligned}$$

since $\mathbf{u}$ is constant on ∂S, which yields together with $n_x^2 + n_y^2 = 1$ that

$$\frac{\partial u_{\mathbf{t}_S}}{\partial \mathbf{n}} = \nabla(\mathbf{u} \cdot \mathbf{t}) \cdot \mathbf{n} = \frac{\partial u_1}{\partial y} - \frac{\partial u_2}{\partial x}.$$

Since $\mathbf{u}$ is divergence-free, one obtains

$$\nabla \cdot \mathbf{u} = 0 \Leftrightarrow \frac{\partial u_1}{\partial x} = -\frac{\partial u_2}{\partial y}.$$

Using this identity together with the easier form of the normal derivative of the tangential velocity yields for $\boldsymbol{\psi} = (1, 0)^T$ that

$$c_d = \frac{2}{\rho \overline{U}^2 D} \int_{\partial S} \rho \mathbf{n}^T\, \mathbb{T}(\mathbf{u}, p) \boldsymbol{\psi},$$

which agrees with the definition of the drag coefficient in [BR03a, Section 11.1]. Inserting the values for $\overline{U}$, ρ and D and applying integration by parts yields

$$\begin{aligned} c_d &= 500 \int\limits_{\partial S} \mathbf{n}^T\, \mathbb{T}(\mathbf{u},p)\boldsymbol{\psi} = 500 \int\limits_{\Omega} \big(-\, \mathbb{T}(\mathbf{u},p)\nabla\overline{\boldsymbol{\psi}} + \nabla\cdot \mathbb{T}(\mathbf{u},p)\overline{\boldsymbol{\psi}} \big) \\ &= 500\,\big(-\nu(\nabla\mathbf{u},\nabla\overline{\boldsymbol{\psi}})_0 + (p\,\mathbb{I},\nabla\overline{\boldsymbol{\psi}})_0 + \nu(\Delta\mathbf{u},\overline{\boldsymbol{\psi}})_0 - (\nabla p,\overline{\boldsymbol{\psi}})_0 \big) \\ &= -500\nu(\nabla\mathbf{u},\nabla\overline{\boldsymbol{\psi}})_0 + 500(p\,\mathbb{I},\nabla\overline{\boldsymbol{\psi}})_0 + (\mathbf{f},\overline{\boldsymbol{\psi}})_0, \end{aligned}$$

where $\overline{\boldsymbol{\psi}}$ is an extension of ψ into the interior of Ω with support along ∂S. Thus, $(\mathbf{f},\overline{\boldsymbol{\psi}})_0 = 0$ and one arrives at the expression

$$c_d = -500\nu(\nabla\mathbf{u},\nabla\overline{\boldsymbol{\psi}})_0 + 500(p\,\mathbb{I},\nabla\overline{\boldsymbol{\psi}})_0 \tag{98}$$

for the drag which does not involve line integrals anymore. As ansatz and test space the finite element pair P_3/P_2 was chosen. When the adaptive algorithm was applied, it was applied using the maximum strategy with the minimum number of cells to be refined set to $\varepsilon_M = 25\%$ together with $\theta = 0.5$. If on a certain level not sufficiently many cells were marked for refinement, the strategy's parameter was relaxed by $\theta \to 0.7\theta$. The used error estimators were energy norm a posteriori error indicator (65) and the dual weighted residual error indicator of Theorem 3.32.

Concerning the dual weighted residual method, the error estimator was evaluated by approximation by difference quotients, see Section 3.2.3. The functional of interest was chosen to be the drag coefficient as given in (98).

6.1 Remark [Discretization of the dual problem]**.** The discretization of the dual problem requires special care in this case because it is in some points different from what was derived for the Stokes equations in Section 2.2.2. Since the drag coefficient (98) is linear in velocity and pressure, one obtains $J'((\mathbf{u},p))(\cdot) = J(\cdot)$. The extension $\overline{\boldsymbol{\psi}}$ is actually of the form $(\overline{\psi},0)^T$, as it extends the vector $(1,0)^T$. Let $q \in Q^h$ be a pressure test function and $\mathbf{v}_1 = v_1\mathbf{e}_1 \in \mathbf{V}^h$ be a velocity test function for the first spatial dimension. One then obtains

$$\begin{aligned} J((\mathbf{v}_1,q)) &= -500\nu \int\limits_{\Omega} \left(\begin{pmatrix} \frac{\partial v_1}{\partial x} & \frac{\partial v_1}{\partial y} \\ 0 & 0 \end{pmatrix} : \begin{pmatrix} \frac{\partial \overline{\psi}}{\partial x} & \frac{\partial \overline{\psi}}{\partial y} \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix} : \begin{pmatrix} \frac{\partial \overline{\psi}}{\partial x} & \frac{\partial \overline{\psi}}{\partial y} \\ 0 & 0 \end{pmatrix} \right) \\ &= -500\nu \int\limits_{\Omega} \left(\frac{\partial v_1}{\partial x}\frac{\partial \overline{\psi}}{\partial x} + \frac{\partial v_1}{\partial y}\frac{\partial \overline{\psi}}{\partial y} \right) + 500 \int\limits_{\Omega} q\frac{\partial \overline{\psi}}{\partial x}. \end{aligned}$$

For the second dimension consider $q \in Q^h$ and $\mathbf{v}_2 = v_2\mathbf{e}_2 \in \mathbf{V}^h$. Analogously, one obtains

$$J((\mathbf{v}_2,q)) = 500 \int\limits_{\Omega} q\frac{\partial \overline{\psi}}{\partial x}.$$

In particular this means that the arising linear system looks like

$$\begin{pmatrix} A & -B^T \\ -B & 0 \end{pmatrix} \begin{pmatrix} \underline{\mathbf{u}} \\ \underline{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \underline{\mathbf{f}}_1 \\ \underline{\mathbf{f}}_2 \end{pmatrix}$$

with

$$(\underline{\mathbf{f}}_1)_{i=1}^{N_u} = -500\nu \int\limits_{\Omega} \left(\frac{\partial v_{1,i}}{\partial x}\frac{\partial \overline{\psi}}{\partial x} + \frac{\partial v_{1,i}}{\partial y}\frac{\partial \overline{\psi}}{\partial y} \right), \quad (\underline{\mathbf{f}}_1)_{i=N_u+1}^{2N_u} = 0\cdot v_{2,i}, \quad (\underline{\mathbf{f}}_2)_{i=1}^{N_p} = 500 \int\limits_{\Omega} q_i\frac{\partial \overline{\psi}}{\partial x},$$

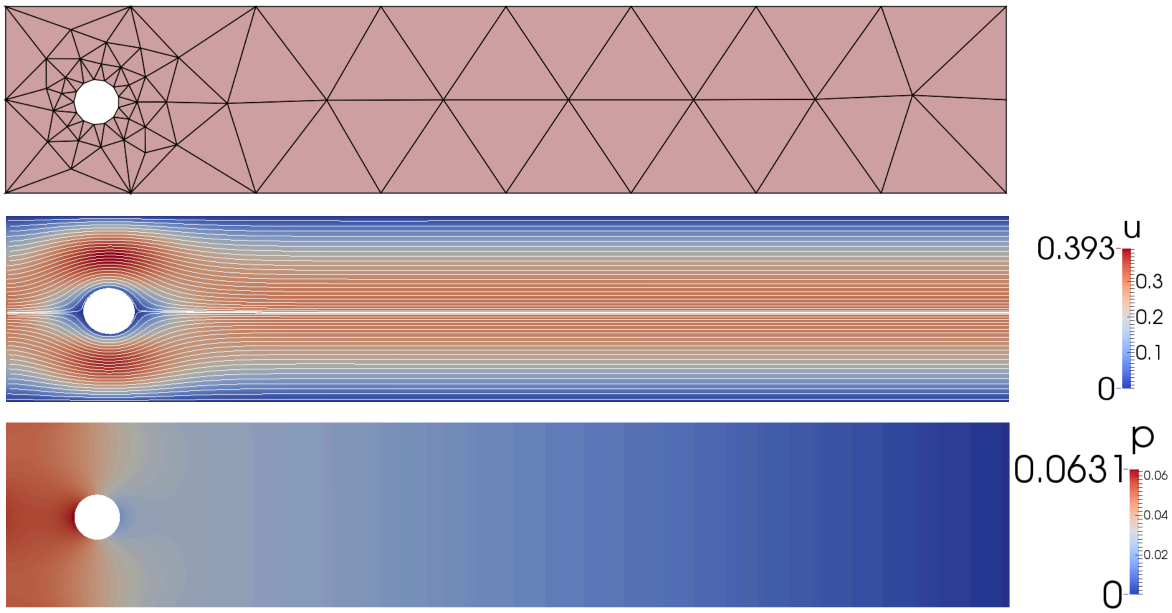


Figure 22: Initial triangulation of the domain (top) with 1056 total degrees of freedom and 99 cells, computed velocity solution (middle) and computed pressure solution (bottom) of the "flow around a cylinder"-example of Section 6.2.2.

where $v_{1,i}\mathbf{e}_1 \in \mathbf{V}^h$, $v_{2,i}\mathbf{e}_2 \in \mathbf{V}^h$ and $q_i \in Q^h$ are the basis functions of the finite-dimensional ansatz and test spaces. Thus, the dual problem is a discretization of the compressible Stokes equations with inverted pressure field and therefore also inverted direction of flow.

The solution of the problem and its initial triangulation can be seen in Figure 22. One can see that the flow is laminar, i.e., a flow without lateral mixing. The pressure is high left of the cylinder and decaying right of the cylinder, indicating a mean flow direction from left to right.

The a posteriori error indicator in the energy norm and the dual weighted residual method both yield a sequence of adaptively refined grids, of which to level 2, 4, 6, and 8 corresponding elements are depicted in Figure 23. One can see that the dual weighted residual error indicator contains the refinement much more around the cylinder, whereas the a posteriori error indicator also refines areas close to the right part of the domain's boundary.

When comparing the corresponding degrees of freedom and number of cells in Table 5, one can see that the dual weighted residual error estimator uses fewer cells per level, but the above statement about the more local refinement still holds: The error estimator (65) yields on level 6 a grid with about 6500 cells and refined in areas close to the right boundary of the domain, whereas the dual weighted residual error estimator yields a more concentrated refinement on level 8 with about 16000 cells.

In Figure 24 one can see that the dual weighted residual error estimator indeed works better than the energy norm estimator (65) and also better than uniform refinement in terms of memory that is required to achieve a certain accuracy of the drag coefficient.

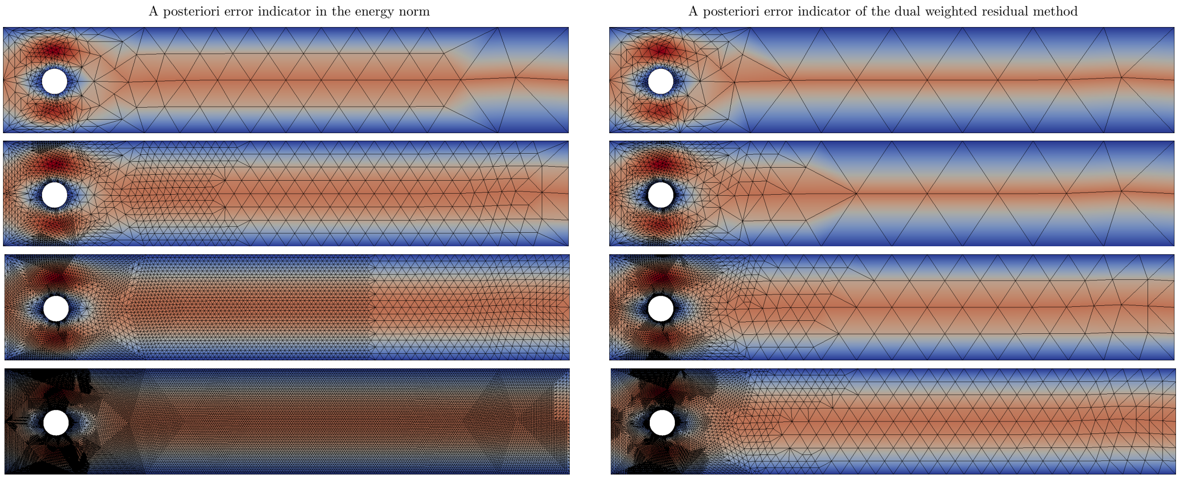


Figure 23: Levels 2, 4, 6 and 8 (top to bottom) of the generated sequence of adaptively refined grids for flow around a cylinder when applying the energy norm estimator (65) on the left and the dual weighted residual error indicator of Theorem 3.32 on the right. The respective number of cells and total degrees of freedom can be found in Table 5.

Table 5: Table showing the number of cells and total degrees of freedom for the grids shown in Figure 23.

	estimator (65)		DWR estimator	
	# cells	total d.o.f.	# cells	total d.o.f.
Level 2	252	2279	232	2104
Level 4	1373	12763	981	9067
Level 6	6467	60765	3989	37093
Level 8	31760	300260	15858	147853

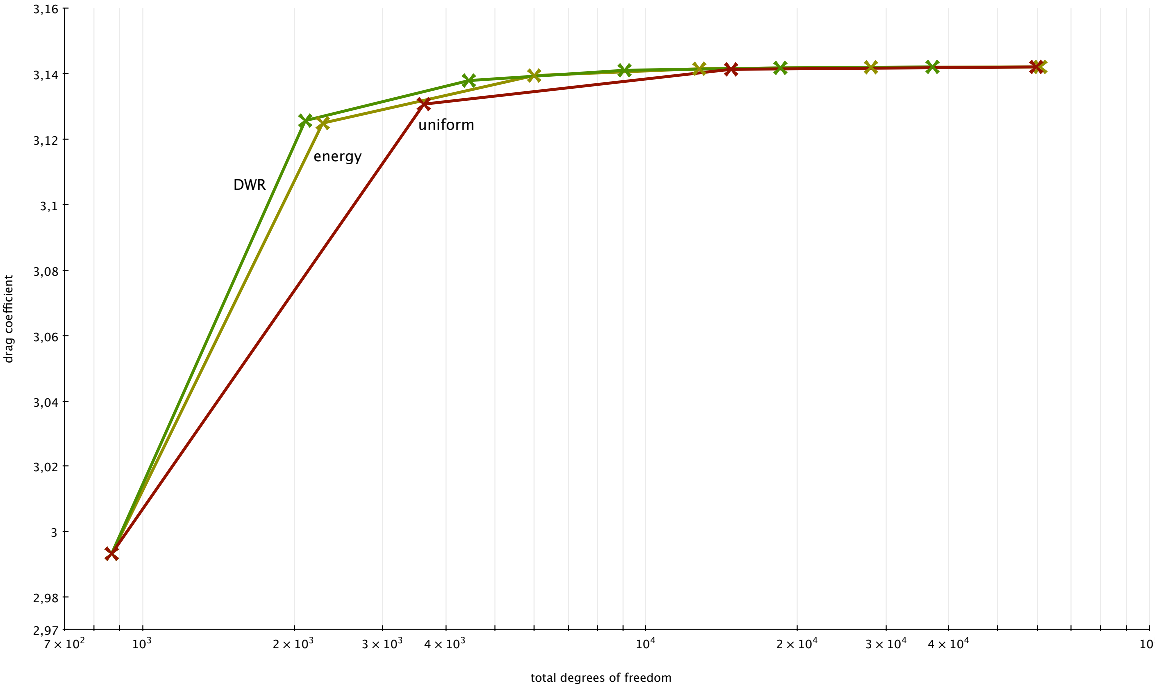


Figure 24: Figure displaying for the example of Section 6.2.2 the computed drag coefficient over the total degrees of freedom for uniform refinement (red), refinement with the energy norm estimator (yellow) and refinement with the dual weighted residual estimator (green). The crosses indicate the current level of the algorithm.

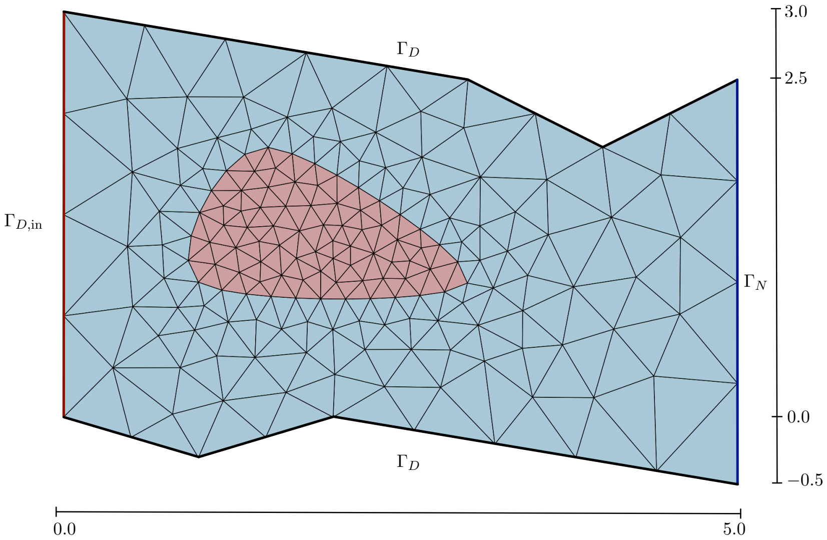


Figure 25: Initial triangulation of the domain of the Stokes–Darcy example of Section 6.3. The blue cells belong to the Stokes subdomain, the red cells belong to the Darcy subdomain.

To investigate further the differences of the three methods, the iteration was stopped as soon as the absolute change of the drag coefficient was smaller than 10^{-4}. In the below table one can see the corresponding number of cells and degrees of freedom as well as the required time:

	c_d	total d.o.f.	# cells	time
energy estimator	3.14219	123764	13135	7.36 s
DWR estimator	3.14221	75282	8080	7.85 s
uniform refinement	3.14224	960814	101376	63.49 s

One can see that the dual weighted residual estimator needs about 60% of the degrees of freedom / cells of the energy estimator and about 8% of the degrees of freedom / cells of uniform refinement to obtain a comparable result. Also the required time on a single-core 3.5 GHz CPU for the dual weighted residual estimator is only slightly higher than for the energy estimator, even though it involves assembling and solving a dual problem.

Altogether the dual weighted residual method seems to be better suited when one is interested in a concrete physical quantity than the energy norm estimator. If the considered problem is nonlinear, the performance of the dual weighted residual method might even increase compared to the other methods because the dual problem is linear and therefore easier to solve.

6.3 Stokes–Darcy problem

This section deals with an example of the coupled Stokes–Darcy problem that was introduced in Section 2.3. Here, the Darcy subdomain $\Omega_p \subset \Omega$ is entirely enclosed in the Stokes subdomain $\Omega_f \subset \Omega$, as sketched in Figure 25.

The outer boundary $\partial\Omega$ consists of relatively open components $\Gamma_D, \Gamma_{D,\text{in}}$ and Γ_N, that correspond to the bold black, red and blue lines in Figure 25, respectively. On Γ_D and Γ_N

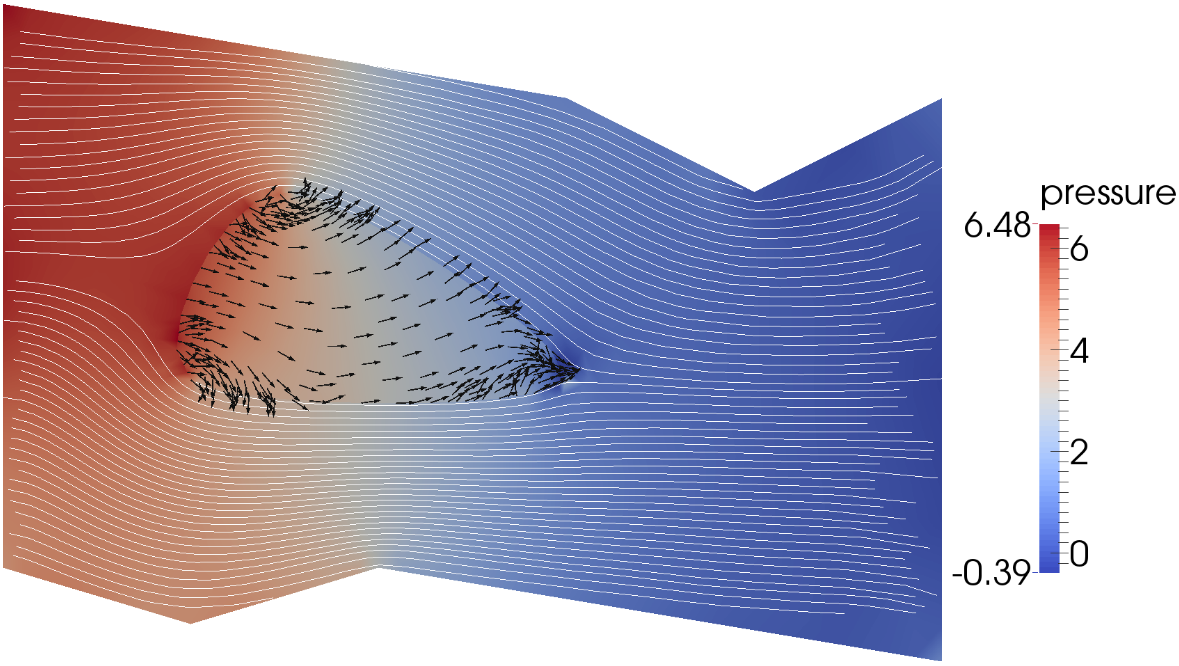


Figure 26: Solution of the Stokes–Darcy problem of Section 6.3. The color indicates the pressure and the Stokes velocities are indicated by the streamlines. The Darcy velocity was recovered by $\mathbf{u}_p = -\mathbb{K}\,\nabla\varphi_p$ and is indicated by unscaled arrows in the Darcy subdomain.

homogeneous Dirichlet and Neumann boundary conditions are prescribed and on $\Gamma_{D,\mathrm{in}} = \{0\} \times [0,3]$ the Stokes velocity is prescribed by

$$\mathbf{u}_f\big|_{\Gamma_{D,\mathrm{in}}}(x,y) = 3y(1-3y).$$

The Darcy subdomain is coupled to the Stokes subdomain by the Neumann–Neumann coupling, see (27). Due to the enclosedness of the Darcy subdomain, its boundary conditions are completely determined by the coupling. The viscosity is set to $\nu = 1$, the permeability is set to $\mathbb{K} = 10^{-3}\,\mathbb{I}$ and the right-hand sides of the equations in (27) are homogeneous. The computations were carried out using P_2/P_1 finite elements for the Stokes subproblem and P_1 finite elements for the Darcy subproblem.

To apply the adaptive algorithm, the errors were estimated separately with the energy norm a posteriori indicator (65) for the Stokes subdomain and the energy norm a posteriori indicator (53) for the Darcy subdomain and then marked for refinement according to two separate marking strategies with $\varepsilon_M = 0.25$, $\theta = 0.5$ and an update of $\theta \to 0.7\theta$ if not sufficiently many cells were marked. If a cell at the interface of one subdomain was marked for refinement, the opposite cell in the other subdomain was not necessarily marked for refinement as well. The numerical solution of the problem can be seen in Figure 26.

One can observe that due to the pressure drop the mean flow direction is from left to right and that the solution changes most in the corners of the Darcy subdomain. The interaction between the Stokes and Darcy subdomains remains local, i.e., the Stokes flow around the Darcy subdomain is still laminar. Most of the refinement is located indeed in and around corners of the Darcy subdomain, even though a refinement at the interface of the Darcy subdomain did not imply a refinement at the interface of the Stokes subdomain, whereas the interior if the Darcy subdomain as well as the portion of the Stokes subdomain right of the Darcy subdomain remain almost untouched, see Figure 27.

This example demonstrates that one can transfer the error estimators of simpler, uncoupled situations to more complex, coupled situations by just applying them separately together with a separate marking strategy. Since all the parameters of the marking strategy are percentage values, the separate marking strategies behave like one marking strategy for the whole domain, just with taking into account the magnitude of the estimated errors in the subdomains.

7 Conclusion and outlook

This thesis discussed adaptive finite element methods for Stokes, Darcy, convection–diffusion equations and the Stokes–Darcy system. In particular, residual based as well as dual weighted residual error estimators have been derived for the Stokes and the Darcy equations. The residual based error estimators were then applied to the coupled Stokes–Darcy problem. The considered error estimators for the convection–diffusion equations were not derived but can be found in the literature.

The numerical examples included two examples for the Stokes equations and for the convection–diffusion equations each, where one example considered a residual based a posteriori error estimator and one example considered a dual weighted residual estimator for both types of equations. It was discussed how parameters ε_M and θ of the maximum cell marking strategy influenced the efficiency of the adaptive algorithm as well as how the residual based error estimators compare to the dual weighted residual error estimators. Further, one example for the coupled Stokes–Darcy problem was discussed.

Considering the “Hemker problem”, which is a convection–diffusion problem, it turned out that values of $\varepsilon_M \approx 25\%$ are advantageous in terms of memory requirements and needed computation time, whereas for the Stokes problem “disc with a crack”, values of $\varepsilon_M \approx 5\%$ and $\theta \approx 90\%$ worked best. However, this is not necessarily caused by the type of considered partial differential equations but is more likely due to the different features of the solutions. In particular, the solution of the Hemker problem contains layers in contrast to the Stokes problem, which possesses a point singularity. The layers required a larger area to be adaptively refined in order to obtain a sufficiently well approximated solution; the point singularity required most of the refinement around that single point. Consequently, a value of $\varepsilon_M \approx 25\%$ seems to be more suitable when larger areas need to be refined. The corresponding value of θ is of little significance because it is dominated by ε_M as long as the error is still concentrated in a relatively small portion of the grid. On the other hand, small values of ε_M and large values of θ cause a increasingly local refinement of the grid which is more suitable for problems like the Stokes problem “disc with a crack”.

The weights of the dual weighted residual method were evaluated by an approximation by difference quotients and applied to the convection–diffusion equations for an example that includes boundary layers and to the Stokes equations with the goal to compute a concrete physical quantity. The functional of interest for the convection–diffusion example was chosen to be the function average value in a small area away from the layers. It turned out that the layers remain unresolved and most of the refinement concentrates around the area of interest, nevertheless a strong over-estimation could be observed which might be due to the implementation of the dual weighted residual method with difference quotients, as reported in the literature, see [BR03a, Section 4.1]. For the computation of the concrete physical quantity in the Stokes example, the dual weighted residual method with a functional of interest representing that exact quantity proved to be superior to the residual based a posteriori

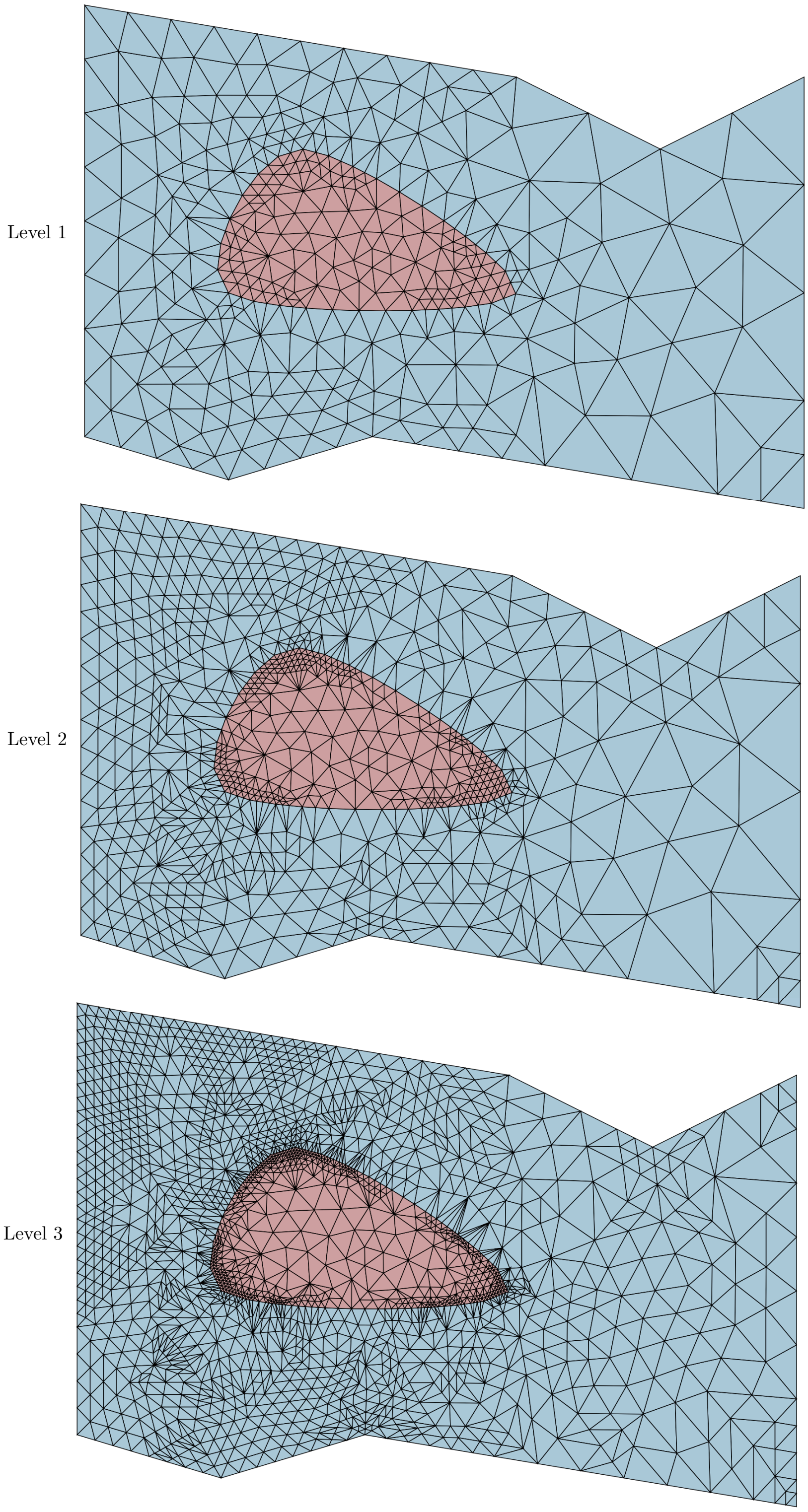


Figure 27: Grids corresponding to levels 1, 2 and 3 of the adaptive algorithm for the Stokes–Darcy example of Section 6.3.

error estimator in the energy norm with respect to required memory in order to obtain a comparable accuracy. With respect to required computational time comparable results were yielded, even though the dual weighted residual method additionally involves assembling and solving a dual problem. The performance might even increase when considering a non-linear problem, as the dual problem is still linear and therefore easier to solve.

Finally, the residual based error estimators in the energy norm for the Darcy equations as well as for the Stokes equations were combined in the more complicated example of the coupled Stokes–Darcy system. It turned out that a possible way of applying them is to apply them separately in their corresponding subdomain and with separate cell marking strategies which were both configured with the same values of ε_M and θ, yielding sensible grids.

Altogether, the adaptive algorithm was successfully applied to Stokes and convection–diffusion equations for residual based a posteriori error-estimators and the dual weighted residual estimators and yielded a significant performance increase. It was also applied for the more complicated situation of a coupled Stokes–Darcy system. The findings in the Section about the Stokes example "disc with a crack" agreed qualitatively with what can be found in [Joh98, Example 11]. It turns out that the dual weighted residual method is more suitable than the considered residual based a posteriori error estimators when one is interested in concrete quantities which are not the solution itself. The optimal choice of ε_M and θ in the marking strategy depends on the considered example.

Further investigations could consider combining the adaptive algorithm with iterative solvers and assessing the performance of the adaptive algorithm when applied to problems with more than two spatial dimensions as well as to non-linear and time-dependent problems. It could also be investigated, how different cell marking strategies compare to each other. Concerning the coupled Stokes–Darcy system, one could derive an error estimator for the whole coupled domain and compare it to the application of the separate error estimators. Such an error estimator for the whole domain in the more complicated case of the Navier–Stokes–Darcy system is investigated in [HAN14].

The dual weighted residual method's weights could be evaluated differently, possibly yielding better results. Also it could be analyzed in how far the grids generated by the dual weighted residual method are eligible for solving the dual problem. A possible extension of the discussed adaptive algorithm is the use of anisotropic mesh refinement, i.e., the use of streched mesh cells along regions of rapid change of the solution like layers which is discussed in, e.g., [Ver13]. The use of these cells can result in more memory-efficient algorithms.

References

[ACF$^+$11] Matthias Augustin, Alfonso Caiazzo, André Fiebach, Jürgen Fuhrmann, Volker John, Alexander Linke, and Rudolf Umla. An assessment of discretizations for convection-dominated convection-diffusion equations. *Computer Methods in Applied Mechanics and Engineering*, 200(47-48):3395–3409, November 2011.

[Bog07] Vladimir I. Bogachev. *Measure Theory*. Number v. 1 in Measure Theory. Springer Berlin Heidelberg, 2007.

[BR03a] Wolfgang Bangerth and Rolf Rannacher. *Adaptive finite element methods for differential equations*. Lectures in Mathematics. Springer, 2003.

[BR03b] Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–103, January 2003.

[BSW83] Randolph E. Bank, Andrew H. Sherman, and Alan Weiser. Some refinement algorithms and data structures for regular local mesh refinement. *Scientific Computing*, 1983.

[BW90] Randolph E. Bank and Bruno D. Welfert. A posteriori error estimates for the Stokes equations: A comparison. *Computer methods in applied mechanics and mechanical engineering*, 82(1-3):323–340, 1990.

[Cia91] Philippe G Ciarlet. Basic error estimates for elliptic problems. In *Finite Element Methods (Part 1)*, pages 17–351. Elsevier, 1991.

[Dar56] Henry Darcy. Les fontaines publiques de la ville de Dijon, 1856.

[Dör96] Willy Dörfler. A Convergent Adaptive Algorithm for Poisson's Equation. *SIAM Journal on Numerical Analysis*, 33(3):1106–1124, June 1996.

[DQ09] Marco Discacciati and Alfio Quarteroni. Navier-stokes/darcy coupling: modeling, analysis, and numerical approximation. *Revista Matemática Complutense*, 22(2), 2009.

[EMT04] Yuli Eidelman, Vitali Milman, and Antonis Tsolomitis. *Functional Analysis: An Introduction*. Graduate studies in mathematics. American Mathematical Society, 2004.

[GHT04] Matthias Grajewski, Jaroslav Hron, and Stefan Turek. Dual Weighted a Posteriori Error Estimation for a New Nonconforming Linear Finite Element on Quadrilaterals, January 2004. Ergebnisberichte des Instituts für Angewandte Mathematik, Nummer 247.

[GKR13] Vivette Girault, Guido Kanschat, and Béatrice Rivière. On the Coupling of Incompressible Stokes or Navier–Stokes and Darcy Flows Through Porous Media. In José A Ferreira, Sílvia Barbeiro, Gonçalo Pena, and Mary F Wheeler, editors, *Modelling and Simulation in Fluid Dynamics in Porous Media*, pages 1–25. Springer, 2013.

[GR81] Vivette Girault and Pierre-Arnaud Raviart. *Finite element approximation of the Navier-Stokes equations*. Springer-Verlag, Berlin, 1981.

[Gru07] Peter M. Gruber. *Convex and Discrete Geometry.* Springer Verlag, June 2007.

[HAN14] M. L. Hadji, A. Assala, and F. Z. Nouri. A posteriori error analysis for Navier–Stokes equations coupled with Darcy problem. *Calcolo*, 52(4):559–576, 2014.

[Hem96] P. W. Hemker. A singularly perturbed model problem for numerical computation. *Journal of Computational and Applied Mathematics*, 76(1-2):277–285, 1996.

[JM04] Volker John and Gunar Matthies. MooNMD – a program package based on mapped finite element methods. *Computing and Visualization in Science*, 6(2-3):163–170, January 2004.

[Joh98] Volker John. A posteriori L^2-error estimates for the nonconforming P_1/P_0-finite element discretization of the Stokes equations. *Journal of Computational and Applied Mathematics*, 1998.

[Joh00] Volker John. A numerical study of a posteriori error estimators for convection–diffusion equations. *Computer Methods in Applied Mechanics and Engineering*, 190(5-7):757–781, November 2000.

[JS14] Volker John and Liesel Schumacher. A study of isogeometric analysis for scalar convection–diffusion equations. *Applied Mathematics Letters*, 2014.

[MS64] Norman G Meyers and James Serrin. H=W. In *Proceedings of the National Academy of Science*, volume 51, pages 1055–1056, 1964.

[Nef02] Patrizio Neff. On Korn's first inequality with non-constant coefficients. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 132(01):221–243, 2002.

[ST96] M. Schäfer and S. Turek. Benchmark computations of laminar flow around a cylinder. In *Flow Simulation with High-Performance Computers II*, volume 48 of *Notes on Numerical Fluid Mechanics (NNFM)*, pages 547–566. Vieweg+Teubner Verlag, 1996.

[Ste08] Olaf Steinbach. *Numerical Approximation Methods for Elliptic Boundary Value Problems.* Finite and Boundary Elements. Springer Science & Business Media, 2008.

[Tem01] Roger Temam. *Navier-Stokes Equations: Theory and Numerical Analysis.* AMS/Chelsea publication. AMS Chelsea Pub., 2001.

[Ver98] Rüdiger Verfürth. A posteriori error estimators for convection-diffusion equations. *Numerische Mathematik*, 80(4):641–663, 1998.

[Ver13] Rüdiger Verfürth. *A Posteriori Error Estimation Techniques for Finite Element Methods*, volume 1. OUP Oxford, Apr 2013.

[Yos80] Kōsaku Yoshida. *Functional Analysis.* Springer Science & Business Media, Berlin, Heidelberg, 1980.