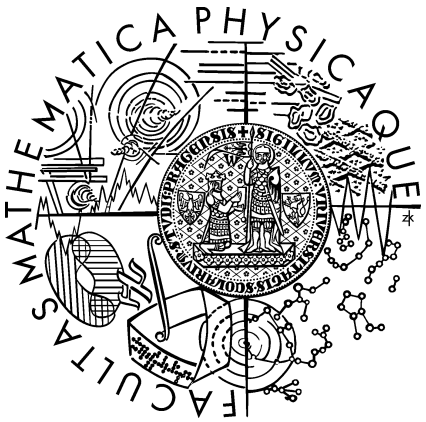
FACULTY OF MATHEMATICS AND PHYSICS
CHARLES UNIVERSITY, PRAGUE



Lecture notes
Course NMNV464

A posteriori numerical analysis based on the method of equilibrated fluxes

Martin VOHRALÍK

April 8, 2024

Preface

These lecture notes were written for the course NM497 **A posteriori error estimates for efficiency and error control in numerical simulations** held at the Sorbonne University (previously Université Pierre et Marie Curie) in the spring semester, academic year 2010/2011 and for the course NMNV464 **A posteriori numerical analysis based on the method of equilibrated fluxes** held at the Faculty of Mathematics and Physics, Charles University, Prague, academic year 2011/2012. Many ideas were taken from the existing books on the subject, namely those of Verfürth [93], Ainsworth and Oden [6], Babuška and Strouboulis [13], Neittaanmäki and Repin [75], Han [63], and Repin [85] and also from recent literature. The principal idea can be traced back to at least Prager and Synge [81] but the presentation is rather independent and done along the lines of the recent work of the author and his collaborators. The material of Chapters 1–8 is presented as self-contained and with as much details as possible; Chapters 9–13 then give an outlook into more complex applications with only the main ideas presented; the details can be found in the cited literature.

Introduction

A large number of environmental and physical phenomena is described by partial differential equations. Unfortunately, in the vast majority of cases, it is not possible to find the analytical, *exact solutions* of these equations. Then numerical methods, mathematically-based algorithms evaluated with the aid of computers, are used as simulation tools.

Numerical methods typically only deliver *approximate solutions*, functions defined in some finite-dimensional spaces, different from the exact solutions. Then two extremely important questions are:

1. How large is the overall error between the exact and approximate solutions?
2. Where in space and in time is the error localized?

Answers to these two questions may be crucial in building bridges and dams, constructing cars and planes, advanced health care techniques, drugs conception, population dynamics simulations, economic and financial predictions, weather forecast, drilling oil and natural gas, depollution of soils and oceans, etc., as a decision is often taken on the basis of the numerical simulation result.

Taking this reflection one step further, the ultimate goal in scientific computing is to design algorithms such that:

1. a precision, given before the simulation start, is attained at the end of the simulation (*precision attainment*);
2. as small as possible amount of computational work is needed (*efficiency*).

The purpose of these lecture notes is to introduce the theory of *a posteriori error estimation*. In particular, the estimates presented in these lecture notes

i) give a fully computable upper bound on the overall error between the unknown exact solution and the known approximate numerical approximation (*error control*);
ii) predict the error at each simulation time and in each part of the simulation domain (*error localization*),

so that they can give answers to the questions **1.**–**2.** above. Moreover, these estimates

i) enable to distinguish and estimate separately the different error components (*error components identification and separation*);
ii) allow to adjust optimally the calculation parameters during the simulation (*adaptivity*),

which leads to algorithms satisfying the properties **1.**–**2.** above. We develop them in a unified framework, applicable to all standard numerical methods.

Contents

List of Figures

Chapter 1

Partial differential equations, numerical methods, and error estimation

We give in this chapter examples of partial differential equations, recall the principle of numerical methods, and introduce the concept of a posteriori error estimation. In view of our interest in numerical methods, we suppose that $\Omega \subset \mathbb{R}^d$, $d \geq 1$, is an open polytope (polygon for $d = 2$ and polyhedron for $d = 3$) throughout these lecture notes; Ω is thus open, bounded, and connected.

1.1 Examples of partial differential equations

We give here several examples of model partial differential equations.

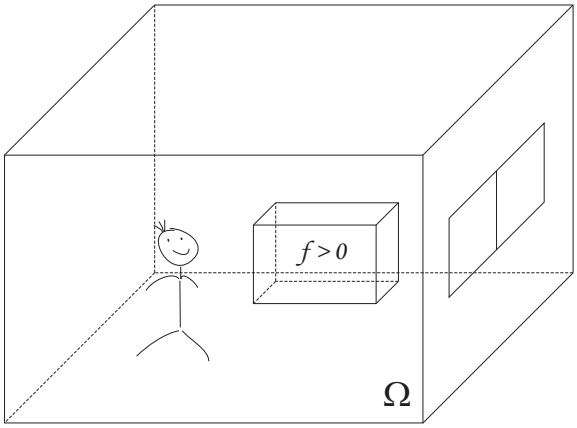


Figure 1.1: A room with a heater

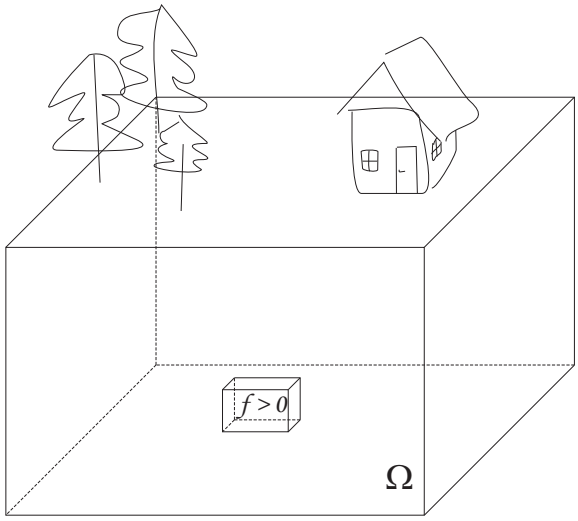


Figure 1.2: Underground with a water well

1.1.1 The Laplace equation

The Poisson problem for the Laplace equation consists in finding, for a given function $f : \Omega \to \mathbb{R}$, the function $u : \Omega \to \mathbb{R}$ such that

$$-\Delta u = f \qquad \text{in } \Omega, \tag{1.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{1.1b}$$

where Δ stands for the Laplacian,

$$\Delta v := \sum_{i=1}^{d} \partial^2_{\mathbf{x}_i} v.$$

Let us mention two examples of physical phenomena modeled by the system (1.1a)–(1.1b):

Example 1.1.1 (Heat flow)**.** *Let Ω be a parallelepiped representing a room, see Figure 1.1. Let f be a function which is zero everywhere in Ω except of the small box indicated, which represents the source of the heat (heater). Then u stands for the heat (temperature) in the room Ω. More precisely, in practice, we switch the heater on at some moment, then the temperature u will start to increase, but will eventually reach an equilibrium, steady state. Model (1.1a)–(1.1b) describes precisely this equilibrium. Let us also remark that (1.1b) means that we suppose that the temperature at the walls of the room (on the boundary of Ω) is zero.*

Example 1.1.2 (Underground water flow)**.** *Suppose that Ω represents a part of the underground as in Figure 1.2. In this model, u represents the so-called piezometric head, and $\boldsymbol{\sigma} := -\nabla u$ is the Darcy velocity of the underground water flow. The function f then stands for the sources: a water well in the present context.*

1.1.2 The advection–diffusion–reaction equation

The problem (1.1a)–(1.1b) only contains the Laplace operator Δ, describing *diffusion*. More precisely, diffusion can be modeled by a term $-\nabla\cdot(\underline{\mathbf{K}}\nabla u)$, where ∇ stands for the gradient,

$$\nabla v := (\partial_{\mathbf{x}_1} v, \dots, \partial_{\mathbf{x}_d} v)^t,$$

$\nabla\cdot$ for the divergence,

$$\nabla\cdot\mathbf{v} = \sum_{i=1}^{d} \partial_{\mathbf{x}_i} \mathbf{v}^i,$$

and where the diffusion–dispersion tensor $\underline{\mathbf{K}} : \Omega \to \mathbb{R}^{d\times d}$ describes the size and orientation of the diffusive effects. Let an advective vector field $\mathbf{w} : \Omega \to \mathbb{R}^d$ and a reaction function $r : \Omega \to \mathbb{R}$ be given. Enriching (1.1a)–(1.1b) by the corresponding *advection* and *reaction* effects, we obtain the advection–diffusion–reaction problem: find the function $u : \Omega \to \mathbb{R}$ such that

$$-\nabla\cdot(\underline{\mathbf{K}}\nabla u) + \nabla\cdot(\mathbf{w}u) + ru = f \qquad \text{in } \Omega, \tag{1.2a}$$
$$u = 0 \qquad \text{on } \partial\Omega. \tag{1.2b}$$

Example 1.1.3 (Contaminant transport)**.** *Suppose that Ω represents a part of the underground as in Figure 1.2. In contrast to Example 1.1.2, f is here the source of a pollution of Ω by some contaminant, whose concentration in Ω is given by the unknown function u. Then equation (1.2a) says that the contaminant is diffused with an intensity and orientation given by the diffusion–dispersion tensor $\underline{\mathbf{K}}$, advected by the velocity field $\mathbf{w}$, and undergoes a reaction described by the reaction function r.*

1.1.3 The Stokes equation

Let a source function $\mathbf{f} : \Omega \to \mathbb{R}^d$ (a vector here) be given. The Stokes problem consists in finding the vector function $\mathbf{u} : \Omega \to \mathbb{R}^d$ and the scalar function $p : \Omega \to \mathbb{R}$ such that

$$-\Delta\mathbf{u} + \nabla p = \mathbf{f} \qquad \text{in } \Omega, \tag{1.3a}$$
$$\nabla\cdot\mathbf{u} = 0 \qquad \text{in } \Omega, \tag{1.3b}$$
$$\mathbf{u} = \mathbf{0} \qquad \text{on } \partial\Omega; \tag{1.3c}$$

by $-\Delta\mathbf{v}$, we understand the componentwise Laplacian,

$$(-\Delta\mathbf{v})^i := -\Delta\mathbf{v}^i \qquad i = 1, \dots, d.$$

In contrast to (1.1a)–(1.1b) and to (1.2a)–(1.2b), (1.3a)–(1.3c) is a system of equations: we are looking for a d-component vector $\mathbf{u}$ and, simultaneously, for the scalar p.

Example 1.1.4 (Stokes flow)**.** *The description of water flow by equations (1.1a)–(1.1b) presented in Example 1.1.2 may not be sufficiently precise in many cases. Then, the Stokes model (1.3a)–(1.3c) can be used, with $\mathbf{u}$ standing for the water velocity and p for the water pressure.*

1.1.4 The heat equation

The models presented so far in Sections 1.1.1–1.1.3 all describe a steady state phenomenon. Let now a final simulation time $T > 0$ be given and consider an unsteady problem on the time interval $(0, T)$: for a given source function $f : \Omega \times (0, T) \to \mathbb{R}$ and for a given initial condition $u_0 : \Omega \to \mathbb{R}$, find the space–time function $u : \Omega \times (0, T) \to \mathbb{R}$ such that

$$\partial_t u - \Delta u = f \qquad \text{in } \Omega \times (0, T), \tag{1.4a}$$
$$u = 0 \qquad \text{on } \partial\Omega \times (0, T), \tag{1.4b}$$
$$u(\cdot, 0) = u_0 \qquad \text{in } \Omega. \tag{1.4c}$$

Example 1.1.5 (Heat problem)**.** *Let, as in Example 1.1.1, Ω be a parallelepiped representing a room, see Figure 1.1. Then (1.4a)–(1.4c) describes the* unsteady heat flow*. In contrast to Example 1.1.1, we can trace the evolution of the temperature u over the time interval $(0,T)$. The initial condition u_0 describes the initial temperature in the room.*

1.1.5 The nonlinear Laplace equation

The models presented so far in Sections 1.1.1–1.1.4 all describe a linear phenomenon. Let us now consider an example of a nonlinear problem. Let $a : \mathbb{R}_+ \to \mathbb{R}$ be a given nonlinear function. Typically, $a(x) = x^{p-2}$ for some real number $p \in (1, +\infty)$. Let $\boldsymbol{\sigma} : \mathbb{R}^d \to \mathbb{R}^d$ take the form

$$\boldsymbol{\sigma}(\boldsymbol{\xi}) = a(|\boldsymbol{\xi}|)\boldsymbol{\xi} \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^d, \tag{1.5}$$

where $|\cdot|$ is the Euclidean norm in $\mathbb{R}^d$. Then, for a given source function $f : \Omega \to \mathbb{R}$, the nonlinear Laplace problem consists in looking for $u : \Omega \to \mathbb{R}$ such that

$$-\nabla\cdot\boldsymbol{\sigma}(\nabla u) = f \qquad \text{in } \Omega, \tag{1.6a}$$
$$u = 0 \qquad \text{on } \partial\Omega. \tag{1.6b}$$

Example 1.1.6 (Nonlinear underground water flow)**.** *The problem (1.6a)–(1.6b) represents, for instance, the extension of the model problem of Example 1.1.2 which takes into account the nonlinear dependence of the Darcy velocity on the pressure head gradient ∇u. Note that (1.6a)–(1.6b) and (1.1a)–(1.1b) coincide, for $a(x) = x^{p-2}$, when $p = 2$.*

1.2 Numerical methods

For all the model problems presented in Sections 1.1.1–1.1.5, it is typically impossible to find the exact solution (u or the couple $(\mathbf{u}, p)$). Thus, numerical methods are used to find an approximate solution. Such methods rely on a notion of a spatial mesh, a partition of the domain Ω into elements that we call K. Herein, we suppose that the elements K are simplices (triangles in two space dimensions and tetrahedra in three space dimensions). We also suppose that the intersection of two elements K and K' is either an empty set, their common vertex, or their common d'-face, $d' = 1, \dots, d-1$ (i.e., edge in two space dimensions and an edge or a face in three space dimensions). The letter h stands for the maximal diameter of the elements and $\mathcal{T}_h$ for the mesh itself. For evolutive problems such as the heat problem of Section 1.1.4, we will also introduce the temporal mesh of the time interval $(0,T)$, consisting of intervals with the maximal size (time step) τ. For each discrete time t^n, $0 \le n \le N$, there is possibly a different mesh $\mathcal{T}_h^n$. We will typically denote the approximate solutions u_h for the problems (1.1a)–(1.1b), (1.2a)–(1.2b), and (1.6a)–(1.6b), by $u_{h\tau}$ for the problem (1.4a)–(1.4c), and by $(\mathbf{u}_h, p_h)$ for the problem (1.3a)–(1.3c).

1.3 A priori error estimates

Traditionally, the quality of numerical solutions is expressed with the aid of *a priori error estimates*. These estimates have typically, for steady problems, the form

$$|||u - u_h||| \le Ch^k, \tag{1.7}$$

where $C > 0$ and $k > 0$ are constants and $|||\cdot|||$ is some norm. We recall that u is the exact solution, u_h the approximate solution, and h the mesh size. Typically a system of

linear algebraic equations needs to be solved in order to obtain u_h and we suppose in (1.7) that this linear system was solved "exactly" (say to machine precision). It can be concluded from (1.7) that the error between u and u_h goes to zero as h goes to zero (with the order k), which justifies the numerical method in question: when we refine the mesh $\mathcal{T}_h$ (decrease the maximal element size h) or improve the precision of the numerical method (increase the order k), the approximate solution approaches the exact one. Unfortunately, the constant C in (1.7) typically depends on the exact solution u, $C = C(u)$, and is unknown. Thus, the quantity Ch^k cannot be evaluated in practice and one cannot obtain a computable upper bound on the error. In particular, property **i)** from the Introduction cannot be achieved. For unsteady problems, the equivalent of (1.7) is

$$|||u - u_{h\tau}||| \leq C(h^k + \tau^l), \tag{1.8}$$

where $C > 0$, $k > 0$, and $l > 0$ and $|||\cdot|||$ is some space-time norm. This justifies the numerical method in question when spatial and temporal approximations are simultaneously improved. Remark finally that either Ch^k or $C(h^k + \tau^l)$ can in fact be evaluated prior to the calculation, without the knowledge of u_h or $u_{h\tau}$, whence the name of this estimate.

1.4 A posteriori error estimates

A posteriori error estimates aim at giving bounds on the error between the known numerical approximation and the unknown exact solution that can be computed in practice, once the approximate solution is known. For a steady problem, they typically take the form

$$|||u - u_h||| \leq \left\{ \sum_{K \in \mathcal{T}_h} \eta_K^q \right\}^{\frac{1}{q}}, \tag{1.9}$$

where $\eta_K = \eta_K(u_h)$ is a quantity linked to the mesh element $K \in \mathcal{T}_h$, computable from u_h. For unsteady problems, the typical form is

$$|||u - u_{h\tau}||| \leq \left\{ \sum_{n=1}^{N} \sum_{K \in \mathcal{T}_h^n} (\eta_K^n)^q \right\}^{\frac{1}{q}}. \tag{1.10}$$

Here $\eta_K^n = \eta_K^n(u_{h\tau})$ is a quantity linked to the discrete time t^n and the mesh element $K \in \mathcal{T}_h^n$, computable from $u_{h\tau}$. The most common choice for q in (1.9) and (1.10) is $q = 2$. The quantities η_K and η_K^n are called *element estimators*. Often, it is supposed that u_h and $u_{h\tau}$ are obtained as "exact" (up to machine precision) solutions of the corresponding linear systems. We will follow such an approach in Chapters 7–12, but one of the purposes of these lecture notes is to show that this assumption can be lifted; we will do so in Chapter 13.

One may formulate the following six properties describing an optimal a posteriori error estimate:

i) ensure that (1.9) or (1.10) holds and that η_K or η_K^n are fully computable from u_h or $u_{h\tau}$ (*guaranteed upper bound*);

ii) for steady problems, ensure that for all $K \in \mathcal{T}_h$, η_K represents a lower bound for the actual error in the vicinity of K, up to a generic constant: this means that there exists a constant $C > 0$ such that

$$\eta_K \leq C|||u - u_h|||_{\mathfrak{T}_K} \qquad \forall K \in \mathcal{T}_h, \tag{1.11}$$

where $\mathfrak{T}_K$ stands for the element K and its neighbors (*local efficiency*); for unsteady problems, ensure that there exists a constant $C > 0$ such that

$$\left\{ \sum_{K \in \mathcal{T}_h^n} (\eta_K^n)^q \right\}^{\frac{1}{q}} \leq C |||u - u_{h\tau}|||_{(t^{n-1}, t^n]} \qquad \forall 1 \leq n \leq N \tag{1.12}$$

(*local-in-time* and *global-in-space efficiency*);

iii) ensure that the effectivity index, given respectively as

$$I_{\text{eff}} := \frac{\left\{ \sum_{K \in \mathcal{T}_h} \eta_K^q \right\}^{\frac{1}{q}}}{|||u - u_h|||} \quad \text{or} \quad I_{\text{eff}} := \frac{\left\{ \sum_{n=1}^N \sum_{K \in \mathcal{T}_h^n} (\eta_K^n)^q \right\}^{\frac{1}{q}}}{|||u - u_{h\tau}|||} \tag{1.13}$$

i.e., as the ratio of the estimated and actual error, goes to one as the computational effort grows (*asymptotic exactness*);

iv) guarantee the three previous properties independently of the parameters of the problem and of their variation (*robustness*);

v) give estimators η_K and η_K^n which can be evaluated locally (only performing calculations in the element K or in its neighborhood $\mathfrak{T}_K$) (*small evaluation cost*);

vi) distinguish and estimate separately the different error components (*error components identification*).

Property **i)** above allows to give a truly computable upper bound on the unknown error $|||u - u_h|||$ or $|||u - u_{h\tau}|||$ and thus the error control in the sense of property **i)** of the Introduction. Property **ii)** enables to predict the error localization in the sense of property **ii)** of the Introduction. In particular, for steady problems, it allows to detect the areas of the computational domain Ω where the error is large, so that one can concentrate more effort therein. Typically, the mesh is refined in such areas, leading to the so-called concept of *adaptive mesh refinement*. The result (1.12) for unsteady problems is somewhat less satisfactory as it justifies theoretically the localization of the error in time but not in space, but seems to be the best currently available. Property **iii)** ensures the optimality of the upper bound; if the error is quite small and the estimator predicts a large value, it may still satisfy properties **i)** and **ii)** but is probably not too useful as it overestimates highly the error. Property **iv)** is one of the most important in practice. In real-life problems, parameters and coefficients such as the domain size and shape, final simulation time, diffusivity, reactivity, advection, or the size of the nonlinearity (respectively Ω, T, the tensor $\underline{\mathbf{K}}$, the function r, the field $\mathbf{w}$, and the function a from Section 1.1) may be large or small or vary over several degrees of magnitude; an estimator satisfying property **iv)** ensures that its results will be equally good in all situations. Next, property **v)** guarantees that the computational cost needed for the evaluation of the estimators η_K or η_K^n will be much smaller than the cost required to obtain the approximate solution u_h or $u_{h\tau}$ itself (recall that usually some kind of a global problem needs to be solved in order to obtain the approximate solution u_h and one such a problem needs to be solved at each time step for implicit time discretizations of unsteady problems to obtain $u_{h\tau}$).

Finally, the numerical error $|||u - u_h|||$ or $|||u - u_{h\tau}|||$ typically consists of several *error components*. The first one is the *discretization error*. For steady problems, the discretization error coincides with the *spatial discretization error*; for unsteady problems, the discretization error is split into the *spatial discretization error* and the *temporal discretization error*. These result respectively from the approximation properties of the numerical scheme and time stepping procedure on the current spatial and temporal meshes. Another typical error component

is the *algebraic error*, linked to the imprecision in the solution of the associated systems of linear algebraic equations by an algebraic solver. For nonlinear problems, the *linearization error*, linked to incomplete convergence of iterative linearizations such as the fixed-point or the Newton method, arises equally. Property **vi)** is essential for the identification of these different error components and for entire *adaptivity*, relying on all adaptive mesh refinement, adaptive time step choice, and adaptive *stopping criteria* for algebraic and linearization solvers. It is at the heart of satisfaction of the properties **i)** and **ii)** of the Introduction.

We will show in the rest of these lecture notes how to derive a posteriori error estimates satisfying as much as possible and as well as possible the six optimal properties **i)**–**vi)** for the model problems of Section 1.1.

Chapter 2

The Laplace equation in one space dimension

Let us, for the sake of clarity, start with the Laplace equation of Section 1.1.1 in one space dimension, i.e., with Ω being an interval. Many concepts will be clear form this simple model case.

Let $f \in L^2(\Omega)$. Rewriting (1.1a)–(1.1b) for $d = 1$ gives

$$-u'' = f \qquad \text{in } \Omega, \tag{2.1a}$$

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

Let us first recall that (2.1a)–(2.1b) does not have a classical solution (i.e., $u \in C^2(\overline{\Omega})$) in general; existence and uniqueness of a solution to (2.1a)–(2.1b) can be ensured using the so-called variational formulation. In order to state it, we first need to recall a fundamental function space, the space $H_0^1(\Omega)$. We follow [61].

2.1 The space $H_0^1(\Omega)$

Let us recall that the space $\mathcal{D}(\Omega)$ is the space of functions from $C^\infty(\overline{\Omega})$ with a compact support in Ω. We first need to introduce the following concept:

Definition 2.1.1 (Weak derivative)**.** *Let a function $v : \Omega \to \mathbb{R}$ be given. We say that v admits a weak derivative if*

1. *$v \in L^2(\Omega)$;*
2. *there exists a function $w : \Omega \to \mathbb{R}$ such that*
   (a) *$w \in L^2(\Omega)$;*
   (b) *$(v, \varphi') = -(w, \varphi) \qquad \forall \varphi \in \mathcal{D}(\Omega)$.*

The function w is called the weak derivative *of v. We use the notation $v' = w$.*

Definition 2.1.2 (The space $H^1(\Omega)$)**.** *The space $H^1(\Omega)$ is the space of all the functions which admit a weak derivative.*

Let us recall from [61] that $H^1(\Omega)$ is a Hilbert space for the scalar product $(u, v)_{H^1(\Omega)} := (u, v) + (u', v')$ and that $H^1(\Omega) \subset C^0(\overline{\Omega})$. This last property enables the following definition:

Definition 2.1.3 (The space $H_0^1(\Omega)$)**.** *The space $H_0^1(\Omega)$ is the space of functions $v \in H^1(\Omega)$ such that $v|_{\partial\Omega} = 0$.*

2.2 Variational formulation

We are now ready to state the variational formulation of (2.1a)–(2.1b):

Definition 2.2.1 (Variational formulation of (2.1a)–(2.1b)). *Find $u \in H_0^1(\Omega)$ such that*

$$(u', v') = (f, v) \qquad \forall v \in H_0^1(\Omega). \tag{2.2}$$

Recall from [61] that there exists a unique solution of (2.2) by the Riesz representation theorem (or by the Lax–Milgram theorem).

Definition 2.2.2 (Flux). *Let u be the solution of (2.2). Set*

$$\sigma := -u'. \tag{2.3}$$

We will call σ the flux.

Theorem 2.2.3 (Properties of the weak solution u of (2.2)). *Let u be the solution of (2.2) and σ the flux given by (2.3). Then*

$$u \in H_0^1(\Omega), \quad \sigma \in H^1(\Omega).$$

Additionally,

$$u \in C^0(\overline{\Omega}), \quad \sigma \in C^0(\overline{\Omega})$$

and

$$\sigma' = f.$$

Proof. We have $u \in H_0^1(\Omega)$ by the definition (2.2) and $H_0^1(\Omega) \subset C^0(\overline{\Omega})$, so that the result for u is immediate. Let us next show that $\sigma \in H^1(\Omega)$. For this, we need to verify the three conditions of Definition 2.1.1 for σ. Recalling that σ is the weak (and not the classical!) derivative of u (with a minus sign), we know that $\sigma \in L^2(\Omega)$, i.e., property 1 is satisfied. The function f is the natural candidate for the weak derivative of σ. As $f \in L^2(\Omega)$ by our assumption, property 2a is satisfied. Finally, we deduce from (2.2), using that $\mathcal{D}(\Omega) \subset H_0^1(\Omega)$, that

$$(\sigma, \varphi') = -(f, \varphi) \qquad \forall \varphi \in \mathcal{D}(\Omega) \tag{2.4}$$

which is nothing but 2b, so that σ admits the weak derivative f and thus $\sigma \in H^1(\Omega)$. The fact that $\sigma \in C^0(\overline{\Omega})$ then follows by the inclusion $H^1(\Omega) \subset C^0(\overline{\Omega})$. □

Remark 2.2.4 (Theorem 2.2.3). *Recall that (2.1a)–(2.1b) is a model for heat flow or for underground water flow, see Examples 1.1.1 and 1.1.2. In these applications, it is physical to have the continuity of u (temperature, pressure): these quantities naturally vary* without jumps *(it is hard to imagine that the temperature between the heater and the surrounding air in Figure 1.1 varies in a discontinuous way). Similarly, the heat or water* flux *σ is a quantity which is naturally and physically* continuous. *Let the domain Ω be divided into two parts Ω_1 and Ω_2. Then the heat flow which* flows out *from Ω_1 to Ω_2 has to be equal to the heat flow which* flows in *to Ω_2 from Ω_1. Theorem 2.2.3 says that in the one-dimensional model given by the variational formulation (2.2), these properties are perfectly maintained. Remark that this is by no means evident at a first sight, especially for the* flux *σ. Remark finally that, fortunately enough, the same situation is repeated in the proper functional setting in multiple space dimensions, see Theorem 7.1.3 below.*

2.3 The finite element method

Let us now introduce the finite element method for approximating the solution of (2.1a)–(2.1b).

Let $\mathcal{T}_h$ be a mesh of Ω, i.e., a division of the interval Ω into subintervals noted as K. Let $\mathbb{P}_k(K)$ stand for the set of polynomials of total degree less than or equal to k on the element $K \in \mathcal{T}_h$. Let, finally,

$$V_h := \{v_h \in C^0(\overline{\Omega});\, v_h|_K \in \mathbb{P}_k(K) \quad \forall K \in \mathcal{T}_h;\, v_h|_{\partial\Omega} = 0\}. \tag{2.5}$$

The formulation of the finite element method is deduced from (2.2). It reads:

Definition 2.3.1 (Finite element method for (2.1a)–(2.1b))**.** *Find $u_h \in V_h$ such that*

$$(u_h', v_h') = (f, v_h) \qquad \forall v_h \in V_h. \tag{2.6}$$

Recall that the existence and uniqueness of u_h follows by the same arguments as that of (2.2). In analogy with Definition 2.2.2, we introduce:

Definition 2.3.2 (Approximate flux)**.** *Let u_h be the solution of (2.6). We will call*

$$-u_h' \tag{2.7}$$

the approximate flux.

The following theorem should be compared to Theorem 2.2.3:

Remark 2.3.3 (Properties of the finite element solution u_h of (2.6))**.** *Let u_h be the solution of (2.6). Then*

$$u_h \in H_0^1(\Omega), \quad -u_h' \notin H^1(\Omega) \text{ in general.}$$

Consequently,

$$u_h \in C^0(\overline{\Omega}), \quad -u_h' \notin C^0(\overline{\Omega}), \text{ and } (-u_h')' \neq f \text{ in general.}$$

Indeed, the fact that $u_h \in H_0^1(\Omega) \cap C^0(\overline{\Omega})$ follows from the fact that $V_h \subset H_0^1(\Omega) \cap C^0(\overline{\Omega})$, whereas the facts that $-u_h' \notin H^1(\Omega)$, $-u_h' \notin C^0(\overline{\Omega})$, and $(-u_h')' \neq f$ in general (recall that u_h' is not the classical but the weak derivative in the sense of Definition 2.1.1) are evident from Example 2.3.4 and Figure 2.1 below.

Example 2.3.4 (Exact and approximate solutions properties)**.** *We illustrate here Remark 2.3.3 on a simple example. Consider (2.1a)–(2.1b) with $\Omega = (0,1)$ and $f = \pi^2 \sin(\pi x)$. Then it is easy to see that the solution of (2.2) is $u = \sin(\pi x)$. This solution, as well as the solution u_h of (2.6) for $k = 1$, are plotted in the left part of Figure 2.1, as an illustration of the fact that $u, u_h \in H_0^1(\Omega)$ and $u, u_h \in C^0(\overline{\Omega})$. The continuity of the exact solution is maintained by the finite element method. The weak derivatives of both u, u_h (i.e., the fluxes σ, $-u_h'$ multiplied by minus one) are then plotted in the right part of Figure 2.1. We have $u' \in H^1(\Omega)$ and $u' \in C^0(\overline{\Omega})$ by Theorem 2.2.3. This property is, however, not repeated on the discrete level for the finite element method anymore, as Remark 2.3.3 states.*

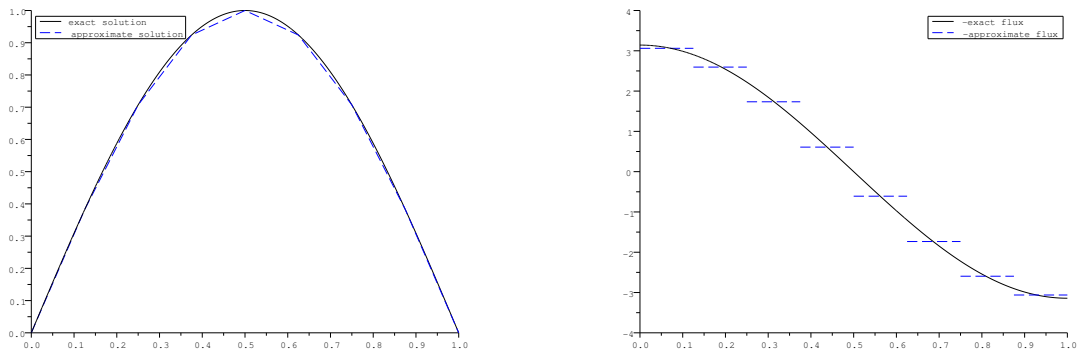


Figure 2.1: Exact and approximate solutions (left) and exact and approximate fluxes (right)

2.4 Energy norm and dual norms

A primordial question for measuring the distance between u and u_h is the choice of the norm ($|||\cdot|||$ in (1.7) and (1.9)). A prominent role between all different possibilities is played by the *energy norm*: this is the norm induced by the scalar product in (2.2):

$$|||u - u_h||| := \|(u - u_h)'\|_{L^2(\Omega)}. \tag{2.8}$$

Below, we will use the simplified notation $\|\cdot\| := \|\cdot\|_{L^2(\Omega)}$, dropping the index $L^2(\Omega)$. This norm admits the following useful characterization:

Theorem 2.4.1 (Energy norm for (2.1a)–(2.1b) as a dual norm)**.** *Let $v \in H^1_0(\Omega)$. Then*

$$\|v'\| = \sup_{\varphi \in H^1_0(\Omega);\, \|\varphi'\|=1} (v', \varphi'). \tag{2.9}$$

Proof. First remark that the proof in the case $v' = 0$ is trivial. Suppose that $v' \neq 0$. We will proceed in two steps.

Step 1. Proof of (2.9) with the sign $\leq$.

By the properties of the $L^2(\Omega)$ scalar product, there holds

$$\|v'\|^2 = (v', v').$$

Thus

$$\|v'\| = \left(v', \frac{v'}{\|v'\|}\right).$$

Set $w := \frac{v}{\|v'\|}$ and remark that $w \in H^1_0(\Omega)$ and that $\|w'\| = 1$. Thus, passing to a supremum, we get

$$\|v'\| = (v', w') \leq \sup_{\varphi \in H^1_0(\Omega);\, \|\varphi'\|=1} (v', \varphi').$$

Step 2. Proof of (2.9) with the sign $\geq$.

Using the Cauchy–Schwarz inequality, we can bound from above the supremum in (2.9),

$$\sup_{\varphi \in H^1_0(\Omega);\, \|\varphi'\|=1} (v', \varphi') \leq \sup_{\varphi \in H^1_0(\Omega);\, \|\varphi'\|=1} \{\|v'\|\|\varphi'\|\} = \|v'\|.$$

□

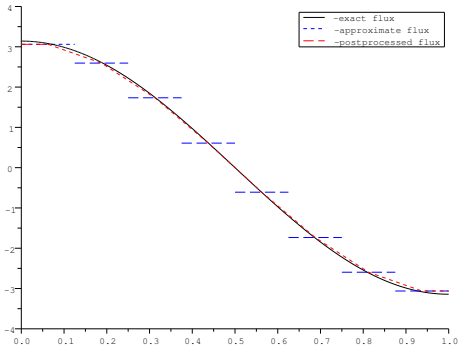


Figure 2.2: Exact, approximate, and reconstructed fluxes

Remark 2.4.2 (Dual norm)**.** *The term appearing on the right-hand side of* (2.9) *is a typical example of a* dual norm *of the function* v*.*

2.5 Flux reconstruction

From Theorem 2.2.3 and Remark 2.3.3 and Example 2.3.4, we see that the approximate flux $-u_h'$ is nonphysical. We will thus introduce its "correction", a flux reconstruction (or a reconstructed flux) σ_h:

Definition 2.5.1 (Flux reconstruction)**.** *Let* u_h *be the solution of* (2.6)*. We will call the* flux reconstruction *any function* σ_h *constructed from* u_h *which satisfies*

$$\sigma_h \in H^1(\Omega). \tag{2.10}$$

We refer to Figure 2.2 for an example of a flux reconstruction σ_h in the context of Example 2.3.4.

2.6 A first a posteriori error estimate

With the notion of the flux reconstruction of Definition 2.5.1 and of the characterization of the energy norm of Theorem 2.4.1, we will now give our first a posteriori estimate on the error between u, the unknown solution of (2.2), and u_h, the known solution of (2.6). The last ingredient that we need is the Friedrichs inequality:

$$\|\varphi\| \leq \frac{h_\Omega}{\pi}\|\varphi'\| \qquad \forall \varphi \in H_0^1(\Omega). \tag{2.11}$$

Theorem 2.6.1 (A first a posteriori error estimate)**.** *Let* u *be the weak solution given by Definition 2.2.1. Let* u_h *be its finite element approximation given by Definition 2.3.1. Let finally* σ_h *be a flux reconstruction following Definition 2.5.1. For any* $K \in \mathcal{T}_h$*, define the* residual estimator *by*

$$\eta_{\mathrm{R},K} := \frac{h_\Omega}{\pi}\|f - \sigma_h'\|_K \tag{2.12}$$

and the flux estimator *by*

$$\eta_{\mathrm{F},K} := \|u_h' + \sigma_h\|_K. \tag{2.13}$$

Then

$$\|(u-u_h)'\| \leq \left\{ \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{R},K}^2 \right\}^{\frac{1}{2}} + \left\{ \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{F},K}^2 \right\}^{\frac{1}{2}}.$$

Proof. Recall Theorem 2.4.1, where we set $v := u - u_h$. Let $\varphi \in H_0^1(\Omega)$ with $\|\varphi'\| = 1$ be fixed. Using the characterization (2.2) of the weak solution, we have

$$((u-u_h)', \varphi') = (f, \varphi) - (u_h', \varphi').$$

Adding and subtracting (σ_h, φ') and using the Green theorem $(\sigma_h, \varphi') = -(\sigma_h', \varphi)$, we have

$$((u-u_h)', \varphi') = (f, \varphi) - (u_h', \varphi') + (\sigma_h, \varphi') - (\sigma_h, \varphi') = (f - \sigma_h', \varphi) - (u_h' + \sigma_h, \varphi').$$

We now bound the two above-resulting terms separately. The Cauchy–Schwarz inequality immediately gives for the second one

$$-(u_h' + \sigma_h, \varphi') \leq \|u_h' + \sigma_h\| \|\varphi'\| = \left\{ \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{F},K}^2 \right\}^{\frac{1}{2}},$$

where we have used the fact that $\|\varphi'\| = 1$. The first term is then bounded using the Cauchy–Schwarz inequality, the Friedrichs inequality (2.11), and the fact that $\|\varphi'\| = 1$ as

$$(f - \sigma_h', \varphi) \leq \|f - \sigma_h'\| \|\varphi\| \leq \|f - \sigma_h'\| \frac{h_\Omega}{\pi} \|\varphi'\| = \left\{ \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{R},K}^2 \right\}^{\frac{1}{2}}.$$

Combining the above developments gives the desired result. □

Remark 2.6.2 (Theorem 2.6.1)**.** *Theorem 2.6.1 gives our first a posteriori error estimate which clearly satisfies property* **i)** *of Section 1.4. Moreover, it may be noted that it is quite general, as it in fact holds true for an arbitrary $u_h \in H_0^1(\Omega)$ (remark that the fact that u_h solves (2.6) was used nowhere in the proof of Theorem 2.6.1). Unfortunately, this result is not fully optimal in the sense that the other properties of Section 1.4 are difficult to satisfy. This is related to the fact that $\eta_{\mathrm{R},K}$ given by (2.12) features h_Ω, the diameter of the whole domain Ω. We will see below in Section 7 how the estimator $\eta_{\mathrm{R},K}$ can be improved to contain h_K instead of h_Ω, which is a much smaller quantity. We will then be able to satisfy most of the optimal properties of Section 1.4. We finally remark that the estimates of the style of Theorem 2.6.1 are those which are developed in the books by Neittaanmäki and Repin [75] and Repin [85].*

Chapter 3

Simplicial meshes

We summarize here all the notation concerning the meshes $\mathcal{T}_h$ used in these lecture notes. Recall that $\mathcal{T}_h$ is a simplicial partition of the polytope Ω, i.e., $\cup_{K\in\mathcal{T}_h}\overline{K} = \overline{\Omega}$, any $K \in \mathcal{T}_h$ is a closed simplex, and the intersection of two different simplices is either empty, a vertex, or an l-dimensional face, $1 \le l \le d-1$.

3.1 Mesh elements

In the context of mesh refinement, we shall consider sequences or families of meshes $\{\mathcal{T}_h\}_h$. We then suppose that these are shape regular in the sense that there exists a constant $\kappa_{\mathcal{T}} > 0$ such that, for all triangulations $\mathcal{T}_h$, $\max_{K\in\mathcal{T}_h} h_K/\varrho_K \le \kappa_{\mathcal{T}}$, where h_K is the diameter of K and ϱ_K is the diameter of the largest ball inscribed in K. In one space dimension, $h_K = \varrho_K$, so that any mesh is shape-regular. In two space dimensions, shape regularity simply means that the smallest angle in any mesh $\mathcal{T}_h$ is uniformly bounded away from zero. Thus, we do not consider so-called anisotropic meshes. On the other hand, shape-regular meshes can be highly graded; for instance, the meshes of Figures 8.4 and 8.5 below are shape-regular. Let $K \in \mathcal{T}_h$. The outward unit normal vector to K is denoted by $\mathbf{n}_K$.

3.2 Mesh faces; jumps, and averages

We denote by $\mathcal{E}_h$ the $(d-1)$-dimensional faces of the mesh $\mathcal{T}_h$, i.e., vertices in one space dimension, edges in two space dimensions, and faces in three space dimensions. We shall shortly speak about faces for any space dimension where there shall arise no confusion. The set $\mathcal{E}_h$ is decomposed into $\mathcal{E}_h^{\rm int}$, the faces lying in the interior of Ω, and $\mathcal{E}_h^{\rm ext}$, the faces lying on the boundary of Ω. To each face $e \in \mathcal{E}_h$, we associate a unit normal vector $\mathbf{n}_e$; the orientation of $\mathbf{n}_e$ is arbitrary for $e \in \mathcal{E}_h^{\rm int}$ and coincides with the outward unit normal vector $\mathbf{n}_\Omega$ of Ω for $e \in \mathcal{E}_h^{\rm ext}$. Let a face $e \in \mathcal{E}_h^{\rm int}$, $e = K \cap K'$ such that $\mathbf{n}_e$ points from K towards K', and a sufficiently regular function v be given. We define the *jump* and *average* of v on e respectively as

$$[\![v]\!]_e := (v|_K)|_e - (v|_{K'})|_e, \tag{3.1}$$

$$\{\!\{v\}\!\}_e := \tfrac{1}{2}((v|_K)|_e + (v|_{K'})|_e). \tag{3.2}$$

We set $[\![v]\!]_e := v|_e$ and $\{\!\{v\}\!\}_e := v|_e$ for $e \in \mathcal{E}_h^{\rm ext}$. This choice for $[\![v]\!]_e$ and $\{\!\{v\}\!\}_e$ on the boundary of Ω is made so as to naturally appear in broken Green formulas, see (4.15) below. Later, we will simply use the notation $[\![v]\!]$ and $\{\!\{v\}\!\}$. For $K \in \mathcal{T}_h$, $\mathcal{E}_K$ denotes the set of faces of K.

3.3 Mesh vertices

We denote the set of vertices by $\mathcal{V}_h$ and decompose it into interior vertices $\mathcal{V}_h^{\text{int}}$ and vertices lying on the boundary $\mathcal{V}_h^{\text{ext}}$. For a vertex $\mathbf{a} \in \mathcal{V}_h$, we will use the notation $\mathcal{T}_\mathbf{a}$ for the patch of the elements of $\mathcal{T}_h$ which share $\mathbf{a}$, and $\omega_\mathbf{a}$ the corresponding polytopic subdomain. Then $\psi_\mathbf{a}$ is the continuous, piecewise affine "hat" function which takes value 1 at the vertex $\mathbf{a}$ and zero at the other vertices. For $K \in \mathcal{T}_h$, $\mathcal{V}_K$ denotes the set of vertices of K.

3.4 Various sets of elements and faces

In addition to the above basic notation, we will in particular in Sections 8.2 and 8.3 need some more sets of elements and faces. For all $K \in \mathcal{T}_h$, we let $\mathfrak{T}_K$ denote all the elements in $\mathcal{T}_h$ sharing at least a vertex with K. Similarly, $\mathfrak{E}_K$ stand for all the faces in $\mathcal{E}_h$ sharing at least a vertex with K, and $\mathfrak{E}_K^{\text{int}}$ its subset collecting those faces lying in the interior of Ω. Those faces of the element $K \in \mathcal{T}_h$ which lie in the interior of Ω are then collected in the set $\mathcal{E}_K^{\text{int}}$. For a face $e \in \mathcal{E}_h$, we will also use the notation $\mathcal{T}_e$ for the (one or two) simplices that share it, and $\mathcal{V}_e$ for all vertices $e \in \mathcal{E}_h$.

Chapter 4

Spaces $H_0^1(\Omega)$, $\mathbf{H}(\mathrm{div},\Omega)$, their broken versions, and useful inequalities

We introduce here some additional functional spaces necessary in order to carry out the analysis in the subsequent sections. We follow Thomas [91] and Allaire [9]. More details can be found in Raviart and Thomas [84] or Adams [2].

4.1 The space $H_0^1(\Omega)$

Let us recall that the space $\mathcal{D}(\Omega)$ is the space of functions from $C^\infty(\overline{\Omega})$ with a compact support in Ω. We first need to introduce the following concept:

Definition 4.1.1 (Weak partial derivative)**.** *Let a scalar function $v : \Omega \to \mathbb{R}$ be given. We say that v admits a weak i-th partial derivative, $1 \le i \le d$, if*

1. *$v \in L^2(\Omega)$;*
2. *there exists a function $w_i : \Omega \to \mathbb{R}$ such that*
   (a) *$w_i \in L^2(\Omega)$;*
   (b) *$(v, \partial_{\mathbf{x}_i}\varphi) = -(w_i, \varphi) \qquad \forall \varphi \in \mathcal{D}(\Omega)$.*

The function w_i is called the weak i-th partial derivative *of v. We use the notation $\partial_{\mathbf{x}_i} v = w_i$.*

Definition 4.1.2 (Weak gradient)**.** *Let a scalar function $v : \Omega \to \mathbb{R}$ be given. We say that v admits a* weak gradient *if v admits the weak i-th partial derivative for all $1 \le i \le d$. We set*

$$\nabla v := (\partial_{\mathbf{x}_1} v, \dots, \partial_{\mathbf{x}_d} v)^t. \tag{4.1}$$

Definition 4.1.3 (The space $H^1(\Omega)$)**.** *The space $H^1(\Omega)$ is the space of all the functions which admit the weak gradient.*

Let us recall from [91, 9, 2] that $H^1(\Omega)$ is a Hilbert space for the scalar product $(u, v)_{H^1(\Omega)} := (u, v) + (\nabla u, \nabla v)$.

Definition 4.1.4 (The space $H_0^1(\Omega)$)**.** *The space $H_0^1(\Omega)$ is the space of functions $v \in H^1(\Omega)$ such that $v|_{\partial\Omega} = 0$.*

Remark 4.1.5 (Trace)**.** *In the above definition, we have used the restriction of a function $v \in H^1(\Omega)$ onto the boundary of Ω, $v|_{\partial\Omega}$. This is by no means evident, as the functions from $H^1(\Omega)$ are a priori only from $L^2(\Omega)$ and thus $v|_{\partial\Omega}$ may not even be defined as $\partial\Omega$ is a set*

of measure zero. Fortunately, $v|_{\partial\Omega}$ can be given a sense introducing the concept of a trace, *which is possible for functions from $H^1(\Omega)$. Namely, $v|_{\partial\Omega}$ in the sense of traces and $v|_{\partial\Omega}$ for $v \in H^1(\Omega) \cap C^0(\overline{\Omega})$ coincide. Recall explicitly that contrarily to Section 2.1, $H^1(\Omega) \not\subset C^0(\overline{\Omega})$ in multiple space dimensions. We refer for all details to [91, 9, 2].*

The two following theorems generalize the Green theorem onto the spaces $H^1(\Omega)$ and $H^1_0(\Omega)$ ($\mathbf{n}_\Omega$ is the unit normal vector of Ω, exterior to Ω):

Theorem 4.1.6 (Green theorem on $H^1(\Omega) \times H^1(\Omega)$)**.** *Let $u, v \in H^1(\Omega)$ and $1 \leq i \leq d$. Then*

$$(\partial_{\mathbf{x}_i} u, v) + (u, \partial_{\mathbf{x}_i} v) = \langle u \mathbf{n}^i_\Omega, v \rangle. \tag{4.2}$$

Theorem 4.1.7 (Green theorem on $H^1_0(\Omega) \times H^1(\Omega)$)**.** *Let $u \in H^1_0(\Omega)$, $v \in H^1(\Omega)$, and $1 \leq i \leq d$. Then*

$$(\partial_{\mathbf{x}_i} u, v) + (u, \partial_{\mathbf{x}_i} v) = 0. \tag{4.3}$$

4.2 The space $\mathbf{H}(\mathrm{div},\Omega)$

Definition 4.2.1 (Weak divergence)**.** *Let a vector function $\mathbf{v} : \Omega \to \mathbb{R}^d$ be given. We say that $\mathbf{v}$ admits a weak divergence if*

1. $\mathbf{v} \in [L^2(\Omega)]^d$;
2. *there exists a function $w : \Omega \to \mathbb{R}$ such that*
   (a) $w \in L^2(\Omega)$;
   (b) $(\mathbf{v}, \nabla\varphi) = -(w, \varphi) \qquad \forall \varphi \in \mathcal{D}(\Omega)$.

The function w is called the weak divergence *of $\mathbf{v}$. We use the notation $\nabla\cdot\mathbf{v} = w$.*

Definition 4.2.2 (The space $\mathbf{H}(\mathrm{div},\Omega)$)**.** *The space $\mathbf{H}(\mathrm{div},\Omega)$ is the space of all the functions which admit the weak divergence.*

Let us recall from [91, 9, 2] that $\mathbf{H}(\mathrm{div},\Omega)$ is a Hilbert space for the scalar product $(\mathbf{u},\mathbf{v})_{\mathbf{H}(\mathrm{div},\Omega)} := (\mathbf{u},\mathbf{v}) + (\nabla\cdot\mathbf{u}, \nabla\cdot\mathbf{v})$.

Remark 4.2.3 (Normal trace)**.** *Similarly to Remark 4.1.5, we do not a priori have a right to speak about $\mathbf{v}\cdot\mathbf{n}_\Omega|_{\partial\Omega}$ for $\mathbf{v} \in \mathbf{H}(\mathrm{div},\Omega)$, as the functions from $\mathbf{H}(\mathrm{div},\Omega)$ may not even be defined on $\partial\Omega$. It turns out that this can be overcome introducing the concept of a* normal trace *on $\mathbf{H}(\mathrm{div},\Omega)$, again generalizing the property for continuous vector fields. We refer for all details to [91, 9, 2].*

The following results will be used many times in these lecture notes:

Theorem 4.2.4 (Green theorem on $H^1(\Omega) \times \mathbf{H}(\mathrm{div},\Omega)$)**.** *Let $v \in H^1(\Omega)$ and $\mathbf{w} \in \mathbf{H}(\mathrm{div},\Omega)$. Then*

$$(\mathbf{w}, \nabla v) + (\nabla\cdot\mathbf{w}, v) = \langle \mathbf{w}\cdot\mathbf{n}_\Omega, v \rangle. \tag{4.4}$$

Theorem 4.2.5 (Green theorem on $H^1_0(\Omega) \times \mathbf{H}(\mathrm{div},\Omega)$)**.** *Let $v \in H^1_0(\Omega)$ and $\mathbf{w} \in \mathbf{H}(\mathrm{div},\Omega)$. Then*

$$(\mathbf{w}, \nabla v) + (\nabla\cdot\mathbf{w}, v) = 0. \tag{4.5}$$

4.3 The spaces $H^1(\mathcal{T}_h)$ and $\mathbf{H}(\mathrm{div}, \mathcal{T}_h)$

Let $\mathcal{T}_h$ be a simplicial mesh of Ω as described in Section 3.1. In the sequel, we will often use the following space:

Definition 4.3.1 (The space $H^1(\mathcal{T}_h)$)**.** *The so-called* broken Sobolev space *is given by*

$$H^1(\mathcal{T}_h) := \{v \in L^2(\Omega);\ v|_K \in H^1(K) \quad \forall K \in \mathcal{T}_h\}. \tag{4.6}$$

The space $H^1(\mathcal{T}_h)$ is thus a collection of independent Sobolev spaces $H^1(K)$ over the individual elements K of the mesh $\mathcal{T}_h$. For a function $v \in H^1(\mathcal{T}_h)$, we introduce the notation $\nabla_h v$ so as to denote the broken weak gradient, $\nabla_h v \in [L^2(\Omega)]^d$,

$$(\nabla_h v)|_K := \nabla(v|_K). \tag{4.7}$$

The following result is simple but important:

Theorem 4.3.2 (Inclusion of $H^1(\Omega)$ in $H^1(\mathcal{T}_h)$)**.** *There holds $H^1(\Omega) \subset H^1(\mathcal{T}_h)$. Moreover, the broken weak gradient coincides for $v \in H^1(\Omega)$ with the weak one, i.e., $(\nabla_h v)|_K = (\nabla v)|_K$ for all $K \in \mathcal{T}_h$.*

Proof. Let $v \in H^1(\Omega)$. Then $v \in L^2(\Omega)$ by the definition of $H^1(\Omega)$, so that the first condition in (4.6) is satisfied. Let now $K \in \mathcal{T}_h$. We need to show that $v|_K \in H^1(K)$ (i.e., the three conditions of Definition 4.1.1 for the domain K and for all $1 \le i \le d$) and that the weak gradient of $v|_K$ coincides with the restriction to K of the weak gradient of v. Let $1 \le i \le d$ be fixed. Condition 1 is obvious, as clearly $v|_K \in L^2(K)$. As for the function w_i, we take $(\partial_{\mathbf{x}_i} v)|_K$, i.e., the restriction of the i-th weak partial derivative of our function $v \in H^1(\Omega)$ to the element K. Obviously, $w_i \in L^2(K)$. We are thus left with showing

$$(v, \partial_{\mathbf{x}_i} \varphi)_K = -(w_i, \varphi)_K \qquad \forall \varphi \in \mathcal{D}(K). \tag{4.8}$$

As $v \in H^1(\Omega)$, we, however, know that

$$(v, \partial_{\mathbf{x}_i} \varphi) = -(w_i, \varphi) \qquad \forall \varphi \in \mathcal{D}(\Omega). \tag{4.9}$$

It is thus enough to extend any function $\varphi \in \mathcal{D}(K)$ from (4.8) by zero outside of K and to use (4.9) for this extension to conclude. □

Remark 4.3.3 (Notation ∇)**.** *As we have just shown that $\nabla_h v = \nabla v$ for all $v \in H^1(\Omega)$, so that ∇_h is a natural extension of the weak gradient from $H^1(\Omega)$ to $H^1(\mathcal{T}_h)$, we will henceforth stick to the unique notation ∇, meaning the weak gradient on $H^1(\Omega)$ and the broken weak gradient on $H^1(\mathcal{T}_h)$.*

The broken space for vectors is:

Definition 4.3.4 (The space $\mathbf{H}(\mathrm{div}, \mathcal{T}_h)$)**.** *The broken divergence space is*

$$\mathbf{H}(\mathrm{div}, \mathcal{T}_h) := \{\mathbf{v} \in [L^2(\Omega)]^d;\ \mathbf{v}|_K \in \mathbf{H}(\mathrm{div}, K) \quad \forall K \in \mathcal{T}_h\}. \tag{4.10}$$

As above, for $\mathbf{v} \in \mathbf{H}(\mathrm{div}, \mathcal{T}_h)$, the notation $\nabla\cdot_h \mathbf{v}$ stands for the broken weak divergence, $\nabla\cdot_h \mathbf{v} \in L^2(\Omega)$,

$$(\nabla\cdot_h \mathbf{v})|_K := \nabla\cdot(\mathbf{v}|_K). \tag{4.11}$$

The following equivalent of Theorem 4.3.2 holds, showing that it is enough to stick to the unique notation $\nabla\cdot$, meaning the weak gradient on $\mathbf{H}(\mathrm{div}, \Omega)$ and the broken weak gradient on $\mathbf{H}(\mathrm{div}, \mathcal{T}_h)$:

Theorem 4.3.5 (Inclusion of $\mathbf{H}(\mathrm{div},\Omega)$ in $\mathbf{H}(\mathrm{div},\mathcal{T}_h)$)**.** *There holds* $\mathbf{H}(\mathrm{div},\Omega)\subset\mathbf{H}(\mathrm{div},\mathcal{T}_h)$. *Moreover, the broken weak divergence coincides for* $\mathbf{v}\in\mathbf{H}(\mathrm{div},\Omega)$ *with the weak one, i.e.,* $(\nabla_h\cdot\mathbf{v})|_K=(\nabla\cdot\mathbf{v})|_K$ *for all* $K\in\mathcal{T}_h$.

Proof. Let $\mathbf{v}\in\mathbf{H}(\mathrm{div},\Omega)$. Then $\mathbf{v}\in[L^2(\Omega)]^d$ by the definition of $\mathbf{H}(\mathrm{div},\Omega)$, so that the first condition in (4.10) is satisfied. Let now $K\in\mathcal{T}_h$. We need to show that $\mathbf{v}|_K\in\mathbf{H}(\mathrm{div},K)$ (i.e., the three conditions of Definition 4.2.1 for the domain K) and that the weak divergence of $\mathbf{v}|_K$ coincides with the restriction to K of the weak divergence of $\mathbf{v}$. Condition 1 is obvious, as clearly $\mathbf{v}|_K\in[L^2(K)]^d$. As for the function w, we take $(\nabla\cdot\mathbf{v})|_K$, i.e., the restriction of the weak divergence of our function $\mathbf{v}\in\mathbf{H}(\mathrm{div},\Omega)$ to the element K. Obviously, $w\in L^2(K)$. We are thus left with showing

$$(\mathbf{v},\nabla\varphi)_K=-(w,\varphi)_K\qquad\forall\varphi\in\mathcal{D}(K). \tag{4.12}$$

As $\mathbf{v}\in\mathbf{H}(\mathrm{div},\Omega)$, we, however, know that

$$(\mathbf{v},\nabla\varphi)=-(w,\varphi)\qquad\forall\varphi\in\mathcal{D}(\Omega). \tag{4.13}$$

It is thus enough to extend any function $\varphi\in\mathcal{D}(K)$ from (4.12) by zero outside of K and to use (4.13) for this extension to conclude. □

4.4 Continuity of traces

We now make a link between the spaces $H^1(\mathcal{T}_h)$ of Definition 4.3.1 and $H^1(\Omega)$ of Definition 4.1.3.

Theorem 4.4.1 (A sufficient condition for $H^1(\Omega)$)**.** *Let* $v\in H^1(\mathcal{T}_h)$ *be such that*

$$[\![v]\!]=0\qquad\forall e\in\mathcal{E}_h^{\mathrm{int}}. \tag{4.14}$$

Then $v\in H^1(\Omega)$ *and* $(\partial_{\mathbf{x}_i}v)|_K=\partial_{\mathbf{x}_i}(v|_K)$ *for all* $1\le i\le d$ *and all* $K\in\mathcal{T}_h$.

Proof. Let $v\in H^1(\mathcal{T}_h)$ satisfying (4.14) be given. We need to show that v admits the weak partial derivatives, i.e., the three conditions of Definition 4.1.1. Condition 1 is obvious, as $v\in H^1(\mathcal{T}_h)$. Let $1\le i\le d$ be fixed. Let us define a function w_i by $w_i|_K:=\partial_{\mathbf{x}_i}(v|_K)$, $K\in\mathcal{T}_h$. This is possible as $v\in H^1(\mathcal{T}_h)$ by our assumption and thus the weak partial derivatives $\partial_{\mathbf{x}_i}(v|_K)$ are well-defined for any $K\in\mathcal{T}_h$. Condition 2a then immediately follows as $w_i|_K$ are square-integrable for all $K\in\mathcal{T}_h$. We are thus left to show 2b. Let $\varphi\in\mathcal{D}(\Omega)$. Decomposing the integral over Ω into a sum of integrals over the mesh elements, using the Green theorem (4.2) in each $K\in\mathcal{T}_h$ (which is possible as $v|_K\in H^1(K)$ for all $K\in\mathcal{T}_h$), rearranging the summation, and finally using that $\varphi\in C^0(\overline{\Omega})$, we obtain

$$\begin{aligned}(v,\partial_{\mathbf{x}_i}\varphi)&=\sum_{K\in\mathcal{T}_h}(v,\partial_{\mathbf{x}_i}\varphi)_K=\sum_{K\in\mathcal{T}_h}\{-(\partial_{\mathbf{x}_i}v,\varphi)_K+\langle v\mathbf{n}_K^i,\varphi\rangle_{\partial K}\}\\&=-\sum_{K\in\mathcal{T}_h}(\partial_{\mathbf{x}_i}v,\varphi)_K+\sum_{e\in\mathcal{E}_h}\langle[\![v]\!]\mathbf{n}_e^i,\varphi\rangle_e.\end{aligned} \tag{4.15}$$

Using our assumption (4.14) and the fact that $\varphi=0$ on $\partial\Omega$, the second term above vanishes. The proof is finished noting that the first term above equals $-(w_i,\varphi)$. □

Similarly, we obtain the following theorem:

Theorem 4.4.2 (A sufficient condition for $H^1_0(\Omega)$)**.** *Let $v \in H^1(\mathcal{T}_h)$ such that*

$$[\![v]\!] = 0 \qquad \forall e \in \mathcal{E}_h.$$

Then $v \in H^1_0(\Omega)$ and $(\partial_{\mathbf{x}_i} v)|_K = \partial_{\mathbf{x}_i}(v|_K)$ for all $1 \le i \le d$ and all $K \in \mathcal{T}_h$.

The following crucial theorem holds in the opposite direction:

Theorem 4.4.3 (Continuity of traces in $H^1_0(\Omega)$)**.** *Let $v \in H^1_0(\Omega)$. Then*

$$[\![v]\!] = 0 \qquad \forall e \in \mathcal{E}_h.$$

Proof. Let $e \in \mathcal{E}_h^{\mathrm{ext}}$. Then $[\![v]\!] = v|_e = 0$ by the definition of $H^1_0(\Omega)$. We now show that $[\![v]\!] = 0$ also for all $e \in \mathcal{E}_h^{\mathrm{int}}$. As $v \in H^1_0(\Omega)$,

$$(v, \partial_{\mathbf{x}_i}\varphi) = -(\partial_{\mathbf{x}_i} v, \varphi) \qquad \forall \varphi \in \mathcal{D}(\Omega). \tag{4.16}$$

A function $v \in H^1_0(\Omega)$ also belongs to $H^1(\mathcal{T}_h)$. We thus from (4.15) infer

$$(v, \partial_{\mathbf{x}_i}\varphi) = -(\partial_{\mathbf{x}_i} v, \varphi) + \sum_{e \in \mathcal{E}_h} \langle [\![v]\!] \mathbf{n}_e^i, \varphi \rangle_e \qquad \forall \varphi \in \mathcal{D}(\Omega) \tag{4.17}$$

for all $1 \le i \le d$. Comparing (4.16) and (4.17) and taking into account that we have already shown that $[\![v]\!] = 0$ for all $e \in \mathcal{E}_h^{\mathrm{ext}}$, we see that

$$\sum_{e \in \mathcal{E}_h^{\mathrm{int}}} \langle [\![v]\!] \mathbf{n}_e^i, \varphi \rangle_e = 0 \qquad \forall \varphi \in \mathcal{D}(\Omega)$$

for all $1 \le i \le d$. Fix a face $e \in \mathcal{E}_h^{\mathrm{int}}$ and denote by $\mathcal{T}_e$ the two simplices that share e. Then the above relation implies

$$\langle [\![v]\!] \mathbf{n}_e^i, \varphi \rangle_e = 0 \qquad \forall \varphi \in \mathcal{D}(\mathcal{T}_e)$$

for all $1 \le i \le d$. There exists at least one $\mathbf{n}_e^i$, $1 \le i \le d$, which is nonzero (at least one component of $\mathbf{n}_e$ is always nonzero). Thus, we obtain that

$$\langle [\![v]\!], \varphi \rangle_e = 0 \qquad \forall \varphi \in \mathcal{D}(\mathcal{T}_e).$$

The assertion follows from the fact that $[\![v]\!]$ as an element of $L^2(e)$ is orthogonal to the traces of all $\varphi \in \mathcal{D}(\mathcal{T}_e)$ on the face e and a density argument. □

Remark 4.4.4 (Continuity of traces in $H^1_0(\Omega)$)**.** *Theorem 4.4.3 means that functions from $H^1_0(\Omega)$, not necessarily continuous (included in $C^0(\overline{\Omega})$), indeed posses a continuity in the sense of traces. Representing in $H^1_0(\Omega)$ the physical variables (temperature, pressure) thus to a certain degree maintains the natural properties of these variables, cf. Remark 2.2.4 for $d = 1$.*

4.5 Continuity of normal traces

Similarly to Section 4.4, we have the following results for the space $\mathbf{H}(\mathrm{div}, \Omega)$:

Theorem 4.5.1 (A sufficient condition for $\mathbf{H}(\mathrm{div}, \Omega)$)**.** *Let $\mathbf{v} \in \mathbf{H}(\mathrm{div}, \mathcal{T}_h)$ with $\mathbf{v}\cdot\mathbf{n}_e|_e \in L^2(e)$ for all $e \in \mathcal{E}_h^{\mathrm{int}}$ be such that*

$$[\![\mathbf{v}]\!]\cdot\mathbf{n}_e = 0 \qquad \forall e \in \mathcal{E}_h^{\mathrm{int}}. \tag{4.18}$$

Then $\mathbf{v} \in \mathbf{H}(\mathrm{div}, \Omega)$.

Proof. The proof follows the same idea as that of Theorem 4.4.1. We need to show the three conditions of Definition 4.2.1. Condition 1 is one of our assumptions. Let us define a function w by $w|_K := \nabla\cdot(\mathbf{v}|_K)$, $K \in \mathcal{T}_h$; then condition 2a follows by the fact that $\mathbf{v}|_K \in \mathbf{H}(\mathrm{div},K)$ and thus its weak divergence exists and is square-integrable for any $K \in \mathcal{T}_h$. We are thus left to show 2b. Let $\varphi \in \mathcal{D}(\Omega)$. Decomposing the integral over Ω into a sum of integrals over the mesh elements, using the Green theorem (4.4) in each $K \in \mathcal{T}_h$ (which is possible as $\mathbf{v}|_K \in \mathbf{H}(\mathrm{div},K)$ for all $K \in \mathcal{T}_h$), rearranging the summation, and finally using that $\varphi \in C^0(\overline{\Omega})$, we obtain

$$\begin{aligned}(\mathbf{v},\nabla\varphi) &= \sum_{K\in\mathcal{T}_h}(\mathbf{v},\nabla\varphi)_K = \sum_{K\in\mathcal{T}_h}\{-(\nabla\cdot\mathbf{v},\varphi)_K + \langle\mathbf{v}\cdot\mathbf{n}_K,\varphi\rangle_{\partial K}\}\\ &= -\sum_{K\in\mathcal{T}_h}(\nabla\cdot\mathbf{v},\varphi)_K + \sum_{e\in\mathcal{E}_h}\langle[\![\mathbf{v}]\!]\cdot\mathbf{n}_e,\varphi\rangle_e.\end{aligned} \tag{4.19}$$

Using our assumption (4.18) and the fact that $\varphi = 0$ on $\partial\Omega$, the second term above vanishes. The proof is finished noting that the first term above equals $-(w,\varphi)$. □

By a similar reasoning, we obtain the following crucial theorem:

Theorem 4.5.2 (Continuity of normal traces in $\mathbf{H}(\mathrm{div},\Omega)$)**.** *Let* $\mathbf{v} \in \mathbf{H}(\mathrm{div},\Omega)$ *satisfy* $\mathbf{v}\cdot\mathbf{n}_e|_e \in L^2(e)$ *for all* $e \in \mathcal{E}_h^{\mathrm{int}}$*. Then*

$$[\![\mathbf{v}]\!]\cdot\mathbf{n}_e = 0 \qquad \forall e \in \mathcal{E}_h^{\mathrm{int}}.$$

Remark 4.5.3 (Continuity of normal traces in $\mathbf{H}(\mathrm{div},\Omega)$)**.** *Theorem 4.5.2 means that functions from* $\mathbf{H}(\mathrm{div},\Omega)$*, whose normal components are not necessarily continuous (included in* $C^0(\overline{\Omega})$*), indeed posses a continuity in the sense of normal traces. Representing in* $\mathbf{H}(\mathrm{div},\Omega)$ *the physical variables (heat or water flux) thus to a certain degree maintains the natural properties of these variables, cf. Remark 2.2.4 for* $d = 1$*.*

4.6 Poincaré, Friedrichs, and trace inequalities

We recall here three basic inequalities that will be often used in the following chapters.

Let $\omega \subset \Omega$ be an open polytope and let h_ω denote its diameter.

Theorem 4.6.1 (Poincaré inequality)**.** *There holds*

$$\|v - v_\omega\|_\omega \le C_{\mathrm{P},\omega}h_\omega\|\nabla v\|_\omega \qquad \forall v \in H^1(\omega), \tag{4.20}$$

where v_ω *is the mean value of* v *over* ω *given by* $v_\omega := (v,1)_\omega/|\omega|$*.*

The constant $C_{\mathrm{P},\omega}$ can be precisely estimated in many cases. Whenever ω is convex, $C_{\mathrm{P},\omega}$ can be taken as $1/\pi$, cf. Payne and Weinberger [78] and Bebendorf [17]. If ω is a simplex (triangle) K, still a more precise estimate using the Bessel function of the first kind can be given, see Laugesen and Siudeja [71]. On the other hand, if ω is not convex, estimates on $C_{\mathrm{P},\omega}$ are little more involved and can be found in Eymard *et al.* [57, 58], Veeser and Verfürth [92], Repin [86], and Šebestová and Vejchodský [88] and the references therein.

Let $\partial\omega_{\mathrm{D}}$ be a simply connected subset of $\partial\omega$ with nonzero co-measure, i.e., $|\partial\omega_{\mathrm{D}}| \neq 0$.

Theorem 4.6.2 (Friedrichs inequality)**.** *There holds*

$$\|v\|_\omega \le C_{\mathrm{F},\omega,\partial\omega_{\mathrm{D}}}h_\omega\|\nabla v\|_\omega \qquad \forall v \in H^1(\omega) \text{ such that } v = 0 \text{ on } \partial\omega_{\mathrm{D}}. \tag{4.21}$$

As long as ω and $\partial\omega_{\mathrm{D}}$ are such that there exists a vector $\mathbf{b} \in \mathbb{R}^d$ such that for almost all $\mathbf{x} \in \omega$, the first intersection of the straight semi-line defined by the origin $\mathbf{x}$ and the vector $\mathbf{b}$ lies in $\partial\omega_{\mathrm{D}}$, the constant $C_{\mathrm{F},\omega,\partial\omega_{\mathrm{D}}}$ can be taken equal to 1, cf. [97, Remark 5.8]. To evaluate $C_{\mathrm{F},\omega,\partial\omega_{\mathrm{D}}}$ in the general case is more complicated but can be done following [28, Section 3], [97, Remark 5.9], [86], and the references therein.

Let finally K be a simplex and let e be one of its faces.

Theorem 4.6.3 (Trace inequality)**.** *There holds*

$$\|v\|_e^2 \le \widetilde{C}_{\mathrm{t},K,e}(h_K^{-1}\|v\|_K^2 + \|v\|_K\|\nabla v\|_K) \qquad \forall v \in H^1(K), \tag{4.22a}$$

$$\|v - v_e\|_e \le \bar{C}_{\mathrm{t},K,e} h_e^{\frac{1}{2}} \|\nabla v\|_K \qquad \forall v \in H^1(K), \tag{4.22b}$$

$$\|v - v_K\|_e \le C_{\mathrm{t},K,e} h_K^{\frac{1}{2}} \|\nabla v\|_K \qquad \forall v \in H^1(K). \tag{4.22c}$$

It follows from Stephansen [89, Lemma 3.12] that the constant $\widetilde{C}_{\mathrm{t},K,e}$ can be evaluated as $|e|h_K/|K|$, see also Carstensen and Funken [28, Theorem 4.1] for $d = 2$. It has been shown in Nicaise [76, Lemma 3.5] that $\bar{C}^2_{\mathrm{t},K,e} = C_{\mathrm{t},d}|e|h_K^2/(|K|h_e)$, where $C_{\mathrm{t},d} \approx 0.77708$ if $d = 2$ and $C_{\mathrm{t},d} \approx 3.84519$ if $d = 3$. Similarly, it follows from the proof of Eymard *et al.* [58, Lemma 9.4] and [97, Lemma 4.1] that $C^2_{\mathrm{t},K,e} = 3dh_K|e|/|K|$.

4.7 Broken Poincaré and Friedrichs inequalities

Let $\mathcal{T}_h$ stand in this subsection for a simplicial partition of ω; $\mathcal{E}_h^{\mathrm{int}}$ then denotes its interior faces. We consider here $H^1(\mathcal{T}_h) := \{v \in L^2(\omega);\, v|_K \in H^1(K) \text{ for all } K \in \mathcal{T}_h\}$. Broken Poincaré and Friedrichs inequalities are the versions of (4.20) and (4.21) valid on the broken space $H^1(\mathcal{T}_h)$.

Let v_e denote the mean value of a function v on the face $v \in \mathcal{E}_h^{\mathrm{int}}$. There in particular holds

$$\|v - v_\omega\|_\omega \le C_{\mathrm{bP},\omega} h_\omega \left\{ \|\nabla v\|_\omega^2 + \sum_{e\in\mathcal{E}_h^{\mathrm{int}}} h_e^{-1}\|([\![v]\!])_e\|_e^2 \right\}^{\frac{1}{2}} \qquad \forall v \in H^1(\mathcal{T}_h) \tag{4.23}$$

and

$$\|v\|_\omega \le C_{\mathrm{bF},\omega} h_\omega \left\{ \|\nabla v\|_\omega^2 + \sum_{e\in\mathcal{E}_h^{\mathrm{int}}} h_e^{-1}\|([\![v]\!])_e\|_e^2 + \langle v, 1\rangle_{\partial\omega}^2 \right\}^{\frac{1}{2}} \qquad \forall v \in H^1(\mathcal{T}_h), \tag{4.24}$$

where $C_{\mathrm{bP},\omega}$ and $C_{\mathrm{bF},\omega}$ are generic constants which can only depend on the shape regularity of the mesh $\mathcal{T}_h$ (on the smallest angle in $\mathcal{T}_h$ for $d = 2$). We refer to Eymard et al. [57], Dolejší et al. [44], Knobloch [69], Brenner [23], and to [97] for details and the values of $C_{\mathrm{bP},\omega}$, $C_{\mathrm{bF},\omega}$.

Chapter 5

Finite-dimensional subspaces of $L^2(\Omega)$, $H_0^1(\Omega)$, and $\mathbf{H}(\mathrm{div}, \Omega)$

We now introduce some finite-dimensional approximations of the spaces $L^2(\Omega)$, $H_0^1(\Omega)$, and $\mathbf{H}(\mathrm{div}, \Omega)$ ($H^1(\mathcal{T}_h)$ and $\mathbf{H}(\mathrm{div}, \mathcal{T}_h)$ respectively) that we shall need later.

5.1 Subspaces of $L^2(\Omega)$

Let $k \geq 0$ and let $\mathbb{P}_k(K)$ for a given mesh element $K \in \mathcal{T}_h$ denote the space of polynomials of total degree at most k on K. We then define

$$\mathbb{P}_k(\mathcal{T}_h) := \{v_h \in L^2(\Omega);\, v_h|_K \in \mathbb{P}_k(K) \quad \forall K \in \mathcal{T}_h\}, \tag{5.1}$$

the space of piecewise polynomials of maximal degree k on each $K \in \mathcal{T}_h$. Remark that there is no requirement on the continuity over the faces between elements, neither any imposing of boundary conditions. The abstract notation Q_h for $\mathbb{P}_k(\mathcal{T}_h)$ will be used often. We will below often employ the $L^2(\Omega)$-orthogonal projection onto Q_h: for each $v \in L^2(\Omega)$, $\Pi_{Q_h} v$ is the element of Q_h such that

$$(\Pi_{Q_h} v - v, q_h) = 0 \qquad \forall q_h \in Q_h. \tag{5.2}$$

5.2 Subspaces of $H^1(\mathcal{T}_h)$ and $H_0^1(\Omega)$

It is easily noted that the space $\mathbb{P}_k(\mathcal{T}_h)$ of (5.1) is also a finite-dimensional subspace of the space $H^1(\mathcal{T}_h)$ from Definition 4.3.1. On the other hand, $\mathbb{P}_k(\mathcal{T}_h)$ is not a subspace of the space $H_0^1(\Omega)$ from Definition 4.1.4.

We shall be using

$$\mathbb{P}_k(\mathcal{T}_h) \cap H_0^1(\Omega) = \{v_h \in H_0^1(\Omega);\, v_h|_K \in \mathbb{P}_k(K) \quad \forall K \in \mathcal{T}_h\} \tag{5.3}$$

as a discrete subspace of $H_0^1(\Omega)$, and often simply denote it as V_h. Recall from Theorem 4.4.2 that in order to make a $H^1(\mathcal{T}_h)$ function $H_0^1(\Omega)$-conforming, we need to make sure that the jumps over all mesh faces are zero. A trace of a k-th degree polynomial on a simplex K on its face e is a k-th degree polynomial on e. Thus, to ensure that the jump is zero, we need to impose the continuity in $\binom{d-1+k}{k}$ ($k+1$ in two space dimensions) points. This is perfectly doable when choosing the degrees of freedom of $\mathbb{P}_k(K)$ as illustrated in Figure 5.1 for $k = 1$ and $k = 2$ in two space dimensions. For these so-called *Lagrange finite elements*, the degrees of

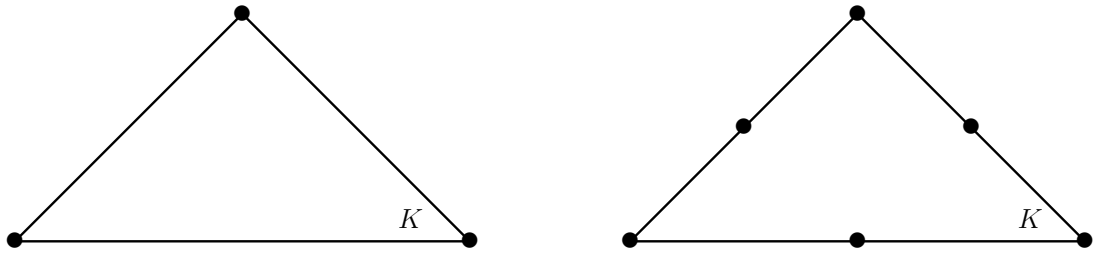


Figure 5.1: Degrees of freedom for the $\mathbb{P}_1(K)$ functions (left) and $\mathbb{P}_2(K)$ functions (right)

freedom are simply the values in a set of points, and there are exactly $\binom{d-1+k}{k}$ such point per mesh face. The hat functions $\psi_{\mathbf{a}}$, already defined in Section 3.3, form a basis of $\mathbb{P}_1(\mathcal{T}_h)\cap H^1_0(\Omega)$ while running over all $\mathbf{a}\in\mathcal{V}_h^{\mathrm{int}}$. We refer for details to classical textbooks, see, e.g., Ciarlet [33, Section 2.2] or Ern and Guermond [47, Section 1.2.3].

5.3 Subspaces of $\mathbf{H}(\mathrm{div},\mathcal{T}_h)$ and $\mathbf{H}(\mathrm{div},\Omega)$

We will in the sequel also extensively use finite-dimensional subspaces of the spaces $\mathbf{H}(\mathrm{div},\mathcal{T}_h)$ from Definition 4.3.4 and $\mathbf{H}(\mathrm{div},\Omega)$ of Definition 4.2.2. Let $K\in\mathcal{T}_h$. The starting point here for us will be the *Raviart–Thomas–Nédélec mixed finite element* space on K,

$$\mathbf{RTN}_k(K) := [\mathbb{P}_k(K)]^d + \mathbf{x}\mathbb{P}_k(K), \tag{5.4}$$

$k\geq 0$. In particular, $\mathbf{v}_h\in\mathbf{RTN}_k(K)$ is such that $\nabla\cdot\mathbf{v}_h\in\mathbb{P}_k(K)$ and $\mathbf{v}_h\cdot\mathbf{n}_e\in\mathbb{P}_k(e)$ for all $e\in\mathcal{E}_K$. The degrees of freedom here are integral moments up to order k of the normal trace on all faces and integral moments up to order $k-1$ on the element itself. Thus, in order to uniquely define a vector $\mathbf{v}_h$ from $\mathbf{RTN}_k(K)$, one can consider $d+1$ scalar functions $v_e\in L^2(e)$, $e\in\mathcal{E}_K$, and one vector function $\mathbf{v}\in L^2(K)$, and prescribe

$$\langle\mathbf{v}_h\cdot\mathbf{n}_e, q_h\rangle_e = \langle v_e, q_h\rangle_e \qquad \forall q_h\in\mathbb{P}_k(e),\ \forall e\in\mathcal{E}_K, \tag{5.5a}$$

$$(\mathbf{v}_h,\mathbf{r}_h)_K = (\mathbf{v},\mathbf{r}_h)_K \qquad \forall\mathbf{r}_h\in[\mathbb{P}_{k-1}(K)]^d. \tag{5.5b}$$

These degrees of freedom for $d=2$ and $k=0$ and $k=1$ are depicted in Figure 5.2, with arrows corresponding to (5.5a) and circles corresponding to (5.5b). Then

$$\mathbf{RTN}_k(\mathcal{T}_h) := \{\mathbf{v}_h\in[L^2(\Omega)]^d;\mathbf{v}_h|_K\in\mathbf{RTN}_k(K)\quad\forall K\in\mathcal{T}_h\}, \tag{5.6}$$

$k\geq 0$, is a finite-dimensional subspace of the broken divergence space $\mathbf{H}(\mathrm{div},\mathcal{T}_h)$.

Shall we think of piecewise vector polynomials from $\mathbf{RTN}_k(\mathcal{T}_h)$ belong to the $\mathbf{H}(\mathrm{div},\Omega)$ space, we know from Theorem 4.5.1 that we need to ensure the continuity of the normal traces over all faces $e\in\mathcal{E}_h^{\mathrm{int}}$. Remark that $\mathbf{v}\cdot\mathbf{n}_e$ are polynomials and thus definitely belong to $L^2(e)$ for each face e. But the degrees of freedom from (5.5a) together with the fact that $\mathbf{v}_h\cdot\mathbf{n}_e\in\mathbb{P}_k(e)$ immediately imply that matching them for two neighboring mesh elements will give a $\mathbf{H}(\mathrm{div},\Omega)$-conforming function. In the lowest-order $k=0$ case, there is, consequently, one basis function $\mathbf{v}_e$ for each face $e\in\mathcal{E}_h$, cf. Figure 5.3 for $d=2$. In general, the Raviart–Thomas–Nédélec finite-dimensional subspace of the space $\mathbf{H}(\mathrm{div},\Omega)$ is

$$\mathbf{RTN}_k := \mathbf{RTN}_k(\mathcal{T}_h)\cap\mathbf{H}(\mathrm{div},\Omega) = \{\mathbf{v}_h\in\mathbf{H}(\mathrm{div},\Omega);\mathbf{v}_h|_K\in\mathbf{RTN}_k(K)\quad\forall K\in\mathcal{T}_h\}, \tag{5.7}$$

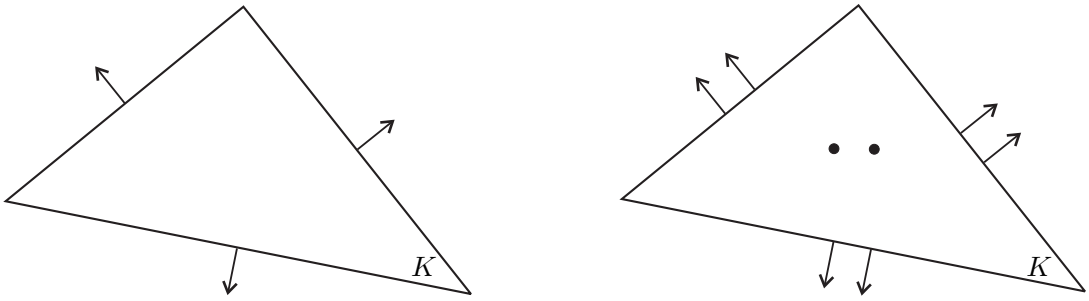


Figure 5.2: Degrees of freedom for the $\mathbf{RTN}_0(K)$ functions (left) and $\mathbf{RTN}_1(K)$ functions (right)

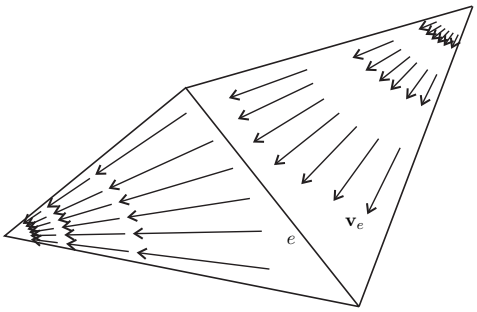


Figure 5.3: The basis function $\mathbf{v}_e$ of $\mathbf{RTN}_0$ associated with $e \in \mathcal{E}_h^{\mathrm{int}}$

often abstractly denoted as $\mathbf{V}_h$. In mixed finite element discretizations, a couple of spaces, $\mathbf{RTN}_k \times \mathbb{P}_k(\mathcal{T}_h)$, $k \geq 0$, or shortly $\mathbf{V}_h \times Q_h$, will be employed. In particular, all the flux reconstructions $\boldsymbol{\sigma}_h$ in these lecture notes will be constructed in the space $\mathbf{RTN}_k$. Details on $\mathbf{H}(\mathrm{div}, \mathcal{T}_h)$- and $\mathbf{H}(\mathrm{div}, \Omega)$-conforming subspaces can be found in Brezzi and Fortin [24] or Roberts and Thomas [87].

Chapter 6

Primal, dual, and dual mixed formulations; minimization, constrained minimization, and saddle-point problems

The construction and analysis of the optimal a posteriori error estimates in the forthcoming chapters will be based on the approximation of local Neumann / Neumann–Dirichlet problems by mixed finite elements. This approximation stems from alternative variational formulations, the so-called dual and dual mixed formulations. We recall in this chapter these alternative formulations and their relations to the classical primal one. The presentation, following the classical textbooks (see Brezzi and Fortin [24], Roberts and Thomas [87], Quarteroni and Valli [83], or Ern and Guermond [47]), will help us to easily understand the structure of the estimates.

6.1 A model problem

Let $\omega \subset \mathbb{R}^d$, $d \geq 1$, be a polytopic (polygonal for $d = 2$, polyhedral for $d = 3$) domain (open, bounded, and connected set). Later on, ω will stand for subsets of the domain Ω, typically for the patches $\omega_{\mathbf{a}}$ of all simplices sharing the given vertex $\mathbf{a}$ of the computational mesh $\mathcal{T}_h$. We consider two different cases. In the first one, a homogeneous Neumann boundary condition will be prescribed on the whole $\partial\omega$. In this case, we set $\partial\omega_{\mathrm{N}} := \partial\omega$ and $\partial\omega_{\mathrm{D}} := \emptyset$. In the second case, we suppose that $\partial\omega$ is divided into two simply connected disjoint parts $\partial\omega_{\mathrm{D}}$ and $\partial\omega_{\mathrm{N}}$ with $|\partial\omega_{\mathrm{D}}| > 0$; a homogeneous Neumann boundary condition will be imposed on $\partial\omega_{\mathrm{N}}$ and a homogeneous Dirichlet boundary condition will be imposed on $\partial\omega_{\mathrm{D}}$. Let $L^2_*(\omega)$ stand for the space of all functions from $L^2(\omega)$ with mean value zero in the first case and for $L^2(\omega)$ in the second one. Let finally $\boldsymbol{\tau} \in [L^2(\omega)]^d$ and $g \in L^2_*(\omega)$ be arbitrary. We consider the problem of finding a function $r : \omega \to \mathbb{R}$, with mean value zero in the first case, such that

$$-\nabla\cdot(\nabla r + \boldsymbol{\tau}) = g \qquad \text{in } \omega, \tag{6.1a}$$

$$-(\nabla r + \boldsymbol{\tau})\cdot\mathbf{n}_\omega = 0 \qquad \text{on } \partial\omega_{\mathrm{N}}, \tag{6.1b}$$

$$r = 0 \qquad \text{on } \partial\omega_{\mathrm{D}}. \tag{6.1c}$$

6.2 Primal, dual, and dual mixed variational formulations

Let $H^1_*(\omega)$ stand for the space of all functions from $H^1(\omega)$ with zero mean value in the first case and for all functions from $H^1(\omega)$ with zero trace on $\partial\omega_{\mathrm{D}}$ in the second one. The first formulation we consider is the classical primal one, compare with Definitions 2.2.1 and 7.1.1:

Definition 6.2.1 (Primal formulation)**.** *Find* $r \in H^1_*(\omega)$ *such that*

$$(\nabla r, \nabla v)_\omega = -(\boldsymbol{\tau}, \nabla v)_\omega + (g, v)_\omega \qquad \forall v \in H^1_*(\omega). \tag{6.2}$$

There exists one and only one solution to (6.2) by the Riesz representation theorem. Indeed, $(\nabla\cdot, \nabla\cdot)_\omega$ is a scalar product on $H^1_*(\omega)$ thanks to the Poincaré inequality

$$\|v\|_\omega \le C_{\mathrm{P},\omega} h_\omega \|\nabla v\|_\omega \qquad \forall v \in H^1_*(\omega)$$

in the first case and thanks to the Friedrichs inequality

$$\|v\|_\omega \le C_{\mathrm{F},\omega} h_\omega \|\nabla v\|_\omega \qquad \forall v \in H^1_*(\omega)$$

in the second case, cf. (4.20)–(4.21), whereas $-(\boldsymbol{\tau}, \nabla\cdot)_\omega + (g, \cdot)_\omega$ is a continuous linear form on $H^1_*(\omega)$. Note that the Neumann compatibility condition

$$-(\boldsymbol{\tau}, \nabla 1)_\omega + (g, 1)_\omega = 0 \tag{6.3}$$

is satisfied in the first case.

Let now $\mathbf{H}_*(\mathrm{div}, \omega)$ stand for $\mathbf{H}(\mathrm{div}, \omega)$ functions with zero normal trace on all $\partial\omega$ in the appropriate sense, i.e.,

$$\mathbf{H}_*(\mathrm{div}, \omega) := \{\mathbf{v} \in \mathbf{H}(\mathrm{div}, \omega);\ (\mathbf{v}, \nabla\varphi)_\omega + (\nabla\cdot\mathbf{v}, \varphi)_\omega = 0 \qquad \forall \varphi \in H^1(\omega)\}$$

in the first case and with zero normal trace only on $\partial\omega_{\mathrm{N}}$ in the appropriate sense, i.e.,

$$\mathbf{H}_*(\mathrm{div}, \omega) := \{\mathbf{v} \in \mathbf{H}(\mathrm{div}, \omega);\ (\mathbf{v}, \nabla\varphi)_\omega + (\nabla\cdot\mathbf{v}, \varphi)_\omega = 0 \qquad \forall \varphi \in H^1_*(\omega)\}$$

in the second one. The two other formulations are:

Definition 6.2.2 (Dual formulation)**.** *Find* $\boldsymbol{\varsigma} \in \mathbf{H}_*(\mathrm{div}, \omega)$ *with* $\nabla\cdot\boldsymbol{\varsigma} = g$ *such that*

$$(\boldsymbol{\varsigma}, \mathbf{v})_\omega = -(\boldsymbol{\tau}, \mathbf{v})_\omega \qquad \forall \mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega) \text{ with } \nabla\cdot\mathbf{v} = 0. \tag{6.4}$$

Definition 6.2.3 (Dual mixed formulation)**.** *Find a couple* $(\boldsymbol{\varsigma}, r) \in \mathbf{H}_*(\mathrm{div}, \omega) \times L^2_*(\omega)$ *such that*

$$(\boldsymbol{\varsigma}, \mathbf{v})_\omega - (r, \nabla\cdot\mathbf{v})_\omega = -(\boldsymbol{\tau}, \mathbf{v})_\omega \qquad \forall \mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega), \tag{6.5a}$$

$$(\nabla\cdot\boldsymbol{\varsigma}, q)_\omega = (g, q)_\omega \qquad \forall q \in L^2_*(\omega). \tag{6.5b}$$

6.3 The relations between the different variational formulations

The existence and uniqueness of the solutions to (6.4) and (6.5) can be proven be means of appropriate variational theories while proceeding as in [24, 87, 83, 47]. It is, however, straightforward to prove them directly. Such an approach has the additional advantage of unveiling the links between the different formulations:

Theorem 6.3.1 (Existence and uniqueness of the dual and dual mixed formulations; equivalence with the primal formulation)**.** *There exists a unique solution* $\boldsymbol{\varsigma}$ *of Definition 6.2.2 and a unique solution couple* $(\boldsymbol{\varsigma}, r)$ *of Definition 6.2.3. Moreover, all the formulations of Definitions 6.2.1, 6.2.2, and 6.2.3 are equivalent, in the sense that* r *from Definitions 6.2.1 and 6.2.3 coincide, that* $\boldsymbol{\varsigma}$ *from Definitions 6.2.2 and 6.2.3 coincide, and that* $\boldsymbol{\varsigma} = -\nabla r - \boldsymbol{\tau}$.

Proof. We proceed in three steps to prove the equivalence of the three Definitions 6.2.1, 6.2.2, and 6.2.3. The existence and uniqueness result for Definitions 6.2.2 and 6.2.3 then follows from that of Definition 6.2.1.

(i) Consider $r \in H^1_*(\omega)$ the solution of (6.2) and set $\boldsymbol{\varsigma} := -\nabla r - \boldsymbol{\tau}$. We check that $\boldsymbol{\varsigma}$ coincides with that of (6.4). We start by verifying that $\boldsymbol{\varsigma} \in \mathbf{H}(\mathrm{div}, \omega)$ with $\nabla{\cdot}\boldsymbol{\varsigma} = g$ according to Definition 4.2.2. First, $\nabla r \in [L^2(\omega)]^d$ and $\boldsymbol{\tau} \in [L^2(\omega)]^d$, so that condition 1 of Definition 4.2.1 is satisfied. The supposed weak divergence of $\boldsymbol{\varsigma}$ is g, which indeed belongs to $L^2(\omega)$, complying with condition 2a. The last condition 2b then follows from (6.2) (note that constant test functions are also authorized in the first case thanks to the Neumann compatibility condition (6.3)). Moreover, (6.2) and the Green theorem give

$$0 = (\boldsymbol{\varsigma}, \nabla v)_\omega + (\nabla{\cdot}\boldsymbol{\varsigma}, v)_\omega = \langle \boldsymbol{\varsigma}{\cdot}\mathbf{n}_\omega, v\rangle_{\partial\omega} \qquad \forall v \in H^1_*(\omega),$$

with $v = 1$ authorized as well in the first case. Thus $\boldsymbol{\varsigma}$ satisfies the requested homogeneous Neumann boundary condition on $\partial\omega_\mathrm{N}$ and belongs to $\mathbf{H}_*(\mathrm{div}, \omega)$. Finally,

$$(\boldsymbol{\varsigma} + \boldsymbol{\tau}, \mathbf{v})_\omega = -(\nabla r, \mathbf{v})_\omega = -\langle \mathbf{v}{\cdot}\mathbf{n}_\omega, r\rangle_{\partial\omega} = 0$$

for any $\mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega)$ with $\nabla{\cdot}\mathbf{v} = 0$ by the Green theorem, so that (6.4) indeed holds true.

(ii) Let $\boldsymbol{\varsigma} \in \mathbf{H}_*(\mathrm{div}, \omega)$ with $\nabla{\cdot}\boldsymbol{\varsigma} = g$ solve (6.4). Define $r \in L^2_*(\omega)$ by

$$(r, \nabla{\cdot}\mathbf{v})_\omega = (\boldsymbol{\varsigma} + \boldsymbol{\tau}, \mathbf{v})_\omega$$

for all $\mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega)$. Then it is immediate that the couple $(\boldsymbol{\varsigma}, r)$ solves (6.5).

(iii) We are left to verify that r from (6.5) solves (6.2) and that $\boldsymbol{\varsigma} = -\nabla r - \boldsymbol{\tau}$. We first check that $r \in H^1(\omega)$ with $\nabla r = -\boldsymbol{\varsigma} - \boldsymbol{\tau}$ according to Definition 4.1.3. We know that $r \in L^2_*(\omega)$, so that condition 1 of Definition 4.1.1 is satisfied. Checking $-\boldsymbol{\varsigma} - \boldsymbol{\tau} \in [L^2(\omega)]^d$ verifies the condition 2a for all $1 \le i \le d$. Finally, let $1 \le i \le d$ and $\varphi \in \mathcal{D}(\omega)$. Setting $\mathbf{v}^i := \varphi$ and $\mathbf{v}^j = 0$, $j \ne i$, $\mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega)$ and (6.5a) implies (2b). Additionally, r is of zero mean value in the first case, so that it actually belongs to $H^1_*(\omega)$. The fact that $r \in H^1_*(\omega)$, i.e., that $r = 0$ on $\partial\omega_\mathrm{D}$ in the second case again follows from the Green theorem and (6.5a): for all $\mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega)$, there holds

$$\langle \mathbf{v}{\cdot}\mathbf{n}_\omega, r\rangle_{\partial\omega} = (\mathbf{v}, \nabla r)_\omega + (\nabla{\cdot}\mathbf{v}, r)_\omega = -(\mathbf{v}, \boldsymbol{\varsigma} + \boldsymbol{\tau})_\omega + (\nabla{\cdot}\mathbf{v}, r)_\omega = 0.$$

Finally, taking in (6.5b) the test function q from the space $H^1_*(\omega)$ and using $\boldsymbol{\varsigma} = -\nabla r - \boldsymbol{\tau} \in \mathbf{H}_*(\mathrm{div}, \omega)$ and the Green theorem gives (6.2). □

6.4 Minimization, constrained minimization, and saddle-point problems

We now state equivalent formulations of Definitions 6.2.1, 6.2.2, and 6.2.3, with a motivation to enable a better insight into their later use in a posteriori error estimation. Let

$$\mathcal{J}(v) := \frac{1}{2}\|\nabla v\|^2_\omega + (\boldsymbol{\tau}, \nabla v)_\omega - (g, v)_\omega \tag{6.6}$$

define the energy functional for $v \in H^1_*(\omega)$.

We start by the following classical result, see, e.g., [33, 83, 47].

Definition 6.4.1 (Energy minimization)**.** *Find r such that*

$$r := \arg \inf_{v \in H^1_*(\omega)} \mathcal{J}(v). \tag{6.7}$$

Lemma 6.4.2 (Equivalence of energy minimization with the primal formulation)**.** *The solution r of the minimization problem of Definition 6.4.1 coincides with that of the primal variational formulation of Definition 6.2.1.*

Proof. For any $r \in H^1_*(\omega)$, $v \in H^1_*(\omega)$, and any real t, develop

$$\begin{aligned} \mathcal{J}(r+tv) = \frac{1}{2}\|\nabla r\|^2_\omega + t(\nabla r, \nabla v)_\omega + \frac{1}{2}t^2\|\nabla v\|^2_\omega + (\boldsymbol{\tau}, \nabla r)_\omega \\ + t(\boldsymbol{\tau}, \nabla v)_\omega - (g, r)_\omega - t(g, v)_\omega. \end{aligned} \tag{6.8}$$

(i) Suppose first that $r \in H^1_*(\omega)$ solves (6.2) and consider $t = 1$ in (6.8). Then the orthogonality condition (6.2) and the fact that $\frac{1}{2}\|\nabla v\|^2_\omega \geq 0$ give

$$\begin{aligned} \mathcal{J}(r+v) &= \frac{1}{2}\|\nabla(r+v)\|^2_\omega + (\boldsymbol{\tau}, \nabla(r+v))_\omega - (g, r+v)_\omega \\ &\geq \frac{1}{2}\|\nabla r\|^2_\omega + (\boldsymbol{\tau}, \nabla r)_\omega - (g, r)_\omega = \mathcal{J}(r) \end{aligned}$$

for any $v \in H^1_*(\omega)$, with equality occurring for $v = 0$.

(ii) Conversely, let r solve (6.7). Then

$$\begin{aligned} \mathcal{J}(r+tv) &= \frac{1}{2}\|\nabla(r+tv)\|^2_\omega + (\boldsymbol{\tau}, \nabla(r+tv))_\omega - (g, r+tv)_\omega \\ &\geq \frac{1}{2}\|\nabla r\|^2_\omega + (\boldsymbol{\tau}, \nabla r)_\omega - (g, r)_\omega = \mathcal{J}(r) \end{aligned}$$

for any $v \in H^1_*(\omega)$ and any real t. Comparing this with the development (6.8), we obtain

$$t(\nabla r, \nabla v)_\omega + \frac{1}{2}t^2\|\nabla v\|^2_\omega + t(\boldsymbol{\tau}, \nabla v)_\omega - t(g, v)_\omega \geq 0.$$

Let $t > 0$. Dividing by t and letting $t \searrow 0$, we obtain

$$(\nabla r, \nabla v)_\omega \geq -(\boldsymbol{\tau}, \nabla v)_\omega + (g, v)_\omega \qquad \forall v \in H^1_*(\omega).$$

Let $t < 0$. Dividing by t and then letting $t \nearrow 0$, we obtain the converse inequality

$$(\nabla r, \nabla v)_\omega \leq -(\boldsymbol{\tau}, \nabla v)_\omega + (g, v)_\omega \qquad \forall v \in H^1_*(\omega).$$

Thus the equality (6.2) is proven. □

Remark 6.4.3 (Comments on Lemma 6.4.2)**.** *Lemma 6.4.2 is the usual result of a minimization of a quadratic functional, see Ciarlet [33, Theorems 1.1.1 and 1.1.2] or Ern and Guermond [47, Proposition 2.4] in context of partial differential equations. Here we have actually given the proof of the existence and uniqueness of the minimization problem (6.7), via its link to (6.2) which is nothing but its Euler optimality conditions. Note also that it follows from this proof that we can actually replace* inf *by* min *in (6.7).*

For any $\mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega)$, define

$$\mathcal{I}(\mathbf{v}) := \frac{1}{2}\|\boldsymbol{\tau} + \mathbf{v}\|_\omega^2 \tag{6.9}$$

the complementary energy, and for any couple $(\mathbf{v}, q) \in \mathbf{H}_*(\mathrm{div}, \omega) \times L_*^2(\omega)$, define

$$\mathcal{L}(\mathbf{v}, q) := \frac{1}{2}\|\boldsymbol{\tau} + \mathbf{v}\|_\omega^2 - (\nabla\cdot\mathbf{v} - g, q)_\omega \tag{6.10}$$

the Lagrangian. The following similar results on the alternative formulations are also rather standard, see [24, 87, 83]:

Definition 6.4.4 (Complementary energy constrained minimization)**.** *Find* $\boldsymbol{\varsigma}$ *such that*

$$\boldsymbol{\varsigma} := \arg \inf_{\mathbf{v}\in\mathbf{H}_*(\mathrm{div},\omega)\,\nabla\cdot\mathbf{v}=g} \mathcal{I}(\mathbf{v}). \tag{6.11}$$

Lemma 6.4.5 (Equivalence of complementary energy constrained minimization with the dual formulation)**.** *The solution* $\boldsymbol{\varsigma}$ *of the constrained minimization problem of Definition 6.4.4 coincides with that of the dual variational formulation of Definition 6.2.2.*

Proof. For any $\boldsymbol{\varsigma} \in \mathbf{H}_*(\mathrm{div}, \omega)$ with $\nabla\cdot\boldsymbol{\varsigma} = g$, any $\mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega)$ with $\nabla\cdot\mathbf{v} = 0$, and any real t, develop

$$\mathcal{I}(\boldsymbol{\varsigma} + t\mathbf{v}) = \frac{1}{2}\|\boldsymbol{\tau} + \boldsymbol{\varsigma} + t\mathbf{v}\|_\omega^2 = \frac{1}{2}\|\boldsymbol{\tau} + \boldsymbol{\varsigma}\|_\omega^2 + t(\boldsymbol{\tau} + \boldsymbol{\varsigma}, \mathbf{v})_\omega + \frac{1}{2}t^2\|\mathbf{v}\|_\omega^2. \tag{6.12}$$

(i) Suppose first that $\boldsymbol{\varsigma} \in \mathbf{H}_*(\mathrm{div}, \omega)$ with $\nabla\cdot\boldsymbol{\varsigma} = g$ solves (6.4) and consider $t = 1$ in (6.12). Then the orthogonality condition (6.4) and the fact that $\frac{1}{2}\|\mathbf{v}\|_\omega^2 \geq 0$ give

$$\mathcal{I}(\boldsymbol{\varsigma} + \mathbf{v}) = \frac{1}{2}\|\boldsymbol{\tau} + \boldsymbol{\varsigma} + \mathbf{v}\|_\omega^2 \geq \frac{1}{2}\|\boldsymbol{\tau} + \boldsymbol{\varsigma}\|_\omega^2 = \mathcal{I}(\boldsymbol{\varsigma})$$

for any $\mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega)$ with $\nabla\cdot\mathbf{v} = 0$, with equality occurring for $\mathbf{v} = \mathbf{0}$.

(ii) Conversely, let $\boldsymbol{\varsigma}$ solve (6.11). Then

$$\mathcal{I}(\boldsymbol{\varsigma} + t\mathbf{v}) = \frac{1}{2}\|\boldsymbol{\tau} + \boldsymbol{\varsigma} + t\mathbf{v}\|_\omega^2 \geq \frac{1}{2}\|\boldsymbol{\tau} + \boldsymbol{\varsigma}\|_\omega^2 = \mathcal{I}(\boldsymbol{\varsigma})$$

for any $\mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega)$ with $\nabla\cdot\mathbf{v} = 0$ and any real t. Comparing this with the development (6.12), we obtain

$$t(\boldsymbol{\tau} + \boldsymbol{\varsigma}, \mathbf{v})_\omega + \frac{1}{2}t^2\|\mathbf{v}\|_\omega^2 \geq 0.$$

Let $t > 0$. Dividing by t and then letting $t \searrow 0$, we obtain

$$(\boldsymbol{\tau} + \boldsymbol{\varsigma}, \mathbf{v})_\omega \geq 0 \qquad \forall \mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega) \text{ with } \nabla\cdot\mathbf{v} = 0.$$

Let $t < 0$. Dividing by t and then letting $t \nearrow 0$, we obtain the converse inequality

$$(\boldsymbol{\tau} + \boldsymbol{\varsigma}, \mathbf{v})_\omega \leq 0 \qquad \forall \mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega) \text{ with } \nabla\cdot\mathbf{v} = 0.$$

Thus the equality (6.4) is proven. □

Definition 6.4.6 (Saddle point search)**.** *Find* $(\boldsymbol{\varsigma}, r) \in \mathbf{H}_*(\mathrm{div}, \omega) \times L_*^2(\omega)$ *such that*

$$\mathcal{L}(\boldsymbol{\varsigma}, q) \leq \mathcal{L}(\boldsymbol{\varsigma}, r) \leq \mathcal{L}(\mathbf{v}, r) \qquad \forall (\mathbf{v}, q) \in \mathbf{H}_*(\mathrm{div}, \omega) \times L_*^2(\omega). \tag{6.13}$$

Lemma 6.4.7 (Equivalence of the saddle point search with the dual mixed formulation)**.** *The solution couple* $(\boldsymbol{\varsigma}, r)$ *of the saddle point problem of Definition 6.4.6 coincides with that of the dual mixed variational formulation of Definition 6.2.3.*

Proof. For any $\boldsymbol{\varsigma} \in \mathbf{H}_*(\mathrm{div}, \omega)$, any $\mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega)$, any $r \in L^2_*(\omega)$, and any real t, develop

$$\mathcal{L}(\boldsymbol{\varsigma} + t\mathbf{v}, r) = \frac{1}{2}\|\boldsymbol{\tau} + \boldsymbol{\varsigma}\|^2_\omega + t(\boldsymbol{\tau} + \boldsymbol{\varsigma}, \mathbf{v})_\omega + \frac{1}{2}t^2\|\mathbf{v}\|^2_\omega - (\nabla{\cdot}\boldsymbol{\varsigma} - g, r)_\omega - t(\nabla{\cdot}\mathbf{v}, r)_\omega. \tag{6.14a}$$

Similarly, for any $\boldsymbol{\varsigma} \in \mathbf{H}_*(\mathrm{div}, \omega)$, any $r \in L^2_*(\omega)$, any $q \in L^2_*(\omega)$, and any real t, develop

$$\mathcal{L}(\boldsymbol{\varsigma}, r + tq) = \frac{1}{2}\|\boldsymbol{\tau} + \boldsymbol{\varsigma}\|^2_\omega - (\nabla{\cdot}\boldsymbol{\varsigma} - g, r)_\omega - t(\nabla{\cdot}\boldsymbol{\varsigma} - g, q)_\omega. \tag{6.14b}$$

(i) Suppose first that $\boldsymbol{\varsigma} \in \mathbf{H}_*(\mathrm{div}, \omega)$ and $r \in L^2_*(\omega)$ solve (6.5). Consider $t = 1$ in (6.14a). Then the orthogonality condition (6.5a) and the fact that $\frac{1}{2}\|\mathbf{v}\|^2_\omega \geq 0$ give

$$\mathcal{L}(\boldsymbol{\varsigma} + \mathbf{v}, r) \geq \frac{1}{2}\|\boldsymbol{\tau} + \boldsymbol{\varsigma}\|^2_\omega - (\nabla{\cdot}\boldsymbol{\varsigma} - g, r)_\omega = \mathcal{L}(\boldsymbol{\varsigma}, r)$$

for any $\mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega)$, with equality occurring for $\mathbf{v} = \mathbf{0}$. Similarly, taking $t = 1$ in (6.14b) and considering the orthogonality condition (6.5b), we obtain

$$\mathcal{L}(\boldsymbol{\varsigma}, r + q) = \frac{1}{2}\|\boldsymbol{\tau} + \boldsymbol{\varsigma}\|^2_\omega - (\nabla{\cdot}\boldsymbol{\varsigma} - g, r)_\omega = \mathcal{L}(\boldsymbol{\varsigma}, r)$$

for all $q \in L^2_*(\omega)$.

(ii) Conversely, let $(\boldsymbol{\varsigma}, r) \in \mathbf{H}_*(\mathrm{div}, \omega) \times L^2_*(\omega)$ solve (6.13). Then

$$\mathcal{L}(\boldsymbol{\varsigma} + t\mathbf{v}, r) \geq \mathcal{L}(\boldsymbol{\varsigma}, r) = \frac{1}{2}\|\boldsymbol{\tau} + \boldsymbol{\varsigma}\|^2_\omega - (\nabla{\cdot}\boldsymbol{\varsigma} - g, r)_\omega$$

for any $\mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega)$ and any real t. Comparing this with the development (6.14a), we obtain

$$t(\boldsymbol{\tau} + \boldsymbol{\varsigma}, \mathbf{v})_\omega + \frac{1}{2}t^2\|\mathbf{v}\|^2_\omega - t(\nabla{\cdot}\mathbf{v}, r)_\omega \geq 0.$$

Let $t > 0$. Dividing by t and then letting $t \searrow 0$, we obtain

$$(\boldsymbol{\tau} + \boldsymbol{\varsigma}, \mathbf{v})_\omega - (\nabla{\cdot}\mathbf{v}, r)_\omega \geq 0 \qquad \forall \mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega).$$

Let $t < 0$. Dividing by t and then letting $t \nearrow 0$, we obtain the converse inequality

$$(\boldsymbol{\tau} + \boldsymbol{\varsigma}, \mathbf{v})_\omega - (\nabla{\cdot}\mathbf{v}, r)_\omega \leq 0 \qquad \forall \mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega).$$

Thus the equality (6.5a) is proven.

Similarly, we know that

$$\mathcal{L}(\boldsymbol{\varsigma}, r + tq) \leq \mathcal{L}(\boldsymbol{\varsigma}, r) = \frac{1}{2}\|\boldsymbol{\tau} + \boldsymbol{\varsigma}\|^2_\omega - (\nabla{\cdot}\boldsymbol{\varsigma} - g, r)_\omega$$

for any $q \in L^2_*(\omega)$ and any real t. By comparison with (6.14b), we have

$$-t(\nabla{\cdot}\boldsymbol{\varsigma} - g, q)_\omega \leq 0.$$

Let $t > 0$. Dividing by t gives

$$-(\nabla{\cdot}\boldsymbol{\varsigma} - g, q)_\omega \leq 0 \qquad \forall q \in L^2_*(\omega).$$

Vice versa, for $t < 0$, the same division leads to

$$-(\nabla{\cdot}\boldsymbol{\varsigma} - g, q)_\omega \geq 0 \qquad \forall q \in L^2_*(\omega).$$

Thus the equality (6.5b) is proven and the proof is finished. □

6.5 Energy (in)equalities

Combining the results of the previous sections, we have the following equalities:

Theorem 6.5.1 (Energy equalities)**.** *Let the couple* $(\boldsymbol{\varsigma}, r)$ *be as specified in Definitions 6.2.1–6.2.3 and 6.4.1–6.4.6. Then*

$$\sup_{\mathbf{v}\in\mathbf{H}_*(\mathrm{div},\omega)\,\nabla\cdot\mathbf{v}=g} -\mathcal{I}(\mathbf{v}) = -\mathcal{I}(\boldsymbol{\varsigma}) = -\mathcal{L}(\boldsymbol{\varsigma}, r) = \mathcal{J}(r) = \inf_{v\in H^1_*(\omega)} \mathcal{J}(v). \tag{6.15}$$

Proof. The first equality is the assertion of Lemma 6.4.5, whereas the last one of Lemma 6.4.2. The second equality follows from the fact that $\nabla\cdot\boldsymbol{\varsigma} = g$, so that the second term of $\mathcal{L}(\cdot,\cdot)$ vanishes. Finally, taking $v = r$ in (6.6), using (6.2), and employing the fact that $\boldsymbol{\varsigma} = -\nabla r - \boldsymbol{\tau}$ from Theorem 6.3.1,

$$\mathcal{J}(r) = \frac{1}{2}\|\nabla r\|^2_\omega + (\boldsymbol{\tau}, \nabla r)_\omega - (g, r)_\omega = -\frac{1}{2}\|\nabla r\|^2_\omega = -\frac{1}{2}\|\boldsymbol{\varsigma} + \boldsymbol{\tau}\|^2_\omega.$$

□

Corollary 6.5.2 (Energy inequality)**.** *Let* $\mathbf{v} \in \mathbf{H}_*(\mathrm{div},\omega)$ *with* $\nabla\cdot\mathbf{v} = g$ *and* $v \in H^1_*(\omega)$ *be arbitrary. Then*

$$-\mathcal{I}(\mathbf{v}) \le \mathcal{J}(v).$$

6.6 Finite-dimensional approximations

We finally take a quick look on the counterparts of the problems of Definitions 6.2.1–6.2.3 and 6.4.1–6.4.6 on a finite-dimensional level. Let V_h be a finite-dimensional subspace of the space $H^1_*(\omega)$ and let $\mathbf{V}_h \times Q_h$ be a finite-dimensional subspace of $\mathbf{H}_*(\mathrm{div},\omega) \times L^2_*(\omega)$ satisfying $\nabla\cdot\mathbf{V}_h = Q_h$. Examples are given in Chapter 5.

We start with the counterpart of Definition 6.2.1:

Definition 6.6.1 (Primal approximation)**.** *Find* $r_h \in V_h$ *such that*

$$(\nabla r_h, \nabla v_h)_\omega = -(\boldsymbol{\tau}, \nabla v_h)_\omega + (g, v_h)_\omega \qquad \forall v_h \in V_h. \tag{6.16}$$

As for Definition 6.2.1, there exists one and only one solution to (6.16) by the Riesz representation theorem.

The finite-dimensional counterparts of Definitions 6.2.2 and 6.2.3 are:

Definition 6.6.2 (Dual approximation)**.** *Find* $\boldsymbol{\varsigma}_h \in \mathbf{V}_h$ *with* $\nabla\cdot\boldsymbol{\varsigma}_h = \Pi_{Q_h} g$ *such that*

$$(\boldsymbol{\varsigma}_h, \mathbf{v}_h)_\omega = -(\boldsymbol{\tau}, \mathbf{v}_h)_\omega \qquad \forall \mathbf{v}_h \in \mathbf{V}_h \text{ with } \nabla\cdot\mathbf{v}_h = 0. \tag{6.17}$$

Definition 6.6.3 (Dual mixed approximation)**.** *Find a couple* $(\boldsymbol{\varsigma}_h, \bar{r}_h) \in \mathbf{V}_h \times Q_h$ *such that*

$$(\boldsymbol{\varsigma}_h, \mathbf{v}_h)_\omega - (\bar{r}_h, \nabla\cdot\mathbf{v}_h)_\omega = -(\boldsymbol{\tau}, \mathbf{v}_h)_\omega \qquad \forall \mathbf{v}_h \in \mathbf{V}_h, \tag{6.18a}$$

$$(\nabla\cdot\boldsymbol{\varsigma}_h, q_h)_\omega = (g, q_h)_\omega \qquad \forall q_h \in Q_h. \tag{6.18b}$$

The equivalences expressed in Theorem 6.3.1 do not completely hold true anymore on the discrete level. In particular, the primal formulation of Definition 6.6.1 is not anymore equivalent to the dual ones of Definitions 6.6.2–6.6.3 and r_h from (6.16) does not equal $\bar{r}_h$ from (6.18). We, however, still have:

Theorem 6.6.4 (Existence and uniqueness of the dual and dual mixed formulations and their equivalence)**.** *There exists a unique solution $\boldsymbol{\varsigma}_h$ of Definition 6.6.2 and a unique solution couple $(\boldsymbol{\varsigma}_h, \bar{r}_h)$ of Definition 6.6.3. Moreover, the formulations of Definitions 6.6.2, and 6.6.3 are equivalent, in the sense that $\boldsymbol{\varsigma}_h$ from Definitions 6.6.2 and 6.6.3 coincide.*

Proof. We first show the equivalence. Let $\boldsymbol{\varsigma}_h \in \mathbf{V}_h$ with $\nabla{\cdot}\boldsymbol{\varsigma}_h = \Pi_{Q_h} g$ solve (6.17). Note that $\nabla{\cdot}\boldsymbol{\varsigma}_h = \Pi_{Q_h} g$ is equivalently stated by (6.18b). Then, using the condition $\nabla{\cdot}\mathbf{V}_h = Q_h$, we simply prescribe $\bar{r}_h \in Q_h$ by

$$(\bar{r}_h, \nabla{\cdot}\mathbf{v}_h)_\omega := (\boldsymbol{\varsigma}_h, \mathbf{v}_h)_\omega + (\boldsymbol{\tau}, \mathbf{v}_h)_\omega \qquad \forall \mathbf{v}_h \in \mathbf{V}_h.$$

To see the converse direction, it suffices to use that (6.18b) is equivalent to $\nabla{\cdot}\boldsymbol{\varsigma}_h = \Pi_{Q_h} g$ and that (6.18a) for a divergence-free $\boldsymbol{\varsigma}_h$ gives (6.17).

We now turn to the existence and uniqueness. Let us show it for the problem (6.18). This is a square linear finite-dimensional system. It thus suffices to prove that setting the right-hand side zero implies that the solution is zero, i.e., that setting $\boldsymbol{\tau} = \mathbf{0}$ and $g = 0$ implies $\boldsymbol{\varsigma}_h = \mathbf{0}$ and $\bar{r}_h = 0$. Let $\boldsymbol{\tau} = \mathbf{0}$ and $g = 0$ and set $q_h = \bar{r}_h$ in (6.18b) and $\mathbf{v}_h = \boldsymbol{\varsigma}_h$ in (6.18a) and sum the equations. This gives $(\boldsymbol{\varsigma}_h, \boldsymbol{\varsigma}_h)_\omega = 0$, whence $\boldsymbol{\varsigma}_h = \mathbf{0}$ follows. Consequently, $(\bar{r}_h, \nabla{\cdot}\mathbf{v}_h)_\omega = 0$ for all $\mathbf{v}_h \in \mathbf{V}_h$, whence $\bar{r}_h = 0$ follows by the assumption $\nabla{\cdot}\mathbf{V}_h = Q_h$. □

Finally, the counterparts of Definitions 6.4.1, 6.4.4, and 6.4.6 are as follows:

Definition 6.6.5 (Discrete energy minimization)**.**

$$r_h := \arg \inf_{v_h \in V_h} \mathcal{J}(v_h). \tag{6.19}$$

Definition 6.6.6 (Discrete complementary energy constrained minimization)**.**

$$\boldsymbol{\varsigma} := \arg \inf_{\mathbf{v}_h \in \mathbf{V}_h} \inf_{\nabla{\cdot}\mathbf{v}_h = \Pi_{Q_h} g} \mathcal{I}(\mathbf{v}_h). \tag{6.20}$$

Definition 6.6.7 (Discrete saddle point search)**.** *Find $(\boldsymbol{\varsigma}_h, \bar{r}_h) \in \mathbf{V}_h \times Q_h$ such that*

$$\mathcal{L}(\boldsymbol{\varsigma}_h, q_h) \le \mathcal{L}(\boldsymbol{\varsigma}_h, \bar{r}_h) \le \mathcal{L}(\mathbf{v}_h, \bar{r}_h) \qquad \forall (\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h. \tag{6.21}$$

Importantly, the results of Lemmas 6.4.2, 6.4.5, and 6.4.7 carry over to the discrete case, which is straightforward to verify. This gives:

Lemma 6.6.8 (Discrete equivalences)**.** *Definitions 6.6.1 and 6.6.5, Definitions 6.6.2 and 6.6.6, and Definitions 6.6.3 and 6.6.7 are mutually equivalent.*

6.7 Extension to inhomogeneous boundary conditions

We finally present an extension of the model problem (6.1) to general inhomogeneous Dirichlet and Neumann boundary conditions. Thus, in the first case, a Neumann boundary condition will be prescribed on the whole $\partial\omega$, with $\partial\omega_\mathrm{N} := \partial\omega$ and $\partial\omega_\mathrm{D} := \emptyset$. In the second case, a Neumann boundary condition will be imposed on $\partial\omega_\mathrm{N}$ and a Dirichlet boundary condition on $\partial\omega_\mathrm{D}$. Let $L^2_*(\omega)$ still stand for the space of all functions from $L^2(\omega)$ with mean value zero in the first case and $L^2(\omega)$ in the second one. Recall also the spaces $H^1_*(\omega)$ and $\mathbf{H}_*(\mathrm{div}, \omega)$ defined in Section 6.2. Let $\boldsymbol{\tau} \in [L^2(\omega)]^d$. We suppose that $r_\mathrm{D} \in H^1(\omega)$, equal to zero in the first case, is given; $r_\mathrm{D}|_{\partial\omega_\mathrm{D}}$ will represent the (inhomogeneous) Dirichlet boundary condition.

Let also $\boldsymbol{\varsigma}_{\mathrm{N}} \in \mathbf{H}(\mathrm{div},\omega)$, equal to zero when $\partial\omega_{\mathrm{N}} = \emptyset$, be given; $\boldsymbol{\varsigma}_{\mathrm{N}}{\cdot}\mathbf{n}_\omega|_{\partial\omega_{\mathrm{N}}}$ will then represent the (inhomogeneous) Neumann boundary condition. We suppose that $\boldsymbol{\varsigma}_{\mathrm{N}}{\cdot}\mathbf{n}_\omega|_{\partial\omega_{\mathrm{N}}} \in L^2(\partial\omega_{\mathrm{N}})$. Finally, the source term $g \in L^2(\omega)$ is supposed to be such that $(g,1)_\omega = \langle \boldsymbol{\varsigma}_{\mathrm{N}}{\cdot}\mathbf{n}_\omega, 1\rangle_{\partial\omega_{\mathrm{N}}}$. The problem is: find a function $r : \omega \to \mathbb{R}$, with mean value zero in the first case, such that

$$-\nabla{\cdot}(\nabla r + \boldsymbol{\tau}) = g \qquad \text{in } \omega, \tag{6.22a}$$

$$-(\nabla r + \boldsymbol{\tau}){\cdot}\mathbf{n}_\omega = \boldsymbol{\varsigma}_{\mathrm{N}}{\cdot}\mathbf{n}_\omega \qquad \text{on } \partial\omega_{\mathrm{N}}, \tag{6.22b}$$

$$r = r_{\mathrm{D}} \qquad \text{on } \partial\omega_{\mathrm{D}}. \tag{6.22c}$$

Definitions 6.2.1–6.2.3 now respectively take the form:

Definition 6.7.1 (Primal formulation)**.** *Find* $r := r_0 + r_{\mathrm{D}}$ *with* $r_0 \in H^1_*(\omega)$ *such that*

$$(\nabla r_0, \nabla v)_\omega = -(\boldsymbol{\tau}, \nabla v)_\omega + (g,v)_\omega - \langle \boldsymbol{\varsigma}_{\mathrm{N}}{\cdot}\mathbf{n}_\omega, v\rangle_{\partial\omega} - (\nabla r_{\mathrm{D}}, \nabla v)_\omega \qquad \forall v \in H^1_*(\omega). \tag{6.23}$$

Definition 6.7.2 (Dual formulation)**.** *Find* $\boldsymbol{\varsigma} := \boldsymbol{\varsigma}_0 + \boldsymbol{\varsigma}_{\mathrm{N}}$ *with* $\boldsymbol{\varsigma}_0 \in \mathbf{H}_*(\mathrm{div},\omega)$ *and* $\nabla{\cdot}\boldsymbol{\varsigma}_0 = g - \nabla{\cdot}\boldsymbol{\varsigma}_{\mathrm{N}}$ *such that*

$$(\boldsymbol{\varsigma}_0, \mathbf{v})_\omega = -(\boldsymbol{\tau}, \mathbf{v})_\omega - (\boldsymbol{\varsigma}_{\mathrm{N}}, \mathbf{v})_\omega - \langle \mathbf{v}{\cdot}\mathbf{n}_\omega, r_{\mathrm{D}}\rangle_{\partial\omega} \qquad \forall \mathbf{v} \in \mathbf{H}_*(\mathrm{div},\omega) \text{ with } \nabla{\cdot}\mathbf{v} = 0. \tag{6.24}$$

Definition 6.7.3 (Dual mixed formulation)**.** *Find a couple* $(\boldsymbol{\varsigma}, r) := (\boldsymbol{\varsigma}_0 + \boldsymbol{\varsigma}_{\mathrm{N}}, r_0 + r_{\mathrm{D}})$ *with* $(\boldsymbol{\varsigma}_0, r_0) \in \mathbf{H}_*(\mathrm{div},\omega) \times L^2_*(\omega)$ *such that*

$$(\boldsymbol{\varsigma}_0, \mathbf{v})_\omega - (r_0, \nabla{\cdot}\mathbf{v})_\omega = -(\boldsymbol{\tau}, \mathbf{v})_\omega - (\boldsymbol{\varsigma}_{\mathrm{N}}, \mathbf{v})_\omega - \langle \mathbf{v}{\cdot}\mathbf{n}_\omega, r_{\mathrm{D}}\rangle_{\partial\omega} + (r_{\mathrm{D}}, \nabla{\cdot}\mathbf{v})_\omega \qquad \forall \mathbf{v} \in \mathbf{H}_*(\mathrm{div},\omega), \tag{6.25a}$$

$$(\nabla{\cdot}\boldsymbol{\varsigma}_0, q)_\omega = (g,q)_\omega - (\nabla{\cdot}\boldsymbol{\varsigma}_{\mathrm{N}}, q)_\omega \qquad \forall q \in L^2_*(\omega). \tag{6.25b}$$

Using the Green theorem, see Theorem 4.2.4, $\langle \boldsymbol{\varsigma}_{\mathrm{N}}{\cdot}\mathbf{n}_\omega, v\rangle_{\partial\omega} = (\boldsymbol{\varsigma}_{\mathrm{N}}, \nabla v)_\omega + (\nabla{\cdot}\boldsymbol{\varsigma}_{\mathrm{N}}, v)_\omega$. Set $\tilde{\boldsymbol{\tau}} := \boldsymbol{\tau} + \boldsymbol{\varsigma}_{\mathrm{N}} + \nabla r_{\mathrm{D}}$ and $\tilde{g} := g - \nabla{\cdot}\boldsymbol{\varsigma}_{\mathrm{N}}$ and note that $(\tilde{g},1)_\omega = 0$. Then Definition 6.7.1 for r_0 takes absolutely the same form as Definition 6.2.1, just with $\tilde{\boldsymbol{\tau}}$ in place of $\boldsymbol{\tau}$ and $\tilde{g}$ in place of g. Similarly, the Green theorem $\langle \mathbf{v}{\cdot}\mathbf{n}_\omega, r_{\mathrm{D}}\rangle_{\partial\omega} = (\mathbf{v}, \nabla r_{\mathrm{D}})_\omega + (\nabla{\cdot}\mathbf{v}, r_{\mathrm{D}})_\omega$ and the assumption $\nabla{\cdot}\mathbf{v} = 0$ show that Definition 6.7.2 for $\boldsymbol{\varsigma}_0$ takes the form of Definition 6.2.2 with $\tilde{\boldsymbol{\tau}}$ and $\tilde{g}$ in place of $\boldsymbol{\tau}$ and g. Finally, there is this same link between Definitions 6.7.3 and 6.2.3. Thus, the existence, uniqueness, and mutual relations results of Theorem 6.3.1, for the homogeneous solutions $\boldsymbol{\varsigma}_0$ and r_0, still hold true. We in particular find $\boldsymbol{\varsigma}_0 = -\nabla r_0 - \tilde{\boldsymbol{\tau}}$, which in turn leads to $\boldsymbol{\varsigma} = -\nabla r - \boldsymbol{\tau}$. Thus, taking into account inhomogeneous boundary conditions can be transformed into considering Definitions 6.2.1–6.2.3 with modified right-hand sides. One particularly important consequence is that all the results presented in Sections 6.4–6.6 carry appropriately over.

Chapter 7

The Laplace equation in multiple space dimensions

Let us recall the Laplace equation (1.1a)–(1.1b): for $f \in L^2(\Omega)$, find $u : \Omega \to \mathbb{R}$ such that

$$-\Delta u = f \qquad \text{in } \Omega, \tag{7.1a}$$
$$u = 0 \qquad \text{on } \partial\Omega. \tag{7.1b}$$

As in Chapter 2, (7.1a)–(7.1b) does not have a classical solution (i.e., $u \in C^2(\overline{\Omega})$) in general. We are thus again led to the variational formulation.

7.1 Variational formulation

In order to define u, we set:

Definition 7.1.1 (Variational formulation of (7.1a)–(7.1b))**.** *Find $u \in H^1_0(\Omega)$ such that*

$$(\nabla u, \nabla v) = (f, v) \qquad \forall v \in H^1_0(\Omega). \tag{7.2}$$

The existence and uniqueness of a solution of (7.2) is ensured by the Riesz representation theorem (or by the Lax–Milgram theorem). The equivalent of Definition 2.2.2 is as follows:

Definition 7.1.2 (Flux)**.** *Let u be the solution of (7.2). Set*

$$\boldsymbol{\sigma} := -\nabla u. \tag{7.3}$$

We will call $\boldsymbol{\sigma}$ the flux*.*

In analogy with Chapter 2 and using the definitions of the spaces $H^1_0(\Omega)$ and $\mathbf{H}(\mathrm{div}, \Omega)$ from Chapter 4, we have:

Theorem 7.1.3 (Properties of the weak solution of (7.1a)–(7.1b))**.** *Let u be the solution of (7.2). Let $\boldsymbol{\sigma}$ be given by (7.3). Then*

$$u \in H^1_0(\Omega), \quad \boldsymbol{\sigma} \in \mathbf{H}(\mathrm{div}, \Omega), \quad \nabla{\cdot}\boldsymbol{\sigma} = f.$$

Proof. The weak solution u belongs to $H^1_0(\Omega)$ by definition. In order to verify that $\boldsymbol{\sigma} \in \mathbf{H}(\mathrm{div}, \Omega)$, we need to check the three conditions of Definition 4.2.1. Condition 1 is obvious, as $u \in H^1_0(\Omega)$ and thus $-\nabla u = \boldsymbol{\sigma}$ is square-integrable. For the function w of condition 2a, choose $w := f$ and note that $f \in L^2(\Omega)$ by assumption. Then condition 2b follows immediately from (7.2) and the fact that $\mathcal{D}(\Omega) \subset H^1_0(\Omega)$. Thus $\nabla{\cdot}\boldsymbol{\sigma} = f$. □

7.2 Approximate solution

In order to make the presentation general, not restricted to any particular numerical method, we are led to suppose here that the approximate solution u_h that we are given satisfies

$$u_h \in H^1(\mathcal{T}_h), \tag{7.4}$$

where $H^1(\mathcal{T}_h)$ is the broken Sobolev space of Definition 4.3.1. Examples of u_h given by various numerical methods are given below in Section 7.13.

In analogy with Definition 2.3.2, we set:

Definition 7.2.1 (Approximate flux)**.** *Let u_h be the approximate solution, cf. (7.4). We will call*

$$-\nabla u_h \tag{7.5}$$

the approximate flux.

The following remark should be compared to Theorem 7.1.3 and to Remark 2.3.3:

Remark 7.2.2 (Properties of the approximate solution u_h of (7.4))**.** *Let u_h be the approximate solution, cf. (7.4). Then*

$$u_h \notin H^1_0(\Omega), \quad -\nabla u_h \notin \mathbf{H}(\mathrm{div},\Omega), \quad \nabla\cdot(-\nabla u_h) \neq f \quad \text{in general.} \tag{7.6}$$

7.3 Energy (semi-)norm and its dual characterization

As in Chapter 2, we will measure the distance between u and u_h in the energy (semi-)norm. This is the norm induced by the scalar product in (7.2):

$$\|\nabla v\|, \qquad v \in H^1_0(\Omega). \tag{7.7}$$

In a complete analogy to Theorem 2.4.1, with the same proof, this norm can be characterized as a dual norm:

Theorem 7.3.1 (Energy norm for (7.1a)–(7.1b) as a dual norm)**.** *Let $v \in H^1_0(\Omega)$. Then*

$$\|\nabla v\| = \sup_{\varphi\in H^1_0(\Omega);\, \|\nabla\varphi\|=1} (\nabla v, \nabla\varphi). \tag{7.8}$$

7.4 Error characterization

Our fundamental result for a posteriori error estimation is the following characterization of the error. It reveals that $\|\nabla(u-u_h)\|$ is given by the *distance* of u_h to the *correct space* for the *primal variable* (potential) u, which is $H^1_0(\Omega)$, plus the *distance* of ∇u_h to the *correct space* for the *dual variable* (flux) $-\nabla u$, which is $\mathbf{H}(\mathrm{div},\Omega)$, subject to a divergence constraint (recall at this occasion Definitions 4.1.4 and 4.2.2).

Theorem 7.4.1 (Error characterization)**.** *Let $u \in H^1_0(\Omega)$ be the weak solution given by Definition 7.1.1 and let $u_h \in H^1(\mathcal{T}_h)$ be arbitrary. Then*

$$\|\nabla(u-u_h)\|^2 = \min_{\substack{\mathbf{v}\in\mathbf{H}(\mathrm{div},\Omega)\\ \nabla\cdot\mathbf{v}=f}} \|\nabla u_h + \mathbf{v}\|^2 + \min_{v\in H^1_0(\Omega)} \|\nabla(u_h - v)\|^2. \tag{7.9}$$

Proof. Let us define a function $s \in H_0^1(\Omega)$ by

$$(\nabla s, \nabla v) = (\nabla u_h, \nabla v) \qquad \forall v \in H_0^1(\Omega). \tag{7.10}$$

There exists one and only one s, which follows from the Riesz representation theorem. Indeed, firstly, $H_0^1(\Omega)$ is a Hilbert space for the scalar product $(\nabla\cdot, \nabla\cdot)$. Secondly, $|(\nabla u_h, \nabla v)| \le \|\nabla u_h\|\|\nabla v\|$ by the Cauchy–Schwarz inequality, so that the right-hand side of (7.10) represents a continuous linear form on $H_0^1(\Omega)$ in view of our assumption (7.4). The function s represents the orthogonal projection of the approximate solution u_h onto the space $H_0^1(\Omega)$ with respect to the scalar product $(\nabla\cdot, \nabla\cdot)$. With the aid of s, we can write the Pythagorean equality

$$\|\nabla(u - u_h)\|^2 = \|\nabla(u - s)\|^2 + \|\nabla(s - u_h)\|^2. \tag{7.11}$$

To prove this shortly, we develop

$$\|\nabla(u - u_h)\|^2 = \|\nabla(u - s + s - u_h)\|^2 = \|\nabla(u - s)\|^2 + \|\nabla(s - u_h)\|^2 + 2(\nabla(u - s), \nabla(s - u_h)),$$

and the last term in the above expression vanishes in view of the orthogonality (7.10), since $u - s$ can be taken as a test function $v \in H_0^1(\Omega)$ in (7.10). We now continue in two steps.

1) Since s is a projection of u_h,

$$\|\nabla(s - u_h)\|^2 = \min_{v \in H_0^1(\Omega)} \|\nabla(v - u_h)\|^2. \tag{7.12}$$

This can again be proven directly from (7.11) used for any function $v \in H_0^1(\Omega)$ in place of $u \in H_0^1(\Omega)$,

$$\|\nabla(v - u_h)\|^2 = \|\nabla(v - s)\|^2 + \|\nabla(s - u_h)\|^2,$$

from where we get

$$\|\nabla(s - u_h)\|^2 = \|\nabla(v - u_h)\|^2 - \|\nabla(v - s)\|^2 \le \|\nabla(v - u_h)\|^2 \qquad \forall v \in H_0^1(\Omega).$$

This handles the second term in (7.11) in the form needed in (7.9).

2) As for the first term in (7.11), we first notice that $u - s \in H_0^1(\Omega)$. Thus (7.8) and (7.10) give

$$\|\nabla(u - s)\| = \sup_{\varphi \in H_0^1(\Omega);\, \|\nabla\varphi\| = 1} (\nabla(u - s), \nabla\varphi) = \sup_{\varphi \in H_0^1(\Omega);\, \|\nabla\varphi\| = 1} (\nabla(u - u_h), \nabla\varphi). \tag{7.13}$$

Let now $\varphi \in H_0^1(\Omega)$ with $\|\nabla\varphi\| = 1$ be fixed. Using the characterization (7.2) of the weak solution, we have

$$(\nabla(u - u_h), \nabla\varphi) = (f, \varphi) - (\nabla u_h, \nabla\varphi). \tag{7.14}$$

Finally, for an arbitrary $\mathbf{v} \in \mathbf{H}(\mathrm{div}, \Omega)$ such that $\nabla\cdot\mathbf{v} = f$, the Green theorem gives

$$(f, \varphi) - (\nabla u_h, \nabla\varphi) = (\nabla\cdot\mathbf{v}, \varphi) - (\nabla u_h, \nabla\varphi) = -(\nabla u_h + \mathbf{v}, \nabla\varphi).$$

Consequently, by the Cauchy–Schwarz inequality,

$$\|\nabla(u - s)\| \le \min_{\substack{\mathbf{v} \in \mathbf{H}(\mathrm{div}, \Omega) \\ \nabla\cdot\mathbf{v} = f}} \|\nabla u_h + \mathbf{v}\|. \tag{7.15}$$

In the rest of the proof, we show that actually

$$\|\nabla(u - s)\| = \min_{\substack{\mathbf{v} \in \mathbf{H}(\mathrm{div}, \Omega) \\ \nabla\cdot\mathbf{v} = f}} \|\nabla u_h + \mathbf{v}\|, \tag{7.16}$$

which handles the first term in (7.11) in the form needed in (7.9).

The argument of the minimum in (7.15) is

$$\boldsymbol{\sigma} := \arg \min_{\substack{\mathbf{v}\in\mathbf{H}(\mathrm{div},\Omega)\\ \nabla\cdot\mathbf{v}=f}} \|\nabla u_h + \mathbf{v}\|$$

and is characterized by the Euler–Lagrange conditions as a function $\boldsymbol{\sigma} \in \mathbf{H}(\mathrm{div},\Omega)$ with $\nabla\cdot\boldsymbol{\sigma} = f$ such that

$$(\boldsymbol{\sigma},\mathbf{v}) = -(\nabla u_h,\mathbf{v}) \qquad \forall \mathbf{v} \in \mathbf{H}(\mathrm{div},\Omega) \text{ with } \nabla\cdot\mathbf{v} = 0.$$

This problem is in turn equivalent to finding $\boldsymbol{\sigma} \in \mathbf{H}(\mathrm{div},\Omega)$ and $r \in L^2(\Omega)$ such that

$$(\boldsymbol{\sigma},\mathbf{v}) - (r,\nabla\cdot\mathbf{v}) = -(\nabla u_h,\mathbf{v}) \qquad \forall \mathbf{v} \in \mathbf{H}(\mathrm{div},\Omega), \tag{7.17}$$

$$(\nabla\cdot\boldsymbol{\sigma},q) = (f,q) \qquad \forall q \in L^2(\Omega). \tag{7.18}$$

Now, (7.17) implies, see Theorem 6.3.1, that $r \in H^1_0(\Omega)$ with $\nabla r = -\boldsymbol{\sigma} - \nabla u_h$. Consequently, by (7.8) and the Green theorem,

$$\begin{aligned}\min_{\substack{\mathbf{v}\in\mathbf{H}(\mathrm{div},\Omega)\\ \nabla\cdot\mathbf{v}=f}} \|\nabla u_h + \mathbf{v}\| = \|\nabla u_h + \boldsymbol{\sigma}\| = \|\nabla r\| &= \sup_{\varphi\in H^1_0(\Omega);\,\|\nabla\varphi\|=1} (\nabla r,\nabla\varphi)\\ &= \sup_{\varphi\in H^1_0(\Omega);\,\|\nabla\varphi\|=1} (-\boldsymbol{\sigma} - \nabla u_h,\nabla\varphi) = \sup_{\varphi\in H^1_0(\Omega);\,\|\nabla\varphi\|=1} \{(f,\varphi) - (\nabla u_h,\nabla\varphi)\},\end{aligned} \tag{7.19}$$

and (7.13)–(7.14) show that (7.16) holds true. Thus (7.11), (7.12), and (7.16) imply the claim (7.9). □

Remark 7.4.2 (A posteriori error estimate by the error characterization of Theorem 7.4.1)**.** *Under the assumptions of Theorem 7.4.1, it follows from* (7.9) *that*

$$\|\nabla(u - u_h)\|^2 \le \|\nabla u_h + \boldsymbol{\sigma}_h\|^2 + \|\nabla(u_h - s_h)\|^2$$

for an arbitrary $\boldsymbol{\sigma}_h \in \mathbf{H}(\mathrm{div},\Omega)$ *with* $\nabla\cdot\boldsymbol{\sigma}_h = f$ *and an arbitrary* $s_h \in H^1_0(\Omega)$*. This is an estimate on the error* $\|\nabla(u - u_h)\|$ *and has been at the origin of a posteriori analysis, when* $u_h \in H^1_0(\Omega)$*, from the fundamental works of Mikhlin [74] or Ladevèze [70]. Imposing* $\nabla\cdot\boldsymbol{\sigma}_h = f$ *stands on the other extreme with respect to Theorem 2.6.1, where* $\nabla\cdot\boldsymbol{\sigma}_h$ *was completely unconstrained. It is practically irrelevant to try to obtain a suitable flux* $\boldsymbol{\sigma}_h$ *which satisfies exactly* $\nabla\cdot\boldsymbol{\sigma}_h = f$ *for general* $f \in L^2(\Omega)$*. It is, though, possible to obtain* $\boldsymbol{\sigma}_h$ *in a finite-dimensional subspace* $\mathbf{V}_h$ *of* $\mathbf{H}(\mathrm{div},\Omega)$*, cf. Section 5.3, such that* $\nabla\cdot\boldsymbol{\sigma}_h = \Pi_{Q_h} f$*. Here* $Q_h \subset L^2(\Omega)$ *and* $\Pi_{Q_h} f$ *is the* $L^2(\Omega)$*-orthogonal projection onto* Q_h*, see* (5.2)*. Then the remaining difference between* f *and* $\Pi_{Q_h} f$ *can be treated, giving rise to the so-called data oscillation.*

In the next sections, we will be inspired by the above result. We will in particular show how to construct a suitable $\boldsymbol{\sigma}_h$ such that $\nabla\cdot\boldsymbol{\sigma}_h = \Pi_{Q_h} f$. Importantly, the construction of $\boldsymbol{\sigma}_h$ will be *local*, over patches of mesh elements, in contrast to some initial developments where a costly global solve over the entire domain Ω was necessary, and similarly for s_h.

7.5 Prager–Synge equality

The following result is due to Prager and Synge [81] and is linked to Theorem 7.4.1 when $u_h \in H^1_0(\Omega)$:

Theorem 7.5.1 (Prager–Synge equality)**.** *Let $u \in H^1_0(\Omega)$ be the solution of (7.2) and let $u_h \in H^1_0(\Omega)$ and $\boldsymbol{\sigma}_h \in \mathbf{H}(\mathrm{div},\Omega)$ with $\nabla{\cdot}\boldsymbol{\sigma}_h = f$ be arbitrary. Then*

$$\|\nabla(u-u_h)\|^2 + \|\nabla u_h + \boldsymbol{\sigma}_h\|^2 = \|\nabla u_h + \boldsymbol{\sigma}_h\|^2. \tag{7.20}$$

Proof. Adding and subtracting ∇u, we develop

$$\begin{aligned}\|\nabla u_h + \boldsymbol{\sigma}_h\|^2 &= \|\nabla(u_h - u) + \nabla u + \boldsymbol{\sigma}_h\|^2 \\ &= \|\nabla(u_h - u)\|^2 + \|\nabla u + \boldsymbol{\sigma}_h\|^2 + 2(\nabla(u_h - u), \nabla u + \boldsymbol{\sigma}_h).\end{aligned}$$

Note from Theorem 7.1.3 that $\nabla u \in \mathbf{H}(\mathrm{div},\Omega)$ with $\nabla{\cdot}(\nabla u) = -f$. Thus $(\nabla u + \boldsymbol{\sigma}_h) \in \mathbf{H}(\mathrm{div},\Omega)$ and in particular $\nabla{\cdot}(\nabla u + \boldsymbol{\sigma}_h) = 0$. Thus, using that $u_h - u \in H^1_0(\Omega)$, the Green theorem, see Theorem 4.2.5, gives

$$(\nabla(u_h - u), \nabla u + \boldsymbol{\sigma}_h) = -(\nabla{\cdot}(\nabla u + \boldsymbol{\sigma}_h), u_h - u) = 0,$$

whence the assertion follows. □

7.6 Potential and flux reconstructions

From Theorem 7.1.3 and Remark 7.2.2, we see that the approximate solution (or approximate potential) u_h and the approximate flux $-\nabla u_h$ can be nonphysical. Starting from Theorem 7.4.1 and developing the ideas of Section 2.5, we will introduce their "corrections", a potential reconstruction s_h and a flux reconstruction $\boldsymbol{\sigma}_h$:

Definition 7.6.1 (Potential reconstruction)**.** *Let u_h be the approximate solution, cf. (7.4). We will call the* potential reconstruction *any function s_h constructed from u_h which satisfies*

$$s_h \in H^1_0(\Omega).$$

In order to improve on the non fully satisfactory result of Theorem 2.6.1, see Remarks 2.6.2 and 7.4.2, not only we will impose that the flux reconstruction $\boldsymbol{\sigma}_h$ lies in the correct functional space, but we will also prescribe a condition on its divergence. This is linked to the fact that on the continuous level, $\nabla{\cdot}\boldsymbol{\sigma} = f$, as we have seen in Theorem 7.1.3:

Definition 7.6.2 (Equilibrated flux reconstruction)**.** *We will call the* equilibrated flux reconstruction *any function $\boldsymbol{\sigma}_h$ constructed from u_h which satisfies*

$$\boldsymbol{\sigma}_h \in \mathbf{H}(\mathrm{div},\Omega), \tag{7.21a}$$

$$(\nabla{\cdot}\boldsymbol{\sigma}_h, 1)_K = (f,1)_K \qquad \forall K \in \mathcal{T}_h. \tag{7.21b}$$

Note that (7.21b) is a weak form of the condition $\nabla{\cdot}\boldsymbol{\sigma} = f$; only the mean values of the divergence of $\boldsymbol{\sigma}_h$ need to coincide with the mean values of f on each mesh element.

7.7 Residual and its dual norm

To get a further theoretical insight, it is useful to define a residual of the variational formulation of Definition 7.1.1. Let a function $u_h \in H^1(\mathcal{T}_h)$ be given.

Definition 7.7.1 (Residual)**.** *Let $u_h \in H^1(\mathcal{T}_h)$. Then $\mathcal{R}(u_h) \in H^{-1}(\Omega)$ is defined by*

$$\langle \mathcal{R}(u_h), \varphi \rangle_{H^{-1}(\Omega),H_0^1(\Omega)} := (f,\varphi) - (\nabla u_h, \nabla\varphi) \qquad \varphi \in H_0^1(\Omega).$$

From (7.2), when $u_h \in H_0^1(\Omega)$, we immediately see that the residual is zero if and only if the function u_h equals to the weak solution u. More precisely, the following important theorem holds for such conforming u_h:

Theorem 7.7.2 (Equivalence between the energy and dual residual norms)**.** *Let u be the weak solution given by Definition 7.1.1. Let $u_h \in H_0^1(\Omega)$ be arbitrary. Then*

$$\|\nabla(u-u_h)\| = \sup_{\varphi\in H_0^1(\Omega);\,\|\nabla\varphi\|=1} \{(f,\varphi) - (\nabla u_h,\nabla\varphi)\} = \|\mathcal{R}(u_h)\|_{H^{-1}(\Omega)}. \tag{7.22}$$

Proof. We have already shown this in the proof of Theorem 7.4.1, but let us redo the proof in this simpler setting where $u_h \in H_0^1(\Omega)$ and not merely $u_h \in H^1(\mathcal{T}_h)$. Take $u-u_h$ in place of v in (7.8). Then using (7.2), we see that $(\nabla(u-u_h),\nabla\varphi) = (f,\varphi) - (\nabla u_h,\nabla\varphi)$ for any $\varphi \in H_0^1(\Omega)$. Definition 7.7.1 then finishes the proof. □

The energy error in $u_h \in H_0^1(\Omega)$ is thus *equal* to the dual norm of the residual of a conforming $u_h \in H_0^1(\Omega)$ for the Laplace equation. Moreover, from (7.19) above, we also see:

Theorem 7.7.3 (Dual residual and $\mathbf{H}(\mathrm{div},\Omega)$ distance norms)**.** *Let u be the weak solution given by Definition 7.1.1. Let $u_h \in H^1(\mathcal{T}_h)$ be arbitrary. Then*

$$\|\mathcal{R}(u_h)\|_{H^{-1}(\Omega)} = \sup_{\varphi\in H_0^1(\Omega);\,\|\nabla\varphi\|=1} \{(f,\varphi) - (\nabla u_h,\nabla\varphi)\} = \min_{\substack{\mathbf{v}\in\mathbf{H}(\mathrm{div},\Omega)\\ \nabla\cdot\mathbf{v}=f}} \|\nabla u_h + \mathbf{v}\|.$$

7.8 A general a posteriori error estimate

We now finally give here our a posteriori error estimate on the distance between u, the unknown solution of (7.2), and u_h, the known approximate solution characterized by (7.4). Note that it gives a guaranteed upper bound in the sense of the property **i)** of Section 1.4. Also note that this estimate is not restricted to any particular numerical method. We will show how to apply it to usual discretization methods in Section 7.13 below.

Theorem 7.8.1 (A general a posteriori error estimate for (7.1a)–(7.1b))**.** *Let u be the weak solution given by Definition 7.1.1. Let u_h be an arbitrary function satisfying (7.4). Let s_h be a potential reconstruction in the sense of Definition 7.6.1 and $\boldsymbol{\sigma}_h$ an equilibrated flux reconstruction in the sense of Definition 7.6.2. For any $K \in \mathcal{T}_h$, define the* residual estimator *by*

$$\eta_{\mathrm{R},K} := \frac{h_K}{\pi}\|f - \nabla\cdot\boldsymbol{\sigma}_h\|_K, \tag{7.23a}$$

the flux estimator *by*

$$\eta_{\mathrm{F},K} := \|\nabla u_h + \boldsymbol{\sigma}_h\|_K, \tag{7.23b}$$

and the nonconformity estimator *by*

$$\eta_{\mathrm{NC},K} := \|\nabla(u_h - s_h)\|_K. \tag{7.23c}$$

Then

$$\|\nabla(u - u_h)\|^2 \le \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{F},K} + \eta_{\mathrm{R},K})^2 + \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{NC},K}^2. \tag{7.24}$$

Proof. If $\nabla\cdot\boldsymbol{\sigma}_h = f$ holds, and not merely (7.21b), then the result follows immediately from Theorem 7.4.1, see Remark 7.4.2. Let us thus suppose that merely (7.21b) holds and let us estimate the first term in (7.9).

From Theorem 7.7.3, we know that

$$\min_{\substack{\mathbf{v}\in\mathbf{H}(\mathrm{div},\Omega)\\ \nabla\cdot\mathbf{v}=f}} \|\nabla u_h + \mathbf{v}\| = \sup_{\varphi\in H_0^1(\Omega);\,\|\nabla\varphi\|=1} \{(f,\varphi) - (\nabla u_h, \nabla\varphi)\}.$$

Let $\varphi \in H_0^1(\Omega)$ with $\|\nabla\varphi\| = 1$ be fixed. Adding and subtracting $(\boldsymbol{\sigma}_h, \nabla\varphi)$, where $\boldsymbol{\sigma}_h$ is the equilibrated flux reconstruction in the sense of Definition 7.6.2, and using the Green theorem $(\boldsymbol{\sigma}_h, \nabla\varphi) = -(\nabla\cdot\boldsymbol{\sigma}_h, \varphi)$ (see Theorem 4.2.5), we have

$$(f,\varphi) - (\nabla u_h, \nabla\varphi) = (f - \nabla\cdot\boldsymbol{\sigma}_h, \varphi) - (\nabla u_h + \boldsymbol{\sigma}_h, \nabla\varphi).$$

The Cauchy–Schwarz inequality gives

$$-(\nabla u_h + \boldsymbol{\sigma}_h, \nabla\varphi) \le \sum_{K\in\mathcal{T}_h} \|\nabla u_h + \boldsymbol{\sigma}_h\|_K \|\nabla\varphi\|_K = \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{F},K} \|\nabla\varphi\|_K,$$

whereas the approximate equilibrium property (7.21b), the Poincaré inequality (4.20), and the Cauchy–Schwarz inequality give

$$\begin{aligned}(f - \nabla\cdot\boldsymbol{\sigma}_h, \varphi) &= \sum_{K\in\mathcal{T}_h} (f - \nabla\cdot\boldsymbol{\sigma}_h, \varphi)_K = \sum_{K\in\mathcal{T}_h} (f - \nabla\cdot\boldsymbol{\sigma}_h, \varphi - \varphi_K)_K \\ &\le \sum_{K\in\mathcal{T}_h} \frac{h_K}{\pi} \|f - \nabla\cdot\boldsymbol{\sigma}_h\|_K \|\nabla\varphi\|_K = \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{R},K} \|\nabla\varphi\|_K.\end{aligned} \tag{7.25}$$

Combining the above results while using the Cauchy–Schwarz inequality gives

$$\begin{aligned}\min_{\substack{\mathbf{v}\in\mathbf{H}(\mathrm{div},\Omega)\\ \nabla\cdot\mathbf{v}=f}} \|\nabla u_h + \mathbf{v}\|^2 &\le \left(\sup_{\varphi\in H_0^1(\Omega);\,\|\nabla\varphi\|=1} \left\{ \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{F},K} + \eta_{\mathrm{R},K}) \|\nabla\varphi\|_K \right\} \right)^2 \\ &\le \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{F},K} + \eta_{\mathrm{R},K})^2,\end{aligned}$$

whence the assertion of the theorem follows. □

Remark 7.8.2 (Theorem 7.8.1)**.** *The three estimators of Theorem 7.8.1* $\eta_{\mathrm{NC},K}$, $\eta_{\mathrm{F},K}$, *and* $\eta_{\mathrm{R},K}$ *reflect respectively the three possible violations of physical properties of the approximate solution* u_h *discussed in Remark 7.2.2. Note that whenever* $u_h \in H_0^1(\Omega)$, *we can set the potential reconstruction* s_h *from Definition 7.6.1 equal to* u_h *and the estimator* $\eta_{\mathrm{NC},K}$ *vanishes. Similarly, shall it happen that* $-\nabla u_h \in \mathbf{H}(\mathrm{div},\Omega)$ *and* $(\nabla\cdot(-\nabla u_h), 1)_K = (f,1)_K$ *for all* $K \in \mathcal{T}_h$ *(we shall indeed meet such cases below), we can set the flux reconstruction* $\boldsymbol{\sigma}_h$ *from Definition 7.6.2 equal to* $-\nabla u_h$ *and the estimator* $\eta_{\mathrm{F},K}$ *vanishes. Shall moreover* $\nabla\cdot(-\nabla u_h) = f$, *then* $\eta_{\mathrm{R},K}$ *vanishes as well.*

7.9 Flux reconstruction via local Neumann mixed finite element problems

This section describes a practical way to obtain the equilibrated flux reconstruction introduced in Definition 7.6.2. We rewrite here equivalently the technique of [22] which generalizes [40], proceeding as in [53, 55]. The equilibration goes over patches of elements $\omega_{\mathbf{a}}$ sharing a vertex $\mathbf{a} \in \mathcal{V}_h$. We will employ the dual mixed finite element method introduced in Definition 6.6.3 for this purpose. Let $\mathbf{V}_h \times Q_h$ stand for the mixed finite element spaces introduced in Section 5.3; whenever $u_h \in \mathbb{P}_k(\mathcal{T}_h)$, $\mathbf{RTN}_k \times \mathbb{P}_k(\mathcal{T}_h)$ is the typical choice. Let $\mathbf{V}_h(\omega_{\mathbf{a}}) \times Q_h(\omega_{\mathbf{a}})$ denote their restriction to $\omega_{\mathbf{a}}$. Recall also the definition of the hat function from Section 5.2.

Definition 7.9.1 (Flux $\boldsymbol{\sigma}_h$)**.** *Let $u_h \in H^1(\mathcal{T}_h)$, and let it satisfy the hat-function orthogonality*

$$(\nabla u_h, \nabla \psi_{\mathbf{a}})_{\omega_{\mathbf{a}}} = (f, \psi_{\mathbf{a}})_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}. \tag{7.26}$$

For each $\mathbf{a} \in \mathcal{V}_h$, prescribe $\boldsymbol{\varsigma}_h^{\mathbf{a}} \in \mathbf{V}_h^{\mathbf{a}}$ and $\bar{r}_h^{\mathbf{a}} \in Q_h^{\mathbf{a}}$ by solving

$$(\boldsymbol{\varsigma}_h^{\mathbf{a}}, \mathbf{v}_h)_{\omega_{\mathbf{a}}} - (\bar{r}_h^{\mathbf{a}}, \nabla\cdot\mathbf{v}_h)_{\omega_{\mathbf{a}}} = -(\boldsymbol{\tau}_h^{\mathbf{a}}, \mathbf{v}_h)_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{v}_h \in \mathbf{V}_h^{\mathbf{a}}, \tag{7.27a}$$

$$(\nabla\cdot\boldsymbol{\varsigma}_h^{\mathbf{a}}, q_h)_{\omega_{\mathbf{a}}} = (g^{\mathbf{a}}, q_h)_{\omega_{\mathbf{a}}} \qquad \forall q_h \in Q_h^{\mathbf{a}}, \tag{7.27b}$$

with the spaces

$$\begin{aligned} &\mathbf{V}_h^{\mathbf{a}} := \{\mathbf{v}_h \in \mathbf{V}_h(\omega_{\mathbf{a}});\ \mathbf{v}_h\cdot\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \text{ on } \partial\omega_{\mathbf{a}}\},\\ &Q_h^{\mathbf{a}} := \{q_h \in Q_h(\omega_{\mathbf{a}});\ (q_h, 1)_{\omega_{\mathbf{a}}} = 0\}, \end{aligned} \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}, \tag{7.28a}$$

$$\begin{aligned} &\mathbf{V}_h^{\mathbf{a}} := \{\mathbf{v}_h \in \mathbf{V}_h(\omega_{\mathbf{a}});\ \mathbf{v}_h\cdot\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \text{ on } \partial\omega_{\mathbf{a}} \setminus \partial\Omega\},\\ &Q_h^{\mathbf{a}} := Q_h(\omega_{\mathbf{a}}), \end{aligned} \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}, \tag{7.28b}$$

and the right-hand sides

$$\boldsymbol{\tau}_h^{\mathbf{a}} := \psi_{\mathbf{a}} \nabla u_h, \tag{7.29a}$$

$$g^{\mathbf{a}} := \psi_{\mathbf{a}} f - \nabla\psi_{\mathbf{a}}\cdot\nabla u_h. \tag{7.29b}$$

Then, set

$$\boldsymbol{\sigma}_h := \sum_{\mathbf{a}\in\mathcal{V}_h} \boldsymbol{\varsigma}_h^{\mathbf{a}}. \tag{7.30}$$

In (7.28), a homogeneous Neumann (no-flux) boundary condition on the whole boundary of the patch $\omega_{\mathbf{a}}$ together with mean value zero is imposed for interior vertices, whereas the no-flux condition is only imposed in the interior of Ω for boundary vertices. Also note that by (7.26), $(g^{\mathbf{a}}, 1)_{\omega_{\mathbf{a}}} = 0$ follows for interior vertices $\mathbf{a}$, which is the Neumann compatibility condition. Existence and uniqueness of the solution to (7.27) have been treated in Section 6.6. We now verify the requirements of Definition 7.6.2:

Lemma 7.9.2 (Properties of $\boldsymbol{\sigma}_h$)**.** *Definition 7.9.1 yields a flux reconstruction $\boldsymbol{\sigma}_h \in \mathbf{H}(\mathrm{div}, \Omega)$ such that*

$$(f - \nabla\cdot\boldsymbol{\sigma}_h, v_h)_K = 0 \qquad \forall v_h \in Q_h(K) \qquad \forall K \in \mathcal{T}_h. \tag{7.31}$$

Proof. It is clear that $\boldsymbol{\sigma}_h \in \mathbf{H}(\mathrm{div}, \Omega)$, as all the individual components $\boldsymbol{\varsigma}_h^{\mathbf{a}}$ belong to $\mathbf{H}(\mathrm{div}, \Omega)$ (when extended by zero outside of ω) (the homogeneous Neumann boundary condition of (7.28) is essential here!), and $\boldsymbol{\sigma}_h$ is their sum by (7.30). Let $\mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}$. Then the facts that $\boldsymbol{\varsigma}_h^{\mathbf{a}}\cdot\mathbf{n}_{\omega_{\mathbf{a}}} = 0$

on $\partial\omega_{\mathbf{a}}$ and $(g^{\mathbf{a}},1)_{\omega_{\mathbf{a}}}=0$ enable us to take the constants as test functions in (7.27b): indeed the Green theorem then immediately gives $(\nabla\cdot\boldsymbol{\varsigma}_h^{\mathbf{a}},1)_{\omega_{\mathbf{a}}}=(g^{\mathbf{a}},1)_{\omega_{\mathbf{a}}}=0$. Thus (7.27b) actually holds for all functions from $Q_h(\omega_{\mathbf{a}})$ and not only for functions with mean value zero. Next crucial argument is that the polynomials in $Q_h(\omega_{\mathbf{a}})$ are discontinuous, see Section 5.1. We thus infer that any $q_h\in Q_h(K)$ for any $K\in\mathcal{T}_{\mathbf{a}}$ can be taken as a test function in (7.27b).

We are now ready to prove (7.31). Fix $K\in\mathcal{T}_h$ and $v_h\in Q_h(K)$. Employing that $\boldsymbol{\sigma}_h|_K=\sum_{\mathbf{a}\in\mathcal{V}_K}\boldsymbol{\varsigma}_h^{\mathbf{a}}|_K$ from (7.30) and (7.27b) with (7.29b),

$$(f-\nabla\cdot\boldsymbol{\sigma}_h,v_h)_K=\sum_{\mathbf{a}\in\mathcal{V}_K}(\psi_{\mathbf{a}}f-\nabla\cdot\boldsymbol{\varsigma}_h^{\mathbf{a}},v_h)_K=\sum_{\mathbf{a}\in\mathcal{V}_K}(\nabla\psi_{\mathbf{a}}\cdot\nabla u_h,v_h)_K=0,$$

where

$$\sum_{\mathbf{a}\in\mathcal{V}_K}\psi_{\mathbf{a}}|_K=1|_K, \tag{7.32}$$

the partition of unity by the hat functions, was also used. □

Remark 7.9.3 (Data oscillation)**.** *The orthogonality* (7.31) *together with the mixed finite element spaces property* $\nabla\cdot\mathbf{V}_h(K)=Q_h(K)$ *for any* $K\in\mathcal{T}_h$ *imply that*

$$\frac{h_K}{\pi}\|f-\nabla\cdot\boldsymbol{\sigma}_h\|_K=\frac{h_K}{\pi}\|f-\Pi_{Q_h}f\|_K,$$

i.e., $\eta_{\mathrm{R},K}$ *from* (7.23a) *is actually a so-called data oscillation term (recall from* (5.2) *that* Π_{Q_h} *is the* $L^2(\Omega)$*-orthogonal projection onto* Q_h*). If the energy error* $\|\nabla(u-u_h)\|$ *converges as* $\mathcal{O}(h^k)$*,* $Q_h=\mathbb{P}_k(\mathcal{T}_h)$*, and* f *is elementwise smooth, this term converges as* $\mathcal{O}(h^{k+2})$*, i.e., by two orders faster.*

Remark 7.9.4 (Local flux minimization)**.** *From* (7.24)*, a possible "best" choice for the equilibrated flux reconstruction would be*

$$\boldsymbol{\sigma}_h:=\arg\min_{\mathbf{v}_h\in\mathbf{V}_h,\,\nabla\cdot\mathbf{v}_h=\Pi_{Q_h}f}\|\nabla u_h+\mathbf{v}_h\|.$$

This, however, represents a global minimization and as such is way too expensive, see property **v)** *and the discussion in Section 1.4. Definition 7.9.1 rather relies on the partition of unity by the hat functions* $\psi_{\mathbf{a}}$ *and finds the following* local minimizers*:*

$$\boldsymbol{\varsigma}_h^{\mathbf{a}}:=\arg\min_{\mathbf{v}_h\in\mathbf{V}_h^{\mathbf{a}},\,\nabla\cdot\mathbf{v}_h=\Pi_{Q_h^{\mathbf{a}}}(\psi_{\mathbf{a}}f-\nabla\psi_{\mathbf{a}}\cdot\nabla u_h)}\|\psi_{\mathbf{a}}\nabla u_h+\mathbf{v}_h\|_{\omega_{\mathbf{a}}}\qquad\forall\mathbf{a}\in\mathcal{V}_h. \tag{7.33}$$

Indeed, it is enough to see Lemma 6.6.8 for the equivalence of the discrete dual mixed formulation (7.27) *with the discrete constrained minimization* (7.33)*.*

Remark 7.9.5 (Flux equilibration)**.** *The advantage of Definition 7.9.1 is that, under assumption* (7.26)*, it is completely generic (works for* $u_h\in H^1(\mathcal{T}_h)$*) and chooses the best flux in the sense of* (7.33)*. Also, as we shall see below, it leads to local efficiency with a constant which is independent of the polynomial degree and which can be fully bounded. On the other hand, local linear systems* (7.27) *need to be implemented and solved. For a given numerical method, different cheaper procedures can be devised; typically, in locally conservative methods, the flux reconstruction can be* prescribed element by element*, without any linear system solve. Such approaches are described below in Section 8.3.*

7.10 Potential reconstruction via local Dirichlet finite element problems

We now turn to the potential reconstruction s_h, necessary when $u_h \notin H^1_0(\Omega)$. Let $W_h \subset H^1_0(\Omega)$ be as the space V_h in Section 5.2. A typical case, for $u_h \in \mathbb{P}_k(\mathcal{T}_h)$, is $W_h := \mathbb{P}_{k+1}(\mathcal{T}_h) \cap H^1_0(\Omega)$. In order to obtain s_h following Definition 7.6.1, we proceed as in [55], similarly to [29]:

Definition 7.10.1 (Potential s_h)**.** *Let $u_h \in H^1(\mathcal{T}_h)$. For each $\mathbf{a} \in \mathcal{V}_h$, set $W_h^{\mathbf{a}} := W_h \cap H^1_0(\omega_{\mathbf{a}})$ and define $s_h^{\mathbf{a}} \in W_h^{\mathbf{a}}$ by*

$$(\nabla s_h^{\mathbf{a}}, \nabla v_h)_{\omega_{\mathbf{a}}} = (\nabla(\psi_{\mathbf{a}} u_h), \nabla v_h)_{\omega_{\mathbf{a}}} \qquad \forall v_h \in W_h^{\mathbf{a}}. \tag{7.34}$$

Then set

$$s_h := \sum_{\mathbf{a} \in \mathcal{V}_h} s_h^{\mathbf{a}}. \tag{7.35}$$

The existence and uniqueness of each $s_h^{\mathbf{a}}$ is straightforward from the Riesz representation theorem, see Section 6.6. Moreover, as each $s_h^{\mathbf{a}} \in H^1_0(\Omega)$ (we again tacitly assume extension by zero outside of $\omega_{\mathbf{a}}$), s_h is indeed in $H^1_0(\Omega)$ by (7.35). Similarly to Remark 7.9.4, we can observe here:

Remark 7.10.2 (Local potential minimization)**.** *From (7.24), a possible "best" choice for the potential reconstruction would be*

$$s_h := \arg \min_{v_h \in W_h} \|\nabla(u_h - v_h)\|$$

for some finite-dimensional subspace W_h of $H^1_0(\Omega)$. This represents a global minimization, too expensive in view of property **v)** *of Section 1.4. Lemma 6.6.8 shows that Definition 7.10.1 actually performs the following partition-of-unity-based* local minimization*:*

$$s_h^{\mathbf{a}} := \arg \min_{v_h \in W_h^{\mathbf{a}}} \|\nabla(\psi_{\mathbf{a}} u_h - v_h)\|_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{a} \in \mathcal{V}_h. \tag{7.36}$$

Let $d = 2$ and let $\mathrm{R}_{\frac{\pi}{2}} := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ be the matrix of rotation by $\frac{\pi}{2}$; then $\mathrm{R}_{\frac{\pi}{2}}\nabla$ stands for the weak broken curl, i.e., the rotated weak broken gradient: for $v \in H^1(\mathcal{T}_h), \mathrm{R}_{\frac{\pi}{2}}\nabla v = (-\partial_y v, \partial_x v)$. An extremely interesting link appears between Definitions 7.9.1 and 7.10.1:

Theorem 7.10.3 (Equivalence of primal and mixed minimizations)**.** *Let $d = 2$. For each $\mathbf{a} \in \mathcal{V}_h$, prescribe $\boldsymbol{\varsigma}_h^{\mathbf{a}} \in \mathbf{V}_h^{\mathbf{a}}$ and $\bar{r}_h^{\mathbf{a}} \in Q_h^{\mathbf{a}}$ by solving*

$$(\boldsymbol{\varsigma}_h^{\mathbf{a}}, \mathbf{v}_h)_{\omega_{\mathbf{a}}} - (\bar{r}_h^{\mathbf{a}}, \nabla\cdot\mathbf{v}_h)_{\omega_{\mathbf{a}}} = -(\boldsymbol{\tau}_h^{\mathbf{a}}, \mathbf{v}_h)_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{v}_h \in \mathbf{V}_h^{\mathbf{a}}, \tag{7.37a}$$

$$(\nabla\cdot\boldsymbol{\varsigma}_h^{\mathbf{a}}, q_h)_{\omega_{\mathbf{a}}} = (g^{\mathbf{a}}, q_h)_{\omega_{\mathbf{a}}} \qquad \forall q_h \in Q_h^{\mathbf{a}}, \tag{7.37b}$$

with the spaces

$$\begin{aligned} \mathbf{V}_h^{\mathbf{a}} &:= \{\mathbf{v}_h \in \mathbf{V}_h(\omega_{\mathbf{a}});\ \mathbf{v}_h\cdot\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \text{ on } \partial\omega_{\mathbf{a}}\}, \\ Q_h^{\mathbf{a}} &:= \{q_h \in Q_h(\omega_{\mathbf{a}});\ (q_h, 1)_{\omega_{\mathbf{a}}} = 0\}, \end{aligned} \tag{7.38}$$

and the right-hand sides

$$\boldsymbol{\tau}_h^{\mathbf{a}} := \mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}} u_h), \tag{7.39a}$$

$$g^{\mathbf{a}} := 0. \tag{7.39b}$$

Set

$$-\mathrm{R}_{\frac{\pi}{2}}\nabla s_h^{\mathbf{a}} := \boldsymbol{\varsigma}_h^{\mathbf{a}}, \tag{7.40a}$$

$$s_h^{\mathbf{a}} := 0 \text{ on } \partial\omega_{\mathbf{a}}. \tag{7.40b}$$

Let W_h be given by Q_h increased by one order and intersected with $H_0^1(\Omega)$. Then $s_h^{\mathbf{a}}$ from (7.34) and (7.40) coincide.

Proof. Note that $\boldsymbol{\varsigma}_h^{\mathbf{a}}\cdot\mathbf{n}_{\omega_{\mathbf{a}}} = 0$ on $\partial\omega_{\mathbf{a}}$ and (7.39b) enable us to take all test functions from $Q_h(\omega_{\mathbf{a}})$ in (7.37b), so that $\nabla\cdot\boldsymbol{\varsigma}_h^{\mathbf{a}} = 0$. Then we conclude from Lemma 6.6.8 that (7.37) is equivalent to

$$\boldsymbol{\varsigma}_h^{\mathbf{a}} := \arg\min_{\mathbf{v}_h\in\mathbf{V}_h^{\mathbf{a}},\,\nabla\cdot\mathbf{v}_h=0} \left\|\mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_h) + \mathbf{v}_h\right\|_{\omega_{\mathbf{a}}} \qquad \forall\mathbf{a}\in\mathcal{V}_h. \tag{7.41}$$

Now, divergence-free functions from $\mathbf{V}_h = \mathbf{RTN}_k$, see (5.7), are rotated gradients of polynomials from $\mathbb{P}_{k+1}$, see [24, Corollary 3.2]. In particular, the continuity of the normal trace of $\boldsymbol{\varsigma}_h^{\mathbf{a}}$ over the interior edges of $\mathcal{T}_{\mathbf{a}}$ implies the continuity of the tangential trace of $\nabla s_h^{\mathbf{a}}$ over the interior edges of $\mathcal{T}_{\mathbf{a}}$, and similarly, the normal trace of $\boldsymbol{\varsigma}_h^{\mathbf{a}}$ being zero on $\partial\omega_{\mathbf{a}}$, $s_h^{\mathbf{a}}$ is constant on $\partial\omega_{\mathbf{a}}$ and we can fix it to zero on $\partial\omega_{\mathbf{a}}$ by (7.40b). Thus, (7.41) together with (7.40a)–(7.40b) gives

$$s_h^{\mathbf{a}} = \arg\min_{v_h\in W_h^{\mathbf{a}}} \left\|\mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_h) - \mathrm{R}_{\frac{\pi}{2}}\nabla v_h\right\|_{\omega_{\mathbf{a}}} \qquad \forall\mathbf{a}\in\mathcal{V}_h.$$

As $\left\|\mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_h - v_h)\right\|_{\omega_{\mathbf{a}}} = \|\nabla(\psi_{\mathbf{a}}u_h - v_h)\|_{\omega_{\mathbf{a}}}$, (7.36) follows. □

A remarkable fact is that the local mixed finite element problem (7.37) is the same as (7.27); only the spaces $\mathbf{V}_h^{\mathbf{a}}$ and $Q_h^{\mathbf{a}}$ differ for boundary vertices, and the right-hand sides $\boldsymbol{\tau}_h^{\mathbf{a}}$ and $g^{\mathbf{a}}$ differ for all vertices. Thus, we can henceforth only study such dual mixed formulations.

Remark 7.10.4 (Alternative potential reconstruction)**.** *Let $d = 2$. Then an alternative potential reconstruction, close to that of [29, Section 6.3] is possible for $u_h \in H^1(\mathcal{T}_h)$ under the assumption*

$$(\nabla u_h, \mathrm{R}_{\frac{\pi}{2}}\nabla\psi_{\mathbf{a}})_{\omega_{\mathbf{a}}} = 0 \qquad \forall\mathbf{a}\in\mathcal{V}_h. \tag{7.42}$$

Set

$$\boldsymbol{\tau}_h^{\mathbf{a}} := \psi_{\mathbf{a}}\mathrm{R}_{\frac{\pi}{2}}\nabla u_h, \tag{7.43a}$$

$$g^{\mathbf{a}} := (\mathrm{R}_{\frac{\pi}{2}}\nabla\psi_{\mathbf{a}})\cdot\nabla u_h, \tag{7.43b}$$

and use (7.37)–(7.38) together with

$$\boldsymbol{\varsigma}_h := \sum_{\mathbf{a}\in\mathcal{V}_h} \boldsymbol{\varsigma}_h^{\mathbf{a}}.$$

This yields $\boldsymbol{\varsigma}_h \in \mathbf{V}_h$ such that $\boldsymbol{\varsigma}_h\cdot\mathbf{n}_\Omega = 0$ on $\partial\Omega$. Moreover, proceeding as in Lemma 7.9.2, one readily checks that $\nabla\cdot\boldsymbol{\varsigma}_h = 0$. Thus, there exists a piecewise polynomial s_h in $H_0^1(\Omega)$ such that

$$-\mathrm{R}_{\frac{\pi}{2}}\nabla s_h = \boldsymbol{\varsigma}_h.$$

The advantage of the choice (7.39) is that condition (7.42) is not needed and that it is linked to Definition 7.10.1 via Theorem 7.10.3. The advantage of the choice (7.43) is that the local efficiency will be proven with a simpler constant; see Remark 7.11.3 below.

Remark 7.10.5 (Potential reconstruction)**.** *Similar observations as in Remark 7.9.5 hold here as well: Definition 7.10.1 is completely generic, chooses the best reconstruction in the sense (7.36), and leads to polynomial degree robustness and guaranteed maximal overestimation. A cheaper generic construction, see Definition 8.3.1 in Section 8.3 below, is an alternative.*

7.11 Polynomial-degree-robust local efficiency

We will investigate in this section the efficiency of the a posteriori error estimate of Theorem 7.8.1 with the reconstructions of Definitions 7.9.1 and 7.10.1. We will show that for shape-regular meshes, see Section 3.1, they also give a (local) lower bound for the error $\|\nabla(u-u_h)\|$, up to a generic constant only depending on the shape-regularity parameter $\kappa_{\mathcal{T}}$.

We proceed in three steps. In Section 7.11.1, we introduce primal continuous problems on patches of mesh elements which are such that the energy norms of their solutions represent lower bounds of the error in the patches. In Section 7.11.2, we show that the local constructions of Sections 7.9 and 7.10 represent lower bounds, up to a polynomial-degree-independent constant, of the energy norms of the continuous solutions from Section 7.11.1. Finally, in Section 7.11.3, elementwise local lower bounds for the actual estimators are derived from the results of Section 7.11.1 and Section 7.11.2.

7.11.1 Continuous-level problems with hat functions on patches

The following important result has been first shown in [27, 21]:

Lemma 7.11.1 (Continuous efficiency, flux reconstruction)**.** *Let u be the weak solution of (7.2) and let $u_h \in H^1(\mathcal{T}_h)$ be arbitrary. Let $\mathbf{a} \in \mathcal{V}_h$ and let $r_{\mathbf{a}} \in H^1_*(\omega_{\mathbf{a}})$ solve*

$$(\nabla r_{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}} = -(\boldsymbol{\tau}_h^{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}} + (g^{\mathbf{a}}, v)_{\omega_{\mathbf{a}}} \qquad \forall v \in H^1_*(\omega_{\mathbf{a}}) \tag{7.44}$$

with the space

$$H^1_*(\omega_{\mathbf{a}}) := \{v \in H^1(\omega_{\mathbf{a}});\ (v,1)_{\omega_{\mathbf{a}}} = 0\}, \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}, \tag{7.45a}$$

$$H^1_*(\omega_{\mathbf{a}}) := \{v \in H^1(\omega_{\mathbf{a}});\ v = 0 \text{ on } \partial\omega_{\mathbf{a}} \cap \partial\Omega\}, \quad \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}, \tag{7.45b}$$

and the right-hand sides $\boldsymbol{\tau}_h^{\mathbf{a}}$ and $g^{\mathbf{a}}$ from (7.29). Then there exists a constant $C_{\mathrm{cont,PF}} > 0$ only depending on the shape-regularity parameter $\kappa_{\mathcal{T}}$ such that

$$\|\nabla r_{\mathbf{a}}\|_{\omega_{\mathbf{a}}} \le C_{\mathrm{cont,PF}} \|\nabla(u-u_h)\|_{\omega_{\mathbf{a}}}. \tag{7.46}$$

Proof. There holds

$$\|\nabla r_{\mathbf{a}}\|_{\omega_{\mathbf{a}}} = \sup_{v \in H^1_*(\omega_{\mathbf{a}});\, \|\nabla v\|_{\omega_{\mathbf{a}}}=1} (\nabla r_{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}}, \tag{7.47}$$

cf. Theorem 2.4.1. Fix $v \in H^1_*(\omega_{\mathbf{a}})$ with $\|\nabla v\|_{\omega_{\mathbf{a}}} = 1$. Definitions (7.44) and (7.29), the fact that $\psi_{\mathbf{a}} v \in H^1_0(\omega_{\mathbf{a}})$, the characterization (7.2) of the weak solution, and the Cauchy–Schwarz inequality imply

$$\begin{aligned}(\nabla r_{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}} &= -(\psi_{\mathbf{a}} \nabla u_h, \nabla v)_{\omega_{\mathbf{a}}} + (\psi_{\mathbf{a}} f - \nabla\psi_{\mathbf{a}} \cdot \nabla u_h, v)_{\omega_{\mathbf{a}}} \\ &= (f, \psi_{\mathbf{a}} v)_{\omega_{\mathbf{a}}} - (\nabla u_h, \nabla(\psi_{\mathbf{a}} v))_{\omega_{\mathbf{a}}} \\ &= (\nabla(u-u_h), \nabla(\psi_{\mathbf{a}} v))_{\omega_{\mathbf{a}}} \\ &\le \|\nabla(u-u_h)\|_{\omega_{\mathbf{a}}} \|\nabla(\psi_{\mathbf{a}} v)\|_{\omega_{\mathbf{a}}}.\end{aligned} \tag{7.48}$$

Next,

$$\begin{aligned}\|\nabla(\psi_{\mathbf{a}} v)\|_{\omega_{\mathbf{a}}} &= \|\nabla\psi_{\mathbf{a}} v + \psi_{\mathbf{a}} \nabla v\|_{\omega_{\mathbf{a}}} \\ &\le \|\nabla\psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}} \|v\|_{\omega_{\mathbf{a}}} + \|\psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}} \|\nabla v\|_{\omega_{\mathbf{a}}} \\ &\le 1 + C_{\mathrm{PF},\omega_{\mathbf{a}}} h_{\omega_{\mathbf{a}}} \|\nabla\psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}},\end{aligned} \tag{7.49}$$

employing $\|\nabla v\|_{\omega_{\mathbf{a}}} = 1$, $\|\psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}} = 1$, the Poincaré inequality (4.20) on the patch $\omega_{\mathbf{a}}$ for $\mathbf{a} \in \mathcal{V}_h^{\rm int}$, the Friedrichs inequality (4.21) on the patch $\omega_{\mathbf{a}}$ for $\mathbf{a} \in \mathcal{V}_h^{\rm ext}$, and setting $C_{{\rm PF},\omega_{\mathbf{a}}} := C_{{\rm P},\omega_{\mathbf{a}}}$ if $\mathbf{a} \in \mathcal{V}_h^{\rm int}$ and $C_{{\rm PF},\omega_{\mathbf{a}}} := C_{{\rm F},\omega_{\mathbf{a}}}$ if $\mathbf{a} \in \mathcal{V}_h^{\rm ext}$. Thus (7.46) follows with $C_{\rm cont,PF} := \max_{\mathbf{a}\in\mathcal{V}_h}\{1 + C_{{\rm PF},\omega_{\mathbf{a}}} h_{\omega_{\mathbf{a}}} \|\nabla\psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}}\}$. □

We now prove a similar estimate for the potential case. The proof hinges on the additional assumption of the continuity in mean of the jumps of the approximate solution u_h (note that this assumption implies (7.42)). We refer to Section 7.13.3 for a more refined analysis when this assumption is not met.

Lemma 7.11.2 (Continuous efficiency, potential reconstruction)**.** *Let $d = 2$. Let u be the weak solution of (7.2) and let $u_h \in H^1(\mathcal{T}_h)$ satisfy*

$$\langle [\![u_h]\!], 1\rangle_e = 0 \qquad \forall e \in \mathcal{E}_h. \tag{7.50}$$

Let $\mathbf{a} \in \mathcal{V}_h$ and let $r_{\mathbf{a}} \in H^1_*(\omega_{\mathbf{a}})$ solve

$$(\nabla r_{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}} = -(\boldsymbol{\tau}_h^{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}} + (g^{\mathbf{a}}, v)_{\omega_{\mathbf{a}}} \qquad \forall v \in H^1_*(\omega_{\mathbf{a}}) \tag{7.51}$$

with the space

$$H^1_*(\omega_{\mathbf{a}}) := \{v \in H^1(\omega_{\mathbf{a}});\ (v,1)_{\omega_{\mathbf{a}}} = 0\} \tag{7.52}$$

and the right-hand sides $\boldsymbol{\tau}_h^{\mathbf{a}}$ and $g^{\mathbf{a}}$ from (7.39). Then there exists a constant $C_{\rm cont,bPF} > 0$ only depending on the shape-regularity parameter $\kappa_{\mathcal{T}}$ such that

$$\|\nabla r_{\mathbf{a}}\|_{\omega_{\mathbf{a}}} \le C_{\rm cont,bPF} \|\nabla(u - u_h)\|_{\omega_{\mathbf{a}}}. \tag{7.53}$$

Proof. We start again from (7.47) and fix $v \in H^1_*(\omega_{\mathbf{a}})$ with $\|\nabla v\|_{\omega_{\mathbf{a}}} = 1$. For an arbitrary $\tilde{u} \in H^1(\omega_{\mathbf{a}})$ such that $(\tilde{u},1)_{\omega_{\mathbf{a}}} = (u_h,1)_{\omega_{\mathbf{a}}}$ if $\mathbf{a} \in \mathcal{V}_h^{\rm int}$ and $\tilde{u} = 0$ on $\partial\omega_{\mathbf{a}} \cap \partial\Omega$ if $\mathbf{a} \in \mathcal{V}_h^{\rm ext}$, we observe that

$$(\mathrm{R}_{\frac{\pi}{2}} \nabla(\psi_{\mathbf{a}} \tilde{u}), \nabla v)_{\omega_{\mathbf{a}}} = 0.$$

This follows easily by the Green Theorem 4.2.4. Thus, using (7.51) with (7.39) and the Cauchy–Schwarz inequality, we arrive at

$$\begin{aligned}(\nabla r_{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}} &= -(\mathrm{R}_{\frac{\pi}{2}} \nabla(\psi_{\mathbf{a}} u_h), \nabla v)_{\omega_{\mathbf{a}}} = (\mathrm{R}_{\frac{\pi}{2}} \nabla(\psi_{\mathbf{a}}(\tilde{u} - u_h)), \nabla v)_{\omega_{\mathbf{a}}} \\ &\le \big\|\mathrm{R}_{\frac{\pi}{2}} \nabla(\psi_{\mathbf{a}}(\tilde{u} - u_h))\big\|_{\omega_{\mathbf{a}}} \|\nabla v\|_{\omega_{\mathbf{a}}} = \|\nabla(\psi_{\mathbf{a}}(\tilde{u} - u_h))\|_{\omega_{\mathbf{a}}}.\end{aligned}$$

We next intend to proceed as in (7.49), with $\tilde{u} - u_h$ in place of v. The difference is that now $\tilde{u} - u_h$ does not belong to $H^1(\omega_{\mathbf{a}})$, with zero trace on $\partial\omega_{\mathbf{a}} \cap \partial\Omega$ for $\mathbf{a} \in \mathcal{V}_h^{\rm ext}$, but is a piecewise H^1 function from $H^1(\mathcal{T}_{\mathbf{a}})$. There is, fortunately, the continuity in mean of the jumps owing to assumption (7.50), and in particular $\langle \tilde{u} - u_h, 1\rangle_e = 0$ for all edges e located in $\partial\omega_{\mathbf{a}} \cap \partial\Omega$ for $\mathbf{a} \in \mathcal{V}_h^{\rm ext}$, as well as $(\tilde{u} - u_h, 1)_{\omega_{\mathbf{a}}} = 0$ for $\mathbf{a} \in \mathcal{V}_h^{\rm int}$. Thus the Poincaré inequality (4.20) and the Friedrichs inequality (4.21) have to be replaced by their broken versions (4.23) and (4.24) respectively, leading to

$$\|\nabla(\psi_{\mathbf{a}}(\tilde{u} - u_h))\|_{\omega_{\mathbf{a}}} \le (1 + C_{{\rm bPF},\omega_{\mathbf{a}}} h_{\omega_{\mathbf{a}}} \|\nabla\psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}}) \|\nabla(\tilde{u} - u_h)\|_{\omega_{\mathbf{a}}}. \tag{7.54}$$

Now it suffices to choose for $\tilde{u}$ the weak solution u shifted on interior patches by a constant such that $(\tilde{u} - u_h, 1)_{\omega_{\mathbf{a}}} = 0$ to infer (7.53) with $C_{\rm cont,bPF} := \max_{\mathbf{a}\in\mathcal{V}_h}\{1 + C_{{\rm bPF},\omega_{\mathbf{a}}} h_{\omega_{\mathbf{a}}} \|\nabla\psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}}\}$. □

Remark 7.11.3 (Efficiency for the potential reconstruction of Remark 7.10.4)**.** *Efficiency for the potential reconstruction of Remark 7.10.4 for $d = 2$ can be shown as above. In particular, problem (7.51) with the right-hand sides $\boldsymbol{\tau}_h^{\mathbf{a}}$ and $g^{\mathbf{a}}$ from Remark 7.10.4 and $H^1_*(\omega_{\mathbf{a}})$ still defined by (7.52) reads: find $r_{\mathbf{a}} \in H^1_*(\omega_{\mathbf{a}})$ such that*

$$(\nabla r_{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}} = (\nabla u_h, \mathrm{R}_{\frac{\pi}{2}} \nabla(\psi_{\mathbf{a}} v))_{\omega_{\mathbf{a}}} \qquad \forall v \in H^1_*(\omega_{\mathbf{a}}).$$

An essential property is that $(\nabla u, \mathrm{R}_{\frac{\pi}{2}} \nabla(\psi_{\mathbf{a}} v))_{\omega_{\mathbf{a}}} = 0$. Thus,

$$(\nabla r_{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}} \le \|\nabla(u - u_h)\|_{\omega_{\mathbf{a}}} \big\|\mathrm{R}_{\frac{\pi}{2}} \nabla(\psi_{\mathbf{a}} v)\big\|_{\omega_{\mathbf{a}}} = \|\nabla(u - u_h)\|_{\omega_{\mathbf{a}}} \|\nabla(\psi_{\mathbf{a}} v)\|_{\omega_{\mathbf{a}}}$$

for any $v \in H^1_*(\omega_{\mathbf{a}})$, and we conclude by (7.49) that

$$\|\nabla r_{\mathbf{a}}\|_{\omega_{\mathbf{a}}} \le C_{\mathrm{cont,P}} \|\nabla(u - u_h)\|_{\omega_{\mathbf{a}}} \tag{7.55}$$

holds in this case, with $C_{\mathrm{cont,P}} := \max_{\mathbf{a} \in \mathcal{V}_h} \{1 + C_{\mathrm{P},\omega_{\mathbf{a}}} h_{\omega_{\mathbf{a}}} \|\nabla \psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}}\}$, thereby requiring only the (usual) Poincaré inequality (4.20).

Remark 7.11.4 (Dual and dual mixed formulations)**.** *For a vertex $\mathbf{a} \in \mathcal{V}_h$, consider the following dual formulation, cf. Definition 6.2.2: Find $\boldsymbol{\varsigma}_{\mathbf{a}} \in \mathbf{H}_*(\mathrm{div}, \omega_{\mathbf{a}})$ with $\nabla{\cdot}\boldsymbol{\varsigma}_{\mathbf{a}} = g^{\mathbf{a}}$ such that*

$$(\boldsymbol{\varsigma}_{\mathbf{a}}, \mathbf{v})_{\omega_{\mathbf{a}}} = -(\boldsymbol{\tau}_h^{\mathbf{a}}, \mathbf{v})_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega_{\mathbf{a}}) \text{ with } \nabla{\cdot}\mathbf{v} = 0. \tag{7.56}$$

Here, $\mathbf{H}_*(\mathrm{div}, \omega_{\mathbf{a}})$ stands for $\mathbf{H}(\mathrm{div}, \omega_{\mathbf{a}})$ functions with zero normal trace in the appropriate sense on $\partial\omega_{\mathbf{a}}$ for $\mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}$ and for $\mathbf{H}(\mathrm{div}, \omega_{\mathbf{a}})$ functions with zero normal trace in the appropriate sense on $\partial\omega_{\mathbf{a}} \setminus \partial\Omega$ for $\mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}$. Similarly, consider the dual mixed formulation, cf. Definition 6.2.3: Find $\boldsymbol{\varsigma}_{\mathbf{a}} \in \mathbf{H}_*(\mathrm{div}, \omega_{\mathbf{a}})$ and $r_{\mathbf{a}} \in L^2_*(\omega_{\mathbf{a}})$ such that

$$(\boldsymbol{\varsigma}_{\mathbf{a}}, \mathbf{v})_{\omega_{\mathbf{a}}} - (r_{\mathbf{a}}, \nabla{\cdot}\mathbf{v})_{\omega_{\mathbf{a}}} = -(\boldsymbol{\tau}_h^{\mathbf{a}}, \mathbf{v})_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{v} \in \mathbf{H}_*(\mathrm{div}, \omega_{\mathbf{a}}), \tag{7.57a}$$

$$(\nabla{\cdot}\boldsymbol{\varsigma}_{\mathbf{a}}, q)_{\omega_{\mathbf{a}}} = (g^{\mathbf{a}}, q)_{\omega_{\mathbf{a}}} \qquad \forall q \in L^2_*(\omega_{\mathbf{a}}). \tag{7.57b}$$

Here, $L^2_*(\omega_{\mathbf{a}})$ is the space of functions from $L^2(\omega_{\mathbf{a}})$ with zero mean value for $\mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}$ and $L^2(\omega_{\mathbf{a}})$ for $\mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}$. Problems (7.56) and (7.57) are equivalent to the primal formulation (7.44), with $\boldsymbol{\varsigma}_{\mathbf{a}} = -\nabla r_{\mathbf{a}} - \boldsymbol{\tau}_h^{\mathbf{a}}$, see Theorem 6.3.1 in Chapter 6. Then, (7.27) is the natural finite element discretization of (7.57), and the same links hold true in the potential reconstruction cases.

7.11.2 Uniform-in-polynomial-degree stability of mixed finite element methods

The following crucial result has been shown in Braess *et al.* [21, Theorem 7], based on Costabel and McIntosh [35, Corollary 3.4] and Demkowicz *et al.* [38, Theorem 7.1]:

Corollary 7.11.5 (Uniform stability of mixed finite element methods)**.** *Let $d = 2$. Let $\mathbf{a} \in \mathcal{V}_h$ and let $\boldsymbol{\tau}_h^{\mathbf{a}}$ and $g^{\mathbf{a}}$ be given either by (7.29) or by (7.39). Suppose that*

$$\boldsymbol{\tau}_h^{\mathbf{a}}|_K \in \mathbf{V}_h(K) \qquad \forall K \in \mathcal{T}_{\mathbf{a}}, \tag{7.58a}$$

$$g^{\mathbf{a}}|_K \in Q_h(K) \qquad \forall K \in \mathcal{T}_{\mathbf{a}}. \tag{7.58b}$$

Let $r_{\mathbf{a}} \in H^1_*(\omega_{\mathbf{a}})$ accordingly solve either (7.44) with $H^1_*(\omega_{\mathbf{a}})$ given by (7.45) or (7.51) with $H^1_*(\omega_{\mathbf{a}})$ given by (7.52). Let finally $\boldsymbol{\varsigma}_h^{\mathbf{a}}$ be the solution of either (7.27) or (7.37). Then there exists a constant $C_{\mathrm{st}} > 0$ only depending on the shape-regularity parameter $\kappa_{\mathcal{T}}$ such that

$$\|\boldsymbol{\varsigma}_h^{\mathbf{a}} + \boldsymbol{\tau}_h^{\mathbf{a}}\|_{\omega_{\mathbf{a}}} \le C_{\mathrm{st}} \|\nabla r_{\mathbf{a}}\|_{\omega_{\mathbf{a}}}. \tag{7.59}$$

Proof. We have from (7.44) or (7.51), using (7.47),

$$\begin{aligned}\|\nabla r_{\mathbf{a}}\|_{\omega_{\mathbf{a}}} &= \sup_{v\in H^1_*(\omega_{\mathbf{a}});\,\|\nabla v\|_{\omega_{\mathbf{a}}}=1}\{-(\boldsymbol{\tau}_h^{\mathbf{a}},\nabla v)_{\omega_{\mathbf{a}}}+(g^{\mathbf{a}},v)_{\omega_{\mathbf{a}}}\}\\ &= \sup_{v\in H^1_*(\omega_{\mathbf{a}});\,\|\nabla v\|_{\omega_{\mathbf{a}}}=1}\left\{\sum_{e\in\mathcal{E}_h,\,\mathbf{a}\in e}\langle\underbrace{[\![-\boldsymbol{\tau}_h^{\mathbf{a}}\cdot\mathbf{n}_e]\!]}_{r_e},v\rangle_e+\sum_{K\in\mathcal{T}_{\mathbf{a}}}(\underbrace{\nabla\cdot\boldsymbol{\tau}_h^{\mathbf{a}}+g^{\mathbf{a}}}_{r_K},v)_K\right\},\end{aligned}$$

so that $\|\nabla r_{\mathbf{a}}\|_{\omega_{\mathbf{a}}}$ in our notation is $\|r\|_{[H^1(\omega)/\mathbb{R}]^*}$ in the notation of [21]. Simultaneously, (7.33) and (7.41) read $\|\boldsymbol{\varsigma}_h^{\mathbf{a}}+\boldsymbol{\tau}_h^{\mathbf{a}}\|_{\omega_{\mathbf{a}}}=\inf_{\mathbf{v}_h\in\mathbf{V}_h^{\mathbf{a}},\,\nabla\cdot\mathbf{v}_h=g^{\mathbf{a}}}\|\mathbf{v}_h+\boldsymbol{\tau}_h^{\mathbf{a}}\|_{\omega_{\mathbf{a}}}$. Setting $\boldsymbol{\delta}_h^{\mathbf{a}}:=\boldsymbol{\varsigma}_h^{\mathbf{a}}+\boldsymbol{\tau}_h^{\mathbf{a}}$, we see that

$$\|\boldsymbol{\varsigma}_h^{\mathbf{a}}+\boldsymbol{\tau}_h^{\mathbf{a}}\|_{\omega_{\mathbf{a}}}=\|\boldsymbol{\delta}_h^{\mathbf{a}}\|_{\omega_{\mathbf{a}}}=\inf_{\mathbf{v}_h\in\mathbf{V}_h^{\mathbf{a}}(\mathcal{T}_{\mathbf{a}}),\,\nabla\cdot\mathbf{v}_h|_K=(\nabla\cdot\boldsymbol{\tau}_h^{\mathbf{a}}+g^{\mathbf{a}})|_K\,\forall K\in\mathcal{T}_{\mathbf{a}}}\|\mathbf{v}_h\|_{\omega_{\mathbf{a}}},$$

where $\mathbf{V}_h^{\mathbf{a}}(\mathcal{T}_{\mathbf{a}})$ is the broken version of $\mathbf{V}_h^{\mathbf{a}}$ with normal jumps imposed by $[\![\boldsymbol{\tau}_h^{\mathbf{a}}\cdot\mathbf{n}_e]\!]$, which is the form employed in [21, Theorem 7]. □

7.11.3 Polynomial-degree-robust local efficiency

We are now ready to prove the main result of this section:

Theorem 7.11.6 (Polynomial-degree-robust local efficiency)**.** *Let $d=2$. Let u be the weak solution of (7.2). Let u_h be a piecewise polynomial and consider Definition 7.9.1 of $\boldsymbol{\sigma}_h$ with the spaces $\mathbf{V}_h$ and Q_h satisfying, for all $\mathbf{a}\in\mathcal{V}_h$,*

$$(\psi_{\mathbf{a}}\nabla u_h)|_K\in\mathbf{V}_h(K)\qquad\forall K\in\mathcal{T}_{\mathbf{a}},\tag{7.60a}$$

$$(\nabla\psi_{\mathbf{a}}\cdot\nabla u_h)|_K\in Q_h(K)\qquad\forall K\in\mathcal{T}_{\mathbf{a}}.\tag{7.60b}$$

Then,

$$\begin{aligned}\|\nabla u_h+\boldsymbol{\sigma}_h\|_K\le C_{\mathrm{st}}C_{\mathrm{cont,PF}}&\sum_{\mathbf{a}\in\mathcal{V}_K}\|\nabla(u-u_h)\|_{\omega_{\mathbf{a}}}\\ &+C_{\mathrm{st}}\sum_{\mathbf{a}\in\mathcal{V}_K}\left\{\sum_{K'\in\mathcal{T}_{\mathbf{a}}}\Big(\frac{h_{K'}}{\pi}\|\psi_{\mathbf{a}}f-\Pi_{Q_h}(\psi_{\mathbf{a}}f)\|_{K'}\Big)^2\right\}^{\frac12},\end{aligned}\tag{7.61}$$

for all $K\in\mathcal{T}_h$, with the constants C_{st} of (7.59) and $C_{\mathrm{cont,PF}}$ of (7.46), respectively. Consider now Definition 7.10.1 of s_h in its equivalent form (7.37)–(7.40) with the space $\mathbf{V}_h$ satisfying, for all $\mathbf{a}\in\mathcal{V}_h$,

$$(\mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_h))|_K\in\mathbf{V}_h(K)\qquad\forall K\in\mathcal{T}_{\mathbf{a}}.\tag{7.62}$$

Assume in addition that u_h verifies the zero-mean condition (7.50). Then,

$$\|\nabla(u_h-s_h)\|_K\le C_{\mathrm{st}}C_{\mathrm{cont,bPF}}\sum_{\mathbf{a}\in\mathcal{V}_K}\|\nabla(u-u_h)\|_{\omega_{\mathbf{a}}}\tag{7.63}$$

for all $K\in\mathcal{T}_h$, with the constants C_{st} of (7.59) and $C_{\mathrm{cont,bPF}}$ of (7.53), respectively.

Proof. (1) We first prove (7.63). Let $K\in\mathcal{T}_h$. Using Definition 7.10.1 and Theorem 7.10.3, the decomposition $s_h|_K=\sum_{\mathbf{a}\in\mathcal{V}_K}s_h^{\mathbf{a}}|_K$, the partition of unity (7.32), and the triangle inequality,

we infer that

$$\begin{aligned}\|\nabla(u_h - s_h)\|_K &= \Big\|\sum_{\mathbf{a}\in\mathcal{V}_K}(\nabla(\psi_{\mathbf{a}}u_h - s_h^{\mathbf{a}}))|_K\Big\|_K \le \sum_{\mathbf{a}\in\mathcal{V}_K}\|\nabla(\psi_{\mathbf{a}}u_h - s_h^{\mathbf{a}})\|_K\\ &= \sum_{\mathbf{a}\in\mathcal{V}_K}\big\|\mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_h - s_h^{\mathbf{a}})\big\|_K = \sum_{\mathbf{a}\in\mathcal{V}_K}\big\|\mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_h) + \boldsymbol{\varsigma}_h^{\mathbf{a}}\big\|_K\\ &\le \sum_{\mathbf{a}\in\mathcal{V}_K}\big\|\mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_h) + \boldsymbol{\varsigma}_h^{\mathbf{a}}\big\|_{\omega_{\mathbf{a}}}.\end{aligned}$$

Noticing that (7.62) is equivalent to (7.58a) (and that $g^{\mathbf{a}} = 0$ in this case, so that condition (7.58b) is trivially satisfied), Corollary 7.11.5 readily yields

$$\big\|\mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_h) + \boldsymbol{\varsigma}_h^{\mathbf{a}}\big\|_{\omega_{\mathbf{a}}} \le C_{\mathrm{st}}\|\nabla r_{\mathbf{a}}\|_{\omega_{\mathbf{a}}}.$$

Lemma 7.11.2 concludes the proof of (7.63).

(2) The proof of (7.61) is similar, with the additional technicality of treating a possibly nonpolynomial source function f. Using Definition 7.9.1, $\boldsymbol{\sigma}_h|_K = \sum_{\mathbf{a}\in\mathcal{V}_K}\boldsymbol{\varsigma}_h^{\mathbf{a}}|_K$, the partition of unity (7.32), and the triangle inequality, we infer that

$$\|\nabla u_h + \boldsymbol{\sigma}_h\|_K = \Big\|\sum_{\mathbf{a}\in\mathcal{V}_K}(\psi_{\mathbf{a}}\nabla u_h + \boldsymbol{\varsigma}_h^{\mathbf{a}})|_K\Big\|_K \le \sum_{\mathbf{a}\in\mathcal{V}_K}\|\psi_{\mathbf{a}}\nabla u_h + \boldsymbol{\varsigma}_h^{\mathbf{a}}\|_{\omega_{\mathbf{a}}}.$$

Note that replacing $\psi_{\mathbf{a}}f$ by $\Pi_{Q_h}(\psi_{\mathbf{a}}f)$ in (7.29b) does not change the solution couple $(\boldsymbol{\varsigma}_h^{\mathbf{a}}, \bar{r}_h^{\mathbf{a}})$ of (7.27). Thus, setting $\tilde{g}^{\mathbf{a}} := \Pi_{Q_h}(\psi_{\mathbf{a}}f) - \nabla\psi_{\mathbf{a}}\cdot\nabla u_h$, assumption (7.60b) implies (7.58b) with $g^{\mathbf{a}}$ replaced by $\tilde{g}^{\mathbf{a}}$, while assumption (7.60a) implies (7.58a). Consequently, Corollary 7.11.5 yields

$$\|\psi_{\mathbf{a}}\nabla u_h + \boldsymbol{\varsigma}_h^{\mathbf{a}}\|_{\omega_{\mathbf{a}}} \le C_{\mathrm{st}}\|\nabla\tilde{r}_{\mathbf{a}}\|_{\omega_{\mathbf{a}}},$$

where $\tilde{r}_{\mathbf{a}}$ solves (7.44) with $g^{\mathbf{a}}$ replaced by $\tilde{g}^{\mathbf{a}}$. We now need to inspect the proof of Lemma 7.11.1. We observe that

$$(\nabla\tilde{r}_{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}} = -(\psi_{\mathbf{a}}\nabla u_h, \nabla v)_{\omega_{\mathbf{a}}} + (\psi_{\mathbf{a}}f - \nabla\psi_{\mathbf{a}}\cdot\nabla u_h, v)_{\omega_{\mathbf{a}}} + (\Pi_{Q_h}(\psi_{\mathbf{a}}f) - \psi_{\mathbf{a}}f, v)_{\omega_{\mathbf{a}}}$$

in place of (7.48). The first two terms on the above right-hand side are treated as in the proof of Lemma 7.11.1, and we are left to bound

$$\begin{aligned}&\sup_{v\in H^1_*(\omega_{\mathbf{a}});\,\|\nabla v\|_{\omega_{\mathbf{a}}}=1}(\Pi_{Q_h}(\psi_{\mathbf{a}}f) - \psi_{\mathbf{a}}f, v)_{\omega_{\mathbf{a}}}\\ &= \sup_{v\in H^1_*(\omega_{\mathbf{a}});\,\|\nabla v\|_{\omega_{\mathbf{a}}}=1}\Big\{\sum_{K'\in\mathcal{T}_{\mathbf{a}}}(\Pi_{Q_h}(\psi_{\mathbf{a}}f) - \psi_{\mathbf{a}}f, v - \Pi^0_{K'}v)_{K'}\Big\}\\ &\le \Big\{\sum_{K'\in\mathcal{T}_{\mathbf{a}}}\Big(\frac{h_{K'}}{\pi}\|\psi_{\mathbf{a}}f - \Pi_{Q_h}(\psi_{\mathbf{a}}f)\|_{K'}\Big)^2\Big\}^{\frac12},\end{aligned}$$

as in (7.25). Combining the above results concludes the proof of (7.61). □

Remark 7.11.7 (Data oscillation)**.** *As in Remark 7.9.3, if f is elementwise smooth enough, the data oscillation term in (7.61) typically converges by two orders of magnitude faster than the energy error.*

Remark 7.11.8 (Robustness for the potential reconstruction of Remark 7.10.4)**.** *Proceeding as in the above proof shows that under the assumptions*

$$(\psi_{\mathbf{a}}\mathrm{R}_{\frac{\pi}{2}}\nabla u_h)|_K \in \mathbf{V}_h(K) \qquad \forall K \in \mathcal{T}_{\mathbf{a}}, \tag{7.64a}$$

$$((\mathrm{R}_{\frac{\pi}{2}}\nabla\psi_{\mathbf{a}})\cdot\nabla u_h)|_K \in Q_h(K) \qquad \forall K \in \mathcal{T}_{\mathbf{a}}, \tag{7.64b}$$

the potential reconstruction of Remark 7.10.4 for $d = 2$ satisfies the bound (7.63) with $C_{\mathrm{cont,bPF}}$ replaced by $C_{\mathrm{cont,P}}$ of (7.55).

Remark 7.11.9 (Examples of choice for the degree of $\mathbf{V}_h$ and Q_h)**.** *For $u_h \in \mathbb{P}_k(\mathcal{T}_h)$, $k \geq 1$, as in many classical numerical methods discussed in Section 7.13 below, the adequate choice for $\mathbf{V}_h \times Q_h$ is $\mathbf{RT}_k \times \mathbb{P}_k(\mathcal{T}_h)$.*

7.12 Maximal overestimation

The previous developments do now alow yet to specify maximal overestimation by our a posteriori estimates, as the constant C_{st} from (7.59) is unknown. We show in this section how guaranteed (local) maximal overestimation can be obtained. The way to the present results has been paved by, among others, Babuška *et al.* [14], Carstensen and Funken [27], Babuška and Strouboulis [13, Section 5.1], Prudhomme *et al.* [82], or Repin [85, Section 4.1.1], see also the references therein, but not necessarily simultaneously with a guaranteed upper bound.

Lemma 7.12.1 (Maximal overestimation)**.** *Let the assumptions of Theorem 7.11.6 be verified, with for simplicity $\psi_{\mathbf{a}} f \in Q_h$ (i.e., with (7.58b) satisfied). Then*

$$\|\nabla u_h + \boldsymbol{\sigma}_h\| \leq 3C_{\mathrm{st}}C_{\mathrm{cont,PF}}\|\nabla(u - u_h)\|,$$
$$\|\nabla(u_h - s_h)\| \leq 3C_{\mathrm{st}}C_{\mathrm{cont,bPF}}\|\nabla(u - u_h)\|.$$

Proof. Employing $\boldsymbol{\sigma}_h|_K = \sum_{\mathbf{a}\in\mathcal{V}_K} \boldsymbol{\varsigma}_h^{\mathbf{a}}|_K$, the partition of unity (7.32), the Cauchy–Schwarz inequality, and proceeding as in the proof of Theorem 7.11.6, we infer that

$$\begin{aligned}
&\|\nabla u_h + \boldsymbol{\sigma}_h\|^2 \\
&= \sum_{K\in\mathcal{T}_h} \|\nabla u_h + \boldsymbol{\sigma}_h\|_K^2 = \sum_{K\in\mathcal{T}_h} \Big\| \sum_{\mathbf{a}\in\mathcal{V}_K} (\psi_{\mathbf{a}}\nabla u_h + \boldsymbol{\varsigma}_h^{\mathbf{a}})|_K \Big\|_K^2 \\
&\leq 3 \sum_{K\in\mathcal{T}_h} \sum_{\mathbf{a}\in\mathcal{V}_K} \|\psi_{\mathbf{a}}\nabla u_h + \boldsymbol{\varsigma}_h^{\mathbf{a}}\|_K^2 = 3 \sum_{\mathbf{a}\in\mathcal{V}_h} \|\psi_{\mathbf{a}}\nabla u_h + \boldsymbol{\varsigma}_h^{\mathbf{a}}\|_{\omega_{\mathbf{a}}}^2 \\
&\leq 3C_{\mathrm{st}}^2 C_{\mathrm{cont,PF}}^2 \sum_{\mathbf{a}\in\mathcal{V}_h} \|\nabla(u - u_h)\|_{\omega_{\mathbf{a}}}^2 = 9C_{\mathrm{st}}^2 C_{\mathrm{cont,PF}}^2 \|\nabla(u - u_h)\|^2.
\end{aligned} \tag{7.65}$$

The bound for $\|\nabla(u_h - s_h)\|$ is similar. □

We finally present a local result indicating additionally how to give a computable upper bound on the value of the unknown constant C_{st} of (7.59):

Lemma 7.12.2 (Guaranteed maximal local overestimation by auxiliary problems)**.** *Let the assumptions of Theorem 7.11.6 be verified, with additionally $\psi_{\mathbf{a}} f \in Q_h$. Fix $\mathbf{a} \in \mathcal{V}_h$ and consider an arbitrary conforming finite element approximation in $W_h^{\mathbf{a}} := \mathbb{P}_{\bar{k}}(\mathcal{T}_{\mathbf{a}}) \cap H^1_*(\omega_{\mathbf{a}})$, $\bar{k} \geq 1$, of (7.44) or (7.51) in the form: find $r_h^{\mathbf{a}} \in W_h^{\mathbf{a}}$ such that*

$$(\nabla r_h^{\mathbf{a}}, \nabla v_h)_{\omega_{\mathbf{a}}} = -(\boldsymbol{\tau}_h^{\mathbf{a}}, \nabla v_h)_{\omega_{\mathbf{a}}} + (g^{\mathbf{a}}, v_h)_{\omega_{\mathbf{a}}} \qquad \forall v_h \in W_h^{\mathbf{a}},$$

with the usual choices (7.29) *or* (7.39) *for the right-hand side. Then,*

$$\|\psi_{\mathbf{a}}\nabla u_h + \boldsymbol{\varsigma}_h^{\mathbf{a}}\|_{\omega_{\mathbf{a}}} \le C_{\mathrm{cont,PF}} \frac{\|\psi_{\mathbf{a}}\nabla u_h + \boldsymbol{\varsigma}_h^{\mathbf{a}}\|_{\omega_{\mathbf{a}}}}{\|\nabla r_h^{\mathbf{a}}\|_{\omega_{\mathbf{a}}}} \|\nabla(u-u_h)\|_{\omega_{\mathbf{a}}}, \tag{7.66a}$$

$$\|\nabla(\psi_{\mathbf{a}}u_h - s_h^{\mathbf{a}})\|_{\omega_{\mathbf{a}}} \le C_{\mathrm{cont,bPF}} \frac{\|\nabla(\psi_{\mathbf{a}}u_h - s_h^{\mathbf{a}})\|_{\omega_{\mathbf{a}}}}{\|\nabla r_h^{\mathbf{a}}\|_{\omega_{\mathbf{a}}}} \|\nabla(u-u_h)\|_{\omega_{\mathbf{a}}}. \tag{7.66b}$$

Proof. As $r_h^{\mathbf{a}}$ is the $(\nabla\cdot,\nabla\cdot)_{\omega_{\mathbf{a}}}$-orthogonal projection of $r_{\mathbf{a}}$ onto $W_h^{\mathbf{a}}$, $\|\nabla r_h^{\mathbf{a}}\|_{\omega_{\mathbf{a}}} \le \|\nabla r_{\mathbf{a}}\|_{\omega_{\mathbf{a}}}$. Thus the results follow respectively from Lemmas 7.11.1 and 7.11.2. □

Remark 7.12.3 (Size of overestimation, comparison with [27])**.** *The above lemma together with Remark 7.11.4 suggest that the constant C_{st} approaches 1 as the polynomial degrees k, $\bar{k}$ are increased. Next, for convex patches $\omega_{\mathbf{a}}$ around interior vertices $\mathbf{a}$, $C_{\mathrm{P},\omega_{\mathbf{a}}} = 1/\pi$, whereas $h_{\omega_{\mathbf{a}}}\|\nabla\psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}} \approx 2$ for "nice" meshes. Thus we may expect $C_{\mathrm{cont,PF}} \approx 1 + 2/\pi$ from the proof of Lemma 7.11.1 in such a case. Then Lemma 7.12.1 gives $3C_{\mathrm{st}}C_{\mathrm{cont,PF}} \approx 4.9$ for the maximal theoretical overestimation factor. In practice, however, the effectivity indices of the present estimates are quite close to the optimal value of one, see [21] and Section 7.14 below. For the conforming finite element method, Carstensen and Funken [27, Example 3.1] obtain a maximal theoretical overestimation factor 2.34 for "nice" meshes, which is roughly twice better than our result. This can be attributed to the localization of the estimators around mesh vertices with a specific use of the partition of unity in [27], see equation (3.7) in this reference and also the next remark, whereas we loose roughly a factor 3 in the estimate* (7.65)*. Note, however, that the upper bound in [27] is, in contrast to the lower one, not guaranteed.*

Remark 7.12.4 (Localization on the patches $\omega_{\mathbf{a}}$)**.** *In [27], see in particular Theorem 3.2 therein, the following local problems similar to* (7.44) *are considered: find $r_{\mathbf{a}} \in \bar{H}^1_*(\omega_{\mathbf{a}})$ such that, with the choice* (7.29) *for the right-hand side,*

$$(\psi_{\mathbf{a}}\nabla r_{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}} = -(\boldsymbol{\tau}_h^{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}} + (g^{\mathbf{a}}, v)_{\omega_{\mathbf{a}}} \qquad \forall v \in \bar{H}^1_*(\omega_{\mathbf{a}}),$$

where $\bar{H}^1_*(\omega_{\mathbf{a}})$ are $\psi_{\mathbf{a}}^{\frac{1}{2}}$-weighted versions of the spaces (7.45)*, and the (unfortunately not computable) a posteriori error estimator is simply $\|\psi_{\mathbf{a}}^{\frac{1}{2}}\nabla r_{\mathbf{a}}\|_{\omega_{\mathbf{a}}}$. Adjusting the equilibration of Definition 7.9.1, its computable upper bound may be constructed via local problems consisting in finding $\boldsymbol{\varsigma}_h^{\mathbf{a}} \in \mathbf{V}_h^{\mathbf{a}}$ and $\bar{r}_h^{\mathbf{a}} \in Q_h^{\mathbf{a}}$ such that*

$$\begin{aligned}(\psi_{\mathbf{a}}\boldsymbol{\varsigma}_h^{\mathbf{a}}, \mathbf{v}_h)_{\omega_{\mathbf{a}}} - (\bar{r}_h^{\mathbf{a}}, \nabla\cdot(\psi_{\mathbf{a}}\mathbf{v}_h))_{\omega_{\mathbf{a}}} &= -(\boldsymbol{\tau}_h^{\mathbf{a}}, \mathbf{v}_h)_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{v}_h \in \mathbf{V}_h^{\mathbf{a}},\\ (\nabla\cdot(\psi_{\mathbf{a}}\boldsymbol{\varsigma}_h^{\mathbf{a}}), q_h)_{\omega_{\mathbf{a}}} &= (g^{\mathbf{a}}, q_h)_{\omega_{\mathbf{a}}} \qquad \forall q_h \in Q_h^{\mathbf{a}}.\end{aligned}$$

7.13 Application to classical discretizations

We show here how to apply our results to common discretizations via the verification of the assumptions (7.26), (7.60), and (7.62). The condition (7.50) or alternatively (7.42) will also be discussed.

7.13.1 Finite element method

Set $V_h := \mathbb{P}_k(\mathcal{T}_h) \cap H^1_0(\Omega)$, $k \ge 1$. The finite element (FE) method reads:

Definition 7.13.1 (FE method for (7.1a)–(7.1b))**.** *Find $u_h \in V_h$ such that*

$$(\nabla u_h, \nabla v_h) = (f, v_h) \qquad \forall v_h \in V_h. \tag{7.67}$$

The application of our framework is straightforward: (7.26) is nothing but the Galerkin orthogonality with respect to the hat basis function $\psi_{\mathbf{a}}$ which follows immediately from (7.67) as $\psi_{\mathbf{a}} \in V_h$ for $\mathbf{a} \in \mathcal{V}_h^{\text{int}}$. Thus the equilibrated flux $\boldsymbol{\sigma}_h$ can be reconstructed following Definition 7.9.1. Condition (7.60) is then satisfied for the choice $\mathbf{V}_h \times Q_h := \mathbf{RT}_k \times \mathbb{P}_k(\mathcal{T}_h)$. The approximate solution u_h is $H_0^1(\Omega)$-conforming, so that we set $s_h := u_h$, the nonconformity estimators $\eta_{\text{NC},K} = \|\nabla(u_h - s_h)\|_K$ disappear, and there is nothing to verify in this respect. The resulting error estimators correspond to those of [40, 22, 21]. We summarize this result in:

Theorem 7.13.2 (Application to the FE method)**.** *All the results of this section hold for the finite element solution u_h of Definition 7.13.1 with $\boldsymbol{\sigma}_h$ constructed following Definition 7.9.1, with $\mathbf{V}_h \times Q_h := \mathbf{RT}_k \times \mathbb{P}_k(\mathcal{T}_h)$ and $s_h = u_h$.*

7.13.2 Nonconforming finite element method

Let V_h stand for functions from $\mathbb{P}_k(\mathcal{T}_h)$, $k \geq 1$, satisfying $\langle [\![u_h]\!], q_h \rangle_e = 0$ for all polynomials q_h on e up to degree $k-1$ and for all $e \in \mathcal{E}_h$. In the lowest-order case $k = 1$, these are piecewise affine functions which are continuous in the barycenters of all interior faces and equal to zero in the barycenters of all boundary faces. The nonconforming finite element method for (7.2), cf. Crouzeix and Raviart [36], Stoyan and Baran [90], or Ainsworth and Rankin [7], reads:

Definition 7.13.3 (NCFE method for (7.1a)–(7.1b))**.** *Find $u_h \in V_h$ such that*

$$(\nabla u_h, \nabla v_h) = (f, v_h) \qquad \forall v_h \in V_h. \tag{7.68}$$

The approximate solution u_h of the nonconforming finite element method (7.68) is precisely such that (7.6) holds: both its approximate potential u_h and its approximate flux $-\nabla u_h$ are nonconforming. We thus need to specify both the potential and flux reconstructions s_h and $\boldsymbol{\sigma}_h$. This is again straightforward: condition (7.26) follows immediately from (7.68) as $\psi_{\mathbf{a}} \in V_h$ for all $\mathbf{a} \in \mathcal{V}_h^{\text{int}}$. The approximate solution u_h also satisfies (7.50) from the definition of the space V_h, so that both constructions of Definition 7.10.1 and that of Remark 7.10.4 are possible. In summary:

Theorem 7.13.4 (Application to the NCFE method)**.** *All the results of this section hold for the nonconforming finite element solution u_h of Definition 7.13.3 with $\boldsymbol{\sigma}_h$ constructed following Definition 7.9.1, s_h constructed following Definition 7.10.1, $\mathbf{V}_h \times Q_h := \mathbf{RT}_k \times \mathbb{P}_k(\mathcal{T}_h)$, and the corresponding $W_h := \mathbb{P}_{k+1}(\mathcal{T}_h) \cap H_0^1(\Omega)$.*

Remark 7.13.5 (Implicit and explicit flux reconstructions)**.** *It has been recently shown in [54] that several seemingly different flux reconstructions for nonconforming finite elements coincide, including that of Definition 7.9.1 with the lowest-order Raviart–Thomas–Nédélec space and the explicit constructions of [39, 3], see Section 8.3.3 below. So, at least in this particular case, this smears the conceptual difference between the implicit estimators presented in this section (where solutions of local problems are necessary) and the, a priori cheaper, explicit (directly computable) ones.*

7.13.3 Discontinuous Galerkin method

Set $V_h := \mathbb{P}_k(\mathcal{T}_h)$, $k \geq 1$, without any continuity requirement. The discontinuous Galerkin (DG) method method, see Di Pietro and Ern [41] and the references therein, is:

Definition 7.13.6 (DG method for (7.1a)–(7.1b)). *Find $u_h \in V_h$ such that*

$$\begin{aligned}&\sum_{K\in\mathcal{T}_h}(\nabla u_h,\nabla v_h)_K-\sum_{e\in\mathcal{E}_h}\{\langle\{\!\{\nabla u_h\}\!\}\cdot\mathbf{n}_e,[\![v_h]\!]\rangle_e+\theta\langle\{\!\{\nabla v_h\}\!\}\cdot\mathbf{n}_e,[\![u_h]\!]\rangle_e\}\\&+\sum_{e\in\mathcal{E}_h}\langle\alpha h_e^{-1}[\![u_h]\!],[\![v_h]\!]\rangle_e=(f,v_h)\qquad\forall v_h\in V_h,\end{aligned}\tag{7.69}$$

where α is a positive stabilization parameter and $\theta\in\{-1,0,1\}$ corresponds respectively to the nonsymmetric, incomplete, and symmetric version.

Here again (7.6) is met, so that we proceed to flux and potential reconstructions following Definitions 7.9.1 and 7.10.1. In the DG context, these have been introduced in [53, 55]. Adjustments for the conditions (7.50) and (7.42) and separate treatments for the different values of θ will be necessary. We undertake it now.

Discrete gradient and flux reconstruction

Introduce the discrete gradient $\mathfrak{G}(u_h):=\nabla u_h-\theta\sum_{e\in\mathcal{E}_h}\mathfrak{l}_e([\![u_h]\!])$ where the lifting operator $\mathfrak{l}_e:L^2(e)\to[\mathbb{P}_0(\mathcal{T}_h)]^2$ is such that $(\mathfrak{l}_e([\![u_h]\!]),\mathbf{v}_h)=\langle\{\!\{\mathbf{v}_h\}\!\}\cdot\mathbf{n}_e,[\![u_h]\!]\rangle_e$ for all $\mathbf{v}_h\in[\mathbb{P}_0(\mathcal{T}_h)]^2$, see [41, Section 4.3]. Observe that $\mathfrak{G}(v)=\nabla v$ for any function v with zero jumps or for any function in $H^1(\mathcal{T}_h)$ if $\theta=0$. Then, taking $v_h=\psi_{\mathbf{a}}$ in (7.69) and since $\psi_{\mathbf{a}}$ has no jumps and $\nabla\psi_{\mathbf{a}}\in[\mathbb{P}_0(\mathcal{T}_h)]^2$, we infer that $(\mathfrak{G}(u_h),\nabla\psi_{\mathbf{a}})_{\omega_{\mathbf{a}}}=(f,\psi_{\mathbf{a}})_{\omega_{\mathbf{a}}}$ for all $\mathbf{a}\in\mathcal{V}_h^{\rm int}$ instead of the hat-function orthogonality (7.26). Thus, Definition 7.9.1 for the flux is possible with right-hand sides $\boldsymbol{\tau}_h^{\mathbf{a}}:=\psi_{\mathbf{a}}\mathfrak{G}(u_h)$ and $g^{\mathbf{a}}:=\psi_{\mathbf{a}}f-\nabla\psi_{\mathbf{a}}\cdot\mathfrak{G}(u_h)$. The guaranteed estimate of Theorem 7.8.1 using the discrete gradient takes the form

$$\|\mathfrak{G}(u-u_h)\|^2\le\sum_{K\in\mathcal{T}_h}\Big(\|\mathfrak{G}(u_h)+\boldsymbol{\sigma}_h\|_K+\frac{h_K}{\pi}\|f-\Pi_{Q_h}f\|_K\Big)^2+\sum_{K\in\mathcal{T}_h}\|\mathfrak{G}(u_h-s_h)\|_K^2,\tag{7.70}$$

and the local efficiency result (7.61) for the flux reconstruction (with $\psi_{\mathbf{a}}f\in Q_h$ for simplicity) takes the form

$$\|\mathfrak{G}(u_h)+\boldsymbol{\sigma}_h\|_K\le C_{\rm st}C_{\rm cont,PF}\sum_{\mathbf{a}\in\mathcal{V}_K}\|\mathfrak{G}(u-u_h)\|_{\omega_{\mathbf{a}}},\tag{7.71}$$

with the polynomial-degree-independent constants $C_{\rm st}$ of (7.59) and $C_{\rm cont,PF}$ of (7.46).

Potential reconstruction for the nonsymmetric and incomplete versions

We use Definition 7.10.1 for the potential reconstruction (observe that condition (7.42) does not hold). As the mean-zero condition (7.50) on the jumps is not satisfied either, we cannot use directly Lemma 7.11.2. The inspection of its proof, however, shows that we merely need to replace the estimate (7.54) by

$$\begin{aligned}\|\nabla(\psi_{\mathbf{a}}\tilde u-\psi_{\mathbf{a}}u_h)\|_{\omega_{\mathbf{a}}}&\le\|\nabla\psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}}\|\tilde u-u_h\|_{\omega_{\mathbf{a}}}+\|\psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}}\|\nabla(\tilde u-u_h)\|_{\omega_{\mathbf{a}}}\\&\le(1+C_{\rm bPF,\omega_{\mathbf{a}}}h_{\omega_{\mathbf{a}}}\|\nabla\psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}})\|\nabla(\tilde u-u_h)\|_{\omega_{\mathbf{a}}}\\&\quad+C_{\rm bPF,\omega_{\mathbf{a}}}h_{\omega_{\mathbf{a}}}\|\nabla\psi_{\mathbf{a}}\|_{\infty,\omega_{\mathbf{a}}}\left\{\sum_{e\in\mathcal{E}_h^{\rm int},\,\mathbf{a}\in e}h_e^{-1}\|\Pi_e^0[\![u_h]\!]\|_e^2\right\}^{\frac12},\end{aligned}$$

with Π_e^0 the $L^2(e)$-orthogonal projection onto constants, using the broken Poincaré–Friedrichs inequalities (4.23)–(4.24) (since $(\tilde{u}-u_h,1)_{\omega_{\mathbf{a}}}=0$ and since $\tilde{u}$ has no jumps). Thus,

$$\|\nabla r_{\mathbf{a}}\|_{\omega_{\mathbf{a}}} \le C_{\mathrm{cont,bPF}}\left(\|\nabla(u-u_h)\|_{\omega_{\mathbf{a}}}+\left\{\sum_{e\in\mathcal{E}_h^{\mathrm{int}},\,\mathbf{a}\in e} h_e^{-1}\|\Pi_e^0[\![u-u_h]\!]\|_e^2\right\}^{\frac12}\right)$$

in place of (7.53). The local efficiency result (7.63) then yields

$$\|\nabla(u_h-s_h)\|_K \le C_{\mathrm{st}}C_{\mathrm{cont,bPF}}\sum_{\mathbf{a}\in\mathcal{V}_K}\left(\|\nabla(u-u_h)\|_{\omega_{\mathbf{a}}}+\left\{\sum_{e\in\mathcal{E}_h^{\mathrm{int}},\,\mathbf{a}\in e} h_e^{-1}\|\Pi_e^0[\![u-u_h]\!]\|_e^2\right\}^{\frac12}\right). \tag{7.72}$$

It is still polynomial-degree robust, but features the additional jump term. The classical option to obtain both upper and lower bounds for the same error measure is to resort to the jumps-augmented energy norm, thereby replacing (7.70) by

$$\begin{aligned}\|\mathfrak{G}(u-u_h)\|^2+\sum_{e\in\mathcal{E}_h} h_e^{-1}\|\Pi_e^0[\![u-u_h]\!]\|_e^2 \le{}& \sum_{K\in\mathcal{T}_h}\left(\|\mathfrak{G}u_h+\boldsymbol{\sigma}_h\|_K+\frac{h_K}{\pi}\|f-\Pi_{Q_h}f\|_K\right)^2\\ &+\sum_{K\in\mathcal{T}_h}\|\mathfrak{G}(u_h-s_h)\|_K^2+\sum_{e\in\mathcal{E}_h} h_e^{-1}\|\Pi_e^0[\![u_h]\!]\|_e^2,\end{aligned} \tag{7.73}$$

using that $[\![u-u_h]\!]=-[\![u_h]\!]$. Then, for the incomplete version, observing that $\nabla(u_h-s_h)=\mathfrak{G}(u_h-s_h)$ and $\nabla(u-u_h)=\mathfrak{G}(u-u_h)$ in (7.72), (7.71) combined with (7.72) yields polynomial-degree-robust local efficiency for the same error measure as in (7.73).

For the nonsymmetric version, we need a bound similar to (7.72), but using the discrete gradient. Since the lifting $\mathfrak{l}$ only includes the neighboring elements and using the triangle inequality, we infer that

$$\|\mathfrak{G}(u_h-s_h)\|_K \le \|\nabla(u_h-s_h)\|_K+\sum_{e\in\mathcal{E}_K}\|\mathfrak{l}_e([\![u_h]\!])\|_K. \tag{7.74}$$

The term $\|\nabla(u_h-s_h)\|_K$ is bounded using (7.72), where on the right-hand side, we further bound $\|\nabla(u-u_h)\|_{\omega_{\mathbf{a}}}$ by $\|\mathfrak{G}(u-u_h)\|_{\omega_{\mathbf{a}}}+\sum_{K\in\mathcal{T}_{\mathbf{a}}}\sum_{e\in\mathcal{E}_K}\|\mathfrak{l}_e([\![u_h]\!])\|_K$. Additionally, relying on the fact that the lifting $\mathfrak{l}$ maps onto piecewise constant functions,

$$\begin{aligned}\|\mathfrak{l}_e([\![u_h]\!])\|_K &\le \sup_{\mathbf{v}_h\in[\mathbb{P}_0(\mathcal{T}_e)]^2;\,\|\mathbf{v}_h\|_{\mathcal{T}_e}=1}(\mathfrak{l}_e([\![u_h]\!]),\mathbf{v}_h)_{\mathcal{T}_e}\\ &= \sup_{\mathbf{v}_h\in[\mathbb{P}_0(\mathcal{T}_e)]^2;\,\|\mathbf{v}_h\|_{\mathcal{T}_e}=1}\langle\{\!\{\mathbf{v}_h\}\!\}\cdot\mathbf{n}_e,\Pi_e^0[\![u-u_h]\!]\rangle_e\\ &\le C_{\kappa_{\mathcal{T}}}h_e^{-\frac12}\|\Pi_e^0[\![u-u_h]\!]\|_e,\end{aligned} \tag{7.75}$$

where $\mathcal{T}_e$ stands for the (one or two) elements sharing the edge e and $C_{\kappa_{\mathcal{T}}}$ uniformly bounds $\frac{h_e}{|K|^{\frac12}}$ and only depends on the mesh-regularity parameter $\kappa_{\mathcal{T}}$. Finally,

$$\|\mathfrak{G}(u_h-s_h)\|_K \le C_{\mathrm{st}}C_{\mathrm{cont,bPF}}\sum_{\mathbf{a}\in\mathcal{V}_K}\|\mathfrak{G}(u-u_h)\|_{\omega_{\mathbf{a}}}+C\left\{\sum_{e\in\mathcal{E}_K^+} h_e^{-1}\|\Pi_e^0[\![u-u_h]\!]\|_e^2\right\}^{\frac12}, \tag{7.76}$$

where C only depends on the mesh-regularity parameter $\kappa_{\mathcal{T}}$ and $\mathcal{E}_K^+:=\{e\in\mathcal{E}_h|\exists\mathbf{a}\in\mathcal{V}_K,\exists K'\in\mathcal{T}_{\mathbf{a}},e\in\mathcal{E}_{K'}\}$, so that (7.71) combined with (7.76) yields polynomial-degree-robust local efficiency for the same error measure as in (7.73).

Potential reconstruction for the symmetric version

A remarkable fact is that the discrete gradient $\mathfrak{G}$ satisfies the following modification of condition (7.42) related to the alternative potential reconstruction from Remark 7.10.4:

$$(\mathfrak{G}(u_h), \mathrm{R}_{\frac{\pi}{2}}\nabla\psi_{\mathbf{a}})_{\omega_{\mathbf{a}}} = 0 \qquad \forall \mathbf{a} \in \mathcal{V}_h. \tag{7.77}$$

Indeed, using the definition of the discrete gradient and the Green theorem, we have

$$\begin{aligned}(\mathfrak{G}(u_h), \mathrm{R}_{\frac{\pi}{2}}\nabla\psi_{\mathbf{a}})_{\omega_{\mathbf{a}}} &= (\nabla u_h, \mathrm{R}_{\frac{\pi}{2}}\nabla\psi_{\mathbf{a}})_{\omega_{\mathbf{a}}} - \theta \sum_{e\in\mathcal{E}_h} \langle \{\!\{ \mathrm{R}_{\frac{\pi}{2}}\nabla\psi_{\mathbf{a}} \}\!\}\cdot\mathbf{n}_e, [\![u_h]\!]\rangle_e \\ &= \sum_{K\in\mathcal{T}_{\mathbf{a}}} \langle u_h, (\mathrm{R}_{\frac{\pi}{2}}\nabla\psi_{\mathbf{a}})\cdot\mathbf{n}_K\rangle_{\partial K} - \theta \sum_{e\in\mathcal{E}_h} \langle \{\!\{ \mathrm{R}_{\frac{\pi}{2}}\nabla\psi_{\mathbf{a}} \}\!\}\cdot\mathbf{n}_e, [\![u_h]\!]\rangle_e,\end{aligned}$$

where $\mathbf{n}_K$ is the outward unit normal vector to K. Now for $\theta = 1$, the two above terms cancel. Thus we can use here the procedure of Remark 7.10.4, where we systematically replace ∇v by $\mathfrak{G}(v)$. The local efficiency result for the flux reconstruction is (7.71) and the one for the potential reconstruction takes the form discussed in Remark 7.11.8,

$$\|\mathfrak{G}(u_h - s_h)\|_K \le C_{\mathrm{st}} C_{\mathrm{cont,P}} \sum_{\mathbf{a}\in\mathcal{V}_K} \|\mathfrak{G}(u - u_h)\|_{\omega_{\mathbf{a}}},$$

with the polynomial-degree-independent constants C_{st} of (7.59) and $C_{\mathrm{cont,P}}$ of (7.55). Note that in this symmetric case, the lifting operator $\mathfrak{l}$ can alternatively be designed as $\mathfrak{l}_e : L^2(e) \to [\mathbb{P}_{k-1}(\mathcal{T}_h)]^2$ with $(\mathfrak{l}_e([\![u_h]\!]), \mathbf{v}_h) = \langle \{\!\{ \mathbf{v}_h \}\!\}\cdot\mathbf{n}_e, [\![u_h]\!]\rangle_e$ for all $\mathbf{v}_h \in [\mathbb{P}_{k-1}(\mathcal{T}_h)]^2$.

Summary

In summary, we have:

Theorem 7.13.7 (Application to the DG method)**.** *With the adjustments explained above, all the results of this section hold for the discontinuous Galerkin finite element solution u_h of Definition 7.13.6 with $\boldsymbol{\sigma}_h$ constructed following Definition 7.9.1, s_h constructed following Definition 7.10.1, $\mathbf{V}_h \times Q_h := \mathbf{RT}_k \times \mathbb{P}_k(\mathcal{T}_h)$, and the corresponding $W_h := \mathbb{P}_{k+1}(\mathcal{T}_h) \cap H^1_0(\Omega)$.*

7.13.4 Mixed finite element method

Let $\mathbf{V}_h \times Q_h$ be given by $\mathbf{RTN}_{k'} \times \mathbb{P}_{k'}(\mathcal{T}_h)$, $k' \ge 0$. The mixed finite element (MFE) method reads:

Definition 7.13.8 (MFE method for (7.1a)–(7.1b))**.** *Find $\boldsymbol{\sigma}_h \in \mathbf{V}_h$ and $\bar{u}_h \in Q_h$ such that*

$$(\boldsymbol{\sigma}_h, \mathbf{v}_h) - (\bar{u}_h, \nabla\cdot\mathbf{v}_h) = 0 \qquad \forall \mathbf{v}_h \in \mathbf{V}_h, \tag{7.78a}$$

$$(\nabla\cdot\boldsymbol{\sigma}_h, q_h) = (f, q_h) \qquad \forall q_h \in Q_h. \tag{7.78b}$$

We have written the formulation explicitly with $\boldsymbol{\sigma}_h$ since this computed flux can serve directly as the equilibrated flux reconstruction of Definition 7.6.2. Flux equilibration following Definition 7.9.1 is useless here (and unfeasible as (7.26) does not hold true in general); remark also that we directly have (7.31) by (7.78b).

The original potential approximation $\bar{u}_h$ has low regularity (it is only piecewise constant in the lowest-order case $k' = 0$); local postprocessing is usually employed to improve it. We now present its general form following Arnold and Brezzi [12] and Arbogast and Chen [10].

The simplest case $k' = 0$ is treated in detail in Section 8.3.5 below. For each couple $\mathbf{V}_h \times Q_h$, there exists a piecewise polynomial space M_h such that $u_h \in M_h$ can be prescribed by

$$\Pi_{Q_h(K)}(u_h|_K) = \bar{u}_h|_K \qquad \forall K \in \mathcal{T}_h, \tag{7.79a}$$

$$\Pi_{\mathbf{V}_h(K)}((-\nabla u_h)|_K) = \boldsymbol{\sigma}_h|_K \qquad \forall K \in \mathcal{T}_h, \tag{7.79b}$$

where $\Pi_{Q_h(K)}$ is the $L^2(K)$-orthogonal projection onto $Q_h(K)$ and $\Pi_{\mathbf{V}_h(K)}$ is the $[L^2(K)]^2$-orthogonal projection onto $\mathbf{V}_h(K)$. Plugging (7.79) into (7.78a), it follows

$$-(\nabla u_h, \mathbf{v}_h) - (u_h, \nabla{\cdot}\mathbf{v}_h) = 0 \qquad \forall \mathbf{v}_h \in \mathbf{V}_h.$$

An immediate consequence of the Green theorem and the structure of $\mathbf{V}_h$ is that

$$\langle [\![u_h]\!], v_h \rangle_e = 0 \qquad \forall e \in \mathcal{E}_h,\ \forall v_h \in \mathbf{V}_h{\cdot}\mathbf{n}_e(e), \tag{7.80}$$

i.e., the jumps of u_h are orthogonal to all polynomials from $\mathbf{V}_h{\cdot}\mathbf{n}$. We let k denote the polynomial degree of functions in M_h, so that u_h, as throughout this section, is a k-th degree piecewise polynomial. With respect to the present a posteriori analysis, the crucial feature is that (7.80) implies (7.50).

For u_h from (7.79), the upper bound of Theorem 7.8.1 holds true, with $\boldsymbol{\sigma}_h$ obtained directly from (7.78) and s_h from Definition 7.10.1 or from Remark 7.10.4. The local lower bound (7.63) holds true but (7.61) cannot be verified, as $\boldsymbol{\sigma}_h$ was not derived from u_h by Definition 7.9.1. This, fortunately, is not obstructive, as $\|\nabla u_h + \boldsymbol{\sigma}_h\|$ by (7.79b) takes small values and can be seen as a numerical quadrature (it is actually zero for $k' = 0$ and the postprocessing of [98]). Alternatively, proceeding as in [100], we may estimate simultaneously the error in both the flux and potential approximations $\boldsymbol{\sigma}_h$ and u_h. This yields

$$\begin{aligned}\|\nabla(u - u_h)\|^2 + \|\nabla u + \boldsymbol{\sigma}_h\|^2 \le{}& \sum_{K\in\mathcal{T}_h} \Big(\|\nabla u_h + \boldsymbol{\sigma}_h\|_K + \frac{h_K}{\pi}\|f - \Pi_{Q_h} f\|_K\Big)^2 \\ &+ \sum_{K\in\mathcal{T}_h} \|\nabla(u_h - s_h)\|_K^2 \\ &+ \sum_{K\in\mathcal{T}_h} \|\nabla s_h + \boldsymbol{\sigma}_h\|_K^2 + \sum_{K\in\mathcal{T}_h} \Big(\frac{h_K}{\pi}\|f - \Pi_{Q_h} f\|_K\Big)^2.\end{aligned}$$

The local efficiency result is then derived by using (7.63) for $\|\nabla(u_h - s_h)\|_K$,

$$\|\nabla s_h + \boldsymbol{\sigma}_h\|_K \le \|\nabla(u_h - s_h)\|_K + \|\nabla u_h + \boldsymbol{\sigma}_h\|_K,$$

and

$$\|\nabla u_h + \boldsymbol{\sigma}_h\|_K \le \|\nabla(u - u_h)\|_K + \|\nabla u + \boldsymbol{\sigma}_h\|_K.$$

Thus, also in this case, we can summarize:

Theorem 7.13.9 (Application to the MFE method)**.** *All the results of this section hold for the mixed finite element solution u_h obtained via (7.79) from $(\boldsymbol{\sigma}_h, \bar{u}_h)$ of Definition 7.13.8. The flux $\boldsymbol{\sigma}_h$ is not constructed by Definition 7.9.1 but comes directly from (7.78), whereas s_h is constructed following Definition 7.10.1. For this, the corresponding space $W_h := \mathbb{P}_{k+1}(\mathcal{T}_h) \cap H_0^1(\Omega)$ needs to be chosen.*

h p	$\|\nabla(u-u_h)\|$	$\|u-u_h\|_{\mathrm{J}}$	$\|u-u_h\|_{\mathrm{DG}}$	$\|\nabla u_h+\boldsymbol{\sigma}_h\|$	$\|\nabla(u_h-s_h)\|$	η_{osc}	η	η_{DG}	I^{eff}	$I^{\mathrm{eff}}_{\mathrm{DG}}$
h_0 1	1.21E+00	4.61E-02	1.21E+00	1.24E+00	1.07E-01	5.56E-02	1.30E+00	1.30E+00	1.07	1.07
$\frac{h_0}{2}$	6.18E-01	1.52E-02	6.19E-01	6.38E-01	5.09E-02	7.02E-03	6.47E-01	6.47E-01	1.05	1.05
	(0.97)	(1.60)	(0.97)	(0.96)	(1.07)	(2.99)	(1.01)	(1.01)		
$\frac{h_0}{4}$	3.12E-01	4.99E-03	3.12E-01	3.22E-01	2.43E-02	8.80E-04	3.24E-01	3.24E-01	1.04	1.04
	(0.99)	(1.61)	(0.99)	(0.99)	(1.07)	(3.00)	(1.00)	(1.00)		
$\frac{h_0}{8}$	1.56E-01	1.68E-03	1.56E-01	1.61E-01	1.18E-02	1.10E-04	1.62E-01	1.62E-01	1.04	1.04
	(1.00)	(1.57)	(1.00)	(1.00)	(1.05)	(3.00)	(1.00)	(1.00)		
h_0 2	1.50E-01	1.49E-02	1.51E-01	1.49E-01	2.76E-02	5.10E-03	1.56E-01	1.57E-01	1.04	1.04
$\frac{h_0}{2}$	3.85E-02	4.03E-03	3.87E-02	3.83E-02	7.99E-03	3.22E-04	3.94E-02	3.96E-02	1.03	1.03
	(1.96)	(1.88)	(1.96)	(1.96)	(1.79)	(3.98)	(1.98)	(1.98)		
$\frac{h_0}{4}$	9.70E-03	1.04E-03	9.75E-03	9.68E-03	2.12E-03	2.02E-05	9.93E-03	9.98E-03	1.02	1.02
	(1.99)	(1.96)	(1.99)	(1.98)	(1.92)	(4.00)	(1.99)	(1.99)		
$\frac{h_0}{8}$	2.43E-03	2.61E-04	2.45E-03	2.43E-03	5.42E-04	1.26E-06	2.49E-03	2.50E-03	1.02	1.02
	(1.99)	(1.99)	(1.99)	(1.99)	(1.96)	(4.00)	(1.99)	(1.99)		
h_0 3	1.32E-02	6.58E-04	1.32E-02	1.29E-02	2.52E-03	3.58E-04	1.35E-02	1.35E-02	1.03	1.03
$\frac{h_0}{2}$	1.67E-03	5.46E-05	1.68E-03	1.65E-03	3.13E-04	1.13E-05	1.70E-03	1.70E-03	1.01	1.01
	(2.98)	(3.59)	(2.98)	(2.97)	(3.01)	(4.99)	(3.00)	(3.00)		
$\frac{h_0}{4}$	2.11E-04	4.48E-06	2.11E-04	2.09E-04	3.83E-05	3.53E-07	2.12E-04	2.12E-04	1.01	1.01
	(2.99)	(3.61)	(2.99)	(2.99)	(3.03)	(5.00)	(3.00)	(3.00)		
$\frac{h_0}{8}$	2.64E-05	3.75E-07	2.64E-05	2.61E-05	4.69E-06	1.10E-08	2.66E-05	2.66E-05	1.01	1.01
	(3.00)	(3.58)	(3.00)	(3.00)	(3.03)	(5.00)	(3.00)	(3.00)		
h_0 4	9.36E-04	8.96E-05	9.40E-04	9.05E-04	2.41E-04	2.12E-05	9.57E-04	9.61E-04	1.02	1.02
$\frac{h_0}{2}$	5.93E-05	6.15E-06	5.96E-05	5.77E-05	1.68E-05	3.36E-07	6.04E-05	6.07E-05	1.02	1.02
	(3.98)	(3.86)	(3.98)	(3.97)	(3.84)	(5.98)	(3.99)	(3.98)		
$\frac{h_0}{4}$	3.72E-06	3.98E-07	3.74E-06	3.63E-06	1.10E-06	5.31E-09	3.80E-06	3.82E-06	1.02	1.02
	(3.99)	(3.95)	(3.99)	(3.99)	(3.94)	(5.98)	(3.99)	(3.99)		
$\frac{h_0}{8}$	2.33E-07	2.52E-08	2.34E-07	2.27E-07	7.02E-08	8.30E-11	2.38E-07	2.39E-07	1.02	1.02
	(4.00)	(3.98)	(4.00)	(4.00)	(3.97)	(6.00)	(4.00)	(3.99)		
h_0 5	5.41E-05	3.04E-06	5.42E-05	5.22E-05	1.38E-05	1.06E-06	5.50E-05	5.50E-05	1.02	1.02
$\frac{h_0}{2}$	1.70E-06	6.44E-08	1.70E-06	1.65E-06	4.39E-07	9.35E-09	1.72E-06	1.72E-06	1.01	1.01
	(4.99)	(5.56)	(5.00)	(4.98)	(4.98)	(6.82)	(5.00)	(5.00)		
$\frac{h_0}{4}$	5.32E-08	1.34E-09	5.32E-08	5.19E-08	1.40E-08	7.67E-11	5.38E-08	5.38E-08	1.01	1.01
	(5.00)	(5.59)	(5.00)	(4.99)	(4.97)	(6.93)	(5.00)	(5.00)		
$\frac{h_0}{8}$	1.66E-09	2.83E-11	1.66E-09	1.62E-09	4.41E-10	5.99E-13	1.68E-09	1.68E-09	1.01	1.01
	(5.00)	(5.57)	(5.00)	(5.00)	(4.99)	(7.00)	(5.00)	(5.00)		

Table 7.1: Numerical results for a smooth solution $\sin(2\pi x)\sin(2\pi y)$ on a unit square and the incomplete interior penalty discontinuous Galerkin method

7.14 Numerical experiments

The following numerical experiment has been kindly calculated by V. Dolejší (Charles University in Prague) and reported in [55]. We consider problem (7.1a) with $\Omega = (0,1)\times(0,1)$ and the right-hand side f such that the exact solution is $u(x,y) = \sin(2\pi x)\sin(2\pi y)$. The discretization is performed via the incomplete interior penalty discontinuous Galerkin method (7.69) with $\theta = 0$ and $\alpha = 20$ (following Dolejší [43]), where we vary the polynomial degree k between 1 and 5. We consider an unstructured triangular mesh of Ω with the initial mesh size $h_0 := 0.168$ that we refine uniformly (every triangle is divided into 4 congruent triangles) three times. The equilibrated flux $\boldsymbol{\sigma}_h$ is obtained via Definition 7.9.1 and the potential s_h via Definition 7.10.1. In both cases, we consider Raviart–Thomas equilibrations of degree k, $\mathbf{V}_h \times Q_h := \mathbf{RT}_k \times \mathbb{P}_k(\mathcal{T}_h)$.

Table 7.1 reports the energy seminorm $\|\nabla(u-u_h)\|$, the jump seminorm $\|u-u_h\|_{\mathrm{J}}^2 :=$

$\sum_{e\in\mathcal{E}_h} h_e^{-1}\|\Pi_e^0[\![u-u_h]\!]\|_e^2$, the full DG norm $\|u-u_h\|_{\mathrm{DG}}^2 := \|\nabla(u-u_h)\|^2 + \|u-u_h\|_{\mathrm{J}}^2$, the estimator η corresponding to (7.24), the full DG estimator $\eta_{\mathrm{DG}}^2 := \eta^2 + \|u_h\|_{\mathrm{J}}^2$ of (7.73), as well as the individual estimators $\|\nabla u_h + \boldsymbol{\sigma}_h\|$, $\|\nabla(u_h - s_h)\|$, and the data oscillation $\eta_{\mathrm{osc}}^2 := \sum_{K\in\mathcal{T}_h}\big(\frac{h_K}{\pi}\|f-\nabla\cdot\boldsymbol{\sigma}_h\|_K\big)^2$. The table also reports the effectivity indices (overestimation factors) $I^{\mathrm{eff}} := \frac{\eta}{\|\nabla(u-u_h)\|}$ and $I_{\mathrm{DG}}^{\mathrm{eff}} := \frac{\eta_{\mathrm{DG}}}{\|u-u_h\|_{\mathrm{DG}}}$ and the corresponding experimental orders of convergence (in parentheses). As predicted by the theory, the estimators η and η_{DG} deliver guaranteed upper bounds on the respective errors, with the effectivity indices robust with respect to the polynomial degree k. Moreover, we experimentally observe asymptotic exactness for this smooth solution case.

Chapter 8

The Laplace equation: complements and different approaches

Various complements to Chapter 7 are treated here. We first in Section 8.1 present the extension to general inhomogeneous Dirichlet and Neumann boundary conditions. Next, Section 8.2 discusses another common a posteriori theory, the so-called residual based one. Then, a simplified version of the equilibrated flux and potential reconstruction theory of Chapter 7 and Section 8.1 is presented in Section 8.3. Herein, no local mixed finite element problems of type (7.27) and (7.37) need to be solved: the degrees of freedom of the reconstructions s_h and $\boldsymbol{\sigma}_h$ are directly prescribed. A link between the equilibration- and residual-based estimates is also made. Finally, Section 8.4 presents a numerical assessment of the simplified prescription equilibrated flux estimates.

8.1 Inhomogeneous Dirichlet and Neumann boundary conditions

Let $\partial\Omega$ be divided into two simply connected parts Γ_D and Γ_N with disjoint interiors. As in Chapter 6, we will distinguish two different cases: either $\Gamma_\mathrm{N} = \partial\Omega$ and $\Gamma_\mathrm{D} = \emptyset$ (Neumann boundary condition on the whole boundary), or $|\Gamma_\mathrm{D}| > 0$ (Dirichlet boundary condition on a set of nonzero $(d-1)$-dimensional measure). Let $f \in L^2(\Omega)$, $u_\mathrm{D} \in H^1(\Gamma_\mathrm{D})$, and $\sigma_\mathrm{N} \in L^2(\Gamma_\mathrm{N})$. In the pure Neumann case, we need the Neumann compatibility condition to be satisfied:

$$\langle \sigma_\mathrm{N}, 1 \rangle_{\Gamma_\mathrm{N}} = (f, 1). \tag{8.1}$$

We consider the Laplace equation (1.1a) equipped with general boundary conditions: find $u : \Omega \to \mathbb{R}$, of mean value zero in the pure Neumann case, such that

$$-\Delta u = f \qquad \text{in } \Omega, \tag{8.2a}$$
$$-\nabla u \cdot \mathbf{n}_\Omega = \sigma_\mathrm{N} \qquad \text{on } \Gamma_\mathrm{N}, \tag{8.2b}$$
$$u = u_\mathrm{D} \qquad \text{on } \Gamma_\mathrm{D}. \tag{8.2c}$$

8.1.1 Variational formulation

Let $H^1_*(\Omega)$ stand for the space of all functions from $H^1(\Omega)$ with zero mean value in the pure Neumann case and for all functions from $H^1(\Omega)$ with zero trace on Γ_D in the Neumann–Dirichlet case. Similarly, $H^1_{*,\mathrm{D}}(\Omega)$ equals $H^1_*(\Omega)$ in the pure Neumann case and denotes all

functions from $H^1(\Omega)$ with trace on Γ_D equal to u_D in the Neumann–Dirichlet case. The variational formulation of (8.2) reads:

Definition 8.1.1 (Variational formulation of (8.2)). *Find $u \in H^1_{*,\mathrm{D}}(\Omega)$ such that*

$$(\nabla u, \nabla v) = (f, v) - \langle \sigma_\mathrm{N}, v\rangle_{\Gamma_\mathrm{N}} \qquad \forall v \in H^1_*(\Omega). \tag{8.3}$$

The existence and uniqueness of a solution of (8.3) is still ensured by the Riesz representation theorem (or by the Lax–Milgram theorem) and an appropriate modification of Theorem 7.1.3 still holds.

8.1.2 Some additional notation

Let $u_h \in H^1(\mathcal{T}_h)$ be the approximate solution as in Section 7.2; we suppose here $(u_h, 1) = 0$ to comply with the mean value condition in the pure Neumann case. Recall that $\mathcal{E}_h^\mathrm{ext}$ stands for the faces lying on the boundary of Ω; we suppose that the interior of each boundary face lies entirely either in Γ_D or Γ_N and denote the corresponding subsets of $\mathcal{E}_h^\mathrm{ext}$ by $\mathcal{E}_h^{\mathrm{ext,D}}$ and $\mathcal{E}_h^{\mathrm{ext,N}}$, respectively. Similarly, $\mathcal{V}_h^{\mathrm{ext,D}}$ ($\mathcal{V}_h^{\mathrm{ext,N}}$) stand for the mesh vertices which lie on some Dirichlet (Neumann) boundary face. Note that $\mathcal{V}_h^{\mathrm{ext,D}} \cap \mathcal{V}_h^{\mathrm{ext,N}}$ is not empty unless $\Gamma_\mathrm{D} = \partial\Omega$ or $\Gamma_\mathrm{N} = \partial\Omega$; vertices on the interface between Γ_D and Γ_N lie both in $\mathcal{V}_h^{\mathrm{ext,D}}$ and $\mathcal{V}_h^{\mathrm{ext,N}}$. Finally, those faces of an element $K \in \mathcal{T}_h$ that lie in Γ_D (Γ_N) are denoted by $\mathcal{E}_K^\mathrm{D}$ ($\mathcal{E}_K^\mathrm{N}$). Recall that $\mathrm{R}_{\frac{\pi}{2}} := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ is the matrix of rotation by $\frac{\pi}{2}$. We use the convention $\mathrm{R}_{\frac{\pi}{2}}\mathbf{n}_\Omega = \mathbf{t}_\Omega$ for the link between exterior normal and tangential vectors, and similarly on subdomains of Ω. In addition to the weak gradient and divergence given respectively by ∇ and $\nabla\cdot$, $\mathrm{R}_{\frac{\pi}{2}}\nabla$ stands for the weak curl in two space dimensions (the rotated gradient), given by $\mathrm{R}_{\frac{\pi}{2}}\nabla v = (-\partial_y v, \partial_x v)$ for $v \in H^1(\mathcal{T}_h)$.

8.1.3 Potential and flux reconstructions

The potential and flux reconstructions are here defined as follows:

Definition 8.1.2 (Potential reconstruction). *Let $u_h \in H^1(\mathcal{T}_h)$, $(u_h, 1) = 0$ in the pure Neumann case, be the approximate solution. We will call the* potential reconstruction *any function s_h constructed from u_h which satisfies*

$$s_h \in H^1(\Omega) \cap C^0(\overline{\Omega}), \tag{8.4a}$$

$$(s_h, 1) = 0 \qquad \text{in the pure Neumann case,} \tag{8.4b}$$

$$s_h(\mathbf{a}) = u_\mathrm{D}(\mathbf{a}) \qquad \forall \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext,D}} \qquad \text{in the Neumann–Dirichlet case.} \tag{8.4c}$$

Definition 8.1.3 (Equilibrated flux reconstruction). *We will call the* equilibrated flux reconstruction *any function $\boldsymbol{\sigma}_h$ constructed from u_h which satisfies*

$$\boldsymbol{\sigma}_h \in \mathbf{H}(\mathrm{div}, \Omega), \tag{8.5a}$$

$$(\nabla\cdot\boldsymbol{\sigma}_h, 1)_K = (f, 1)_K \qquad \forall K \in \mathcal{T}_h, \tag{8.5b}$$

$$\boldsymbol{\sigma}_h\cdot\mathbf{n}_e|_e \in L^2(e) \qquad \forall e \in \mathcal{E}_h^{\mathrm{ext,N}}, \tag{8.5c}$$

$$\langle \boldsymbol{\sigma}_h\cdot\mathbf{n}_\Omega, 1\rangle_e = \langle \sigma_\mathrm{N}, 1\rangle_e \qquad \forall e \in \mathcal{E}_h^{\mathrm{ext,N}}. \tag{8.5d}$$

The continuity of s_h imposed in (8.4a) is needed in (8.4c) to take punctual values; similarly, the requirement on normal trace of $\boldsymbol{\sigma}_h$ to belong to $L^2(e)$ in (8.5c) is needed in (8.5d). In practice, s_h and $\boldsymbol{\sigma}_h$ will again be constructed in finite-dimensional (piecewise polynomial) spaces as in Chapter 7.

8.1.4 A general a posteriori error estimate

Theorem 7.8.1 from Chapter 7 for problem (8.2) takes here the following form:

Theorem 8.1.4 (A general a posteriori error estimate for (8.2))**.** *Let u be the weak solution given by Definition 8.1.1. Let $u_h \in H^1(\mathcal{T}_h)$ be an arbitrary approximation, with $(u_h, 1) = 0$ in the pure Neumann case. Let s_h be a potential reconstruction in the sense of Definition 8.1.2 and $\boldsymbol{\sigma}_h$ an equilibrated flux reconstruction in the sense of Definition 8.1.3. Then*

$$\begin{aligned}
&\|\nabla(u-u_h)\|^2 \\
&\le \sum_{K\in\mathcal{T}_h} \left(\underbrace{\|\nabla u_h + \boldsymbol{\sigma}_h\|_K}_{\text{constitutive rel.}} + \underbrace{\frac{h_K}{\pi}\|f - \nabla\cdot\boldsymbol{\sigma}_h\|_K}_{\text{equilibrium}} + \sum_{e\in\mathcal{E}_K^{\mathrm{N}}} \Big(\underbrace{\bar{C}_{\mathrm{t},K,e} h_e^{\frac{1}{2}} \|\boldsymbol{\sigma}_h\cdot\mathbf{n}_\Omega - \sigma_{\mathrm{N}}\|_e}_{\text{Neumann BC}} \Big) \right)^2 \\
&\quad + \sum_{K\in\mathcal{T}_h} \Big(\underbrace{\|\nabla(u_h - s_h)\|_K}_{\text{pot. nonconformity}} + \underbrace{\min_{\substack{v\in H^1(K),\\ v|_{\partial K\cap\Gamma_{\mathrm{D}}} = u_{\mathrm{D}} - s_h \\ v|_{\partial K\setminus\Gamma_{\mathrm{D}}} = 0}} \|\nabla v\|_K}_{\text{Dirichlet BC, if } |\partial K\cap\Gamma_{\mathrm{D}}|>0} \Big)^2 .
\end{aligned} \tag{8.6}$$

Proof. As in the proof of Theorem 7.8.1, let $s \in H^1_{*,\mathrm{D}}(\Omega)$ be given by

$$(\nabla s, \nabla v) = (\nabla u_h, \nabla v) \qquad \forall v \in H^1_*(\Omega). \tag{8.7}$$

Then again, we can write the Pythagorean equality

$$\|\nabla(u-u_h)\|^2 = \|\nabla(u-s)\|^2 + \|\nabla(s-u_h)\|^2. \tag{8.8}$$

The potential nonconformity term satisfies

$$\|\nabla(s-u_h)\|^2 = \min_{w\in H^1_{*,\mathrm{D}}(\Omega)} \|\nabla(w-u_h)\|^2.$$

We cannot bound it directly by

$$\min_{w\in H^1_{*,\mathrm{D}}(\Omega)} \|\nabla(w-u_h)\|^2 \le \|\nabla(s_h-u_h)\|^2,$$

except for the pure Neumann case (thanks to condition (8.4b)) or unless $s_h|_{\Gamma_{\mathrm{D}}} = u_{\mathrm{D}}$ in the Neumann–Dirichlet case. Consider the Neumann–Dirichlet case with $s_h|_{\Gamma_{\mathrm{D}}} \ne u_{\mathrm{D}}$. Proceeding as in [67, Section 4.1], we have

$$\begin{aligned}
\min_{w\in H^1_{*,\mathrm{D}}(\Omega)} \|\nabla(w-u_h)\|^2 &\le \min_{w\in H^1_{*,\mathrm{D}}(\Omega)} \sum_{K\in\mathcal{T}_h} \big(\|\nabla(w-s_h)\|_K + \|\nabla(s_h-u_h)\|_K \big)^2 \\
&\le \sum_{K\in\mathcal{T}_h} \left(\min_{\substack{w\in H^1(K),\\ w|_{\partial K\cap\Gamma_{\mathrm{D}}} = u_{\mathrm{D}} \\ w|_{\partial K\setminus\Gamma_{\mathrm{D}}} = s_h}} \|\nabla(w-s_h)\|_K + \|\nabla(s_h-u_h)\|_K \right)^2 .
\end{aligned}$$

The first estimate above follows by localization on mesh elements and by the triangle inequality. The second one is then possible by constraining the global minimum over all $w \in H^1_{*,\mathrm{D}}(\Omega)$ to elementwise minima over functions $w \in H^1(K)$ with values on ∂K fixed respectively to u_{D} or s_h, thanks to conditions (8.4a) and (8.4c). Note that the first terms in the last sum are

only nonzero on elements of the Dirichlet boundary. Thus we have come to the second sum of (8.6).

The first term in (8.8) allows the following equivalent rewriting, relying on the fact that $(u-s)\in H^1_*(\Omega)$:

$$\begin{aligned}\|\nabla(u-s)\| &= \sup_{\varphi\in H^1_*(\Omega);\,\|\nabla\varphi\|=1}(\nabla(u-s),\nabla\varphi), && \text{dual norm, cf. (7.8)}\\ &= \sup_{\varphi\in H^1_*(\Omega);\,\|\nabla\varphi\|=1}(\nabla(u-u_h),\nabla\varphi), && \text{by (8.7)}\\ &= \sup_{\varphi\in H^1_*(\Omega);\,\|\nabla\varphi\|=1}\{(f,\varphi)-\langle\sigma_{\mathrm{N}},\varphi\rangle_{\Gamma_{\mathrm{N}}}-(\nabla u_h,\nabla\varphi)\}. && \text{by (8.3)}\end{aligned}$$

Remark that the last expression above is nothing but the dual norm of the residual of u_h, $\mathcal{R}(u_h)\in(H^1_*(\Omega))'$,

$$\langle\mathcal{R}(u_h),\varphi\rangle_{(H^1_*(\Omega))',H^1_*(\Omega)}:=(f,\varphi)-\langle\sigma_{\mathrm{N}},\varphi\rangle_{\Gamma_{\mathrm{N}}}-(\nabla u_h,\nabla\varphi)\qquad\varphi\in H^1_*(\Omega),$$

compare with Definition 7.7.1. Let now $\varphi\in H^1_*(\Omega)$ with $\|\nabla\varphi\|=1$ be fixed. Adding and subtracting $(\boldsymbol{\sigma}_h,\nabla\varphi)$, where $\boldsymbol{\sigma}_h$ is the equilibrated flux reconstruction in the sense of Definition 8.1.3, and using the Green theorem, we have

$$\langle\mathcal{R}(u_h),\varphi\rangle_{(H^1_*(\Omega))',H^1_*(\Omega)}=(f-\nabla\cdot\boldsymbol{\sigma}_h,\varphi)+\langle\boldsymbol{\sigma}_h\cdot\mathbf{n}_\Omega-\sigma_{\mathrm{N}},\varphi\rangle_{\Gamma_{\mathrm{N}}}-(\nabla u_h+\boldsymbol{\sigma}_h,\nabla\varphi).$$

The first and last terms above are treated exactly as in the proof of Theorem 7.8.1, using in particular the equilibration (8.5b). For the middle one, we have

$$\begin{aligned}\langle\boldsymbol{\sigma}_h\cdot\mathbf{n}_\Omega-\sigma_{\mathrm{N}},\varphi\rangle_{\Gamma_{\mathrm{N}}} &= \sum_{K\in\mathcal{T}_h}\sum_{e\in\mathcal{E}_K^{\mathrm{N}}}\langle\boldsymbol{\sigma}_h\cdot\mathbf{n}_\Omega-\sigma_{\mathrm{N}},\varphi\rangle_e\\ &= \sum_{K\in\mathcal{T}_h}\sum_{e\in\mathcal{E}_K^{\mathrm{N}}}\langle\boldsymbol{\sigma}_h\cdot\mathbf{n}_\Omega-\sigma_{\mathrm{N}},\varphi-\varphi_e\rangle_e && \text{by (8.5d)}\\ &\le \sum_{K\in\mathcal{T}_h}\sum_{e\in\mathcal{E}_K^{\mathrm{N}}}\big\{\|\boldsymbol{\sigma}_h\cdot\mathbf{n}_\Omega-\sigma_{\mathrm{N}}\|_e\bar{C}_{\mathrm{t},K,e}h_e^{\frac12}\|\nabla\varphi\|_K\big\}, && \text{CS and (4.22b)}\end{aligned}$$

where φ_e is the mean value of the function φ on the face e. Combining these results while using the Cauchy–Schwarz inequality and the constraint $\|\nabla\varphi\|=1$ gives the first sum of (8.6). □

8.1.5 Inhomogeneous Dirichlet boundary condition

The expression for the general Dirichlet boundary condition error from Theorem 8.1.4 is not fully computable. We now give a computable estimate following [29, Theorem 5.1].

Theorem 8.1.5 (Inhomogeneous Dirichlet boundary condition estimate)**.** *Let $|\Gamma_{\mathrm{D}}|>0$ and let $K\in\mathcal{T}_h$ such that $|\partial K\cap\Gamma_{\mathrm{D}}|>0$ be given. Let $\mathbf{x}_K$ denote the barycenter of K. For each $e\in\mathcal{E}_K^{\mathrm{D}}$, consider the polar coordinates r,θ centered at $\mathbf{x}_K$, where the simplex K_e given by the face e and the point $\mathbf{x}_K$ is described by $\theta\in[\alpha_e,\beta_e]$ and $r\in[0,R_e(\theta)]$; $R_e(\theta)$ is thus the distance of $\mathbf{x}_K$ and $\mathbf{x}_\theta\in e$, see Figure 8.1. Set $g_e(\theta):=(u_{\mathrm{D}}-s_h)(\mathbf{x}_\theta)$ and denote by $'$ the differentiation with respect to θ. Then*

$$\begin{aligned}&\min_{v\in H^1(K),\ \substack{v|_{\partial K\cap\Gamma_{\mathrm{D}}}=u_{\mathrm{D}}-s_h\\ v|_{\partial K\setminus\Gamma_{\mathrm{D}}}=0}}\|\nabla v\|_K\\ &\le\sum_{e\in\mathcal{E}_K^{\mathrm{D}}}\left\{\frac12\int_{\alpha_e}^{\beta_e}\big\{[g_e(\theta)]^2+[(g_e'(\theta)R_e(\theta)-g_e(\theta)R_e'(\theta))/R_e(\theta)]^2\big\}\,\mathrm{d}\theta\right\}^{\frac12}.\end{aligned}\tag{8.9}$$

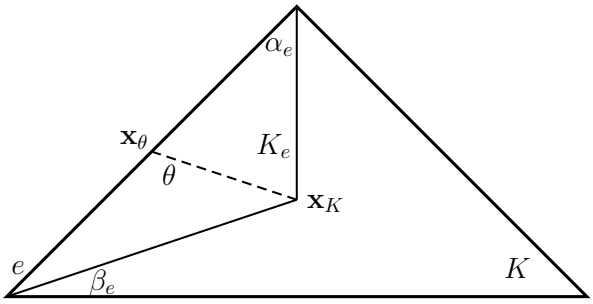


Figure 8.1: Notation for the inhomogeneous Dirichlet boundary condition estimate

Proof. Let $e \in \mathcal{E}_K^{\mathrm{D}}$. Extend the function g_e on the whole subsimplex K_e by $g_e(r,\theta) := g_e(\theta)r/R_e(\theta)$. This is a function given by the Dirichlet boundary misfit $(u_{\mathrm{D}} - s_h)(\mathbf{x}_\theta)$ on the face e, decreasing linearly with respect to r to take the value 0 in $\mathbf{x}_K$. By condition (8.4c), g_e is also zero on $\partial K_e \setminus e$, and thus we can further extend it by 0 to a function $g_e \in H^1(K)$ defined on the whole K. By proceeding similarly for all Dirichlet boundary faces of K, we obtain $g := \sum_{e\in\mathcal{E}_K^{\mathrm{D}}} g_e$, $g \in H^1(K)$, $g|_{\partial K\cap\Gamma_{\mathrm{D}}} = u_{\mathrm{D}} - s_h$, and $g|_{\partial K\setminus\Gamma_{\mathrm{D}}} = 0$. Thus the minimum on the left-hand side of (8.9) can be estimated by $\|\nabla g\|_K$ and further by the triangle inequality by $\sum_{e\in\mathcal{E}_K^{\mathrm{D}}}\|\nabla g_e\|_K$. Finally, in order to easily compute $\|\nabla g_e\|_K$, we develop

$$
\begin{aligned}
\|\nabla g_e\|_K^2 &= \int_K |\nabla g_e(\mathbf{x})|^2 \,\mathrm{d}\mathbf{x} = \int_{\alpha_e}^{\beta_e}\int_0^{R_e(\theta)} |\nabla g_e(r,\theta)|^2 r \,\mathrm{d}r\,\mathrm{d}\theta \\
&= \int_{\alpha_e}^{\beta_e}\int_0^{R_e(\theta)} \{[\partial_r g_e(r,\theta)]^2 + [\partial_\theta g_e(r,\theta)/r]^2\} r \,\mathrm{d}r\,\mathrm{d}\theta \\
&= \int_{\alpha_e}^{\beta_e}\int_0^{R_e(\theta)} \{[g_e(\theta)/R_e(\theta)]^2 + [(g_e'(\theta)R_e(\theta) - g_e(\theta)R_e'(\theta))/R_e^2(\theta)]^2\} r \,\mathrm{d}r\,\mathrm{d}\theta \\
&= \frac{1}{2}\int_{\alpha_e}^{\beta_e} \left\{[g_e(\theta)]^2 + [(g_e'(\theta)R_e(\theta) - g_e(\theta)R_e'(\theta))/R_e(\theta)]^2\right\} \,\mathrm{d}\theta,
\end{aligned}
$$

which finishes the proof. □

This computable estimate is of higher order whenever u_{D} has enough regularity, see the discussions in [67, 29].

8.1.6 Flux reconstruction via local Neumann mixed finite element problems

Definition 7.9.1 for the general boundary conditions (8.2b)–(8.2c) takes the following form:

Definition 8.1.6 (Flux $\boldsymbol{\sigma}_h$, inhomogeneous boundary conditions)**.** *Let u_h satisfy the hat-function orthogonality*

$$
(\nabla u_h, \nabla\psi_{\mathbf{a}})_{\omega_{\mathbf{a}}} = (f,\psi_{\mathbf{a}})_{\omega_{\mathbf{a}}} - \langle\sigma_{\mathrm{N}},\psi_{\mathbf{a}}\rangle_{\Gamma_{\mathrm{N}}} \qquad \forall \mathbf{a} \in \mathcal{V}_h \setminus \mathcal{V}_h^{\mathrm{ext,D}}. \tag{8.10}
$$

For each $\mathbf{a} \in \mathcal{V}_h$, prescribe $\boldsymbol{\varsigma}_h^{\mathbf{a}} \in \mathbf{V}_{h,\mathrm{N}}^{\mathbf{a}}$ and $\bar{r}_h^{\mathbf{a}} \in Q_h^{\mathbf{a}}$ by solving

$$
\begin{aligned}
(\boldsymbol{\varsigma}_h^{\mathbf{a}}, \mathbf{v}_h)_{\omega_{\mathbf{a}}} - (\bar{r}_h^{\mathbf{a}}, \nabla\cdot\mathbf{v}_h)_{\omega_{\mathbf{a}}} &= -(\psi_{\mathbf{a}}\nabla u_h, \mathbf{v}_h)_{\omega_{\mathbf{a}}} && \forall \mathbf{v}_h \in \mathbf{V}_h^{\mathbf{a}}, && \text{(8.11a)}\\
(\nabla\cdot\boldsymbol{\varsigma}_h^{\mathbf{a}}, q_h)_{\omega_{\mathbf{a}}} &= (\psi_{\mathbf{a}} f - \nabla\psi_{\mathbf{a}}\cdot\nabla u_h, q_h)_{\omega_{\mathbf{a}}} && \forall q_h \in Q_h^{\mathbf{a}} && \text{(8.11b)}
\end{aligned}
$$

with the spaces

$$\begin{aligned} \mathbf{V}_{h,\mathrm{N}}^{\mathbf{a}} &:= \mathbf{V}_h^{\mathbf{a}} := \{\mathbf{v}_h \in \mathbf{V}_h(\omega_{\mathbf{a}});\ \mathbf{v}_h{\cdot}\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \ \textit{on}\ \partial\omega_{\mathbf{a}}\}, \\ Q_h^{\mathbf{a}} &:= \{q_h \in Q_h(\omega_{\mathbf{a}});\ (q_h, 1)_{\omega_{\mathbf{a}}} = 0\}, \end{aligned} \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}, \tag{8.12a}$$

$$\begin{aligned} \mathbf{V}_h^{\mathbf{a}} &:= \{\mathbf{v}_h \in \mathbf{V}_h(\omega_{\mathbf{a}});\ \mathbf{v}_h{\cdot}\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \ \textit{on}\ \partial\omega_{\mathbf{a}} \setminus \partial\Omega, \\ &\qquad \mathbf{v}_h{\cdot}\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \ \textit{on}\ \partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{N}}\}, \\ \mathbf{V}_{h,\mathrm{N}}^{\mathbf{a}} &:= \{\mathbf{v}_h \in \mathbf{V}_h(\omega_{\mathbf{a}});\ \mathbf{v}_h{\cdot}\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \ \textit{on}\ \partial\omega_{\mathbf{a}} \setminus \partial\Omega, \\ &\qquad \mathbf{v}_h{\cdot}\mathbf{n}_{\omega_{\mathbf{a}}} = \Pi_{\mathbf{V}_h{\cdot}\mathbf{n}}(\psi_{\mathbf{a}}\sigma_{\mathrm{N}}) \ \textit{on}\ \partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{N}}\}, \end{aligned} \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}, \tag{8.12b}$$

$$Q_h^{\mathbf{a}} := \{q_h \in Q_h(\omega_{\mathbf{a}});\ (q_h, 1)_{\omega_{\mathbf{a}}} = 0\}, \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}} \setminus \mathcal{V}_h^{\mathrm{ext,D}}, \tag{8.12c}$$

$$Q_h^{\mathbf{a}} := Q_h(\omega_{\mathbf{a}}), \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext,D}}. \tag{8.12d}$$

Then, set

$$\boldsymbol{\sigma}_h := \sum_{\mathbf{a} \in \mathcal{V}_h} \boldsymbol{\varsigma}_h^{\mathbf{a}}. \tag{8.13}$$

The above local problems only differ from those of Definition 7.9.1 on vertices which lie on the Neumann boundary Γ_{N}. There an inhomogeneous Neumann boundary condition is encoded in the space $\mathbf{V}_{h,\mathrm{N}}^{\mathbf{a}}$ of (8.12b): $\boldsymbol{\varsigma}_h^{\mathbf{a}}{\cdot}\mathbf{n}_\Omega$ on $\partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{N}}$ is imposed by the polynomial projection of $\psi_{\mathbf{a}}\sigma_{\mathrm{N}}$. Problem (8.11) is pure Neumann when $\partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{N}} = \partial\omega_{\mathbf{a}} \cap \partial\Omega$, i.e., the whole boundary of $\omega_{\mathbf{a}}$ lying on $\partial\Omega$ is the Neumann boundary. This happens if and only if $\mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}} \setminus \mathcal{V}_h^{\mathrm{ext,D}}$. The Neumann compatibility condition then requests

$$(\psi_{\mathbf{a}} f - \nabla\psi_{\mathbf{a}}{\cdot}\nabla u_h, 1)_{\omega_{\mathbf{a}}} = \langle \Pi_{\mathbf{V}_h{\cdot}\mathbf{n}}(\psi_{\mathbf{a}}\sigma_{\mathrm{N}}), 1\rangle_{\partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{N}}}.$$

Noting that $\langle \Pi_{\mathbf{V}_h{\cdot}\mathbf{n}}(\psi_{\mathbf{a}}\sigma_{\mathrm{N}}), 1\rangle_{\partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{N}}} = \langle \psi_{\mathbf{a}}\sigma_{\mathrm{N}}, 1\rangle_{\Gamma_{\mathrm{N}}}$, this is nothing but (8.10) for $\mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}} \setminus \mathcal{V}_h^{\mathrm{ext,D}}$. Shall $|\partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{D}}| > 0$, we have a local Neumann–Dirichlet problem, with the normal trace of $\boldsymbol{\varsigma}_h^{\mathbf{a}}$ not prescribed on $\partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{D}}$. Lemma 7.9.2 still holds. Moreover, we have:

Lemma 8.1.7 (Normal flux of $\boldsymbol{\sigma}_h$ on Γ_{N})**.** *There holds*

$$(\boldsymbol{\sigma}_h{\cdot}\mathbf{n}_\Omega)|_{\Gamma_{\mathrm{N}}} = \Pi_{\mathbf{V}_h{\cdot}\mathbf{n}}(\sigma_{\mathrm{N}}) \quad \textit{i.e.} \quad \langle \boldsymbol{\sigma}_h{\cdot}\mathbf{n}_\Omega - \sigma_{\mathrm{N}}, v_h\rangle_e = 0 \qquad \forall v_h \in (\mathbf{V}_h{\cdot}\mathbf{n})|_e,\ \forall e \in \mathcal{E}_h^{\mathrm{ext,N}}. \tag{8.14}$$

Thus, in particular, (8.5d) *is satisfied.*

Proof. Let $e \in \mathcal{E}_h^{\mathrm{ext,N}}$ and let v_h be a polynomial on the face e from the discrete normal trace space $(\mathbf{V}_h{\cdot}\mathbf{n})|_e$ (k-degree polynomial for $\mathbf{V}_h$ being the Raviart–Thomas–Nédélec space of degree k, see Section 5.3). Employing that $\boldsymbol{\sigma}_h|_e = \sum_{\mathbf{a} \in \mathcal{V}_e} \boldsymbol{\varsigma}_h^{\mathbf{a}}$ and using the normal trace condition imposed on $\mathbf{V}_{h,\mathrm{N}}^{\mathbf{a}}$ in (8.12b),

$$\langle \boldsymbol{\sigma}_h{\cdot}\mathbf{n}_\Omega, v_h\rangle_e = \sum_{\mathbf{a} \in \mathcal{V}_e} \langle \boldsymbol{\varsigma}_h^{\mathbf{a}}{\cdot}\mathbf{n}_\Omega, v_h\rangle_e = \sum_{\mathbf{a} \in \mathcal{V}_e} \langle \psi_{\mathbf{a}}\sigma_{\mathrm{N}}, v_h\rangle_e = \langle \sigma_{\mathrm{N}}, v_h\rangle_e$$

is easily inferred. □

Finally, the following equivalent of Remark 7.9.4 holds true, following Lemma 6.6.8:

Remark 8.1.8 (Local flux minimization)**.** *Definition 8.1.6 can be equivalently stated as:*

$$\boldsymbol{\varsigma}_h^{\mathbf{a}} := \arg \min_{\mathbf{v}_h \in \mathbf{V}_{h,\mathrm{N}}^{\mathbf{a}},\ \nabla{\cdot}\mathbf{v}_h = \Pi_{Q_h^{\mathbf{a}}}(\psi_{\mathbf{a}} f - \nabla\psi_{\mathbf{a}}{\cdot}\nabla u_h)} \|\psi_{\mathbf{a}}\nabla u_h + \mathbf{v}_h\|_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{a} \in \mathcal{V}_h. \tag{8.15}$$

8.1.7 Potential reconstruction via local Dirichlet finite element problems

We adjust here Definition 7.10.1 for inhomogeneous boundary conditions. We use the equivalent form of Theorem 7.10.3 but remark directly that a primal version is also in place, see (8.21).

Definition 8.1.9 (Potential s_h, inhomogeneous boundary conditions)**.** *For each* $\mathbf{a} \in \mathcal{V}_h$, *prescribe* $\boldsymbol{\varsigma}_h^{\mathbf{a}} \in \mathbf{V}_{h,\mathrm{N}}^{\mathbf{a}}$ *and* $\bar{r}_h^{\mathbf{a}} \in Q_h^{\mathbf{a}}$ *by solving*

$$(\boldsymbol{\varsigma}_h^{\mathbf{a}}, \mathbf{v}_h)_{\omega_{\mathbf{a}}} - (\bar{r}_h^{\mathbf{a}}, \nabla\cdot\mathbf{v}_h)_{\omega_{\mathbf{a}}} = -(\mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_h), \mathbf{v}_h)_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{v}_h \in \mathbf{V}_h^{\mathbf{a}}, \tag{8.16a}$$

$$(\nabla\cdot\boldsymbol{\varsigma}_h^{\mathbf{a}}, q_h)_{\omega_{\mathbf{a}}} = 0 \qquad \forall q_h \in Q_h^{\mathbf{a}}, \tag{8.16b}$$

with the spaces

$$\begin{aligned} \mathbf{V}_{h,\mathrm{N}}^{\mathbf{a}} &:= \mathbf{V}_h^{\mathbf{a}} := \{\mathbf{v}_h \in \mathbf{V}_h(\omega_{\mathbf{a}});\ \mathbf{v}_h\cdot\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \text{ on } \partial\omega_{\mathbf{a}}\}, \\ Q_h^{\mathbf{a}} &:= \{q_h \in Q_h(\omega_{\mathbf{a}});\ (q_h, 1)_{\omega_{\mathbf{a}}} = 0\}, \end{aligned} \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}, \tag{8.17a}$$

$$\begin{aligned} \mathbf{V}_h^{\mathbf{a}} &:= \{\mathbf{v}_h \in \mathbf{V}_h(\omega_{\mathbf{a}});\ \mathbf{v}_h\cdot\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \text{ on } \partial\omega_{\mathbf{a}} \setminus \partial\Omega, \\ &\qquad \mathbf{v}_h\cdot\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \text{ on } \partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{D}}\}, \\ \mathbf{V}_{h,\mathrm{N}}^{\mathbf{a}} &:= \{\mathbf{v}_h \in \mathbf{V}_h(\omega_{\mathbf{a}});\ \mathbf{v}_h\cdot\mathbf{n}_{\omega_{\mathbf{a}}} = 0 \text{ on } \partial\omega_{\mathbf{a}} \setminus \partial\Omega, \\ &\qquad \mathbf{v}_h\cdot\mathbf{n}_{\omega_{\mathbf{a}}} = \Pi_{\mathbf{V}_h\cdot\mathbf{n}}(\nabla(\psi_{\mathbf{a}}u_{\mathrm{D}})\cdot\mathbf{t}_\Omega) \text{ on } \partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{D}}\}, \end{aligned} \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}, \tag{8.17b}$$

$$Q_h^{\mathbf{a}} := \{q_h \in Q_h(\omega_{\mathbf{a}});\ (q_h, 1)_{\omega_{\mathbf{a}}} = 0\}, \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}} \setminus \mathcal{V}_h^{\mathrm{ext,N}}, \tag{8.17c}$$

$$Q_h^{\mathbf{a}} := Q_h(\omega_{\mathbf{a}}), \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext,N}}. \tag{8.17d}$$

Then set

$$-\mathrm{R}_{\frac{\pi}{2}}\nabla s_h^{\mathbf{a}} := \boldsymbol{\varsigma}_h^{\mathbf{a}}, \tag{8.18a}$$

$$s_h^{\mathbf{a}} := 0 \text{ on } \partial\omega_{\mathbf{a}} \setminus \partial\Omega, \tag{8.18b}$$

$$s_h := \sum_{\mathbf{a}\in\mathcal{V}_h} s_h^{\mathbf{a}} \qquad \textit{in the Neumann–Dirichlet case}, \tag{8.18c}$$

$$s_h := \sum_{\mathbf{a}\in\mathcal{V}_h} s_h^{\mathbf{a}} - \left(\sum_{\mathbf{a}\in\mathcal{V}_h} s_h^{\mathbf{a}}, 1\right)|\Omega|^{-1} \qquad \textit{in the pure Neumann case}. \tag{8.18d}$$

Note that for boundary vertices, $\mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}} \setminus \mathcal{V}_h^{\mathrm{ext,N}}$ if and only if $\partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{D}} = \partial\omega_{\mathbf{a}} \cap \partial\Omega$, i.e., the whole boundary of $\omega_{\mathbf{a}}$ lying on $\partial\Omega$ is the Dirichlet boundary. Then (8.16) is a pure Neumann problem: the normal trace of $\boldsymbol{\varsigma}_h^{\mathbf{a}}$ on $\partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{D}}$ is imposed by the (polynomial projection of the) tangential trace of $\nabla(\psi_{\mathbf{a}}u_{\mathrm{D}})$. The Neumann compatibility condition then requests

$$0 = \langle \Pi_{\mathbf{V}_h\cdot\mathbf{n}}(\nabla(\psi_{\mathbf{a}}u_{\mathrm{D}})\cdot\mathbf{t}_\Omega), 1\rangle_{\partial\omega_{\mathbf{a}}\cap\Gamma_{\mathrm{D}}}.$$

This is immediate developing the above right-hand term as

$$\begin{aligned} \langle \Pi_{\mathbf{V}_h\cdot\mathbf{n}}(\nabla(\psi_{\mathbf{a}}u_{\mathrm{D}})\cdot\mathbf{t}_\Omega), 1\rangle_{\partial\omega_{\mathbf{a}}\cap\Gamma_{\mathrm{D}}} &= -\langle \mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_{\mathrm{D}})\cdot\mathbf{n}_\Omega, 1\rangle_{\partial\omega_{\mathbf{a}}\cap\Gamma_{\mathrm{D}}} = -\langle \mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_{\mathrm{D}})\cdot\mathbf{n}_{\omega_{\mathbf{a}}}, 1\rangle_{\partial\omega_{\mathbf{a}}} \\ &= -(\nabla\cdot(\mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_{\mathrm{D}})), 1)_{\omega_{\mathbf{a}}} - (\mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_{\mathrm{D}}), \nabla 1)_{\omega_{\mathbf{a}}} = 0, \end{aligned}$$

for any smooth enough extension u_{D} of the Dirichlet boundary condition u_{D}. Shall $|\partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{N}}| > 0$, we have a local Neumann–Dirichlet problem, with the normal trace of $\boldsymbol{\varsigma}_h^{\mathbf{a}}$ not prescribed on $\partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{N}}$.

As in Section 7.10, the potential reconstruction satisfies (8.4a), and, by (8.18d), the mean value condition (8.4b) trivially follows in the pure Neumann case. Moreover, the treatment of Dirichlet boundary conditions in Definition 8.1.9 is coherent, as the following lemma shows:

Lemma 8.1.10 (Boundary value of s_h on Γ_D)**.** *Condition* (8.4c) *is satisfied. Moreover, there holds*

$$(\nabla s_h\cdot\mathbf{t}_\Omega)|_{\Gamma_\mathrm{D}} = \Pi_{\mathbf{V}_h\cdot\mathbf{n}}(\nabla u_\mathrm{D}\cdot\mathbf{t}_\Omega)|_{\Gamma_\mathrm{D}}. \tag{8.19}$$

Proof. We start by showing (8.19). Let $e\in\mathcal{E}_h^{\mathrm{ext,D}}$ and let v_h be a polynomial on the face e from the discrete normal trace space $(\mathbf{V}_h\cdot\mathbf{n})|_e$. Note that $-\mathrm{R}_{\frac{\pi}{2}}\nabla s_h = \sum_{\mathbf{a}\in\mathcal{V}_h}\boldsymbol{\varsigma}_h^{\mathbf{a}}$, whence $(-\mathrm{R}_{\frac{\pi}{2}}\nabla s_h)|_e = \sum_{\mathbf{a}\in\mathcal{V}_e}\boldsymbol{\varsigma}_h^{\mathbf{a}}|_e$. Using the normal trace condition imposed on $\mathbf{V}_{h,\mathrm{N}}^{\mathbf{a}}$ in (8.17b),

$$\langle -\mathrm{R}_{\frac{\pi}{2}}\nabla s_h\cdot\mathbf{n}_\Omega, v_h\rangle_e = \sum_{\mathbf{a}\in\mathcal{V}_e}\langle\boldsymbol{\varsigma}_h^{\mathbf{a}}\cdot\mathbf{n}_\Omega, v_h\rangle_e = \sum_{\mathbf{a}\in\mathcal{V}_e}\langle\nabla(\psi_{\mathbf{a}}u_\mathrm{D})\cdot\mathbf{t}_\Omega, v_h\rangle_e = \langle\nabla u_\mathrm{D}\cdot\mathbf{t}_\Omega, v_h\rangle_e$$

is easily inferred. Thus (8.19) follows by the fact that $-\mathrm{R}_{\frac{\pi}{2}}\nabla s_h\cdot\mathbf{n}_\Omega = \nabla s_h\cdot\mathbf{t}_\Omega$.

To show (8.4c), we reason as follows: for each $\mathbf{a}\in\mathcal{V}_h^{\mathrm{ext,D}}$, $\nabla s_h^{\mathbf{a}}\mathbf{t}_\Omega$ preserves sidewise moments of $\nabla(\psi_{\mathbf{a}}u_\mathrm{D})\cdot\mathbf{t}_\Omega$ on $\partial\omega_{\mathbf{a}}\cap\Gamma_\mathrm{D}$, i.e., $(\nabla s_h^{\mathbf{a}}\cdot\mathbf{t}_\Omega)|_e = \Pi_{\mathbf{V}_h\cdot\mathbf{n}}(\nabla(\psi_{\mathbf{a}}u_\mathrm{D})\cdot\mathbf{t}_\Omega)|_e$ for all $e\in\mathcal{E}_h^{\mathrm{ext,D}}$ contained in $\partial\omega_{\mathbf{a}}$. This follows as above by (8.18a) and (8.17b). Moreover, by (8.18b), $s_h^{\mathbf{a}}(\mathbf{a}') = 0 = (\psi_{\mathbf{a}}u_\mathrm{D})(\mathbf{a}')$ for the other vertices $\mathbf{a}'$ of $\omega_{\mathbf{a}}$ lying on $\partial\omega_{\mathbf{a}}\cap\Gamma_\mathrm{D}$. Thus $s_h^{\mathbf{a}}(\mathbf{a}) = (\psi_{\mathbf{a}}u_\mathrm{D})(\mathbf{a}) = u_\mathrm{D}(\mathbf{a})$ and the conclusion follows by (8.18c). □

As in Remark 7.10.2, we have, following Lemma 6.6.8:

Remark 8.1.11 (Local potential minimization)**.** *Definition 8.1.9 can be equivalently written as*

$$\boldsymbol{\varsigma}_h^{\mathbf{a}} := \arg\min_{\mathbf{v}_h\in\mathbf{V}_{h,\mathrm{N}}^{\mathbf{a}},\,\nabla\cdot\mathbf{v}_h=0}\big\|\mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}}u_h)+\mathbf{v}_h\big\|_{\omega_{\mathbf{a}}}\qquad\forall\mathbf{a}\in\mathcal{V}_h. \tag{8.20}$$

Moreover, a discrete primal formulation is

$$s_h^{\mathbf{a}} := \arg\min_{v_h\in V_{h,\mathrm{D}}^{\mathbf{a}}}\|\nabla(\psi_{\mathbf{a}}u_h - v_h)\|_{\omega_{\mathbf{a}}}\qquad\forall\mathbf{a}\in\mathcal{V}_h, \tag{8.21}$$

where $V_{h,\mathrm{D}}^{\mathbf{a}}$ denotes piecewise $(k+1)$-th degree polynomials on $\mathcal{T}_{\mathbf{a}}$ (for $\mathbf{V}_h$ the Raviart–Thomas–Nédélec space of degree k), with Dirichlet boundary condition on $\partial\omega_{\mathbf{a}}\setminus\Gamma_\mathrm{N}$ given by

$$\begin{aligned}(\nabla v_h\cdot\mathbf{t}_\Omega)|_e &= \Pi_{\mathbf{V}_h\cdot\mathbf{n}}(\nabla(\psi_{\mathbf{a}}u_\mathrm{D})\cdot\mathbf{t}_\Omega)|_e\qquad\forall e\in\mathcal{E}_h^{\mathrm{ext,D}},\ e\subset\partial\omega_{\mathbf{a}},\\ v_h(\mathbf{a}) &= u_\mathrm{D}(\mathbf{a}),\\ v_h|_{\partial\omega_{\mathbf{a}}\setminus\partial\Omega} &= 0.\end{aligned}$$

Let $V_h^{\mathbf{a}}$ be as $V_{h,\mathrm{D}}^{\mathbf{a}}$, with a homogeneous Dirichlet boundary condition everywhere on $\partial\omega_{\mathbf{a}}\setminus\Gamma_\mathrm{N}$. Then (8.21) is further equivalent to finding $s_h^{\mathbf{a}}\in V_{h,\mathrm{D}}^{\mathbf{a}}$ such that

$$(\nabla s_h^{\mathbf{a}},\nabla v_h) = (\nabla(\psi_{\mathbf{a}}u_h),\nabla v_h)\qquad\forall v_h\in V_h^{\mathbf{a}}.$$

Finally, inhomogeneous boundary conditions (8.2b)–(8.2c) for the alternative potential reconstruction of Remark 7.10.4 can be derived similarly.

8.1.8 Local efficiency

The generalization of the results presented in Section 7.11 to the inhomogeneous boundary conditions (8.2b)–(8.2c) can be summarized as follows:

Theorem 8.1.12 (Polynomial-degree-robust local efficiency, inhomogeneous boundary conditions)**.** *Let u be the weak solution given by (8.3). Let u_h be a piecewise polynomial and consider Definition 8.1.6 of $\boldsymbol{\sigma}_h$ with the spaces $\mathbf{V}_h$ and Q_h satisfying (7.60) for all $\mathbf{a} \in \mathcal{V}_h$. Then,*

$$\begin{aligned}\|\nabla u_h + \boldsymbol{\sigma}_h\|_K \le C_{\mathrm{st}} C_{\mathrm{cont,PF}} &\sum_{\mathbf{a}\in\mathcal{V}_K} \|\nabla(u-u_h)\|_{\omega_\mathbf{a}} \\ &+ C_{\mathrm{st}} \sum_{\mathbf{a}\in\mathcal{V}_K} \left\{ \sum_{K'\in\mathcal{T}_\mathbf{a}} \left(\frac{h_{K'}}{\pi} \|\psi_\mathbf{a} f - \Pi_{Q_h}(\psi_\mathbf{a} f)\|_{K'} \right)^2 \right\}^{\frac12} \\ &+ C_{\mathrm{st}} \sum_{\mathbf{a}\in\mathcal{V}_K} \left\{ \sum_{K'\in\mathcal{T}_\mathbf{a}} \left\{ \sum_{e\in\mathcal{E}_K^{\mathrm{N}}} \left(\bar{C}_{\mathrm{t},K,e} h_e^{\frac12} \|\psi_\mathbf{a}\sigma_{\mathrm{N}} - \Pi_{\mathbf{V}_h\cdot\mathbf{n}}(\psi_\mathbf{a}\sigma_{\mathrm{N}})\|_e \right) \right\}^2 \right\}^{\frac12}\end{aligned} \tag{8.22}$$

for all $K \in \mathcal{T}_h$, with the constants C_{st} of (7.59) and $C_{\mathrm{cont,PF}}$ of (7.46), respectively. Consider now Definition 8.1.9 of s_h with the space $\mathbf{V}_h$ satisfying (7.62) for all $\mathbf{a} \in \mathcal{V}_h$. Assume in addition that u_h verifies the zero-mean condition (7.50), where the jump on Dirichlet boundary faces is given by $u_h - u_{\mathrm{D}}$. Then,

$$\begin{aligned}&\|\nabla(u_h - s_h)\|_K \\ &\le C_{\mathrm{st}} C_{\mathrm{cont,bPF}} \sum_{\mathbf{a}\in\mathcal{V}_K} \|\nabla(u-u_h)\|_{\omega_\mathbf{a}} \\ &\quad + C_{\mathrm{st}} \sum_{\mathbf{a}\in\mathcal{V}_K} \left\{ \sum_{K'\in\mathcal{T}_\mathbf{a}} \left\{ \sum_{e\in\mathcal{E}_K^{\mathrm{D}}} \left(\bar{C}_{\mathrm{t},K,e} h_e^{\frac12} \|\nabla(\psi_\mathbf{a} u_{\mathrm{D}})\cdot\mathbf{t}_\Omega - \Pi_{\mathbf{V}_h\cdot\mathbf{n}}(\nabla(\psi_\mathbf{a} u_{\mathrm{D}})\cdot\mathbf{t}_\Omega)\|_e \right) \right\}^2 \right\}^{\frac12}\end{aligned} \tag{8.23}$$

for all $K \in \mathcal{T}_h$, with the constants C_{st} of (7.59) and $C_{\mathrm{cont,bPF}}$ of (7.53), respectively.

Proof. For interior vertices $\mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}$, the assertion coincides with that of Theorem 7.11.6 (note that the contributions on boundary faces $e \in \mathcal{E}_K^{\mathrm{N}}$ and $e \in \mathcal{E}_K^{\mathrm{D}}$ in (8.22) and (8.23), respectively, are discarded by the hat function $\psi_\mathbf{a}$).

For boundary vertices $\mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}$, one needs to treat the inhomogeneous Neumann boundary conditions imposed on the spaces $\mathbf{V}_{h,\mathrm{N}}^\mathbf{a}$ in (8.12b) and (8.17b). The proof follows the treatment of data oscillation (nonpolynomial source function f) in Theorem 7.11.6. First, the continuous-level problem (7.44) will appear with an inhomogeneous Neumann boundary condition $g_{\mathrm{N}}^\mathbf{a}$ on $\partial\omega_\mathbf{a} \cap \Gamma_{\mathrm{N}}$, with $r_\mathbf{a} \in H^1_*(\omega_\mathbf{a})$ satisfying

$$(\nabla r_\mathbf{a}, \nabla v)_{\omega_\mathbf{a}} = -(\boldsymbol{\tau}_h^\mathbf{a}, \nabla v)_{\omega_\mathbf{a}} + (g^\mathbf{a}, v)_{\omega_\mathbf{a}} - \langle g_{\mathrm{N}}^\mathbf{a}, v\rangle_{\partial\omega_\mathbf{a}\cap\Gamma_{\mathrm{N}}} \qquad \forall v \in H^1_*(\omega_\mathbf{a})$$

with $g_{\mathrm{N}}^\mathbf{a} := \psi_\mathbf{a}\sigma_{\mathrm{N}}$, $\boldsymbol{\tau}_h^\mathbf{a} := \psi_\mathbf{a}\nabla u_h$, and $g^\mathbf{a} := \psi_\mathbf{a} f - \nabla\psi_\mathbf{a}\cdot\nabla u_h$, and with

$$\begin{aligned} H^1_*(\omega_\mathbf{a}) &:= \{v \in H^1(\omega_\mathbf{a});\ (v,1)_{\omega_\mathbf{a}} = 0\}, && \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}} \setminus \mathcal{V}_h^{\mathrm{ext,D}}, \\ H^1_*(\omega_\mathbf{a}) &:= \{v \in H^1(\omega_\mathbf{a});\ v = 0 \text{ on } \partial\omega_\mathbf{a} \cap \partial\Gamma_{\mathrm{D}}\}, && \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext,D}}. \end{aligned}$$

Then the local discrete formulation (8.11) will lead us to the study of the above problem with a polynomial Neumann term given by $\tilde{g}_{\mathrm{N}}^\mathbf{a} := \Pi_{\mathbf{V}_h\cdot\mathbf{n}}(\psi_\mathbf{a}\sigma_{\mathrm{N}})$, and we will have to bound the misfit

$$\sup_{v\in H^1_*(\omega_\mathbf{a});\, \|\nabla v\|_{\omega_\mathbf{a}}=1} \langle g_{\mathrm{N}}^\mathbf{a} - \tilde{g}_{\mathrm{N}}^\mathbf{a}, v\rangle_{\partial\omega_\mathbf{a}\cap\Gamma_{\mathrm{N}}},$$

which leads to the third term on the right-hand side in (8.22) as in the proof of Theorem 8.1.4.

Similarly, problem (7.51) will appear with an inhomogeneous Neumann boundary condition $g_{\mathrm{N}}^{\mathbf{a}}$ on $\partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{D}}$ (Dirichlet boundary condition on Γ_{D} appears here as a Neumann boundary condition on $\partial\omega_{\mathbf{a}} \cap \Gamma_{\mathrm{D}}$)

$$(\nabla r_{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}} = -(\boldsymbol{\tau}_h^{\mathbf{a}}, \nabla v)_{\omega_{\mathbf{a}}} + (g^{\mathbf{a}}, v)_{\omega_{\mathbf{a}}} - \langle g_{\mathrm{N}}^{\mathbf{a}}, v\rangle_{\partial\omega_{\mathbf{a}}\cap\Gamma_{\mathrm{D}}} \qquad \forall v \in H^1_*(\omega_{\mathbf{a}}),$$

with $g_{\mathrm{N}}^{\mathbf{a}} := \nabla(\psi_{\mathbf{a}} u_{\mathrm{D}})\cdot\mathbf{t}_\Omega$, $\boldsymbol{\tau}_h^{\mathbf{a}} := \mathrm{R}_{\frac{\pi}{2}}\nabla(\psi_{\mathbf{a}} u_h)$, and $g^{\mathbf{a}} := 0$, and with

$$\begin{aligned} H^1_*(\omega_{\mathbf{a}}) &:= \{v \in H^1(\omega_{\mathbf{a}});\ (v,1)_{\omega_{\mathbf{a}}} = 0\}, && \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}} \setminus \mathcal{V}_h^{\mathrm{ext,N}},\\ H^1_*(\omega_{\mathbf{a}}) &:= \{v \in H^1(\omega_{\mathbf{a}});\ v = 0 \text{ on } \partial\omega_{\mathbf{a}} \cap \partial\Gamma_{\mathrm{N}}\}, && \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext,N}}. \end{aligned}$$

Again, the local discrete formulation (8.16) will lead us to the study of the above problem with a polynomial Neumann term given by $\tilde{g}_{\mathrm{N}}^{\mathbf{a}} := \Pi_{\mathbf{V}_h\cdot\mathbf{n}}(\nabla(\psi_{\mathbf{a}} u_{\mathrm{D}})\cdot\mathbf{t}_\Omega)$, and we will have to bound the misfit

$$\sup_{v\in H^1_*(\omega_{\mathbf{a}});\, \|\nabla v\|_{\omega_{\mathbf{a}}}=1} \langle g_{\mathrm{N}}^{\mathbf{a}} - \tilde{g}_{\mathrm{N}}^{\mathbf{a}}, v\rangle_{\partial\omega_{\mathbf{a}}\cap\Gamma_{\mathrm{D}}}.$$

This leads to the second term on the right-hand side in (8.23) and concludes the proof. □

8.2 Residual-based a posteriori error estimators

We introduce in this section the so-called residual-based a posteriori error estimators following the books of Verfürth [93] and Babuška and Strouboulis [13] for conforming discretizations. Nonconforming discretizations are treated following Dari *et al.* [37], Achdou *et al.* [1], and Karakashian and Pascal [66], see also the references therein.

Recall the notation of Section 3.4. Let us then introduce the classical *residual indicators* for problem (7.1a)–(7.1b). They are given by, for $K \in \mathcal{T}_h$,

$$\eta_{\mathrm{res},K} := \left\{\sum_{K'\in\mathfrak{T}_K} h_{K'}^2 \|f + \Delta u_h\|_{K'}^2\right\}^{\frac12} + \left\{\sum_{e\in\mathfrak{E}_K^{\mathrm{int}}} h_e \|[\![\nabla u_h]\!]\cdot\mathbf{n}_e\|_e^2\right\}^{\frac12}, \tag{8.24a}$$

$$|u_h|_{\mathrm{J},K} := \left\{\sum_{e\in\mathfrak{E}_K} h_e^{-1} \|[\![u_h]\!]\|_e^2\right\}^{\frac12}. \tag{8.24b}$$

The first term of (8.24a) is called the *element residual*, the second term of (8.24a) the *face residual*, and $|u_h|_{\mathrm{J},K}$ from (8.24b) the *jump residual*. It will be useful to introduce in this section the following assumption:

Assumption 8.2.1 (Setting for residual-based estimates)**.** *We suppose that*

1. *the mesh $\mathcal{T}_h$ is shape-regular with a constant $\kappa_{\mathcal{T}} > 0$ in the sense of Section 3.1;*
2. *for a fixed integer $k \geq 1$*
   a) *the approximate solution u_h is in the space $\mathbb{P}_k(\mathcal{T}_h)$;*
   b) *the datum f is in the space $\mathbb{P}_k(\mathcal{T}_h)$.*

8.2.1 Reliability

The following result follows by [93, 37, 13, 1, 66]:

Theorem 8.2.2 (Residual-based a posteriori error estimate)**.** *Let u be the weak solution given by Definition 7.1.1. Let Assumption 8.2.1 1–2a) hold. Let finally u_h satisfy the hat-function orthogonality (7.26). Then there exists a generic constant $C_{\rm res}$ only depending on the space dimension d, mesh shape regularity parameter $\kappa_{\mathcal{T}}$, and on the polynomial degree k such that*

$$\|\nabla(u-u_h)\| \le C_{\rm res}\left(\left\{\sum_{K\in\mathcal{T}_h}\eta_{{\rm res},K}^2\right\}^{\frac12}+\left\{\sum_{K\in\mathcal{T}_h}|u_h|_{{\rm J},K}^2\right\}^{\frac12}\right). \tag{8.25}$$

8.2.2 Efficiency of element and face residuals via the bubble functions technique

We show in this section the local efficiency (cf. property **ii)** of Section 1.4) of the a posteriori error estimates of Theorem 8.2.2. Henceforth, we use $A \lesssim B$ when there exists a positive constant C that can only depend on the space dimension d, the shape-regularity parameter $\kappa_{\mathcal{T}}$, the polynomial degree k, and the parameter α in the case of the discontinuous Galerkin method such that $A \le CB$.

Lemma 8.2.3 (Efficiency of the element residuals)**.** *Let Assumption 8.2.1 hold. Let $K \in \mathcal{T}_h$. Then*

$$h_K\|f+\Delta u_h\|_K \lesssim \|\nabla(u-u_h)\|_K. \tag{8.26}$$

Proof. The proof follows Verfürth [93]. Set

$$v_K := (f+\Delta u_h)|_K. \tag{8.27}$$

Let ψ_K be the bubble function on K given by the product of the $d+1$ barycentric coordinates $\psi_{\mathbf{a}}$, $\mathbf{a}\in\mathcal{V}_K$, of K (recall that the barycentric coordinate $\psi_{\mathbf{a}}$, or the "hat" function, is the affine function on K which takes the value one at the vertex $\mathbf{a}$ of K and zero at all other vertices of K). Note that $\psi_K|_{\partial K}=0$. Also note that both v_K and ψ_K are polynomials (cf. Assumption 8.2.1). By equivalence of norms on finite-dimensional spaces, there holds

$$(v_K,v_K)_K \lesssim (v_K,\psi_K v_K)_K. \tag{8.28}$$

Using the inverse inequality (cf. Quarteroni and Valli [83, Proposition 6.3.2]), we obtain

$$h_K\|\nabla(\psi_K v_K)\|_K \lesssim \|\psi_K v_K\|_K. \tag{8.29}$$

Finally, from the definition of the bubble function ψ_K, there holds

$$\|\psi_K v_K\|_K \le \|\psi_K\|_{\infty,K}\|v_K\|_K \le \|v_K\|_K. \tag{8.30}$$

Thus, using (8.27) and (8.28), noting that $\psi_K v_K \in H_0^1(K)$ and using (7.2), employing the Green theorem, the Cauchy–Schwarz inequality, (8.29), and (8.30)

$$\begin{aligned}\|v_K\|^2 &\lesssim (v_K,\psi_K v_K)_K = (f+\Delta u_h,\psi_K v_K)_K = (\nabla(u-u_h),\nabla(\psi_K v_K))_K\\ &\le \|\nabla(u-u_h)\|_K\|\nabla(\psi_K v_K)\|_K \lesssim \|\nabla(u-u_h)\|_K h_K^{-1}\|v_K\|_K.\end{aligned}$$

Therefrom, the assertion of the lemma easily follows. □

Lemma 8.2.4 (Efficiency of the face residuals)**.** *Let Assumption 8.2.1 hold. Let* $e \in \mathcal{E}_h^{\mathrm{int}}$. *Then*

$$h_e^{\frac{1}{2}} \|[\![\nabla u_h]\!]\cdot\mathbf{n}_e\|_e \lesssim \|\nabla(u-u_h)\|_{\mathcal{T}_e}.$$

Proof. The proof follows again Verfürth [93]. Let

$$v_e := [\![\nabla u_h]\!]\cdot\mathbf{n}_e|_e. \tag{8.31}$$

Recall that $\mathcal{T}_e$ denotes the two simplices that share the face e. Let ψ_e be the bubble function on $\mathcal{T}_e$ given by the product of the barycentric coordinates with vertices in e and remark that $\psi_e|_{\partial\mathcal{T}_e} = 0$. Then, by equivalence of norms on finite-dimensional spaces, there holds

$$\langle v_e, v_e\rangle_e \lesssim \langle v_e, \psi_e v_e\rangle_e. \tag{8.32}$$

Let us keep the same notation for the extension of the function v_e, originally only defined on the face e, to a function defined on the two simplices $\mathcal{T}_e$. The extension is done by constant values in the direction of the barycenter of e–opposite vertex. Then we also have the estimate

$$\|v_e\|_{\mathcal{T}_e} \lesssim h_e^{\frac{1}{2}} \|v_e\|_e. \tag{8.33}$$

Finally, from the definition of the bubble function ψ_e, there holds

$$\|\psi_e v_e\|_{\mathcal{T}_e} \le \|\psi_e\|_{\infty,\mathcal{T}_e}\|v_e\|_{\mathcal{T}_e} \le \|v_e\|_{\mathcal{T}_e}, \tag{8.34}$$

whereas the inverse inequality and the shape-regularity of the mesh $\mathcal{T}_h$ yield

$$\|\nabla(\psi_e v_e)\|_{\mathcal{T}_e} \lesssim h_e^{-1}\|\psi_e v_e\|_{\mathcal{T}_e}. \tag{8.35}$$

Thus, using (8.31), (8.32), the Green theorem, (7.2) after noting that $\psi_e v_e \in H_0^1(\mathcal{T}_e)$, the Cauchy–Schwarz inequality, (8.33), (8.34), and (8.35),

$$\begin{aligned}
\|v_e\|_e^2 &\lesssim \langle v_e, \psi_e v_e\rangle_e = \langle [\![\nabla u_h]\!]\cdot\mathbf{n}_e, \psi_e v_e\rangle_e \\
&= (f + \Delta u_h, \psi_e v_e)_{\mathcal{T}_e} + (\nabla(u_h - u), \nabla(\psi_e v_e))_{\mathcal{T}_e} \\
&\le \|f + \Delta u_h\|_{\mathcal{T}_e}\|\psi_e v_e\|_{\mathcal{T}_e} + \|\nabla(u-u_h)\|_{\mathcal{T}_e}\|\nabla(\psi_e v_e)\|_{\mathcal{T}_e} \\
&\lesssim \|f + \Delta u_h\|_{\mathcal{T}_e}\|v_e\|_{\mathcal{T}_e} + \|\nabla(u-u_h)\|_{\mathcal{T}_e} h_e^{-1}\|v_e\|_{\mathcal{T}_e} \\
&\lesssim (h_e\|f + \Delta u_h\|_{\mathcal{T}_e} + \|\nabla(u-u_h)\|_{\mathcal{T}_e}) h_e^{-\frac{1}{2}}\|v_e\|_e.
\end{aligned}$$

Combining this result with (8.26) and using the shape-regularity of the mesh $\mathcal{T}_h$ yields the assertion of the lemma. □

8.2.3 Efficiency of jumps terms via local Neumann problems

Lemma 8.2.5 (Efficiency of the jump residuals)**.** *Let Assumption 8.2.1 1 hold. Let* $e \in \mathcal{E}_h$ *and let*

$$\langle [\![u_h]\!], 1\rangle_e = 0. \tag{8.36}$$

Then

$$h_e^{-\frac{1}{2}} \|[\![u_h]\!]\|_e \lesssim \|\nabla(u-u_h)\|_{\mathcal{T}_e}.$$

Proof. The proof follows Achdou *et al.* [1]. Recall that $\mathcal{T}_e$ denotes the one or two simplices that share the face e. On each $K \in \mathcal{T}_e$, consider the following local Neumann problem: find φ_K such that

$$-\Delta\varphi_K = 0 \qquad \text{in } K, \tag{8.37a}$$
$$\nabla\varphi_K\cdot\mathbf{n}_e = [\![u_h]\!]|_e \qquad \text{on } \partial K \cap e, \tag{8.37b}$$
$$\nabla\varphi_K\cdot\mathbf{n}_K = 0 \qquad \text{on } \partial K \setminus e, \tag{8.37c}$$
$$(\varphi_K, 1)_K = 0. \tag{8.37d}$$

Note that it follows from (8.37b)–(8.37c) which prescribe the Neumann boundary conditions, from (8.36), and from the fact that the right-hand faces in (8.37a) are zero that the mean values of the Neumann boundary conditions on ∂K are equal to the mean values of the source terms, so that (8.37a)–(8.37d) lead to well-posed weak formulations with unique solutions. For any $K \in \mathcal{T}_e$, these are characterized by: find $\varphi_K \in H^1(K)$ with $(\varphi_K, 1)_K = 0$ such that

$$(\nabla\varphi_K, \nabla v)_K = \langle [\![u_h]\!], v\rangle_e \mathbf{n}_K\cdot\mathbf{n}_e \qquad \forall v \in H^1(K) \text{ such that } (v,1)_K = 0.$$

Set $\varphi \in H^1(\mathcal{T}_e)$ by $\varphi|_K := \varphi_K$, $K \in \mathcal{T}_e$. We have, by the fact that the exact solution u belongs to $H^1_0(\Omega)$, Theorem 4.4.3, the Green theorem, (8.37a), and the Cauchy–Schwarz inequality

$$\begin{aligned}\|[\![u_h]\!]\|_e^2 = \langle [\![u-u_h]\!], \nabla\varphi\cdot\mathbf{n}_e\rangle_e &= (\nabla(u-u_h), \nabla\varphi)_{\mathcal{T}_e} + (u-u_h, \Delta\varphi)_{\mathcal{T}_e} \\ &= (\nabla(u-u_h), \nabla\varphi)_{\mathcal{T}_e} \le \|\nabla(u-u_h)\|_{\mathcal{T}_e}\|\nabla\varphi\|_{\mathcal{T}_e}.\end{aligned} \tag{8.38}$$

We now bound $\|\nabla\varphi\|_{\mathcal{T}_e}$. Let $K \in \mathcal{T}_e$. We develop

$$\|\nabla\varphi\|_K^2 = (\nabla\varphi, \nabla\varphi)_K = -(\Delta\varphi, \varphi)_K + \langle \nabla\varphi\cdot\mathbf{n}_K, \varphi\rangle_{\partial K} = \langle \nabla\varphi\cdot\mathbf{n}_K, \varphi\rangle_e \le \|[\![u_h]\!]\|_e\|\varphi\|_e,$$

using respectively the Green theorem, (8.37a), (8.37c), the Cauchy–Schwarz inequality, and (8.37b). As also (8.37d) holds true, the trace inequality (4.22c) and the shape-regularity of $\mathcal{T}_h$ give

$$\|\varphi\|_e \lesssim h_e^{\frac12}\|\nabla\varphi\|_K.$$

Combining the two above bounds, we come to

$$\|\nabla\varphi\|_{\mathcal{T}_e} \lesssim h_e^{\frac12}\|[\![u_h]\!]\|_e. \tag{8.39}$$

Now combining (8.38) and (8.39) gives the desired result. □

8.3 Reconstructions by direct prescription

Potential and flux reconstructions s_h and $\boldsymbol{\sigma}_h$ according to Definitions 7.6.1 and 7.6.2 may be obtained differently than as described in Sections 7.9 and 7.10. The interest of the constructions we now present is that they are carried out *locally*, mesh element by mesh element, via a *direct prescription* of the corresponding degrees of freedom. Thus no local Neumann/Dirichlet problems on the patches of elements as those in Definitions 7.9.1, 7.10.1, or Remark 7.10.4 need to be assembled and solved. On the other hand, the constructions become *scheme-dependent* and it is not clear whether the local efficiency is still polynomial-degree-robust or not. In this section, while proving efficiency, we also make link of the reconstruction-based estimators with the residual ones of Section 8.2.

8.3.1 Averaging operator

Definition 7.10.1 from Chapter 7 gave us a generic way to construct the potential reconstruction s_h of Definition 7.6.1. Here we present its simpler and cheaper antecedent: the averaging operator $\mathcal{I}_{\rm av} : \mathbb{P}_k(\mathcal{T}_h) \to \mathbb{P}_k(\mathcal{T}_h) \cap H^1_0(\Omega)$, $k \geq 1$, following Achdou *et al.* [1], Karakashian and Pascal [66], and Burman and Ern [25]. This operator associates to a piecewise k-th order discontinuous polynomial a piecewise k-th order polynomial which is $H^1_0(\Omega)$-conforming, i.e., in particular, continuous. Recall from Section 5.2 that the Lagrangian degrees of freedom of $\mathbb{P}_k(\mathcal{T}_h) \cap H^1_0(\Omega)$ are punctual values in points herein called nodes. In order to specify a function in $\mathbb{P}_k(\mathcal{T}_h) \cap H^1_0(\Omega)$, we thus need to fix these degrees of freedom. We will do so by prescribing at each Lagrangian node $\mathbf{a}$ of $\mathbb{P}_k(\mathcal{T}_h) \cap H^1_0(\Omega)$ the average of the values of the original polynomial at this node. As a particular consequence, when the node $\mathbf{a}$ lies in the interior of some $K \in \mathcal{T}_h$, the value is unchanged. Finally, at boundary nodes, 0 is imposed. Denote by $\mathcal{T}_\mathbf{a}$ all the elements sharing a given node (degree of freedom) $\mathbf{a}$ and $|\mathcal{T}_\mathbf{a}|$ their number. Then we define:

Definition 8.3.1 (Nodewise averaging)**.** *Let $v_h \in \mathbb{P}_k(\mathcal{T}_h)$. Define $\mathcal{I}_{\rm av}(v_h) \in \mathbb{P}_k(\mathcal{T}_h) \cap H^1_0(\Omega)$ by*

$$\mathcal{I}_{\rm av}(v_h)(\mathbf{a}) := \frac{1}{|\mathcal{T}_\mathbf{a}|} \sum_{K \in \mathcal{T}_\mathbf{a}} v_h|_K(\mathbf{a}) \qquad \mathbf{a} \in \Omega,$$
$$\mathcal{I}_{\rm av}(v_h)(\mathbf{a}) := 0 \qquad \mathbf{a} \in \partial\Omega.$$

8.3.2 A general local efficiency result

Recall the definition of the residual indicators $\eta_{{\rm res},K}$ and $|u_h|_{{\rm J},K}$ of (8.24a)–(8.24b). In order to proceed generally at this point, without the specification of a particular numerical method, we now make the following assumption. Will verify it later and use it for concluding the local efficiency for each numerical method:

Assumption 8.3.2 (Approximation property for (7.1a)–(7.1b))**.** *We assume that the potential and flux reconstructions s_h and $\boldsymbol{\sigma}_h$ that we shall construct in this section are piecewise polynomials at most of order k such that, for all $K \in \mathcal{T}_h$,*

$$\eta_{{\rm NC},K} + \eta_{{\rm F},K} \lesssim \eta_{{\rm res},K} + |u_h|_{{\rm J},K}. \tag{8.40}$$

We then have:

Theorem 8.3.3 (Efficiency of the estimate of Theorem 7.8.1)**.** *Let u be the weak solution given by Definition 7.1.1, let u_h be as in (7.4), let $\eta_{{\rm R},K}$, $\eta_{{\rm F},K}$, and $\eta_{{\rm NC},K}$ be given respectively by (7.23a), (7.23b), and (7.23c), and let finally Assumptions 8.2.1 and 8.3.2 be satisfied. Then, for all $K \in \mathcal{T}_h$,*

$$\eta_{{\rm NC},K} + \eta_{{\rm R},K} + \eta_{{\rm F},K} \lesssim \|\nabla(u - u_h)\|_{\mathfrak{T}_K} + |u - u_h|_{{\rm J},K}.$$

Proof. We first observe that $\eta_{{\rm NC},K} + \eta_{{\rm F},K} \lesssim \eta_{{\rm res},K} + |u_h|_{{\rm J},K}$ directly by Assumption 8.3.2, whereas, for $\eta_{{\rm R},K}$, the triangle and inverse inequalities yield

$$\begin{aligned}\eta_{{\rm R},K} &\leq \frac{h_K}{\pi}\|f + \Delta u_h\|_K + \frac{h_K}{\pi}\|\Delta u_h + \nabla\cdot\boldsymbol{\sigma}_h\|_K\\ &\lesssim h_K\|f + \Delta u_h\|_K + \|\nabla u_h + \boldsymbol{\sigma}_h\|_K\\ &\lesssim \eta_{{\rm res},K} + |u_h|_{{\rm J},K},\end{aligned}$$

owing to Assumptions 8.2.1 and 8.3.2. Then combining Lemmas 8.2.3 and 8.2.4 with the shape-regularity of the mesh $\mathcal{T}_h$, we obtain $\eta_{\mathrm{res},K} \lesssim \|\nabla(u-u_h)\|_{\mathfrak{T}_K}$. The result follows by noticing that $|u_h|_{\mathrm{J},K} = |u-u_h|_{\mathrm{J},K}$, as $[\![u]\!] = 0$ for all $e \in \mathcal{E}_h$ by Theorem 4.4.3. □

Remark 8.3.4 (Equivalence result). *If u_h is in $H^1_0(\Omega)$, the jump seminorms $|u-u_h|_{\mathrm{J},K}$ vanish according to Theorem 4.4.3. If the jumps of u_h have zero mean values, i.e., if (8.36) holds for all $e \in \mathcal{E}_h$, Lemma 8.2.5 yields $|u_h|_{\mathrm{J},K} \lesssim \|\nabla(u-u_h)\|_{\mathfrak{T}_K}$. Thus, in these two cases, Theorem 8.3.3 actually gives*

$$\eta_{\mathrm{NC},K} + \eta_{\mathrm{R},K} + \eta_{\mathrm{F},K} \lesssim \|\nabla(u-u_h)\|_{\mathfrak{T}_K} \tag{8.41}$$

for all $K \in \mathcal{T}_h$. Note that (7.24) together with (8.41) gives simultaneously the guaranteed upper bound *and* local efficiency *in the sense of the properties* **i)** *and* **ii)** *of Section 1.4.*

In the general case, using again $|u_h|_{\mathrm{J},K} = |u-u_h|_{\mathrm{J},K}$, the following equivalence result, satisfying both **i)** *and* **ii)** *of Section 1.4, holds true:*

$$\|\nabla(u-u_h)\|^2 + \sum_{K\in\mathcal{T}_h} |u-u_h|^2_{\mathrm{J},K} \le \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{F},K} + \eta_{\mathrm{R},K})^2 + \sum_{K\in\mathcal{T}_h} \eta^2_{\mathrm{NC},K} + \sum_{K\in\mathcal{T}_h} |u_h|^2_{\mathrm{J},K},$$
$$\eta_{\mathrm{NC},K} + \eta_{\mathrm{R},K} + \eta_{\mathrm{F},K} + |u_h|_{\mathrm{J},K} \lesssim \|\nabla(u-u_h)\|_{\mathfrak{T}_K} + |u-u_h|_{\mathrm{J},K}.$$

We now come back to the averaging operator of Section 8.3.1. The following results have been proved in Karakashian and Pascal [66, Theorem 2.2] and Burman and Ern [25, Lemmas 3.2 and 5.3 and Remark 3.2]:

Lemma 8.3.5 (Averaging operator). *Let Assumption 8.2.1 1 hold. Let $v_h \in \mathbb{P}_k(\mathcal{T}_h)$, $k \ge 1$. Then*

$$\|\nabla(v_h - \mathcal{I}_{\mathrm{av}}(v_h))\|_K \lesssim |v_h|_{\mathrm{J},K}$$

for all $K \in \mathcal{T}_h$.

We admit the following result, which can be shown using the equivalence of norms on finite-dimensional spaces, the Piola transformation, and scaling arguments for all $\mathbf{v}_h \in \mathbf{RTN}_k(K)$, $K \in \mathcal{S}_h$, $k \ge 0$:

$$\|\mathbf{v}_h\|_K \lesssim \left\{ \sum_{e\in\mathcal{E}_K} h_e \|\mathbf{v}_h\cdot\mathbf{n}_e\|^2_e + \left(\sup_{\mathbf{r}_h\in[\mathbb{P}_{k-1}(K)]^d} \frac{(\mathbf{v}_h,\mathbf{r}_h)_K}{\|\mathbf{r}_h\|_K} \right)^2 \right\}^{\frac12}. \tag{8.42}$$

Note that for $k=0$, it in particular gives

$$\|\mathbf{v}_h\|_K \lesssim \left\{ \sum_{e\in\mathcal{E}_K} h_e \|\mathbf{v}_h\cdot\mathbf{n}_e\|^2_e \right\}^{\frac12}.$$

It follows from Theorem 8.3.3 that, in order to apply the local efficiency results to a given numerical method, we merely have to verify Assumptions 8.2.1 and 8.3.2. Assumption 8.2.1 is technical and will always be satisfied. Taking into account that we either set $s_h := u_h$ or $s_h := \mathcal{I}_{\mathrm{av}}(u_h)$ for the potential reconstruction and the result of Lemma 8.3.5, we also have

$$\eta_{\mathrm{NC},K} \lesssim |u_h|_{\mathrm{J},K}$$

for all $K \in \mathcal{T}_h$. We are thus left to verify

$$\eta_{\mathrm{F},K} \lesssim \eta_{\mathrm{res},K} + |u_h|_{\mathrm{J},K}$$

for all $K \in \mathcal{T}_h$. We do so now separately for the different numerical methods.

8.3.3 Crouzeix–Raviart nonconforming finite element method

Let V_h be the nonconforming finite element space defined in Section 7.13.2 for $k = 1$. This variant is called after Crouzeix and Raviart [36]. For simplicity, we suppose in this section that $u_h \in V_h$ solves (7.68) with $f \in \mathbb{P}_0(\mathcal{T}_h)$, i.e., that the source function f is piecewise constant on $\mathcal{T}_h$.

Proceeding following Destuynder and Métivet [39], Ainsworth [3], Braess [20], and [53], the potential reconstruction is simply obtained via

$$s_h := \mathcal{I}_{\rm av}(u_h), \tag{8.43}$$

where $\mathcal{I}_{\rm av}$ is the averaging operator of Section 8.3.1. We can use $\mathcal{I}_{\rm av}$ for $k = 1$, but better numerical results are obtained when we consider $v_h \in V_h$ as a function from $\mathbb{P}_2(\mathcal{T}_h)$ and reconstruct s_h in the space $\mathbb{P}_2(\mathcal{T}_h) \cap H^1_0(\Omega)$. The flux reconstruction $\boldsymbol{\sigma}_h$ is then obtained as follows. For $K \in \mathcal{T}_h$, let $\mathbf{x}_K$ denote its barycenter. Denote by $\mathbf{f}_h$ the piecewise affine vector function given on each element $K \in \mathcal{T}_h$ by $\frac{f|_K}{d}(\mathbf{x} - \mathbf{x}_K)$. Recall that $\{\!\{\cdot\}\!\}$ denotes the average operator defined by (3.2). Then:

Definition 8.3.6 (Elementwise flux prescription for the NCFE method)**.** *Let u_h be given by Definition 7.13.3 with $k = 1$. Then prescribe $\boldsymbol{\sigma}_h \in \mathbf{RTN}_0(K)$ for all $K \in \mathcal{T}_h$ by*

$$\boldsymbol{\sigma}_h|_K := -\nabla u_h|_K + \mathbf{f}_h|_K. \tag{8.44}$$

With these constructions, we have:

Lemma 8.3.7 (Potential and flux reconstructions in the NCFE method)**.** *Let $f \in \mathbb{P}_0(\mathcal{T}_h)$ and let u_h be given by Definition 7.13.3. Then s_h prescribed by (8.43) and $\boldsymbol{\sigma}_h$ prescribed by Definition 8.3.6 satisfy respectively Definitions 7.6.1 and 7.6.2.*

Proof. There is nothing to show for the potential reconstruction s_h and also (7.21b) is obvious, taking into account that $-\nabla u_h$ is piecewise constant and thus its divergence is zero, whereas the divergence of $\mathbf{f}_h$ is precisely f. So we are left with verifying the assumption (7.21a), i.e., the fact that $\boldsymbol{\sigma}_h$ given by (8.44) belongs to $\mathbf{RTN}_0(\mathcal{T}_h)$. According to Theorem 4.5.1, we need to show that $[\![\boldsymbol{\sigma}_h]\!]\cdot\mathbf{n}_e = 0$ for all faces $e \in \mathcal{E}_h^{\rm int}$. Let $e \in \mathcal{E}_h^{\rm int}$ be given, let K and K' be the two elements that share the face e, and consider the line in (7.68) associated with $v_h = \psi_e$, where ψ_e is the basis function associated with the face e. This is a function that takes value 1 in the barycenter of e and value 0 in all other face barycenters. Taking into account that $-\nabla u_h$ is piecewise constant, the above properties of the basis function ψ_e, and the Green theorem gives

$$\begin{aligned}(\nabla u_h, \nabla\psi_e)_{K\cup K'} &= \langle \nabla u_h\cdot\mathbf{n}_K, \psi_e\rangle_{\partial K} + \langle \nabla u_h\cdot\mathbf{n}_{K'}, \psi_e\rangle_{\partial K'}\\ &= \langle \nabla u_h\cdot\mathbf{n}_K, 1\rangle_e + \langle \nabla u_h\cdot\mathbf{n}_{K'}, 1\rangle_e.\end{aligned}$$

Similarly, using the simple property $(\mathbf{x}, \nabla\psi_e)_K = (\mathbf{x}_K, \nabla\psi_e)_K$ for all $K \in \mathcal{T}_h$, the Green theorem, and the fact that the normal component is sidewise constant for the $\mathbf{RTN}_0(K)$ functions, we rewrite the term on the right-hand side of (7.68) as

$$\begin{aligned}(f, \psi_e)_{K\cup K'} &= (\mathbf{f}_h, \nabla\psi_e)_K + (\mathbf{f}_h, \nabla\psi_e)_{K'} + (f, \psi_e)_K + (f, \psi_e)_{K'}\\ &= \langle \mathbf{f}_h\cdot\mathbf{n}_K, \psi_e\rangle_{\partial K} + \langle \mathbf{f}_h\cdot\mathbf{n}_{K'}, \psi_e\rangle_{\partial K'}\\ &= \langle \mathbf{f}_h\cdot\mathbf{n}_K, 1\rangle_e + \langle \mathbf{f}_h\cdot\mathbf{n}_{K'}, 1\rangle_e.\end{aligned}$$

The assertion follows by combining the two above identities with (7.68) and the definition (8.44) of $\boldsymbol{\sigma}_h$. □

The efficiency result is then:

Lemma 8.3.8 (Efficiency of the NCFE method for (7.1a)–(7.1b)). *Let u_h be given by Definition 7.13.3 and $\boldsymbol{\sigma}_h$ by (8.44). Let Assumption 8.2.1 be satisfied. Then*

$$\eta_{\mathrm{F},K} \lesssim \eta_{\mathrm{res},K} \tag{8.45}$$

for all $K \in \mathcal{T}_h$.

Proof. Let $K \in \mathcal{T}_h$ be given. Note that $\nabla u_h + \boldsymbol{\sigma}_h = \mathbf{f}_h$ by the definition (8.44). Using that there holds $\Delta u_h = 0$ since u_h is piecewise affine and the assumption that the source term f is piecewise constant, we come to

$$\begin{aligned}\eta_{\mathrm{F},K} = \|\mathbf{f}_h\|_K &= \frac{1}{d}\left\{\int_K f^2|\mathbf{x}-\mathbf{x}_K|^2\,\mathrm{d}\mathbf{x}\right\}^{\frac12}\\ &\lesssim h_K\|f\|_K = h_K\|f-\Delta u_h\|_K,\end{aligned}$$

using that $|\mathbf{x}-\mathbf{x}_K| \le h_K$, whence the assertion follows using the definition (8.24a) of $\eta_{\mathrm{res},K}$. □

It is to be noted that in the Crouzeix–Raviart nonconforming method, any function satisfies (8.36) and we actually have (8.41) as the final efficiency result, see Remark 8.3.4.

8.3.4 Discontinuous Galerkin method

Let $V_h := \mathbb{P}_k(\mathcal{T}_h)$ as in Section 7.13.3 and let $u_h \in V_h$ be given by (7.69). For the potential reconstruction be prescription, we again set

$$s_h := \mathcal{I}_{\mathrm{av}}(u_h). \tag{8.46}$$

Concerning the equilibrated flux reconstruction by direct prescription, we follow [15, 4, 67, 34, 50, 49, 8] and more precisely Kim [68] and [48]. Set $w_e := \frac12$ for $e \in \mathcal{E}_h^{\mathrm{int}}$, $w_e := 1$ for $e \in \mathcal{E}_h^{\mathrm{ext}}$, and either $l = k-1$ or $l = k$. We will construct $\boldsymbol{\sigma}_h$ in the space $\mathbf{RTN}_l(\mathcal{T}_h)$, cf. (5.7):

Definition 8.3.9 (Elementwise flux prescription for the DG method). *Let u_h be given by Definition 7.13.6. For all $K \in \mathcal{T}_h$, specify the degrees of freedom of $\boldsymbol{\sigma}_h \in \mathbf{RTN}_l(\mathcal{T}_h)$ by setting, for all $e \in \mathcal{E}_K$ and all $q_h \in \mathbb{P}_l(e)$,*

$$\langle \boldsymbol{\sigma}_h\cdot\mathbf{n}_e, q_h\rangle_e = \langle -\{\!\{\nabla u_h\}\!\}\cdot\mathbf{n}_e + \alpha h_e^{-1}[\![u_h]\!], q_h\rangle_e, \tag{8.47a}$$

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

$$(\boldsymbol{\sigma}_h, \mathbf{r}_h)_K = -(\nabla u_h, \mathbf{r}_h)_K + \theta\sum_{e\in\mathcal{E}_K} w_e\langle \mathbf{r}_h\cdot\mathbf{n}_e, [\![u_h]\!]\rangle_e. \tag{8.47b}$$

These developments imply:

Lemma 8.3.10 (Potential and flux reconstructions in the DG method). *Let u_h be given by Definition 7.13.6. Then s_h prescribed by (8.46), and $\boldsymbol{\sigma}_h$ prescribed by Definition 8.3.9 satisfy respectively Definitions 7.6.1 and 7.6.2. Moreover,*

$$(f - \nabla\cdot\boldsymbol{\sigma}_h, v_h)_K = 0 \qquad \forall v_h \in \mathbb{P}_l(K) \qquad \forall K \in \mathcal{T}_h. \tag{8.48}$$

Proof. First note that $\boldsymbol{\sigma}_h$ is indeed in $\mathbf{RTN}_l(\mathcal{T}_h)$ and thus the requirement (7.21a) is satisfied, cf. Theorem 4.5.1, as the normal components over the interior faces are by (8.47a) univalent. We next show (7.21b), or, more precisely, (8.48). Let $K \in \mathcal{T}_h$ and $v_h \in \mathbb{P}_l(K)$ be fixed. The Green theorem gives

$$(f - \nabla\cdot\boldsymbol{\sigma}_h, v_h)_K = (f, v_h)_K + (\boldsymbol{\sigma}_h, \nabla v_h)_K - \langle \boldsymbol{\sigma}_h\cdot\mathbf{n}_K, v_h\rangle_{\partial K} =: T_1 + T_2 + T_3.$$

Since $\nabla v_h \in [\mathbb{P}_{l-1}(K)]^d$, (8.47b) gives

$$T_2 = -(\nabla u_h, \nabla v_h)_K + \theta \sum_{e\in\mathcal{E}_K} w_e \langle \nabla v_h\cdot\mathbf{n}_e, [\![u_h]\!]\rangle_e.$$

Furthermore, the fact that $v_h|_e \in \mathbb{P}_l(e)$ for all $e \in \mathcal{E}_K$ and (8.47a) yield

$$-T_3 = \sum_{e\in\mathcal{E}_K} \{\mathbf{n}_K\cdot\mathbf{n}_e\langle -\{\!\{\nabla u_h\}\!\}\cdot\mathbf{n}_e + \alpha h_e^{-1}[\![u_h]\!], v_h\rangle_e\}.$$

Extend v_h by 0 outside of K. Using the above identities and the definition (7.69) of the discontinuous Galerkin scheme gives (8.48). □

We have:

Lemma 8.3.11 (Efficiency of the DG method for (7.1a)–(7.1b))**.** *Let u_h be given by Definition 7.13.6 and $\boldsymbol{\sigma}_h$ by (8.47a)–(8.47b). Let Assumption 8.2.1 be satisfied. Then*

$$\eta_{\mathrm{F},K} \lesssim \eta_{\mathrm{res},K} + |u_h|_{\mathrm{J},K} \tag{8.49}$$

for all $K \in \mathcal{T}_h$.

Proof. Let $K \in \mathcal{T}_h$. Set $\mathbf{v}_h := \nabla u_h + \boldsymbol{\sigma}_h$ and remark that $\mathbf{v}_h \in \mathbf{RTN}_l(K)$. By (8.47a),

$$\mathbf{v}_h\cdot\mathbf{n}_e = (1 - w_e)[\![\nabla u_h]\!]\cdot\mathbf{n}_e + \alpha h_e^{-1}\Pi_{l,e}([\![u_h]\!]),$$

where $\Pi_{l,e}$ is the $L^2(e)$-orthogonal projection onto $\mathbb{P}_l(e)$, and thus

$$\|\mathbf{v}_h\cdot\mathbf{n}_e\|_e \lesssim (1 - w_e)\|[\![\nabla u_h]\!]\cdot\mathbf{n}_e\|_e + \alpha h_e^{-1}\|[\![u_h]\!]\|_e. \tag{8.50}$$

By (8.47b), the Cauchy–Schwarz inequality, and the inverse inequality $\|\mathbf{r}_h\|_e \lesssim h_e^{-\frac{1}{2}}\|\mathbf{r}_h\|_K$,

$$(\mathbf{v}_h, \mathbf{r}_h)_K = \theta \sum_{e\in\mathcal{E}_K} w_e\langle \mathbf{r}_h\cdot\mathbf{n}_e, [\![u_h]\!]\rangle_e \lesssim \|\mathbf{r}_h\|_K \sum_{e\in\mathcal{E}_K} h_e^{-\frac{1}{2}}\|[\![u_h]\!]\|_e. \tag{8.51}$$

Combining (8.50) and (8.51) and using the bound (8.42), we arrive at

$$\|\nabla u_h + \boldsymbol{\sigma}_h\|_K \lesssim \left\{\sum_{e\in\mathcal{E}_K^{\mathrm{int}}} h_e\|[\![\nabla u_h]\!]\cdot\mathbf{n}_e\|_e^2\right\}^{\frac{1}{2}} + \left\{\sum_{e\in\mathcal{E}_K}\left((1+\alpha^2)h_e^{-1}\right)\|[\![u_h]\!]\|_e^2\right\}^{\frac{1}{2}}, \tag{8.52}$$

wherefrom (8.49) follows using the definitions (8.24a) of $\eta_{\mathrm{res},K}$ and (8.24b) of $|u_h|_{\mathrm{J},K}$. □

8.3.5 Mixed finite element method

Define $\mathbf{V}_h \times Q_h$ as in Section 7.13.4, with $k' = 0$ for simplicity. We immediately have:

Definition 8.3.12 (Flux reconstruction for the MFE method)**.** *Let $\bar{u}_h \in Q_h$ and $\boldsymbol{\sigma}_h \in \mathbf{V}_h$ be given by Definition 7.13.8. Take $\boldsymbol{\sigma}_h$ directly for the equilibrated flux reconstruction.*

As $\bar{u}_h$ is only piecewise constant in the lowest-order mixed finite element method (7.78a)–(7.78b), there holds $\nabla\bar{u}_h = 0$, where, recall ∇ is the broken weak gradient, see (4.7) and Remark 4.3.3. Consequently, it does not give much sense to estimate the energy error $\|\nabla(u - \bar{u}_h)\|$, as this is equal to $\|\nabla u\|$. For this reason, we first postprocess $\bar{u}_h \in \mathbb{P}_0(\mathcal{T}_h)$ into a more regular, higher-order polynomial function u_h, following [98, Section 4.1]. First define the space $\mathbb{P}_{1,2}(\mathcal{T}_h)$ as the space $\mathbb{P}_1(\mathcal{T}_h)$ enriched elementwise by the parabolas $\sum_{i=1}^d x_i^2$. Then set

$$-\nabla u_h|_K = \boldsymbol{\sigma}_h|_K \qquad \forall K \in \mathcal{T}_h, \tag{8.53a}$$

$$\frac{(u_h, 1)_K}{|K|} = \bar{u}_h|_K, \qquad \forall K \in \mathcal{T}_h. \tag{8.53b}$$

We consider henceforth $u_h \in \mathbb{P}_2(\mathcal{T}_h)$ as the approximate solution yielded by the mixed finite element method (7.78a)–(7.78b). This postprocessing is in general not included in $H_0^1(\Omega)$. We are thus lead to specify $s_h := \mathcal{I}_{\rm av}(u_h)$ for the potential reconstruction. With these constructions, we have:

Lemma 8.3.13 (Potential and flux reconstructions in the MFE method)**.** *Let $\bar{u}_h \in Q_h$ and $\boldsymbol{\sigma}_h \in \mathbf{V}_h$ be given by Definition 7.13.8, let $\boldsymbol{\sigma}_h$ be given by Definition 8.3.12, let u_h be given by (8.53a)–(8.53b), and let finally $s_h := \mathcal{I}_{\rm av}(u_h)$. Then s_h and $\boldsymbol{\sigma}_h$ satisfy respectively Definitions 7.6.1 and 7.6.2.*

The following important remark stems from the construction of our postprocessing u_h: by (8.53a), the flux estimators $\eta_{{\rm F},K}$ of (7.23b) are zero. This is once again in agreement with the "flux-conforming" nature of the mixed finite element method.

The postprocessing (8.53a)–(8.53b) also leads to the following observation: fix one face $e \in \mathcal{E}_h$ and choose the basis function $\mathbf{v}_e$ of $\mathbf{V}_h$ having nonzero normal trace only across this face, cf. Figure 5.3. It follows from Definition 8.3.12, (8.53a)–(8.53b), (7.78a), the fact that $\nabla{\cdot}\mathbf{v}_e$ is piecewise constant, and the fact that $\mathbf{v}_e$ is supported on the elements $\mathcal{T}_e$ sharing e that

$$-(\nabla u_h, \mathbf{v}_e)_{\mathcal{T}_e} - (u_h, \nabla{\cdot}\mathbf{v}_e)_{\mathcal{T}_e} = 0.$$

Using the Green theorem and the facts that $\mathbf{v}_e{\cdot}\mathbf{n}_e$ is constant on e and that $\mathbf{v}_e{\cdot}\mathbf{n}_{e'} = 0$ on $e' \neq e$, we arrive from this equality at

$$\langle [\![u_h]\!], 1\rangle_e = 0 \qquad \forall e \in \mathcal{E}_h. \tag{8.54}$$

This means that the postprocessed potential u_h has the mean value of the jump equal to zero on all the faces of $\mathcal{T}_h$. Alternatively, we can say that u_h has means of traces continuous on the interior faces of $\mathcal{T}_h$ and means of traces equal to zero on the boundary faces of $\mathcal{T}_h$. This result is very much useful for the efficiency analysis, see Lemma 8.2.5, which altogether gets trivial:

Lemma 8.3.14 (Efficiency of the MFE method for (7.1a)–(7.1b))**.** *Let $\bar{u}_h$ and $\boldsymbol{\sigma}_h$ be given by Definition 7.13.8, $\boldsymbol{\sigma}_h$ by Definition 8.3.12, and u_h by (8.53a)–(8.53b). Then*

$$\eta_{{\rm F},K} = 0 \tag{8.55}$$

for all $K \in \mathcal{T}_h$.

Note that thanks to (8.54), we actually have (8.41), see Remark 8.3.4.

8.3.6 Cell-centered finite volume method

Set $V_h := \mathbb{P}_0(\mathcal{T}_h)$. The cell-centered finite volume (CCFV) method reads:

Definition 8.3.15 (CCFV method for (7.1a)–(7.1b))**.** *Find $\bar{u}_h \in V_h$ such that*

$$\sum_{e \in \mathcal{E}_K} F_{K,e} = (f, 1)_K \qquad \forall K \in \mathcal{T}_h. \tag{8.56}$$

Here, $F_{K,e}$ is the approximate normal flux through the face e of an element K, expressed linearly from the elementwise values $\bar{u}_h$.

The above definition covers a broad spectrum of different finite volume methods, in that the form of the normal flux $F_{K,e}$ needs not be specified. The important point is that no additional condition is necessary in order to apply our a posteriori error estimation framework.

We next proceed similarly as for the mixed finite element method above. We first define, following Eymard *at al.* [59]:

Definition 8.3.16 (Elementwise flux prescription for the CCFV method)**.** *Let $\bar{u}_h \in V_h$ be given by Definition 8.3.15. Then prescribe $\boldsymbol{\sigma}_h \in \mathbf{RTN}_0(\mathcal{T}_h)$ by*

$$\langle \boldsymbol{\sigma}_h \cdot \mathbf{n}_K, 1 \rangle_e := F_{K,e} \qquad \forall K \in \mathcal{T}_h, \forall e \in \mathcal{E}_K. \tag{8.57}$$

Note that (8.57) and (8.56) together with the Green theorem imply (7.21b). Next we specify $u_h \in \mathbb{P}_{1,2}(\mathcal{T}_h)$ by (8.53a)–(8.53b). Note that, consequently, the flux estimators $\eta_{\mathrm{F},K}$ of (7.23b) are zero, as for the mixed finite element method. Finally, as the resulting approximation $u_h \notin H^1_0(\Omega)$, we set $s_h := \mathcal{I}_{\mathrm{av}}(u_h)$. It is interesting to note that in contrast to mixed finite elements, we do not have (8.54) here in general.

Altogether, these developments lead to:

Lemma 8.3.17 (Potential and flux reconstructions in the CCFV method)**.** *Let $\bar{u}_h \in V_h$ be given by Definition 8.3.15, let $\boldsymbol{\sigma}_h$ be given by Definition 8.3.16, let u_h be given by (8.53a)–(8.53b), and let finally $s_h := \mathcal{I}_{\mathrm{av}}(u_h)$. Then s_h and $\boldsymbol{\sigma}_h$ satisfy respectively Definitions 7.6.1 and 7.6.2.*

The efficiency result for the finite volume method is likewise straightforward:

Lemma 8.3.18 (Efficiency of the CCFV method for (7.1a)–(7.1b))**.** *Let $\bar{u}_h$ be given by Definition 8.3.15, $\boldsymbol{\sigma}_h$ by (8.57), and u_h by (8.53a)–(8.53b). Then*

$$\eta_{\mathrm{F},K} = 0 \tag{8.58}$$

for all $K \in \mathcal{T}_h$.

Unfortunately, in contrast to the mixed finite element case, one does not have here in general (8.54), although this property holds for $f = 0$. Thus, the final efficiency result is that of Theorem 8.3.3.

8.3.7 Vertex-centered finite volume method

In addition to $\mathcal{T}_h$, we will now also need a dual mesh $\mathcal{D}_h$. Every dual volume $D \in \mathcal{D}_h$ is associated with one vertex of $\mathcal{T}_h$ and constructed around this vertex by joining the face, edge, and element barycenters as indicated in Figure 8.2, left, for $d = 2$. By $\mathcal{D}_h^{\mathrm{int}}$, we denote those

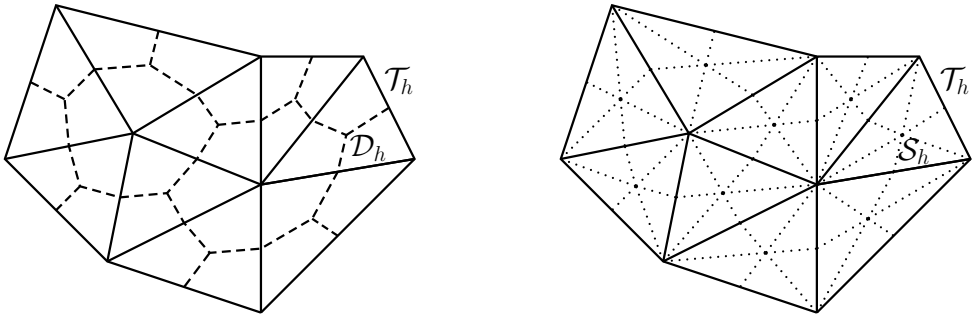


Figure 8.2: Simplicial mesh $\mathcal{T}_h$ and the dual mesh $\mathcal{D}_h$ (left); simplicial submesh $\mathcal{S}_h$ (right)

dual volumes associated with the interior vertices of $\mathcal{T}_h$ and by $\mathcal{D}_h^{\text{ext}}$ those associated with the boundary vertices of $\mathcal{T}_h$. We will then also introduce a second simplicial mesh $\mathcal{S}_h$, a submesh of both $\mathcal{T}_h$ and $\mathcal{D}_h$, constructed using the barycenters of the elements, faces, and edges of the mesh $\mathcal{T}_h$ as indicated in Figure 8.2, right, for $d = 2$. We will use the notation $\mathcal{S}_D$ for the submesh of the dual volume D by the simplices of $\mathcal{S}_h$, $\partial\mathcal{S}_D^{\text{int}}$ for the interior faces of $\mathcal{S}_D$, and $\partial\mathcal{S}_D^{\text{ext}}$ for the boundary faces of $\mathcal{S}_D$.

Set $V_h := \mathbb{P}_1(\mathcal{T}_h) \cap H_0^1(\Omega)$. The vertex-centered finite volume (VCFV) method reads:

Definition 8.3.19 (VCFV method for (7.1a)–(7.1b))**.** *Find $u_h \in V_h$ such that*

$$-\langle \nabla u_h \cdot \mathbf{n}_D, 1\rangle_{\partial D} = (f,1)_D \qquad \forall D \in \mathcal{D}_h^{\text{int}}. \tag{8.59}$$

In the vertex-centered finite volume method, the approximate potential u_h is conforming, $u_h \in H_0^1(\Omega)$, so that we simply set $s_h := u_h$ for the potential reconstruction. Note that, consequently, the nonconformity estimators $\eta_{\text{NC},K}$ of (7.23c) are zero, which is in agreement with this conforming nature of the vertex-centered finite volume method.

As for the flux reconstruction, we will construct $\boldsymbol{\sigma}_h \in \mathbf{RTN}_0(\mathcal{S}_h)$, where, recall, $\mathcal{S}_h$ is the simplicial submesh of both $\mathcal{T}_h$ and $\mathcal{D}_h$ depicted in Figure 8.2, right, for $d = 2$ and the space $\mathbf{RTN}_0(\mathcal{S}_h)$ is defined in Section 5.3. We proceed following Luce and Wohlmuth [72] and [99, 101]. For a given dual volume $D \in \mathcal{D}_h$, recall that $\mathcal{S}_D$ stands for the submesh of the dual volume D by the simplices of $\mathcal{S}_h$ and $\partial\mathcal{S}_D^{\text{ext}}$ for the faces of $\mathcal{S}_D$ lying in ∂D. Define the space

$$\mathbf{RTN}_0^{\text{N}}(\mathcal{S}_D) := \{\mathbf{v}_h \in \mathbf{RTN}_0(\mathcal{S}_D);\ \mathbf{v}_h \cdot \mathbf{n}_e = -\nabla u_h \cdot \mathbf{n}_e \quad \forall e \in \partial\mathcal{S}_D^{\text{ext}},\ e \not\subset \partial\Omega\}. \tag{8.60}$$

This is the space of Raviart–Thomas–Nédélec vector functions over the mesh $\mathcal{S}_D$ which are such that their normal components over those faces of $\partial\mathcal{S}_D^{\text{ext}}$ which do not lie at the boundary of Ω are given by the piecewise constant function $-\nabla u_h \cdot \mathbf{n}_e$. Note that $-\nabla u_h \cdot \mathbf{n}_e$ is univalued on such faces (since they always lie in the interior of some simplex K from the original mesh $\mathcal{T}_h$); consequently, any function $\mathbf{v}_h$ such that $\mathbf{v}_h|_D \in \mathbf{RTN}_0^{\text{N}}(\mathcal{S}_D)$ for all $D \in \mathcal{D}_h$ belongs to the space $\mathbf{RTN}_0(\mathcal{S}_h)$, as its normal component is continuous in the whole domain Ω, cf. Theorem 4.5.1. Thus assumption (7.21a) holds. Remark also that it follows from (8.59) that any such function $\mathbf{v}_h$ satisfies $\langle \mathbf{v}_h \cdot \mathbf{n}_D, 1\rangle_{\partial D} = (f,1)_D$ for all $D \in \mathcal{D}_h^{\text{int}}$ and, consequently, by the Green theorem

$$(\nabla \cdot \mathbf{v}_h, 1)_D = (f,1)_D \qquad \forall D \in \mathcal{D}_h^{\text{int}}. \tag{8.61}$$

Note that (8.61) is a local conservation property which is precisely behind the philosophy of the vertex-centered finite volume method (8.59). It follows in particular from (8.61) that $\mathbf{v}_h$

such that $\mathbf{v}_h|_D \in \mathbf{RTN}_0^{\mathrm{N}}(\mathcal{S}_D)$ "almost" satisfies assumption (7.21b), almost in the sense that the desired property holds for all interior dual volumes $D \in \mathcal{D}_h^{\mathrm{int}}$ but not for all elements K of the original mesh $\mathcal{T}_h$. We also observe that in the space $\mathbf{RTN}_0(\mathcal{S}_h)$, there are many additional degrees of freedom which have not been fixed yet. These are the degrees of freedom of the interior faces of $\mathcal{S}_D$, $D \in \mathcal{D}_h$, and the faces of $\mathcal{S}_h$ lying on the boundary. We will do so now, with the double objective to satisfy (7.21b) and to choose the remaining degrees of freedom in the best possible way. We actually proceed similarly to the developments of Section 7.9 but we remind that direct elementwise prescription is also possible following [101, 4.3.1].

Let $f_h \in \mathbb{P}_0(\mathcal{S}_h)$ be given by $(f,1)_K/|K|$ for all $K \in \mathcal{S}_h$. We then define:

Definition 8.3.20 (Flux reconstruction for the VCFV method)**.** *Let u_h be given by Definition 8.3.19. Then prescribe $\boldsymbol{\sigma}_h \in \mathbf{RTN}_0(\mathcal{S}_h)$ on each $D \in \mathcal{D}_h$ by*

$$\boldsymbol{\sigma}_h|_D := \arg \inf_{\mathbf{v}_h \in \mathbf{RTN}_0^{\mathrm{N}}(\mathcal{S}_D),\, \nabla\cdot\mathbf{v}_h = f_h} \|\nabla u_h + \mathbf{v}_h\|_D. \tag{8.62}$$

Problem (8.62) is again a complementary energy minimization problem, as that of Definition 7.9.1. Noting that $\nabla\cdot\boldsymbol{\sigma}_h = f_h$, i.e., $\nabla\cdot\boldsymbol{\sigma}_h|_K = f_h|_K$ for all $K \in \mathcal{S}_h$, the property (7.21b) immediately follows. From (8.62), we see that we impose a constraint that the residual estimators (7.23a) will be very small, as $f - \nabla\cdot\boldsymbol{\sigma}_h = f - f_h$; they will eventually disappear when $f = f_h$, i.e., whenever the source function f is piecewise constant on the mesh $\mathcal{S}_h$. Finally, the equilibrated flux $\boldsymbol{\sigma}_h$ that we find by (8.62) can be seen as a minimization of the flux estimators (7.23b) (with the constraint $\nabla\cdot\boldsymbol{\sigma}_h = f_h$). We summarize the above developments in the following:

Lemma 8.3.21 (Potential and flux reconstructions in the VCFV method)**.** *Let u_h be given by Definition 8.3.19. Then $s_h := u_h$ and $\boldsymbol{\sigma}_h$ prescribed by Definition 8.3.20 satisfy respectively Definitions 7.6.1 and 7.6.2.*

In order to practically compute, on a given dual volume $D \in \mathcal{D}_h$, the equilibrated flux $\boldsymbol{\sigma}_h$ of (8.62), we proceed as follows. We first define a new space $\mathbf{RTN}_0^{\mathrm{N},0}(\mathcal{S}_D)$ as the space $\mathbf{RTN}_0^{\mathrm{N}}(\mathcal{S}_D)$ of (8.60) but with the normal flux condition $\mathbf{v}_h\cdot\mathbf{n}_e = 0$ on all $e \in \partial\mathcal{S}_D^{\mathrm{ext}}$, $e \not\subset \partial\Omega$, for all the functions $\mathbf{v}_h$ from this space. For $D \in \mathcal{D}_h^{\mathrm{ext}}$, we then let $\mathbb{P}_0^*(\mathcal{S}_D)$ be spanned by piecewise constants on $\mathcal{S}_D$. For $D \in \mathcal{D}_h^{\mathrm{int}}$, the space $\mathbb{P}_0^*(\mathcal{S}_D)$ is spanned by piecewise constants on $\mathcal{S}_D$, with imposed zero mean value on the volume D. Then it is easy to show that (8.62) is equivalent to finding $\boldsymbol{\sigma}_h \in \mathbf{RTN}_0^{\mathrm{N}}(\mathcal{S}_D)$ and $r_h \in \mathbb{P}_0^*(\mathcal{S}_D)$ such that

$$(\boldsymbol{\sigma}_h, \mathbf{v}_h)_D - (r_h, \nabla\cdot\mathbf{v}_h)_D = -(\nabla u_h, \mathbf{v}_h)_D \qquad \forall \mathbf{v}_h \in \mathbf{RTN}_0^{\mathrm{N},0}(\mathcal{S}_D), \tag{8.63a}$$

$$(\nabla\cdot\boldsymbol{\sigma}_h, q_h)_D = (f, q_h)_D \qquad \forall q_h \in \mathbb{P}_0^*(\mathcal{S}_D). \tag{8.63b}$$

Note that (8.63a)–(8.63b) is the lowest-order Raviart–Thomas–Nédélec mixed finite element approximation of a local Neumann problem on the interior dual volumes $D \in \mathcal{D}_h^{\mathrm{int}}$; the Neumann boundary condition is given by $-\nabla u_h\cdot\mathbf{n}_e$. Note in particular that the function $-\nabla u_h\cdot\mathbf{n}_e$ on the boundary of each $D \in \mathcal{D}_h^{\mathrm{int}}$ by (8.59) satisfies the Neumann compatibility condition with the source term f, whence we can take all $q_h \in \mathbb{P}_0(\mathcal{S}_D)$ as test functions in (8.63b) and $\nabla\cdot\boldsymbol{\sigma}_h = f_h$ follows. On the boundary dual volumes $D \in \mathcal{D}_h^{\mathrm{ext}}$, (8.63a)–(8.63b) is the lowest-order Raviart–Thomas–Nédélec mixed finite element approximation of a local problem where the same Neumann boundary condition is imposed on that part of the boundary of D which lies inside Ω and the homogeneous Dirichlet boundary condition is imposed on the remaining part of the boundary of D.

The efficiency for the vertex-centered finite volume method is:

Lemma 8.3.22 (Efficiency of the VCFV method for (7.1a)–(7.1b)). *Let u_h be given by Definition 8.3.19 and $\boldsymbol{\sigma}_h$ by (8.63a)–(8.63b). Let Assumption 8.2.1 be satisfied, with more precisely $f \in \mathbb{P}_0(\mathcal{T}_h)$. Then*

$$\eta_{\mathrm{F},K} \lesssim \eta_{\mathrm{res},K} \tag{8.64}$$

for all $K \in \mathcal{T}_h$.

Proof. Let $D \in \mathcal{D}_h$ and recall that $\partial\mathcal{S}_D^{\mathrm{int}}$ stands for all the interior faces of $\mathcal{S}_D$. We will show

$$\|\nabla u_h + \boldsymbol{\sigma}_h\|_D \lesssim \left\{\sum_{K\in\mathcal{S}_D} h_K^2 \|f + \Delta u_h\|_K^2\right\}^{\frac12} + \left\{\sum_{e\in\partial\mathcal{S}_D^{\mathrm{int}}} h_e \|[\![\nabla u_h]\!]\cdot\mathbf{n}_e\|_e^2\right\}^{\frac12}. \tag{8.65}$$

Therefrom, (8.64) easily follows.

Let $D \in \mathcal{D}_h$ and let $\boldsymbol{\sigma}_h \in \mathbf{RTN}_0^{\mathrm{N}}(\mathcal{S}_D)$ and $r_h \in \mathbb{P}_0^*(\mathcal{S}_D)$ be given by (8.63a)–(8.63b). Following Arnold and Brezzi [12], Arbogast and Chen [10], and [98, Section 4.1], we define a postprocessing $\tilde r_h$ of the scalar function r_h such that

$$-\nabla\tilde r_h|_K = (\boldsymbol{\sigma}_h + \nabla u_h)|_K \qquad \forall K \in \mathcal{S}_D, \tag{8.66a}$$

$$\frac{(\tilde r_h, 1)_K}{|K|} = r_h|_K, \qquad \forall K \in \mathcal{S}_D. \tag{8.66b}$$

It follows from (8.66a)–(8.66b) and (8.63a) that

$$(\nabla\tilde r_h, \mathbf{v}_h)_D + (\tilde r_h, \nabla\cdot\mathbf{v}_h)_D = 0 \qquad \forall \mathbf{v}_h \in \mathbf{RTN}_0^{\mathrm{N},0}(\mathcal{S}_D).$$

Fixing one face $e \in \partial\mathcal{S}_D^{\mathrm{int}}$, choosing the basis functions of $\mathbf{RTN}_0^{\mathrm{N},0}(\mathcal{S}_D)$ having nonzero normal trace only across this face, and using the Green theorem, we arrive at

$$\langle[\![\tilde r_h]\!], 1\rangle_e = 0. \tag{8.67}$$

This means that the postprocessed function $\tilde r_h$ has the mean value of the jump equal to zero on the interior faces of $\mathcal{S}_D$. Alternatively, we can say that $\tilde r_h$ has means of traces continuous on the interior faces of $\mathcal{S}_D$. If $D \in \mathcal{D}_h^{\mathrm{ext}}$, we arrive similarly at

$$\langle\tilde r_h, 1\rangle_e = 0 \tag{8.68}$$

for all $e \in \partial\mathcal{S}_D^{\mathrm{ext}}$ such that $e \subset \partial\Omega$. Thus, on exterior faces of $\mathcal{S}_D$ belonging to $\partial\Omega$, the mean value of $\tilde r_h$ is zero. Finally, for $D \in \mathcal{D}_h^{\mathrm{int}}$, we have that $(r_h, 1)_D = 0$ from the definition of $\mathbb{P}_0^*(\mathcal{S}_D)$. From this fact and (8.66b), we deduce that

$$(\tilde r_h, 1)_D = 0 \tag{8.69}$$

on all $D \in \mathcal{D}_h^{\mathrm{int}}$. Thus, on dual volumes not touching the boundary, the mean value of $\tilde r_h$ is zero.

We denote by $M(\mathcal{S}_D) \subset \mathbb{P}_{1,2}(\mathcal{S}_D)$ the corresponding space of polynomials verifying (8.67), (8.68), and (8.69). Using the above developments, we have

$$\|\nabla u_h + \boldsymbol{\sigma}_h\|_D = \sup_{m_h\in M(\mathcal{S}_D),\, \|\nabla m_h\|_D=1} (\nabla u_h + \boldsymbol{\sigma}_h, \nabla m_h)_D. \tag{8.70}$$

We now develop the right-hand side of (8.70). Using the Green theorem, the fact that $\nabla\cdot\boldsymbol{\sigma}_h = f_h = f$ for all $K \in \mathcal{S}_D$, see (8.62) or (8.63b), (8.67) (with $\tilde r_h$ replaced by m_h) and the facts that

$((\nabla u_h+\boldsymbol{\sigma}_h)\cdot\mathbf{n}_e)|_e$ is in $\mathbb{P}_0(e)$ and that $[\![\boldsymbol{\sigma}_h]\!]\cdot\mathbf{n}_e|_e=0$ for all faces $e\in\partial\mathcal{S}_D^{\rm int}$ as $\boldsymbol{\sigma}_h\in\mathbf{RTN}_0^{\rm N}(\mathcal{S}_D)$, we arrive at

$$\begin{aligned}&(\nabla u_h+\boldsymbol{\sigma}_h,\nabla m_h)_D\\ &=\sum_{K\in\mathcal{S}_D}\{-(m_h,\nabla\cdot(\nabla u_h+\boldsymbol{\sigma}_h))_K+\langle(\nabla u_h+\boldsymbol{\sigma}_h)\cdot\mathbf{n}_K,m_h\rangle_{\partial K}\}\\ &=-\sum_{K\in\mathcal{S}_D}(m_h,f+\Delta u_h)_K+\sum_{e\in\partial\mathcal{S}_D^{\rm int}}\langle[\![\nabla u_h]\!]\cdot\mathbf{n}_e,m_h\rangle_e.\end{aligned}\tag{8.71}$$

We have also used that $\boldsymbol{\sigma}_h\cdot\mathbf{n}_e=-\nabla u_h\cdot\mathbf{n}_e$ for all boundary faces e of $\mathcal{S}_D$ not included in $\partial\Omega$ since $\boldsymbol{\sigma}_h\in\mathbf{RTN}_0^{\rm N}(\mathcal{S}_D)$, and (8.68) for all boundary faces e of $\mathcal{S}_D$ included in $\partial\Omega$. By the Cauchy–Schwarz inequality, the inverse inequality $\|m_h\|_e\lesssim h_e^{-\frac12}\|m_h\|_K$, and the shape-regularity of $\mathcal{T}_h$ (and consequently $\mathcal{S}_h$), we can further estimate

$$\begin{aligned}&(\nabla u_h+\boldsymbol{\sigma}_h,\nabla m_h)_D\\ &\le\left\{\sum_{K\in\mathcal{S}_D}h_K^{-2}\|m_h\|_K^2\right\}^{\frac12}\left\{\sum_{K\in\mathcal{S}_D}h_K^2\|f+\Delta u_h\|_K^2\right\}^{\frac12}\\ &\quad+\left\{\sum_{e\in\partial\mathcal{S}_D^{\rm int}}h_e^{-1}\|m_h\|_e^2\right\}^{\frac12}\left\{\sum_{e\in\partial\mathcal{S}_D^{\rm int}}h_e\|[\![\nabla u_h]\!]\cdot\mathbf{n}_e\|_e^2\right\}^{\frac12}\\ &\lesssim h_D^{-1}\|m_h\|_D\left\{\sum_{K\in\mathcal{S}_D}h_K^2\|f+\Delta u_h\|_K^2+\sum_{e\in\partial\mathcal{S}_D^{\rm int}}h_e\|[\![\nabla u_h]\!]\cdot\mathbf{n}_e\|_e^2\right\}^{\frac12}.\end{aligned}$$

Recall that, as $m_h\in M(\mathcal{S}_D)$, we have (8.68) or (8.69) for m_h. Thus, the broken Poincaré or Friedrichs inequalities (4.23) and (4.24), that we apply on the mesh $\mathcal{S}_D$ of D, give

$$\|m_h\|_D\lesssim h_D\|\nabla m_h\|_D.$$

Consequently, (8.65) follows from the above estimates and (8.70). □

Remark 8.3.23 (Equilibrated flux reconstruction for lowest-order conforming finite elements)**.** *It can be shown, see the references in [101] or Lemmas 3.8 and 3.11 in [101], that the finite element method of Definition 7.13.1 and the vertex-centered finite volume method of Definition 8.3.19 coincide when $k=1$ and $f\in\mathbb{P}_0(\mathcal{T}_h)$. In order to obtain an equilibrated flux reconstruction $\boldsymbol{\sigma}_h$ for lowest-order conforming finite elements, we can thus also proceed as in this section.*

8.4 Numerical examples

We present finally the results of several numerical experiments illustrating the performance of the simplified reconstruction estimators of Section 8.3. We define the effectivity index by

$$I_{\rm eff}:=\frac{\left\{\sum_{K\in\mathcal{T}_h}(\eta_{{\rm F},K}+\eta_{{\rm R},K})^2+\sum_{K\in\mathcal{T}_h}\eta_{{\rm NC},K}^2\right\}^{\frac12}}{\|\nabla(u-u_h)\|}.$$

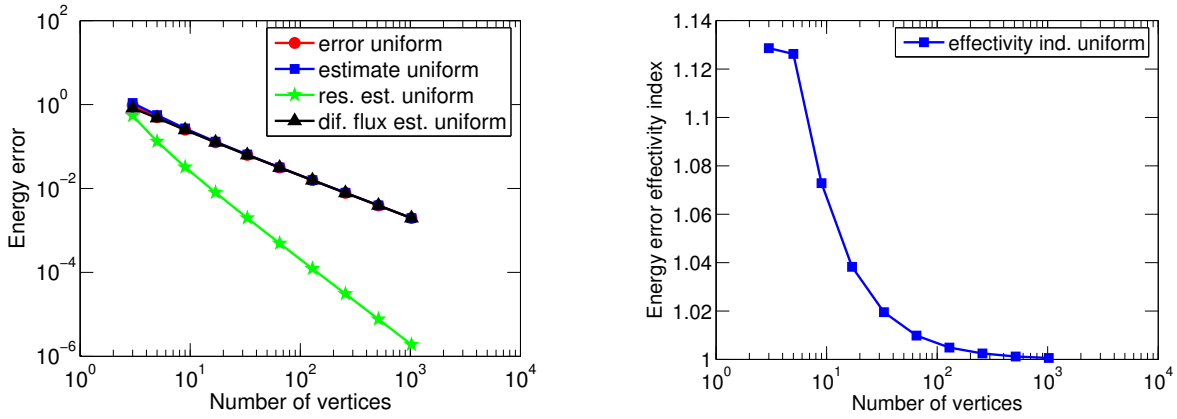


Figure 8.3: Estimated and actual energy error and the corresponding effectivity index, vertex-centered finite volume method, $d = 1$

8.4.1 Vertex-centered finite volume method in one space dimension

Let us first consider the one-dimensional case, i.e., (2.1a)–(2.1b). We will reuse Example 2.3.4, i.e., we take $\Omega = (0,1)$ and $f = \pi^2 \sin(\pi x)$, which leads to $u = \sin(\pi x)$. We consider the vertex-centered finite volume method of Definition 8.3.19 and report the results in Figure 8.3. In its left part, we see that the estimate is indeed guaranteed, which illustrates that property **i)** of Section 1.4 is satisfied. Indeed, the effectivity index, given in the right part of Figure 8.3, is above one. We can also remark that in this one-dimensional case, we have asymptotic exactness, i.e., property **iii)** of Section 1.4 is satisfied. As discussed in Section 8.3.7, the nonconformity estimators $\eta_{\mathrm{NC},K}$ of (7.23c) are zero. In correspondence with the fact that $\nabla{\cdot}\boldsymbol{\sigma}_h = f_h$, the residual estimators $\eta_{\mathrm{R},K}$ of (7.23a) are very small here. In fact, as $f \in H^1(\mathcal{T}_h)$, we obtain

$$\frac{h_K}{\pi}\|f - f_h\|_K \leq \frac{h_K^2}{\pi^2}\|\nabla f\|_K$$

by the Poincaré inequality (4.20), so that these estimators converge as $\mathcal{O}(h^2)$, which is illustrated in the left part of Figure 8.3. Consequently, $\eta_{\mathrm{R},K}$ are negligible on refined meshes and the principal component of the a posteriori estimate are the flux estimators $\eta_{\mathrm{F},K}$ of (7.23b).

8.4.2 Cell-centered finite volume method

We consider here the following slight modification of problem (7.1a)–(7.1b): for $g \in H^{\frac{1}{2}}(\partial\Omega)$, find u such that

$$-\nabla{\cdot}(\underline{\mathbf{K}}\nabla u) = 0 \qquad \text{in } \Omega, \tag{8.72a}$$

$$u = g \qquad \text{on } \partial\Omega. \tag{8.72b}$$

We take $\Omega = (-1,1)\times(-1,1)$, divided into four subdomains Ω_i along the Cartesian axes, with $\underline{\mathbf{K}}|_{\Omega_i} = a^i\,\mathbb{I}$, where $\mathbb{I}$ is the identity matrix. We take either $a^1 = a^3 = 5$, $a^2 = a^4 = 1$ or $a^1 = a^3 = 100$, $a^2 = a^4 = 1$, so that a weak solution of (8.72a)–(8.72b) has a singularity in the origin.

We consider the cell-centered finite volume method of Definition 8.3.15. We know that in this case both the residual estimators $\eta_{\mathrm{R},K}$ of (7.23a) and the flux estimators $\eta_{\mathrm{F},K}$ of (7.23b) are zero, so that the only component of the a posteriori estimate are the nonconformity estimators

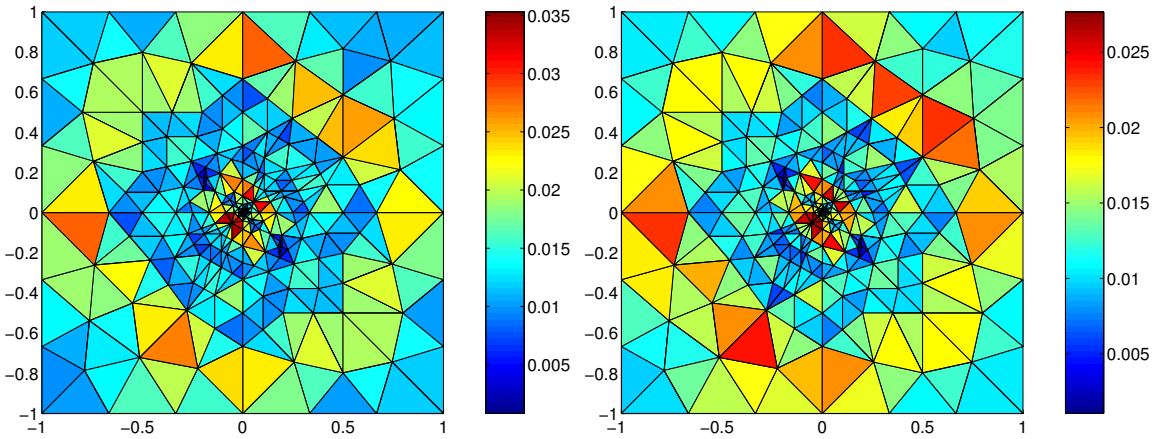


Figure 8.4: Estimated (left) and actual (right) energy error distribution, cell-centered finite volume method

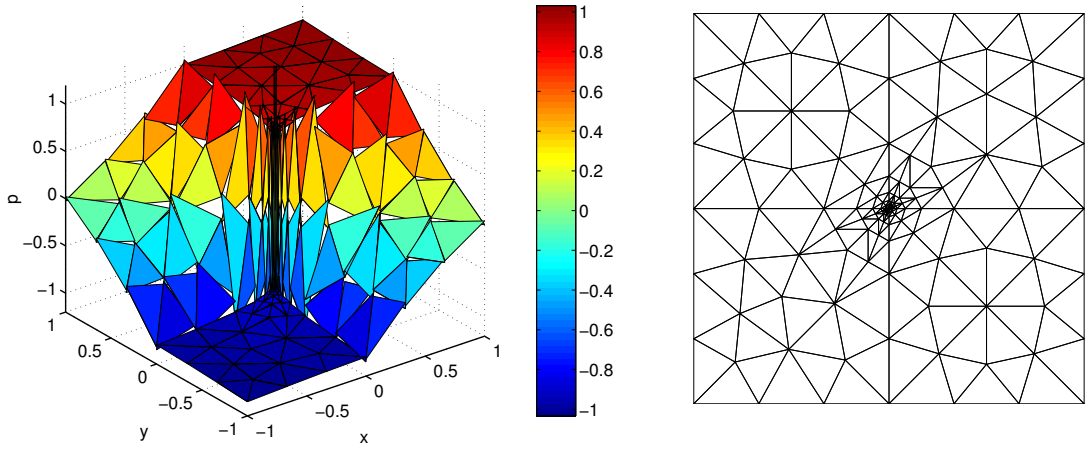


Figure 8.5: Approximate solution and the corresponding adaptively refined mesh, cell-centered finite volume method

$\eta_{\mathrm{NC},K}$ which take here the form $\eta_{\mathrm{NC},K} := \|\underline{\mathbf{K}}^{\frac{1}{2}}\nabla(u_h - s_h)\|_K$ instead of (7.23c). We show in the left part of Figure 8.4 ($a^1 = a^3 = 5$) the estimated error distribution, i.e., the values $\eta_{\mathrm{NC},K}$ for each $K \in \mathcal{T}_h$, and in the right part of Figure 8.4 the exact error distribution, i.e., the values $\|\underline{\mathbf{K}}^{\frac{1}{2}}\nabla(u - u_h)\|_K$ for each $K \in \mathcal{T}_h$. We see that the two plots match nicely, which is a numerical evidence of the local efficiency, property **ii)** of Section 1.4. We thus can do the adaptive refinement of only those mesh elements where the error is increased. The approximate solution on a refined mesh and the corresponding mesh are given in Figure 8.5 ($a^1 = a^3 = 100$). We can see that we can efficiently approximate the singularity at the origin, which would not be possible on a uniformly refined mesh. Finally, in Figure 8.6 ($a^1 = a^3 = 5$), we plot the dependence of the error and estimates on the number of mesh elements for both the uniform and adaptive mesh refinement. We can see that the error decreases with much higher (optimal) rate in the adaptive regime. Right part of Figure 8.6 then shows that also in multiple space dimensions, the effectivity indices of our a posteriori error estimates are quite close to the optimal value of one, i.e, close to the asymptotic exactness.

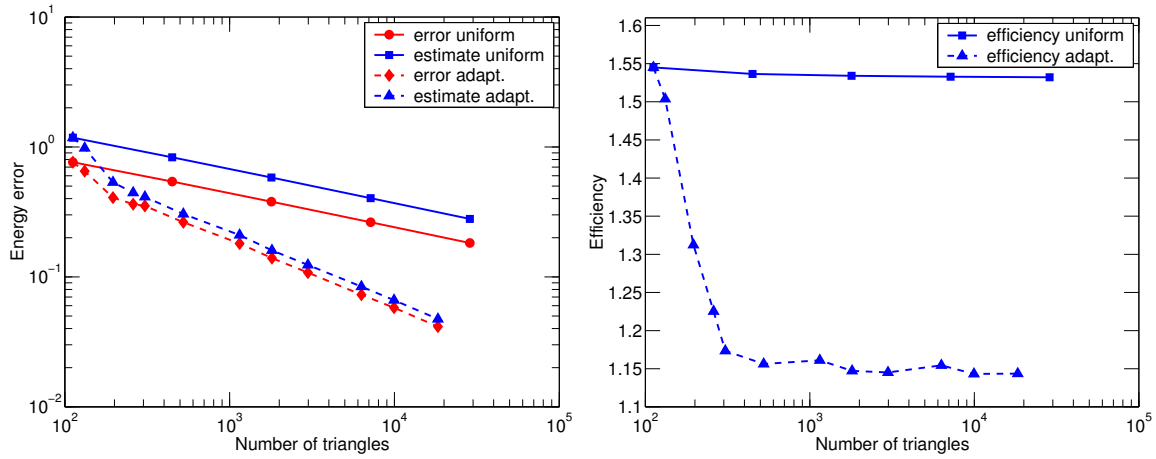


Figure 8.6: Estimated and actual energy errors and the corresponding effectivity indices, cell-centered finite volume method

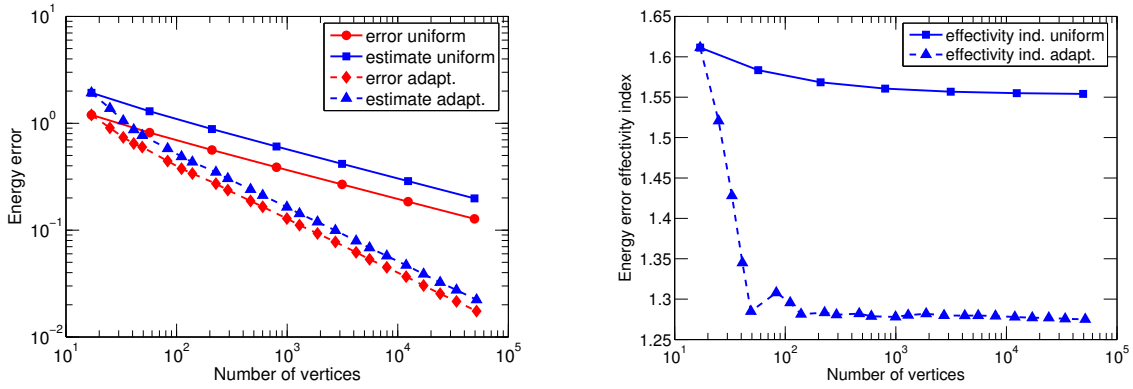


Figure 8.7: Estimated and actual energy error and the corresponding effectivity index, finite element method, $a^1 = a^3 = 5$

8.4.3 Finite element method

We show in Figure 8.7 the counterpart of Figure 8.6 for the finite element method of Definition 7.13.1. We compute the flux reconstruction following Remark 8.3.23. Similar conclusions as in the previous cases can be drawn. In addition to this situation $a^1 = a^3 = 5$, we present in Figure 8.8 the same results for the case with the increased contrast in the coefficient $\underline{\mathbf{K}}$, $a^1 = a^3 = 100$. Although for the adaptive mesh refinement with a sufficient number of mesh elements, the effectivity index is once again close to 1, this is not anymore the case for uniform mesh refinement. This is a typical example of non robustness, where property **iv)** of Section 1.4 is not satisfied.

8.4.4 Conclusions

The presented numerical experiments testify that our estimates satisfy properties **i)**, **ii)**, (approximately) **iii)**, and **v)** of Section 1.4. The robustness property **iv)** is not satisfied with respect to the variations in the diffusion tensor $\underline{\mathbf{K}}$ (remedies can be found in [101]). Altogether, the presented estimates enable the error control and error localization in the sense of

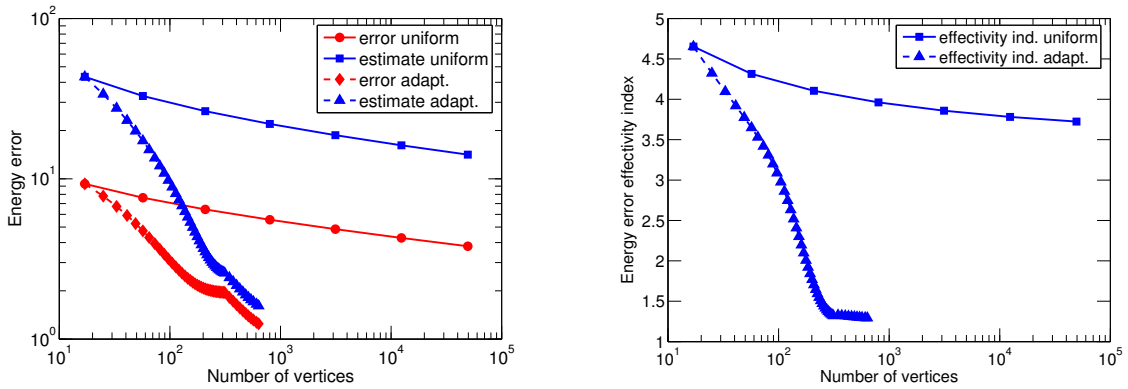


Figure 8.8: Estimated and actual energy error and the corresponding effectivity index, finite element method, $a^1 = a^3 = 100$

the properties **i)**–**ii)** and particularly precision attainment and efficiency in the sense of the properties **1.**–**2.** of the Introduction.

Chapter 9

The advection–diffusion–reaction equation

We investigate here the advection–diffusion–reaction equation (1.2a)–(1.2b). It reads: for $f \in L^2(\Omega)$, $r \in L^\infty(\Omega)$, and $\mathbf{w} \in [W^{1,\infty}(\Omega)]^d$ such that $\frac{1}{2}\nabla\cdot\mathbf{w} + r \geq 0$ and symmetric $\underline{\mathbf{K}} \in [L^\infty(\Omega)]^{d\times d}$ with uniformly positive smallest eigenvalue, find u such that

$$-\nabla\cdot(\underline{\mathbf{K}}\nabla u) + \nabla\cdot(\mathbf{w}u) + ru = f \qquad \text{in } \Omega, \tag{9.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{9.1b}$$

9.1 Variational formulation

In order to pose properly (9.1a)–(9.1b), we are led to the variational formulation:

Definition 9.1.1 (Variational formulation of (9.1a)–(9.1b))**.** *Find $u \in H_0^1(\Omega)$ such that*

$$(\underline{\mathbf{K}}\nabla u, \nabla v) - (\mathbf{w}u, \nabla v) + (ru, v) = (f, v) \qquad \forall v \in H_0^1(\Omega). \tag{9.2}$$

The existence and uniqueness of a solution of (9.2) is ensured by the Lax–Milgram theorem.

We now proceed as in Chapter 7. We first make the following equivalent of Definition 7.1.2:

Definition 9.1.2 (Flux)**.** *Let u be the solution of (9.2). Set*

$$\boldsymbol{\sigma} := -\underline{\mathbf{K}}\nabla u + \mathbf{w}u. \tag{9.3}$$

We will call $\boldsymbol{\sigma}$ the flux*.*

In analogy with Theorem 7.1.3, we have:

Theorem 9.1.3 (Properties of the weak solution of (9.1a)–(9.1b))**.** *Let u be the solution of (9.2). Let $\boldsymbol{\sigma}$ be given by (9.3). Then*

$$u \in H_0^1(\Omega), \quad \boldsymbol{\sigma} \in \mathbf{H}(\mathrm{div}, \Omega), \quad \nabla\cdot\boldsymbol{\sigma} = f - ru.$$

Proof. The weak solution u belongs to $H_0^1(\Omega)$ by definition. In order to verify that $\boldsymbol{\sigma} \in \mathbf{H}(\mathrm{div}, \Omega)$, we need to check the three conditions of Definition 4.2.1. Condition 1 is obvious, as $u \in H_0^1(\Omega)$, $\underline{\mathbf{K}} \in [L^\infty(\Omega)]^{d\times d}$, and $\mathbf{w} \in [W^{1,\infty}(\Omega)]^d$, so that $-\underline{\mathbf{K}}\nabla u + \mathbf{w}u = \boldsymbol{\sigma}$ is square-integrable. For the function w of condition 2a, choose $w := f - ru$ and note that $f \in L^2(\Omega)$ by our assumption and $ru \in L^2(\Omega)$ as $r \in L^\infty(\Omega)$ and $u \in H_0^1(\Omega)$, so that $w \in L^2(\Omega)$. Then condition 2b follows immediately from (9.2) and the fact that $\mathcal{D}(\Omega) \subset H_0^1(\Omega)$. □

9.2 Approximate solution

In order to make the presentation general as in Chapter 7, we are led to suppose in this section that the approximate solution u_h that we are given satisfies

$$u_h \in H^1(\mathcal{T}_h), \tag{9.4}$$

where $H^1(\mathcal{T}_h)$ is the broken Sobolev space of Definition 4.3.1.

In analogy with Definition 7.2.1, we set:

Definition 9.2.1 (Approximate flux)**.** *Let u_h be the approximate solution, cf. (9.4). We will call*

$$-\underline{\mathbf{K}}\nabla u_h + \mathbf{w}u_h \tag{9.5}$$

the approximate flux*.*

The following remark should be compared to Theorem 9.1.3:

Remark 9.2.2 (Properties of the approximate solution u_h of (9.4))**.** *Let u_h be the approximate solution, cf. (9.4). Then*

$$u_h \notin H^1_0(\Omega), \quad -\underline{\mathbf{K}}\nabla u_h + \mathbf{w}u_h \notin \mathbf{H}(\mathrm{div},\Omega), \quad \nabla\cdot(-\underline{\mathbf{K}}\nabla u_h + \mathbf{w}u_h) \neq f - ru_h \quad \textit{in general.}$$

9.3 Potential and flux reconstructions

From Theorem 9.1.3 and Remark 9.2.2, we see that the approximate solution (or approximate potential) u_h and the approximate flux $-\underline{\mathbf{K}}\nabla u_h + \mathbf{w}u_h$ can be nonphysical. As in Section 7.6, we will introduce their "corrections", a potential reconstruction s_h and an equilibrated flux reconstruction $\boldsymbol{\sigma}_h$:

Definition 9.3.1 (Potential reconstruction)**.** *Let u_h be the approximate solution, cf. (9.4). We will call the* potential reconstruction *any function s_h constructed from u_h which satisfies*

$$s_h \in H^1_0(\Omega).$$

Definition 9.3.2 (Equilibrated flux reconstruction)**.** *We will call the* equilibrated flux reconstruction *any function $\boldsymbol{\sigma}_h$ constructed from u_h which satisfies*

$$\boldsymbol{\sigma}_h \in \mathbf{H}(\mathrm{div},\Omega), \tag{9.6a}$$

$$(\nabla\cdot\boldsymbol{\sigma}_h + ru_h, 1)_K = (f,1)_K \qquad \forall K \in \mathcal{T}_h. \tag{9.6b}$$

9.4 Energy (semi-)norm augmented by a dual norm and its equivalence with the dual norm of the residual

Define two bilinear forms

$$\mathcal{B}_{\mathrm{S}}(u,v) := (\underline{\mathbf{K}}\nabla u, \nabla v) + ((\tfrac{1}{2}\nabla\cdot\mathbf{w} + r)u, v), \tag{9.7a}$$

$$\mathcal{B}_{\mathrm{A}}(u,v) := -(\mathbf{w}u, \nabla v) - ((\tfrac{1}{2}\nabla\cdot\mathbf{w})u, v). \tag{9.7b}$$

The energy (semi-)norm for the problem (9.1a)–(9.1b) is given by

$$|||v||| := \mathcal{B}_{\mathrm{S}}(v,v)^{\frac{1}{2}} = \left\{ \sum_{K\in\mathcal{T}_h} |||v|||_K^2 \right\}^{\frac{1}{2}}, \qquad v \in H^1(\mathcal{T}_h), \tag{9.8a}$$

$$|||v|||_K := \left\{ \|\underline{\mathbf{K}}^{\frac{1}{2}}\nabla v\|_K^2 + \|(\tfrac{1}{2}\nabla\cdot\mathbf{w} + r)^{\frac{1}{2}} v\|_K^2 \right\}^{\frac{1}{2}}. \tag{9.8b}$$

It appears that, in contrast to Chapter 7, it is not obvious to give optimal (in the sense of Section 1.4, in particular in what concerns property **iv)**) a posteriori error estimates in the energy norm (9.8a). We thus, following Verfürth [96] and [49] introduce the following augmented (semi-)norm:

$$|||v|||_\oplus := |||v||| + \sup_{\varphi\in H_0^1(\Omega),\, |||\varphi|||=1} \mathcal{B}_{\mathrm{A}}(v,\varphi) \qquad v \in H^1(\mathcal{T}_h). \tag{9.9}$$

Let us denote by $\mathcal{B}$ the bilinear form appearing in (9.2), i.e.,

$$\mathcal{B}(u,v) := (\underline{\mathbf{K}}\nabla u, \nabla v) - (\mathbf{w}u, \nabla v) + (ru, v).$$

Note that it follows from (9.7a)–(9.7b) that

$$\mathcal{B}(u,v) = \mathcal{B}_{\mathrm{S}}(u,v) + \mathcal{B}_{\mathrm{A}}(u,v). \tag{9.10}$$

With this notation, we have the following important result on relating the augmented norm of (9.9) with the dual norm generated by the form $\mathcal{B}$:

Theorem 9.4.1 (Equivalence of the augmented norm and of the dual norm of $\mathcal{B}$)**.** *Let $v \in H_0^1(\Omega)$. Then*

$$\sup_{\varphi\in H_0^1(\Omega);\, |||\varphi|||=1} \mathcal{B}(v,\varphi) \le |||v|||_\oplus \le 3 \sup_{\varphi\in H_0^1(\Omega);\, |||\varphi|||=1} \mathcal{B}(v,\varphi).$$

Proof. Let $\varphi \in H_0^1(\Omega)$ be given. Then

$$\mathcal{B}(v,\varphi) = \mathcal{B}_{\mathrm{S}}(v,\varphi) + \mathcal{B}_{\mathrm{A}}(v,\varphi) \le |||v||| \, |||\varphi||| + \mathcal{B}_{\mathrm{A}}(v,\varphi)$$

by (9.10), (9.7a), the Cauchy–Schwarz inequality, and (9.8a). Consequently,

$$\sup_{\varphi\in H_0^1(\Omega);\, |||\varphi|||=1} \mathcal{B}(v,\varphi) \le |||v||| + \sup_{\varphi\in H_0^1(\Omega);\, |||\varphi|||=1} \mathcal{B}_{\mathrm{A}}(v,\varphi) = |||v|||_\oplus.$$

Conversely,

$$|||v|||^2 = \mathcal{B}_{\mathrm{S}}(v,v) = \mathcal{B}(v,v),$$

which follows from (9.8a) and from the fact that

$$\mathcal{B}_{\mathrm{A}}(v,v) = 0.$$

Thus

$$|||v||| \le \sup_{\varphi\in H_0^1(\Omega);\, |||\varphi|||=1} \mathcal{B}(v,\varphi). \tag{9.11}$$

Let next $\varphi \in H_0^1(\Omega)$ with $|||\varphi||| = 1$ be given. Then by the same reasoning as above,

$$\mathcal{B}_{\mathrm{A}}(v,\varphi) = \mathcal{B}(v,\varphi) - \mathcal{B}_{\mathrm{S}}(v,\varphi) \le \mathcal{B}(v,\varphi) + |||v||| \, |||\varphi||| = \mathcal{B}(v,\varphi) + |||v|||.$$

Thus

$$\sup_{\varphi\in H_0^1(\Omega);\, |||\varphi|||=1} \mathcal{B}_{\mathrm{A}}(v,\varphi) \le \sup_{\varphi\in H_0^1(\Omega);\, |||\varphi|||=1} \mathcal{B}(v,\varphi) + |||v|||. \tag{9.12}$$

Combining (9.11) with (9.12) and the definition (9.9) gives the assertion of the theorem. □

Let us now generalize Definition 7.7.1 to the advection–diffusion–reaction setting:

Definition 9.4.2 (Residual)**.** *Let $v_h \in H^1_0(\Omega)$. Then $\mathcal{R}(v_h) \in H^{-1}(\Omega)$ is defined by*

$$\langle \mathcal{R}(v_h), \varphi \rangle_{H^{-1}(\Omega),H^1_0(\Omega)} := (f,\varphi) - \mathcal{B}(v_h,\varphi) \qquad \varphi \in H^1_0(\Omega).$$

Taking $u - v_h$ in place of v in Theorem 9.4.1 and noting that $\mathcal{B}(u - v_h, \varphi)$ equals $(f,\varphi) - \mathcal{B}(v_h,\varphi)$ by (9.2) leads to the following generalization of Theorem 7.7.2:

Corollary 9.4.3 (Equivalence between the augmented energy and dual residual norms)**.** *Let u be the weak solution given by Definition 9.1.1. Let $v_h \in H^1_0(\Omega)$ be arbitrary. Then*

$$\sup_{\varphi\in H^1_0(\Omega);\,|||\varphi|||=1} \{(f,\varphi) - \mathcal{B}(v_h,\varphi)\} \le |||u - v_h|||_\oplus \le 3 \sup_{\varphi\in H^1_0(\Omega);\,|||\varphi|||=1} \{(f,\varphi) - \mathcal{B}(v_h,\varphi)\}. \quad (9.13)$$

It is to be noted that in contrast to (7.22), here the norm on the test functions φ is the energy norm (9.8a) and not the $H^1_0(\Omega)$ seminorm. In conclusion, it is in the augmented norm $|||\cdot|||_\oplus$ that one obtains the *equivalence* with the dual norm of the residual.

9.5 A general posteriori error estimate

In the sequel, let for simplicity $(\frac{1}{2}\nabla\cdot\mathbf{w} + r)$ be piecewise constant. For any $K \in \mathcal{T}_h$, we will need the two following constants:

$$m_K := \min(C_{\mathrm{P},K} c_{\underline{\mathbf{K}},K}^{-\frac{1}{2}} h_K, (\tfrac{1}{2}\nabla\cdot\mathbf{w} + r)|_K^{-\frac{1}{2}}),$$
$$\widetilde{m}_K := 2(1 + C_{\mathrm{P},K}) c_{\underline{\mathbf{K}},K}^{-\frac{1}{2}} m_K,$$

with $c_{\underline{\mathbf{K}},K}$ the smallest eigenvalue of $\underline{\mathbf{K}}$ on K and $C_{\mathrm{P},K}$ the constant from the Poincaré inequality (4.20). Recall also the notation Π_0 for the $L^2(\Omega)$-orthogonal projection onto $\mathbb{P}_0(\mathcal{T}_h)$ and the notation I for the identity operator. Then we have the following a posteriori error estimate:

Theorem 9.5.1 (A general a posteriori error estimate for (9.1a)–(9.1b))**.** *Let u be the weak solution given by Definition 9.1.1. Let u_h be an arbitrary function satisfying (9.4). Let finally s_h be a potential reconstruction in the sense of Definition 9.3.1 and $\boldsymbol{\sigma}_h$ an equilibrated flux reconstruction in the sense of Definition 9.3.2. For any $K \in \mathcal{T}_h$, define the* residual estimator *by*

$$\eta_{\mathrm{R},K} := m_K \|f - \nabla\cdot\boldsymbol{\sigma}_h - r u_h\|_K, \quad (9.14)$$

the flux estimator *by*

$$\eta_{\mathrm{F},K} := \min(\eta_{\mathrm{F},1,K}, \eta_{\mathrm{F},2,K}), \quad (9.15a)$$
$$\eta_{\mathrm{F},1,K} := \left\{ \|\underline{\mathbf{K}}^{-\frac{1}{2}} \mathbf{a}_h\|_K^2 + \|(\tfrac{1}{2}\nabla\cdot\mathbf{w} + r)^{-\frac{1}{2}} (\tfrac{1}{2}\nabla\cdot\mathbf{w})(u_h - s_h)\|_K^2 \right\}^{\frac{1}{2}}, \quad (9.15b)$$
$$\eta_{\mathrm{F},2,K} := m_K \|(I - \Pi_0)\nabla\cdot\mathbf{a}_h\|_K + \widetilde{m}_K^{\frac{1}{2}} \sum_{e\in\mathcal{E}_K} \widetilde{C}_{\mathrm{t},K,e}^{\frac{1}{2}} \|\mathbf{a}_h\cdot\mathbf{n}_K\|_e$$
$$+ \|(\tfrac{1}{2}\nabla\cdot\mathbf{w} + r)^{-\frac{1}{2}} (\tfrac{1}{2}\nabla\cdot\mathbf{w})(u_h - s_h)\|_K, \quad (9.15c)$$

with $\mathbf{a}_h := \boldsymbol{\sigma}_h + \underline{\mathbf{K}}\nabla u_h - \mathbf{w}s_h$ *and* $\widetilde{C}_{\mathrm{t},K,e}$ *the constant from* (4.22a)*, and the* nonconformity estimators *by*

$$\eta_{\mathrm{NC},K} := |||u_h - s_h|||_K, \tag{9.16a}$$

$$\widetilde{\eta}_{\mathrm{NC},K} := \min(\widetilde{\eta}_{\mathrm{NC},1,K}, \widetilde{\eta}_{\mathrm{NC},2,K}), \tag{9.16b}$$

$$\widetilde{\eta}_{\mathrm{NC},1,K} := \left\{ \|\underline{\mathbf{K}}^{-\frac{1}{2}}\mathbf{b}_h\|_K^2 + \|(\tfrac{1}{2}\nabla\cdot\mathbf{w} + r)^{-\frac{1}{2}}(\tfrac{1}{2}\nabla\cdot\mathbf{w})(u_h - s_h)\|_K^2 \right\}^{\frac{1}{2}}, \tag{9.16c}$$

$$\begin{aligned}\widetilde{\eta}_{\mathrm{NC},2,K} :=& m_K\|(I - \Pi_0)\nabla\cdot\mathbf{b}_h\|_K + \widetilde{m}_K^{\frac{1}{2}} \sum_{e\in\mathcal{E}_K} \widetilde{C}_{\mathrm{t},K,e}^{\frac{1}{2}}\|\mathbf{b}_h\cdot\mathbf{n}_K\|_e \\ &+ \|(\tfrac{1}{2}\nabla\cdot\mathbf{w} + r)^{-\frac{1}{2}}(\tfrac{1}{2}\nabla\cdot\mathbf{w})(u_h - s_h)\|_K,\end{aligned} \tag{9.16d}$$

with $\mathbf{b}_h := \mathbf{w}(u_h - s_h)$*. Then*

$$\begin{aligned}|||u - u_h|||_\oplus \le 4 &\left\{ \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{NC},K}^2 \right\}^{\frac{1}{2}} + \left\{ \sum_{K\in\mathcal{T}_h} \widetilde{\eta}_{\mathrm{NC},K}^2 \right\}^{\frac{1}{2}} \\ &+ 3\left\{ \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{R},K} + \eta_{\mathrm{F},K})^2 \right\}^{\frac{1}{2}}.\end{aligned} \tag{9.17}$$

Proof. The triangle inequality gives

$$|||u - u_h|||_\oplus \le |||u - s_h|||_\oplus + |||s_h - u_h|||_\oplus,$$

whereas Theorem 9.4.1 for $v := u - s_h$ yields

$$|||u - s_h|||_\oplus \le 3 \sup_{\varphi\in H_0^1(\Omega);\, |||\varphi|||=1} \mathcal{B}(u - s_h, \varphi).$$

We finally employ that

$$\begin{aligned}&\sup_{\varphi\in H_0^1(\Omega);\, |||\varphi|||=1} \mathcal{B}(u - s_h, \varphi) \\ =& \sup_{\varphi\in H_0^1(\Omega);\, |||\varphi|||=1} \{\mathcal{B}(u - u_h, \varphi) + \mathcal{B}_{\mathrm{A}}(u_h - s_h, \varphi) + \mathcal{B}_{\mathrm{S}}(u_h - s_h, \varphi)\} \\ \le& |||u_h - s_h||| + \sup_{\varphi\in H_0^1(\Omega);\, |||\varphi|||=1} \{\mathcal{B}(u - u_h, \varphi) + \mathcal{B}_{\mathrm{A}}(u_h - s_h, \varphi)\}.\end{aligned}$$

Combining the three above bounds and the definition (9.9) of the augmented norm yields

$$\begin{aligned}|||u - u_h|||_\oplus \le 4|||u_h - s_h||| &+ \sup_{\varphi\in H_0^1(\Omega), |||\varphi|||=1} \mathcal{B}_{\mathrm{A}}(u_h - s_h, \varphi) \\ &+ 3 \sup_{\varphi\in H_0^1(\Omega), |||\varphi|||=1} \{\mathcal{B}(u - u_h, \varphi) + \mathcal{B}_{\mathrm{A}}(u_h - s_h, \varphi)\}.\end{aligned} \tag{9.18}$$

The first term on the right-hand side of (9.18) gives immediately rise to the first estimator in (9.17). As for the second one, we can write

$$\mathcal{B}_{\mathrm{A}}(u_h - s_h, \varphi) = \sum_{K\in\mathcal{T}_h} \{-(\mathbf{w}(u_h - s_h), \nabla\varphi)_K - ((\tfrac{1}{2}\nabla\cdot\mathbf{w})(u_h - s_h), \varphi)_K\}.$$

Let $K \in \mathcal{T}_h$. The Cauchy–Schwarz inequality and the definition (9.8b) of the energy norm on the one hand yield

$$\begin{aligned}
&- (\mathbf{b}_h, \nabla\varphi)_K - ((\tfrac{1}{2}\nabla\cdot\mathbf{w})(u_h - s_h), \varphi)_K \\
&\le \left\{ \|\underline{\mathbf{K}}^{-\frac{1}{2}}\mathbf{b}_h\|_K^2 + \|(\tfrac{1}{2}\nabla\cdot\mathbf{w} + r)^{-\frac{1}{2}}(\tfrac{1}{2}\nabla\cdot\mathbf{w})(u_h - s_h)\|_K^2 \right\}^{\frac{1}{2}} \\
&\quad \times \left\{ \|\underline{\mathbf{K}}^{\frac{1}{2}}\nabla\varphi\|_K^2 + \|(\tfrac{1}{2}\nabla\cdot\mathbf{w} + r)^{\frac{1}{2}}\varphi\|_K^2 \right\}^{\frac{1}{2}} \\
&= \widetilde{\eta}_{\mathrm{NC},1,K} |||\varphi|||_K.
\end{aligned}$$

The Poincaré inequality (4.20), the trace inequality (4.22a), and (9.8b) give, for $K \in \mathcal{T}_h$ and $e \in \mathcal{E}_K$, cf. [49],

$$\begin{aligned}
\|\varphi - \varphi_K\|_K &\le m_K |||\varphi|||_K \\
\|\varphi - \varphi_K\|_e &\le \widetilde{m}_K^{\frac{1}{2}} \widetilde{C}_{\mathrm{t},K,e}^{\frac{1}{2}} |||\varphi|||_K.
\end{aligned}$$

Thus, noting that $(\nabla\varphi)|_K = \nabla(\varphi - \varphi_K)|_K$ and integrating by parts on K leads to

$$\begin{aligned}
&- (\mathbf{b}_h, \nabla\varphi)_K - ((\tfrac{1}{2}\nabla\cdot\mathbf{w})(u_h - s_h), \varphi)_K \\
&= ((I - \Pi_0)\nabla\cdot\mathbf{b}_h, \varphi - \varphi_K)_K - \sum_{e\in\mathcal{E}_K} (\mathbf{b}_h\cdot\mathbf{n}_K, \varphi - \varphi_K)_e - ((\tfrac{1}{2}\nabla\cdot\mathbf{w})(u_h - s_h), \varphi)_K \\
&\le m_K \|(I - \Pi_0)\nabla\cdot\mathbf{b}_h\|_K |||\varphi|||_K + \widetilde{m}_K^{\frac{1}{2}} \sum_{e\in\mathcal{E}_K} \widetilde{C}_{\mathrm{t},K,e}^{\frac{1}{2}} \|\mathbf{b}_h\cdot\mathbf{n}_e\|_e |||\varphi|||_K \\
&\quad + \|(\tfrac{1}{2}\nabla\cdot\mathbf{w} + r)^{-\frac{1}{2}}(\tfrac{1}{2}\nabla\cdot\mathbf{w})(u_h - s_h)\|_K |||\varphi|||_K = \widetilde{\eta}_{\mathrm{NC},2,K} |||\varphi|||_K.
\end{aligned}$$

Above, we have subtracted the projection Π_0 as the term $\|(I - \Pi_0)\nabla\cdot\mathbf{b}_h\|_K$ may be much smaller than $\|\nabla\cdot\mathbf{b}_h\|_K$. Altogether,

$$\mathcal{B}_{\mathrm{A}}(u_h - s_h, \varphi) \le \sum_{K\in\mathcal{T}_h} \widetilde{\eta}_{\mathrm{NC},K} |||\varphi|||_K \le \left\{ \sum_{K\in\mathcal{T}_h} \widetilde{\eta}_{\mathrm{NC},K}^2 \right\}^{\frac{1}{2}} |||\varphi|||.$$

Finally, for the third term on the right-hand side of (9.18), we observe that

$$\begin{aligned}
\mathcal{B}(u - u_h, \varphi) + \mathcal{B}_{\mathrm{A}}(u_h - s_h, \varphi) &= (f - \nabla\cdot\boldsymbol{\sigma}_h - r u_h, \varphi) - (\mathbf{a}_h, \nabla\varphi) - ((\tfrac{1}{2}\nabla\cdot\mathbf{w})(u_h - s_h), \varphi) \\
&\le \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{R},K} + \eta_{\mathrm{F},K}) |||\varphi|||_K,
\end{aligned}$$

using the definition (9.2) of the weak solution, adding and subtracting the term $(\boldsymbol{\sigma}_h, \nabla\varphi)$ and using the Green theorem, and finally employing the assumption (9.6b) for the residual term and proceeding for $\mathbf{a}_h$ and the term with $(\frac{1}{2}\nabla\cdot\mathbf{w})$ as for $\mathbf{b}_h$. □

9.6 Applications and efficiency

One can apply Theorem 9.5.1 to various discretization methods as in Section 7.13. A general efficiency result in the spirit of Theorem 8.3.3 can likewise be obtained; the conceptual difference with Theorem 8.3.3 is that here the efficiency is global and not local (as the augmented norm (9.9) is global). Most importantly, it is the concept of the augmented norm (9.9) which allows to show the robustness with respect to the model parameter $\mathbf{w}$, i.e., to satisfy property **iv)** of Section 1.4. We refer to [96] and [49, 52] for the details.

Chapter 10

The Stokes equation

Recall the Stokes problem (1.3a)–(1.3c) from Chapter 1: for $\mathbf{f} \in [L^2(\Omega)]^d$, find $\mathbf{u}$ and p such that

$$-\Delta\mathbf{u} + \nabla p = \mathbf{f} \qquad \text{in } \Omega, \tag{10.1a}$$
$$\nabla\cdot\mathbf{u} = 0 \qquad \text{in } \Omega, \tag{10.1b}$$
$$\mathbf{u} = \mathbf{0} \qquad \text{on } \partial\Omega. \tag{10.1c}$$

10.1 Variational formulation

Denote by $L^2_0(\Omega)$ the space of $L^2(\Omega)$ functions having zero mean value over Ω. The weak formulation of (10.1a)–(10.1c) reads:

Definition 10.1.1 (Variational formulation of (10.1a)–(10.1c))**.** *Find* $(\mathbf{u}, p) \in [H^1_0(\Omega)]^d \times L^2_0(\Omega)$ *such that*

$$(\nabla\mathbf{u}, \nabla\mathbf{v}) - (\nabla\cdot\mathbf{v}, p) = (\mathbf{f}, \mathbf{v}) \qquad \forall \mathbf{v} \in [H^1_0(\Omega)]^d, \tag{10.2a}$$
$$-(\nabla\cdot\mathbf{u}, q) = 0 \qquad \forall q \in L^2_0(\Omega). \tag{10.2b}$$

Problem (10.2a)–(10.2b) is well-posed (cf. [60]) due to the inf–sup condition (we systematically assume the arguments nonzero)

$$\inf_{q\in L^2_0(\Omega)} \sup_{\mathbf{v}\in[H^1_0(\Omega)]^d} \frac{(q, \nabla\cdot\mathbf{v})}{\|\nabla\mathbf{v}\|\,\|q\|} = \beta, \tag{10.3}$$

where β is a positive constant.

Remark 10.1.2 (Alternative variational formulation)**.** *Alternative equivalent of the variational formulation of Definition 10.1.1 is: find* $\mathbf{u} \in [H^1_0(\Omega)]^d$ *with* $\nabla\cdot\mathbf{u} = 0$ *such that*

$$(\nabla\mathbf{u}, \nabla\mathbf{v}) = (\mathbf{f}, \mathbf{v}) \qquad \forall \mathbf{v} \in [H^1_0(\Omega)]^d \text{ with } \nabla\cdot\mathbf{v} = 0, \tag{10.4}$$

see, e.g., [83, Section 9.1].

As previously, we can here also introduce the concept of flux (called stress in the Stokes setting):

Definition 10.1.3 (Stress). *Let* $(\mathbf{u}, p)$ *be the solution of* (10.2a)–(10.2b). *Set*

$$\underline{\boldsymbol{\sigma}} := \nabla \mathbf{u} - p\underline{\boldsymbol{I}}. \tag{10.5}$$

We will call $\underline{\boldsymbol{\sigma}}$ *the* stress.

The weak solution can be, by exactly the same arguments as in the proof of Theorem 7.1.3, shown to posses the following property, as usual mimicking the physical setting:

Theorem 10.1.4 (Properties of the weak solution of (10.2a)–(10.2b)). *Let* $(\mathbf{u}, p)$ *be the solution of* (10.2a)–(10.2b). *Let* $\underline{\boldsymbol{\sigma}}$ *be given by* (10.5). *Then*

$$\mathbf{u} \in [H_0^1(\Omega)]^d, \quad \underline{\boldsymbol{\sigma}} \in [\mathbf{H}(\mathrm{div}, \Omega)]^d, \quad \nabla\cdot\underline{\boldsymbol{\sigma}} = -\mathbf{f}.$$

10.2 Approximate solution

In order to make the presentation general, not restricted to any particular numerical method, we are led to suppose as before that the approximate solution $(\mathbf{u}_h, p_h)$ that we are given merely satisfies

$$\mathbf{u}_h \in [H^1(\mathcal{T}_h)]^d, \quad p_h \in L_0^2(\Omega). \tag{10.6}$$

Note that p_h is here a conforming approximation, belonging to the same space $L_0^2(\Omega)$ as the weak solution p; it is hard to imagine a numerical approximation which would not be square-integrable, whereas the condition $(p_h, 1) = 0$ is typically satisfied.

In analogy with Definition 10.1.3, we set:

Definition 10.2.1 (Approximate stress). *Let* $(\mathbf{u}_h, p_h)$ *be the approximate solution, cf.* (10.6). *We will call*

$$\nabla \mathbf{u}_h - p_h\underline{\boldsymbol{I}} \tag{10.7}$$

the approximate stress.

The following remark should be compared to Theorem 10.1.4:

Remark 10.2.2 (Properties of the approximate solution $(\mathbf{u}_h, p_h)$ of (10.6)). *Let* $(\mathbf{u}_h, p_h)$ *be the approximate solution, cf.* (10.6). *Then*

$$\mathbf{u}_h \notin [H_0^1(\Omega)]^d, \quad \nabla \mathbf{u}_h - p_h\underline{\boldsymbol{I}} \notin [\mathbf{H}(\mathrm{div}, \Omega)]^d, \quad \nabla\cdot(\nabla \mathbf{u}_h - p_h\underline{\boldsymbol{I}}) \neq -\mathbf{f} \quad \textit{in general.}$$

10.3 Velocity and stress reconstructions

From Theorem 10.1.4 and Remark 10.2.2, we see that the approximate solution (velocity) $\mathbf{u}_h$ and the approximate stress $\nabla \mathbf{u}_h - p_h\underline{\boldsymbol{I}}$ can be nonphysical. Developing the ideas of the previous chapters, we will introduce their "corrections", a velocity reconstruction $\mathbf{s}_h$ and a stress reconstruction $\underline{\boldsymbol{\sigma}}_h$ (here $\mathbf{e}_i$ stands for the i-th Euclidean vector):

Definition 10.3.1 (Velocity reconstruction). *We will call the* velocity reconstruction *any function* $\mathbf{s}_h$ *constructed from* $\mathbf{u}_h$ *which satisfies*

$$\mathbf{s}_h \in [H_0^1(\Omega)]^d.$$

Definition 10.3.2 (Equilibrated stress reconstruction). *We will call the* equilibrated stress reconstruction *any function* $\underline{\boldsymbol{\sigma}}_h$ *constructed from* $\mathbf{u}_h$ *which satisfies*

$$\underline{\boldsymbol{\sigma}}_h \in [\mathbf{H}(\mathrm{div}, \Omega)]^d, \tag{10.8a}$$

$$-(\nabla\cdot\underline{\boldsymbol{\sigma}}_h, \mathbf{e}_i)_K = (\mathbf{f}, \mathbf{e}_i)_K \qquad i = 1, \dots, d, \quad \forall K \in \mathcal{T}_h. \tag{10.8b}$$

10.4 A general a posteriori error estimate

We can now prove our a posteriori error estimate.

Theorem 10.4.1 (A general a posteriori error estimate for (10.1a)–(10.1c))**.** *Let* $(\mathbf{u}, p)$ *be the weak solution given by Definition 10.1.1. Let* $(\mathbf{u}_h, p_h)$ *be arbitrary functions satisfying* (10.6)*. Let* $\mathbf{s}_h$ *be a velocity reconstruction in the sense of Definition 10.3.1 and* $\underline{\boldsymbol{\sigma}}_h$ *an equilibrated stress reconstruction in the sense of Definition 10.3.2. For any* $K \in \mathcal{T}_h$*, define the* residual estimator *by*

$$\eta_{\mathrm{R},K} := C_{\mathrm{P},K} h_K \|\nabla\cdot\underline{\boldsymbol{\sigma}}_h + \mathbf{f}\|_K, \tag{10.9}$$

the flux estimator *by*

$$\eta_{\mathrm{F},K} := \|\nabla\mathbf{u}_h - p_h\underline{\boldsymbol{I}} - \underline{\boldsymbol{\sigma}}_h\|_K, \tag{10.10}$$

the nonconformity estimator *by*

$$\eta_{\mathrm{NC},K} := \|\nabla(\mathbf{u}_h - \mathbf{s}_h)\|_K, \tag{10.11}$$

and the divergence estimator *by*

$$\eta_{\mathrm{D},K} := \frac{\|\nabla\cdot\mathbf{s}_h\|_K}{\beta}. \tag{10.12}$$

Then

$$\|\nabla(\mathbf{u} - \mathbf{u}_h)\|^2 \le \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{R},K} + \eta_{\mathrm{F},K})^2 + \left\{ \left\{ \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{D},K}^2 \right\}^{1/2} + \left\{ \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{NC},K}^2 \right\}^{1/2} \right\}^2, \tag{10.13a}$$

$$\|p - p_h\| \le \frac{1}{\beta} \left(\left\{ \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{R},K} + \eta_{\mathrm{F},K})^2 \right\}^{1/2} + \left\{ \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{D},K}^2 \right\}^{1/2} + \left\{ \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{NC},K}^2 \right\}^{1/2} \right). \tag{10.13b}$$

Proof. The proof follows [45, 5]. We first bound $\|\nabla(\mathbf{u} - \mathbf{u}_h)\|$, by proceeding similarly to the proof of Theorem 7.8.1 in Chapter 7. Let $\mathbf{s} \in [H_0^1(\Omega)]^d$ with $\nabla\cdot\mathbf{s} = 0$ be the solution of

$$(\nabla\mathbf{s}, \nabla\mathbf{v}) = (\nabla\mathbf{u}_h, \nabla\mathbf{v}) \qquad \forall \mathbf{v} \in [H_0^1(\Omega)]^d \text{ with } \nabla\cdot\mathbf{v} = 0. \tag{10.14}$$

This problem has a unique solution, cf. Remark 10.1.2. We again have the Pythagorean equality

$$\|\nabla(\mathbf{u} - \mathbf{u}_h)\|^2 = \|\nabla(\mathbf{u} - \mathbf{s})\|^2 + \|\nabla(\mathbf{s} - \mathbf{u}_h)\|^2, \tag{10.15}$$

which follows from the fact that

$$\|\nabla(\mathbf{u} - \mathbf{u}_h)\|^2 = \|\nabla(\mathbf{u} - \mathbf{s} + \mathbf{s} - \mathbf{u}_h)\|^2 = \|\nabla(\mathbf{u} - \mathbf{s})\|^2 + \|\nabla(\mathbf{s} - \mathbf{u}_h)\|^2 + 2(\nabla(\mathbf{u} - \mathbf{s}), \nabla(\mathbf{s} - \mathbf{u}_h)),$$

with the last term disappearing thanks the fact that $\nabla\cdot(\mathbf{u} - \mathbf{s}) = 0$ and to (10.14). We estimate the two terms in (10.15) separately.

In a complete analog of Theorem 7.3.1, we have here

$$\|\nabla(\mathbf{u} - \mathbf{s})\| = \sup_{\boldsymbol{\varphi}\in[H_0^1(\Omega)]^d;\, \nabla\cdot\boldsymbol{\varphi}=0,\, \|\nabla\boldsymbol{\varphi}\|=1} (\nabla(\mathbf{u} - \mathbf{s}), \nabla\boldsymbol{\varphi}).$$

Let thus $\boldsymbol{\varphi} \in [H_0^1(\Omega)]^d$ with $\nabla\cdot\boldsymbol{\varphi} = 0$ and $\|\nabla\boldsymbol{\varphi}\| = 1$ be fixed. Employing the definitions (10.4) and (10.14), we have

$$(\nabla(\mathbf{u} - \mathbf{s}), \nabla\boldsymbol{\varphi}) = (\mathbf{f}, \boldsymbol{\varphi}) - (\nabla\mathbf{u}_h, \nabla\boldsymbol{\varphi}). \tag{10.16}$$

Next, using that $0 = (p_h, \nabla\cdot\boldsymbol{\varphi}) = (p_h\underline{\boldsymbol{I}}, \nabla\boldsymbol{\varphi})$, adding and subtracting $(\underline{\boldsymbol{\sigma}}_h, \nabla\boldsymbol{\varphi})$, and using the Green theorem (Theorem 4.2.5, component by component), we get

$$(\nabla(\mathbf{u}-\mathbf{s}), \nabla\boldsymbol{\varphi}) = (\nabla\cdot\underline{\boldsymbol{\sigma}}_h + \mathbf{f}, \boldsymbol{\varphi}) - (\nabla\mathbf{u}_h - p_h\underline{\boldsymbol{I}} - \underline{\boldsymbol{\sigma}}_h, \nabla\boldsymbol{\varphi}).$$

We have, for any $K \in \mathcal{T}_h$,

$$(\nabla\cdot\underline{\boldsymbol{\sigma}}_h + \mathbf{f}, \boldsymbol{\varphi})_K = (\nabla\cdot\underline{\boldsymbol{\sigma}}_h + \mathbf{f}, \boldsymbol{\varphi} - \boldsymbol{\varphi}_K)_K \le \eta_{\mathrm{R},K}\|\nabla\boldsymbol{\varphi}\|_K$$

using (10.8b), whereas the estimate

$$(\nabla\mathbf{u}_h - p_h\underline{\boldsymbol{I}} - \underline{\boldsymbol{\sigma}}_h, \nabla\boldsymbol{\varphi})_K \le \eta_{\mathrm{F},K}\|\nabla\boldsymbol{\varphi}\|_K$$

is immediate by the Cauchy–Schwarz inequality. Thus the Cauchy–Schwarz inequality gives

$$(\nabla(\mathbf{u}-\mathbf{s}), \nabla\boldsymbol{\varphi}) \le \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{R},K} + \eta_{\mathrm{F},K})\|\nabla\boldsymbol{\varphi}\|_K \le \left\{\sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{R},K} + \eta_{\mathrm{F},K})^2\right\}^{1/2}. \tag{10.17}$$

We now treat the term $\|\nabla(\mathbf{s}-\mathbf{u}_h)\|$. We have

$$\|\nabla(\mathbf{s}-\mathbf{u}_h)\|^2 = (\nabla(\mathbf{s}-\mathbf{u}_h), \nabla(\mathbf{s}-\mathbf{u}_h)) = (\nabla(\mathbf{s}-\mathbf{s}_h), \nabla(\mathbf{s}-\mathbf{u}_h)) + (\nabla(\mathbf{s}_h-\mathbf{u}_h), \nabla(\mathbf{s}-\mathbf{u}_h)).$$

As in Remark 10.1.2 with respect to Definition 10.1.1, the following equivalent formulation of (10.14) can be given: find $(\mathbf{s}, w) \in [H^1_0(\Omega)]^d \times L^2_0(\Omega)$ such that

$$(\nabla\mathbf{s}, \nabla\mathbf{v}) - (\nabla\cdot\mathbf{v}, w) = (\nabla\mathbf{u}_h, \nabla\mathbf{v}) \qquad \forall\mathbf{v} \in [H^1_0(\Omega)]^d, \tag{10.18a}$$

$$-(\nabla\cdot\mathbf{s}, q) = 0 \qquad \forall q \in L^2_0(\Omega). \tag{10.18b}$$

Thus we have, as $\mathbf{s}-\mathbf{s}_h \in [H^1_0(\Omega)]^d$ can be taken as a test function in (10.18a),

$$(\nabla(\mathbf{s}-\mathbf{u}_h), \nabla(\mathbf{s}-\mathbf{s}_h)) = (\nabla\cdot(\mathbf{s}-\mathbf{s}_h), w) = -(\nabla\cdot\mathbf{s}_h, w) \le \|\nabla\cdot\mathbf{s}_h\|\|w\|.$$

We have also used the fact that $\nabla\cdot\mathbf{s} = 0$ and the Cauchy–Schwarz inequality. To estimate $\|w\|$, we will rely on the inf–sup condition (10.3):

$$\|w\| \le \frac{1}{\beta}\sup_{\mathbf{v}\in[H^1_0(\Omega)]^d} \frac{(w, \nabla\cdot\mathbf{v})}{\|\nabla\mathbf{v}\|} = \frac{1}{\beta}\sup_{\mathbf{v}\in[H^1_0(\Omega)]^d} \frac{(\nabla(\mathbf{s}-\mathbf{u}_h), \nabla\mathbf{v})}{\|\nabla\mathbf{v}\|} \le \frac{1}{\beta}\|\nabla(\mathbf{s}-\mathbf{u}_h)\|,$$

where we have employed (10.18a) and the Cauchy–Schwarz inequality. We thus arrive at

$$(\nabla(\mathbf{s}-\mathbf{s}_h), \nabla(\mathbf{s}-\mathbf{u}_h)) \le \frac{\|\nabla\cdot\mathbf{s}_h\|}{\beta}\|\nabla(\mathbf{s}-\mathbf{u}_h)\|,$$

whence

$$\|\nabla(\mathbf{s}-\mathbf{u}_h)\| \le \frac{\|\nabla\cdot\mathbf{s}_h\|}{\beta} + \|\nabla(\mathbf{s}_h-\mathbf{u}_h)\| = \left\{\sum_{K\in\mathcal{T}_h} \eta^2_{\mathrm{D},K}\right\}^{1/2} + \left\{\sum_{K\in\mathcal{T}_h} \eta^2_{\mathrm{NC},K}\right\}^{1/2}. \tag{10.19}$$

Finally, the term $\|p-p_h\|$ is treated through the inf–sup condition (10.3), which in particular gives

$$\|p-p_h\| \le \frac{1}{\beta}\sup_{\boldsymbol{\varphi}\in[H^1_0(\Omega)]^d;\,\|\nabla\boldsymbol{\varphi}\|=1} (p-p_h, \nabla\cdot\boldsymbol{\varphi}).$$

Fix $\boldsymbol{\varphi} \in [H^1_0(\Omega)]^d$ with $\|\nabla\boldsymbol{\varphi}\| = 1$. The weak solution characterization (10.2a) gives

$$(p, \nabla\cdot\boldsymbol{\varphi}) = (\nabla\mathbf{u}, \nabla\boldsymbol{\varphi}) - (\mathbf{f}, \boldsymbol{\varphi}).$$

Thus using also $(p_h, \nabla\cdot\boldsymbol{\varphi}) = (p_h\underline{\boldsymbol{I}}, \nabla\boldsymbol{\varphi})$, adding and subtracting $(\underline{\boldsymbol{\sigma}}_h, \nabla\boldsymbol{\varphi})$ as well as $(\nabla\mathbf{u}_h, \nabla\boldsymbol{\varphi})$, and using the Green theorem, we arrive at

$$(p - p_h, \nabla\cdot\boldsymbol{\varphi}) = (\nabla(\mathbf{u} - \mathbf{u}_h), \nabla\boldsymbol{\varphi}) - (\nabla\cdot\underline{\boldsymbol{\sigma}}_h + \mathbf{f}, \boldsymbol{\varphi}) + (\nabla\mathbf{u}_h - p_h\underline{\boldsymbol{I}} - \underline{\boldsymbol{\sigma}}_h, \nabla\boldsymbol{\varphi}).$$

The two last terms on the above right-hand side could be estimated as in (10.17) and the first one could be bounded by $\|\nabla(\mathbf{u} - \mathbf{u}_h)\|$ and consequently by (10.13a). Such a straightforward bound can, however, be substantially improved while proceeding as in [5]. Let $\boldsymbol{\varphi}_{\mathrm{C}} \in [H^1_0(\Omega)]^d$ with $\nabla\cdot\boldsymbol{\varphi}_{\mathrm{C}} = 0$ be the solution of

$$(\nabla\boldsymbol{\varphi}_{\mathrm{C}}, \nabla\mathbf{v}) = (\nabla\boldsymbol{\varphi}, \nabla\mathbf{v}) \qquad \forall\mathbf{v} \in [H^1_0(\Omega)]^d \text{ with } \nabla\cdot\mathbf{v} = 0.$$

Let $\boldsymbol{\varphi}_{\mathrm{NC}} := \boldsymbol{\varphi} - \boldsymbol{\varphi}_{\mathrm{C}}$; note that

$$(\nabla\boldsymbol{\varphi}_{\mathrm{NC}}, \nabla\mathbf{v}) = 0 \qquad \forall\mathbf{v} \in [H^1_0(\Omega)]^d \text{ with } \nabla\cdot\mathbf{v} = 0. \tag{10.20}$$

Then, as in (10.15), we immediately have

$$\|\nabla\boldsymbol{\varphi}\|^2 = \|\nabla\boldsymbol{\varphi}_{\mathrm{C}}\|^2 + \|\nabla\boldsymbol{\varphi}_{\mathrm{NC}}\|^2, \tag{10.21}$$

as $(\nabla\boldsymbol{\varphi}_{\mathrm{NC}}, \nabla\boldsymbol{\varphi}_{\mathrm{C}}) = 0$. Now

$$(p - p_h, \nabla\cdot\boldsymbol{\varphi}) = (p - p_h, \nabla\cdot\boldsymbol{\varphi}_{\mathrm{NC}})$$

and

$$(p - p_h, \nabla\cdot\boldsymbol{\varphi}_{\mathrm{NC}}) = (\nabla(\mathbf{u} - \mathbf{u}_h), \nabla\boldsymbol{\varphi}_{\mathrm{NC}}) - (\nabla\cdot\underline{\boldsymbol{\sigma}}_h + \mathbf{f}, \boldsymbol{\varphi}_{\mathrm{NC}}) + (\nabla\mathbf{u}_h - p_h\underline{\boldsymbol{I}} - \underline{\boldsymbol{\sigma}}_h, \nabla\boldsymbol{\varphi}_{\mathrm{NC}}).$$

We can estimate

$$-(\nabla\cdot\underline{\boldsymbol{\sigma}}_h + \mathbf{f}, \boldsymbol{\varphi}_{\mathrm{NC}}) + (\nabla\mathbf{u}_h - p_h\underline{\boldsymbol{I}} - \underline{\boldsymbol{\sigma}}_h, \nabla\boldsymbol{\varphi}_{\mathrm{NC}}) \le \left\{\sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{R},K} + \eta_{\mathrm{F},K})^2\right\}^{1/2}$$

as in (10.17), since $\|\nabla\boldsymbol{\varphi}_{\mathrm{NC}}\| \le \|\nabla\boldsymbol{\varphi}\| = 1$ by (10.21). The gain is that

$$(\nabla(\mathbf{u} - \mathbf{u}_h), \nabla\boldsymbol{\varphi}_{\mathrm{NC}}) = (\nabla(\mathbf{s} - \mathbf{u}_h), \nabla\boldsymbol{\varphi}_{\mathrm{NC}}),$$

as $(\nabla(\mathbf{u} - \mathbf{s}), \nabla\boldsymbol{\varphi}_{\mathrm{NC}}) = 0$ by (10.20) since $\mathbf{u} - \mathbf{s} \in [H^1_0(\Omega)]^d$ and $\nabla\cdot(\mathbf{u} - \mathbf{s}) = 0$. Now the Cauchy–Schwarz inequality and (10.19) together with $\|\nabla\boldsymbol{\varphi}_{\mathrm{NC}}\| \le \|\nabla\boldsymbol{\varphi}\| = 1$ give

$$(\nabla(\mathbf{u} - \mathbf{u}_h), \nabla\boldsymbol{\varphi}_{\mathrm{NC}}) \le \left\{\sum_{K\in\mathcal{T}_h} \eta^2_{\mathrm{D},K}\right\}^{1/2} + \left\{\sum_{K\in\mathcal{T}_h} \eta^2_{\mathrm{NC},K}\right\}^{1/2}.$$

Combining the above results gives (10.13b). □

10.5 Application to classical discretization methods and local efficiency

As in Chapter 7, the estimate of Theorem 10.4.1 can be used for many numerical methods upon specifying the velocity and stress reconstructions $\mathbf{s}_h$ and $\underline{\boldsymbol{\sigma}}_h$. Similarly to Section 7.11 in Chapter 7, local efficiency can then be shown.

Chapter 11

The heat equation

We give in this chapter a few results on a model unsteady problem, the heat equation (1.4a)–(1.4c). It reads: for $f \in L^2(\Omega \times (0,T))$, $u_0 \in L^2(\Omega)$, and $T > 0$, find u such that

$$\partial_t u - \Delta u = f \quad \text{in } \Omega \times (0,T), \tag{11.1a}$$
$$u = 0 \quad \text{on } \partial\Omega \times (0,T), \tag{11.1b}$$
$$u(\cdot,0) = u_0 \quad \text{in } \Omega. \tag{11.1c}$$

11.1 Variational formulation

Set

$$X := L^2(0,T;H_0^1(\Omega)), \tag{11.2a}$$
$$Y := \{v \in X; \partial_t v \in X'\}, \tag{11.2b}$$

where $X' = L^2(0,T;H^{-1}(\Omega))$. In order to properly define u, we use:

Definition 11.1.1 (Variational formulation of (11.1a)–(11.1c))**.** *Find $u \in Y$ such that $u(\cdot,0) = u_0$ and such that, for a.e. $t \in (0,T)$,*

$$\langle \partial_t u, \varphi \rangle_{H^{-1}(\Omega),H_0^1(\Omega)}(t) + (\nabla u, \nabla \varphi)(t) = (f,\varphi)(t) \qquad \forall \varphi \in H_0^1(\Omega). \tag{11.3}$$

As usual, we will now introduce the flux:

Definition 11.1.2 (Flux)**.** *Let u be the solution of (11.3). Set*

$$\boldsymbol{\sigma} := -\nabla u. \tag{11.4}$$

We will call $\boldsymbol{\sigma}$ the flux*.*

Also in the unsteady context, we have the following physical-relevance result:

Theorem 11.1.3 (Properties of the weak solution of (11.1a)–(11.1c))**.** *Let $u_0 \in H_0^1(\Omega)$ and let u be the solution of (11.3). Let $\boldsymbol{\sigma}$ be given by (11.4). Then*

$$u \in X \cap C(0,T;L^2(\Omega)), \quad \partial_t u \in L^2(0,T;L^2(\Omega)), \tag{11.5}$$
$$\boldsymbol{\sigma} \in L^2(0,T;\mathbf{H}(\mathrm{div},\Omega)), \quad \nabla\cdot\boldsymbol{\sigma} = f - \partial_t u. \tag{11.6}$$

Proof. The weak solution u belongs to X by definition. Moreover, it follows from $u \in Y$ that $u \in C(0,T;L^2(\Omega))$. The additional regularity on the initial condition, $u_0 \in H^1_0(\Omega)$, then implies $\partial_t u \in L^2(0,T;L^2(\Omega))$, cf., e.g., Evans [56]. As for $\boldsymbol{\sigma}$, we now check the three conditions of Definition 4.2.1 for a.e. $t \in (0,T)$. Condition 1 is obvious, as $u(\cdot,t) \in H^1_0(\Omega)$, so that $\boldsymbol{\sigma}(\cdot,t) \in [L^2(\Omega)]^d$. For the function w of condition 2a, choose $w := f(\cdot,t) - \partial_t u(\cdot,t)$ and note that $w \in L^2(\Omega)$. Finally, condition 2b follows immediately from (11.3) and the fact that $\mathcal{D}(\Omega) \subset H^1_0(\Omega)$. The conclusion follows from the fact that $\boldsymbol{\sigma} \in L^2(0,T;[L^2(\Omega)]^d)$ as $u \in X$ and $\nabla\cdot\boldsymbol{\sigma} \in L^2(0,T;L^2(\Omega))$ as both f and $\partial_t u$ belong to $L^2(0,T;L^2(\Omega))$. □

11.2 Space-time meshes and spaces

In comparison with the previous chapters, we now also have the time variable t, so that we need to introduce some more notation. We consider an increasing sequence of discrete times $\{t^n\}_{0\le n\le N}$ such that $t^0 = 0$ and $t^N = T$ and introduce the time intervals $I_n := (t^{n-1}, t^n]$ and the time steps $\tau^n := t^n - t^{n-1}$ for all $1 \le n \le N$. The spatial meshes are allowed to vary in time; we denote by $\mathcal{T}_h^n$ the mesh used to march in time from t^{n-1} to t^n, for all $1 \le n \le N$, and by $\mathcal{T}_h^0$ the initial mesh. We also denote by $P^1_\tau(H^1_0(\Omega))$ the space of functions that are piecewise affine and continuous in time and $H^1_0(\Omega)$ in space. Note that $P^1_\tau(H^1_0(\Omega)) \subset X \cap C(0,T;L^2(\Omega))$ and any $v_{h\tau} \in P^1_\tau(H^1_0(\Omega))$ satisfies $\partial_t v_{h\tau} \in L^2(0,T;L^2(\Omega))$. Similarly, we denote by $P^0_\tau(\mathbf{H}(\mathrm{div},\Omega))$ the space of functions that are piecewise constant in time and $\mathbf{H}(\mathrm{div},\Omega)$ in space and note that $P^0_\tau(\mathbf{H}(\mathrm{div},\Omega)) \subset L^2(0,T;\mathbf{H}(\mathrm{div},\Omega))$. Finally, let $P^1_\tau(H^1(\mathcal{T}_h))$ be the space of functions $v_{h\tau}$ that are piecewise affine and continuous in time, given by functions $v_h^n := v_{h\tau}(\cdot,t^n)$ from $H^1(\mathcal{T}_h^n)$ for all $0 \le n \le N$. We also set

$$\partial_t^n v_{h\tau} := \frac{(v_h^n - v_h^{n-1})}{\tau^n},$$

$1 \le n \le N$. For the sake of simplicity, we suppose that f is piecewise constant in time and denote $f^n := f|_{I_n}$. For $\mathbf{v}_{h\tau} \in P^0_\tau(\mathbf{H}(\mathrm{div},\Omega))$, we define $\mathbf{v}_h^n := \mathbf{v}_{h\tau}|_{I_n}$.

11.3 Approximate solution

In order to make, as usual, the presentation general, not restricted to any particular numerical method, we are led to suppose in this chapter that the approximate solution $u_{h\tau}$ that we are given satisfies

$$u_{h\tau} \in P^1_\tau(H^1(\mathcal{T}_h)). \tag{11.7}$$

We now define the approximate flux:

Definition 11.3.1 (Approximate flux)**.** *Let $u_{h\tau}$ be the approximate solution, cf. (11.7). We will call*

$$-\nabla u_{h\tau} \tag{11.8}$$

the approximate flux.

As usual, the following remark should be compared to Theorem 11.1.3:

Remark 11.3.2 (Properties of the approximate solution $u_{h\tau}$ of (11.7))**.** *Let $u_{h\tau}$ be the approximate solution, cf. (11.7). Then*

$$u_{h\tau} \notin X, \quad -\nabla u_{h\tau} \notin L^2(0,T;\mathbf{H}(\mathrm{div},\Omega)), \quad \nabla\cdot(-\nabla u_{h\tau}) \ne f - \partial_t u_{h\tau} \quad \text{in general.}$$

Note however that we do have $u_{h\tau} \in C(0,T;L^2(\Omega))$ and $\partial_t u_{h\tau} \in L^2(0,T;L^2(\Omega))$, so that the nonconformity in $u_{h\tau}$ is only with respect to space and not with respect to time.

11.4 Potential and flux reconstructions

Based on the preceding considerations, we are one again lead to introduce a potential reconstruction and an equilibrated flux reconstruction. Herein, these are space–time functions.

Definition 11.4.1 (Potential reconstruction)**.** *Let $u_{h\tau}$ be the approximate solution, cf.* (11.7)*. We will call the* potential reconstruction *any function $s_{h\tau}$ constructed from $u_{h\tau}$ which satisfies*

$$s_{h\tau} \in P^1_\tau(H^1_0(\Omega)), \tag{11.9a}$$
$$(\partial_t^n s_{h\tau}, 1)_K = (\partial_t^n u_{h\tau}, 1)_K \qquad \forall 1 \le n \le N, \quad \forall K \in \mathcal{T}_h^n. \tag{11.9b}$$

Note that in contrast to the previous chapters, we require in (11.9b) that the mean values of the time derivative of $u_{h\tau}$ are preserved by the potential reconstruction $s_{h\tau}$.

Definition 11.4.2 (Equilibrated flux reconstruction)**.** *We will call the* equilibrated flux reconstruction *any function $\boldsymbol{\sigma}_{h\tau}$ constructed from $u_{h\tau}$ which satisfies*

$$\boldsymbol{\sigma}_{h\tau} \in P^0_\tau(\mathbf{H}(\mathrm{div}, \Omega)), \tag{11.10a}$$
$$(\partial_t^n u_{h\tau} + \nabla\cdot\boldsymbol{\sigma}_h^n, 1)_K = (f^n, 1)_K \qquad \forall 1 \le n \le N, \quad \forall K \in \mathcal{T}_h^n. \tag{11.10b}$$

11.5 Energy (semi-)norm augmented by a dual norm and its equivalence with the dual norm of the residual

Let $v \in X$. The space–time energy norm for (11.1a)–(11.1c) is given by

$$\|v\|_X := \left\{ \int_0^T \|\nabla v\|^2(t)\,\mathrm{d}t \right\}^{\frac{1}{2}}. \tag{11.11}$$

Following Verfürth [95], we augment the energy norm by the dual norm of the time derivative, forming a norm on the space Y. In particular, for $v \in Y$, we set

$$\|v\|_Y := \|v\|_X + \|\partial_t v\|_{X'}, \tag{11.12}$$

where

$$\|\partial_t v\|_{X'} := \left\{ \int_0^T \|\partial_t v\|^2_{H^{-1}(\Omega)}(t)\,\mathrm{d}t \right\}^{\frac{1}{2}}.$$

Let $t \in (0,T)$. One can compute $\|\partial_t v\|_{H^{-1}(\Omega)}$ at the time t by introducing the following elliptic problem: find $w(\cdot,t) \in H^1_0(\Omega)$ such that

$$(\nabla w(\cdot,t), \nabla\varphi) = \langle \partial_t v, \varphi\rangle_{H^{-1}(\Omega),H^1_0(\Omega)}(t) \qquad \forall \varphi \in H^1_0(\Omega).$$

Then $\|\partial_t v\|_{H^{-1}(\Omega)}(t) = \|\nabla w(\cdot,t)\|$ by Theorem 7.3.1, and, consequently, $\|w\|_X = \|\partial_t v\|_{X'}$. This also leads to the following useful characterization of the norm $\|\partial_t v\|_{X'}$:

$$\|\partial_t v\|_{X'} = \sup_{\varphi\in X,\, \|\varphi\|_X = 1} \int_0^T \langle \partial_t v, \varphi\rangle_{H^{-1}(\Omega),H^1_0(\Omega)}(t)\,\mathrm{d}t. \tag{11.13}$$

Similarly as, for example, in the advection–reaction–diffusion equation, Theorem 9.4.1, we have the following crucial equivalence result:

Theorem 11.5.1 (Equivalence of the $\|\cdot\|_Y$ norm and of a dual norm)**.** *Let $v \in Y$. Then*

$$\|v\|_Y \le 3 \sup_{\varphi\in X,\, \|\varphi\|_X=1} \int_0^T \{\langle \partial_t v, \varphi\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} + (\nabla v, \nabla\varphi)\}(t)\,\mathrm{d}t + 2^{1/2}\|v(\cdot,0)\|, \tag{11.14a}$$

$$\sup_{\varphi\in X,\, \|\varphi\|_X=1} \int_0^T \{\langle \partial_t v, \varphi\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} + (\nabla v, \nabla\varphi)\}(t)\,\mathrm{d}t \le \|v\|_Y. \tag{11.14b}$$

Proof. We start by (11.14a). Since v is in Y, there holds (see, e.g., [56, Theorem 5.9.3])

$$\tfrac{1}{2}\|v(\cdot,T)\|^2 = \tfrac{1}{2}\|v(\cdot,0)\|^2 + \int_0^T \langle \partial_t v, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}(t)\,\mathrm{d}t.$$

As a result,

$$\begin{aligned}\|v\|_X^2 &\le \tfrac{1}{2}\|v(\cdot,T)\|^2 + \|v\|_X^2\\ &= \tfrac{1}{2}\|v(\cdot,0)\|^2 + \int_0^T \{\langle \partial_t v, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} + (\nabla v, \nabla v)\}(t)\,\mathrm{d}t.\end{aligned}$$

Passing to the supremum gives

$$\|v\|_X^2 \le \sup_{\varphi\in X,\, \|\varphi\|_X=1} \int_0^T \{\langle \partial_t v, \varphi\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} + (\nabla v, \nabla\varphi)\}(t)\,\mathrm{d}t\|v\|_X + \tfrac{1}{2}\|v(\cdot,0)\|^2.$$

Since $x^2 \le ax + b^2$ implies $x \le a + b$ for non-negative a and b, it is inferred that

$$\|v\|_X \le \sup_{\varphi\in X,\, \|\varphi\|_X=1} \int_0^T \{\langle \partial_t v, \varphi\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} + (\nabla v, \nabla\varphi)\}(t)\,\mathrm{d}t + 2^{-1/2}\|v(\cdot,0)\|.$$

Let now $\varphi \in X$ with $\|\varphi\|_X = 1$ and observe that

$$\int_0^T \langle \partial_t v, \varphi\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}(t)\,\mathrm{d}t = \int_0^T \{\langle \partial_t v, \varphi\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} + (\nabla v, \nabla\varphi) - (\nabla v, \nabla\varphi)\}(t)\,\mathrm{d}t,$$

whence, from (11.13),

$$\|\partial_t v\|_{X'} \le \sup_{\varphi\in X,\, \|\varphi\|_X=1} \int_0^T \{\langle \partial_t v, \varphi\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} + (\nabla v, \nabla\varphi)\}(t)\,\mathrm{d}t + \|v\|_X,$$

so that (11.14a) follows.

As for (11.14b), it is an immediate consequence of the definitions (11.12) and (11.13). □

Let us now, as in Chapters 7 and 9, define the residual of a function $v_{h\tau} \in Y$:

Definition 11.5.2 (Residual)**.** *Let $v_{h\tau} \in Y$. Then $\mathcal{R}(v_{h\tau}) \in X'$ is defined by*

$$\langle \mathcal{R}(v_{h\tau}), \varphi\rangle_{X',X} := \int_0^T \{(f,\varphi) - \langle \partial_t v_{h\tau}, \varphi\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} - (\nabla v_{h\tau}, \nabla\varphi)\}(t)\,\mathrm{d}t \qquad \varphi \in X. \tag{11.15}$$

Again the weak formulation (11.3) implies that the residual is zero if and only if the function $v_{h\tau}$ equals to the weak solution u, provided $v_{h\tau}$ satisfies the initial condition. More precisely, Theorem 11.5.1 implies the following important corollary, compare again with Theorem 7.7.2:

Corollary 11.5.3 (Equivalence between the Y and dual residual norms)**.** *Let u be the weak solution given by Definition 11.1.1. Let $v_{h\tau} \in Y$ be arbitrary. Then*

$$\|u - v_{h\tau}\|_Y \leq 3\|\mathcal{R}(v_{h\tau})\|_{X'} + 2^{\frac{1}{2}}\|(u - v_{h\tau})(\cdot, 0)\|, \tag{11.16a}$$
$$\|\mathcal{R}(v_{h\tau})\|_{X'} \leq \|u - v_{h\tau}\|_Y, \tag{11.16b}$$

Proof. The dual norm of the residual is given by

$$\|\mathcal{R}(v_{h\tau})\|_{X'} := \sup_{\varphi \in X,\, \|\varphi\|_X = 1} \langle \mathcal{R}(v_{h\tau}), \varphi \rangle_{X',X}. \tag{11.17}$$

By (11.3), we have

$$\langle \mathcal{R}(v_{h\tau}), \varphi \rangle_{X',X} = \int_0^T \{\langle \partial_t(u - v_{h\tau}), \varphi \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} + (\nabla(u - v_{h\tau}), \nabla\varphi)\}(t)\, \mathrm{d}t.$$

Thus, it is enough to take $v = u - v_{h\tau}$ in Theorem 11.5.1. □

Thus, as in Theorem 7.7.2 and Corollary 9.4.3, we still have the equivalence of the norm $\|u - v_{h\tau}\|_Y$ with the dual norm of the residual $\mathcal{R}(v_{h\tau})$, up to the error in the initial data.

11.6 A general a posteriori error estimate

We state here our main upper bound result for the heat equation. In order to, as usual, proceed generally, we assume:

Assumption 11.6.1 (Potential and flux reconstructions for (11.1a)–(11.1c))**.** *We suppose that $s_{h\tau}$ is a potential reconstruction in the sense of Definition 11.4.1 and $\boldsymbol{\sigma}_{h\tau}$ an equilibrated flux reconstruction in the sense of Definition 11.4.2.*

For $u - u_{h\tau}$, which is not in Y in general (cf. (11.7) and Remark 11.3.2), extend the definition (11.12), where the gradient is understood in the broken sense (cf. Definition 4.3.1). Then $\|u - u_{h\tau}\|_Y$ is a seminorm only in general. We have:

Theorem 11.6.2 (A general a posteriori error estimate for (11.1a)–(11.1c))**.** *Let u be the weak solution given by Definition 11.1.1. Let $u_{h\tau}$ be an arbitrary function satisfying (11.7). Let Assumption 11.6.1 be satisfied. Let finally $1 \leq n \leq N$ and $K \in \mathcal{T}_h^n$ and define the* residual estimator *by*

$$\eta_{\mathrm{R},K}^n := C_{\mathrm{P},K} h_K \|f^n - \partial_t^n s_{h\tau} - \nabla\cdot\boldsymbol{\sigma}_h^n\|_K, \tag{11.18}$$

the flux estimator *by*

$$\eta_{\mathrm{F},K}^n(t) := \|\nabla s_{h\tau}(t) + \boldsymbol{\sigma}_h^n\|_K, \tag{11.19}$$

the nonconformity estimators *by*

$$\eta_{\mathrm{NC},1,K}^n(t) := \|\nabla(s_{h\tau} - u_{h\tau})(t)\|_K, \tag{11.20a}$$
$$\eta_{\mathrm{NC},2,K}^n := C_{\mathrm{P},K} h_K \|\partial_t^n(s_{h\tau} - u_{h\tau})\|_K, \tag{11.20b}$$

and the initial condition estimator *by*

$$\eta_{\mathrm{IC}} := 2^{\frac{1}{2}}\|s_{h\tau}(\cdot, 0) - u_0\|. \tag{11.21}$$

Then

$$\|u - u_{h\tau}\|_Y \le 3\left\{\sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h^n} (\eta_{\mathrm{R},K}^n + \eta_{\mathrm{F},K}^n(t))^2 \,\mathrm{d}t\right\}^{\frac12} + \eta_{\mathrm{IC}}$$
$$+ \left\{\sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h^n} (\eta_{\mathrm{NC},1,K}^n)^2(t) \,\mathrm{d}t\right\}^{\frac12} + \left\{\sum_{n=1}^N \tau^n \sum_{K\in\mathcal{T}_h^n} (\eta_{\mathrm{NC},2,K}^n)^2\right\}^{\frac12}.$$

In order to prove Theorem 11.6.2, we proceed in three steps.

Lemma 11.6.3 (Abstract $\|\cdot\|_Y$-norm error estimate)**.** *Let $s_{h\tau}$ be as in Assumption 11.6.1. Then*

$$\|u - u_{h\tau}\|_Y \le \|s_{h\tau} - u_{h\tau}\|_Y + 3\|\mathcal{R}(s_{h\tau})\|_{X'} + 2^{\frac12}\|s_{h\tau}(\cdot,0) - u_0\|. \tag{11.22}$$

Proof. By the triangle inequality and (11.16a). □

The dual norm $\|\mathcal{R}(s_{h\tau})\|_{X'}$ in the abstract error estimate (11.22) is not easily computable. We are now going to infer a computable upper bound for this quantity, introducing the flux reconstruction $\boldsymbol{\sigma}_{h\tau}$ and making use of Assumption 11.6.1.

Lemma 11.6.4 (Computable upper bound on $\|\mathcal{R}(s_{h\tau})\|_{X'}$)**.** *Let Assumption 11.6.1 hold. Then*

$$\|\mathcal{R}(s_{h\tau})\|_{X'} \le \left\{\sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h^n} (\eta_{\mathrm{R},K}^n + \eta_{\mathrm{F},K}^n(t))^2 \,\mathrm{d}t\right\}^{\frac12}.$$

Proof. Let $\varphi \in X$ with $\|\varphi\|_X = 1$. Then, using the definition (11.15), noting that $\partial_t s_{h\tau} \in L^2(0,T;L^2(\Omega))$, adding and subtracting $(\boldsymbol{\sigma}_{h\tau}, \nabla\varphi)$ in the integrand for a.e. $t \in (0,T)$, and using the Green theorem (Theorem 4.2.5) yields

$$\langle \mathcal{R}(s_{h\tau}), \varphi\rangle_{X',X} = \int_0^T \{(f - \partial_t s_{h\tau} - \nabla\cdot\boldsymbol{\sigma}_{h\tau}, \varphi) - (\nabla s_{h\tau} + \boldsymbol{\sigma}_{h\tau}, \nabla\varphi)\}(t)\,\mathrm{d}t$$
$$=: T_1 + T_2.$$

Owing to Assumption 11.6.1, there holds $s_{h\tau} \in P_\tau^1(H_0^1(\Omega))$ and $\boldsymbol{\sigma}_{h\tau} \in P_\tau^0(\mathbf{H}(\mathrm{div},\Omega))$, so that

$$T_1 = \sum_{n=1}^N \int_{I_n} (f^n - \partial_t^n s_{h\tau} - \nabla\cdot\boldsymbol{\sigma}_h^n, \varphi(t))\,\mathrm{d}t.$$

For all $1 \le n \le N$, owing to (11.9b) and (11.10b),

$$(f^n - \partial_t^n s_{h\tau} - \nabla\cdot\boldsymbol{\sigma}_h^n, 1)_K = 0 \qquad \forall K \in \mathcal{T}_h^n.$$

Hence, for a.e. $t \in I_n$,

$$(f^n - \partial_t^n s_{h\tau} - \nabla\cdot\boldsymbol{\sigma}_h^n, \varphi(t)) = (f^n - \partial_t^n s_{h\tau} - \nabla\cdot\boldsymbol{\sigma}_h^n, \varphi(t) - \varphi_K(t))$$
$$\le \sum_{K\in\mathcal{T}_h^n} C_{\mathrm{P},K} h_K \|f^n - \partial_t^n s_{h\tau} - \nabla\cdot\boldsymbol{\sigma}_h^n\|_K \|\nabla\varphi\|_K(t),$$

where we have used the Poincaré inequality (4.20) on each $K \in \mathcal{T}_h^n$. Moreover,

$$T_2 \le \sum_{n=1}^{N} \int_{I_n} \sum_{K \in \mathcal{T}_h^n} \|\nabla s_{h\tau}(t) + \boldsymbol{\sigma}_h^n\|_K \|\nabla \varphi\|_K(t) \,\mathrm{d}t.$$

Collecting the above estimates yields using the Cauchy–Schwarz inequality

$$|T_1 + T_2| \le \left\{ \sum_{n=1}^{N} \int_{I_n} \sum_{K \in \mathcal{T}_h^n} (\eta_{\mathrm{R},K}^n + \eta_{\mathrm{F},K}^n(t))^2 \,\mathrm{d}t \right\}^{\frac{1}{2}},$$

which concludes the proof. □

Owing to the definition (11.21) of η_{IC}, the last step is to derive a computable upper bound on $\|s_{h\tau} - u_{h\tau}\|_Y$.

Lemma 11.6.5 (Computable upper bound on $\|s_{h\tau} - u_{h\tau}\|_Y$)**.** *Let Assumption 11.6.1 hold. Then,*

$$\|s_{h\tau} - u_{h\tau}\|_Y \le \left\{ \sum_{n=1}^{N} \int_{I_n} \sum_{K \in \mathcal{T}_h^n} (\eta_{\mathrm{NC},1,K}^n)^2(t) \,\mathrm{d}t \right\}^{\frac{1}{2}} + \left\{ \sum_{n=1}^{N} \tau^n \sum_{K \in \mathcal{T}_h^n} (\eta_{\mathrm{NC},2,K}^n)^2 \right\}^{\frac{1}{2}}.$$

Proof. It follows from the definition of the energy norm (11.11) that

$$\|s_{h\tau} - u_{h\tau}\|_X = \left\{ \sum_{n=1}^{N} \int_{I_n} \sum_{K \in \mathcal{T}_h^n} (\eta_{\mathrm{NC},1,K}^n)^2(t) \,\mathrm{d}t \right\}^{\frac{1}{2}}.$$

Let now $\varphi \in X$ with $\|\varphi\|_X = 1$. Observe that since both $s_{h\tau}$ and $u_{h\tau}$ are piecewise affine and continuous in time,

$$\langle \partial_t(s_{h\tau} - u_{h\tau}), \varphi \rangle_{X',X} = \sum_{n=1}^{N} \int_{I_n} (\partial_t^n(s_{h\tau} - u_{h\tau}), \varphi(t)) \,\mathrm{d}t.$$

For all $1 \le n \le N$, it is inferred from (11.9b) that the quantity $\partial_t^n(s_{h\tau} - u_{h\tau})$ has zero mean on each element $K \in \mathcal{T}_h^n$. Hence, using the Poincaré inequality (4.20) yields

$$\begin{aligned} \langle \partial_t(s_{h\tau} - u_{h\tau}), \varphi \rangle_{X',X} &= \sum_{n=1}^{N} \int_{I_n} (\partial_t^n(s_{h\tau} - u_{h\tau}), \varphi(t) - \varphi_K(t)) \,\mathrm{d}t \\ &\le \sum_{n=1}^{N} \int_{I_n} \sum_{K \in \mathcal{T}_h^n} \eta_{\mathrm{NC},2,K}^n \|\nabla \varphi\|_K(t) \,\mathrm{d}t, \end{aligned}$$

whence the desired estimate is inferred using the Cauchy–Schwarz inequality. □

11.7 Application to classical discretization methods and efficiency

As in Chapter 7, the estimate of Theorem 11.6.2 can be used for many numerical methods upon specifying the potential and flux reconstructions $s_{h\tau}$ and $\boldsymbol{\sigma}_{h\tau}$. Similarly to Section 7.11 in Chapter 7, efficiency can then be shown. The efficiency is local in time but only global in space, owing to the global-in-space nature of the norm $\|\cdot\|_Y$, see (11.12). Most importantly, the efficiency is robust with respect to the final time; in other words, the overestimation of the error by our estimates does not depend on the final simulation time T. We refer for details to [51].

11.8 Numerical examples

We give here some illustrative examples of the performance of the above estimates. The vertex-centered finite volume method (cf. Section 8.3.7) has been employed.

In Figures 11.1 and 11.2, we compare the estimated and actual $\|u - u_{h\tau}\|_Y$ errors for two different final simulation times: $T = 1.5$ and $T = 3$. One should in particular notice that the overall error and estimates increase significantly when doubling the simulation time. Contrarily, the effectivity indices stay almost unchanged, which is the numerical evidence of the robustness (property **iv)** of Section 1.4) with respect to the final simulation time. The asymptotic exactness is neither achieved nor approached here, as the effectivity indices take the value of around 5.5.

For illustration, we also show the performance of the a posteriori error estimates for an unsteady advection–diffusion–reaction extension of (11.1a)–(11.1c), representing the propagation of a concentration plume in the underground (cf. Example 1.1.3). In Figure 11.3, we see that we can predict the spatial error distribution on a given time level, as for steady problems in Sections 7.14 and 8.4. In Figure 11.4, we then illustrate the effect of adaptive mesh refinement: it is enough to refine the mesh locally and move it adaptively following the plume in order to attain a considerably better resolution (note the difference in the scales of the two pictures). We refer to [64] for the details.

Altogether, efficiency with precision attainment in the sense of the properties **1.**–**2.** of the Introduction can be achieved also for model unsteady problems.

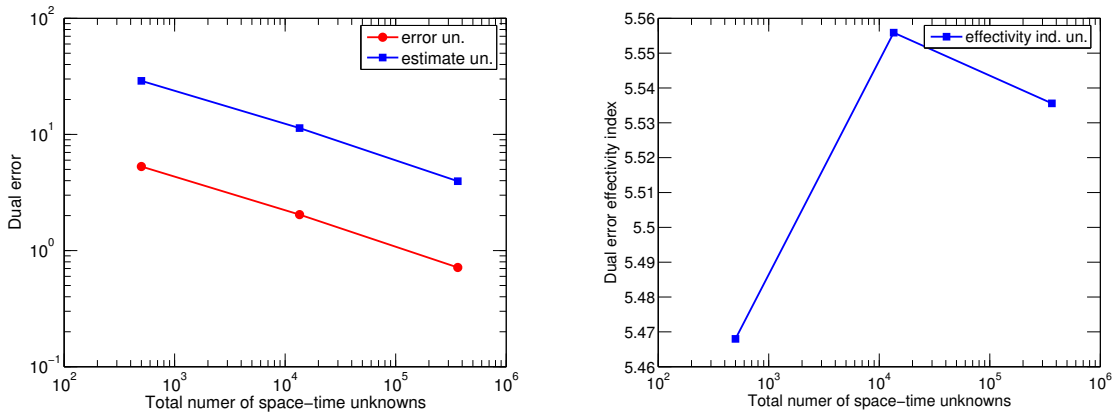


Figure 11.1: Estimated and actual $\|u - u_{h\tau}\|_Y$ error and corresponding effectivity index, final time $T = 1.5$

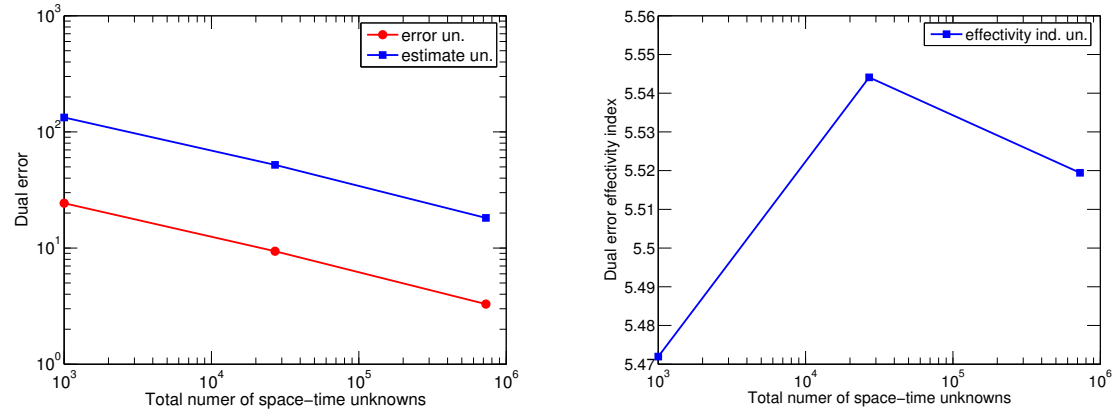


Figure 11.2: Estimated and actual $\|u - u_{h\tau}\|_Y$ error and corresponding effectivity index, final time $T = 3$

Figure 11.3: Estimated (left) and actual (right) energy error distribution for an unsteady advection–diffusion–reaction problem

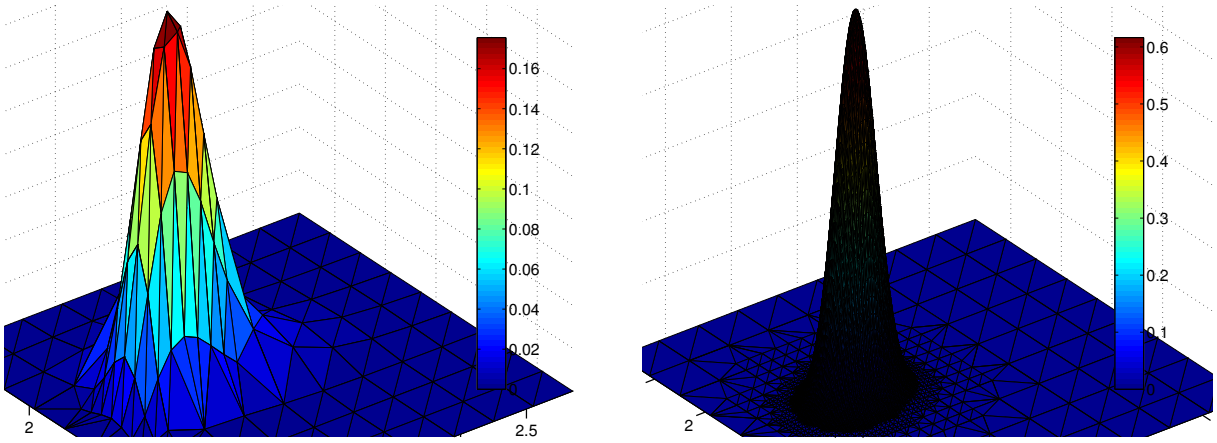


Figure 11.4: Examples of simulated concentration plumes based on space–time adaptivity, two (left) and four (right) levels of refinement maximum

Chapter 12

The nonlinear Laplace equation

We take here a brief look on the nonlinear problem (1.6a)–(1.6b). Recall the definition (1.5) of the nonlinear function $\boldsymbol{\sigma}(\nabla u)$ and let q be the dual exponent of p given by the relation $\frac{1}{p}+\frac{1}{q}=1$. The nonlinear Laplace equation reads: for $f \in L^q(\Omega)$, find u such that

$$-\nabla\cdot\boldsymbol{\sigma}(\nabla u) = f \qquad \text{in } \Omega, \tag{12.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{12.1b}$$

12.1 Variational formulation

The definition of the weak solution relies on the Sobolev space $W_0^{1,p}(\Omega)$, see [2, 9]:

Definition 12.1.1 (Variational formulation of (12.1a)–(12.1b))**.** *Find $u \in W_0^{1,p}(\Omega)$ such that*

$$(\boldsymbol{\sigma}(\nabla u), \nabla v) = (f, v) \qquad \forall v \in W_0^{1,p}(\Omega). \tag{12.2}$$

Also in this nonlinear case, we make the following definition:

Definition 12.1.2 (Flux)**.** *Let u be the solution of (12.2). Set*

$$\boldsymbol{\sigma} := -\boldsymbol{\sigma}(\nabla u). \tag{12.3}$$

We will call $\boldsymbol{\sigma}$ the flux*.*

Introducing the equivalent of the space $\mathbf{H}(\mathrm{div},\Omega)$ of Definition 4.2.2 in the $[L^q(\Omega)]^d$ setting, $\mathbf{H}^q(\mathrm{div},\Omega) := \{\mathbf{v} \in [L^q(\Omega)]^d;\ \nabla\cdot\mathbf{v} \in L^q(\Omega)\}$, where the divergence is to be taken in the weak sense, we have the following result:

Theorem 12.1.3 (Properties of the weak solution of (12.1a)–(12.1b))**.** *Let u be the solution of (12.2). Let $\boldsymbol{\sigma}$ be given by (12.3). Then*

$$u \in W_0^{1,p}(\Omega), \quad \boldsymbol{\sigma} \in \mathbf{H}^q(\mathrm{div},\Omega), \quad \nabla\cdot\boldsymbol{\sigma} = f.$$

The proof of this result follows exactly the same path as that of Theorem 7.1.3.

12.2 Approximate solution

We will once again suppose a general approximate solution u_h. For the sake of simplicity, we, however, suppose in this chapter that u_h is conforming, i.e.,

$$u_h \in W_0^{1,p}(\Omega). \tag{12.4}$$

In analogy with the previous chapters, we set:

Definition 12.2.1 (Approximate flux)**.** *Let u_h be the approximate solution, cf.* (12.4)*. Then*

$$-\boldsymbol{\sigma}(\nabla u_h) \tag{12.5}$$

will be called the approximate flux*.*

The following remark should be once again compared to Theorem 12.1.3:

Remark 12.2.2 (Properties of the approximate solution u_h of (12.4))**.** *Let u_h be the approximate solution, cf.* (12.4)*. Then*

$$-\boldsymbol{\sigma}(\nabla u_h) \notin \mathbf{H}^q(\mathrm{div},\Omega), \quad \nabla\cdot(-\boldsymbol{\sigma}(\nabla u_h)) \neq f \quad \textit{in general.}$$

12.3 Flux reconstruction

Taking into account that we suppose in this chapter for simplicity that $u_h \in W_0^{1,p}(\Omega)$, we are led to only introduce an equilibrated flux reconstruction $\boldsymbol{\sigma}_h$:

Definition 12.3.1 (Equilibrated flux reconstruction)**.** *We will call the* equilibrated flux reconstruction *any function $\boldsymbol{\sigma}_h$ constructed from u_h which satisfies*

$$\boldsymbol{\sigma}_h \in \mathbf{H}^q(\mathrm{div},\Omega), \tag{12.6a}$$

$$(\nabla\cdot\boldsymbol{\sigma}_h, 1)_K = (f,1)_K \qquad \forall K \in \mathcal{T}_h. \tag{12.6b}$$

12.4 Dual flux norm, the dual norm of the residual

Define also here the residual:

Definition 12.4.1 (Residual)**.** *Let $v_h \in W_0^{1,p}(\Omega)$. Then the residual of v_h, $\mathcal{R}(v_h) \in W_0^{1,p}(\Omega)'$, is defined by*

$$\langle \mathcal{R}(v_h), \varphi\rangle_{W_0^{1,p}(\Omega)',W_0^{1,p}(\Omega)} := (f,\varphi) - (\boldsymbol{\sigma}(\nabla v_h), \nabla\varphi) \qquad \varphi \in W_0^{1,p}(\Omega). \tag{12.7}$$

It appears advantageous in this nonlinear case to measure the distance between the exact solution u and $v_h \in W_0^{1,p}(\Omega)$ directly by the dual norm of the residual

$$\|\mathcal{R}(v_h)\|_{W_0^{1,p}(\Omega)'} := \sup_{\varphi\in W_0^{1,p}(\Omega);\, \|\nabla\varphi\|_p=1} \langle \mathcal{R}(v_h), \varphi\rangle_{W_0^{1,p}(\Omega)',W_0^{1,p}(\Omega)}. \tag{12.8}$$

Using (12.2), we also have

$$\|\mathcal{R}(v_h)\|_{W_0^{1,p}(\Omega)'} = \sup_{\varphi\in W_0^{1,p}(\Omega);\, \|\nabla\varphi\|_p=1} (\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla v_h), \nabla\varphi), \tag{12.9}$$

so that $\|\mathcal{R}(v_h)\|_{W_0^{1,p}(\Omega)'}$ is also a dual norm for the difference of the fluxes $\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla v_h)$. Remark that (12.9) takes precisely the form (7.8) (with $v = u - v_h$) known from the linear case. Whereas the equality (7.22) holds in the linear case, $\|\mathcal{R}(v_h)\|_{W_0^{1,p}(\Omega)'}$ is not equal to $\|\nabla(u - v_h)\|_p$ in the nonlinear one.

12.5 A general a posteriori error estimate

In order to state our a posteriori error estimate, we will need the following $W_0^{1,p}(\Omega)$-equivalent of the Poincaré inequality (4.20):

$$\|\varphi - \varphi_K\|_{p,K} \leq C_{\mathrm{P},p,K} h_K \|\nabla\varphi\|_{p,K} \qquad \forall \varphi \in W^{1,p}(K). \tag{12.10}$$

Here, owing to the convexity of simplices, $C_{\mathrm{P},p,K} = \pi^{-\frac{2}{p}} d^{\frac{1}{2}-\frac{1}{p}}$ for $p \geq 2$, see Verfürth [94], and $C_{\mathrm{P},p,K} = p^{\frac{1}{p}} 2^{\frac{(p-1)}{p}}$ for all $p \in (1, +\infty)$, see Chua and Wheeden [32].

In the following, we shall as usual assume:

Assumption 12.5.1 (Flux reconstruction for (12.1a)–(12.1b))**.** *We suppose that $\boldsymbol{\sigma}_h$ is an equilibrated flux reconstruction in the sense of Definition 12.3.1.*

Our general upper bound result is:

Theorem 12.5.2 (A general a posteriori error estimate for (12.1a)–(12.1b))**.** *Let u be the weak solution given by Definition 12.1.1. Let u_h be an arbitrary function satisfying (12.4). Let finally Assumption 12.5.1 be satisfied. For any $K \in \mathcal{T}_h$, define the* residual estimator *by*

$$\eta_{\mathrm{R},K} := C_{\mathrm{P},p,K} h_K \|f - \nabla\cdot\boldsymbol{\sigma}_h\|_{q,K}, \tag{12.11}$$

where $C_{\mathrm{P},p,K}$ is the constant from (12.10), and the flux estimator *by*

$$\eta_{\mathrm{F},K} := \|\boldsymbol{\sigma}(\nabla u_h) + \boldsymbol{\sigma}_h\|_{q,K}. \tag{12.12}$$

Then

$$\|\mathcal{R}(u_h)\|_{W_0^{1,p}(\Omega)'} \leq \left\{ \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{F},K} + \eta_{\mathrm{R},K})^q \right\}^{\frac{1}{q}}. \tag{12.13}$$

Proof. Consider a fixed $\varphi \in W_0^{1,p}(\Omega)$ with $\|\nabla\varphi\|_p = 1$ in (12.7), with $v_h = u_h$. Adding and subtracting $(\boldsymbol{\sigma}_h, \nabla\varphi)$ and using the Green theorem $(\boldsymbol{\sigma}_h, \nabla\varphi) = -(\nabla\cdot\boldsymbol{\sigma}_h, \varphi)$ (this is an equivalent of Theorem 4.2.5 on $W_0^{1,p}(\Omega) \times \mathbf{H}^q(\mathrm{div}, \Omega)$), we have

$$\langle \mathcal{R}(u_h), \varphi \rangle_{W_0^{1,p}(\Omega)', W_0^{1,p}(\Omega)} = (f - \nabla\cdot\boldsymbol{\sigma}_h, \varphi) - (\boldsymbol{\sigma}(\nabla u_h) + \boldsymbol{\sigma}_h, \nabla\varphi).$$

The Hölder inequality gives

$$-(\boldsymbol{\sigma}(\nabla u_h) + \boldsymbol{\sigma}_h, \nabla\varphi) \leq \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{F},K} \|\nabla\varphi\|_{p,K},$$

and the approximate equilibrium property (12.6b), the generalized Poincaré inequality (12.10), and the Hölder inequality give

$$\begin{aligned}(f - \nabla\cdot\boldsymbol{\sigma}_h, \varphi) &= \sum_{K\in\mathcal{T}_h} (f - \nabla\cdot\boldsymbol{\sigma}_h, \varphi)_K = \sum_{K\in\mathcal{T}_h} (f - \nabla\cdot\boldsymbol{\sigma}_h, \varphi - \varphi_K)_K \\ &\leq \sum_{K\in\mathcal{T}_h} C_{\mathrm{P},p,K} h_K \|f - \nabla\cdot\boldsymbol{\sigma}_h\|_{q,K} \|\nabla\varphi\|_{p,K} = \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{R},K} \|\nabla\varphi\|_{p,K}.\end{aligned}$$

Combining these results while using the Hölder inequality and $\|\nabla\varphi\|_p = 1$,

$$\langle \mathcal{R}(u_h), \varphi \rangle_{W_0^{1,p}(\Omega)', W_0^{1,p}(\Omega)} \leq \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{F},K} + \eta_{\mathrm{R},K}) \|\nabla\varphi\|_{p,K} \leq \left\{ \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{F},K} + \eta_{\mathrm{R},K})^q \right\}^{\frac{1}{q}},$$

which concludes the proof in view of the definition (12.8) of the dual norm of the residual $\|\mathcal{R}(u_h)\|_{W_0^{1,p}(\Omega)'}$. □

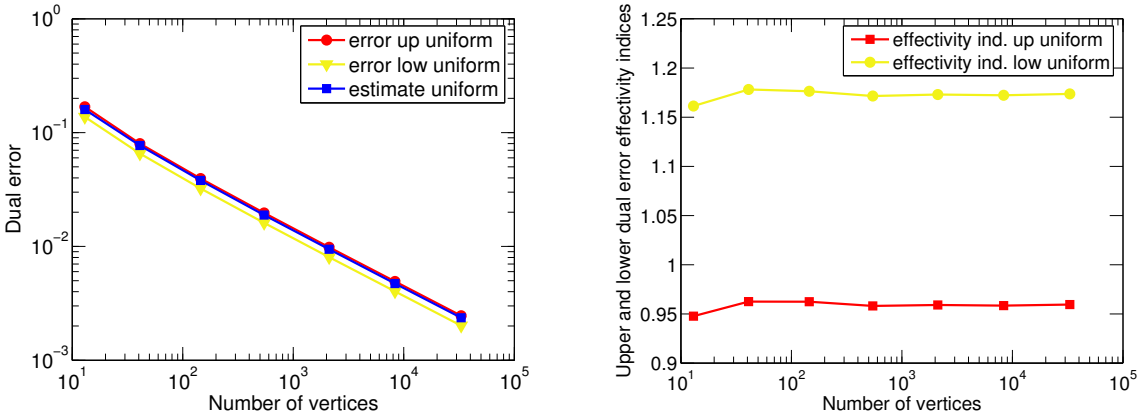


Figure 12.1: Estimated and actual dual errors and corresponding effectivity indices, nonlinear Laplace equation, $p = 1.4$

12.6 Application to classical discretization methods, efficiency, and robustness

As in Chapter 7, the estimate of Theorem 12.5.2 can be used for many numerical methods upon specifying the flux reconstruction $\boldsymbol{\sigma}_h$. Efficiency can then be shown: it is global only for the (global) norm $\|\mathcal{R}(u_h)\|_{W_0^{1,p}(\Omega)'}$ of (12.8). A local efficiency result can also be obtained. It holds true for the following upper bound on $\|\mathcal{R}(u_h)\|_{W_0^{1,p}(\Omega)'}$:

$$\|\mathcal{R}(u_h)\|_{W_0^{1,p}(\Omega)'} \leq \|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|_q. \tag{12.14}$$

Most importantly, these efficiencies are robust with respect to the nonlinear function a in (1.5) and with respect to the exponent p; property **iv)** of Section 1.4 is satisfied. This means that the overestimation of the error by our estimates and thus their quality does not depend on the size of the nonlinearity. We refer for details to [46].

12.7 Numerical examples

We give here a couple of examples of the performance of the above estimates. The finite element method (cf. Section 7.13.1), equivalent in the present case to the vertex-centered finite volume one (cf. Section 8.3.7), has been used in order to discretize (12.2) with a of (1.5) given as $a(x) = x^{p-2}$.

We first illustrate the robustness with respect to the nonlinearity (exponent p). In Figures 12.1 and 12.2, we plot the upper bound (12.14) and the lower bound

$$\|\mathcal{R}(u_h)\|_{W_0^{1,p}(\Omega)'} \geq \frac{|(\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h), \nabla(u - u_h))|}{\|\nabla(u - u_h)\|_p} \tag{12.15}$$

on the dual norm of the residual $\|\mathcal{R}(u_h)\|_{W_0^{1,p}(\Omega)'}$ of (12.8), the estimates of Theorem 12.5.2, and the corresponding effectivity indices. The effectivity index for $\|\mathcal{R}(u_h)\|_{W_0^{1,p}(\Omega)'}$, lying between these two, is thus independent of p, as predicted by the robustness efficiency result. Moreover, they are quite close to the optimal value of 1, i.e., to the asymptotic exactness (property **iii)** of Section 1.4).

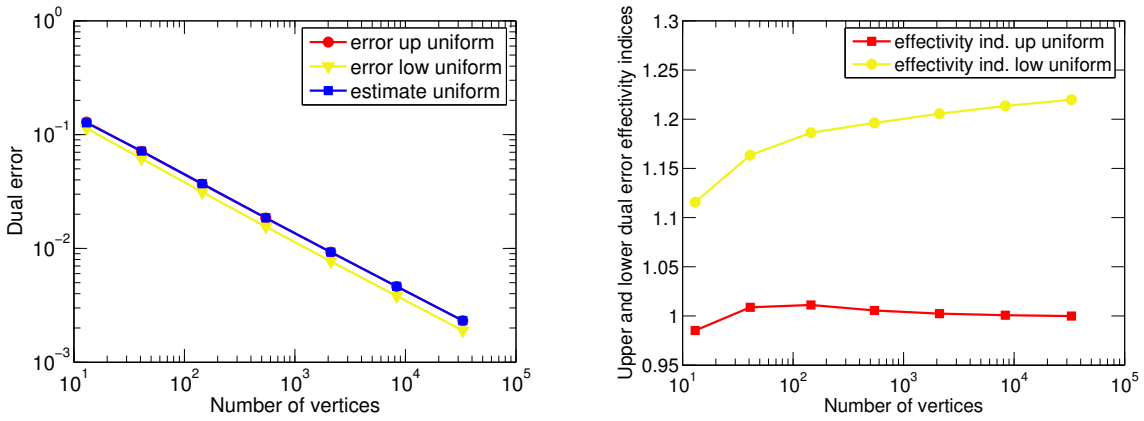


Figure 12.2: Estimated and actual dual errors and corresponding effectivity indices, nonlinear Laplace equation, $p = 3$

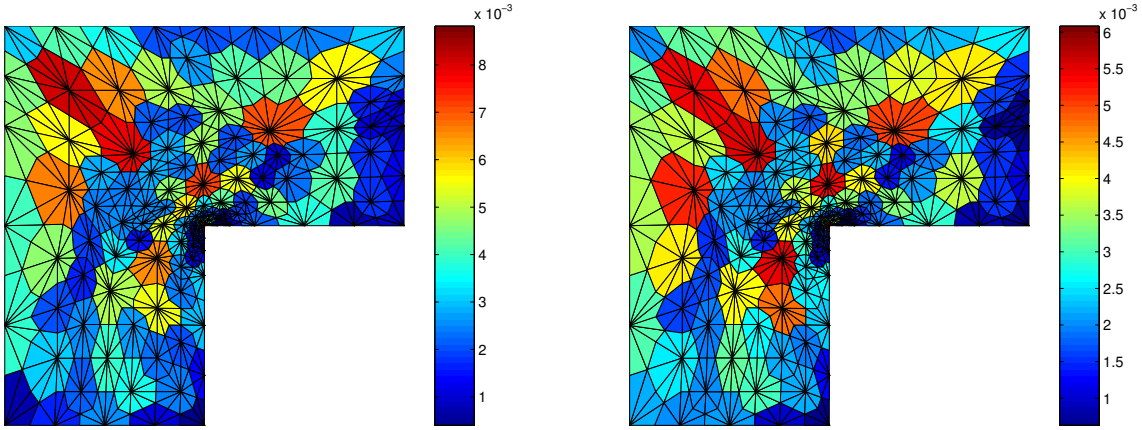


Figure 12.3: Estimated (left) and actual (right) error distribution, nonlinear Laplace equation

We next consider a test case with a singular solution. We show in the left part of Figure 12.3 the estimated error distribution for the dual volumes around each vertex, and in the right part of Figure 12.3 the exact error distribution, or, more precisely, its upper bound (12.14) (the values $\|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|_{q,D}$ for each dual volume D). We see that the two plots match nicely, which is a numerical evidence of the local efficiency. Our estimates obviously represent a very good prediction of the $[L^q(\Omega)]^d$-norm difference of the exact and approximate flux. Adaptive mesh refinement has been used in Figure 12.3 and allows for the resolution of the singularity residing in the reentrant corner.

Finally, Figure 12.4 shows that using this adaptive refinement strategy, much higher precision can be achieved for the given number of unknowns. As in the linear case, see Sections 7.14 and 8.4, the error decreases much faster (optimally in the function of number of unknowns) in the adaptive refinement, whereas in the uniform refinement, we can only achieve the decrease of the error given by the regularity of the weak solution. We refer for more details to [46].

Altogether, efficiency with precision attainment in the sense of the properties **1.**–**2.** of the Introduction can be achieved also for the nonlinear Laplace equation.

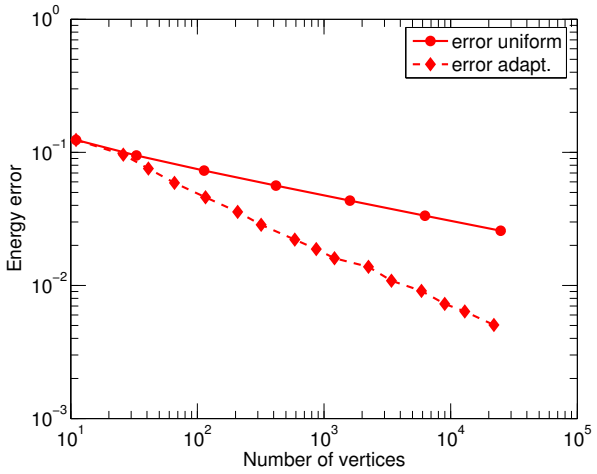


Figure 12.4: Energy errors $\|\nabla(u - u_h)\|_p$ on uniformly/adaptively refined meshes, nonlinear Laplace equation

Chapter 13

Stopping criteria for linear and nonlinear solvers and balancing different error components

A typical numerical simulation includes many sources of the error. Consequently, the overall error is composed of many *error components*. So far, in Chapters 7–12, we have focused on the *discretization error*, related to the numerical scheme chosen and the mesh $\mathcal{T}_h$. We now discuss other error components usually present in a numerical simulation and show how to identify them by a posteriori error estimates. Such a knowledge reveals in particular crucial for deriving optimal *stopping criteria* for various iterative algorithms and for balancing error components such as the spatial and temporal ones.

13.1 Algebraic error and stopping criteria for iterative algebraic solvers

In the numerical approximation of all the model problems of Chapters 7–12, there arise systems of linear algebraic equations. In order to obtain an approximate solution, these systems need to be solved. So far, we have supposed that these systems have been solved exactly. This is in our setting related to the fact that we have supposed the flux reconstructions to be equilibrated, satisfying exactly (7.21b)/(9.6b)/(10.8b)/(11.10b)/(12.6b).

In [65], a posteriori error estimates enabling to take into account the error stemming from the fact that a linear system is not solved exactly were derived. More precisely, two components of the error, the discretization error and the *algebraic error* were distinguished. Then, for an iterative algebraic solver, a reasonable stopping criterion is to cease the iterations whenever the algebraic error does not contribute significantly to the overall error. The major idea, see also Becker et al. [18], Patera and Rønquist [77], Arioli et al. [11], or Picasso [80], is explained on the example of Figure 13.1.

In this figure, we consider the problem (8.72a)–(8.72b) for two different tensors $\underline{\mathbf{K}}$ discretized by the cell-centered finite volume method (cf. Section 8.3.6). The mesh $\mathcal{T}_h$ is fixed and we plot the evolution of the energy error as a function of the number of iterations of the preconditioned conjugate gradients iterative solver. The behavior is characteristic: in first cca 23 iterations, the overall error decreases, but it stagnates for all successive iterations. At the beginning (we start from a zero initial vector), the algebraic error (estimated by $\eta_{\mathrm{AE}}^{(3)}$) dominates. Then the algebraic error gets small in comparison with the discretization one (estimated

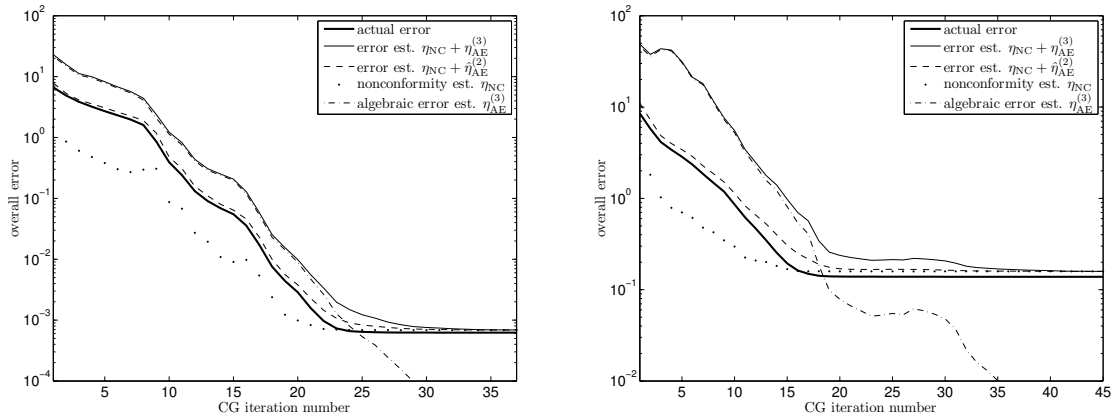


Figure 13.1: Energy error, overall estimators, and the algebraic and discretization estimators as a function of the number of iterations of the conjugate gradients iterative solver, problem (8.72a)–(8.72b)

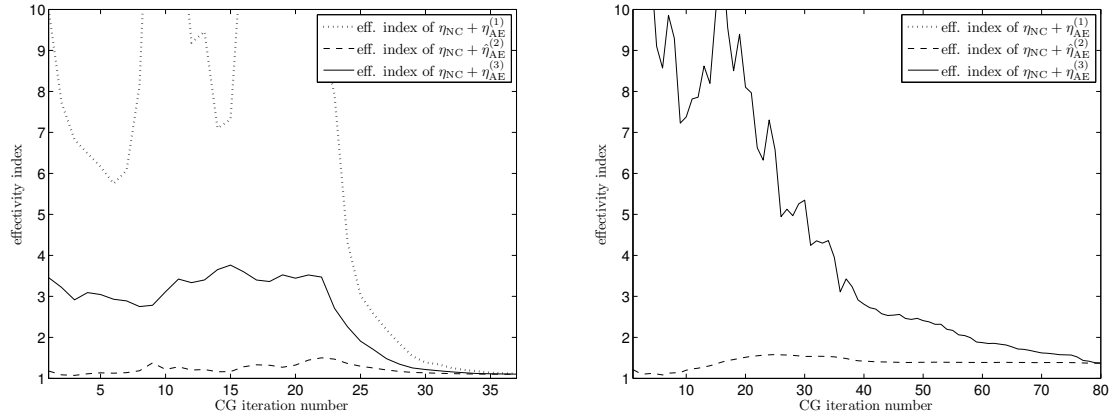


Figure 13.2: Effectivity indices for a posteriori error estimates including the algebraic error

by η_{NC}), and the overall error stagnates, as the discretization error becomes dominant.

The stopping criterion that we advocate roughly says that we should cease the algebraic solver iteration when the curves of η_{NC} and $\eta_{AE}^{(3)}$ in Figure 13.1 cross. An important number of the algebraic solver iterations, where the overall error does not improve anymore and where the CPU time is literally wasted, may be sparred. In Figure 13.1, we also plot two different overall error estimators ($\eta_{NC}+\eta_{AE}^{(3)}$ and $\eta_{NC}+\hat{\eta}_{AE}^{(2)}$) showing our final error estimate including the algebraic error. The corresponding effectivity indices are reported in Figure 13.2. We see that also in presence of the algebraic error, they are nicely close to the optimal value of one, so that we are close to the asymptotic exactness.

We present in Section 13.3 below a way how to prove such type of results, in a larger context of an inexact Newton method.

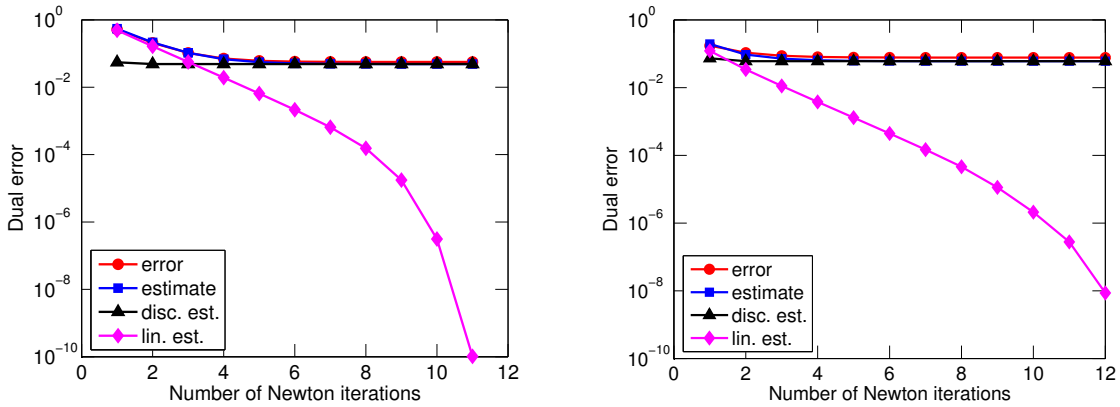


Figure 13.3: $[L^q(\Omega)]^d$-error of the fluxes, overall estimator, and the linearization and discretization estimators as a function of the number of iterations of the Newton iterative solver, the nonlinear Laplacian, $p = 10$ (left), $p = 50$ (right)

13.2 Linearization error and stopping criteria for iterative linearization solvers

One particularity of the nonlinear case of Chapter 12 is that in numerical approximations, some kind of iterative linearization, like the fixed-point or Newton ones, is usually employed. Then one is not only interested in the overall error between u and u_h, but also in the *linearization error*, as a second error component in addition to the discretization error. A natural stopping criterion for iterative linearization is to cease the iterations whenever the linearization error does not contribute significantly to the overall error. The major idea, see also Han [62] or Chaillou and Suri [30, 31], is explained on the example of Figure 13.3 from [46].

In this figure, we plot the evolution of the error as a function of the number of iterations of the Newton iterative nonlinear solver for the model problem (12.1a)–(12.1b) discretized by the finite element (vertex-centered finite volume) method, cf. Section 12.7. The behavior is characteristic: in first cca 5 iterations, the error decreases, but it stagnates for all successive iterations. At the beginning, the linearization error dominates. Then, however, the linearization error gets small in comparison with the discretization one, and the overall error stagnates, as the discretization error becomes dominant.

The stopping criterion that we advocate roughly says that we should stop the linearization solver iteration when the curves of the discretization error estimator and linearization error estimator in Figure 13.3 cross. An important number of the linearization solver iterations, where the overall error does not improve anymore and where the CPU time is literally wasted, may be sparred, which is evident from Figure 13.3.

We present in the following section a way how to prove such type of results, in a larger context of an inexact Newton method.

13.3 An adaptive inexact Newton method

We consider here the nonlinear Laplace equation (12.1a)–(12.1b) of Chapter 12. We show how to simultaneously take into account the linearization and algebraic errors and derive stopping criteria for iterative linearizations and iterative algebraic solvers.

Suppose that some numerical method has been applied to the discretization of (12.1a)–(12.1b). This gives rise to a system of nonlinear algebraic equations written in the form: find a vector $U \in \mathbb{R}^N$ such that

$$\mathcal{A}(U) = F, \tag{13.1}$$

where $\mathcal{A} : \mathbb{R}^N \to \mathbb{R}^N$ is a discrete nonlinear operator and $F \in \mathbb{R}^N$ a given vector. We will do our presentation in the framework of an adaptive inexact form of the Newton method proposed in [53]. Here $\eta_{\mathrm{disc}}^{k,i}$, $\eta_{\mathrm{lin}}^{k,i}$, $\eta_{\mathrm{alg}}^{k,i}$, and $\eta_{\mathrm{rem}}^{k,i}$ are estimators discussed subsequently and γ_{rem}, γ_{alg}, and γ_{lin} are positive user-given weights, typically of order 0.1.

Algorithm 13.3.1 (Adaptive inexact Newton method)**.**

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

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

$$\mathbb{A}^k U^k = F^k. \tag{13.2}$$

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

   (b) Perform a step of a chosen iterative linear solver for the solution of the linear system (13.2), starting from the vector $U^{k,i-1}$. This yields an approximation $U^{k,i}$ to U^k which satisfies

$$\mathbb{A}^k U^{k,i} = F^k - R^{k,i}, \tag{13.3}$$

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

   (c) Perform $\nu > 0$ additional steps of the iterative linear solver yielding an approximation $U^{k,i+\nu}$ to U^k which satisfies

$$\mathbb{A}^k U^{k,i+\nu} = F^k - R^{k,i+\nu}, \tag{13.4}$$

   where $R^{k,i+\nu} \in \mathbb{R}^N$ is the algebraic residual vector on step $i+\nu$. The parameter ν is progressively increased until

$$\eta_{\mathrm{rem}}^{k,i} \le \gamma_{\mathrm{rem}} \max\{\eta_{\mathrm{disc}}^{k,i}, \eta_{\mathrm{lin}}^{k,i}, \eta_{\mathrm{alg}}^{k,i}\}. \tag{13.5}$$

   (d) Check the stopping criterion for the linear solver in the form

$$\eta_{\mathrm{alg}}^{k,i} \le \gamma_{\mathrm{alg}} \max\{\eta_{\mathrm{disc}}^{k,i}, \eta_{\mathrm{lin}}^{k,i}\}. \tag{13.6}$$

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

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

$$\eta_{\mathrm{lin}}^{k,i} \le \gamma_{\mathrm{lin}} \eta_{\mathrm{disc}}^{k,i}. \tag{13.7}$$

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

As usual, we rely on an equilibrated flux reconstruction. Weakening (12.6b) from Definition 12.3.1, we are lead to suppose

$$\boldsymbol{\sigma}_h^{k,i} \in \mathbf{H}^q(\mathrm{div}, \Omega), \tag{13.8a}$$

$$(\nabla\cdot\boldsymbol{\sigma}_h^{k,i}, 1)_K = (f - \rho_h^{k,i}, 1)_K \qquad \forall K \in \mathcal{T}_h, \tag{13.8b}$$

where $\rho_h^{k,i}$ is a small enough (in the sense of (13.5), see below) piecewise polynomial function called *algebraic remainder*. Our key idea is to decompose $\boldsymbol{\sigma}_h^{k,i}$ in the following way, where $\mathbf{d}_h^{k,i}$ is meant to approximate the *discretization* flux, $\mathbf{l}_h^{k,i}$ represents the *linearization error*, and $\mathbf{a}_h^{k,i}$ the *algebraic error*.

Assumption 13.3.2 (Discretization, linearization error, and algebraic error fluxes)**.** *There exist fluxes* $\mathbf{d}_h^{k,i}, \mathbf{l}_h^{k,i}, \mathbf{a}_h^{k,i} \in [L^q(\Omega)]^d$ *such that*

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

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

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

Theorem 12.5.2, the triangle inequality, and the $W_0^{1,p}(\Omega)$-version of the Friedrichs inequality (4.21) then give:

Theorem 13.3.3 (An a posteriori error estimate for (12.1a)–(12.1b) distinguishing the discretization, linearization, and algebraic error components)**.** *Let* u *be the weak solution given by Definition 12.1.1. Let* $u_h^{k,i} \in W_0^{1,p}(\Omega)$ *be an approximate solution associated with* $U^{k,i}$ *on linearization step* k *and algebraic step* i *of Algorithm 13.3.1. Let finally* (13.8a)–(13.8b) *together with Assumption 13.3.2 be satisfied. For any* $K \in \mathcal{T}_h$*, define the* discretization estimator *by*

$$\eta_{\mathrm{disc},K}^{k,i} := C_{\mathrm{P},p,K} h_K \|f - \nabla\cdot\boldsymbol{\sigma}_h^{k,i} - \rho_h^{k,i}\|_{q,K} + \|\boldsymbol{\sigma}(\nabla u_h^{k,i}) + \mathbf{d}_h^{k,i}\|_{q,K}, \tag{13.9}$$

the linearization estimator

$$\eta_{\mathrm{lin},K}^{k,i} := \|\mathbf{l}_h^{k,i}\|_{q,K}, \tag{13.10}$$

the algebraic estimator

$$\eta_{\mathrm{alg},K}^{k,i} := \|\mathbf{a}_h^{k,i}\|_{q,K}, \tag{13.11}$$

and the algebraic remainder estimator

$$\eta_{\mathrm{rem},K}^{k,i} := h_\Omega \|\rho_h^{k,i}\|_{q,K}. \tag{13.12}$$

Define global versions of these estimators as $\eta_\cdot^{k,i} := \left\{\sum_{K\in\mathcal{T}_h} \left(\eta_{\cdot,K}^{k,i}\right)^q\right\}^{1/q}$*. Then*

$$\|\mathcal{R}(u_h)\|_{W_0^{1,p}(\Omega)'} \le \eta_{\mathrm{disc}}^{k,i} + \eta_{\mathrm{lin}}^{k,i} + \eta_{\mathrm{alg}}^{k,i} + \eta_{\mathrm{rem}}^{k,i}. \tag{13.13}$$

Practically, one proceeds as follows. We first construct the flux $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})$, in $\mathbf{H}^q(\mathrm{div},\Omega)$, using the given numerical scheme as in Section 7.13. We next repeat the same construction on the algebraic solver step $i+\nu$ of (13.4) and define

$$\mathbf{a}_h^{k,i} := (\mathbf{d}_h^{k,i+\nu} + \mathbf{l}_h^{k,i+\nu}) - (\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i}), \tag{13.14a}$$

$$\rho_h^{k,i} := r_h^{k,i+\nu}, \tag{13.14b}$$

where $r_h^{k,i+\nu}$ is a piecewise polynomial constructed from the residual vector $R^{k,i+\nu}$ of (13.4). Finally, the flux $\mathbf{l}_h^{k,i}$ is constructed from the given linearization. Concerning the efficiency and robustness, the same comments as in Section 12.6 hold true here as well. We refer for this, all the details, and extensions to nonconforming methods to [53].

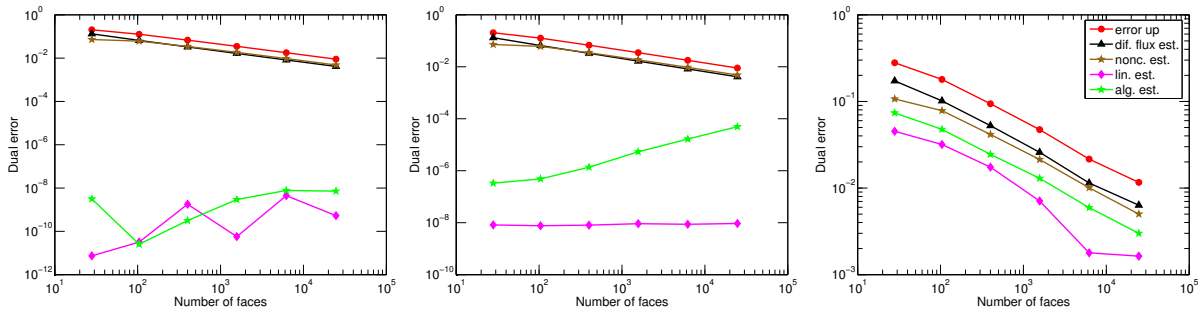


Figure 13.4: Error and estimators on uniformly refined meshes. Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

We conclude this section by some numerical illustration. We consider (12.1a) with $a(x) = x^{p-2}$, $p = 10$, $\Omega := (0,1) \times (0,1)$, and $f := 2$. In place of (12.1b), we prescribe the Dirichlet boundary condition by the exact solution

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

We employ the Crouzeix–Raviart nonconforming finite element method (cf. Section 7.13.2) for the discretization, the Newton linearization, and the conjugate gradient algebraic solver with diagonal preconditioning.

We compare three different stopping criteria in Algorithm 13.3.1, leading to three different solution approaches:

- In the *Full Newton (FN) method*, both the nonlinear and linear solvers are iterated to "almost" convergence. More precisely, we use the stopping criteria $\eta_{\mathrm{alg}}^{k,i} \leq 10^{-8}$ and $\eta_{\mathrm{lin}}^{k,i} \leq 10^{-8}$. This is the classical approach.
- In the *Inexact Newton (IN) method*, the only difference with FN is that a fixed number of preconditioned CG iterations is performed on each Newton linearization step. These values were chosen respectively as 2, 3, 5, 8, 10, 15 on each level of uniform mesh refinement. This approach typically reduces the computational requirements of the previous one and is very popular in the engineering practice.
- Finally, in the *Adaptive Inexact Newton (AIN) method* of this section, we rely on the stopping criteria (13.5), (13.6), and (13.7) with $\gamma_{\mathrm{rem}} = \gamma_{\mathrm{alg}} = \gamma_{\mathrm{lin}} = 0.3$.

We start by Figure 13.4 which displays the curves of the overall error ($\|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h^{k,i})\|_q$ augmented by a jump seminorm), of the estimator of Theorem 12.5.2, and of the nonconformity estimator as a function of the number of mesh faces. We observe that the three methods (FN, IN, and AIN) almost do not differ for these quantities. Figure 13.4 also displays the curves of the linearization estimator $\eta_{\mathrm{lin}}^{k,i}$ and of the algebraic estimator $\eta_{\mathrm{alg}}^{k,i}$ of Theorem 13.3.3. The conceptual difference between the three methods lies in the size and behavior of these two estimators: both take values below 10^{-8} for FN; $\eta_{\mathrm{alg}}^{k,i}$ takes larger values for IN; both $\eta_{\mathrm{alg}}^{k,i}$ and $\eta_{\mathrm{lin}}^{k,i}$ take larger values that are just sufficiently small so as not to influence the error and estimators for AIN.

Figure 13.5 then focuses on the 6th level uniformly refined mesh and tracks the dependence on the Newton iterations. Typically, the error and all the estimators except $\eta_{\mathrm{lin}}^{k,i}$ start to

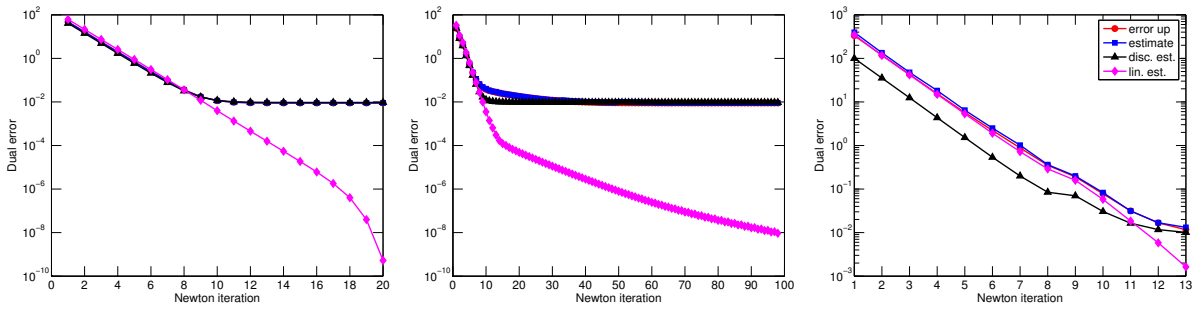


Figure 13.5: Error and estimators as a function of Newton iterations, 6th level mesh. Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

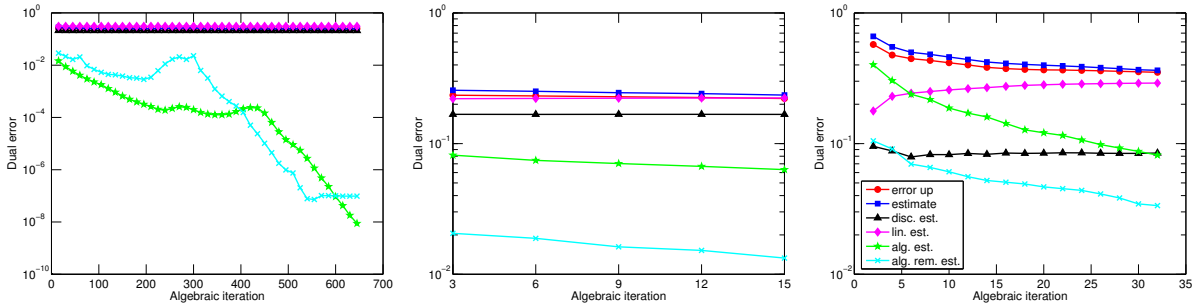


Figure 13.6: Error and estimators as a function of preconditioned CG iterations, 6th level mesh. Newton, 6th step (left), inexact Newton, 6th step (middle), and adaptive inexact Newton, 8th step (right)

stagnate after the linearization error ceases to dominate. This is precisely the point where the nonlinear iteration is stopped in AIN, whereas both FN and IN perform many unnecessary additional iterations.

Figure 13.6 further analyzes the situation on one chosen Newton iteration from Figure 13.5. We see that almost no decrease of the error can be observed during the almost 650 iterations of the preconditioned CG method in the FN case. The fixed 15 CG iterations in the IN case are, on the contrary, not completely sufficient to decrease significantly the error. In the adaptive approach, just the sufficient, "online-decided" number of CG iterations is performed.

Figure 13.7 illustrates the overall performance of the three approaches. We can see that the number of Newton iterations is smaller in the adaptive case than for FN, whereas it is significantly higher for IN. On one Newton iteration (example for the 6th level refined mesh), the number of CG iterations also varies significantly between the three approaches; in particular, AIN picks up the number that is "just necessary." The total number of necessary CG iterations per refinement level is finally displayed in the right part of Figure 13.7. The adaptive approach yields an economy by a factor of roughly 5 with respect to IN and roughly 30 with respect to FN in terms of total number of iterations.

To conclude, Figure 13.8 displays the distribution of the overall error estimator and of the error on the 2nd level uniformly refined mesh for AIN. Even in presence of algebraic and linearization errors, the overall error distribution is very well predicted.

More details on all the presented developments can be found in [53].

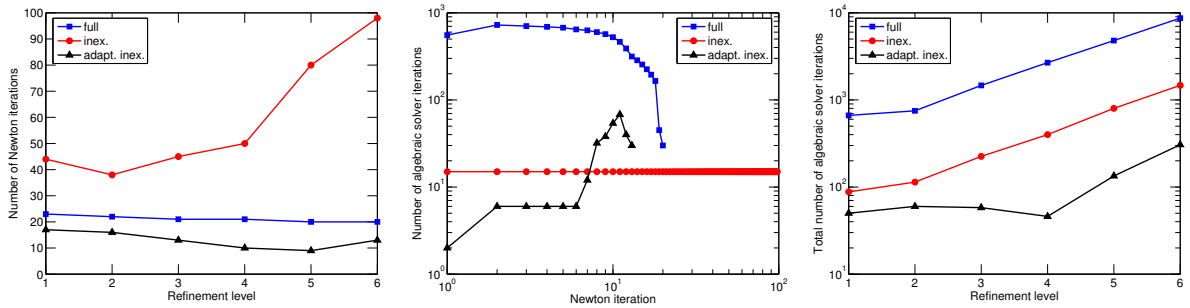


Figure 13.7: Number of Newton iterations per refinement level (left), number of linear solver iterations per Newton step on the 6th level mesh (middle), and total number of linear solver iterations per refinement level (right)

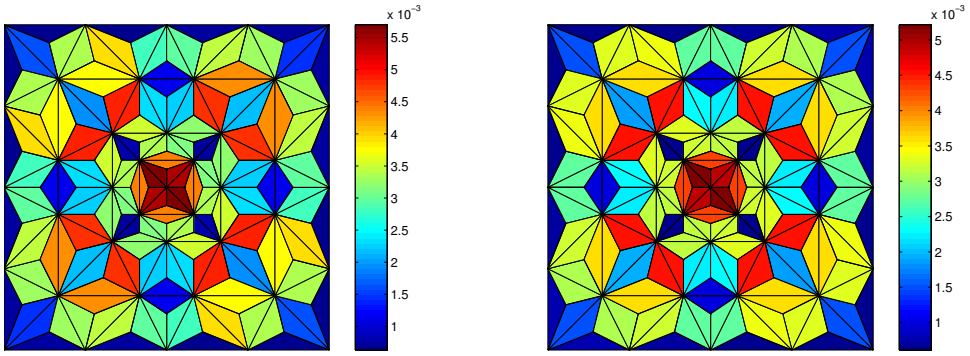


Figure 13.8: Estimated (left) and actual (right) error distribution, 2nd level uniformly refined mesh, adaptive inexact Newton

13.4 Balancing spatial and temporal errors

We now turn to unsteady problems. Similarly as in Sections 13.1 and 13.2, two main error components may be identified herein: the *spatial discretization error* and the *temporal discretization error*. The first one is obviously connected to the spatial discretization scheme chosen and the mesh elements sizes h_K, whereas the second one to the temporal discretization scheme chosen and the time steps τ^n.

In the heat equation setting of Chapter 11, error components identification can be achieved via the triangle inequality in Theorem 11.6.2. Define, for all $1 \leq n \leq N$,

$$(\eta_{\mathrm{sp}}^n)^2 := \sum_{K \in \mathcal{T}_h^n} 3 \left\{ \tau^n (9(\eta_{\mathrm{R},K}^n + \eta_{\mathrm{F},1,K}^n)^2 + (\eta_{\mathrm{NC},2,K}^n)^2) + \int_{I_n} (\eta_{\mathrm{NC},1,K}^n)^2(t) \, \mathrm{d}t \right\}, \tag{13.15}$$

$$(\eta_{\mathrm{tm}}^n)^2 := \sum_{K \in \mathcal{T}_h^n} 3\tau^n \|\nabla(s_h^n - s_h^{n-1})\|_K^2, \tag{13.16}$$

where

$$\eta_{\mathrm{F},1,K}^n := \|\nabla s_h^n + \boldsymbol{\sigma}_h^n\|_K. \tag{13.17}$$

We then have:

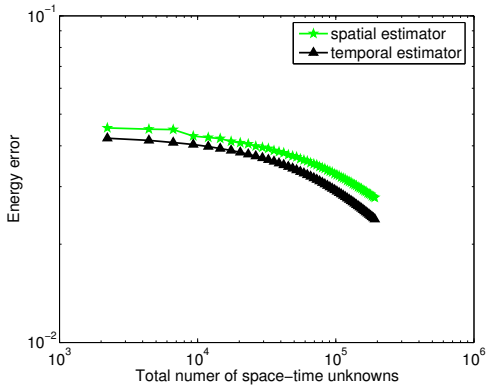


Figure 13.9: Spatial estimators η_{sp}^n and temporal estimators η_{tm}^n equilibrated, as a function of the total number of space–time unknowns

Theorem 13.4.1 (Distinguishing the spatial and temporal errors for (11.1a)–(11.1c))**.** *Under the assumptions of Theorem 11.6.2, there holds*

$$\|u - u_{h\tau}\|_Y \leq \left\{\sum_{n=1}^{N} (\eta_{\mathrm{sp}}^n)^2\right\}^{\frac{1}{2}} + \left\{\sum_{n=1}^{N} (\eta_{\mathrm{tm}}^n)^2\right\}^{\frac{1}{2}} + \eta_{\mathrm{IC}}.$$

The goal in an efficient numerical simulation is to achieve the equilibration of the spatial and temporal error components,

$$\eta_{\mathrm{sp}}^n \approx \eta_{\mathrm{tm}}^n \qquad \forall 1 \leq n \leq N, \tag{13.18}$$

as well as the equilibration of the spatial error in all mesh elements,

$$\eta_{\mathrm{sp},K}^n \approx \eta_{\mathrm{sp},K'}^n \qquad \forall 1 \leq n \leq N,\ \forall K, K' \in \mathcal{T}_h^n,$$

where

$$\eta_{\mathrm{sp},K}^n := \eta_{\mathrm{R},K}^n + \eta_{\mathrm{F},1,K}^n + \eta_{\mathrm{NC},2,K}^n + \eta_{\mathrm{NC},1,K}^n.$$

This can be achieved via adaptive choice of the time step and adaptive mesh refinement through algorithms such as those presented in Picasso [79], Makridakis and Nochetto [73], Verfürth [95], Bergam et al. [19], or [51, 64].

We present a quick illustration from [64] that the space–time error equilibration (13.18) is indeed advantageous. Figure 13.9 presents a result of an adaptive calculation where (13.18) is satisfied. In Figure 13.10, in contrast, we overrefine in time while choosing much finer time steps, and in Figure 13.11, we overrefine in space while choosing much finer spatial meshes. In the left parts of these figures, we can see that the spatial and temporal estimators are now disequilibrated. Much worse precision for a given computational effort is now achieved in comparison with the equilibrated case, as we can see in the right parts of the Figures 13.10 and 13.11.

13.5 A fully adaptive algorithm for unsteady nonlinear problems

We now finally summarize and generalize the developments of Sections 13.1–13.4.

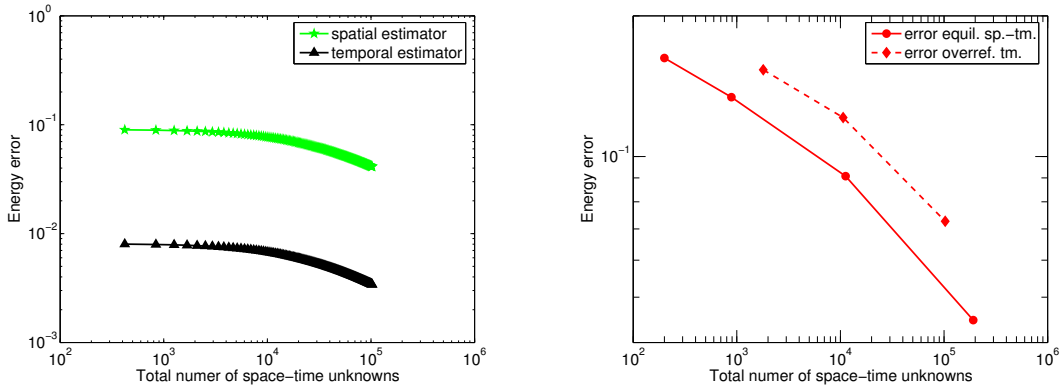


Figure 13.10: Spatial and temporal estimators for overrefinement in time (left) and comparison of the corresponding energy error with the equilibrated case (right)

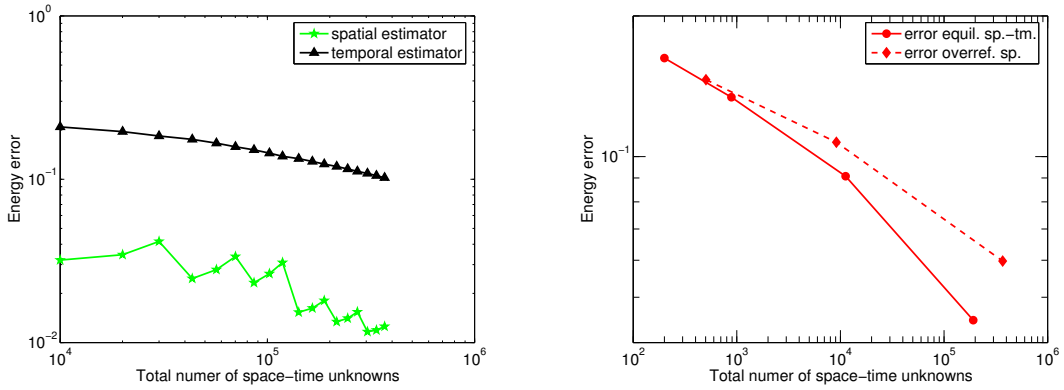


Figure 13.11: Spatial and temporal estimators for overrefinement in space (left) and comparison of the corresponding energy error with the equilibrated case (right)

Many real-life problems are described by unsteady nonlinear problems. A numerical algorithm for such a problem typically involves several iterative procedures; there is usually a loop over time steps, a linearization iteration, and an algebraic solver iteration. For as efficient as possible numerical algorithm, one should intend, at each moment of the calculation:

i) distinguish and estimate separately the different error components (*error components identification and separation*);

ii) classify the error components into two groups: *substantial error components* (crucial for the calculation, those errors which will always be present (e.g., spatial discretization error, temporal discretization error)) and *subsidiary error components* (side for the calculation, those errors which are in general made small or even zero for a sufficient number of iterations (e.g., linearization error, linear algebraic solver error));

iii) stop the different iterative algorithms whenever the corresponding subsidiary errors drop to the level at which they do not affect significantly the overall error (*stopping criteria*);

iv) adjust the calculation parameters (e.g., space meshes and time steps) such that the substantial errors are equally distributed and of comparable size (*error components equilibration*).

We have seen in Sections 13.3 and 13.4 how to achieve this in the context of the model problems of Chapters 12 and 11, respectively. Extensions to (coupled) unsteady nonlinear degenerate problems have recently been performed in [26] for the two-phase flow in porous media and in [42] for the Stefan problem.

We end up these lecture notes by a citation from Baxter and Iserles [16, p. 273]: "The purpose of computation is not to produce a solution with least error but to produce reliably, robustly and affordably a solution which is within a user-specified tolerance." Hopefully, an attentive reader can now see that the a posteriori error estimates indeed enable to achieve such a goal, or more precisely the efficiency with precision attainment in the sense of the properties **1.**–**2.** of the Introduction. Adaptive strategies satisfying properties **i)**–**iv)** appear crucial in this respect.

Bibliography

[1] Achdou, Y., Bernardi, C., and Coquel, F. A priori and a posteriori analysis of finite volume discretizations of Darcy's equations. *Numer. Math. 96*, 1 (2003), 17–42.

[2] Adams, R. A. *Sobolev spaces.* Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 65.

[3] Ainsworth, M. Robust a posteriori error estimation for nonconforming finite element approximation. *SIAM J. Numer. Anal. 42*, 6 (2005), 2320–2341.

[4] Ainsworth, M. A posteriori error estimation for discontinuous Galerkin finite element approximation. *SIAM J. Numer. Anal. 45*, 4 (2007), 1777–1798.

[5] Ainsworth, M., Allendes, A., Barrenechea, G. R., and Rankin, R. On the adaptive selection of the parameter in stabilized finite element approximations. *SIAM J. Numer. Anal. 51*, 3 (2013), 1585–1609.

[6] Ainsworth, M., and Oden, J. T. *A posteriori error estimation in finite element analysis.* Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2000.

[7] Ainsworth, M., and Rankin, R. Fully computable bounds for the error in nonconforming finite element approximations of arbitrary order on triangular elements. *SIAM J. Numer. Anal. 46*, 6 (2008), 3207–3232.

[8] Ainsworth, M., and Rankin, R. Constant free error bounds for nonuniform order discontinuous Galerkin finite-element approximation on locally refined meshes with hanging nodes. *IMA J. Numer. Anal. 31*, 1 (2011), 254–280.

[9] Allaire, G. *Analyse numérique et optimisation.* Editions Ecole Polytechnique, Palaiseau, France, 2005.

[10] Arbogast, T., and Chen, Z. On the implementation of mixed methods as nonconforming methods for second-order elliptic problems. *Math. Comp. 64*, 211 (1995), 943–972.

[11] Arioli, M., Loghin, D., and Wathen, A. J. Stopping criteria for iterations in finite element methods. *Numer. Math. 99*, 3 (2005), 381–410.

[12] Arnold, D. N., and Brezzi, F. Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. *RAIRO Modél. Math. Anal. Numér. 19*, 1 (1985), 7–32.

[13] Babuška, I., and Strouboulis, T. *The finite element method and its reliability.* Numerical Mathematics and Scientific Computation. The Clarendon Press Oxford University Press, New York, 2001.

[14] Babuška, I., Strouboulis, T., and Gangaraj, S. K. Guaranteed computable bounds for the exact error in the finite element solution. I. One-dimensional model problem. *Comput. Methods Appl. Mech. Engrg. 176*, 1-4 (1999), 51–79. New advances in computational methods (Cachan, 1997).

[15] Bastian, P., and Rivière, B. Superconvergence and $H(\mathrm{div})$ projection for discontinuous Galerkin methods. *Internat. J. Numer. Methods Fluids 42*, 10 (2003), 1043–1057.

[16] Baxter, B. J. C., and Iserles, A. On the foundations of computational mathematics. In *Handbook of Numerical Analysis, Vol. XI.* North-Holland, Amsterdam, 2003, pp. 3–34.

[17] Bebendorf, M. A note on the Poincaré inequality for convex domains. *Z. Anal. Anwendungen 22*, 4 (2003), 751–756.

[18] Becker, R., Johnson, C., and Rannacher, R. Adaptive error control for multigrid finite element methods. *Computing 55*, 4 (1995), 271–288.

[19] Bergam, A., Bernardi, C., and Mghazli, Z. A posteriori analysis of the finite element discretization of some parabolic equations. *Math. Comp. 74*, 251 (2005), 1117–1138.

[20] Braess, D. An a posteriori error estimate and a comparison theorem for the nonconforming P_1 element. *Calcolo 46*, 2 (2009), 149–155.

[21] Braess, D., Pillwein, V., and Schöberl, J. Equilibrated residual error estimates are p-robust. *Comput. Methods Appl. Mech. Engrg. 198*, 13-14 (2009), 1189–1197.

[22] Braess, D., and Schöberl, J. Equilibrated residual error estimator for edge elements. *Math. Comp. 77*, 262 (2008), 651–672.

[23] Brenner, S. C. Poincaré-Friedrichs inequalities for piecewise H^1 functions. *SIAM J. Numer. Anal. 41*, 1 (2003), 306–324.

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

[25] Burman, E., and Ern, A. Continuous interior penalty hp-finite element methods for advection and advection-diffusion equations. *Math. Comp. 76*, 259 (2007), 1119–1140.

[26] Cancès, C., Pop, I. S., and Vohralík, M. An a posteriori error estimate for vertex-centered finite volume discretizations of immiscible incompressible two-phase flow. *Math. Comp. 83*, 285 (2014), 153–188.

[27] Carstensen, C., and Funken, S. A. Fully reliable localized error control in the FEM. *SIAM J. Sci. Comput. 21*, 4 (1999/00), 1465–1484.

[28] Carstensen, C., and Funken, S. A. Constants in Clément-interpolation error and residual based a posteriori error estimates in finite element methods. *East-West J. Numer. Math. 8*, 3 (2000), 153–175.

[29] Carstensen, C., and Merdon, C. Computational survey on a posteriori error estimators for nonconforming finite element methods for the Poisson problem. *J. Comput. Appl. Math. 249* (2013), 74–94.

[30] Chaillou, A. L., and Suri, M. Computable error estimators for the approximation of nonlinear problems by linearized models. *Comput. Methods Appl. Mech. Engrg. 196*, 1-3 (2006), 210–224.

[31] Chaillou, A. L., and Suri, M. A posteriori estimation of the linearization error for strongly monotone nonlinear operators. *J. Comput. Appl. Math. 205*, 1 (2007), 72–87.

[32] Chua, S.-K., and Wheeden, R. L. Estimates of best constants for weighted Poincaré inequalities on convex domains. *Proc. London Math. Soc. (3) 93*, 1 (2006), 197–226.

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

[34] Cochez-Dhondt, S., and Nicaise, S. Equilibrated error estimators for discontinuous Galerkin methods. *Numer. Methods Partial Differential Equations 24*, 5 (2008), 1236–1252.

[35] Costabel, M., and McIntosh, A. On Bogovskiĭ and regularized Poincaré integral operators for de Rham complexes on Lipschitz domains. *Math. Z. 265*, 2 (2010), 297–320.

[36] Crouzeix, M., and Raviart, P.-A. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge 7*, R-3 (1973), 33–75.

[37] Dari, E., Durán, R., Padra, C., and Vampa, V. A posteriori error estimators for nonconforming finite element methods. *RAIRO Modél. Math. Anal. Numér. 30*, 4 (1996), 385–400.

[38] Demkowicz, L., Gopalakrishnan, J., and Schöberl, J. Polynomial extension operators. Part III. *Math. Comp. 81*, 279 (2012), 1289–1326.

[39] Destuynder, P., and Métivet, B. Explicit error bounds for a nonconforming finite element method. *SIAM J. Numer. Anal. 35*, 5 (1998), 2099–2115.

[40] Destuynder, P., and Métivet, B. Explicit error bounds in a conforming finite element method. *Math. Comp. 68*, 228 (1999), 1379–1396.

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

[42] Di Pietro, D. A., Vohralík, M., and Yousef, S. Adaptive regularization, linearization, and discretization and a posteriori error control for the two-phase Stefan problem. *Math. Comp. 84*, 291 (2015), 153–186.

[43] Dolejší, V. Semi-implicit interior penalty discontinuous Galerkin methods for viscous compressible flows. *Commun. Comput. Phys. 4*, 2 (2008), 231–274.

[44] Dolejší, V., Feistauer, M., and Felcman, J. On the discrete Friedrichs inequality for nonconforming finite elements. *Numer. Funct. Anal. Optim. 20*, 5-6 (1999), 437–447.

[45] Dörfler, W., and Ainsworth, M. Reliable a posteriori error control for nonconformal finite element approximation of Stokes flow. *Math. Comp. 74*, 252 (2005), 1599–1619.

[46] El Alaoui, L., Ern, A., and Vohralík, M. Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems. *Comput. Methods Appl. Mech. Engrg. 200*, 37-40 (2011), 2782–2795.

[47] Ern, A., and Guermond, J.-L. *Theory and practice of finite elements*, vol. 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[48] Ern, A., Nicaise, S., and Vohralík, M. An accurate $\mathbf{H}(\mathrm{div})$ flux reconstruction for discontinuous Galerkin approximations of elliptic problems. *C. R. Math. Acad. Sci. Paris 345*, 12 (2007), 709–712.

[49] Ern, A., Stephansen, A. F., and Vohralík, M. Guaranteed and robust discontinuous Galerkin a posteriori error estimates for convection-diffusion-reaction problems. *J. Comput. Appl. Math. 234*, 1 (2010), 114–130.

[50] Ern, A., and Vohralík, M. Flux reconstruction and a posteriori error estimation for discontinuous Galerkin methods on general nonmatching grids. *C. R. Math. Acad. Sci. Paris 347*, 7-8 (2009), 441–444.

[51] Ern, A., and Vohralík, M. A posteriori error estimation based on potential and flux reconstruction for the heat equation. *SIAM J. Numer. Anal. 48*, 1 (2010), 198–223.

[52] Ern, A., and Vohralík, M. A unified framework for a posteriori error estimation in elliptic and parabolic problems with application to finite volumes. In *Finite Volumes for Complex Applications VI* (Berlin, Heidelberg, 2011), J. Fořt, J. Fürst, J. Halama, R. Herbin, and F. Hubert, Eds., Springer-Verlag, pp. 821–837.

[53] Ern, A., and Vohralík, M. Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs. *SIAM J. Sci. Comput. 35*, 4 (2013), A1761–A1791.

[54] Ern, A., and Vohralík, M. Four closely related equilibrated flux reconstructions for nonconforming finite elements. *C. R. Math. Acad. Sci. Paris 351*, 1-2 (2013), 77–80.

[55] Ern, A., and Vohralík, M. Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations. *SIAM J. Numer. Anal.* (2015). Accepted for publication.

[56] Evans, L. C. *Partial differential equations*, vol. 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998.

[57] Eymard, R., Gallouët, T., and Herbin, R. Convergence of finite volume schemes for semilinear convection diffusion equations. *Numer. Math. 82*, 1 (1999), 91–116.

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

[59] Eymard, R., Gallouët, T., and Herbin, R. Finite volume approximation of elliptic problems and convergence of an approximate gradient. *Appl. Numer. Math. 37*, 1-2 (2001), 31–53.

[60] Girault, V., and Raviart, P.-A. *Finite element methods for Navier-Stokes equations*, vol. 5 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986. Theory and algorithms.

[61] Godlewski, E., and Le Dret, H. *Bases des méthodes numériques*. Université Pierre et Marie Curie, 2010. Lecture notes.

[62] Han, W. A posteriori error analysis for linearization of nonlinear elliptic problems and their discretizations. *Math. Methods Appl. Sci. 17*, 7 (1994), 487–508.

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

[64] Hilhorst, D., and Vohralík, M. A posteriori error estimates for combined finite volume–finite element discretizations of reactive transport equations on nonmatching grids. *Comput. Methods Appl. Mech. Engrg. 200*, 5-8 (2011), 597–613.

[65] Jiránek, P., Strakoš, Z., and Vohralík, M. A posteriori error estimates including algebraic error and stopping criteria for iterative solvers. *SIAM J. Sci. Comput. 32*, 3 (2010), 1567–1590.

[66] Karakashian, O. A., and Pascal, F. A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems. *SIAM J. Numer. Anal. 41*, 6 (2003), 2374–2399.

[67] Kim, K.-Y. A posteriori error analysis for locally conservative mixed methods. *Math. Comp. 76*, 257 (2007), 43–66.

[68] Kim, K.-Y. A posteriori error estimators for locally conservative methods of nonlinear elliptic problems. *Appl. Numer. Math. 57*, 9 (2007), 1065–1080.

[69] Knobloch, P. Uniform validity of discrete Friedrichs' inequality for general nonconforming finite element spaces. *Numer. Funct. Anal. Optim. 22*, 1-2 (2001), 107–126.

[70] Ladevèze, P. *Comparaison de modèles de milieux continus*. Ph.D. thesis, Université Pierre et Marie Curie (Paris 6), 1975.

[71] Laugesen, R. S., and Siudeja, B. A. Minimizing Neumann fundamental tones of triangles: an optimal Poincaré inequality. *J. Differential Equations 249*, 1 (2010), 118–135.

[72] Luce, R., and Wohlmuth, B. I. A local a posteriori error estimator based on equilibrated fluxes. *SIAM J. Numer. Anal. 42*, 4 (2004), 1394–1414.

[73] Makridakis, C., and Nochetto, R. H. Elliptic reconstruction and a posteriori error estimates for parabolic problems. *SIAM J. Numer. Anal. 41*, 4 (2003), 1585–1594.

[74] Mikhlin, S. G. *Variational methods in mathematical physics*. Translated by T. Boddington; editorial introduction by L. I. G. Chambers. A Pergamon Press Book. The Macmillan Co., New York, 1964.

[75] Neittaanmäki, P., and Repin, S. *Reliable methods for computer simulation*, vol. 33 of *Studies in Mathematics and its Applications*. Elsevier Science B.V., Amsterdam, 2004. Error control and a posteriori estimates.

[76] Nicaise, S. A posteriori error estimations of some cell-centered finite volume methods. *SIAM J. Numer. Anal. 43*, 4 (2005), 1481–1503.

[77] Patera, A. T., and Rønquist, E. M. A general output bound result: application to discretization and iteration error estimation and control. *Math. Models Methods Appl. Sci. 11*, 4 (2001), 685–712.

[78] Payne, L. E., and Weinberger, H. F. An optimal Poincaré inequality for convex domains. *Arch. Rational Mech. Anal. 5* (1960), 286–292.

[79] Picasso, M. Adaptive finite elements for a linear parabolic problem. *Comput. Methods Appl. Mech. Engrg. 167*, 3-4 (1998), 223–237.

[80] Picasso, M. A stopping criteria for the conjugate gradient algorithm in the framework of adaptive finite elements. *Comm. Numer. Meth. Eng. 25*, 4 (2009), 339–355.

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

[82] Prudhomme, S., Nobile, F., Chamoin, L., and Oden, J. T. Analysis of a subdomain-based error estimator for finite element approximations of elliptic problems. *Numer. Methods Partial Differential Equations 20*, 2 (2004), 165–192.

[83] Quarteroni, A., and Valli, A. *Numerical approximation of partial differential equations*, vol. 23 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1994.

[84] Raviart, P.-A., and Thomas, J.-M. *Introduction à l'analyse numérique des équations aux dérivées partielles*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1983.

[85] Repin, S. I. *A posteriori estimates for partial differential equations*, vol. 4 of *Radon Series on Computational and Applied Mathematics*. Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

[86] Repin, S. I. Computable majorants of constants in the Poincaré and Friedrichs inequalities. *J. Math. Sci. (N. Y.) 186*, 2 (2012), 307–321. Problems in mathematical analysis. No. 66.

[87] Roberts, J. E., and Thomas, J.-M. Mixed and hybrid methods. In *Handbook of Numerical Analysis, Vol. II*. North-Holland, Amsterdam, 1991, pp. 523–639.

[88] Šebestová, I., and Vejchodský, T. Two-sided bounds for eigenvalues of differential operators with applications to Friedrichs, Poincaré, trace, and similar constants. *SIAM J. Numer. Anal. 52*, 1 (2014), 308–329.

[89] Stephansen, A. F. *Méthodes de Galerkine discontinues et analyse d'erreur a posteriori pour les problèmes de diffusion hétérogène*. Ph.D. thesis, Ecole Nationale des Ponts et Chaussées, 2007.

[90] Stoyan, G., and Baran, Á. Crouzeix-Velte decompositions for higher-order finite elements. *Comput. Math. Appl. 51*, 6-7 (2006), 967–986.

[91] Thomas, J.-M. *Sur l'analyse numérique des méthodes d'éléments finis hybrides et mixtes*. Ph.D. dissertation, Université Pierre et Marie Curie (Paris 6), May 1977.

[92] Veeser, A., and Verfürth, R. Poincaré constants for finite element stars. *IMA J. Numer. Anal. 32*, 1 (2012), 30–47.

[93] Verfürth, R. *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. Teubner-Wiley, Stuttgart, 1996.

[94] Verfürth, R. A note on polynomial approximation in Sobolev spaces. *M2AN Math. Model. Numer. Anal. 33*, 4 (1999), 715–719.

[95] Verfürth, R. A posteriori error estimates for finite element discretizations of the heat equation. *Calcolo 40*, 3 (2003), 195–212.

[96] Verfürth, R. Robust a posteriori error estimates for stationary convection-diffusion equations. *SIAM J. Numer. Anal. 43*, 4 (2005), 1766–1782.

[97] Vohralík, M. On the discrete Poincaré–Friedrichs inequalities for nonconforming approximations of the Sobolev space H^1. *Numer. Funct. Anal. Optim. 26*, 7-8 (2005), 925–952.

[98] Vohralík, M. A posteriori error estimates for lowest-order mixed finite element discretizations of convection-diffusion-reaction equations. *SIAM J. Numer. Anal. 45*, 4 (2007), 1570–1599.

[99] Vohralík, M. A posteriori error estimation in the conforming finite element method based on its local conservativity and using local minimization. *C. R. Math. Acad. Sci. Paris 346*, 11-12 (2008), 687–690.

[100] Vohralík, M. Unified primal formulation-based a priori and a posteriori error analysis of mixed finite element methods. *Math. Comp. 79*, 272 (2010), 2001–2032.

[101] Vohralík, M. Guaranteed and fully robust a posteriori error estimates for conforming discretizations of diffusion problems with discontinuous coefficients. *J. Sci. Comput. 46*, 3 (2011), 397–438.

Index