FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Filip Roskovec

Goal-oriented a posteriori error estimates and adaptivity for the numerical solution of partial differential equations

Department of Numerical Mathematics

Supervisor of the doctoral thesis:	prof. RNDr. Vít Dolejší, Ph.D., DSc.
Study programme:	Mathematics
Study branch:	Scientific and Technical Calculations

Prague 2019

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date .... signature of the author

At the end of this work, I would like to thank all people who supported me during my studies.

First of all, I am very grateful to my supervisor Vít Dolejší, for the countless advices he gave me, the hundreds of hours he spent supporting my studies and for the patience he always had with me. Without his precious suggestions, many of the obstacles I faced during my research would not have been overcome.

I would like to express my thankfulness to all my colleagues and especially other PhD students at our department for many cheerful discussion focused not only on our research.

Moreover, a special thank goes to my whole family and friends for their unreserved support. Finally, I would like to thank you Radka, for always standing by my side helping me to overcome all what comes.

Title: Goal-oriented a posteriori error estimates and adaptivity for the numerical solution of partial differential equations

Author: Filip Roskovec

Department: Department of Numerical Mathematics

Supervisor: prof. RNDr. Vít Dolejší, Ph.D., DSc., Department of Numerical Mathematics

Abstract: A posteriori error estimation is an inseparable component of any reliable numerical method for solving partial differential equations. The aim of the goal-oriented a posteriori error estimates is to control the computational error directly with respect to some quantity of interest, which makes the method very convenient for many engineering applications. The resulting error estimates may be employed for mesh adaptation which enables to find a numerical approximation of the quantity of interest under some given tolerance in a very efficient manner. In this thesis, the goal-oriented error estimates are derived for discontinuous Galerkin discretizations of the linear scalar model problems, as well as of the Euler equations describing inviscid compressible flows. It focuses on several aspects of the goal-oriented error estimation method, in particular, higher order reconstructions, adjoint consistency of the discretizations, control of the algebraic errors arising from iterative solutions of both algebraic systems, and linking the estimates with the *hp*-anisotropic mesh adaptation. The computational performance is demonstrated by numerical experiments.

Keywords: a posteriori error estimates, discontinuous Galerkin method, Euler equations, goal-oriented error estimates, quantity of interest

Contents

Introduction

Partial differential equations (PDE) describe many phenomena of the world around us. Apart the classical physical problems of fluid dynamics, elasticity, sound, electrodynamics or quantum mechanics, many problems in more humanistic areas such as economics or social sciences can be modeled as PDEs.

Numerical solution of partial differential equations

Unfortunately, only a minor fragment of these equations can be solved analytically. Therefore, using numerical methods for approximation of the solution is inevitable. With the development of computers in the second half of the twentieth century new possibilities arose. The finite element method (FEM), finite volumes (FV) and discontinuous Galerkin method (DG) are the most popular methods using the variational formulation for solving boundary-value problems, see, e.g., Ciarlet [1979], Godlewski and Raviart [1996], Eymard et al. [2000], Di Pietro and Ern [2012], Dolejší and Feistauer [2015].

Unlike standard finite elements, the *discontinuous Galerkin method* does not require continuity between neighboring elements. That makes it more convenient for problems where even the exact solution possesses discontinuity, e.g., in computational fluid dynamics. The solution process starts with dividing the computational domain into finite number of elements (triangles or quadrilaterals in 2D, and tetrahedra or hexahedra in 3D) and the solution is approximated in the space of piecewise polynomial functions.

In order to provide trustworthy results, these numerical solution has to be complemented with a relevant estimate on the error. To estimate the error of the numerical solution, i.e., its distance from the exact solution, is seemingly impossible task since the true solution is not known (if it was reachable, the numerical solution would not have much use), so the distance $|u - u_h|$ (u denotes the exact solution and u_h its numerical approximation) has to be bounded only using u_h and the data of the solved problem. Many methods for computing the so-called *a posteriori error estimates* of numerical solution of PDEs were developed through last 50 years, see, e.g., Verfürth [1996], Ainsworth and Oden [2000], Neittaanmäki and Repin [2004], Verfürth [2013], Vohralík [2010] and the reference cited therein.

The a posteriori error estimates are used also for the automatic mesh adaptation. Therefore the estimates need to be divided into localized error indicators approximating the computational error on a small part of the domain (e.g., one element). These may be later used to drive the automatic mesh adaptation which offers a very powerful tool for the efficient, accurate and robust numerical algorithms for solving partial differential equations, Park et al. [2016].

Goal-oriented error estimates

Usually, methods for a posteriori error estimation measure the error of the approximate solution in a norm which typically arises from the mathematical formulation of the problem. On the other hand, in engineering application there may exist some *quantity of interest* (e.g., lift of an airplane wing), computation of which is the main goal of the simulation. Therefore, the so-called *goal-oriented error estimating* techniques focusing on the error measured with respect to this quantity of interest exhibit a powerful tool for the numerical approximation of problems where we are not interested in the approximate solution itself but rather in a certain solution-dependent target quantity. The quantity of interest is usually represented by a (linear) functional $J(u)$, where u is the exact solution of the given problem. Further, the error between the exact value $J(u)$ and its approximation $J(u_h)$ is estimated.

In order to connect the target functional with the solved problem, the so-called adjoint problem (corresponding to the given equation) is used. It contains the adjoint operator to the differential operator of the original (primal) problem and the target functional on the right-hand side. The adjoint (or dual in other literature) problem is used for numerical analysis since 1970s for a priori analysis of the error (e.g., the well-known Aubin-Nitsche trick, Brenner and Scott [1994]), in optimization or when post-processing a numerical solution, see e.g., Giles and Süli [2002].

We refer to Becker and Rannacher [1995, 1996] where the so-called *dual weighted residual* (DWR) estimates dealing with this subject were introduced. For the first time, this method was formulated for finite element method applied to linear elliptic partial differential equations. Since then it was intensively explored and developed for many other problems. In Bötcher and Rannacher [1996] application of the DWR method on ordinary differential equations is compared to other a posteriori estimates for one-step numerical methods. In Rannacher and Suttmeier [1997] the DWR method is introduced for linear elasticity. We recommend latter publications, e.g., Bangerth and Rannacher [2003], Becker and Rannacher [2001], Giles and Süli [2002], where the method is thoroughly studied.

The DWR method uses the adjoint problem in a specific way to weight local residua of the numerical solution. In other words, it measures how sensitive the target functional is on a potential error localized in certain area (e.g., one mesh element). It was demonstrated in many numerical experiments, see e.g., Bangerth and Rannacher [2003], Becker and Rannacher [2001], that this weighting of the local residua enables obtaining approximations of $J(u)$ much more efficiently compared to other methods.

In its theoretical formulation the method relies on computing the exact solution z of the adjoint problem. Unfortunately, in most cases this is equivalently difficult (or impossible) as obtaining the true solution of the original problem. Therefore, in practical computations the goal oriented error estimates require a sufficiently accurate approximation of the solution of the adjoint problem. One possibility is to solve the adjoint problem on a globally refined mesh and/or with polynomial approximation of higher degree, see, e.g., Šolín and Demkowicz [2004], Hartmann and Houston [2006a,b], Harriman et al. [2003]. This method usually gives quite reliable results, but it is computationally very time-consuming, since the computational demanding of the discrete adjoint problem exceeds the primal one. Alternatively, the adjoint solution may be computed in the same discrete function space as the primal one and then reconstructed locally and hence much faster.

Discretization of the primal and adjoint problems leads to two linear algebraic systems which are usually solved by a suitable iterative technique. Therefore the resulting discrete solutions (and also their error estimates) are influenced by the algebraic error resulting from the inexact solution of both algebraic systems (primal and adjoint). The goal-oriented estimates can be naturally extended by estimates measuring the impact of these algebraic errors on the error of the quantity of interest, see, e.g., Meidner et al.

[2009], Rannacher and Vihharev [2013], Arioli et al. [2013].

The DWR method can be also generalized for nonlinear PDEs. Its derivation for a general nonlinear variational problem was developed in Becker and Rannacher [2001]. The method, introduced via the Euler-Lagrange method of constrained optimization, may be employed for many nonlinear differential equations. In Heuveline and Rannacher [2001] it is applied to a generalized eigenvalue problem, in Becker et al. [1998] for optimization problems, in Bangerth and Rannacher [1999] for hyperbolic problems and in Hartmann and Houston [2006a] for compressible Navier-Stokes equations.

For nonlinear problems the adjoint problem is introduced for the linearized original problem and hence it is always linear. Therefore, its solution is usually much faster than the solution of the original problem. Typically the approximation of the adjoint solution corresponds to one iteration of the Newton method used for primal problem, see, e.g., Hartmann and Houston [2006a].

The application of DWR method for DG discretizations is described in detail, e.g., in Kanschat and Rannacher [2002] for elliptic partial differential equations, in Hartmann and Houston [2002] for Euler equations and in Hartmann and Houston [2006b], Hartmann and Houston [2006a] for Navier-Stokes equations. Since the DG method belongs to the family of the nonconforming methods, the suitability of the discretization for the adjoint problem is more delicate (compared to standard FEM). The so-called *adjoint consistency* of the discretization, i.e., whether the solution of the adjoint problem satisfies the discrete adjoint problem, has to be verified. The importance of this property was firstly pointed out in Harriman et al. [2004] and then further explored in, e.g., Hartmann [2006, 2007].

Adaptive refinement

The aim of the goal-oriented mesh adaptation is, based on a posteriori error estimates, to reduce the error of the quantity of interest under some given tolerance using as small number of degrees of freedom (DoF) as possible. Starting with an initial coarse mesh, sizes of the mesh elements are iteratively updated according to the goal-oriented error indicators.

Moreover, even the element shapes may be optimized during the adaptation process. The so-called anisotropic adaptation method which generates anisotropic meshes consisting of possibly thin and long triangular elements, has been exhibited to be a very efficient tool for the reduction of the number of degrees of freedom. Its use is highly advantageous for many applications (namely in fluid dynamics), as was demonstrated, see, e.g., Belme et al. [2012], Ait-Ali-Yahia et al. [2002], Aguilar and Goodman [2006], Frey and Alauzet [2005], Habashi et al. [2000], Dompierre et al. [2002], Dolejší and Felcman [2002], Simpson [1994], Breuss et al. [2006], Dolejší [1998] and the references mentioned therein.

Most of the anisotropic mesh adaptation techniques mentioned above are based on the interpolation error estimates of the numerical solution. To use it for goal-oriented error estimates, it is necessary to link the anisotropy of the elements with respect to a solution-dependent target functional, we refer to Balan et al. [2016], Loseille et al. [2010], Venditti and Darmofal [2002], Dolejší et al. [2017], for a survey and visions of goal-oriented mesh adaptations, see Fidkowski and Darmofal [2011], Park et al. [2016]. However, these techniques are mostly heuristic. Rigorous goal-oriented error estimates taking into account the anisotropy (= size, aspect ratio and orientation) of mesh elements were derived in Alauzet et al. [2009], Formaggia and Perotto [2003] for purely diffusive problems and in Carpio et al. [2013] for a convection-diffusion-reaction equation. However, these results deal only with a piecewise linear approximation.

Further, the anisotropic adaptation may be coupled with the *hp*-adaptation method, which enables the polynomial approximation degree to vary among the mesh elements. The *hp*-finite element methods have a prominent place among other adaptive methods since, under some assumptions, they give an exponential rate of convergence, cf. Gui and Babuška [1986], Babuška et al. [1997], Babuška and Strouboulis [2001], Schwab [1998], Demkowicz [2007]. Therefore, the use of *hp*-anisotropic meshes, with possibly high polynomial approximation degrees, offers enough of flexibility to an efficient and accurate numerical approximation of PDEs. This was also identified as one of the key enabling technologies in the Workshop on High-Order CFD Methods, Wang et al. [2013].

Novelty of the results presented in the thesis

This thesis contributes to the research dealing with goal-oriented error estimates and anisotropic *hp*-adaptation for linear and nonlinear partial differential equations. The novelty of the results can be summarized in the following way.

- *error estimates including algebraic error* – in Meidner et al. [2009], there was presented an adaptive technique technique which controls that the algebraic errors are bounded by the discretization ones. However, it is not clear if this technique is optimal in the sense that an "over-solution" is performed. We study these aspects in Chapter 2.

- *higher order reconstruction technique* – we develop a new higher order reconstruction technique for arbitrary polynomial approximation degrees on unstructured grids. Numerical experiments show that the reconstruction is stable even for anisotropic meshes.

- *anisotropic hp-mesh adaptation* – combination of *hp*-method and anisotropic mesh adaptation in the framework of the goal-oriented error estimates is a completely original work. A rigorous analysis was performed for a scalar linear problem and then extended to nonlinear problems including compressible Euler equations. We develop a little different approach for the compressible Euler equations in comparison to Hartmann and Houston [2006b], Hartmann [2007].

Outline

The content of the thesis is divided into four chapters.

In Chapter 1 the goal-oriented estimation method is introduced for prototypical linear and nonlinear problems. The important aspects, advantages and drawbacks are discussed. Finally, these are illustrated by several exemplary numerical experiments.

In Chapter 2 we introduce the DG discretization of convection-diffusion-reaction equation. The computational error is estimated in the framework of the DWR method. Further, we focus on the control of the algebraic errors arising from iterative solutions of algebraic systems corresponding to both primal and adjoint problems. Moreover, we present two different reconstruction techniques allowing an efficient evaluation of the error estimators. We propose a complex algorithm which controls discretization and algebraic errors and it drives the adaptation of the mesh.

In Chapter 3 we introduce the goal-oriented error indicators enabling anisotropic adaptation of the mesh, which provide upper bounds to the estimates derived in previous chapter. The concept of anisotropy of triangles and polynomial functions is presented. Further, the size, shape and polynomial degree optimization of each elements are discussed. These estimates are based on the concepts of anisotropic adaptation from Dolejší [2014], Dolejší [2015] which were later generalized for goal-oriented error estimation in Dolejší et al. [2018], Dolejší et al. [2019], Bartoš et al. [2019]. Finally, we develop a *hp*-variant of the goal-oriented anisotropic mesh adaptation algorithm. Numerical experiments, presented at the end of the chapter, demonstrate the efficiency of the proposed algorithm.

In Chapter 4 we introduce the DG discretization for the Euler equations describing the flow of an inviscid compressible fluid. The goal-oriented error estimates are introduced for the drag, lift and momentum coefficients, which represent the most significant quantities of interest for such problem. Special attention is paid to the adjoint consistency of the discretization of the adjoint problem. Furthermore, $hp-$anisotropic error indicators based on the approach from Chapter 3 are presented. Several numerical experiments document the performance of the method.

We note that some parts of this text were already published in several articles. Chapter 2 follows, only with minor modifications, the article Dolejší and Roskovec [2017]. The content of the Chapter 3 is a unification of the results published in Dolejší et al. [2019] and Bartoš et al. [2019].

1. Goal-oriented error estimates for model problems

In this chapter we introduce the basic ideas of the goal-oriented error estimates. We follow the Dual Weighted Residual method (DWR) as originally developed in Becker and Rannacher [1996] and further described in Becker and Rannacher [2001] and Bangerth and Rannacher [2003].

1.1 DWR method for linear problems

As a model problem for the linear case we consider the Poisson equation on a polygonal ($d = 2$) or polyhedral ($d = 3$) domain $\Omega \in \mathbb{R}^d$ with homogeneous Dirichlet conditions

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

where f is some given function.

Except the solution of the given (primal) problem (1.1), the goal-oriented error estimates require to solve and additional so-called adjoint (or dual) problem. If we denote the operator $\mathscr{L} := -\Delta$ in (1.1) the adjoint problem in its strong form reads

$$\mathscr{L}^* z = J, \tag{1.2}$$

where $\mathscr{L}^* = -\Delta$ is the adjoint operator to $\mathscr{L}$ and J is the target functional representing the quantity of interest.

1.1.1 Weak formulation and discretization of the primal problem

Let us start with introducing the following notation. For a domain $D \subset \mathbb{R}^d$, $L^2(D)$ is the Lebesgue space of square-integrable functions on D. This space is equipped with the norm defined by

$$\|v\|_D = \left(\int_D |v|^2 \, \mathrm{d}x \right)^{1/2}$$

and scalar product

$$(v, w)_D := \int_D vw \, \mathrm{d}x.$$

We introduce the *multi-index* notation. A multi-index α is an n-tuple of non-negative integers and its length is given by $|\alpha| = \sum_{j=1}^{d} \alpha_j$. Further, let $k \geq 1$, we define the Sobolev spaces

$$H^k(D) := \{ v \in L^2(D); D^\alpha v \in L^2(D) \; \forall \alpha, |\alpha| \leq k \},$$

where $D^\alpha v$ denotes the weak partial derivative of v given by

$$D^\alpha v = \left(\frac{\partial}{\partial x_1} \right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_d} \right)^{\alpha_d} v.$$

For $v \in H^k(D)$ we define the Sobolev-norm and seminorm, respectively

$$\|v\|_{H^k(D)} := \left(\sum_{|\alpha| \le k} \|D_\alpha v\|^2_{L^2(D)} \right)^{1/2}, \qquad |v|_{H^k(D)} := \left(\sum_{|\alpha| = k} \|D_\alpha v\|^2_{L^2(D)} \right)^{1/2}.$$

Further, we define the Sobolev space

$$H^1_0(D) := \{v \in H^1(D); w\big|_{\partial D} = 0\}.$$

Due to the Poincaré inequality

$$\|v\|_{L^2(D)} \le c\, \|\nabla v\|_{L^2(D)}\,, \, v \in H^1_0(D),$$

hence the H^1-seminorm is a norm on $H^1_0(D)$. If $D = \Omega$, we usually omit the subscript Ω in the notation of norms or scalar products.

Definition 1.1. *Let $f \in L^2(\Omega)$ and $V = H^1_0(\Omega)$, we say that a function $u \in V$ is a weak solution of problem* (1.1)*, if*

$$a(u, \varphi) = (f, \varphi)_\Omega \qquad \forall \varphi \in V, \tag{1.3}$$

where

$$a(u, \varphi) = \int_\Omega \nabla u \cdot \nabla \varphi \,\mathrm{d}x. \tag{1.4}$$

The existence and uniqueness of the weak solution u can be proven using the Lax-Milgram Lemma, see Brenner and Scott [1994][Theorem 2.7.7].

Let us denote $\mathscr{T}_h$ a partition covering $\overline{\Omega}$ consisting of finite number of closed d-dimensional simplices K with mutually disjoint interiors. The boundary of the element $K \in \mathscr{T}_h$ will be denoted by ∂K, its diameter $h_K = \mathrm{diam}(K)$ and $|K|$ its d-dimensional Lebesgue measure.

Let $P^p(K)$ denote the space of polynomial functions up to degree p with support on element K. We introduce the notation V_h, $h > 0$ for a general finite element space of functions containing piecewise polynomial functions. For conforming finite elements V_h is given by

$$V^c_h = \{v \in C(\overline{\Omega}); v\big|_K \in P^p(K)\, \forall K \in \mathscr{T}_h\}, \tag{1.5}$$

while for the discontinuous Galerkin method (DG) no continuity across mesh elements is required, i.e.,

$$V^d_h = \{v \in L^2(\Omega); v\big|_K \in P^p(K)\, \forall K \in \mathscr{T}_h\}, \tag{1.6}$$

Further we define a function space W_h such that both $u \in W_h$ and $V_h \subset W_h$. Since $V_h \subset V$ for the conforming finite element method we may simply put $W_h = V$. However, for nonconforming methods, where $V_h \not\subset V$, the choice of W_h is not trivial. For the discontinuous Galerkin method we set

$$W_h := H^2(\Omega, \mathscr{T}_h) = \{v \in L^2(\Omega); v\big|_K \in H^2(K) \forall K \in \mathscr{T}_h\}. \tag{1.7}$$

Definition 1.2. *We say that $u_h \in V_h$ is the discrete solution of the primal problem* (1.3) *if it satisfies*

$$a_h(u_h, \varphi_h) = \ell_h(\varphi_h) \qquad \forall \varphi_h \in V_h, \tag{1.8}$$

where $a_h(\cdot,\cdot)$ and $\ell_h(\cdot)$ denote the discrete forms coming from the discretization of problem (1.3).

We note that $a_h = a$, $\ell_h(\cdot) = (f, \cdot)_\Omega$ for conforming FEM and the detailed definition of a_h and ℓ_h will be introduced in (2.6). We assume that the problem (1.8) is well-posed, for more detailed analysis of the assumptions on the data and properties of the solution u_h see e.g., Ciarlet [1979], Brenner and Scott [1994] (for conforming FEM) and e.g., Dolejší and Feistauer [2015] (for DG).

Further, we assume that the discrete problem (1.8) is *consistent*, i.e. the weak solution $u \in V$ of (1.3) satisfies also

$$a_h(u, \varphi) = \ell_h(\varphi) \qquad \forall \varphi \in W_h. \tag{1.9}$$

This implies the so-called *Galerkin orthogonality* of the error $e_h = u - u_h$ of the primal problem

$$a_h(e_h, \varphi_h) = 0 \qquad \forall \varphi_h \in V_h. \tag{1.10}$$

Finally, we define the residual of the discrete primal problem by

$$r_h(u_h)(\varphi) = \ell_h(\varphi) - a_h(u_h, \varphi). \tag{1.11}$$

1.1.2 Quantity of interest

In many applications the solution itself may not be as important as some feature reliant on it. Then the main goal of the computation is to provide a sufficiently accurate approximation of the quantity of interest rather than the solution itself. In this case it is advantageous if also the a posteriori control of the error can be considered with respect to the target functional $J \in V'$, where the value $J(u)$ represents the given quantity of interest.

Typically J is given as a volume integral over some $\Omega_J \subset \Omega$ or a boundary integral over some part of the boundary $\Gamma_J \subset \partial\Omega$. We present a few possible examples:

- point value $u(a)$ for an $a \in \Omega$: since point values are not well defined for Sobolev functions in V these have to be approximated by regularized point value $J(u) = \frac{1}{|B_\varepsilon(a)|}\int_{B_\varepsilon(a)} u \,\mathrm{d}x$, where $B_\varepsilon(a)$ is a ball with center at point $a \in \Omega$ and diameter $\varepsilon \ll 1$, for more detailed analysis see Kanschat and Rannacher [2002].

- mean normal flux: $J(u) = \frac{1}{|E|}\int_E \boldsymbol{n} \cdot \nabla u \,\mathrm{d}S$, where $\boldsymbol{n}$ denotes the unit outer normal of $\partial\Omega$ and $E \subset \partial\Omega$. We refer to Section 1.1.9 where the setting of the boundary conditions for the adjoint problem for boundary target functionals is explained.

- energy norm error $\|\nabla(u - u_h)\|$: this choice of target quantity may be used, see Remark at the end of Section 1.1.4, to show the equivalence between DWR method with
$$J(\varphi) = \frac{1}{\|\nabla(u - u_h)\|}\int_\Omega \nabla\varphi \cdot \nabla(u - u_h)\,\mathrm{d}x$$
and the well-known energy-norm a posteriori error estimates, see, e.g., Brenner and Scott [1994].

1.1.3 Adjoint problem and abstract error identities

The goal-oriented error estimates are based on the relation between the residual (1.11) of the primal problem and the error with respect to the target functional $J(u) - J(u_h)$. We follow the line of reasoning from Šolín and Demkowicz [2004].

Let $J \in W_h'$ and $r_h(u_h) \in W_h'$ then we may express this relationship between $r_h(u_h)(\cdot)$ and $J(\cdot)$ by defining $G \in W_h''$ such that

$$G(r_h(u_h)) = J(e_h).$$

Since the function space W_h is a Hilbert space, it is reflexive and the functional G can be related to some function $\tilde{z} \in W_h$. We may write

$$G(r_h(u_h)) = r_h(u_h)(\tilde{z}). \tag{1.12}$$

Further, using the properties of residual we obtain

$$\begin{aligned} r_h(u_h)(\tilde{z}) &= \ell_h(\tilde{z}) - a_h(u_h, \tilde{z}) \\ &= a_h(u, \tilde{z}) - a_h(u_h, \tilde{z}) = a_h(e_h, \tilde{z}). \end{aligned} \tag{1.13}$$

Now, it is obvious that the function $\tilde{z} \in W_h$ can be defined as the solution of

$$a_h(\psi, \tilde{z}) = J(\psi) \qquad \forall \psi \in W_h. \tag{1.14}$$

Notice that the problem (1.14) was defined employing the discrete bilinear form a_h. However, if we add the assumption that the numerical scheme is *adjoint consistent*, i.e.

$$a_h(\psi, z) = J(\psi) \qquad \forall \psi \in W_h, \tag{1.15}$$

where z is the solution of the adjoint problem (1.2), then $\tilde{z} = z$.

Now, employing (1.13) and (1.11) and the Galerkin orthogonality (1.10) we obtain the *primal abstract error identity* for the error of the target quantity

$$J(e_h) = a_h(e_h, z - \varphi_h) = r_h(u_h)(z - \varphi_h) \qquad \forall \varphi_h \in V_h. \tag{1.16}$$

Moreover, we may introduce the discrete adjoint problem.

Definition 1.3. *We say that $z_h \in V_h$ is the discrete adjoint solution if it satisfies*

$$a(\psi_h, z_h) = J(\psi_h) \qquad \forall \psi_h \in V_h. \tag{1.17}$$

Similarly as in the primal case, the adjoint consistency (1.15) implies the Galerkin orthogonality of the error of the adjoint problem, i.e., denoting $e_h^* = z - z_h$

$$a(\psi_h, e_h^*) = 0 \qquad \forall \psi_h \in V_h. \tag{1.18}$$

Finally, introducing the residual of the adjoint problem by

$$r_h^*(z_h)(\psi) := J(\psi) - a_h(\psi, z_h). \tag{1.19}$$

we get also the *adjoint abstract error identity*

$$\begin{aligned} J(e_h) &= a(e_h, z) = a(u - u_h, z - z_h) \\ &= a(u - \psi_h, z - z_h) = J(u - \psi_h) - a(u - \psi_h, z_h) \\ &= r_h^*(z_h)(u - \psi_h) \qquad \forall \psi_h \in V_h. \end{aligned} \tag{1.20}$$

Further, there holds the following error equivalence

$$J(u - u_h) = J(u) - J(u_h) = \ell_h(z) - \ell_h(z_h) = \ell_h(z - z_h), \tag{1.21}$$

since due to (1.9), (1.15) and (1.8), (1.17) we have that

$$\ell_h(z) = a_h(u, z) = J(u), \qquad \ell_h(z_h) = a_h(u_h, z_h) = J(u_h). \tag{1.22}$$

Therefore, the difference $\ell_h(z - u_h)$ expresses the error of the quantity of interest as well.

Due to (1.16) and (1.20) we see that the error of the target functional can be equivalently expressed by the primal and adjoint residuals, respectively. We may even take the arithmetic average of (1.16) and (1.20)

$$J(u) - J(u_h) = \frac{1}{2}\left(r_h(u_h)(z - \varphi_h) + r_h^*(z_h)(u - \psi_h)\right), \qquad \varphi_h, \psi_h \in V_h. \tag{1.23}$$

Remark. While trivial for conforming finite elements (since $a_h = a$), the adjoint consistency (1.15) is crucial for nonconforming discretizations such as the discontinuous Galerkin method. For an adjoint inconsistent discretization (e.g., the nonsymetric interior penalty Galerkin (NIPG) method, see Dolejší and Feistauer [2015]) the function $\tilde{z} \in W_h$ in (1.14) differs from z and it is not smooth on the edges of the mesh $\mathscr{T}_h$, which causes a slower convergence rate of the error $J(e_h)$. In Experiment 1.3.3 an example of such nonsmooth adjoint solution (of a problem with a known solution) is shown.

1.1.4 Classical dual weighted residual method

Here, we follow the DWR approach as was introduced in Bangerth and Rannacher [2003]. It should help to clear up how the method got its name “Dual Weighted Residual”.

We assume that the problem (1.1) was discretized by the conforming finite element method, i.e., $a_h(\cdot,\cdot) = a(\cdot,\cdot)$, $\ell_h(\cdot) = (f,\cdot)_\Omega$ and $V_h = V_h^c$ is the space of continuous piecewise polynomial functions. Integrating the error identity (1.16) by parts on each mesh element we obtain

$$\begin{aligned} r_h(u_h)(z - \varphi_h) &= \sum_{K \in \mathscr{T}_h} \left((f + \Delta u_h, z - \varphi_h)_K - (\nabla u_h \cdot \boldsymbol{n}, z - \varphi_h)_{\partial K}\right) \\ &= \sum_{K \in \mathscr{T}_h} \left((f + \Delta u_h, z - \varphi_h)_K + \frac{1}{2}([\![\nabla u_h \cdot \boldsymbol{n}]\!], z - \varphi_h)_{\partial K \setminus \partial \Omega}\right) \end{aligned} \tag{1.24}$$

where $[\![\nabla u_h \cdot \boldsymbol{n}]\!]$ denotes the jump of $\nabla u_h \cdot \boldsymbol{n}$ across an element edge, i.e., for two neighboring elements $K, K' \in \mathscr{T}_h$ with common edge Γ and unit normal $\boldsymbol{n}$ pointing from K to K', we define $[\![\nabla u_h \cdot \boldsymbol{n}]\!] := (\nabla u_h\big|_{K' \cap \Gamma} - \nabla u_h\big|_{K \cap \Gamma}) \cdot \boldsymbol{n}$.

Then we obtain the a posteriori error representation of the error

$$J(e_h) = \sum_{K\in\mathscr{T}_h} \big((R_h, z-\varphi_h)_K + (r_h, z-\varphi_h)_{\partial K\backslash\partial\Omega}\big) \qquad \forall\varphi_h \in V_h, \tag{1.25}$$

where

$$R_h\big|_K := f + \Delta u_h, \qquad r_h\big|_\Gamma := \begin{cases} \frac{1}{2}[\![\nabla u_h \cdot n]\!] & \text{for } \Gamma \subset \partial K\backslash\partial\Omega, \\ 0 & \text{for } \Gamma \subset \partial\Omega. \end{cases} \tag{1.26}$$

Then using the triangle inequality and Cauchy-Schwarz inequality we get

$$|J(e_h)| \le \sum_{K\in\mathscr{T}_h} |(R_h, z-\varphi_h)_K + (r_h, z-\varphi_h)_{\partial K}| \tag{1.27a}$$

$$\le \sum_{K\in\mathscr{T}_h} (\|R_h\|_K \|z-\varphi_h\|_K + \|r_h\|_{\partial K} \|z-\varphi_h\|_{\partial K}) \qquad \forall\varphi_h \in V_h, \tag{1.27b}$$

Finally, due to the Hölder inequality we obtain the following a posteriori error estimate of the error with respect to the target functional

$$|J(u-u_h)| \le \sum_{K\in\mathscr{T}_h} \rho_K \omega_K, \tag{1.28}$$

with the element residuals ρ_K and weights ω_K given by

$$\rho_K := \left(\|R_h\|_K^2 + h_K^{-1} \|r_h\|_{\partial K}^2\right)^{1/2}, \tag{1.29a}$$

$$\omega_K := \left(\|z-\varphi_h\|_K^2 + h_K \|z-\varphi_h\|_{\partial K}^2\right)^{1/2}, \tag{1.29b}$$

for an arbitrary $\varphi_h \in V_h$.

Unlike in (1.16) the estimate (1.28) is not independent from the choice of φ_h and hence the choice of φ_h strongly influences the tightness of the error estimate (1.28). For FEM we may choose $\varphi_h := I_h z$, where $I_h : V \to V_h$ is the Lagrange interpolation, see Brenner and Scott [1994].

Moreover, every step of (1.27a), (1.27b), (1.28) possibly increases the gap between $J(e_h)$ and the error estimate. Even though it is a bit counter intuitive and very unusual when thinking about adaptive solution of PDEs, even a very poor approximate solution u_h may lead to $J(e_h) = 0$ since the target functional J does not possess the standard additive property of norms and hence the individual summands of $J(e_h)$ in (1.25) may change signs. We call this property the global orthogonality of the error components. This property is lost when the triangle inequality is used in (1.27a).

Further artificial growth in the estimate may appear due to the application of the Cauchy-Schwarz inequality in (1.27b) since some of the summands belonging to an element $K \in \mathscr{T}_h$ may be close to orthogonal and then

$$(R_h, z-\varphi_h)_K + (r_h, z-\varphi_h)_{\partial K} \ll \|R_h\|_K \|z-\varphi_h\|_K + \|r_h\|_{\partial K} \|z-\varphi_h\|_{\partial K}.$$

Remark. In order to show that the standard energy-norm error estimates, see, e.g., Babuška and Rheinboldt [1978], Brenner and Scott [1994], can be derived also in the goal-oriented setting, we choose the (bit artificial) target functional

$$J(\varphi) = \frac{1}{\|\nabla e_h\|}(\nabla\varphi, \nabla e_h)_\Omega,$$

where $e_h = u - u_h$ is considered as a fixed quantity.

Then the corresponding adjoint problem for conforming FEM reads: find $z \in V$ such that

$$(\nabla\psi, \nabla z)_\Omega = \frac{(\nabla\psi, \nabla e_h)_\Omega}{\|\nabla e_h\|} \qquad \forall\psi \in V. \tag{1.30}$$

It is clear that the adjoint solution is given by $z = \frac{e_h}{\|e_h\|}$. Then from the estimate (1.28) we get

$$|J(e_h)| = \|\nabla e_h\| \leq \sum_{K\in\mathscr{T}_h} \rho_K \omega_K \leq \left(\sum_{K\in\mathscr{T}_h} h_K^2 \rho_K^2\right)^{1/2} \left(\sum_{K\in\mathscr{T}_h} h_K^{-2} \omega_K^2\right)^{1/2}. \tag{1.31}$$

Employing the standard interpolation error estimate, see, e.g., Brenner and Scott [1994],

$$\inf_{\varphi_h \in V_h} \left(\sum_{K\in\mathscr{T}_h} h_K^{-2} \|z - \varphi_h\|_K^2 + h_K^{-1} \|z - \varphi_h\|_{\partial K}^2\right)^{1/2} \leq C_I \|\nabla z\|, \tag{1.32}$$

in (1.31) we obtain the following estimate

$$\|\nabla e_h\| \leq \left(\sum_{K\in\mathscr{T}_h} h_K^2 \rho_K^2\right)^{1/2} C_I \|\nabla z\| \leq C_I \left(\sum_{K\in\mathscr{T}_h} h_K^2 \rho_K^2\right)^{1/2} \tag{1.33}$$

since evidently $\|\nabla z\| \leq 1$. Finally, recalling the definition of ρ_K in (1.29a) we come to the well-known energy-norm error estimate, for details of this method see, e.g., [Brenner and Scott, 1994, Section 9.2],

$$\|\nabla e_h\| \leq C_I \left(\sum_{K\in\mathscr{T}_h} h_K^2 \|f + \Delta u_h\|_K^2 + \frac{h_K}{2} \|[\![\nabla u_h \cdot n]\!]\|_{\partial K \setminus \partial\Omega}^2\right)^{1/2}. \tag{1.34}$$

1.1.5 Computable error estimates

In previous section we have presented the error identities (1.16), (1.20), (1.23) which can equivalently express the error of the target functional $J(e_h)$. Unfortunately, these are only of theoretical use since they contain unknown functions u and z. In order to obtain a computable error estimates these functions have to be approximated.

Omitting the very few cases when z can be found exactly or estimated a priori (such as in Remark 1.1.4), these have to be computed by some numerical method.

It is evident that the approximation $z \approx z_h$ is not satisfactory since due to the Galerkin orthogonality (1.10) $r_h(u_h)(z_h) = 0$. Hence the function z has to be approximated by some z_h^+ in $V_h^{(+)}$, where $V_h^{(+)}$ need to be a space richer than V_h (and similarly $u \approx u_h^+$, where $u_h^+ \in V_h^{(+)}$).

The error identities (1.16) and (1.20) can be rewritten as

$$J(e_h) = r_h(u_h)(z - \varphi_h) = r_h(u_h)(z - z_h^+) + r_h(u_h)(z_h^+ - \varphi_h), \tag{1.35a}$$

$$J(e_h) = r_h^*(z_h)(u - \psi_h) = r_h^*(z_h)(u - u_h^+) + r_h^*(z_h)(u_h^+ - \psi_h). \tag{1.35b}$$

Usually the terms $r_h(u_h)(z - z_h^+)$, $r_h^*(z_h)(u - u_h^+)$ are neglected and we choose $\varphi_h := \Pi z_h^+$ and $\psi_h := \Pi u_h^+$, where $\Pi : W_h \to V_h$ is some interpolation (for conforming FEM) or projection (for DG) operator satisfying

$$\|\varphi - \Pi\varphi\| \leq Ch^{p+1} \|\varphi\|_{H^{p+1}(\Omega)} \qquad \forall \varphi \in H^{p+1}(\Omega). \tag{1.36}$$

Further, we introduce the primal and dual error estimators

$$\eta_{\mathrm{S}} := r_h(u_h)(z_h^+ - \Pi z_h^+), \qquad \eta_{\mathrm{S}}^* := r_h^*(z_h)(u_h^+ - \Pi u_h^+). \tag{1.37}$$

Then we obtain the computable estimates

$$J(e_h) \approx \eta_{\mathrm{S}}, \qquad J(e_h) \approx \eta_{\mathrm{S}}^* \qquad J(e_h) \approx \frac{1}{2}\left(\eta_{\mathrm{S}} + \eta_{\mathrm{S}}^*\right). \tag{1.38}$$

Any of these can be used to approximate the error of the target quantity and further also to derive error indicators for adaptive mesh refinement, see Section 1.1.6.

Basically, there are two options how z_h^+ (or u_h^+, respectively) can be obtained – the discrete adjoint problem (1.17) can be either directly solved in $V_h^{(+)}$ (approximation by a higher order method) or this approximation may be obtained using some local post-processing of z_h given by $z_h^+ := \mathscr{R}(z_h)$ (approximation by a higher-order reconstruction).

The first approach is used in Šolín and Demkowicz [2004], where the so-called reference primal and adjoint solutions u_{ref}, z_{ref} are computed, on a globally refined mesh with $h/2$ and with one degree higher polynomial approximation, i.e., $u_{ref}, z_{ref} \in V_{h/2}^{p+1}$. It leads to very precise results since the reference solution u_{ref}, z_{ref} approximate the unknown solutions very accurately, c.f. Mitchell and McClain [2014] where several adaptive methods for solving elliptic PDEs are compared. On the other hand, the solution of the algebraic systems corresponding to the reference solutions is much more time-consuming compared to the original problems for u_h and z_h. Quite similar approach which computes the adjoint problem with higher polynomial degree (but on the same mesh) is used in Hartmann and Houston [2006b].

The main advantage of the higher-order reconstruction technique is that it does not significantly increase the computational costs, since it is computed locally for each $K \in \mathscr{T}_h$ and hence can be easily performed in parallel. Moreover, the algebraic systems corresponding to the discrete primal and adjoint problems utilize the same matrix. More precisely, we obtain

$$\mathbb{A}\boldsymbol{u}_h = \boldsymbol{b} \quad \text{and} \quad \mathbb{A}^{\mathrm{T}}\boldsymbol{z}_h = \boldsymbol{c}, \tag{1.39}$$

where N_h denotes the dimension of V_h, $\mathbb{A} \in \mathbb{R}^{N_h \times N_h}$ is the matrix coming from the discretization of (1.1), $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^{N_h}$ correspond to the right-hand sides of (1.8) and (1.17), respectively, and $\boldsymbol{u}_h, \boldsymbol{z}_h \in \mathbb{R}^{N_h}$ denote the algebraic representations of the discrete solution u_h and z_h, respectively. That can be beneficial in practical computations, see Dolejší and Tichý [2019], where both of these systems are solved simultaneously by the BiCG iterative method. On the other hand, even though such reconstruction may preform well in numerical experiments, it is usually difficult to theoretically provide guaranteed a priori proof that the reconstructed functions z_h^+ and u_h^+, respectively, should have better approximation properties than the original discrete solution z_h.

For conforming FEM mostly reconstructions based on some patch-wise higher-order interpolation are used. We mention the pioneering work Zienkiewicz and Zhu

[1992a,b] and later works Meidner et al. [2009], Richter and Wick [2015], Rannacher and Vihharev [2013] where those reconstructions are used for the goal-oriented error estimates, e.g. in Richter and Wick [2015] linear functions are reconstructed to quadratic using the patch-wise structure of quadrilateral meshes. Further, we refer to the work Carpio et al. [2013] where a patch-wise, higher-order interpolation recovery extensible to finite elements of arbitrary order is used. It is capable to evaluate the weights of the error estimator on unstructured meshes composed of anisotropic triangles. The extension of any of those approaches to the discontinuous Galerkin method is questionable, since, e.g., nodal values of functions, used in most of the techniques, are ambiguous on element edges due to the discontinuity of functions in V_h^d.

In Chapter 2 we present two possible ways how the reconstruction can be computed for the DG method on simplicial meshes which were firstly introduced in articles Dolejší and Solin [2016] and Dolejší and Roskovec [2017]. Apart from those we are not aware of any paper where a local reconstruction of the DG solution would be used to goal-oriented estimates, even though, for instance, the reconstruction based on orthogonal polynomials from Huynh [2009] may be applicable on quadrilateral meshes.

1.1.6 Adaptive algorithm

The main aim of the goal-oriented error estimation method is to obtain an approximation of the quantity of interest $J(u_h)$ with its error under some given tolerance TOL. Given an initial mesh $\mathscr{T}_h^{(0)}$, we iteratively adapt the mesh until the error estimate (1.38) decreases under the tolerance TOL. Employing the goal-oriented error estimate (1.38) may considerably decrease the computational effort if it is used for adaptive mesh refinement. In order to benefit from the estimate (1.38) for mesh adaptation η_{S} and η_{S}^* need to be to localized into positive error indicators describing local error contributions.

In conforming FEM this is usually done by plugging some partition of unity into (1.37) (see, e.g., Richter and Wick [2015]). In DG discretizations we simply define element-wise contributions of (1.37) for each $K \in \mathscr{T}_h$ by

$$\eta_{\mathrm{S,K}} = |r_h(u_h)((z_h^+ - \Pi z_h^+)\,\chi_K)|, \quad \eta_{\mathrm{S,K}}^* = |r_h^*(z_h)((u_h^+ - \Pi u_h^+)\,\chi_K)|, \tag{1.40}$$

which corresponds to a partition of unity formed of the characteristic functions of mesh elements, i.e., $1 = \sum_{K \in \mathscr{T}_h} \chi_K$, plugged into (1.37). Either or both of those can be used as a local error indicator for mesh refinement.

The functional J generally does not have the additive property such as norms and it can attain both positive and negative values on different elements, c.f. Section 1.1.4. Therefore, we cannot expect that the sum of the local error indicators would sharply approximate the total error $J(e_h)$.

Remark. Although the primal and adjoint residuals are theoretically equivalent, see (1.16), (1.20), in the following way

$$r_h(u_h)(z - \varphi_h) = r_h^*(z_h)(u - \psi) \quad \forall \varphi_h, \psi_h \in V_h, \tag{1.41}$$

their localizations (1.40) can differ notably (even if $z_h^+ := z$ and $u_h^+ := u$) and may lead to differently refined meshes.

More precise description of a general goal-oriented mesh adaptation algorithm is given in the following algorithm.

Algorithm 1: Goal-oriented adaptive algorithm

```
let TOL $> 0$ be the given tolerance and $\mathscr{T}_h^{(0)}$ be the initial coarse mesh
for $n = 0, 1, \ldots$ do
    evaluate $u_h^{(n)}$ and $z_h^{(n)}$ by solving (1.8) and (1.17) on $\mathscr{T}_h^{(n)}$
    reconstruct $u_h^{+,(n)}$ and $z_h^{+,(n)}$
    compute $\eta_h^{(n)} = \frac{1}{2}(\eta_{\mathrm{S}}^{(n)} + \eta_{\mathrm{S}}^{*,(n)})$
    if $\eta_h^{(n)} \leq$ TOL then
        STOP computations
    else
        using indicators $\eta_{\mathrm{S,K}}^{(n)}$, $\eta_{\mathrm{S,K}}^{*,(n)}$, $\forall K \in \mathscr{T}_h^{(n)}$ generate new mesh $\mathscr{T}_h^{(n+1)}$
    end
end
```

1.1.7 Guaranteed upper bounds

One of the biggest drawbacks of the goal-oriented error estimates as presented in previous Sections is that the estimate (1.38) is not a guaranteed upper bound of the error quantity $J(e_h)$ due to neglecting of the "higher-order" terms in (1.35).

In this section we assume that the higher-order solutions $u_h^+, z_h^+ \in V_h^{(+)}$ are computed by solving the problems (1.8) and (1.17), respectively, with polynomials of higher degree on a globally refined mesh as in Šolín and Demkowicz [2004] ($V_h^{(+)} = V_{h/2}^{p+1}$). Further, we denote the two parts of the error terms in the primal error identity (1.35) by

$$J(e_h) = r_h(u_h)(z - \varphi_h) = r_h(u_h)(z_h^+ - \varphi_h) + r_h(u_h)(z - z_h^+) =: \eta_h + \varepsilon_h. \tag{1.42}$$

In general, we expect that ε_h is "higher order term" than η_h.

If there exist an "energy" norm $|||\cdot|||$ such that the bilinear form a_h can be bounded in the following way

$$a_h(\varphi, \psi) \leq |||\varphi||| \, |||\psi|||, \qquad \varphi, \psi \in V,$$

we may exploit the Galerkin orthogonality of z_h^+

$$a_h(\psi_h^{(+)}, z - z_h^+) = 0 \qquad \forall \psi_h^{(+)} \in V_h^{(+)}$$

and formally write

$$\begin{aligned} \varepsilon_h &= r_h(u_h)(z - z_h^+) = a_h(u - u_h, z - z_h^+) \\ &= a_h(u - u_h^+, z - z_h^+) \leq |||u - u_h^+||| \, |||z - z_h^+|||. \end{aligned} \tag{1.43}$$

Generally, such norm bounding the bilinear form a_h is clearly linked to the concrete problem and discretization method used, see, e.g., [Dolejší and Feistauer, 2015, Lemma 2.28], for discontinuous Galerkin method applied to Poisson equation. Then under sufficient assumptions on the regularity of the primal and adjoint problems we get

$$|||u - u_h^+||| = O((h/2)^{p+1}), \qquad |||z - z_h^+||| = O((h/2)^{p+1})$$

and hence asymptotically we have that $\eta_h + \varepsilon_h \to \eta_h$.

On the other hand, on a coarse mesh the term ε_h may be superior to the estimate itself. In [Nochetto et al., 2009, Section 2] a numerical example was presented showing that the classical DWR method (neglecting of the approximation error ε_h) may be unreliable. More precisely, for some given tolerance TOL the adaptive algorithm based on η_h stops after computation on the first mesh since $\eta_h <$ TOL even though the true error is still higher than the prescribed tolerance (and $\varepsilon_h \gg \eta_h$).

We briefly summarize the main ideas given in Nochetto et al. [2009] for a design of a more reliable variant of the goal-oriented estimate not suffering of the problems arising from neglecting ε_h which they call "safeguarded DWR estimator".

In order to obtain a guaranteed upper bound of the error of the target quantity $J(e_h)$ the term ε_h has to be estimated somehow. The most straightforward method would be to estimate the higher-order therm $|||u - u_h^+|||$ and $|||z - z_h^+|||$ from (1.43) by global energy norm estimates, e.g., Babuška and Rheinboldt [1978]. Unfortunately, that requires to solve an additional (and larger) algebraic system of equations to compute u_h^+ and further the estimate may be very pessimistic due to a large overestimation in (1.43). Therefore the authors of Nochetto et al. [2009] suggest a slightly different approach. Due to the relation

$$\begin{aligned}\varepsilon_h &= r_h(u_h)(z - z_h^+) = a_h(u - u_h, z - z_h^+) \\ &= a_h(u - \psi^{(+)}, z - z_h^+) = r_h^*(z_h^+)(u - \psi^{(+)}) \qquad \forall \psi^{(+)} \in V_h^{(+)}.\end{aligned} \tag{1.44}$$

computation of u_h^+ can be avoided and instead the term $r_h^*(z_h^+)(u - \psi^{(+)})$ can be estimated by some less or more standard methods.

Later several other papers struggling with this unreliability of DWR method were published. Among others we mention the articles Ainsworth and Rankin [2012], where the quantities of interest of arbitrary order finite element approximations in the context of a linear second-order elliptic problem are bounded by a fully computable error estimates, Mozolevski and Prudhomme [2015], where the method based on equilibrated flux reconstruction technique, cf. Vohralík [2010], is employed to derive guaranteed asymptotically exact estimates of the target quantity, Ladeveze et al. [2013], where the authors use the Saint-Venant's principle to obtain guaranteed and accurate error estimates for FEM discretizations of linear elasticity problems.

1.1.8 Goal oriented error estimates including algebraic errors

Unfortunately, due to algebraic errors neither the "exact" discrete solution u_h of (1.8) nor the solution z_h of (1.17) are available in practical computations. Instead, we compute a sequence of their approximations $u_{h,A}^{(k)} \in V_h$ and $z_{h,A}^{(k)} \in V_h$, $k = 0, 1, \ldots$ resulting from a finite number of iterations of an algebraic iterative solver. We note that even when using direct solvers the results are affected by rounding errors and hence u_h and z_h are still not attainable.

Here we shortly introduce the goal-oriented error estimates including algebraic errors following the ideas from Meidner et al. [2009], Dolejší and Roskovec [2017] and Dolejší and Tichý [2019]. Further this concept will be described in detail in Section

2.3. Considering the algebraically inexact discrete solution $u_{h,A}^{(k)}$ the Galerkin orthogonality (1.10) and (1.18) do not hold anymore. Hence we must add an additional term measuring the deviation from the Galerkin orthogonality due to algebraic errors. For the primal error identity (1.16) using the triangle inequality we have

$$\begin{aligned} J(u-u_{h,A}^{(k)}) &= a_h(u-u_{h,A}^{(k)},z) = a_h(u-u_{h,A}^{(k)},z-\varphi_h) + a_h(u-u_{h,A}^{(k)},\varphi_h) \\ &= r_h(u_{h,A}^{(k)})(z-\varphi_h) + r_h(u_{h,A}^{(k)})(\varphi_h) \qquad \forall \varphi_h \in V_h. \end{aligned} \tag{1.45}$$

Selecting $\varphi_h := z_{h,A}^{(k)}$ in (1.45) we get

$$J(u-u_{h,A}^{(k)}) = r_h(u_{h,A}^{(k)})(z-z_{h,A}^{(k)}) + r_h(u_{h,A}^{(k)})(z_{h,A}^{(k)}) = e_{S,n} + e_{A,n}, \tag{1.46}$$

where the quantity $e_{S,n} := r_h(u_{h,A}^{(k)})(z-z_{h,A}^{(k)})$ represents the discretization error of the primal problem and the quantity $e_{A,n} := r_h(u_{h,A}^{(k)})(z_{h,A}^{(k)})$ represents the algebraic error of the primal problem. We note that $e_{S,n}$ matches (1.8) if $u_{h,A}^{(k)} = u_h$ since $e_{A,n}$ disappears in that case.

In order to derive the counter-part to the adjoint error identity (1.20) including the algebraic errors we exploit the error equivalence (1.21). Then we obtain the following adjoint error identity including algebraic errors

$$\begin{aligned} \ell_h(z-z_{h,A}^{(k)}) &= a_h(u,z-z_{h,A}^{(k)}) = a_h(u-\psi_h,z-z_{h,A}^{(k)}) + a_h(\psi_h,z-z_{h,A}^{(k)}) \\ &= r_h^*(z_{h,A}^{(k)})(u-\psi_h) + r_h^*(z_{h,A}^{(k)})(\psi_h) \qquad \forall \psi_h \in V_h. \end{aligned} \tag{1.47}$$

Similarly to the primal case we choose $\psi_h := u_{h,A}^{(k)}$ and we obtain

$$\ell_h(z-z_{h,A}^{(k)}) = r_h^*(z_{h,A}^{(k)})(u-u_{h,A}^{(k)}) + r_h(u_{h,A}^{(k)})(u_{h,A}^{(k)}) = e_{S,n}^* + e_{A,n}^*, \tag{1.48}$$

where $e_{S,n}^* := r_h^*(z_{h,A}^{(k)})(u-u_{h,A}^{(k)})$ represents the discretization error of the adjoint problem and $e_{A,n}^* := r_h^*(z_{h,A}^{(k)})(u_{h,A}^{(k)})$ stands for the algebraic error of the primal problem.

We use the quantity $\ell_h(z-z_{h,A}^{(k)})$ as analogue to $J(u-u_{h,A}^{(k)})$, since both terms coincide when $u_{h,A}^{(k)} = u_h$ and $z_{h,A}^{(k)} = z_h$, but strictly speaking the identity (1.48) is not equivalent to the error of the quantity of interest since due to algebraic errors generally

$$J(u-u_{h,A}^{(k)}) \neq \ell_h(z-z_{h,A}^{(k)}). \tag{1.49}$$

Nevertheless, it is shown in Dolejší and Tichý [2019] that if the bi-conjugate gradient (BiCG) method is used to compute the approximations $u_{h,A}^{(k)}$ and $z_{h,A}^{(k)}$ then

$$J(u-u_{h,A}^{(k)}) = \ell_h(z-z_{h,A}^{(k)}) \qquad k = 1,2,\dots \tag{1.50}$$

1.1.9 Adjoint problem for a general linear problem

Let us briefly present a technique which enables verifying whether a given target functional J is compatible with a general linear problem and to define the boundary conditions for the adjoint problem. This method originates from Giles and Pierce [1997] and it was further utilized for goal-oriented error estimates in Hartmann [2006].

Let $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$, we consider a linear problem

$$\mathscr{L}u = f \text{ in } \Omega, \tag{1.51}$$

$$\mathscr{B}u = g \text{ on } \partial\Omega, \tag{1.52}$$

where $\mathscr{L}$ and $\mathscr{B}$ denote linear differential operators defined on Ω and on $\partial\Omega$, respectively. Let J be the target functional given by

$$J(u) = \int_\Omega j_\Omega u \,\mathrm{d}x + \int_{\partial\Omega} j_\Gamma \mathscr{C}u \,\mathrm{d}S, \tag{1.53}$$

where $\mathscr{C}$ is some differential operator on the boundary and j_Ω, j_Γ are some given functions defined on Ω and $\partial\Omega$, respectively.

The adjoint problem to (1.51) is reads

$$\mathscr{L}^* z = j_\Omega \quad \text{in } \Omega, \tag{1.54}$$

$$\mathscr{B}^* z = j_\Gamma \quad \text{on } \partial\Omega, \tag{1.55}$$

where $\mathscr{L}^*$ is the adjoint operator to $\mathscr{L}$ and $\mathscr{B}^*$ is some differential boundary operator ($\mathscr{B}^*$ is not adjoint operator to $\mathscr{B}$). We may further define the dual functional

$$J^*(z) = \int_\Omega fz \,\mathrm{d}x + \int_{\partial\Omega} g\mathscr{C}^* z \,\mathrm{d}S, \tag{1.56}$$

even though it is not directly needed for the error estimation.

We say that the target functional (1.53) is *compatible* with the problem (1.51) if it holds

$$\begin{aligned} J(u) &= (j_\Omega, u)_\Omega + (j_\Gamma, \mathscr{C}u)_{\partial\Omega} = (\mathscr{L}^* z, u)_\Omega + (\mathscr{B}^* z, \mathscr{C}u)_{\partial\Omega} \\ &= (z, \mathscr{L}u)_\Omega + (\mathscr{C}^* z, \mathscr{B}u)_{\partial\Omega} = (f, z)_\Omega + (g, \mathscr{C}^* z)_{\partial\Omega} = J^*(z). \end{aligned} \tag{1.57}$$

Here we use the notation $(w,v)_\Omega = \int_\Omega wv \,\mathrm{d}x$ and $(w,v)_{\partial\Omega} = \int_\Omega wv \,\mathrm{d}S$ for better clarity.

The first two and the last two equalities hold directly from the definitions of the problems (1.51), (1.54) and the functionals (1.53), (1.56), hence the critical step is to prove that

$$(\mathscr{L}^* z, u)_\Omega + (\mathscr{B}^* z, \mathscr{C}u)_{\partial\Omega} = (z, \mathscr{L}u)_\Omega + (\mathscr{C}^* z, \mathscr{B}u)_{\partial\Omega}. \tag{1.58}$$

Integrating the left-hand side of the equality (1.51) by parts we obtain

$$(\mathscr{L}u, z)_\Omega = (\mathscr{L}^* z, u)_\Omega + (\mathscr{A}_1 z, \mathscr{A}_2 u)_{\partial\Omega}, \tag{1.59}$$

where $\mathscr{L}^*$ is the adjoint differential operator to $\mathscr{L}$ and $\mathscr{A}_1, \mathscr{A}_2$ the differential operators on the boundary $\partial\Omega$ originating from the integration by parts.

Therefore, target functional J is compatible with the problem (1.51), only if the operators $\mathscr{B}^*$ and $\mathscr{C}^*$ satisfy

$$(\mathscr{A}_1 z, \mathscr{A}_2 u)_{\partial\Omega} = (\mathscr{B}^* z, \mathscr{C}u)_{\partial\Omega} - (\mathscr{C}^* z, \mathscr{B}u)_{\partial\Omega}. \tag{1.60}$$

In Section 2.1.3 we employ this method to verify the compatibility of a given target functional and determine the boundary conditions of the adjoint problem for the linear convection-diffusion equation.

The adjoint consistency of a discretization (1.15) also depends on the choice of the target functional $J(\cdot)$. In some cases, the functional is not in a suitable form and it is necessary to modify J (see Hartmann [2007] for more details) in the following way

$$\tilde{J}(v) := J(i(v)) + \int_{\Omega} r_J(v)\,\mathrm{d}S, \tag{1.61}$$

where $i(\cdot)$, $r_J(\cdot)$ are appropriate functions. We say that a modification of target functional (1.61) is *consistent* if $\tilde{J}(u) = J(u)$ for u being the exact solution of problem (1.3). Thus the consistency of the modification (1.61) is guaranteed if u satisfies $i(u) = u$ and $r_J(u) = 0$. Even though the exact value of the target functional $J(u)$ remains unchanged, $\tilde{J}(u_h) \neq J(u_h)$. Moreover, $\tilde{J}$ is not linear anymore and hence the approach for nonlinear problems, which will be explained in Section 1.2, has to be adapted to this case.

1.2 Goal-oriented error estimates for nonlinear problems

In this section we extend the method for obtaining goal-oriented error estimates to nonlinear problem. We proceed in similar way to the procedure described in [Bangerth and Rannacher, 2003, Chapter 6] and in Hartmann [2006].

Let $\Omega \in \mathbb{R}^d$ be a bounded open domain. We consider the following nonlinear problem

$$\mathscr{A}(u) = 0 \text{ in } \Omega, \qquad \mathscr{B}(u) = 0 \text{ on } \partial\Omega. \tag{1.62}$$

For linear problems the connection of the target quantity and the solved problem is established using the adjoint (or dual) problem. Contrarily, there is no adjoint operator to $\mathscr{A}$ and hence the adjoint problem to (1.62) cannot be defined directly. First, problem (1.62) has to be linearized and then the adjoint operator will be defined with its linearization.

We start with introducing the Fréchet derivative of an operator.

Definition 1.4. *Let V, W be normed vector spaces and $U \subset V$ be an open subset of V. Then a function $f : U \to W$ is called* Fréchet differentiable *at $x \in U$ if there exist a continuous linear operator $f'[x] : V \to W$ such that*

$$\lim_{\|h\|_V \to 0} \frac{\|f(x+h) - f(x) - f'[x](h)\|_W}{\|h\|_V} = 0. \tag{1.63}$$

We call $f'[x] \in \mathscr{L}(V,W)$ the Fréchet derivative of f at the point x.

The function in the bracket $[\cdot]$ represents the state where the linearization is taken about and hence $f'[x](h)$ is the directional derivative at x with direction h.

We assume that both $\mathscr{A}$, $\mathscr{B}$ are differentiable operators and that the quantity of interest can be expressed as a nonlinear functional

$$J(u) = \int_{\Omega} j_{\Omega}(u)\,\mathrm{d}x + \int_{\partial\Omega} j_{\Gamma}(u)\,\mathrm{d}S \tag{1.64}$$

with Fréchet derivative

$$J'[u](v) = \int_\Omega j'_\Omega[u](v)\,\mathrm{d}x + \int_{\partial\Omega} j'_\Gamma[u](v)\,\mathrm{d}S. \tag{1.65}$$

The (strong) adjoint problem to (1.62) reads: find a function z such that

$$(\mathscr{A}'[u])^* z = j'_\Omega[u] \quad \text{in } \Omega, \qquad (\mathscr{B}'[u])^* z = j'_\Gamma[u] \quad \text{on } \partial\Omega, \tag{1.66}$$

where $(\mathscr{A}'[u])^*$ and $(\mathscr{B}'[u])^*$ are the adjoint operators to $\mathscr{A}'[u]$ and $\mathscr{B}'[u]$, respectively.

We consider Hilbert spaces V, W_h and the finite dimensional space $V_h \subset W_h$, consisting of piecewise polynomial functions on a given mesh $\mathscr{T}_h$, see (1.5), (1.6). Further $a_h(\cdot;\cdot)$ represents a form corresponding to some discretization of the problem (1.62) (FEM or DG). In this case a_h is nonlinear with respect to its first argument, but linear with respect to the second one. Now we can introduce the discretization of the problem (1.62).

Definition 1.5. *We say that a function $u_h \in V_h$ is the discrete solution of the problem* (1.62) *if it satisfies*

$$a_h(u_h;\varphi_h) = 0 \qquad \forall \varphi_h \in V_h. \tag{1.67}$$

We assume that both problems (1.62) and (1.67) are well-posed and their solutions u and u_h exist uniquely. The discretization (1.67) is said to be *consistent*, c.f. (1.9), if the exact solution u of the primal problem (1.62) satisfies

$$a_h(u;\varphi) = 0 \qquad \forall \varphi \in W_h. \tag{1.68}$$

where W_h is a suitable infinite dimensional space such that $V_h \subset W_h$ and $u \subset W_h$.

Further, let $a_h : W_h \times W_h \to \mathbb{R}$ be Fréchet differentiable. Then for its directional derivative in the direction $(\varphi, \psi) \in W_h \times W_h$ we may write

$$\begin{aligned} D_{(\varphi,\psi)} a_h(u;w) &= \lim_{t\to 0} \frac{a_h(u+t\varphi; w+t\psi) - a_h(u;w)}{t} \\ &= \lim_{t\to 0} \frac{a_h(u+t\varphi;w) + t a_h(u+t\varphi;\psi) - a_h(u;w)}{t} \\ &= \lim_{t\to 0} \frac{a_h(u+t\varphi;w) - a_h(u;w)}{t} + a_h(u;\psi) \\ &= D_{(\varphi,0)} a_h(u;w) + a_h(u;\psi). \end{aligned} \tag{1.69}$$

For the integrity of notation we use ${a_h}'[u](\varphi,w) := D_{(\varphi,0)} a_h(u;w)$ in the rest of the text, denoting the directional derivative of a_h with respect to its first (nonlinear) variable. Then, similarly to (1.66), we can introduce the discrete adjoint (dual) problem.

Definition 1.6. *We say that a function $z_h \in V_h$ is the discrete adjoint solution of problem* (1.66) *if it satisfies*

$${a_h}'[u_h](\psi_h, z_h) = J'[u_h](\psi_h) \qquad \forall \psi_h \in V_h. \tag{1.70}$$

The discretization (1.67) is said to be *adjoint consistent*, c.f. (1.15), if z (the exact solution of (1.66)) satisfies

$${a_h}'[u](\psi;z) = J'[u](\psi) \qquad \forall \psi \in W_h. \tag{1.71}$$

In other words, the associated discrete adjoint (dual) problem is a consistent discretization of the continuous adjoint problem (1.66). Similarly to the linear case, see Section 1.1.9, only some choices of target functionals are compatible with an adjoint consistent discretization. For some functionals it is necessary to replace the target functional by slightly adjusted satisfying $\tilde{J}(u) = J(u)$.

Remark. We note that the adjoint continuous and discrete problems (1.66) and (1.70) differ more than in the linear case, since while in (1.66) the linearization is done around u, the linearization in (1.70) is done around u_h.

1.2.1 Error estimation for nonlinear problems

Unlike the linear case, it is not obvious how should the adjoint problem help in estimation of the error of the quantity of interest $J(u) - J(u_h)$. The basic relation between the adjoint problem and the error measured with respect to the target functional is presented in the Theorem 1.7. For these estimates we introduce the following notation. For arbitrary functions φ, $\psi \in V$ we denote

$$r_h(u_h)(\varphi) := -a_h(u_h;\varphi) \tag{1.72a}$$

$$r_h^*(u_h,z_h)(\psi) := J'[u_h](\psi) - a_h{}'[u_h](\psi,z_h) \tag{1.72b}$$

the primal and adjoint residuals, respectively. Further, we denote $e_h = u - u_h$, $e_h^* = z - z_h$ the primal and adjoint errors, respectively.

Theorem 1.7. *Let $J : W_h \to \mathbb{R}$ and $a_h : W_h \times W_h \to \mathbb{R}$ be three-times differentiable and let* (1.67) *and* (1.70) *be consistent and adjoint consistent discretization of* (1.67)*. Then for the solutions u solving* (1.62) *and u_h solving* (1.67)*, it holds*

$$J(u) - J(u_h) = \frac{1}{2} r_h(u_h)(z - \varphi_h) + \frac{1}{2} r_h^*(u_h,z_h)(u - \psi_h) + \mathscr{R}_h^{(3)} \qquad \forall \varphi_h, \psi_h \in V_h. \tag{1.73}$$

The dependency of the expression $\mathscr{R}_h^{(3)}$ on the error e_h is qubic and it is given by

$$\begin{aligned} \mathscr{R}_h^{(3)} \;\; = \;\; & \frac{1}{2}\int_0^1 \big\{J'''[u_h+te_h](e_h,e_h,e_h) - a_h{}'''[u_h+te_h](e_h,e_h,e_h,z_h+te_h^*) \\ & -3a_h{}''[u_h+te_h](e_h,e_h,e_h^*)\big\}\, t(t-1)\,\mathrm{d}t. \end{aligned} \tag{1.74}$$

Proof. Our proof is inspired by the proof of [Bangerth and Rannacher, 2003, Proposition 6.1]. Unlike in their approach we avoid introduction of the Euler-Lagrange method since in our opinion it is in this case a bit artificial technique, which complicates the understanding of the underlying relations.

Due to the primal consistency (1.68), formulation (1.67), the integral representation formula and (1.69) it holds for error of the target functional

$$\begin{aligned} J(u) - J(u_h) &= J(u) - J(u_h) - a_h(u;z) + a_h(u_h;z_h) \qquad (1.75) \\ &= \int_0^1 J'[u_h+te_h](e_h) - D_{(e_h,e_h^*)}a_h(u_h+te_h;z_h+te_h^*)\,\mathrm{d}t. \\ &= \int_0^1 J'[u_h+te_h](e_h) - \big(a_h{}'[u_h+te_h](e_h,z_h+te_h^*) + a_h(u_h+te_h;e_h^*)\big)\,\mathrm{d}t. \end{aligned}$$

From the Trapezoidal error formula we have

$$\int_0^1 f(t)\,\mathrm{d}t = \frac{1}{2}\left(f(0)+f(1)\right) + \frac{1}{2}\int_0^1 f''(s)s(s-1)\,\mathrm{d}t$$

for any function $f \in C^2([0,1])$.

Employing this error representation and further the primal consistency (1.68) and adjoint consistency (1.71) we get

$$\begin{aligned} J(u) - J(u_h) &= \frac{1}{2}\left[J'[u_h](e_h) - {a_h}'[u_h](e_h, z_h) - a_h(u_h; e_h^*)\right. \\ &\quad \left. + (J'[u](e_h) - {a_h}'[u](e_h, z)) - a_h(u; e_h^*)\right] + \mathscr{R}_h^{(3)} \\ &= \frac{1}{2}\left(r_h^*(u_h, z_h)(e_h) + r_h(u_h)(e_h^*)\right) + \mathscr{R}_h^{(3)} \end{aligned} \tag{1.76}$$

where $\mathscr{R}_h^{(3)}$ is the error term from the trapezoidal rule. The statement of the theorem is the a direct consequence of relations

$$\begin{aligned} r_h(u_h)(e_h^*) &= r_h(u_h)(z - \varphi_h) \quad \forall \varphi_h \in V_h, \\ r_h^*(u_h, z_h)(e_h) &= r_h^*(u_h, z_h)(u - \psi_h) \quad \forall \psi_h \in V_h. \end{aligned}$$

□

Furthermore, we can get also a simplified error representation employing only the primal residuum (1.72a).

Theorem 1.8. *The error of the target functional satisfies also*

$$J(u) - J(u_h) = r_h(u_h)(z - \varphi_h) + \mathscr{R}_h^{(2)} \qquad \forall \varphi_h \in V_h, \tag{1.77}$$

where $\mathscr{R}_h^{(2)}$ *depends quadratically on the error* e_h *and it is given by*

$$\mathscr{R}_h^{(2)} \;=\; \int_0^1 \left({a_h}''[u_h + se_h](e_h, e_h, z) - J''[u_h + se_h](e_h, e_h)\right) t\,\mathrm{d}t \tag{1.78}$$

Proof. Similarly to the previous proof we employ the integral representation theorem and then thanks to the integration by parts we get

$$\begin{aligned} J(u) - J(u_h) - r_h(u_h)(z - \psi_h) &= J(u) - J(u_h) - a_h(u; z) + a_h(u_h; z) \\ &= \int_0^1 (J'[u_h + te_h](e_h) - {a_h}'[u_h + te_h](e_h, z))1\,\mathrm{d}t \\ &= -\int_0^1 \left(J''[u_h + te_h](e_h, e_h) - {a_h}''[u_h + te_h](e_h, e_h, z)\right) t\,\mathrm{d}t. \end{aligned} \tag{1.79}$$

We note that the boundary terms from integration by parts disappear due to the adjoint consistency (i.e., $(J'[u](e_h) - {a_h}'[u](e_h, z)) \cdot 1 = 0$). □

1.2.2 Alternative approach

In the following we present a slightly different approach to a posteriori error estimation from Hartmann [2005], Hartmann and Houston [2006b]. This approach enables deriving of the error representation without the terms coming from the linearization, hence theoretically it provides a more direct relation between the error and the residual of the primal problem. The cost for that is that the adjoint problem is derived an integral form and in order to obtain a computable quantity the integral representation has to be replaced by its approximation (linearization). That is done by the authors without formulating precisely what the error term caused by this linearization looks like.

We consider the (adjoint consistent) DG discretization of the general nonlinear problem (1.62). Further, we assume that the target functional $J(\cdot)$ is differentiable and hence we can define

$$\overline{J}(u,u_h;\psi) := \int_0^1 J'[tu+(1-t)u_h](\psi)\,\mathrm{d}t, \tag{1.80}$$

where $J'[w](\cdot)$ denotes the Fréchet derivative of J evaluated at some $w \in V$. Then it holds $\overline{J}(u,u_h;u-u_h) = J(u)-J(u_h)$.

Here, V is some suitably chosen function space such that $V_h \subset V$. Similarly, we define the mean-value linerization of the form a_h by

$$M(u,u_h;\psi,\varphi) := \int_0^1 {a_h}'[tu+(1-t)u_h](\psi,\varphi)\,\mathrm{d}t \tag{1.81}$$

for all $\varphi \in V$. Here, ${a_h}'[w](\cdot,\varphi)$ denotes the Fréchet derivative of $u \mapsto a_h(u,\varphi)$ for $\varphi \in V$ fixed at some $w \in V$. We denote (once more) that the linearization defined in (1.81) is only a formal manipulation, since for concrete problems ${a_h}'[w](\cdot,\cdot)$ may not exist in general and hence its suitable approximation has to be used instead, cf. Hartmann [2005], where finite difference quotients are employed to this approximation, or Hartmann and Houston [2006a], where the approximation is defined for the compressible Navier-Stokes equations.

We have

$$M(u,u_h;u-u_h,\varphi) = a_h(u;\varphi) - a_h(u_h;\varphi).$$

Then we can introduce the adjoint problem in the following shape: find $z \in V$ such that

$$M(u,u_h;\varphi,z) = \overline{J}(u,u_h;\varphi) \qquad \forall \varphi \in V. \tag{1.82}$$

Then we directly obtain the following error representation

$$\begin{aligned} J(u)-J(u_h) &= \overline{J}(u,u_h;u-u_h) = M(u,u_h;u-u_h,z) \\ &= a_h(u;z) - a_h(u_h;z) = -a_h(u_h;z-\varphi_h) \quad \forall \varphi_h \in V_h. \end{aligned} \tag{1.83}$$

In general we assume that the problem (1.82) possesses an unique solution, although in a concrete application this clearly depends on the definition of $M(u,u_h;\cdot,\cdot)$ and the choice of the considered target functional J.

In order to obtain a computable a posteriori error estimate the unknown solution z needs to be replaced by some suitable numerical approximation $\tilde{z}$. To this end the adjoint problem is linearized which directly leads to (1.71) and its discretized counterpart equals (1.70).

1.2.3 Differences from linear case

Similarly to the linear case u, z in (1.73), or at least z in (1.77), have to be replaced by some computable quantities in order to obtain computable estimates of the error of the quantity of interest. Fortunately, all of the techniques presented in Section 1.1.5 can be used as well for nonlinear problems. Moreover, the remainder terms $\mathscr{R}_h^{(3)}$ and $\mathscr{R}_h^{(2)}$ are usually neglected, which again leads to estimates which are not guaranteed upper bounds. Altogether we arrive to the approximate error representations

$$J(u) - J(u_h) \approx \frac{1}{2}\left(r_h(u_h)(z_h^+ - \varphi_h) + r_h^*(u_h, z_h)(u_h^+ - \psi_h)\right) \tag{1.84}$$

$$J(u) - J(u_h) \approx r_h(u_h)(z_h^+ - \varphi_h) \tag{1.85}$$

based on (1.73) and (1.77), respectively.

In the linear case, the primal and adjoint residuals coincide, see (1.41). That is no longer true for the nonlinear case, but the deviation from this relation can be estimated using

$$\Delta r = r_h^*(u_h, z_h)(u - \psi_h) - r_h(u_h)(z - \varphi_h) \qquad \forall \varphi_h, \psi_h \in V_h, \tag{1.86}$$

where

$$\Delta r = \int_0^1 a_h{}''[u_h + se_h](e_h, e_h, z_h + te_h^*) - J''[u_h + se_h](e_h, e_h)\,\mathrm{d}t. \tag{1.87}$$

The proof of (1.86) can be found in Bangerth and Rannacher [2003].

1.2.4 Solution of the nonlinear discrete primal problem

The discrete problem (1.67) forms a nonlinear system of algebraic equations, which has to be solved by some iterative method. The most widely used method for solving such system is the Newton method. Compared to linear problem it is obviously much more computationally expensive since the Newton method requires a solution of a linear problem in each step. On the other hand, solving the adjoint problem (1.70) for the converged primal solution corresponds to one additional step of the Newton method. Therefore the goal-oriented error estimates require relatively small overhead in the computation effort. Here, we briefly introduce the algorithm of the Newton method.

Let N_h denote the dimension of the space V_h and let $\mathrm{B}_h = \{\varphi_i(x),\ i = 1, \dots, N_h\}$ denote a set of linearly independent functions forming a basis of V_h. Any function $u_h \in V_h$ can be expressed in the form

$$u_h(x) = \sum_{j=1}^{N_h} \xi^j \varphi_j(x) \in V_h \longleftrightarrow \boldsymbol{\xi} = (\xi^j)_{j=1}^{N_h} \in \mathbb{R}^{N_h}, \tag{1.88}$$

where $\xi^j \in \mathbb{R},\ j = 1, \dots, N_h$ are its basis coefficients with respect to B_h. Obviously, (1.88) defines an isomorphism between $u_h \in V_h$ and $\boldsymbol{\xi} \in \mathbb{R}^{N_h}$. We call $\boldsymbol{\xi}_k$ the *algebraic representation* of u_h.

In order to rewrite the nonlinear system (1.67) to its algebraic representation, we define the vector-valued function $\boldsymbol{F}_h : \mathbb{R}^{N_h} \to \mathbb{R}^{N_h}$ by

$$\boldsymbol{F}_h(\boldsymbol{\xi}) = (a_h(u_h; \varphi_i))_{i=1}^{N_h}, \tag{1.89}$$

where $\boldsymbol{\xi} \in \mathbb{R}^{N_h}$ is the algebraic representation of $\boldsymbol{U}_h \in V_h$.

Therefore, the algebraic representation of the systems (1.67) reads: Find $\boldsymbol{\xi} \in \mathbb{R}^{N_h}$ such that

$$\boldsymbol{F}_h(\boldsymbol{\xi}) = \mathbf{0}. \tag{1.90}$$

Further, we denote the $N_h \times N_h$ Jacobi matrix

$$\frac{D\boldsymbol{F}_h(\bar{\boldsymbol{\xi}})}{D\boldsymbol{\xi}} = \left(a_h{}'[\bar{u}_h](\varphi_j, \varphi_i)\right)_{i,j=1}^{N_h} \tag{1.91}$$

where $\varphi_i \in \mathrm{B}_h,\ i = 1,\dots,N_h$ and $\bar{\boldsymbol{\xi}} \in \mathbb{R}^{N_h}$ is the algebraic representation of $\bar{u}_h \in V_h$.

Now we introduce the damped Newton method solving (1.90). This method generates a sequence $\boldsymbol{\xi}^l,\ l = 0,1,\dots$ of approximations of the true numerical solution $\boldsymbol{\xi}$. For $l = 0,...$ the $l-$th iterative approximation $\boldsymbol{\xi}^l \in \mathbb{R}^{N_h}$ is updated by

$$\boldsymbol{\xi}^{l+1} = \boldsymbol{\xi}^l + \lambda^l \boldsymbol{\delta}^l, \tag{1.92}$$

where the update vector $\boldsymbol{\delta}^l \in \mathbb{R}^{N_h}$ is given as the solution of the problem

$$\frac{D\boldsymbol{F}_h(\bar{\boldsymbol{\xi}}^l)}{D\boldsymbol{\xi}}\boldsymbol{\delta}^l = -\boldsymbol{F}_h(\boldsymbol{\xi}^l). \tag{1.93}$$

The so-called damping parameter λ^l improves the convergence of the method, see Deuflhard [2004]. After solving linear system (1.93), we check, whether it holds

$$\kappa^l = \frac{\left\|\boldsymbol{F}_h(\boldsymbol{\xi}^{l+1})\right\|}{\left\|\boldsymbol{F}_h(\boldsymbol{\xi}^l)\right\|} < 1. \tag{1.94}$$

When this condition is satisfied, we proceed to the next Newton-like iteration, otherwise, we put $\lambda^l = \lambda^l/2$ and set $\boldsymbol{\xi}^{l+1}$ with this new λ^l and check whether (1.94) is now satisfied. This process is iterated until (1.94) holds or some minimal level of the step-length $\lambda_{\mathrm{MIN}} > 0$ is achieved. The damping procedure provides convergence of the method under weaker assumptions on the function $\boldsymbol{F}$ than the standard Newton method, see Deuflhard [2004].

1.2.5 Goal-oriented error estimates including algebraic errors

Similarly to Section 1.1.8, we comment on how the algebraic errors arising from the inexact solution of the problems (1.67) and (1.70) can be included to the goal-oriented error estimates.

The whole iteration process (1.92) is terminated when some error criterion decreases under a given tolerance $\mathrm{TOL}_{\mathrm{nl}}$. In general case it is usually used

$$\left\|\boldsymbol{F}_h(\boldsymbol{\xi}^l)\right\| < \mathrm{TOL}_{\mathrm{nl}} \text{ or } \left\|\boldsymbol{\delta}^l\right\| < \mathrm{TOL}_{\mathrm{nl}}. \tag{1.95}$$

Choosing appropriate tolerance $\mathrm{TOL}_{\mathrm{nl}}$ and error criterion for terminating the iterative process (1.92) has a key role for efficiency of the adaptive algorithm.

Since we are interested in estimating the error $J(u)-J(u_h)$ it suggests itself to use some criterion connected with estimates of the error with respect to the quantity of interest. Following the analysis of Rannacher and Vihharev [2013] we revisit the error estimates from Section 1.2.1 assuming now that the solution of the discrete problems (1.67) and (1.70) are only approximated.

Lemma 1.9. *Let* $u_{h,A}^{(l)}, z_{h,A}^{(l)} \in V_h$ *be the inexact solutions of the problems* (1.67) *and* (1.70)*, respectively. Then it holds*

$$\begin{aligned} J(u)-J(u_{h,A}^{(l)}) =& \frac{1}{2} r_h(u_{h,A}^{(l)})(z-z_{h,A}^{(l)}) + \frac{1}{2} r_h^*(u_{h,A}^{(l)}, z_{h,A}^{(l)})(u-u_{h,A}^{(l)}) \\ &+ r_h(u_{h,A}^{(l)})(z_{h,A}^{(l)}) + \mathscr{R}_h^{(3)}. \end{aligned} \tag{1.96}$$

Proof. We proceed similarly to (1.75), with the difference that $a_h(u_{h,A}^{(l)}; z_{h,A}^{(l)}) \neq 0$.

Hence we obtain

$$J(u)-J(u_{h,A}^{(l)}) = J(u)-J(u_{h,A}^{(l)}) - a_h(u;z) + a_h(u_{h,A}^{(l)}; z_{h,A}^{(l)}) - a_h(u_{h,A}^{(l)}; z_{h,A}^{(l)}). \tag{1.97}$$

The rest of the proof follows exactly the progression of the proof of the Theorem 1.7. □

Hence we may replace the criterion for the error of the Newton solver (1.95) by

$$|r_h(u_{h,A}^{(l)})(z_{h,A}^{(l)})| < \mathrm{TOL}_{\mathrm{nl}}. \tag{1.98}$$

Further, even the linear system (1.93) does not have to be solved exactly, but only up to a given tolerance $\mathrm{TOL}_{\mathrm{lin}}$ using some iterative method instead. A goal-oriented adaptive algorithm controlling all of these sources of errors was presented in Rannacher and Vihharev [2013].

1.3 Demonstration experiments

We present two numerical experiments which should give some basic intuition on how the adjoint problems behave for both linear and nonlinear problem. While it can be usually quite easily seen for linear problems from the parameters of the primal problem and the definition of the quantity of interest, it is often quite unclear for nonlinear problems, where the adjoint problem depends also on the primal solution u which is a priori unknown. Finally we compare an adjoint consistent and adjoint inconsistent DG discretizations for a linear problem where the exact primal and adjoint solution equal.

1.3.1 Linear convection-diffusion example

On square domain $\Omega = (0,1)^2$ we consider a convection-diffusion equation

$$\mathscr{L}u := -\varepsilon \Delta u + \nabla \cdot (\boldsymbol{b}u) = 0, \tag{1.99}$$

where the convection is given by vector $\boldsymbol{b} = (-x_2, x_1)$ and the parameter prescribing the amount of diffusion is $\varepsilon = 10^{-6}$. We prescribe homogeneous Neumann boundary

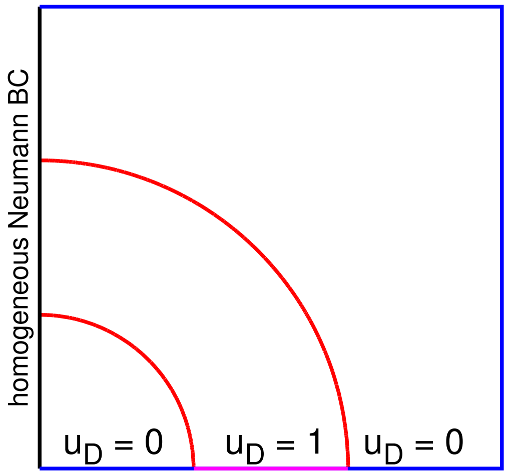

Figure 1.1: Linear convection-diffusion: boundary condition

conditions on $\Gamma_N = \{(x_1,x_2) \in \partial\Omega; x_1 = 0\}$ and Dirichlet boundary condition on $\Gamma_D = \partial\Omega \backslash \Gamma_N$. There, we put

$$u = \begin{cases} 1 & \text{if } x_1 \in (\frac{1}{3}, \frac{2}{3}) \text{ and } x_2 = 0, \\ 0 & \text{elsewhere on } \Gamma_D. \end{cases} \tag{1.100}$$

The setting of the boundary conditions is depicted in Figure 1.1.

Multiplying the equation by a test function, integrating over Ω and finally employing integration by parts we determine that the corresponding adjoint operator is given by

$$\mathscr{L}^* z = -\varepsilon \Delta z - \boldsymbol{b} \cdot \nabla z. \tag{1.101}$$

While the diffusive part of the operator $\mathscr{L}^*$ remains the same, convective flow of $\mathscr{L}^*$ propagates in the reversed direction than in $\mathscr{L}$.

We set Ω_J as square with vertices $[0.25, 0.625]$, $[0.3125, 0.6875]$, $[0.375, 0.625]$, $[0.3125, 0.5625]$ and we choose

$$J(\varphi) := \int_{\Omega_J} \varphi \, \mathrm{d}x. \tag{1.102}$$

In Figure 1.2 we sketched the primal and adjoint solutions, respectively, together with Ω_J. Due to the small amount of diffusion these are not discontinuous as Figure 1.2 depicts, but smoothened over the edges in a layer proportional to the amount of diffusion ε.

First, we show a typical result of adaptive algorithm based on a standard method of a posteriori error estimation (no quantity of interest involved). For this purpose we used the (RES) method from Dolejší [2013] estimating the dual norm of the residual of the numerical solution.

In Figure 1.3 the adaptively refined mesh and isocurves of the corresponding numerical solution u_h are shown after several iterations of the adaptive algorithm (RES). We see that the mesh is strongly refined along both waves where the solution u possesses steep gradients.

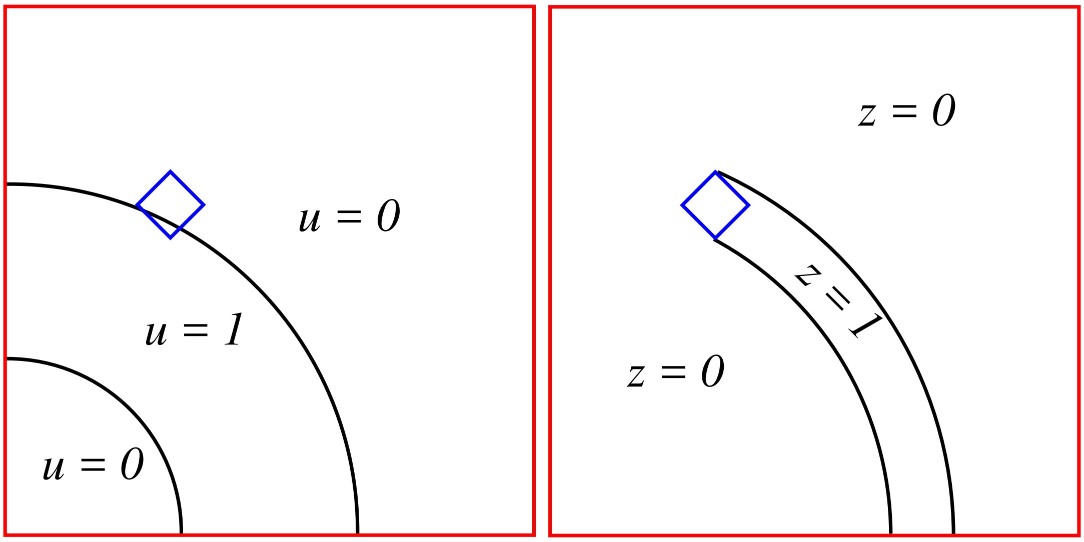

Figure 1.2: Linear convection-diffusion: sketch of primal (left) and adjoint solution (right)

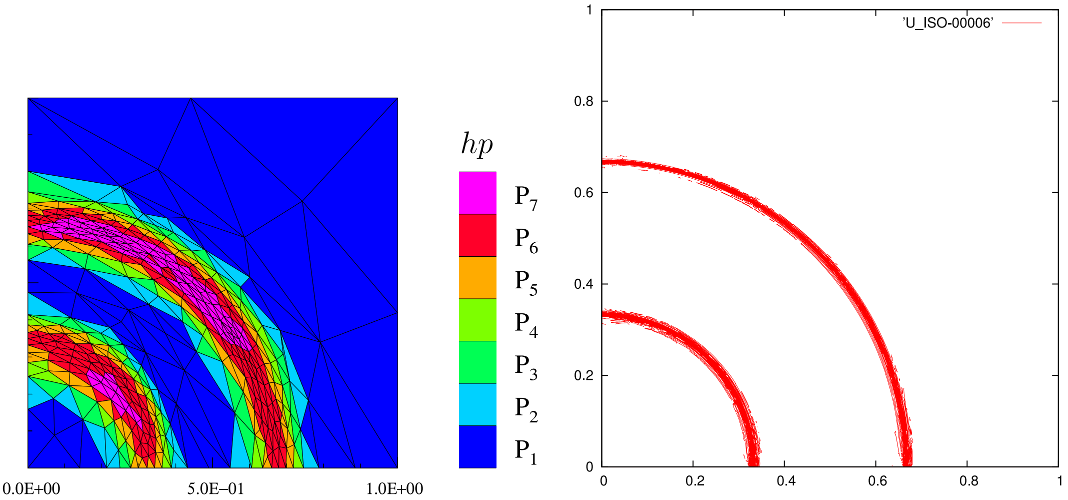

Figure 1.3: Linear convection-diffusion: adaptively refined mesh and isocurves of the numerical solution obtained by (RES) algorithm (not goal-oriented)

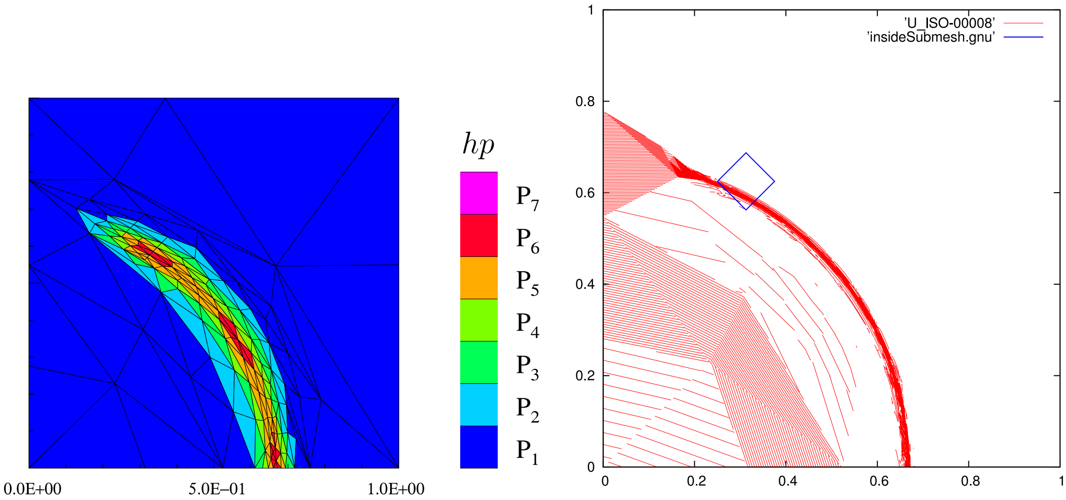

Figure 1.4: Linear convection-diffusion: adaptively refined mesh and isocurves of the numerical solution obtained by (DWR) algorithm

In Figure 1.4 the analogous results are shown after several adaptive steps based on the goal-oriented error estimation (DWR) as introduced in Algorithm 11. Unlike the previous case the mesh is only refined in the vicinity of the "wave" passing through Ω_J and only in the direction opposite to the convection of $\mathscr{L}$.

We note that the quantity of interest $J(u)$ is approximated with very similar precision for both methods, but the DWR method attains it with half the amount of degrees of freedom and hence much faster.

1.3.2 Nonlinear equation

Recalling the argumentation in Section 1.2 we note that the adjoint problem for nonlinear problems is formulated using the linearization of the problem around its solution u (or u_h for its discrete counterpart). Generally, the relation of the adjoint solution with the original is rather unclear for nonlinear problems, since it not influenced only by the definition of the primal problem, but also by the shape of the solution u itself.

Let us consider the nonlinear scalar problem

$$-\varepsilon\nabla\cdot(\mathbb{K}(u)\nabla u)+\nabla(\boldsymbol{b}u)=f \text{ in } \Omega=(0,1)^2, \tag{1.103}$$

$$u=g_D \text{ on } \Gamma_D=(x_1,x_2)\in\partial\Omega, \tag{1.104}$$

$$\nabla\cdot u=g_N \text{ in } \Gamma_N=\partial\Omega\backslash\Gamma_D, \tag{1.105}$$

where the coefficients are given by $\varepsilon=10^{-2}$, $\boldsymbol{b}=(1,0)^{\mathrm{T}}$, $\mathbb{K}(u)=\mathbb{I}|u|^{\gamma}$ and the constant $\gamma\geq 0$ will be specified later. We note that the question of existence and uniqueness of the solution of the problem with (1.103) general data f, g_D, g_N would require a deeper analysis. In this experiment, the functions f, g_D, g_N are chosen such that the exact solution has the form

$$u=\arctan(\alpha(x_1-\beta))+\frac{\pi}{2}, \tag{1.106}$$

with $\alpha=-25$ and $\beta=0.4$. The quantity is of interest is chosen as the mean value of u over rectangle region Ω_J with corners $[0.75,0.375]$, $[0.875,0.375]$, $[0.875,0.625]$,

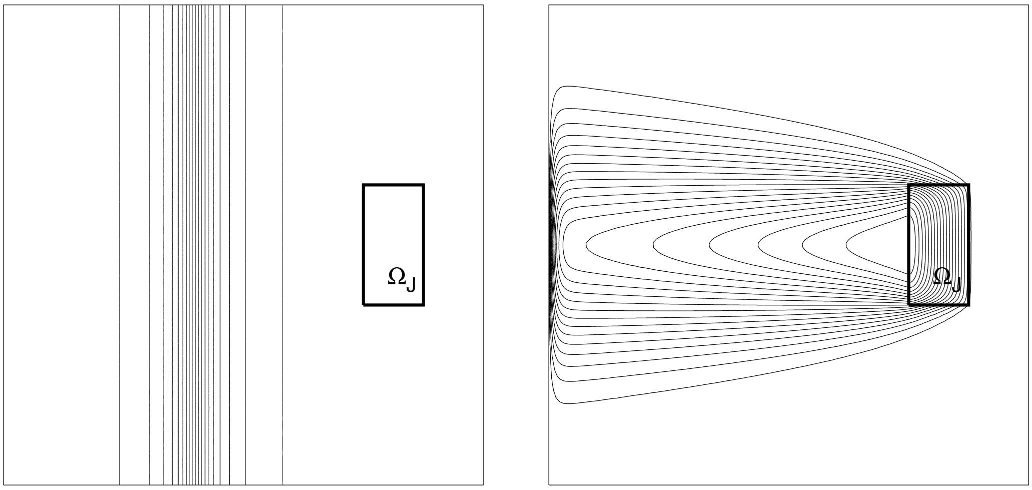

Figure 1.5: Linear case ($\gamma = 0$): isolines of the primal solution (left) and adjoint solution (right)

$[0.75, 0.625]$,i.e.,

$$J(u) := \frac{1}{|\Omega_J|} \int_{\Omega_J} u \, \mathrm{d}x. \tag{1.107}$$

The linearized adjoint problem to (1.103) in its strong form reads

$$-\varepsilon \nabla \cdot (\mathbb{K}(u)\nabla z + \mathbb{K}'[u](z)\nabla u) - \boldsymbol{b} \cdot \nabla z = \frac{1}{|\Omega_J|} \chi_{\Omega_J}. \tag{1.108}$$

Although the problem (1.103) is a bit artificial, we believe that it is illustrative for visualizing the relation between primal and adjoint solutions for nonlinear problems (similarly to Example 1.3.1 for the linear case). Generally, the relation of the adjoint solution with the original is rather unclear, but in this case we can exploit the knowledge of the primal solution to comment on the shape of the adjoint solution z, see right parts of the Figures 1.5 and 1.6.

If we choose the parameter $\gamma = 0$ (in the definition of $\mathbb{K}$), we obtain a linear convection-diffusion problem similar to Example 1.3.1. In Figure 1.5 we see that the adjoint solution z is linearly "spreading" with the increasing distance from Ω_J in the direction of $-\boldsymbol{b}$, which is caused by the constant diffusion contained in the problem.

When we choose $\gamma = 2$, the diffusion of the problem (1.103) becomes nonlinear. We may imagine that u stands for temperature and the quantity of interest represents its mean value over Ω_J in an environment where the diffusion of u depends on the value of u while the convective flow remains constant. Then the adjoint problem helps us to estimate the dependence of the value of u in Ω_J on its values in different parts of the domain Ω which can be later used to conduct the adaptive mesh refinement. Since $u \approx 0$ on the right half of the domain Ω, convection (with opposite direction as in the primal problem) dominates there. The character of the problem (1.108) changes to convection-diffusion over the transition zone around $x_1 = 0.4$.

This "physical" interpretation also helps us to understand why the adjoint problem is linear even for nonlinear primal problems. The nonlinearity in u expresses the dependence of the model on the value of u. While that is not a priori known for the primal problem, it is already resolved when solving the adjoint problem.

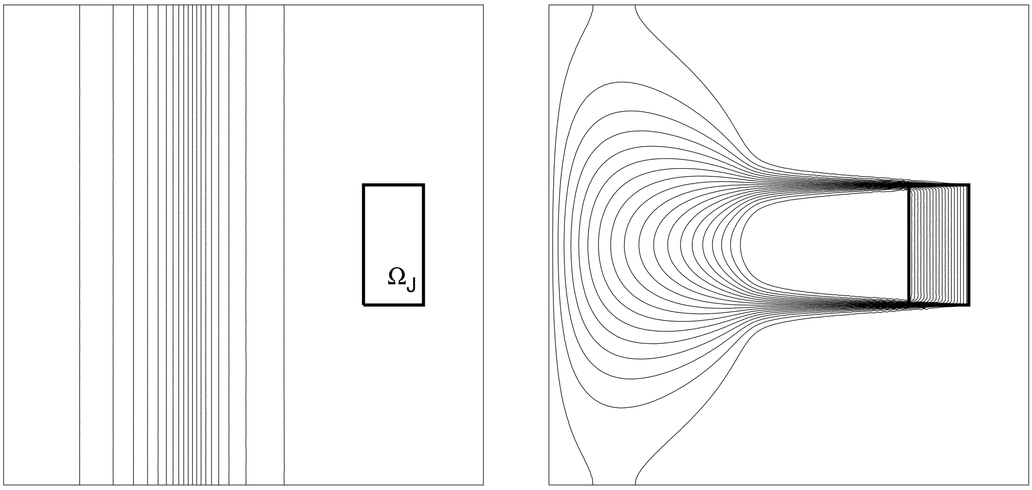

Figure 1.6: Nonlinear case ($\gamma = 2$): isolines of the primal solution (left) and adjoint solution (right)

1.3.3 Importance of adjoint consistency

We present an numerical experiment illustrating the importance of adjoint consistency of the discretization. We set $\Omega = (0,1)^2$ and we consider the Poisson problem with homogeneous Dirichlet boundary condition

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{1.109}$$

We choose $f = -32x(1-x)y(1-y)$ so the exact solution equals

$$u = 16x(1-x)y(1-y). \tag{1.110}$$

Further, we set

$$J(u) := \int_\Omega f u \,\mathrm{d}x, \tag{1.111}$$

so the adjoint problem is equivalent to the primal one (and $z = u$). In this example we use linear polynomial ($p = 1$) on a regular mesh containing 125 triangles.

In Figure 1.7 the results obtained by the symmetrical (SIPG) and nonsymmetrical (NIPG) interior penalty Galerkin method, respectively, are compared. Only the SIPG version is adjoint consistent, see Lemma 2.5 in Chapter 2, while NIPG is not. We can see that while the primal and adjoint solution coincide for SIPG method, the adjoint solution obtained by NIPG method contains nonphysical mesh-dependent oscillations, caused by the inconsistent terms in the discretization scheme.

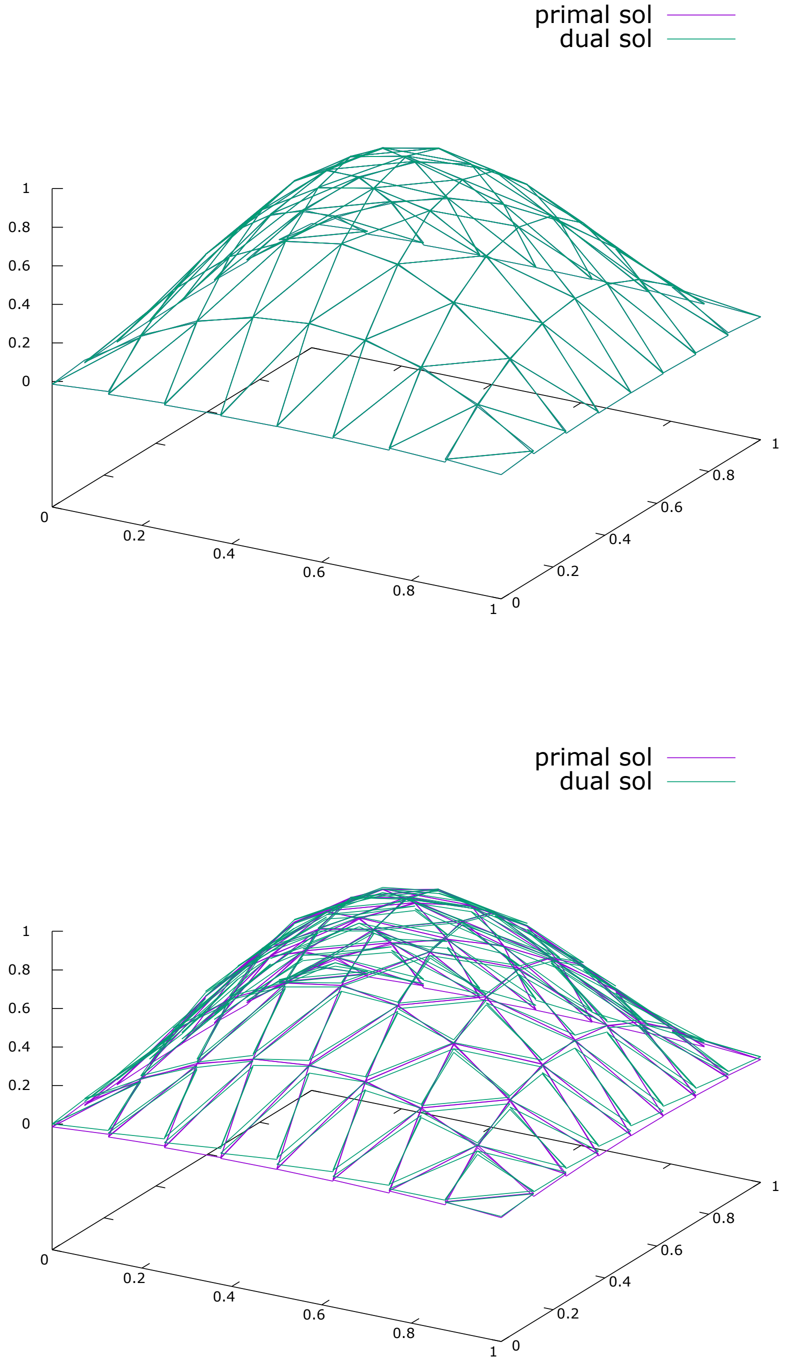

Figure 1.7: Poisson problem, comparison between the primal and adjoint solutions obtained by the SIPG (top) and NIPG (bottom) methods

2. Linear convection-diffusion problem

In this chapter we deal with *discontinuous Galerkin* discretization of a linear convection-reaction-diffusion equation and the corresponding a posteriori error estimates of $J(u) - J(u_h)$, where u_h is the approximate solution.

The goal oriented error estimates require a sufficiently accurate approximation of the solution of the (continuous) adjoint problem. One possibility is to solve the adjoint problem on globally refined mesh which is time-consuming. We present two different reconstruction techniques allowing an efficient and accurate approximation of the solution of the adjoint problem. This way of post-processing is commonly used for finite element computations, see Richter and Wick [2015], but in DG discretizations most of the methods for goal-oriented error estimation described in literature, e.g., Hartmann and Houston [2006a], Harriman et al. [2003], are based on globally higher-order solution of the adjoint problem.

The discretization of the primal and adjoint problems leads to two linear algebraic systems, which are usually solved by a suitable iterative technique. Therefore the error of the resulting solution and its error estimate are influenced by the error resulting from inexact solution of both algebraic systems. Following the ideas from Arioli et al. [2013], we take into account also the *algebraic error* resulting from inaccurate solution of the algebraic systems mentioned above. This aspect was considered in Meidner et al. [2009] with the emphasis on the multigrid methods for conforming finite element methods. The novelty of our approach is the consideration of the algebraic error of the adjoint problem, which was not taken into account in Meidner et al. [2009]. Then we are able to balance the discretization and algebraic errors for the primal as well as for the adjoint problem.

Further, we propose an adaptive algorithm including stopping criteria for the iterative solutions of the primal and adjoint algebraic problems.

Finally, numerical experiments are presented where the decrease of the algebraic errors, when employing the algebraic estimators, is demonstrated.

2.1 Problem description

Let $\Omega \subset \mathbb{R}^d$ be a bounded polygonal domain with Lipschitz boundary. Moreover, let the vector valued function $\boldsymbol{b} = \{b_i\}_{i=1}^d$ be a linear convection coefficient whose entries b_i are Lipschitz continuous real-valued functions in Ω, c denotes the reaction coefficient and $\mathbb{A} = \{a_{i,j}\}_{i,j=1}^d$ is a symmetric diffusion tensor with bounded piece-wise continuous real-valued entries, satisfying the elliptic property

$$\boldsymbol{\zeta}^T \mathbb{A}(x) \boldsymbol{\zeta} \geq 0 \quad \forall \boldsymbol{\zeta} \in \mathbb{R}^d, \text{a.e. } x \in \Omega.$$

By $\boldsymbol{n}(x)$ we denote the unit outward normal vector to $\partial\Omega$ at $x \in \partial\Omega$. We define a disjoint decomposition of the boundary $\partial\Omega$ by

$$\begin{aligned}
\Gamma_0 &:= \{x \in \partial\Omega : \boldsymbol{n}(x)^{\mathrm{T}} \mathbb{A}(x) \boldsymbol{n}(x) > 0\}, \\
\Gamma_- &:= \{x \in \partial\Omega \backslash \Gamma_0 : \boldsymbol{b}(x) \cdot \boldsymbol{n}(x) < 0\}, \\
\Gamma_+ &:= \{x \in \partial\Omega \backslash \Gamma_0 : \boldsymbol{b}(x) \cdot \boldsymbol{n}(x) \geq 0\}.
\end{aligned}$$

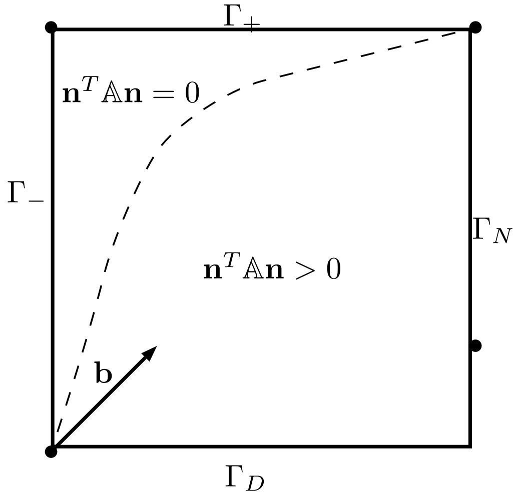
Γ_+
$\mathbf{n}^T\mathbb{A}\mathbf{n} = 0$
Γ_-
Γ_N
$\mathbf{n}^T\mathbb{A}\mathbf{n} > 0$
$\mathbf{b}$
Γ_D

Figure 2.1: Example of the division of the boundary $\partial\Omega$ into Γ_-, Γ_+, and $\Gamma_0 = \Gamma_D \cup \Gamma_N$.

Obviously, these sets are disjoint and $\partial\Omega = \Gamma_0 \cup \Gamma_- \cup \Gamma_+$. Further, we divide Γ_0 into two disjoint subset Γ_D and Γ_N, see Figure 2.1. We assume that $\Gamma_- \cup \Gamma_D \neq \emptyset$ and that $\boldsymbol{b} \cdot \boldsymbol{n} \geq 0$ on Γ_N whenever $\Gamma_N \neq \emptyset$.

We consider the following linear convection-diffusion-reaction model problem

$$\mathscr{L}u := -\nabla \cdot \mathbb{A}\nabla u + \nabla \cdot (\boldsymbol{b}u) + cu = f \qquad \text{in } \Omega, \tag{2.1a}$$

$$u = u_D \qquad \text{on } \Gamma_D \cup \Gamma_-, \tag{2.1b}$$

$$\mathbb{A}\nabla u \cdot \boldsymbol{n} = g_N \qquad \text{on } \Gamma_N, \tag{2.1c}$$

where $u : \Omega \to \mathbb{R}$ is an unknown scalar function. Since the diffusion may degenerate in some parts of Ω, problem (2.1) has to be considered as a first-order PDE in those parts and hence no boundary condition can be set on Γ_+. This kind of problems is termed as "partial differential equations with nonnegative characteristic form" in Houston et al. [2002].

We assume that the data satisfy $f \in L^2(\Omega)$, u_D is trace of some $u^* \in H^1(\Omega)$ on $\Gamma_D \cup \Gamma_-$, $g_N \in L^2(\Gamma_N)$, $c \in L^\infty(\Omega)$.

We proceed to the weak formulation of (2.1).

Definition 2.1. *The function $u \in H^1(\Omega)$ is called the* weak solution *of* (2.1) *if $u - u^* \in H^1_D(\Omega) := \{v \in H^1(\Omega); v|_{\Gamma_D \cup \Gamma_-} = 0\}$ and*

$$a(u, \varphi) = \ell(\varphi) \quad \forall \varphi \in H^1_D(\Omega), \tag{2.2}$$

where

$$a(u,\varphi) := \int_\Omega \mathbb{A}\nabla u \cdot \nabla\varphi \,\mathrm{d}x - \int_\Omega (u\boldsymbol{b} \cdot \nabla\varphi - cu\varphi)\,\mathrm{d}x + \int_{\Gamma_+ \cup \Gamma_N} \boldsymbol{b} \cdot \boldsymbol{n}\, u\varphi \,\mathrm{d}S,$$

$$\ell(\varphi) := \int_\Omega f\varphi \,\mathrm{d}x + \int_{\Gamma_N} g_N \varphi \,\mathrm{d}S, \qquad u, \varphi \in H^1(\Omega).$$

The well-posedness of the boundary value problem (2.2), in the case of homogeneous boundary conditions, is shown in Houston et al. [2000]. The well-posedness of

(2.2) with the non-homogeneous Dirichlet boundary condition can be proved by the standard technique based on the decomposition $u = u_0 + u^*$ where $u_0 \in H_0^1(\Omega)$ and solving $a(u_0, \varphi) = \ell(\varphi) - a(u^*, \varphi)\ \forall \varphi \in H_0^1(\Omega)$.

2.1.1 DG discretization of the problem

For the DG discretization we introduce a partition $\mathscr{T}_h$ covering $\overline{\Omega}$ consisting of finite number of closed d-dimensional simplices K with mutually disjoint interiors. The boundary of the element $K \in \mathscr{T}_h$ will be denoted by ∂K, its diameter $h_K = \mathrm{diam}(K)$ and $|K|$ its d-dimensional Lebesgue measure.

By $\mathscr{F}_h$ we denote the union of all faces contained in the partition $\mathscr{T}_h$ and $\mathscr{F}_h^I$, $\mathscr{F}_h^D$ the union of interior and Dirichlet boundary faces, respectively. Further, let $\mathscr{F}_h^{ID} := \mathscr{F}_h^I \cup \mathscr{F}_h^D$. For each face $\Gamma \subset \mathscr{F}_h^I$ there exist two neighboring elements $K_L, K_R \in \mathscr{T}_h$ such that $\Gamma = K_L \cap K_R$. It is possible to define a unit normal vector $\boldsymbol{n} = (n_1, \dots, n_d)$ at almost every point of $\mathscr{F}_h$. The orientation of $\boldsymbol{n}$ can be chosen arbitrarily for the interior faces, so we can assume that $\boldsymbol{n} = \boldsymbol{n}_{K_L} = -\boldsymbol{n}_{K_R}$. Further, for $K \in \mathscr{T}_h$ we set

$$\partial K^- := \{x \in \partial K; \boldsymbol{b} \cdot \boldsymbol{n}(x) < 0\}, \quad \partial K^+ := \{x \in \partial K; \boldsymbol{b} \cdot \boldsymbol{n}(x) \geq 0\}. \tag{2.3}$$

We assume that there exists $h_0 > 0$ such that $\{\mathscr{T}_h\}_{h \in (0,h_0)}$ is a system of triangulations is *shape-regular* and *locally quasi-uniform*, see Dolejší and Feistauer [2015]. We do not require the conforming properties known from finite element methods. Therefore the triangulations $\mathscr{T}_h$ could contain so called *hanging nodes*. Over the triangulation $\mathscr{T}_h$ we define the so-called *broken Sobolev space*, c.f. (1.7), over the triangulation $\mathscr{T}_h$ as

$$H^s(\Omega, \mathscr{T}_h) = \{v \in L^2(\Omega), v\big|_K \in H^s(K) \forall K \in \mathscr{T}_h\} \tag{2.4}$$

with the norm and the semi-norm $\|v\|_{H^s(\Omega,\mathscr{T}_h)} = \left(\sum_{K \in \mathscr{T}_h} \|v\|^2_{H^s(K)}\right)^{\frac{1}{2}}$ and $|v|_{H^s(\Omega,\mathscr{T}_h)} = \left(\sum_{K \in \mathscr{T}_h} |v|_{H^s(K)}{}^2\right)^{\frac{1}{2}}$, respectively.

Discontinuous Galerkin method is very convenient for hp-adaptation. Therefore, to each $K \in \mathscr{T}_h$ we assign its local polynomial degree p_K. Then we define set $\boldsymbol{p} := \{p_K; K \in \mathscr{T}_h\}$ and the finite dimensional space

$$S_h^{\boldsymbol{p}} = \{v \in L^2(\Omega); v\big|_K \in P^{p_K}(K) \forall K \in \mathscr{T}_h\}. \tag{2.5}$$

The dimension of $S_h^{\boldsymbol{p}}$ corresponding to the number of degrees of freedom can be calculated as $N_h^{\boldsymbol{p}} := \mathrm{dim} S_h^{\boldsymbol{p}} = \sum_{K \in \mathscr{T}_h} \binom{p_k+d}{d}$. The pair $\{\mathscr{T}_h, \boldsymbol{p}\} =: \mathscr{T}_{h,\boldsymbol{p}}$ is called the *hp-mesh*.

Let $\Gamma \subset \mathscr{F}_h^I$, $v \in H^1(\Omega, \mathscr{T}_h)$, we introduce the notation $v_L =$ trace of $v\big|_{K_L}$ on Γ, and $v_R =$ trace of $v\big|_{K_R}$ on Γ, Further, we denote the jump of v on Γ by $[\![v]\!] = v_L - v_R$ and its mean value $\langle v \rangle = \frac{1}{2}(v_L + v_R)$. On $\Gamma \subset \mathscr{F}_h^D$ we set $[\![v]\!] = \langle v \rangle = v_L$, where K_L is such element, that $\Gamma = K_L \cap \partial\Omega$. Given an element $K \in \mathscr{T}_h$ we denote by v^- the exterior trace of v defined on $\partial K \backslash \partial\Omega$, the interior trace on ∂K will be denoted simply by v.

We discretize the equation (2.2) using the interior penalty Galerkin method (IPG), see, e.g., [Dolejší and Feistauer, 2015, Section 4.6] or Houston et al. [2002]. For

$u, \varphi \in H^2(\Omega, \mathscr{T}_h)$ we define the forms

$$A_h(u,\varphi) := \sum_{K \in \mathscr{T}_h} \int_K \mathbb{A}\nabla u \cdot \nabla \varphi \, \mathrm{d}x \tag{2.6a}$$

$$- \sum_{\Gamma \in \mathscr{F}_h^{ID}} \int_\Gamma \langle \mathbb{A}\nabla u \rangle \cdot \boldsymbol{n} \llbracket \varphi \rrbracket + \theta \langle \mathbb{A}\nabla \varphi \rangle \cdot \boldsymbol{n} \llbracket u \rrbracket \, \mathrm{d}S,$$

$$J_h^\delta(u,\varphi) := \sum_{\Gamma \in \mathscr{F}_h^{ID}} \int_\Gamma \delta \llbracket u \rrbracket \llbracket \varphi \rrbracket \, \mathrm{d}S, \tag{2.6b}$$

$$B_h(u,\varphi) := \sum_{K \in \mathscr{T}_h} \Big(- \int_K u \boldsymbol{b} \cdot \nabla \varphi - c u \varphi \, \mathrm{d}x \tag{2.6c}$$

$$+ \int_{\partial K^+} \boldsymbol{b} \cdot \boldsymbol{n}_K u \varphi \, \mathrm{d}S + \int_{\partial K^- \setminus \partial\Omega} \boldsymbol{b} \cdot \boldsymbol{n}_K u^- \varphi \, \mathrm{d}S \Big)$$

$$\ell_h(\varphi) := \int_\Omega f \varphi \, \mathrm{d}x + \int_{\Gamma_N} g_N \varphi \, \mathrm{d}S - \sum_{K \in \mathscr{T}_h} \int_{\partial K^- \cap \partial\Omega} (\boldsymbol{b} \cdot \boldsymbol{n}) u_D \varphi \, \mathrm{d}S \tag{2.6d}$$

$$+ \sum_{\Gamma \in \mathscr{F}_h^D} \int_\Gamma (\delta \varphi - \theta \mathbb{A} \nabla \varphi \cdot \boldsymbol{n}) \, u_D \, \mathrm{d}S.$$

The choice of $\theta \in \{-1, 0, 1\}$ leads to the nonsymmetric (NIPG), incomplete (IIPG), and symmetric (SIPG) variant of the discontinuous Galerkin method. The penalty parameter δ is chosen by $\delta\big|_\Gamma = \delta_\Gamma = \frac{\varepsilon C_W p_\Gamma^2}{h_\Gamma}, \Gamma \in \mathscr{F}_h^{ID}$, where ε denotes the amount of diffusivity ($\approx |\mathbb{A}|$), $h_\Gamma = \mathrm{diam}\Gamma$, $p_\Gamma = \max(p_K, p_{K'})$ for $\Gamma \subset K \cap K'$ and $C_W > 0$ has to be chosen large enough to guarantee convergence of the method, see [Dolejší and Feistauer, 2015, Chapter 2]. Further, we introduce the DG-norm, c.f. Chapter 1.43,

$$|||v||| := \sum_{K \in \mathscr{T}_h} \Big(\left\| \mathbb{A}^{1/2} \nabla v \right\|_K^2 + \frac{1}{2} \|v\|^2_{\partial K^- \cap (\Gamma_D \cup \Gamma_-)} + \frac{1}{2} \|\llbracket v \rrbracket\|^2_{\partial K^- \setminus \partial\Omega} + \frac{1}{2} \|v\|^2_{\partial K^+ \cap \partial\Omega}$$

$$+ \|c_0 v\|_K^2 \Big) + \int_{\Gamma \in \mathscr{F}_h^{ID}} \left(\delta \llbracket v \rrbracket^2 + \frac{1}{\delta} \langle \mathbb{A} \nabla v \cdot \boldsymbol{n} \rangle^2 \right) \mathrm{d}S. \tag{2.7}$$

We use the convention that the edges Γ where $\boldsymbol{n}^T \mathbb{A} \boldsymbol{n} = 0$ are omitted from the integration in the form $J_h^\delta(\cdot,\cdot)$ and in the DG-norm.

Finally, we put

$$a_h(u,\varphi) := A_h(u,\varphi) + J_h^\delta(u,\varphi) + B_h(u,\varphi), \qquad u, \varphi \in H^2(\Omega, \mathscr{T}_h). \tag{2.8}$$

We are ready to define the discrete problem.

Definition 2.2. *We say that $u_h \in S_h^{\boldsymbol{p}}$ is the* discrete solution *of* (2.2) *obtained by discontinuous Galerkin method if*

$$a_h(u_h, \varphi_h) = \ell_h(\varphi_h) \quad \forall \varphi_h \in S_h^{\boldsymbol{p}}. \tag{2.9}$$

Lemma 2.3. *The discrete problem* (2.9) *is* consistent *with the weak formulation* (2.2)*, i.e., the exact solution $u \in H^2(\Omega)$ satisfies*

$$a_h(u,\varphi) = \ell_h(\varphi) \quad \forall \varphi \in H^2(\Omega, \mathscr{T}_h). \tag{2.10}$$

Proof. See, e.g., [Dolejší and Feistauer, 2015, Chapters 2 and 3], Harriman et al. [2003]. □

That gives us the Galerkin orthogonality of the exact and discrete solutions, c.f. (1.10),

$$a_h(u-u_h,\varphi_h)=0 \qquad \forall\varphi_h\in S_h^p, \tag{2.11}$$

which is a crucial property in goal-oriented estimates.

2.1.2 Quantity of interest

The goal of the whole computation process is to determine the value of the *quantity of interest* $J(u)$, where J is a linear functional defined for the weak as well as the approximate solutions. It was shown in Hartmann [2007], that the primal problem (2.1), the corresponding adjoint problem and target functional $J(u)$ have to satisfy the so-called *compatibility condition* which together with the *consistency* of the numerical method and the *adjoint consistency* guarantee the regularity of the adjoint solution and then the optimal order of convergence. The low regularity of the solution of the adjoint problem, c.f. Example 1.3.3, causes a suboptimal convergence rate of the DWR error estimate, see Hartmann [2007], Harriman et al. [2004].

We consider the functional J in the form

$$J(u)=\int_\Omega j_\Omega(x)u(x)\,\mathrm{d}x+\int_{\Gamma_D} j_{\Gamma_D}\mathbb{A}\nabla u\cdot\boldsymbol{n}\,\mathrm{d}S+\int_{\Gamma_+\cup\Gamma_N} j_{\Gamma_N}u\,\mathrm{d}S, \tag{2.12}$$

where $j_{\Gamma_D}, j_{\Gamma_N}\in L^2(\partial\Omega)$ and $j_\Omega\in L^2(\Omega)$ are given functions, typically characteristic functions of some subdomain in $\partial\Omega$ or Ω, respectively.

2.1.3 Derivation of the adjoint problem

We derive the adjoint problem to the problem (2.1) with the target functional given by (2.12) and verify the compatibility of the target function with the primal problem (2.1). Recalling the reasoning of the Section 1.1.9 we multiply the left-hand side of equation (2.1a) by a function $z\in H^2(\Omega)$ and integrate by parts over Ω

$$\begin{aligned}\int_\Omega \mathscr{L}uz\,\mathrm{d}x&=\int_\Omega -\nabla\cdot\mathbb{A}\nabla uz+\nabla\cdot(\boldsymbol{b}u)z+cuz\,\mathrm{d}x \\ &=\int_\Omega(\mathbb{A}\nabla u)\cdot\nabla z-\boldsymbol{b}\cdot\nabla zu+cuz\,\mathrm{d}x+\int_{\partial\Omega}-(\mathbb{A}\nabla u)\cdot\boldsymbol{n}z+(\boldsymbol{b}\cdot\boldsymbol{n})uz\,\mathrm{d}S \\ &=\int_\Omega\nabla\cdot(\mathbb{A}\nabla z)u-\boldsymbol{b}\cdot\nabla zu+cuz\,\mathrm{d}x \\ &\quad+\int_{\partial\Omega}-(\mathbb{A}\nabla u)\cdot\boldsymbol{n}z+(\boldsymbol{b}\cdot\boldsymbol{n})uz+(\mathbb{A}\nabla z)\cdot\boldsymbol{n}u\,\mathrm{d}S\end{aligned} \tag{2.13}$$

We directly see that the adjoint operator to $\mathscr{L}$ is defined by

$$\mathscr{L}^*z=-\nabla\cdot\mathbb{A}\nabla z-\boldsymbol{b}\cdot\nabla z+cz.$$

The boundary operator $\mathscr{B}^*$ has to be derived separately on each individual part of the boundary according to (1.60). Recalling the notation from Section 1.1.9 we have

$$\int_{\partial\Omega}\mathscr{A}_1z\mathscr{A}_2u\,\mathrm{d}S=\int_{\partial\Omega}-(\mathbb{A}\nabla u)\cdot\boldsymbol{n}z+(\boldsymbol{b}\cdot\boldsymbol{n})uz+(\mathbb{A}\nabla z)\cdot\boldsymbol{n}u\,\mathrm{d}S. \tag{2.14}$$

On Γ_D we have $\mathscr{B}u = u, \mathscr{C}u = \mathbb{A}\nabla u \cdot \boldsymbol{n}$ and hence $\mathscr{B}^* z = -z$. On the Neumann part of the boundary Γ_N we have $\mathscr{B}u = \mathbb{A}\nabla u \cdot \boldsymbol{n}, \mathscr{C}u = u$ and hence for the boundary condition of the adjoint problem we get $\mathscr{B}^* z = \mathbb{A}\nabla z \cdot \boldsymbol{n} + \boldsymbol{b} \cdot \boldsymbol{n} z$. On Γ_+ we have $\mathscr{B}u = 0, \mathscr{C}u = u$ and hence $\mathscr{B}^* z = \boldsymbol{b} \cdot \boldsymbol{n} z$ and finally $\mathscr{B}^* = 0$ on Γ_- since there holds $\mathscr{B}u = u$ and J vanishes there.

Therefore target functional (2.12) is compatible with the equation (2.1) and the corresponding adjoint problem reads in its strong formulation: Find a function $z : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} -\nabla \cdot \mathbb{A}\nabla z - \boldsymbol{b} \cdot \nabla z + cz = j_\Omega \quad & \text{in } \Omega, \\ z = -j_{\Gamma_D} \quad & \text{on } \Gamma_D, \\ \mathbb{A}\nabla z \cdot \boldsymbol{n} + \boldsymbol{b} \cdot \boldsymbol{n} z = j_{\Gamma_N} \quad & \text{on } \Gamma_N, \\ \boldsymbol{b} \cdot \boldsymbol{n} z = j_{\Gamma_N} \quad & \text{on } \Gamma_+. \end{aligned} \tag{2.15}$$

The adjoint problem (2.15) contains a Newton boundary condition on Γ_N, but since $\boldsymbol{b} \cdot \boldsymbol{n} \geq 0$ on Γ_N this boundary condition will contribute to the coercivity of the problem and the problem is well-posed.

Definition 2.4. *We say that a function $z_h \in S_h^{\boldsymbol{p}}$ is the* discrete adjoint solution *if it satisfies*

$$a_h(\psi_h, z_h) = J(\psi_h) \qquad \forall \psi_h \in S_h^{\boldsymbol{p}}. \tag{2.16}$$

We recall that a discretization (2.9) of the problem (2.2) with the target function J given by (2.12) is *adjoint consistent* if the exact solution $z \in H^2(\Omega)$ of (2.15) satisfies (2.16), i.e.

$$a_h(\psi, z) = J(\psi) \qquad \forall \psi \in H^2(\Omega, \mathscr{T}_h). \tag{2.17}$$

2.1.4 Adjoint consistency

In the following, we deal with the adjoint consistency of the discrete adjoint problem (2.16). We show that in order to guarantee the adjoint consistency, the right-hand side of (2.16) has to be slightly modified.

Following the approach from Hartmann [2007] we rewrite (2.16) element-wise and by integration by parts and the definition of the forms (2.6) we get that the solution of (2.16) satisfies

$$\begin{aligned} r_h^*(z_h)(\psi_h) = \sum_{K \in \mathscr{T}_h} \int_K & r_{K,\mathrm{V}}^*(z_h) \psi_h \,\mathrm{d}x \\ & + \int_{\partial K} r_{K,\mathrm{B}}^*(z_h) \psi_h + r_{K,\mathrm{D}}^*(z_h) \nabla \mathbb{A} \psi_h \cdot \boldsymbol{n} \,\mathrm{d}S = 0 \qquad \forall \psi_h \in S_h^{\boldsymbol{p}}, \end{aligned} \tag{2.18}$$

where the adjoint residuals consist of

$$r_{K,\mathrm{V}}^*(z_h) := j_\Omega + \nabla \cdot (\mathbb{A}\nabla z_h) + \boldsymbol{b} \cdot \nabla z_h - c z_h, \tag{2.19}$$

$$r^*_{K,\mathrm{B}}(z_h) := \begin{cases} -\frac{1}{2}[\![\mathbb{A}\nabla z_h]\!]\boldsymbol{n} + (1-\theta)\,\langle \mathbb{A}\nabla z_h\rangle \cdot \boldsymbol{n} - (\delta \boldsymbol{n}\cdot\boldsymbol{n}_K + \boldsymbol{b}\cdot\boldsymbol{n})[\![z_h]\!] & \\ & \text{on } \partial K^+ \setminus \partial\Omega, \\ -\frac{1}{2}[\![\mathbb{A}\nabla z_h]\!]\boldsymbol{n} + (1-\theta)\,\langle \mathbb{A}\nabla z_h\rangle \cdot \boldsymbol{n} - \delta \boldsymbol{n}\cdot\boldsymbol{n}_K[\![z_h]\!] & \\ & \text{on } \partial K^- \setminus \partial\Omega, \\ -(1-\theta)\mathbb{A}\nabla z_h\cdot\boldsymbol{n} - \delta z_h & \text{on } \partial K^- \cap \Gamma_D, \\ -(1-\theta)\mathbb{A}\nabla z_h\cdot\boldsymbol{n} - (\delta + \boldsymbol{b}\cdot\boldsymbol{n}) z_h & \text{on } \partial K^+ \cap \Gamma_D, \\ j_{\Gamma_N} - \mathbb{A}\nabla z_h\cdot\boldsymbol{n} - \boldsymbol{b}\cdot\boldsymbol{n} z_h & \text{on } \partial K \cap \Gamma_N, \\ j_{\Gamma_N} - \boldsymbol{b}\cdot\boldsymbol{n} z_h & \text{on } \partial K \cap \Gamma_+, \\ 0 & \text{on } \partial K \cap \Gamma_-, \end{cases}$$

$$r^*_{K,\mathrm{D}}(z_h) := \begin{cases} \frac{1}{2}[\![z_h]\!] & \text{on } \partial K \setminus \partial\Omega, \\ j_{\Gamma_D} + z_h & \text{on } \partial K \cap \Gamma_D, \\ 0 & \text{on } \partial K \cap (\partial\Omega \setminus \Gamma_D). \end{cases}$$

Considering the symmetric variant of DG ($\theta = 1$) we see that if $z \in H^2(\Omega)$ is the solution of the problem (2.15), it nullifies the volume residual $r^*_{K,\mathrm{V}}$ and also all residuals on interior edges and boundary edges except Γ_D. On Γ_D we have $z = -j_{\Gamma_D}$ from $r^*_{K,\mathrm{D}}$, but also $\delta z + \boldsymbol{b}\cdot\boldsymbol{n} z = 0$ on $\partial K^+ \cap \Gamma_D$ and $\delta z = 0$ on $\partial K^- \cap \Gamma_D$, which are in conflict unless $j_{\Gamma_D} = 0$.

This problem can be overcome by a small modification of the target functional according to the method introduced in (1.61). We define

$$\mathscr{N}(u_h) := \int_{\Gamma_D} r_J(v)\,\mathrm{d}S, \quad r_J(v) = \begin{cases} -\delta(v - u_D) j_{\Gamma_D} & \text{on } \partial K^- \cap \Gamma_D \\ -(\delta + \boldsymbol{b}\cdot\boldsymbol{n})(v - u_D) j_{\Gamma_D} & \text{on } \partial K^+ \cap \Gamma_D \end{cases} \tag{2.20}$$

and then

$$\tilde{J}(v) := J(v) + \mathscr{N}(u_h), \tag{2.21}$$

The modification is designed such that $\tilde{J}(u) = J(u)$ for u being the exact solution of the problem (2.1). Further, since $\tilde{J}(v)$ is affine, it holds $\tilde{J}(u) - \tilde{J}(u_h) = \tilde{J}'_u(u - u_h)$, where

$$\tilde{J}'_u(v) = J(v) - \sum_{K\in\mathscr{T}_h} \left(\int_{\Gamma_D \cap \partial K^-} v \delta j_{\Gamma_D}\,\mathrm{d}S - \int_{\Gamma_D \cap \partial K^+} v(\delta + \boldsymbol{b}\cdot\boldsymbol{n}) j_{\Gamma_D}\,\mathrm{d}S \right) \tag{2.22}$$

is the Gateaux derivative of $\tilde{J}$ in direction v. In order to guarantee the adjoint consistency of the adjoint problem, we can replace the adjoint problem (2.16) by

$$a_h(\psi_h, z_h) = \tilde{J}'_u(\psi_h) \qquad \forall \psi_h \in S_h^{\boldsymbol{p}}. \tag{2.23}$$

From definition of the modification of the target functional we deduce that

$$\begin{aligned} J(u) - J(u_h) &= \mathscr{N}(u_h) + \tilde{J}'_u(e_h) \\ &= \mathscr{N}(u_h) + a_h(e_h, z_h) \end{aligned} \tag{2.24}$$

So the term $\mathscr{N}(u_h)$ may be viewed as the "amount of violation" of the Dirichlet boundary condition.

All the derivations presented in Subsection 2.1.4 can be summarized into the following result.

Lemma 2.5. *The SIPG variant of* (2.23) *is an* adjoint consistent *discretization of the problem* (2.1) *with target functionals defined according to* (2.21)*,i.e.,*

$$r_h^*(z)(\psi) = 0 \qquad \forall \psi \in H^2(\Omega, \mathscr{T}_h) \tag{2.25}$$

Moreover, it provides the Galerkin orthogonality for the adjoint solutions z and z_h

$$a_h(\psi_h, z - z_h) = 0 \quad \forall \psi_h \in S_h^p. \tag{2.26}$$

On the other hand for nonsymmetric variants ($\theta \in \{-1, 0\}$) the adjoint discretization is surely not adjoint consistent with (2.15) due to $\langle \mathbb{A} \nabla z \rangle \neq 0$ in (2.19). Therefore we limit our further steps only for SIPG variant.

In the following, we will restrict ourselves for the sake of simplicity to the case when $j_{\Gamma_D} = 0$, i.e., when the modification of J is not needed, but all our further derivations may be easily generalized also to the case when $j_{\Gamma_D} \neq 0$ using the relation (2.24).

2.1.5 Goal-oriented error estimates

In the framework presented in Section 1.1.3 we introduce the goal-oriented error estimates. Using the adjoint consistency (2.17), the consistency (2.10), the Galerkin orthogonality of the error (2.11), we get the *primal error identity* (c.f. (1.16)) for the error of the quantity of interest

$$\begin{aligned} J(e_h) &= a_h(u - u_h, z) = \ell_h(z) - a_h(u_h, z) =: r_h(u_h)(z) \\ &= r_h(u_h)(z - \varphi_h) \quad \forall \varphi_h \in S_h^p \end{aligned} \tag{2.27}$$

where $r_h(u_h)(\cdot)$ denotes the residual of the problem (2.9). Let us note that the Galerkin orthogonality was used only in the last step, i.e, the identity $J(u - u_h) = r_h(u_h)(z)$ is valid also for u_h violating the Galerkin orthogonality, which is the case of the approximate solution suffering from algebraic errors.

Similarly, exploiting (2.26) we get the *adjoint error identity* (c.f. (1.20))

$$\begin{aligned} J(u - u_h) &= a_h(u - u_h, z - z_h) = a_h(u - \psi_h, z - z_h) \\ &= J(u - \psi_h) - a_h(u - \psi_h, z_h) \\ &=: r_h^*(z_h)(u - \psi_h) \quad \forall \psi_h \in S_h^p, \end{aligned} \tag{2.28}$$

where $r_h^*(z_h)(\cdot)$ denotes the residual of the adjoint problem (2.16).

Hence, similarly to (1.21) the residuals $r_h(u_h)(\cdot)$ and $r_h^*(z_h)(\cdot)$ are equivalent in the following way

$$r_h(u_h)(z - \varphi_h) = r_h^*(z_h)(u - \psi_h) \quad \forall \varphi_h, \psi_h \in S_h^p. \tag{2.29}$$

2.2 Reconstruction of the discrete solutions

Except for a very few examples, neither u nor z are a priori known. Therefore, they must replaced by some computable quantities in (2.27) and (2.28). Following the concept from Section 1.1.5 we define

$$\eta_S := r_h(u_h)(z_h^+ - \Pi z_h^+), \qquad \eta_S^* := r_h^*(z_h)(u_h^+ - \Pi u_h^+), \tag{2.30}$$

where $\Pi : L^2(\Omega) \to S_h^p$ is an arbitrary projection to S_h^p. Obviously the functions z_h^+ and u_h^+ must be from a richer space than S_h^p otherwise the residuals would degenerate, since $r_h(u_h)(\varphi_h) = r_h^*(z_h)(\varphi_h) = 0$ for all $\varphi_h \in V_h$, c.f. Section 1.1.5.

As in (1.35) we get the following equality for the error (primal formulation)

$$\begin{aligned} J(u-u_h) = r_h(u_h)(z-\phi_h) &= r_h(u_h)(z_h^+ - \phi_h) + r_h(u_h)(z - z_h^+) \\ := \eta_h + \varepsilon_h \qquad & \forall \phi_h \in S_h^p, \end{aligned} \tag{2.31}$$

where the second term is usually neglected, see Section 1.1.7. Motivated by the equality (2.29), c.f. (1.38), we employ the arithmetic average of η_S and η_S^* and define the following a posteriori error estimate

$$J(e_h) \approx \eta^\mathrm{I} := \frac{1}{2}\left(\eta_\mathrm{S} + \eta_\mathrm{S}^*\right). \tag{2.32}$$

In order to localize (2.30) into positive error indicators describing local error contributions we proceed according to the Section 1.1.6 and introduce element-wise contributions of (2.30)

$$\eta_\mathrm{S,K} = r_h(u_h)((z_h^+ - \phi_h)\chi_K), \quad \eta_\mathrm{S,K}^* = r_h^*(z_h)((u_h^+ - \varphi_h)\chi_K), \quad K \in \mathscr{T}_h, \tag{2.33}$$

where χ_K denotes the characteristic function of element K. Further, we define

$$\eta_K^\mathrm{I} := \frac{1}{2}\left(\eta_\mathrm{S,K} + \eta_\mathrm{S,K}^*\right). \tag{2.34}$$

Evidently, $\eta^\mathrm{I} = \sum_{K\in\mathscr{T}_h} \eta_K^\mathrm{I}$.

The absolute value of $|\eta_K^\mathrm{I}|$ can be used as a local error indicator for mesh refinement, namely the setting of the sizes in the new mesh, the elements with too high value of $|\eta_K^\mathrm{I}|$ (or $|\eta_\mathrm{S,K}|$ or $|\eta_\mathrm{S,K}^*|$, if only one kind of the estimates is used) are refined whereas the element with too small $|\eta_K^\mathrm{I}|$ may be coarsened. We note that $\eta_\mathrm{S} \neq \sum_{K\in\mathscr{T}_h} |\eta_\mathrm{S,K}|$ and the sum of the indicators may strongly overestimate the error $J(e_h)$, c.f. Section 1.1.6.

We present two local reconstructions of the discrete solutions applicable for DG of an arbitrary degree (even hp-variant). Both of these methods do not require any patch-wise structure of the mesh. This is very favorable since we aim for the combination of the goal-oriented estimates with the anisotropic mesh generator Dolejší [2000]. We present the ideas for reconstruction of the discrete solution u_h, computation of z_h^+ is done alike using function z_h. Another standard way of obtaining u_h^+, z_h^+ is to compute the adjoint problem on a finer mesh and/or with higher polynomial degree as introduced in Section 1.1.5.

2.2.1 Weighted least-square method

First, we employ the method developed in Dolejší and Solin [2016]. For the purpose of the presented reconstruction we define the space

$$S_h^{p+1} := \{v \in L^2(\Omega); v\big|_K \in P^{p_k+1}(K) \forall K \in \mathscr{T}_h.\}$$

Obviously $S_h^p \subset S_h^{p+1} \subset H^2(\Omega, \mathscr{T}_h)$.

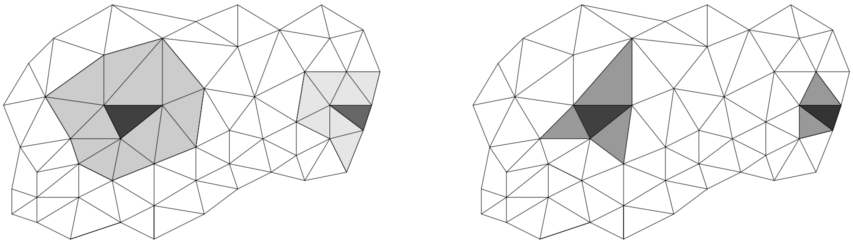

Figure 2.2: Examples of patches $\mathscr{D}_K$ corresponding to interior and boundary elements, large (left) and small (right) patches.

Let $u_h \in S_h^p$ be the approximate solution of (2.9). For the reconstruction $u_h^+ \in S_h^{p+1}$ on element $K \in \mathscr{T}_h$ we use a weighted least square approximation from the elements sharing at least a vertex with K, see Figure 2.2, left. We denote this patch of elements $\mathscr{D}_K = \{K' \in \mathscr{T}_h; K' \cap K \neq \emptyset\}$.

We compute the function $u_K^+ \in P^{p_K+1}(\mathscr{D}_K)$ by

$$u_K^+ = \operatorname*{arg\,min}_{U_h \in P^{p_K+1}(\mathscr{D}_K)} \sum_{K' \in \mathscr{D}_K} \omega_{K'} \|U_h - u_h\|^2_{H^1(K')} . \tag{2.35}$$

Then we assemble the higher-order reconstruction u_h^+ as an element-wise composition of $u_K^+|_K$, i.e., $u_h^+ = \sum_{K \in \mathscr{T}_h} u_K^+|_K$. In the following we will refer to this method as the LS reconstruction.

When choosing the values of the weights $\omega_{K'}$, we distinguish between elements sharing a face and elements having only a common vertex. We set $\omega_{K'} = 1$ if $K' = K$ or if K, K' share a face and $\omega_{K'} = \varepsilon$ if K, K' share only a vertex. The parameter ε helps to stabilize the reconstruction when local polynomial degrees are too varying on $\mathscr{D}_K$. Hence we choose

$$\varepsilon := \overline{\varepsilon} \max(0, \Delta p_K - 1), \text{ where } \Delta p_k = \max_{K' \in \mathscr{D}_K} p_{K'} - \min_{K' \in \mathscr{D}_K} p_{K'} \tag{2.36}$$

where $\overline{\varepsilon} := 0.02$ was empirically chosen. Consequently, the small patches, see Figure 2.2 right, are used when $\Delta p_k \leq 1$.

This method is actually independent of the solved problem. This can be viewed as a disadvantage since approximation tailored specifically for the solved problem may work more accurately, but on the other hand such specialized technique may not be available for complex problems.

As shown in Dolejší et al. [2017], this reconstruction can be advantageously used also to determine the anisotropic hp-adaptation of the mesh. Although we cannot prove theoretically that $\|u - u_h\| \approx \left\|u_h^+ - u_h\right\|$, it was numerically verified on several examples in Dolejší and Solin [2016].

2.2.2 Solving local problems

Another common method for computing a reconstruction u_h^+ in FEM computations is based on the solution of local problems defined on patches of elements, see Babuska and Rheinboldt [1978], Bank and Weiser [1985]. For conforming FEM applied to Poisson problem ($\mathscr{L} := -\Delta$) the authors of Babuska and Rheinboldt [1978] suggest to solve the auxiliary problems

$$\mathscr{L}u_i^+ = f \quad \text{in } \Omega_i := \operatorname{supp}\psi_i, \quad u_i^+ = u_h \text{ on } \partial\Omega_i, \tag{2.37}$$

where $\{\psi_i\}_{i=1}^M$ is a partition of unity satisfying $\sum_{i=1}^M \psi(x) = 1$ for all $x \in \Omega$ and each $\psi_i \geq 0$.

For solution of (2.37) we propose to employ again the DG method, which includes the Dirichlet boundary condition only by the penalty terms. Since no inter-element continuity is required in DG, we can define these problems even element-wise setting simply $\psi_i := \chi_K$, $K \in \mathscr{T}_h$, where χ_K is the characteristic function of the element K. Namely, for each $K \in \mathscr{T}_h$ we define the function $u_K^+ : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} &\text{(i)} \quad u_K^+|_{K'} := u_h|_{K'} \text{ for all } K' \neq K, \\ &\text{(ii)} \quad u_K^+|_K \in P^{p_K+1}(K), \\ &\text{(iii)} \quad a_h(u_K^+, \varphi_h) = \ell_h(\varphi_h) \quad \forall \varphi_h \in P^{p_K+1}(K), \end{aligned} \tag{2.38}$$

where a_h is the form given by (2.8). Since evidently $u_K^+ \in S_h^{p+1}$, we finally define $u_h^+ \in S_h^{p+1}$ by $u_h^+|_K := u_K^+ \quad \forall K \in \mathscr{T}_h$. In the rest of the paper we will refer to this kind of reconstruction as the LOC reconstruction.

In the following we show that it is not necessary to assemble and to solve problem (2.38) for each K explicitly, when we use the residual based approach from Dolejší [2013]. We denote $N_K = (p_K+1)(p_K+2)/2$ the number of degrees of freedom attached to the element $K \in \mathscr{T}_h$ and $\boldsymbol{\varphi}_K = \{\varphi_{h,K}^i\}_{i=1}^{N_K}$ the basis of the space $P^{p_K}(K)$. The basis of S_h^p, denoted by $\boldsymbol{\varphi} = \{\varphi_h^i\}_{i=1}^{N_h^p}$, $N_h^p = \dim S_h^p$, can be assembled by the functions from $\boldsymbol{\varphi}_K$ for all $K \in \mathscr{T}_h$ extended by zero outside K. Due to the discontinuity of the functions in S_h^p across the element edges, we can write u_h in the element-wise components $\boldsymbol{u}_K \in \mathbb{R}^{N_K}$ corresponding to $K \in \mathscr{T}_h$, i.e.,

$$u_h = \sum_{i=1}^{N_h^p} U^i \varphi_h^i = \sum_{K \in \mathscr{T}_h} \boldsymbol{u}_K \cdot \boldsymbol{\varphi}_K.$$

Denoting $\boldsymbol{f}_K := \{\ell_h(\varphi_{h,K}^i)\}_{i=1}^{N_K}$, the problem (2.9) can be rewritten in the block-matrix form (one block-row for each $K \in \mathscr{T}_h$)

$$\mathbb{A}_{K,K}\boldsymbol{u}_K + \sum\nolimits_{K' \in N(K)} \mathbb{A}_{K,K'}\boldsymbol{u}_K' = \boldsymbol{f}_K \quad \forall K \in \mathscr{T}_h, \tag{2.39}$$

where $\mathbb{A}_{K,K}$ are diagonal blocks (corresponding to a_h) of size $N_K \times N_K$, $\mathbb{A}_{K,K'}$ are the off-diagonal blocks of size $N_K \times N_{K'}$ and $N(K)$ is the set of elements sharing an edge with $K \in \mathscr{T}_h$.

For each $K \in \mathscr{T}_h$, we can write $u_K^+ = u_h + \tilde{u}_K$, where u_h is the approximate solution given by (2.9) and $\tilde{u}_K \in P^{p_K+1}(K)$ can be considered as a local higher order correction. Obviously, due to condition (i) in (2.38), we have $\tilde{u}_K = 0$ on all $K' \neq K$, $K' \in \mathscr{T}_h$.

Let $\varphi_{h,K} \in P^{p_K+1}(K)$. Using the linearity of a_h, condition (iii) in (2.38) and (2.27), we have

$$\begin{aligned} a_h(\tilde{u}_K, \varphi_{h,K}) &= a_h(u_K^+, \varphi_{h,K}) - a_h(u_h, \varphi_{h,K}) = \ell_h(\varphi_{h,K}) - a_h(u_h, \varphi_{h,K}) \\ &= r_h(u_h)(\varphi_{h,K}). \end{aligned} \tag{2.40}$$

Hence we have to solve

$$a(\tilde{u}_K, \varphi_{h,K}) = r_h(u_h)(\varphi_{h,K}) \quad \forall \varphi_{h,K} \in P^{p_K+1}(K) \tag{2.41}$$

for each $K \in \mathscr{T}_h$. We denote $N_K^+ = \dim P^{p_K+1}(K) = (p_K+2)(p_K+3)/2$ and we choose a basis $\varphi_{h,K}^1, \dots, \varphi_{h,K}^{N_K}, \dots, \varphi_{h,K}^{N_K^+}$ of P^{p_K+1} as hierarchical extension of the basis $\boldsymbol{\varphi}_K$. Then (2.41) can be written in similar form to (2.39), where the off-diagonal terms are vanishing since $\tilde{u}_K = 0$ on all $K' \neq K$, namely

$$\mathbb{A}_{K,K}^+ \tilde{\boldsymbol{u}}_K = \boldsymbol{r}, \tag{2.42}$$

where $\mathbb{A}_{K,K}^+ \in \mathbb{R}^{N_K^+ \times N_K^+}$ is the matrix $\mathbb{A}_{K,K}$ enlarged by $N_K^+ - N_K$ rows and columns, $\boldsymbol{r} \in \mathbb{R}^{N_K^+}$ is the vector with components $r_i = r_h(u_h)(\varphi_{h,K}^i), i = 1, \dots, N_K^+$ and $\tilde{\boldsymbol{u}}_K$ is the vector of basis coefficients defining the function $\tilde{u}_K$ on K. Let us note that first N_K components of $\boldsymbol{r}$ are vanishing up to the algebraic errors.

Therefore, in order to find the reconstruction (2.38) for each $K \in \mathscr{T}_h$, we have to assemble the block-diagonal block $\mathbb{A}_{K,K}^+$, evaluate the residual (2.27) for all basis functions of $P^{p_K+1} \backslash P^{p_K}$ and solve the linear algebraic system (2.42). Finally, we put $u_h^+ = u_h + \sum_{K \in \mathscr{T}_h} \tilde{u}_K$.

Remark. This method can be used even for nonlinear problems, but in that case the computation of the update $\tilde{u}_K$ has to be iterated several times, c.f. Section 4.4.1.

Remark. For the reconstruction based on the solution of the local problems we have (in exact arithmetics) due to (2.9), (2.16) and (2.40) that

$$\begin{aligned} \eta_{\mathrm{S,K}} = r_h(u_h)(z_h^+|_K) = r_h(u_h)(\tilde{z}_K) = a_h(\tilde{u}_K, \tilde{z}_K) \\ = r_h^*(z_h)(\tilde{u}_K) = r_h^*(z_h)(u_h^+|_K) = \eta_{\mathrm{S,K}}^* \end{aligned} \tag{2.43}$$

Hence we get not only the global equivalence corresponding to (2.29), but even the local error indicators $\eta_{\mathrm{S,K}}$ and $\eta_{\mathrm{S,K}}^*$ are equivalent for this reconstruction.

On the contrary, the LS reconstruction is not connected with the solved problem and the error estimates η_{S} and η_{S}^* may differ both locally and globally.

Remark. We may also solve the $a_h(u,v) = r_h(u_h)(v)$ reconstruction on patches of element having one common vertex. This would be connected with the partition of unity using the piece-wise linear "hat" functions.

2.3 Algebraic errors

As already discussed in the Section 2.3 for a general linear problems, due to algebraic errors neither the "exact" discrete solution u_h of (2.9) nor the solution z_h of (2.16) are available in practical computations. Instead, we compute their approximations $u_{h,A}^{(k)}$ and $z_{h,A}^{(k)}$ resulting from a finite number of iterations of an algebraic iterative solver.

Considering the algebraically inexact discrete solution $u_{h,A}^{(k)}$ the Galerkin orthogonality (2.11) and (2.26) do not hold anymore. Hence we must add an additional term measuring the deviation from the Galerkin orthogonality due to algebraic errors. For the primal error identity (2.27) using the triangle inequality we have

$$J(u - u_{h,A}^{(k)}) = r_h(u_{h,A}^{(k)})(z) = r_h(u_{h,A}^{(k)})(z - \varphi_h) + r_h(u_{h,A}^{(k)})(\varphi_h) \quad \forall \varphi_h \in S_h^{\boldsymbol{p}}. \tag{2.44}$$

Regarding the revision of adjoint estimate (2.28) we proceed similarly. Using the definitions of residuals r_h and r_h^* in (2.27) and (2.28), respectively, and the triangle inequality, we get

$$\begin{aligned} r_h(u_{h,A}^{(k)})(z - z_{h,A}^{(k)}) &= a_h(u - u_{h,A}^{(k)}, z - z_{h,A}^{(k)}) \\ &= a_h(u - \psi_h, z - z_{h,A}^{(k)}) + a_h(\psi_h - u_{h,A}^{(k)}, z - z_{h,A}^{(k)}) \\ &= r_h^*(z_{h,A}^{(k)})(u - \psi_h) + r_h^*(z_{h,A}^{(k)})(\psi_h - u_{h,A}^{(k)}) \qquad \forall \psi_h \in S_h^{\boldsymbol{p}}. \end{aligned} \tag{2.45}$$

Then putting $\varphi_h := z_{h,A}^{(k)}$ in (2.44) and using (2.45), we obtain

$$J(u - u_{h,A}^{(k)}) = r_h^*(z_{h,A}^{(k)})(u - \psi_h) + r_h^*(z_{h,A}^{(k)})(\psi_h - u_{h,A}^{(k)}) + r_h(u_{h,A}^{(k)})(z_{h,A}^{(k)}) \quad \forall \psi_h \in S_h^{\boldsymbol{p}}. \tag{2.46}$$

The impact of algebraic errors in goal-oriented estimates was studied in Meidner et al. [2009], where the equivalence (2.29) is mentioned but only the estimates based on the primal residual are considered. Since this equivalence is not relevant for algebraically inexact solutions, we use both of these estimates and compare their accuracy in concrete computations (see Section 2.4).

The primal and adjoint part of the error identity in (2.27) can be separated, c.f. (1.43) for general linear problems or later (3.4) for the problem (2.1). Due to this separation, tightness of the estimates is strongly dependent on the choice of φ_h. Contrarily, in (2.27) the choice of φ_h is irrelevant but when those errors are taken into account as in (2.44), then the choice of φ_h may again influence the computation process. Therefore, we present three variants of (2.30)

$$\overline{\eta}_{\mathrm{S}}^{(k)} := r_h(u_{h,A}^{(k)})(z_h^+), \qquad \overline{\eta}_{\mathrm{S}}^{*,(k)} := r_h^*(z_{h,A}^{(k)})(u_h^+), \tag{2.47a}$$

$$\tilde{\eta}_{\mathrm{S}}^{(k)} := r_h(u_{h,A}^{(k)})(z_h^+ - z_{h,A}^{(k)}), \qquad \tilde{\eta}_{\mathrm{S}}^{*,(k)} := r_h^*(z_{h,A}^{(k)})(u_h^+ - u_{h,A}^{(k)}), \tag{2.47b}$$

$$\hat{\eta}_{\mathrm{S}}^{(k)} := r_h(u_{h,A}^{(k)})(z_h^+ - P_h^p z_h^+), \qquad \hat{\eta}_{\mathrm{S}}^{*,(k)} := r_h^*(z_{h,A}^{(k)})(u_h^+ - P_h^p u_h^+). \tag{2.47c}$$

Here P_h^p denotes the L^2-orthogonal projection to $S_h^{\boldsymbol{p}}$, i.e., for any $v \in L^2(\Omega)$ it satisfies $\int_\Omega P_h^p v \varphi_h \, \mathrm{d}x = \int_\Omega v \varphi_h \, \mathrm{d}x, \quad \forall \varphi_h \in S_h^{\boldsymbol{p}}$. Furthermore, we introduce the *primal* and *adjoint algebraic* error estimates

$$\overline{\eta}_{\mathrm{A}}^{(k)} = \tilde{\eta}_{\mathrm{A}}^{(k)} := r_h(u_{h,A}^{(k)})(z_{h,A}^{(k)}), \qquad \overline{\eta}_{\mathrm{A}}^{*,(k)} = \tilde{\eta}_{\mathrm{A}}^{*,(k)} := r_h^*(z_{h,A}^{(k)})(u_{h,A}^{(k)}) \tag{2.48a}$$

$$\hat{\eta}_{\mathrm{A}}^{(k)} := r_h(u_{h,A}^{(k)})(P_h^p z_h^+), \qquad \hat{\eta}_{\mathrm{A}}^{*,(k)} := r_h^*(z_{h,A}^{(k)})(P_h^p u_h^+). \tag{2.48b}$$

Since the exact specification is usually not necessary we will talk generally about $\eta_{\mathrm{A}}^{(k)} \in \{\overline{\eta}_{\mathrm{A}}^{(k)}, \tilde{\eta}_{\mathrm{A}}^{(k)}, \hat{\eta}_{\mathrm{A}}^{(k)}\}$ and $\eta_{\mathrm{A}}^{*,(k)} \in \{\overline{\eta}_{\mathrm{A}}^{*,(k)}, \tilde{\eta}_{\mathrm{A}}^{*,(k)}, \hat{\eta}_{\mathrm{A}}^{*,(k)}\}$. Let us note that if $u_{h,A}^{(k)}$ and $z_{h,A}^{(k)}$ satisfy the Galerkin orthogonality (2.11) and (2.26), respectively, then

$$\overline{\eta}_{\mathrm{S}}^{(k)} = \tilde{\eta}_{\mathrm{S}}^{(k)} = \hat{\eta}_{\mathrm{S}}^{(k)}, \qquad \overline{\eta}_{\mathrm{S}}^{*,(k)} = \tilde{\eta}_{\mathrm{S}}^{*,(k)} = \hat{\eta}_{\mathrm{S}}^{*,(k)}, \qquad \eta_{\mathrm{A}}^{(k)} = \eta_{\mathrm{A}}^{*,(k)} = 0.$$

Remark. We may express the reconstruction of the adjoint solution with respect to an orthogonal basis of the space $S_h^{\boldsymbol{p}+1}$, i.e., $z_h^+ = \sum_{j=1}^{N_h^{\boldsymbol{p}+1}} z_j \phi_j$ where $N_h^{\boldsymbol{p}+1} = \sum_{K \in \mathscr{T}_h} \binom{p_K + d}{d}$.

Then for $\varphi_h = P_h^p z$ we get

$$J(u - u_{h,A}^{(k)}) \approx r_h(u_{h,A}^{(k)})((I - P_h^p)z_h^+) + r_h(u_{h,A}^{(k)})(P_h^p z_h^+)$$

$$= \underbrace{\sum_{j=N_h^p+1}^{N_h^{p+1}} z_j r_h(u_{h,A}^{(k)})(\phi_j)}_{\text{discretization error}} + \underbrace{\sum_{j=1}^{N_h^p} z_j r_h(u_{h,A}^{(k)})(\phi_j)}_{\text{algebraic error}}. \tag{2.49}$$

Then the second term $\overline{\eta}_{\mathrm{A}}^{(k)}$ measures deviation of $u_{h,A}^{(k)}$ from u_h with respect to the target quantity while the first one $\overline{\eta}_{\mathrm{S}}^{(k)}$ measures the discretization error weighted by the oscillations of the adjoint solution of degree $p+1$. The algebraic errors represent the oscillations of the lower degrees which have more global behavior and hence may strengthen the oscillations (changing signs) of the global discretization estimate.

The reconstruction of the adjoint solution z_h^+ used in $\eta_{\mathrm{S}}^{(n)}$ is affected by algebraic errors as well. In order to take these into account in practical computations, we monitor the value of $\eta_{\mathrm{A}}^{*,(k)}$ in error estimates based on the primal error identity (2.44) too.

2.3.1 Adaptive algorithm

We recall the notation $e_h = u - u_h$ and using the error estimates (2.47) and error indicators (2.33) we propose the following adaptive algorithm.

Algorithm 2: Adaptive algorithm balancing discretization and algebraic errors

```
initialization: set $\eta = 2\mathrm{TOL}$;
while $\eta > \mathrm{TOL}$ do
    while $\eta_{\mathrm{A}}^{(k)} > C_A^{(1)}\eta_{\mathrm{S}}^{(n)}$ and $\eta_{\mathrm{A}}^{*,(k)} > C_A^{(1)}\eta_{\mathrm{S}}^{*,(n)}$ do
        perform GMRES iterations for primal problem (2.9);
        perform GMRES iterations for adjoint problem (2.16);
    end
    if $\eta_{\mathrm{A}}^{(k)} < C_A^{(1)}\eta_{\mathrm{S}}^{(n)}$ then
        perform GMRES iterations for adjoint problem until $\eta_{\mathrm{A}}^{*,(k)} < C_A^{(2)}\eta_{\mathrm{S}}^{(n)}$;
        use $\eta := \eta_{\mathrm{S}}^{(n)}$ , $\eta_K := \eta_{\mathrm{S,K}}^{(n)}$;
    else
        perform GMRES iterations for primal problem until $\eta_{\mathrm{A}}^{(k)} < C_A^{(2)}\eta_{\mathrm{S}}^{*,(n)}$;
        use $\eta := \eta_{\mathrm{S}}^{*,(n)}$, $\eta_K := \eta_{\mathrm{S,K}}^{*,(n)}$;
    end
    according to error indicators $\eta_K$ refine $\mathscr{T}_h$;
end
```

The purpose of the safety constants $C_A^{(1)}, C_A^{(2)} \leq 1$ is to suppress the impact of the algebraic errors on the discretization estimates since otherwise error indicators η_K would not produce a reasonable mesh refinement. From the numerical experiments, it seems that primal error estimate $\eta_{\mathrm{S}}^{(n)}$ is more sensitive to algebraic errors in primal problem (and vice versa for $\eta_{\mathrm{S}}^{*,(n)}$), hence we set $C_A^{(1)} = 0.01$ and $C_A^{(2)} = 0.1$, but in many numerical experiments even the value $C_A^{(2)} = 1$ leads to stable results.

The refinement of the mesh $\mathscr{T}_h$ in the step 13 is by done by refining 20% of the elements with the largest error (HG) in the following experiment (Section 2.4.1). This refinement leads to meshes with hanging nodes, see Figure 2.4. For standard FEM method the mesh adaptation with hanging nodes is a bit tricky due to the continuity of the functions in the discrete function space. Therefore hanging nodes are usually avoided for triangular meshes in this case. On the contrary, there is no continuity requirement on function from S_h^p for the DG method and hence its implementation is pretty straightforward.

A more complex mesh adaptation technique based on the *hp*-anisotropic mesh refinement strategy (AMA) will be introduced in the Chapter 3. This method balances the element error contributions η_K over the whole mesh and both size and shape of each element K are optimized. Moreover the local polynomial degrees p_K may change during adaptation cycles.

Remark. It seems temping to select the more promising of the estimates η_{S} and η_{S}^* (as early as possible) and stop computing the other one. Unfortunately, having in mind the curves mapping the size of the residual for GMRES,cf. Greenbaum et al. [1996], which can be almost constant and then decrease to zero in one iteration, gives us the clue that it may not be possible.

2.4 Numerical experiments

We illustrate the presented approach by one numerical experiment. We focus on the two presented techniques for the discrete solution reconstruction introduced in Section 2.2. We compare the performance of the proposed local reconstructions to the globally higher order adjoint solution. Further, the influence of the algebraic errors is studied. It is demonstrated how the presented approach based on controlling the ratio between discretization and algebraic errors helps to obtain the discrete solution reliably and efficiently. More experiments focusing on the mesh adaptation will be presented at the end of the next chapter, see Section 3.4, after the *hp*-anisotropic mesh adaptation method will be introduced.

2.4.1 Elliptic problem on a "cross" domain

We examine the performance of the reconstructions for linear Poisson equation

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{2.50}$$

in the cross shaped domain $\Omega = (-2,2) \times (-1,1) \cup (-1,1) \times (-2,2)$. We set

$$J(u) = \frac{1}{|\Omega_J|} \int_{\Omega} j_{\Omega}(x) u(x) \,\mathrm{d}x,$$

where j_{Ω} is the characteristic function of the square $\Omega_J = [1.2, 1.4] \times [0.2, 0.4],$. see Figure 2.4 The exact value of $J(u)$ is unknown hence we use the reference value 0.407617863684 which was computed in Ainsworth and Rankin [2012] on an adaptively refined mesh with more than 15 million triangles.

First we compare the quality of the presented reconstructions – primal and adjoint estimate based on the LS reconstruction (2.35) denoted η_S^{LS} and $\eta_S^{*,LS}$, estimate based

$p = 1$					
N_h	$J(e_h)$	η_S^+	η_S^{loc}	η_S^{LS}	$\eta_S^{*,LS}$
290	1.24×10^{-2}	1.21×10^{-2}	6.39×10^{-3}	1.01×10^{-2}	9.62×10^{-3}
i_{eff}		(0.98)	(0.51)	(0.81)	(0.78)
1160	4.47×10^{-3}	4.36×10^{-3}	2.29×10^{-3}	3.54×10^{-3}	3.45×10^{-3}
i_{eff}		(0.97)	(0.51)	(0.79)	(0.77)
4640	1.64×10^{-3}	1.60×10^{-3}	8.31×10^{-4}	1.29×10^{-3}	1.28×10^{-3}
i_{eff}		(0.97)	(0.51)	(0.79)	(0.78)
18560	6.18×10^{-4}	5.97×10^{-4}	3.07×10^{-4}	4.82×10^{-4}	4.80×10^{-4}
i_{eff}		(0.97)	(0.50)	(0.78)	(0.77)
74240	2.35×10^{-4}	2.19×10^{-4}	1.17×10^{-4}	1.83×10^{-4}	1.83×10^{-4}
i_{eff}		(0.93)	(0.50)	(0.78)	(0.78)
$p = 2$					
N_h	$J(e_h)$	η_S^+	η_S^{loc}	η_S^{LS}	$\eta_S^{*,LS}$
290	1.78×10^{-3}	1.27×10^{-3}	8.36×10^{-4}	4.54×10^{-4}	4.99×10^{-4}
i_{eff}		(0.71)	(0.46)	(0.25)	(0.28)
1160	7.02×10^{-4}	4.98×10^{-4}	3.27×10^{-4}	1.75×10^{-4}	1.79×10^{-4}
i_{eff}		(0.71)	(0.47)	(0.25)	(0.25)
4640	2.80×10^{-4}	1.99×10^{-4}	1.29×10^{-4}	7.03×10^{-5}	7.09×10^{-5}
i_{eff}		(0.71)	(0.46)	(0.25)	(0.25)
18560	1.15×10^{-4}	7.49×10^{-5}	5.09×10^{-5}	2.80×10^{-5}	2.82×10^{-5}
i_{eff}		(0.65)	(0.46)	(0.25)	(0.25)

Table 2.1: Elliptic problem – error estimates of the target quantity for $p = 1, 2$ on uniformly refined meshes.

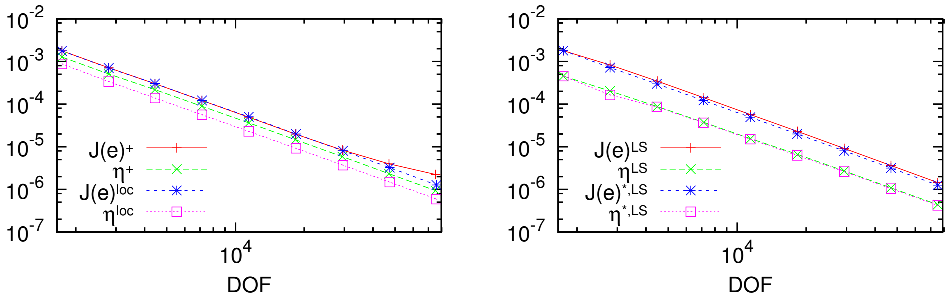

Figure 2.3: Elliptic problem – decrease of $J(e_h)$ and its estimates η_S for $p = 2$ on adaptively refined meshes.

on the LOC reconstruction (2.38) (only primal, see (2.43)) denoted η_S^{loc} and lastly the computation when the adjoint problem is solved with globally increased polynomial degree $p+1$ denoted by η_S^+.

In Table 2.1 the actual error measured with respect to the quantity of interest is compared to the discretization error estimates with effectivity indices measuring the ratio of $\eta_S / J(e_h)$. We see that although the effectivity indices are bellow one, they maintain at the same level.

Moreover, the Figure 2.3 shows the decrease of the error $J(e_h)$ and estimates η_S when adaptive refinement is used and the final mesh for η_S^{loc} is shown in Figure 2.4. It seems that although the estimates based on the local reconstructions underestimate the true error, the resulting error indicators are not worse than those obtained by global higher order solution of the adjoint problem. On the contrary, especially for the finer meshes they perform even better since the algebraic error can be more easily suppressed using the estimates (2.48a).

Further, we focus on the impact of the algebraic errors on the computation. The

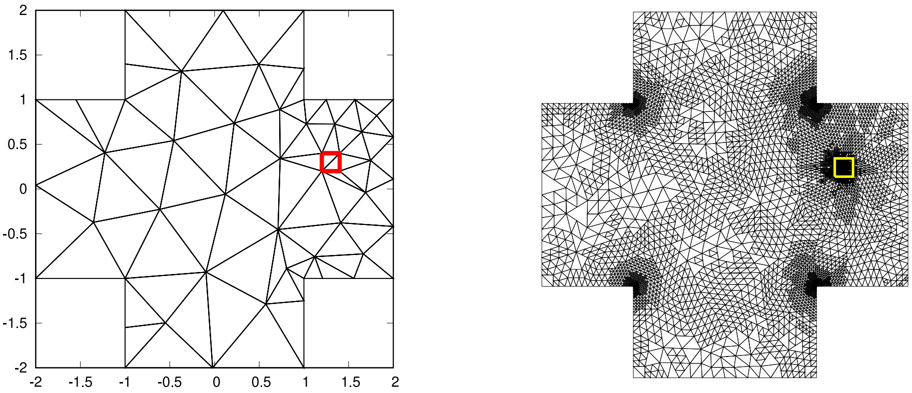

Figure 2.4: Elliptic problem – initial mesh (left) and final mesh (right), containing 14 417 triangles obtained by adaptive refinement based on the LOC reconstruction, with Ω_J highlighted.

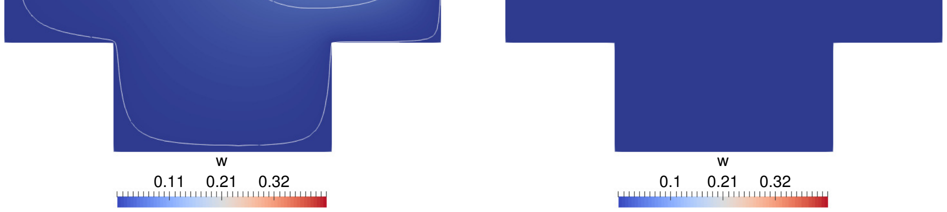

Figure 2.5: Elliptic problem – algebraically precise adjoint solution z_h (left) and its approximation after 30 GMRES iterations (right).

solution is computed with piecewise linear approximation on uniformly refined mesh with 4640 triangles.

Figure 2.5 shows the algebraically precise discrete solution z_h (left) and its approximation $z_h^{(k)}$ spoiled by algebraic errors obtained by 30 GMRES iterations. The widest contour line represents the value 10^{-4} so we see that the adjoint solution $z_h^{(k)}$ steadily equals to zero in the major part of the domain Ω unlike z_h.

That is caused by the local character of the quantity of interest. The right-hand side of the problem is nonzero only for basis functions having support in Ω_J and if we take $z_h^{(0)} = 0$ then it takes many GMRES iterations to spread the information through the whole computation domain. Since the local reconstruction of a steady zero would be again a zero function, the resulting error indicators would lead to refinement only around Ω_J and not in surroundings of the reentrant corners where the refinement is deserving due to the irregularity of the primal solution.

In Figure 2.6 the differences in the mesh refinement are exhibited if 20% of the elements with largest indicators were to refine – blue triangles would be refined due to algebraic errors while the yellow one should be refined instead. Especially, on very fine meshes this phenomenon may occur if the algebraic error was not controlled by (2.48a). A suitable preconditioning may help to overcome this phenomenon.

The dependence of the error estimates on the choice of $\eta_S^{(k)} \in \{\overline{\eta}_S^{(k)}, \tilde{\eta}_S^{(k)}, \hat{\eta}_S^{(k)}\}$, cf. (2.47), is documented in Table 2.2 and in Figure 2.7. Table 2.2 shows the number of differently (incorrectly) refined elements (column #) due to the algebraic errors in $\eta_S^{(k)}, \eta_S^{*,(k)}$ and Figure 2.7 shows the decrease of the error estimates for the least-square reconstruction. Each iteration *iter* corresponds to 50 iterations of GMRES for the primal problem and 30 iterations for the adjoint problem, respectively.

The estimates $\tilde{\eta}_S^{(k)}, \hat{\eta}_S^{(k)}$ seem to be better for the least squares reconstruction than $\overline{\eta}_S^{(k)}$ which is very sensitive to algebraic errors. Moreover it can be seen that the primal estimate $\overline{\eta}_S^{(k)}$ is more sensitive to algebraic errors in primal problem while $\overline{\eta}_S^{*,(k)}$ is more sensitive to errors in adjoint problem, which is in agreement with experiments performed in Dolejší and Roskovec [2016]. Estimates $\tilde{\eta}_S^{(k)}$ work similarly to $\hat{\eta}_S^{(k)}$ for LS reconstruction and similarly $\overline{\eta}_S^{(k)}$ for LOC reconstruction. The bold zeros in Table 2.2 mark the step where Algorithm 15 would stop. Altogether, estimates $\hat{\eta}_S^{(k)}$ and $\hat{\eta}_S^{*,(k)}$ seem to be the most robust with respect to algebraic errors and can be used equivalently, cf. Table 2.2.

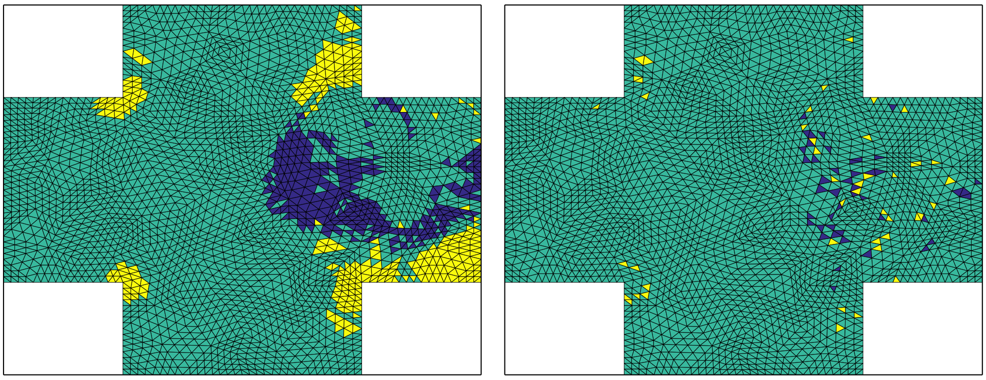

Figure 2.6: Elliptic problem – differences in refinement indicators based on $\overline{\eta}_{\mathrm{S}}^{(k)}$ after 30 (left) and 180 (right) GMRES iterations using the LS reconstruction (yellow triangles would be refined instead of the blue ones if the algebraic errors were suppressed).

iter	$\#\overline{\eta}_S$	$\overline{\eta}_A/\overline{\eta}_S$	$\overline{\eta}_A^*/\overline{\eta}_S$	$\#\overline{\eta}_S^*$	$\overline{\eta}_A/\overline{\eta}_S^*$	$\overline{\eta}_A^*/\overline{\eta}_S^*$
2	464	3.67	1.76×10^{1}	815	2.47×10^{-1}	1.18
4	349	1.32	3.08×10^{1}	836	4.56×10^{-2}	1.06
6	45	8.80×10^{-2}	1.25×10^{1}	809	7.43×10^{-3}	1.06
8	5	4.22×10^{-3}	3.53	665	1.04×10^{-3}	8.73×10^{-1}
10	2	2.00×10^{-4}	9.47×10^{-1}	414	1.09×10^{-4}	5.13×10^{-1}
12	1	9.37×10^{-6}	2.60×10^{-1}	130	7.51×10^{-6}	2.08×10^{-1}
14	**0**	3.48×10^{-7}	7.21×10^{-2}	18	3.23×10^{-7}	6.69×10^{-2}
16	0	2.31×10^{-8}	2.16×10^{-2}	4	2.24×10^{-8}	2.10×10^{-2}
18	0	1.91×10^{-8}	5.13×10^{-3}	**0**	1.88×10^{-8}	5.04×10^{-3}

iter	$\#\hat{\eta}_S$	$\hat{\eta}_A/\hat{\eta}_S$	$\hat{\eta}_A^*/\hat{\eta}_S$	$\#\hat{\eta}_S^*$	$\hat{\eta}_A/\hat{\eta}_S^*$	$\hat{\eta}_A^*/\hat{\eta}_S^*$
2	132	5.67×10^{1}	2.72×10^{2}	129	5.53×10^{1}	2.65×10^{2}
4	38	2.06	4.80×10^{1}	35	2.03	4.73×10^{1}
6	10	9.03×10^{-2}	1.29×10^{1}	11	8.91×10^{-2}	1.27×10^{1}
8	4	4.22×10^{-3}	3.53	3	4.17×10^{-3}	3.48
10	2	2.00×10^{-4}	9.47×10^{-1}	1	1.98×10^{-4}	9.34×10^{-1}
12	1	9.37×10^{-6}	2.60×10^{-1}	0	9.24×10^{-6}	2.57×10^{-1}
14	**0**	3.48×10^{-7}	7.21×10^{-2}	**0**	3.43×10^{-7}	7.12×10^{-2}
16	0	2.31×10^{-8}	2.16×10^{-2}	0	2.28×10^{-8}	2.13×10^{-2}
18	0	1.91×10^{-8}	5.13×10^{-3}	0	1.88×10^{-8}	5.06×10^{-3}

Table 2.2: Elliptic problem – number of incorrectly marked elements due to algebraic errors (LS reconstruction).

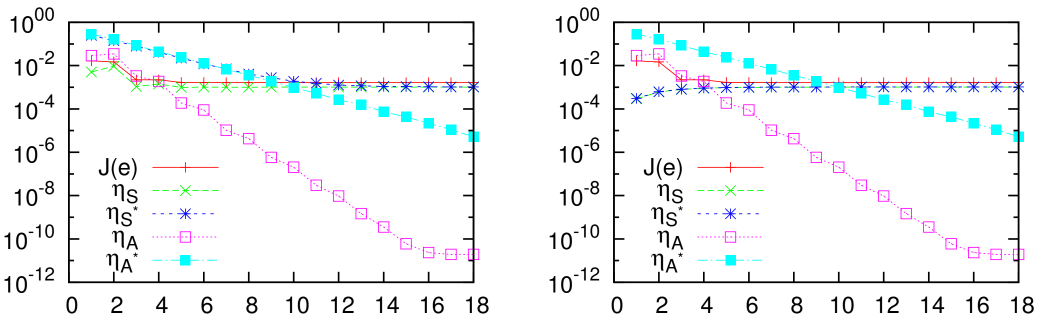

Figure 2.7: Elliptic problem – error decrease during GMRES iterations for the estimates based the least-squares reconstruction, $\overline{\eta}_{\mathrm{S}}^{(k)}$ (left), $\hat{\eta}_{\mathrm{S}}^{(k)}$ (right).

3. Anisotropic *hp*-refinement

In this chapter, we introduce the *hp*-anisotropic mesh adaptation method for DG solution of the problem (2.1) using goal-oriented error estimates. We start with the goal-oriented error estimate (2.32) and we derive its upper bound enabling setting optimal size, shape and local polynomial approximation degree for element.

The majority ideas presented here adopt the results which we have already published in Dolejší et al. [2019] and Bartoš et al. [2019]. In Dolejší et al. [2019], we derived goal oriented *a posteriori* error estimates for problem (2.1) taking into account anisotropy of the elements with arbitrary high (but constant) polynomial approximation degree $p \geq 1$. Later in Bartoš et al. [2019] we extended Dolejší et al. [2019] for the *hp*-adaptation method. Additionally, we developed a goal-oriented variant of the *p*-decision criterion from Dolejší et al. [2018], where the *p*-adaptation was based on the interpolation error estimates.

3.1 Goal-oriented error estimates in residual form

For the purpose of the anisotropic error estimate, we further estimate the residuals $r_h(u_h)(\cdot)$ and $r_h^*(z_h)(\cdot)$ of the primal and adjoint problems given by (2.9) and (2.16), respectively. In (2.19) we have already prepared the starting point for deriving the "weighted residual" estimate for the adjoint estimate η_{S}^*. Similarly to (2.19) we may rewrite the primal discrete problem in the element-wise residual form. After some manipulation we obtain the identity (cf. Hartmann [2007])

$$r_h(u_h)(\varphi) = \sum_{K \in \mathscr{T}_h} \int_K r_{K,\mathrm{V}}(u_h)\varphi \,\mathrm{d}x + \int_{\partial K} r_{K,\mathrm{B}}(u_h)\varphi + r_{K,\mathrm{D}}(u_h)\mathbb{A}\nabla\varphi \cdot \boldsymbol{n} \,\mathrm{d}x \qquad \forall \varphi \in H^2(\Omega, \mathscr{T}_h) \tag{3.1}$$

where

$$r_{K,\mathrm{V}}(u_h) := f + \nabla \cdot \mathbb{A}\nabla u_h - \nabla \cdot (\boldsymbol{b} u_h) - c u_h, \tag{3.2}$$

$$r_{K,\mathrm{B}}(u_h) := \begin{cases} -\delta[\![u_h]\!]\boldsymbol{n}\cdot\boldsymbol{n}_K - \frac{1}{2}[\![\mathbb{A}\nabla u_h]\!]\cdot\boldsymbol{n} & \text{on } \partial K^+ \setminus \partial\Omega, \\ -\delta[\![u_h]\!]\boldsymbol{n}\cdot\boldsymbol{n}_K - \frac{1}{2}[\![\mathbb{A}\nabla u_h]\!]\cdot\boldsymbol{n} + \boldsymbol{b}\cdot\boldsymbol{n}[\![u_h]\!] & \text{on } \partial K^- \setminus \partial\Omega, \\ (\delta - \boldsymbol{b}\cdot\boldsymbol{n})(u_D - u_h) & \text{on } \partial K^- \cap \Gamma_D, \\ \delta(u_D - u_h) & \text{on } \partial K^+ \cap \Gamma_D, \\ g_N - \mathbb{A}\nabla u_h \cdot \boldsymbol{n} & \text{on } \partial K \cap \Gamma_N, \\ 0 & \text{on } \partial K \cap \Gamma_+, \\ -\boldsymbol{b}\cdot\boldsymbol{n}(u_D - u_h) & \text{on } \partial K \cap \Gamma_-, \end{cases}$$

$$r_{K,\mathrm{D}}(u_h) := \begin{cases} \theta\frac{1}{2}[\![u_h]\!] & \text{on } \partial K \setminus \partial\Omega, \\ -\theta(u_D - u_h) & \text{on } \partial K \cap \Gamma_D, \\ 0 & \text{on } \partial K \cap (\partial\Omega \setminus \Gamma_D). \end{cases}$$

Obviously, $\boldsymbol{n}\cdot\boldsymbol{n}_K = \pm 1$ depending on the orientation of the edge normal vector $\boldsymbol{n} = \boldsymbol{n}_\gamma$. The forms $r_{K,\mathrm{V}}$, $r_{K,\mathrm{B}}$ and $r_{K,\mathrm{D}}$ represent the element residuals. If $u \in H^2(\Omega)$

is the weak solution of (2.2) then $r_{K,\mathrm{V}}(u) = r_{K,\mathrm{B}}(u) = r_{K,\mathrm{D}}(u) = 0\ \forall K \in \mathscr{T}_h$ imply $r_h(u_h)(\varphi) = 0\ \forall \varphi \in H^2(\Omega, \mathscr{T}_h)$, which is in agreement with the consistency of the DG method (2.10).

Finally, we introduce the notation

$$R_{K,\mathrm{V}} := \left\| r_{K,\mathrm{V}}(u_h) \right\|_K, \quad R_{K,\mathrm{B}} := \left\| r_{K,\mathrm{B}}(u_h) \right\|_{\partial K}, \quad R_{K,\mathrm{D}} := \left\| r_{K,\mathrm{D}}(u_h) \right\|_{\partial K}, \ K \in \mathscr{T}_h. \tag{3.3}$$

Then using the Cauchy inequality in (3.1) we obtain the estimate

$$|r_h(u_h)(\varphi)| \leq \sum_{K \in \mathscr{T}_h} R_{K,\mathrm{V}} \, \|\varphi\|_K + R_{K,\mathrm{B}} \, \|\varphi\|_{\partial K} + R_{K,\mathrm{D}} \, \|\mathbb{A}\nabla\varphi\|_{\partial K}. \tag{3.4}$$

Similarly, we introduce the following notation for the norms of the adjoint residual terms given in (2.18)

$$R^*_{K,\mathrm{V}} := \left\| r^*_{K,\mathrm{V}}(z_h) \right\|_K, \quad R^*_{K,\mathrm{B}} := \left\| r^*_{K,\mathrm{B}}(z_h) \right\|_{\partial K}, \quad R^*_{K,\mathrm{D}} := \left\| r^*_{K,\mathrm{D}}(z_h) \right\|_{\partial K}, \ K \in \mathscr{T}_h. \tag{3.5}$$

Now, applying the Cauchy inequality on the adjoint residual form (2.18) we obtain the estimate of the adjoint residual

$$|r^*_h(z_h)(\psi)| \leq \sum_{K \in \mathscr{T}_h} R^*_{K,\mathrm{V}} \, \|\psi\|_K + R^*_{K,\mathrm{B}} \, \|\psi\|_{\partial K} + R^*_{K,\mathrm{D}} \, \|\mathbb{A}\nabla\psi\|_{\partial K}. \tag{3.6}$$

Finally, we insert (3.4) and (3.6) into approximation of the error of the quantity of interest (2.32). We obtain

$$|\eta^{\mathrm{I}}| \leq \eta^{\mathrm{II}}, \qquad \eta^{\mathrm{II}} := \sum_{K \in \mathscr{T}_h} \eta^{\mathrm{II}}_K, \tag{3.7}$$

where

$$\begin{aligned} \eta^{\mathrm{II}}_K := \frac{1}{2}\big(& R_{K,\mathrm{V}} \left\| z_h^+ - \Pi z_h^+ \right\|_K + R^*_{K,\mathrm{V}} \left\| u_h^+ - \Pi u_h^+ \right\|_K \\ & + R_{K,\mathrm{B}} \left\| z_h^+ - \Pi z_h^+ \right\|_{\partial K} + R^*_{K,\mathrm{B}} \left\| u_h^+ - \Pi u_h^+ \right\|_{\partial K} \\ & + R_{K,\mathrm{D}} \left\| \mathbb{A}\nabla(z_h^+ - \Pi z_h^+) \right\|_{\partial K} + R^*_{K,\mathrm{D}} \left\| \mathbb{A}\nabla(u_h^+ - \Pi u_h^+) \right\|_{\partial K} \big), \end{aligned} \tag{3.8}$$

The terms $u_h^+ - \Pi u_h^+$ and $z_h^+ - \Pi z_h^+$ measured in the norms appearing in (3.8) are called the *weights* and the estimate (3.7) is the so-called *dual weighted residual* (DWR) estimate, cf. Becker and Rannacher [2001], Bangerth and Rannacher [2003].

The weighted-residual estimate (3.7) often overestimates the true error $J(e_h)$, c.f. Chapter 1. Hence in our approach it serves only as a auxiliary result for obtaining the anisotropic error indicators for mesh adaptation and the error of the target functional is still approximated by η^{I}, which should more closely approximate $J(e_h)$.

3.2 Goal-oriented error estimates enabling anisotropic refinement

In this section, we further estimate the weighting terms from (3.8) in a way convenient for choosing not only a the proper size of each element but also its shape and orientation. First, we introduce the anisotropy of triangles $K \in \mathscr{T}_h$, then we derive estimates of the interpolation error $(w - \Pi w)|_K$, $w \in P^{p+1}(K)$, $p \in \mathbb{N}$, which take into account the anisotropy of the element K.

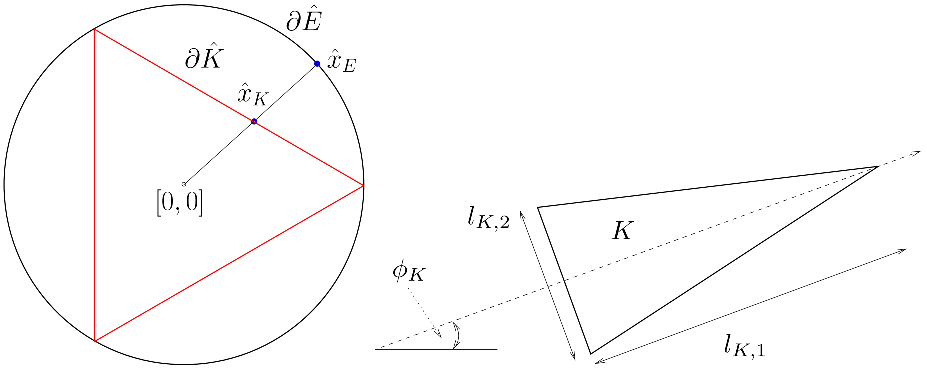

Figure 3.1: The reference triangle $\hat{K}$ and reference ellipse E (left) and a general triangle K with its orientation ϕ_K and aspect ratio $\sigma_K = \frac{l_{K,1}}{l_{K,2}}$ (right).

3.2.1 Anisotropy of mesh elements

Let $\hat{K}$ be a reference equilateral triangle with vertices $(1,0)$, $(-\frac{1}{2},+\frac{\sqrt{3}}{2})$, $(-\frac{1}{2},-\frac{\sqrt{3}}{2})$, see Figure 3.1. Let $K \in \mathscr{T}_h$ be an arbitrary but fixed triangle. There exists an affine mapping $F_K : \hat{K} \to K$ which maps $\hat{K}$ onto K given by

$$F_K\hat{x} = \mathbb{M}_K\hat{x} + x_K \quad \hat{x} \in \hat{K}, \tag{3.9}$$

where $x_K \in \mathbb{R}^2$ is the barycenter of K, $\mathbb{M}_K$ is a 2×2 matrix defining size and shape of K. Further, the matrix $\mathbb{M}_K$ may be rewritten using its singular value decomposition into

$$\mathbb{M}_K\hat{x} = \mathbb{Q}_{\phi_K}\mathbb{L}_K\mathbb{Q}^{\mathrm{T}}_{\psi_K}, \tag{3.10}$$

where the matrices $\mathbb{Q}_{\phi_K}$ and $\mathbb{Q}_{\psi_K}$ are the rotation matrices through angles ϕ_K and ψ_K counterclockwise, respectively, given by

$$\mathbb{Q}_\phi := \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}. \tag{3.11}$$

Furthermore, putting $\lambda_K := \sqrt{l_{K,1}l_{K,2}} > 0$ and $\sigma_K := \sqrt{l_{K,1}/l_{K,2}} \geq 1$, the matrix $\mathbb{L}_K$ equals

$$\mathbb{L}_K = \begin{pmatrix} l_{K,1} & 0 \\ 0 & l_{K,2} \end{pmatrix} = \lambda_K \begin{pmatrix} \sigma_K & 0 \\ 0 & \sigma_K^{-1} \end{pmatrix} =: \lambda_K\mathbb{S}_K. \tag{3.12}$$

We note that the area of K can then be computed as $|K| = l_{K,1}l_{K,2}|\hat{K}| = \lambda_K^2|\hat{K}|$, where $|\hat{K}| = 3\sqrt{3}/4$.

Definition 3.1. *Using* (3.9), (3.10), (3.12)*, we can characterize each element $K \in \mathscr{T}_h$ by the quintet $\{x_K, \lambda_K, \sigma_K, \phi_K, p_K\}$, where $x_K \in \Omega$ is the barycenter of K, $\lambda_K > 0$ is called the* size *of K, $\sigma_K \geq 1$ is called the* aspect ratio *of K, the angle $\phi_K \in [0, 2\pi)$ is the* orientation *of K and $p_K \in \mathbb{N}$ is the local polynomial approximation degree on K. Moreover, we call the pair $\{\sigma_K,\ \phi_K\}$ the* shape *of K and the triplet $\{\lambda_K,\ \sigma_K,\ \phi_K\}$ the* anisotropy *of K.*

For any hp-mesh $\mathscr{T}_{h,\boldsymbol{p}} = \{\mathscr{T}_h, \boldsymbol{p}\}$ (given by the mesh $\mathscr{T}_h$ and the set of element polynomial degrees p_K, $K \in \mathscr{T}_h$) we can evaluate the quintets $\{x_K, \lambda_K, \sigma_K, \phi_K, p_K\}$ for each element $K \in \mathscr{T}_h$. On the other hand, there exist algorithms and softwares (e.g., Dolejší [2000]) which are able to generate the corresponding hp-mesh by the least square technique, for the given set of quintets $\{x_l, \lambda_l, \sigma_l, \phi_l, p_l\}$, $l = 1, \ldots, r$, (satisfying $x_l \in \Omega$, $\lambda_l > 0$, $\sigma_l \geq 1$, $\phi_l \in [0, 2\pi)$ and $p_l \in \mathbb{N}$), see Dolejší et al. [2018].

Remark. Let $\hat{E} = \{x \in \mathbb{R}^2, |x|^2 \leq 1\}$ be the unit ball. Obviously, $\hat{K} \subset \hat{E}$. Moreover, applying mapping F_K on $\hat{E}$, we obtain the ellipse $E_K := \mathbb{M}_K \hat{E} + x_K$, whose center is the barycenter of K, $K \subset E_K$, $l_{K,1}$ and $l_{K,2}$ are its semimajor and semiminor axes, respectively, and ϕ_K is the orientation of the main axis of E_K. Therefore, the triplet $\{\lambda_K, \sigma_K, \phi_K\}$ also denotes the anisotropy of the ellipse E_K.

3.2.2 Anisotropy of polynomial functions

Now we introduce the anisotropy of polynomial functions following the approach from Dolejší [2014].

Definition 3.2. *Let $\bar{x} = (\bar{x}_1, \bar{x}_2) \in \Omega$ and $q \in \mathbb{N}$ be given. We say that a polynomial function $\mu : \Omega \to \mathbb{R}$ is a q-*function located at $\bar{x}$ *if*

$$\mu(x) = \textstyle\sum_{l=0}^{q} \alpha_l (x_1 - \bar{x}_1)^l (x_2 - \bar{x}_2)^{q-l}, \quad x = (x_1, x_2) \in \Omega, \tag{3.13}$$

where $\alpha_l \in \mathbb{R}$, $l = 0, \ldots, q$.

In other words any q-function is a polynomial function containing only terms of degree q. Any q-function may be written in the polar coordinates with the origin at $\bar{x}$ as

$$\mu(r,t) = r^q \textstyle\sum_{l=0}^{q} \alpha_i \cos^l(t) \sin^{q-l}(t), \quad r > 0,\ t \in [0, 2\pi). \tag{3.14}$$

The following result, derived in [Dolejší, 2014, Lemma 3.12], is the essential component for the error estimates enabling anisotropic refinement.

Lemma 3.3. *Let $\mu : \Omega \to \mathbb{R}$ be a q-function located at $\bar{x}$, $q \geq 2$. Then there exist values $A \geq 0$, $\rho \geq 1$ and $\varphi \in [0, 2\pi)$ such that*

$$|\mu(x)| \leq A \left((x - \bar{x})^{\mathrm{T}} \mathbb{Q}_{\varphi} \mathbb{D}_{\rho}^{q} \mathbb{Q}_{\varphi}^{\mathrm{T}} (x - \bar{x}) \right)^{q/2} \ \forall x \in \Omega, \tag{3.15}$$

where $A > 0$, $\mathbb{Q}_{\varphi}$ is the 2×2 rotation matrix through angle φ counterclockwise, and $\mathbb{D}_{\rho}^{q} := \mathrm{diag}(1, \rho^{-2/q})$ (symbol q in $\mathbb{D}_{\rho}^{q}$ means only the superscript, not the power of $\mathbb{D}$).

Definition 3.4. *Let $\mu : \Omega \to \mathbb{R}$ be a q-function located at $\bar{x}$, $q \geq 2$. The values $A \geq 0$, $\rho \geq 1$ and $\varphi \in [0, 2\pi)$ from* (3.15) *define the* size, *the* aspect ratio *and the* orientation *of the q-function μ, respectively. Moreover, the triplet $\{A, \rho, \varphi\}$ is called the* anisotropy *of the q-function μ.*

3.2.3 Estimate of the interpolation error function

Let $w \in P^{p+1}(\Omega)$ with $p \in \mathbb{N}$ and $\bar{x} = (\bar{x}_1, \bar{x}_2) \in \Omega$ be given. Using the Taylor expansion of degree $p+1$ at $\bar{x}$, we have

$$w(x) = \Pi_{\bar{x},p} w(x) + w^{\text{int}}_{\bar{x},p}(x),\ x \in \Omega, \tag{3.16}$$

where $\Pi_{\bar{x},p} w$ is a p-degree polynomial approximation of w given by

$$\Pi_{\bar{x},p} w(x) := \sum_{k=0}^{p} \frac{1}{k!} \left(\sum_{l=0}^{k} \left(\binom{k}{l} \right) \frac{\partial^k w(\bar{x})}{\partial x_1^l \partial x_2^{k-l}} (x_1 - \bar{x}_1)^l (x_2 - \bar{x}_2)^{k-l} \right),\ x \in \Omega, \tag{3.17}$$

with $\left(\binom{k}{l}\right) = \frac{k!}{l\,!(k-l)!}$ and

$$w - \Pi_{\bar{x},p} w = w^{\text{int}}_{\bar{x},p}(x) := \textstyle\sum_{l=0}^{p+1} \alpha_l (x_1 - \bar{x}_1)^l (x_2 - \bar{x}_2)^{p+1-l},\ x \in \Omega, \tag{3.18}$$

where $\alpha_l = \frac{1}{(p+1)!} \left(\binom{p+1}{l} \right) \frac{\partial^{p+1} w(\bar{x})}{\partial x_1^l \partial x_2^{p+1-l}}$, $l = 0, \ldots, p+1$.

We call $w^{\text{int}}_{\bar{x},p}$ the *interpolation error function* of degree p located at $\bar{x}$.

Obviously, the function $w^{\text{int}}_{\bar{x},p}$ is a $(p+1)$-function in the sense of Definition 3.2 and therefore, using (3.15) , there exist $A_w \geq 0$, $\rho_w \geq 1$ and $\varphi_w \in [0, 2\pi)$ such that

$$|w^{\text{int}}_{\bar{x},p}(x)| \leq A_w \left((x - \bar{x})^{\mathrm{T}} \mathbb{Q}_{\varphi_w} \mathbb{D}^{p+1}_{\rho_w} \mathbb{Q}^{\mathrm{T}}_{\varphi_w} (x - \bar{x}) \right)^{\frac{p+1}{2}},\ x \in \Omega. \tag{3.19}$$

Remark. Let us note that the value A_w represents the maximal value of the $(\bar{p}+1)^{\text{th}}$-order scaled directional derivative of w at $\bar{x}$, φ_w is the angle of the direction of the maximal derivative and ρ_w corresponds to the ratio between A_w and the $(\bar{p}+1)^{\text{th}}$-order scaled directional derivative along the perpendicular direction. Further, the value A_w depends on $\alpha_l = \frac{1}{l!(q-l)!} \frac{\partial^q w(x)}{\partial x_1^l \partial x_2^{q-l}}$, $l = 0, \ldots, q$ and can be evaluated numerically in such a way that estimate (3.19) is sharp, for more details see Dolejší [2014].

Now, we further develop the estimate of $w^{\text{int}}_{\bar{x},p}$ (3.19) for the barycenter of K, i.e., $\bar{x} := x_K$.

Lemma 3.5. *Let $K \in \mathscr{T}_h$ be a triangle with the anisotropy $\{\lambda_K, \sigma_K, \phi_K\}$ (cf. Definition 3.1), $w \in P^{p+1}(K)$ and $w^{\text{int}}_{\bar{x},p}$ be the corresponding interpolation error function with the anisotropy $\{A_w, \rho_w, \varphi_w\}$ given by 3.19. Then*

$$\left\| w^{\text{int}}_{\bar{x},p} \right\|^2_K \leq A_w^2 \lambda_K^{2(p+2)} \frac{1}{2p+4} \boldsymbol{G}(p+1, p+1, \rho_w, \varphi_w; \sigma_K, \phi_K), \tag{3.20a}$$

$$\left\| w^{\text{int}}_{\bar{x},p} \right\|^2_{\partial K} \leq A_w^2 \lambda_K^{2p+3} \sigma_K \boldsymbol{G}(p+1, p+1, \rho_w, \varphi_w; \sigma_K, \phi_K), \tag{3.20b}$$

where

$$\boldsymbol{G}(q_1, q_2, \rho, \varphi; \sigma, \phi) = \int_0^{2\pi} \left(\mathbb{G}_{11} \cos^2 t + 2\mathbb{G}_{12} \sin t \cos t + \mathbb{G}_{22} \sin^2 t \right)^{q_1} \mathrm{d}t, \tag{3.21}$$

$$\begin{aligned} \textit{with}\quad \mathbb{G}_{11} &= \sigma^2 [\cos^2(\phi - \varphi) + \rho^{-2/q_2} \sin^2(\phi - \varphi)], \\ \mathbb{G}_{12} &= -\sin(\phi - \varphi) \cos(\phi - \varphi)(1 - \rho^{-2/q_2}), \\ \mathbb{G}_{22} &= \sigma^{-2} [\sin^2(\phi - \varphi) + \rho^{-2/q_2} \cos^2(\phi - \varphi)]. \end{aligned}$$

Proof. Let F_K be the affine function (3.9) mapping $\hat{K}$ onto K and let $\mathbb{M}_K$ denote the corresponding Jacobian matrix. Moreover, for any function$f(x): K \to \mathbb{R}$ we define $\hat{f}: \hat{K} \to \mathbb{R}$ by $\hat{f}(\hat{x}) = f(F_K\hat{x})$. Using the substitution theorem, (3.9), (3.10) and (3.12), we have

$$\int_K f(x)\,\mathrm{d}x = \int_{\hat{K}} \hat{f}(\hat{x})|\det \mathbb{M}_K|\,\mathrm{d}\hat{x} = \lambda_K^2 \int_{\hat{K}} \hat{f}(\hat{x})\,\mathrm{d}\hat{x} \tag{3.22}$$

Furthermore, we can write

$$x - x_K = \lambda_K \mathbb{Q}_{\phi_K} \mathbb{S}_K \mathbb{Q}_{\psi_K}^{\mathrm{T}} \hat{x} \quad \Rightarrow \quad (x - x_K)^{\mathrm{T}} = \lambda_K \hat{x}^{\mathrm{T}} \mathbb{Q}_{\psi_K} \mathbb{S}_K \mathbb{Q}_{\phi_K}^{\mathrm{T}}. \tag{3.23}$$

Using (3.19), (3.22) and (3.23), we obtain

$$\begin{aligned}\left\|w_{\bar{x},p}^{\mathrm{int}}\right\|_K^2 &= \int_K \left|w_{\bar{x},p}^{\mathrm{int}}(x)\right|^2 \mathrm{d}x \le \int_K A_w^2 \left((x-x_K)^{\mathrm{T}} \mathbb{Q}_{\varphi_w} \mathbb{D}_{\rho_w}^{p+1} \mathbb{Q}_{\varphi_w}^{\mathrm{T}} (x-x_K)\right)^{p+1} \mathrm{d}x \\ &= A_w^2 \lambda_K^{2(p+2)} \int_{\hat{K}} \left(\hat{x}^{\mathrm{T}} \mathbb{Q}_{\psi_K} \mathbb{G} \mathbb{Q}_{\psi_K}^{\mathrm{T}} \hat{x}\right)^{p+1} \mathrm{d}\hat{x},\end{aligned} \tag{3.24}$$

where

$$\mathbb{G} := \mathbb{S}_K \mathbb{Q}_{\phi_K}^{\mathrm{T}} \mathbb{Q}_{\varphi_w} \mathbb{D}_{\rho_w}^{p+1} \mathbb{Q}_{\varphi_w}^{\mathrm{T}} \mathbb{Q}_{\phi_K} \mathbb{S}_K. \tag{3.25}$$

Let $\hat{E} = \{\hat{x} \in \mathbb{R}^2 : \ |\hat{x}| \le 1\}$ be the unit ball. Obviously, $\hat{K} \subset \hat{E}$ and then we can replace the domain of integration in the last integral (3.24) by

$$\left\|w_{\bar{x},p}^{\mathrm{int}}\right\|_K^2 \le A_w^2 \lambda_K^{2(p+2)} \int_{\hat{E}} \left(\hat{x}^{\mathrm{T}} \mathbb{G} \hat{x}\right)^{p+1} \mathrm{d}\hat{x}, \tag{3.26}$$

since $\mathbb{Q}_{\psi_K}$ is a rotation matrix and hence the transformation $\hat{x} \to \mathbb{Q}_{\psi_K}^{\mathrm{T}} \hat{x}$ maps $\hat{E}$ onto itself. Now we evaluate the matrix $\mathbb{G}$ from (3.25) and then using it together with the identity $\mathbb{Q}_{\varphi_w}^{\mathrm{T}} \mathbb{Q}_{\phi_K} = \mathbb{Q}_{\phi_K - \varphi_w} =: \mathbb{Q}_{\tau_K}$ (i.e., $\tau_K := \phi_K - \varphi_w$), we get

$$\mathbb{G} = \begin{pmatrix} \sigma_K^2(\cos^2\tau_K + \rho_w^{-\frac{2}{p+1}} \sin^2\tau_K) & -\sin\tau_K \cos\tau_K (1 - \rho_w^{-\frac{2}{p+1}}) \\ -\sin\tau_K \cos\tau_K (1 - \rho_w^{-\frac{2}{p+1}}) & \sigma_K^{-2}(\sin^2\tau_K + \rho_w^{-\frac{2}{p+1}} \cos^2\tau_K) \end{pmatrix}. \tag{3.27}$$

Since $\mathbb{G}$ is independent of x_K, we evaluate the last integral in (3.26) using the polar coordinates, namely

$$\left\|w_{\bar{x},p}^{\mathrm{int}}\right\|_K^2 \le A_w^2 \lambda_K^{2(p+2)} \int_0^1 r^{2p+3} \left(\int_0^{2\pi} g(t)^{p+1}\,\mathrm{d}t\right) \mathrm{d}r, \tag{3.28}$$

where $g(t) = \mathbb{G}_{11}\cos^2 t + 2\mathbb{G}_{12}\sin t \cos t + \mathbb{G}_{22}\sin^2 t$. Since $g(t)$ is independent of r, we change the order of integration and computing $\int_0^1 r^{2p+3}\,\mathrm{d}r = 1/(2p+4)$, yields (3.20a).

In order to prove (3.20b), we proceed similarly. Using path integration, we have

$$\int_{\partial K} f(x)\,\mathrm{d}S = \int_{\partial\hat{K}} \hat{f}(\hat{x})|\mathbb{M}_K \cdot \boldsymbol{t}|\,\mathrm{d}\hat{S} \le \int_{\partial\hat{K}} \hat{f}(\hat{x})\|\mathbb{M}_K\|_2\,\mathrm{d}\hat{S}, \tag{3.29}$$

where $\boldsymbol{t}$ is the unit tangent vector to $\partial\hat{K}$, $\mathbb{M}_k$ is the matrix from (3.9) and $\|\mathbb{M}_K\|_2 = r(\mathbb{M}_K^T\mathbb{M}_K)^{1/2} = \lambda_K\sigma_K$ is the matrix Eucledian norm where $r(\cdot)$ denotes the spectral radius of its argument. Using (3.29) and a similar manipulation as in (3.24), we obtain

$$\left\|w_{\bar{x},p}^{\mathrm{int}}\right\|_{\partial K}^2 \leq A_w^2\lambda_K^{2p+3}\sigma_K\int_{\partial\hat{K}}\left(\hat{x}^{\mathrm{T}}\mathbb{Q}_{\psi_K}\mathbb{G}\mathbb{Q}_{\psi_K}^{\mathrm{T}}\hat{x}\right)^{p+1}\,\mathrm{d}\hat{x}. \tag{3.30}$$

The integrand of the last integral is an increasing function of $|\hat{x}|$. Since the $\hat{x}_E \in \partial\hat{E}$ is circumscribed circle of $\hat{x}_K \in \partial\hat{K}$, see Figure 3.1, for each $\hat{x}_K \in \partial\hat{K}$ there exists $\hat{x}_E \in \partial\hat{E}$, where the value of integrand is higher. Moreover, $|\partial\hat{K}| < |\partial\hat{E}|$ and hence when we replace the domain of integration in (3.30), we obtain

$$\left\|w_{\bar{x},p}^{\mathrm{int}}\right\|_{\partial K}^2 \leq A_w^2\lambda_K^{2p+3}\sigma_K\int_{\partial\hat{E}}\left(\hat{x}^{\mathrm{T}}\mathbb{Q}_{\psi_K}\mathbb{G}\mathbb{Q}_{\psi_K}^{\mathrm{T}}\hat{x}\right)^{p+1}\,\mathrm{d}\hat{x}. \tag{3.31}$$

Now, analogously as in (3.28) converting (3.31) to polar coordinates and using

$$\int_{\partial\hat{E}}\left(\hat{x}^{\mathrm{T}}\mathbb{G}\hat{x}\right)^{p+1}\mathrm{d}\hat{x} = \int_0^{2\pi} g(t)^{p+1}\mathrm{d}t.$$

leads to (3.20b). □

In virtue of (3.8), now it only remains to derive the estimate of

$$\left\|\mathbb{A}\nabla(w-\Pi_{\bar{x},p}w)\right\|_{L^2(\partial K)} = \left\|\mathbb{A}\nabla w_{\bar{x},p}^{\mathrm{int}}\right\|_{L^2(\partial K)},$$

where $\mathbb{A} = \mathbb{A}(x)$, $x \in \Omega$ is the diffusion matrix from (2.1a), $\bar{x} = x_K$ is the barycenter of K and $p = p_K$ is the local polynomial degree.

Let $\bar{\mathbb{A}}$ be a constant approximation of $\mathbb{A}$ on K, e.g., $\bar{\mathbb{A}}$ be the mean value of $\mathbb{A}$. In the following, we estimate the term $\|\bar{\mathbb{A}}\nabla w_{\bar{x},p}^{\mathrm{int}}\|_{L^2(\partial K)}$. The term $\|(\mathbb{A}-\bar{\mathbb{A}})\nabla w_{\bar{x},p}^{\mathrm{int}}\|_{L^2(\partial K)}$ is neglected since it is a higher order term. If $\mathbb{A}$ is constant on K then this term vanishes completely.

Lemma 3.6. *Let $w_{\bar{x},p}^{\mathrm{int}}$ be the interpolation error function* (3.18) *and $\bar{\mathbb{A}}$ the constant approximation of the diffusive matrix$\mathbb{A}$. Then $|\bar{\mathbb{A}}\nabla w_{\bar{x},p}^{\mathrm{int}}(x)|^2$ is a $2p$-function (cf. Definition 3.2) satisfying*

$$|\bar{\mathbb{A}}\nabla w_{\bar{x},p}^{\mathrm{int}}(x)|^2 = \textstyle\sum_{i=0}^{2p}\delta_i(x_1-\bar{x}_1)^i(x_2-\bar{x}_2)^{2p-i}, \tag{3.32}$$

where δ_i, $i = 0,\ldots,2p$ are given by

$$\begin{aligned}
\delta_i &:= \textstyle\sum_{j=0}^{i}(\beta_j^{(1)}\beta_{i-j}^{(1)} + \beta_j^{(2)}\beta_{i-j}^{(2)}), \quad i = 0,\ldots,p,\\
\delta_{2p-i} &:= \textstyle\sum_{j=0}^{i}(\beta_{p-j}^{(1)}\beta_{p-(i-j)}^{(1)} + \beta_{p-j}^{(2)}\beta_{p-(i-j)}^{(2)}), \quad i = 0,\ldots,p.\\
\beta_l^{(1)} &:= \bar{\mathbb{A}}_{11}(l+1)\,\alpha_{l+1} + \bar{\mathbb{A}}_{12}(p+1-l)\,\alpha_l, \quad l = 0,\ldots,p,\\
\beta_l^{(2)} &:= \bar{\mathbb{A}}_{21}(l+1)\,\alpha_{l+1} + \bar{\mathbb{A}}_{22}(p+1-l)\,\alpha_l, \quad l = 0,\ldots,p,
\end{aligned} \tag{3.33}$$

Proof. We evaluate the square of the magnitude of $\bar{\mathbb{A}}\nabla w_{\bar{x},p}^{\mathrm{int}}$. Using (3.18) and a direct manipulation (the detailed procedure can be found in [Dolejší, 2015, Section 3.3]), we obtain

$$|\bar{\mathbb{A}}\nabla w_{\bar{x},p}^{\rm int}(x)|^2 = \sum_{i=1}^{2}\left(\sum_{j=1}^{2}\bar{\mathbb{A}}_{ij}\frac{\partial}{\partial x_j}\sum_{l=0}^{p+1}\alpha_l(x_1-\bar{x}_1)^l(x_2-\bar{x}_2)^{p+1-l}\right)^2 \tag{3.34}$$
$$= \left(\sum_{l=0}^{p}\beta_l^{(1)}\xi_1^l\xi_2^{p-l}\right)^2 + \left(\sum_{l=0}^{p}\beta_l^{(2)}\xi_1^l\xi_2^{p-l}\right)^2 = \sum_{i=0}^{2p}\delta_i\xi_1^i\xi_2^{2p-i},$$

where $\xi_i = x_i - \bar{x}_i,\ i = 1,2$ □

Therefore, in virtue of Lemma 3.3, there exist values $\tilde{A}_w \geq 0$, $\tilde{\rho}_w \geq 1$ and $\tilde{\varphi}_w \in [0,2\pi)$ such that

$$\left|\bar{\mathbb{A}}\nabla w_{\bar{x},p}^{\rm int}(x)\right|^2 \leq \tilde{A}_w\left((x-\bar{x})^{\rm T}\mathbb{Q}_{\tilde{\varphi}_w}\mathbb{D}_{\tilde{\rho}_w}^{2p}\mathbb{Q}_{\tilde{\varphi}_w}^{\rm T}(x-\bar{x})\right)^p \ \forall x\in\Omega, \tag{3.35}$$

where $\mathbb{Q}_{\tilde{\varphi}_w}$ is a rotation matrix and $\mathbb{D}_{\tilde{\rho}_w}^{2p} = \mathrm{diag}(1,\tilde{\rho}_w^{-1/p})$. Finally, we extend the Lemma 3.5 by the following result.

Lemma 3.7. *Let $K\in\mathscr{T}_h$ be a triangle with the anisotropy $\{\lambda_K,\,\sigma_K,\,\phi_K\}$ (cf. Definition 3.1), $w\in P^{p+1}(K)$, $w_{\bar{x},p}^{\rm int}$ be the corresponding interpolation error function* (3.18) *and let $|\bar{\mathbb{A}}\nabla w_{\bar{x},p}^{\rm int}(x)|^2$ have the anisotropy $\{\tilde{A}_w,\,\tilde{\rho}_w,\,\tilde{\varphi}_w\}$ given by* (3.35). *Then*

$$\left\|\bar{\mathbb{A}}\nabla w_{\bar{x},p}^{\rm int}\right\|_{\partial K}^2 \leq \tilde{A}_w\lambda_K^{2p+1}\sigma_K\boldsymbol{G}(p,2p,\tilde{\rho}_w,\tilde{\varphi}_w;\sigma_K,\phi_K), \tag{3.36}$$

where $\boldsymbol{G}$ is defined by (3.21).

Proof. Using the definition of the Eucledian norm, we have

$$\left\|\bar{\mathbb{A}}\nabla w_{\bar{x},p}^{\rm int}\right\|_{\partial K}^2 = \int_{\partial K}|\bar{\mathbb{A}}\nabla w_{\bar{x},p}^{\rm int}|^2\,\mathrm{d}S. \tag{3.37}$$

Now, applying the estimate (3.35) on the right-hand side of (3.37) and using an analogous procedure as in the proof of (3.20b), we obtain (3.36). □

3.2.4 Anisotropic goal-oriented error estimates

We employ the estimates (3.20) and (3.36) in order to derive estimates of $\eta_K^{\rm II}$, $K\in\mathscr{T}_h$ given by (3.8) which take into account the size and shape of mesh elements. We consider the operator $\Pi : S_h^{p+1} \to S_h^p$ given by $\Pi\big|_K := \Pi_{x_K,p}$, $K\in\mathscr{T}_h$, see (3.17), and we estimate the individual weighting terms ($u_h^+ - \Pi u_h^+$ and $z_h^+ - \Pi z_h^+$) appearing in (3.8). We recall that $u_h^+\in S_h^{p+1}$ and $z_h^+\in S_h^{p+1}$ are the higher-order reconstructions of the primal and adjoint solutions of (2.9) and (2.23), respectively.

Let $K\in\mathscr{T}_h$ be an arbitrary mesh element and let $\{\lambda_K,\sigma_K,\phi_K\}$ denote its anisotropy. Relation (3.18) implies that $(u_h^+ - \Pi u_h^+)|_K$ and $(z_h^+ - \Pi z_h^+)|_K$ are (p_K+1)-functions, cf. Definition 3.2. Similarly, $|\bar{\mathbb{A}}\nabla(u_h^+ - \Pi u_h^+)|_K|^2$ and $|\bar{\mathbb{A}}\nabla(z_h^+ - \Pi z_h^+)|_K|^2$ are $2p_K$-functions. Therefore, we apply (3.19) and (3.35) for introducing

$$\begin{aligned}\{A_u,\,\rho_u,\,\varphi_u\}&\ \text{the anisotropy of } (u_h^+ - \Pi u_h^+)|_K, \\ \{\tilde{A}_u,\,\tilde{\rho}_u,\,\tilde{\varphi}_u\}&\ \text{the anisotropy of } |\bar{\mathbb{A}}\nabla(u_h^+ - \Pi u_h^+)|_K|^2, \\ \{A_z,\,\rho_z,\,\varphi_z\}&\ \text{the anisotropy of } (z_h^+ - \Pi z_h^+)|_K, \\ \{\tilde{A}_z,\,\tilde{\rho}_z,\,\tilde{\varphi}_z\}&\ \text{the anisotropy of } |\bar{\mathbb{A}}\nabla(z_h^+ - \Pi z_h^+)|_K|^2,\end{aligned} \tag{3.38}$$

which depend on $(p+1)^{\rm th}$-derivatives of u_h^+ and z_h^+.

Remark. We note that the second power in the definition of the anisotropy $|\bar{\mathbb{A}}\nabla(u_h^+ - \Pi u_h^+)|_K|^2$ (and $|\bar{\mathbb{A}}\nabla(z_h^+ - \Pi z_h^+)|_K|^2$, respectively) is not a mistake, because while $|\bar{\mathbb{A}}\nabla(u_h^+ - \Pi u_h^+)|_K|^2$ is a polynomial of degree $2p_K$, $|\bar{\mathbb{A}}\nabla(u_h^+ - \Pi u_h^+)|_K|$ is not even polynomial function.

Finally, applying (3.20) and (3.36) together with (3.38) gives the anisotropic weighting terms

$$
\begin{aligned}
\|u_h^+ - \Pi u_h^+\|_K &\le \Big(\frac{A_u^2 \lambda_K^{2(p_K+2)}}{2p_K+4} \boldsymbol{G}(p_K+1, p_K+1, \rho_u, \varphi_u; \sigma_K, \phi_K)\Big)^{1/2} &&=: \theta_{K,\mathrm{V}},\\
\|z_h^+ - \Pi z_h^+\|_K &\le \Big(\frac{A_z^2 \lambda_K^{2(p_K+2)}}{2p_K+4} \boldsymbol{G}(p_K+1, p_K+1, \rho_z, \varphi_z; \sigma_K, \phi_K)\Big)^{1/2} &&=: \theta^*_{K,\mathrm{V}},\\
\|u_h^+ - \Pi u_h^+\|_{\partial K} &\le \Big(A_u^2 \lambda_K^{2p_K+3} \sigma_K \boldsymbol{G}(p_K+1, p_K+1, \rho_u, \varphi_u; \sigma_K, \phi_K)\Big)^{1/2} &&=: \theta_{K,\mathrm{B}},\\
\|z_h^+ - \Pi z_h^+\|_{\partial K} &\le \Big(A_z^2 \lambda_K^{2p_K+3} \sigma_K \boldsymbol{G}(p_K+1, p_K+1, \rho_z, \varphi_z; \sigma_K, \phi_K)\Big)^{1/2} &&=: \theta^*_{K,\mathrm{B}},\\
\|\bar{\mathbb{A}}\nabla(u_h^+ - \Pi u_h^+)\|_{\partial K} &\le \Big(\tilde{A}_u \lambda_K^{2p_K+1} \sigma_K \boldsymbol{G}(p_K, 2p_K, \tilde{\rho}_u, \tilde{\varphi}_u; \sigma_K, \phi_K)\Big)^{1/2} &&=: \theta_{K,\mathrm{D}},\\
\|\bar{\mathbb{A}}\nabla(z_h^+ - \Pi z_h^+)\|_{\partial K} &\le \Big(\tilde{A}_z \lambda_K^{2p_K+1} \sigma_K \boldsymbol{G}(p_K, 2p_K, \tilde{\rho}_z, \tilde{\varphi}_z; \sigma_K, \phi_K)\Big)^{1/2} &&=: \theta^*_{K,\mathrm{D}}.
\end{aligned}
\tag{3.39}
$$

Now, applying (3.39) on the terms appearing in (3.8), we get $\eta_K^{\mathrm{II}} \le \eta_K^{\mathrm{III}}$, where

$$
\eta_K^{\mathrm{III}} := \frac{1}{2}\big(R_{K,\mathrm{V}}\theta^*_{K,\mathrm{V}} + R_{K,\mathrm{B}}\theta^*_{K,\mathrm{B}} + R_{K,\mathrm{D}}\theta^*_{K,\mathrm{D}} + R^*_{K,\mathrm{V}}\theta_{K,\mathrm{V}} + R^*_{K,\mathrm{B}}\theta_{K,\mathrm{B}} + R^*_{K,\mathrm{D}}\theta_{K,\mathrm{D}}\big)
\tag{3.40}
$$

with the residuals $R_{K,\mathrm{V}}$, $R_{K,\mathrm{B}}$, $R_{K,\mathrm{D}}$, $R^*_{K,\mathrm{V}}$, $R^*_{K,\mathrm{B}}$, $R^*_{K,\mathrm{D}}$ defined by (3.3) and (3.5). Punctually written the estimate η_K^{III} depends on all twelve the anisotropy parameters A_u, ρ_u, ... from (3.38) and also on the element anisotropy parameters $\lambda_K, \sigma_K, \phi_K$ and the polynomial approximation degree p_K, but for the sake of clarity this dependence is omitted in the notation.

The previous derivations can be summarized into the following results.

Theorem 3.8. *Let u_h and z_h be the approximate solutions of the primal and adjoint problems given by* (2.9) *and* (2.23)*, respectively. Further, let u_h^+ and z_h^+ be higher-order approximations of the primal and adjoint solutions u and z reconstructed from u_h and z_h, respectively. Finally, let $\eta^{\mathrm{I}} \approx J(e_h)$ be given by* (2.32)*. Then*

$$
|J(e_h)| \approx |\eta^{\mathrm{I}}| \le \eta^{\mathrm{II}} \le \sum_{K \in \mathscr{T}_h} \eta_K^{\mathrm{III}}
\tag{3.41}
$$

where η^{II} is given by (3.7) *and η_K^{III}, $K \in \mathscr{T}_h$ are computable quantities given by* (3.40).

3.3 Anisotropic *hp*-mesh adaptation algorithm

Standard mesh adaption algorithms simply split the elements with highest error estimates into several (four in 2D) sub-elements. This procedure is repeated until condition (3.42) is achieved. On the other hand, the anisotropic mesh adaption technique allows to adapt the size of elements continuously, i.e., not only the cutting of elements, but optimizing size and shape of each element given by the anisotropic parameters given in Definition 3.1.

Based on the estimates (2.32), (2.32) and (3.41), we now present the goal-oriented anisotropic *hp*-mesh adaptation algorithm. First, we starts with its main steps, in order to present it compactly. Later, its individual parts are described in detail in the subsequents sections.

The goal of the algorithm is to generate iteratively a sequence of *hp*-meshes $\mathscr{T}_{h,\boldsymbol{p}}^n$, $n = 0,\ldots,N$ and compute the corresponding approximate primal and adjoint solution $u_h^{(n)}$ and $z_h^{(n)}$ of (2.9) and (2.23), respectively, such that

$$|\eta^{\mathrm{I}}(u_h^{(N)}, z_h^{(N)})| \leq \mathrm{TOL}, \tag{3.42}$$

where η^{I} is the error estimate given by (2.32) and $\mathrm{TOL} > 0$ is the prescribed tolerance. Obviously, the number of the iterative cycles N is a priori unknown. Moreover, we require that the number of degrees of freedom of $\mathscr{T}_{h,\boldsymbol{p}}^N$ is as small as possible.

Algorithm 3: Goal-oriented anisotropic *hp*-mesh adaptation algorithm

```
1  initialization: let $\mathrm{TOL} > 0$ be given, let $\mathscr{T}_{h,\boldsymbol{p}}^0$ be the given initial hp-mesh
     (coarse with low polynomial approximation degrees $p_K$)
2  for $n = 0,1,\ldots$ do
3      construct the DG spaces $S_h^{\boldsymbol{p}}$ and $S_h^{\boldsymbol{p}+1}$ corresponding to $\mathscr{T}_{h,\boldsymbol{p}}^n = \{\mathscr{T}_h^{(n)}, \boldsymbol{p}^n\}$
4      solve primal and adjoint solution problems (2.9) and (2.23), respectively, set
         $u_h^{(n)} \in S_h^{\boldsymbol{p}}$ and $z_h^{(n)} \in S_h^{\boldsymbol{p}}$
5      compute reconstructions $u_h^{+,(n)} = \mathscr{R}(u_h^{(n)}) \in S_h^{\boldsymbol{p}+1}$ and
         $z_h^{+,(n)} = \mathscr{R}(z_h^{(n)}) \in S_h^{\boldsymbol{p}+1}$
6      evaluate $\eta^{\mathrm{I}}(u_h^{(n)}, z_h^{(n)})$ and $\eta_K^{\mathrm{I}}$, $K \in \mathscr{T}_h^{(n)}$
7      if $\eta^{\mathrm{I}}(u_h^{(n)}, z_h^{(n)}) \leq \mathrm{TOL}$ then
8          STOP the mesh adaptation process
9      else
10         foreach $K \in \mathscr{T}_h^{(n)}$ do
11             set better element size $\tilde{\lambda}_K$
12             set better element shape $\tilde{\sigma}_K$ and $\tilde{\phi}_K$
13             set better polynomial approximation degree $\tilde{p}_K$
14         end
15     end
16     from the set of quintets $\{x_K, \tilde{\lambda}_K, \tilde{\sigma}_K, \tilde{\phi}_K, \tilde{p}_K\}$, $K \in \mathscr{T}_h^{(n)}$ generate a new
         hp-mesh $\mathscr{T}_{h,\boldsymbol{p}}^{n+1} = \{\mathscr{T}_h^{(n+1)}, \boldsymbol{p}^{n+1}\}$
17 end
```

We introduce the mesh adaptation algorithm in detail in Algorithm 3. The step 5 is based on one of the reconstructions presented in Sections 2.2.1–2.2.2. In step 16, the in-house code ANGENER Dolejší [2000] is employed. In the following sections, we describe the steps 11–13 in detail. Let us note that in practice, steps 12 and 13 are carried out together.

3.3.1 Element size

There are several possibilities for the optimization of the elements sizes. In Dolejší and Solin [2016], an approach was developed based on the so-called equidistribution principle (cf. [Dolejší and Solin, 2016, Corollary 5.4]). The size of each element was adapted such that

$$|\eta_K^{\mathrm{I}}| \approx \frac{\mathrm{TOL}}{\#\mathscr{T}_h^{(n)}}, \tag{3.43}$$

where TOL is the tolerance from (3.42) and $\#\mathscr{T}_h^{(n)}$ denotes the number of elements of mesh $\mathscr{T}_h^{(n)}$. Unfortunately, the increase of the number of degrees of freedom (DoF) based on the element size adaptation (3.43) is strongly non-monotone and moreover (3.43) requires defining a security constant which is problem-dependent.

A different approach was developed in Dolejší et al. [2018] for the mesh optimization with respect to the interpolation error measured in the L^q-norm ($q \in [1,\infty]$). There the number of DoF was a priori prescribed for each adaptation level of Algorithm 3.

Here, we present another technique developed in Bartoš et al. [2019] which uses the equidistribution principle from Dolejší and Solin [2016], but (usually) leads to a monotone increase of DoF and no problem-dependent security constant are required.

For each adaptation level $n = 0, 1, \dots$ of Algorithm 3, we prescribe the loop tolerance $\tau_n > 0$ which is successively decreasing. More precisely, we put

$$\tau_0 := \frac{1}{10}\left|\eta^{\mathrm{I}}(u_h^{(0)}, u_h^{(0)})\right|, \qquad \tau_n = \tau_0\zeta^n,\ n = 1, 2, \dots, \quad \zeta \in (0,1), \tag{3.44}$$

hence the initial loop tolerance τ_0 is set as one tenth of the error estimate of the solution computed on the initial hp-mesh $\mathscr{T}_{h,p}^0$ and this tolerance is successively decreased in each loop of Algorithm 3 by a constant factor $\zeta \in (0,1)$. This constant represents the rate of increase of number of degrees of freedom (DoF) which is obviously higher for smaller ζ. Values of ζ close to 1 leads to many cycles of Algorithm 3 necessary for the achieving the given tolerance TOL. On the other hand, a small value $\zeta \lessapprox 0.05$ gives a final hp-mesh with more DoF. In practical computations, we use $\zeta = 0.25$. However, values in the range $0.15 \le \zeta \le 0.5$ would lead to similar results. Let us note that there is not any explicit relation between TOL and τ_n, $n = 0, 1, \dots$.

Having prescribed the loop tolerance τ_n for the given level of mesh adaptation $n = 0, 1, \dots$, we set new sizes $\tilde{\lambda}_K$ ($\sim |K|^{1/2}$) of $K \in \mathscr{T}_h^{(n)}$ using the equidistribution principle (cf. [Dolejší and Solin, 2016, Corollary 5.4]), i.e., we intend to have approximately the same size of the error estimate η_K^{I} for each triangle. Since $\eta^{\mathrm{I}}(u_h^{(n)}, z_h^{(n)}) = \sum_{K\in\mathscr{T}_h^{(n)}} \eta_K^{\mathrm{I}}$ we require

$$|\eta_K^{\mathrm{I}}| \approx \frac{\tau_n}{\#\mathscr{T}_h^{(n)}} =: \bar{\tau}_n = \text{const.} \quad \forall K \in \mathscr{T}_h^{(n)}, \tag{3.45}$$

where $\bar{\tau}_n$ can be considered as the tolerance for the local loop following from (3.44). Obviously, if $\eta_K^{\mathrm{I}} > \bar{\tau}_n$ then we have to decrease the element size and wise versa. Particularly, we define a new element size $\tilde{\lambda}_K$ by

$$\tilde{\lambda}_K = \lambda_K q_K, \quad K \in \mathscr{T}_h^{(n)} \tag{3.46}$$

where $q_K > 0$ is the size adaptation factor. Obviously, if $\eta_K^{\mathrm{I}} = \bar{\tau}_n$ then it is natural to put $q_K = 1$, for $\eta_K^{\mathrm{I}} > \bar{\tau}_n$ we have to put $q_K \in (0,1)$ and for $\eta_K^{\mathrm{I}} < \bar{\tau}_n$ we set $q_K > 1$.

In Dolejší et al. [2019], where only the h-variant was developed, the element size adaptation factor was given by

$$q_K = \left(\frac{\bar{\tau}_n}{|\eta_K^{\mathrm{I}}|} \right)^{1/\alpha_K}, \quad K \in \mathscr{T}_h^{(n)}, \tag{3.47}$$

where $\alpha_K > 0$ is a suitably chosen value corresponding to the rate of convergence of the used method. Relation (3.47) follows from the assumed rate of convergence $\eta_K^{\mathrm{I}} = O(\lambda_K^{\alpha_K})$. Based on theoretical results [Georgoulis et al., 2007, Theorem 4.5], in the context of DWR error estimates $\alpha_K = 2p_K$, see Dolejší et al. [2019]. It was shown in Dolejší et al. [2019] that this setting works relatively well for the h-variant ($p_K =$ const. $\forall K \in \mathscr{T}_h$). Unfortunately, for varying p_K, $K \in \mathscr{T}_h$, it leads to several repeating refinements and coarsenings during the mesh adaptation process which lengthens the computational time, see Figure 3.6, center. This is caused by the fact that q_K given by (3.47) changes rapidly for $\eta_K^{\mathrm{I}} \approx \bar{\tau}_n$, i.e., $\frac{d}{d\eta_K^{\mathrm{I}}} q_K(\bar{\tau}_n)$ is large, see Figure 3.2, left.

In order to avoid such defective behavior, (3.47) has to be replaced by a different relation where q_K would be almost constant in vicinity of $\eta_K^{\mathrm{I}} = \bar{\tau}_n$. Based on the ideas from [Balan et al., 2016, Section 4.3] an improved definition of q_K was introduced in Bartoš et al. [2019]. First, the maximal and minimal value of the estimator η_K^{I}, $K \in \mathscr{T}_h$ are defined by

$$\eta_{\max}^{\mathrm{I}} = \max_{K \in \mathscr{T}_h^{(n)}} |\eta_K^{\mathrm{I}}|, \quad \eta_{\min}^{\mathrm{I}} = \min_{K \in \mathscr{T}_h^{(n)}} |\eta_K^{\mathrm{I}}|. \tag{3.48}$$

Further, two user-defined parameters are chosen, namely the maximal refinement factor $r_{\max} \in (0,1)$ and the maximal coarsening factor $c_{\max} > 1$. Finally, the element size adaptation factor q_K is given by

$$q_K = \begin{cases} 1 + (r_{\max} - 1)\xi_K^2, \quad \xi_K := \frac{\log(|\eta_K^{\mathrm{I}}|) - \log(\bar{\tau}_n)}{\log(\eta_{\max}^{\mathrm{I}}) - \log(\bar{\tau}_n)} & \text{for } \eta_K^{\mathrm{I}} \geq \bar{\tau}_n, \\ \\ 1 + (c_{\max} - 1)\xi_K^2, \quad \xi_K := \frac{\log(|\eta_K^{\mathrm{I}}|) - \log(\bar{\tau}_n)}{\log(\eta_{\min}^{\mathrm{I}}) - \log(\bar{\tau}_n)} & \text{for } \eta_K^{\mathrm{I}} < \bar{\tau}_n. \end{cases} \tag{3.49}$$

Obviously, q_K given by (3.49) is almost equal to one in vicinity of $\eta_K^{\mathrm{I}} \approx \bar{\tau}_n$, see Figure3.2.

The efficiency of the whole algorithm can be still slightly improved, see Bartoš et al. [2019], by a small modification of (3.49) in such a way that q_K is almost constant in vicinity of the limit values $\eta_{\min}^{\mathrm{I}}$ and $\eta_{\max}^{\mathrm{I}}$. Namely, we put

$$q_K = \begin{cases} \frac{1}{2}(1 - r_{\max})(\cos(\pi\xi_K) + 1) + r_{\max}, & \xi_K := \frac{\log(|\eta_K^{\mathrm{I}}|) - \log(\bar{\tau}_n)}{\log(\eta_{\max}^{\mathrm{I}}) - \log(\bar{\tau}_n)} \quad \text{if } \eta_K^{\mathrm{I}} \geq \bar{\tau}_n, \\ \\ \frac{1}{2}(c_{\max} - 1)(\cos(\pi(\xi_K + 1)) + 1) + 1, & \xi_K := \frac{\log(|\eta_K^{\mathrm{I}}|) - \log(\bar{\tau}_n)}{\log(\eta_{\min}^{\mathrm{I}}) - \log(\bar{\tau}_n)} \quad \text{if } \eta_K^{\mathrm{I}} < \bar{\tau}_n. \end{cases} \tag{3.50}$$

Figure 3.2, left, shows an example of the dependence of the size adapted factor q_K on the error estimate $|\eta_K^{\mathrm{I}}|$ given by (3.47), (3.49) and (3.50). As mentioned above, the functions given by (3.49) and (3.50) are almost constant for $|\eta_K^{\mathrm{I}}| \approx \bar{\tau}_n$. Therefore the convergence of the algorithm is much more monotone.

Moreover, Figure 3.2, right, compares techniques (3.49) and (3.50) for different values of $\eta^{\mathrm{I}}_{\max}$ and $\eta^{\mathrm{I}}_{\min}$. The following observation can be made. Let $\eta^{\mathrm{I}}_{K} = 10^{-10}$. If $\eta^{\mathrm{I}}_{\min} = 10^{-10}$ then $q_K = 2.5$ for both (3.49) and (3.50). However, if $\eta^{\mathrm{I}}_{\min} = 10^{-12}$ then $q_K \approx 1.75$ for (3.49) but $q_K \approx 2.25$ for (3.50) . Therefore, relation (3.50) is less sensitive to the variation of $\eta^{\mathrm{I}}_{\min}$ than relation (3.49). Similarly, we can observe less sensitivity of (3.50) with respect to $\eta^{\mathrm{I}}_{\max}$. Hence, the maximal or minimal value of η^{I}_{K}, $K \in \mathscr{T}_h^{(n)}$ (achieved on one element) has lower influence on the size modification of the others elements.

Finally, let us note that only two values $r_{\max}$ and $c_{\max}$ are the user-defined parameters and the algorithm is not too sensitive to their choice. In the standard mesh adaption algorithms, where the triangles with highest error estimates are split by halving onto four similar sub-elements, these values corresponds to $r_{\max} = 1/2$ and $c_{\max} = 2$. In [Bartoš et al., 2019, Section 6.1.1.] we investigated the influence of the choice of the parameters $r_{\max} \in (0,1)$ and $c_{\max} > 1$ from (3.50) on the convergence of the hp-adaptive method. These experiments have shown that the convergence of the method changes only slightly for any choice of $c_{\max} > 1$ and $r_{\max} \lessapprox 0.1$. Based on these experiments, we use the values $r_{\max} = 0.1$ and $c_{\max} = 2.5$ in all numerical experiments.

In 3.4, we present several examples comparing the performance of the element size settings (3.47), (3.49) and (3.50). We will see that the modification (3.49) and (3.50) indeed improve the converge of the adaptive algorithm compared to (3.47) in some cases. Moreover, the setting (3.50) is a little more efficient than (3.49).

3.3.2 Element shape

In this section, we describe the optimization of the element shape for the given polynomial approximation degree $\bar{p} \in \mathbb{N}$. In Section 3.3.3 we will use the following technique for $\bar{p} \in \{p_K - 1, p_K, p_K + 1\}$.

In order to determine a better shaped element, we employ the error estimate (3.41). We evaluate η^{III}_{K} and set new aspect ratio and orientation $\tilde{\sigma}_K$ and $\tilde{\phi}_K$ for each $K \in \mathscr{T}_h^{(n)}$ by minimizing η^{III}_{K} w.r.t. $\{\sigma_K, \phi_K\} \in [1,\infty) \times [0,\pi]$. It means that we set

$$\{\sigma_K^{\bar{p}}, \phi_K^{\bar{p}}\} = \operatorname*{arg\,min}_{\sigma \geq 1, \phi \in [0,\pi)} \eta^{\mathrm{III}}_{K}(\lambda_K, \sigma, \phi, \bar{p}), \quad K \in \mathscr{T}_h^{(n)}, \tag{3.51}$$

where η^{III}_{K} is given by (3.40).

Relation (3.21) implies that $\boldsymbol{G}$, and consequently $\mathscr{G}$, are π-periodic with respect to ϕ. Furthermore, using a similar technique as in [Dolejší, 2014, §3.3.1], we can prove that $\boldsymbol{G} \to \infty$ for $\sigma \to \infty$ (and thus $\mathscr{G} \to \infty$ for $\sigma \to \infty$). Hence, the continuity of $\mathscr{G}$ with respect to σ and ϕ implies the existence of at least one minimum of (3.51).

Although the uniqueness of the minimum is not guaranteed, we carried out hundreds of numerical experiments for different input parameters and we did not observe an existence of more than one local minimum. The only exception is the case $\sigma = 1$ since $\mathscr{G}(\cdot, 1, \phi) = \text{const.}$ for any $\phi \in [0, \pi)$. However, when the minimum is attained for $\sigma = 1$ then the element is isotropic and its orientation ϕ is not relevant.

The minimal value of (3.51) is solved numerically. We start with the value $\{\sigma_K, \phi_K\}$ corresponding to the anisotropy of $K \in \mathscr{T}_h^{(n)}$. With the chosen steps $\Delta\sigma > 1$ and $\Delta\phi \geq 0$, we test successively the values of $\eta^{\mathrm{III}}_{K}(\tilde{\lambda}_K, \sigma, \phi \pm \Delta\phi, \bar{p})$ and $\eta^{\mathrm{III}}_{K}(\tilde{\lambda}_K, \sigma\Delta\sigma^{\pm 1}, \phi, \bar{p})$. From them we choose the pair σ, ϕ giving the smallest value of η^{III}_{K} and repeat the

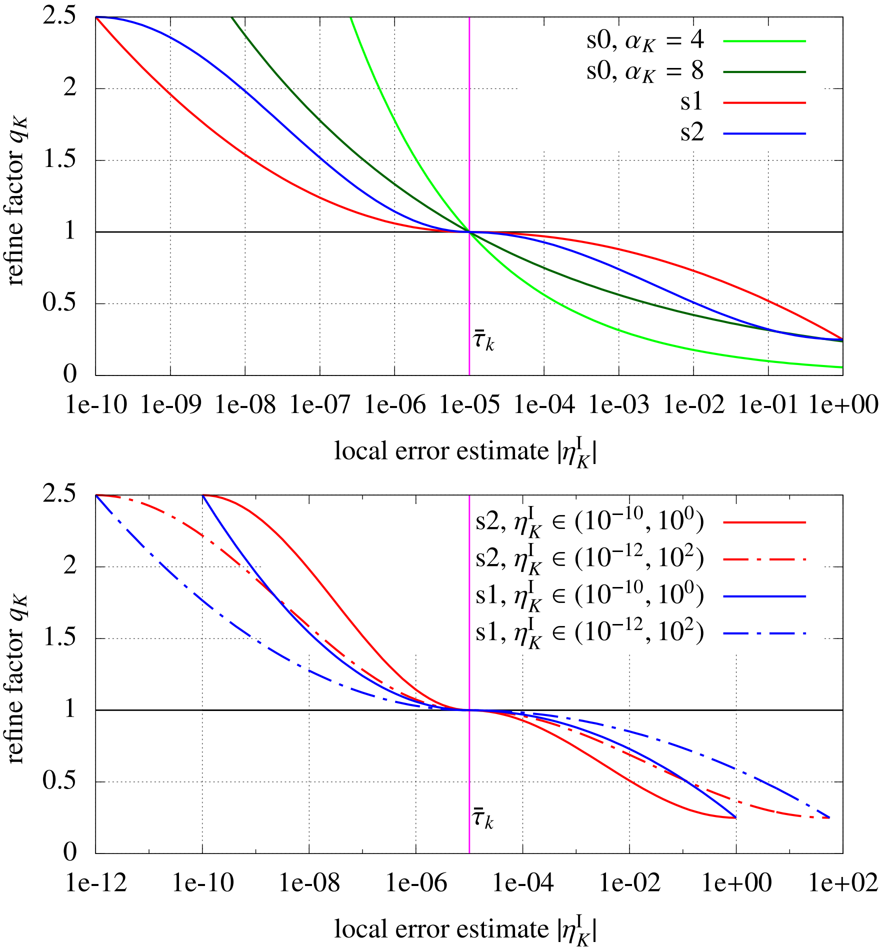

Figure 3.2: The dependence of the size adapted factor q_K on $|\eta_K^{\mathrm{I}}|$ for the values $\bar{\tau}_n = 10^{-5}$, $r_{\max} = 0.25$ and $c_{\max} = 2.5$; s0 – relation (3.47) for $\alpha_K = 4$ and $\alpha_K = 8$, s1 – relation (3.49) and s2 – relation (3.50); (a) all techniques for $\eta^{\mathrm{I}}_{\max} = 1$ and $\eta^{\mathrm{I}}_{\min} = 10^{-10}$; (b) s1 and s2 techniques for $\eta^{\mathrm{I}}_{\max} = 1$, $\eta^{\mathrm{I}}_{\min} = 10^{-10}$ and $\eta^{\mathrm{I}}_{\max} = 10^2$, $\eta^{\mathrm{I}}_{\min} = 10^{-12}$.

search. When we find σ, ϕ such that

$$\eta_K^{\mathrm{III}}(\lambda_K, \sigma, \phi, \bar{p}) \leq \eta_K^{\mathrm{III}}(\lambda_K, \sigma, \phi \pm \Delta\phi, \bar{p}),$$
$$\eta_K^{\mathrm{III}}(\lambda_K, \sigma, \phi, \bar{p}) \leq \eta_K^{\mathrm{III}}(\lambda_K, \sigma\Delta\sigma^{\pm 1}, \phi, \bar{p}),$$

we decrease the steps $\Delta\sigma := \sqrt{\Delta\sigma}$ and $\Delta\phi := \Delta\phi/2$ and repeat the searching until the required tolerances $\omega_\sigma > 0$ and $\omega_\phi > 0$ for the accuracy of σ and ϕ are achieved.

The decrease of step $\Delta\phi$ is done arithmetically since it represents an absolute tolerance for $\phi > 0$. On the other hand, the decrease of $\Delta\sigma > 1$ is done geometrically since the difference $\sigma - 1$ represents the relative tolerance for σ. We put the relative tolerance for σ equal to 1% ($\omega_\sigma = 1.01$) and the absolute tolerance for ϕ equal to 1° ($\omega_\phi = \pi/180$). For the initial steps, we use the values $\Delta\sigma = 1.5$ and $\Delta\phi = \pi/10$. This technique is illustrated by Figure 3.3. Algorithm 4 describes more precisely the individual steps of this process.

For the initial levels of the adaptation Algorithm 3 (low n), the shape optimization Algorithm 4 requires several dozens of loops to achieve to the given accuracy but the number of elements of $\mathscr{T}_h^{(n)}$ is low. On the other hand, on finer meshes $\mathscr{T}_h^{(n)}$ (higher n), the initial approximations of σ and ϕ are already very good (from the previous mesh adaptation level), and hence only few loops in Algorithm 4 needs to be performed.

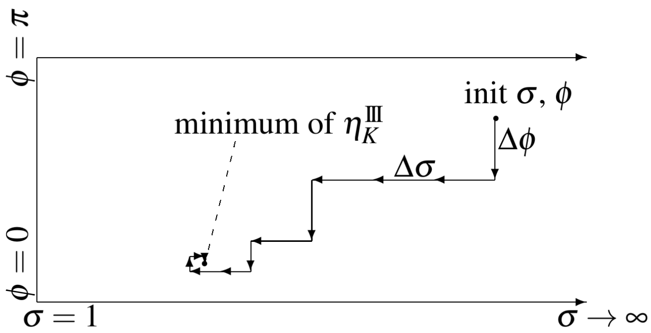

Figure 3.3: Illustration of the iterative process seeking the minimum of η_K^{III}.

3.3.3 Element polynomial approximation degree

Now we describe the process of choosing the suitable polynomial degree in the step 13 of Algorithm 3. We adopt the technique from Dolejší et al. [2018]. We exploit so-called *density of degrees of freedom*, see Dolejší [2014] or Dolejší et al. [2018] for details, which represents the the ratio of the number of degrees of freedom DoF and the volume of the area of K.

In order to determine which polynomial approximation degree should be chosen for each element $K \in \mathscr{T}_h^{(n)}$, we take three candidates for $\tilde{p}_K$, namely $p_K + i$, $i \in \{-1, 0, 1\}$. Then we set the new element size proposals $\lambda_K^{(p_K+i)}$, $i \in \{-1, 0, 1\}$ such that the corre-

Algorithm 4: Element shape optimization – minimization of $\eta_K^{\mathrm{III}}(\tilde{\lambda}_K, \sigma, \phi, \bar{p})$

```
**initialization:** let the required tolerances $\omega_\sigma > 0$ and $\omega_\phi > 0$ be given
set initial steps $\Delta\sigma > 1$ and $\Delta\phi \geq 0$ (multiplies of $\omega_\sigma$ and $\omega_\phi$)
let initial pair $\sigma, \phi$ be given
**while** $\Delta\sigma > \omega_\sigma$ or $\Delta\phi > \omega_\phi$ **do**
    **for** $i = 0, 1, \dots$ **do**
        $g^0 := \eta_K^{\mathrm{III}}(\tilde{\lambda}_K, \sigma, \phi, \bar{p})$, $g^+ := \eta_K^{\mathrm{III}}(\tilde{\lambda}_K, \sigma, \phi + \Delta\phi, \bar{p})$,
          $g^- := \eta_K^{\mathrm{III}}(\tilde{\lambda}_K, \sigma, \phi - \Delta\phi, \bar{p})$
        **while** $g^0 > \min\{g^+, g^-\}$ **do**
            **if** $g^0 > g^+$ and $g^- > g^+$ **then**
                $\phi := \phi + \Delta\phi$
            **else**
                $\phi := \phi - \Delta\phi$
            **end**
        **end**
        $g^0 := \eta_K^{\mathrm{III}}(\tilde{\lambda}_K, \sigma, \phi, \bar{p})$, $g^+ := \eta_K^{\mathrm{III}}(\tilde{\lambda}_K, \sigma\Delta\sigma, \phi, \bar{p})$,
          $g^- := \eta_K^{\mathrm{III}}(\tilde{\lambda}_K\tilde{\lambda}_K, \sigma/\Delta\sigma, \phi, \bar{p})$
        **while** $g^0 > \min\{g^+, g^-\}$ **do**
            **if** $g^0 > g^+$ and $g^- > g^+$ **then**
                $\sigma := \sigma \cdot \Delta\sigma$
            **else**
                $\sigma := \sigma/\Delta\sigma$
            **end**
        **end**
        $\Delta\phi := \Delta\phi/2$; $\Delta\sigma := \sqrt{\Delta\sigma}$
    **end**
**end**
```

sponding density of degrees of freedom is the same, i.e.,

$$\lambda_K^{(p_K-1)} = \lambda_K \left(\frac{p_K(p_k+1)}{(p_K+1)(p_K+2)} \right)^{1/2}, \tag{3.52}$$

$$\lambda_K^{(p_K)} = \lambda_K, \tag{3.53}$$

$$\lambda_K^{(p_K+1)} = \lambda_K \left(\frac{(p_K+2)(p_k+3)}{(p_K+1)(p_K+2)} \right)^{1/2}, \tag{3.54}$$

where λ_K corresponds to the size of $K \in \mathscr{T}_h^{(n)}$ given by Definition 3.1. We remind that $|K| = 3\sqrt{3}/4\lambda_K^2$.

Furthermore, for each of the candidates $p_K + i$, $i \in \{-1, 0, 1\}$, we find the optimal shape using the technique described in Section 3.3.2, i.e., we determine $\sigma_K^{(p_K+i)}$ and $\phi_K^{(p_K+i)}$, $i \in \{-1, 0, 1\}$ such that

$$\{\sigma_K^{(p_K+i)}, \phi_K^{(p_K+i)}\} = \underset{\sigma \geq 1, \phi \in [0,\pi)}{\arg\min} \; \eta_K^{\mathrm{III}}(\lambda_K^{(p_K+i)}, \sigma, \phi, p_K + i), \quad i \in \{-1, 0, 1\}. \tag{3.55}$$

Finally, from the three candidates $p_K+i,\ i\in\{-1,0,1\}$ we choose the one, which has the minimal value of the estimate, i.e.,

$$\{\tilde{\sigma}_K, \tilde{\phi}_K, \tilde{p}_K\} = \operatorname*{arg\,min}_{i\in\{-1,0,1\}} \eta_K^{\mathrm{III}}(\lambda_K^{(p_K+i)}, \sigma_K^{(p_K+i)}, \phi_K^{(p_K+i)}, p_K+i). \tag{3.56}$$

Remark. The evaluation of $\eta_K^{\mathrm{III}}(\cdot,\cdot,\cdot,p_K+i)$, $i\in\{-1,0,1\}$ requires the approximation of the partial derivatives of degree p_K+i+1. For that we employ the least-square reconstruction technique from Section 2.2.1.

3.4 Numerical experiments

In this section, we present several numerical experiments, which demonstrate the efficiency, accuracy and robustness of Algorithm 3. We present a comparison with *hp*-isotropic mesh adaptation and the *h*-anisotropic variant of the presented technique.

We deal with three problems:

- purely elliptic problem on a "cross" domain with four corner singularities,
- mixed hyperbolic-elliptic problem,
- convection-diffusion equation with a dominating convection with three different target functionals.

For each case, we carried out several variants of computations:

- *hp*-AMA_s0: full *hp*-anisotropic mesh adaptation Algorithm 3 with the element size setting given by (3.47),
- *hp*-AMA_s1: full *hp*-anisotropic mesh adaptation Algorithm 3 with the element size setting given by (3.49),
- *hp*-AMA_s2: full *hp*-anisotropic mesh adaptation Algorithm 3 with the element size setting given by (3.50),
- *hp*-IMA: *hp*-isotropic mesh adaptation, we apply Algorithm 3 with the element size setting given by (3.50) and the aspect ratios are fixed $\sigma_K = 1$ for all $K\in\mathscr{T}_h^{(n)}$, $n=0,1,2,\dots$
- *h*-AMA: *h*-anisotropic mesh adaptation, we apply Algorithm 3 with the element size setting given by (3.50) and the polynomial approximation degrees are fixed $p_K=3$ for all $K\in\mathscr{T}_h^{(n)}$, $n=0,1,2,\dots$
- *hp*-ISO: *hp*-isotropic mesh adaptation where, at each adaptation level, the fixed per cent q of elements having the highest $|\eta_K^{\mathrm{I}}|$ are either split into 4 similar subtriangles or the polynomial approximation degree is increased by 1. Therefore, the meshes $\mathscr{T}_h^{(n)}$, $n=0,1,\dots$ are nested and no re-meshing is performed. In [Bartoš et al., 2019, Section 6.1.2.] we have examined the sensitivity of the convergence of the algorithm with respect to the choice of the percentage q of refined elements. The convergence of the error estimate $\eta^{\mathrm{I}} = \eta^{\mathrm{I}}(u_h^{(n)}, z_h^{(n)})$ was compared with respect to DoF and with respect to the computational time for the values $q\in\{10\%, 20\%, 30\%\}$. Based on these experiments we choose $q=20\%$, which seemed as a reasonable compromise giving the most stable results.

- *hp*-ideal: *hp*-isotropic mesh adaptation the same as the *hp*-ISO approach above but a priori knowledge of the regularity of the primal/adjoint solutions is used for the decision criterion between *h*- or *p*-refinement. This technique can be used only for problems where such information is available, here only the first example in Section3.4.1.

We compare these adaptive techniques for the problems mentioned above. The subsequent figures show the decay of the error estimate η^{I} with respect to $\mathrm{DoF}^{1/3}$ in logarithmic-linear scale. The reason of that is to demonstrate an exponential rate of convergence with respect to the DoF in the sense of the result from Gui and Babuška [1986], Demkowicz et al. [2002], Chaillou and Suri [2007]

$$|J(u) - J(u_h)| \approx C \exp\left(-b\,\mathrm{DoF}^{1/3}\right), \tag{3.57}$$

where $C > 0$ and $b > 0$ are constants independent of DoF. Geometrically, the exponential rate of convergence corresponds to a straight line in logarithmic-linear scale.

Moreover, we plot also the decay of the error estimate η^{I} with respect to the computational time. Let us note that the used code is not optimized with respect to the computational time, hence these figures have only an informative character.

3.4.1 Elliptic problem on a "cross" domain

We revisit the elliptic problem on a "cross" domain from Section 2.4.1. We recall the Figure 2.4 (left) which shows the computational domain with the initial mesh and the domain of interest Ω_J. Although there is no anisotropic feature, it serve as a good test example for *hp*-adaptation techniques due to the presence of singularities at interior corners.

Our goal is to compare all adaptive techniques mentioned at the beginning of Section 3.4. For the *hp*-ideal method, we perform the *h*-refinement if a vertex of a marked element is the interior corner of Ω, otherwise the *p*-adaptation is applied.

The differences among the *hp*-variants are almost negligible since this example has no anisotropic feature. Even the *hp*-AMA_s0 variant evinces very good convergence, c.f. Section 3.4.2 where the difference between *hp*-AMA_s0 and *hp*-AMA_s1 or *hp*-AMA_s2, respectively, in not negligible. A very interesting comparison is that of the *hp*-AMA_s2 variant with the *hp*-ideal one. The latter one achieves the given tolerance with smaller number of DoF (Figure 3.4, left) but the computational time is practically the same for both techniques (Figure 3.4, center). It is caused by the fact that the *hp*-ideal variant needs 18 cycles of adaptations whereas the *hp*-AMA_s2 one only 12 cycles.

Furthermore, Figure 3.4, right shows the comparison of the error estimates η^{I} and η^{II} with the actual error $J(u - u_h)$ for the *hp*-AMA_s2 method. Obviously, the error is underestimated but both estimates give good approximation of the magnitude of the error. Therefore, in this case, the estimate (3.7) does not cause an essential increase of the estimate.

Finally, Figure 3.5 shows the final *hp*-mesh, total view and its zoom of the corner singularity at $x = (1,1)$. We observe high polynomial degrees in the major part of Ω, only in the vicinity of the corner singularity a strong *h*-refinement with a low polynomial degree is presented. This demonstrates that the *p*-decision criterion proposed in Section 3.3.3 is able to detect the solution singularity.

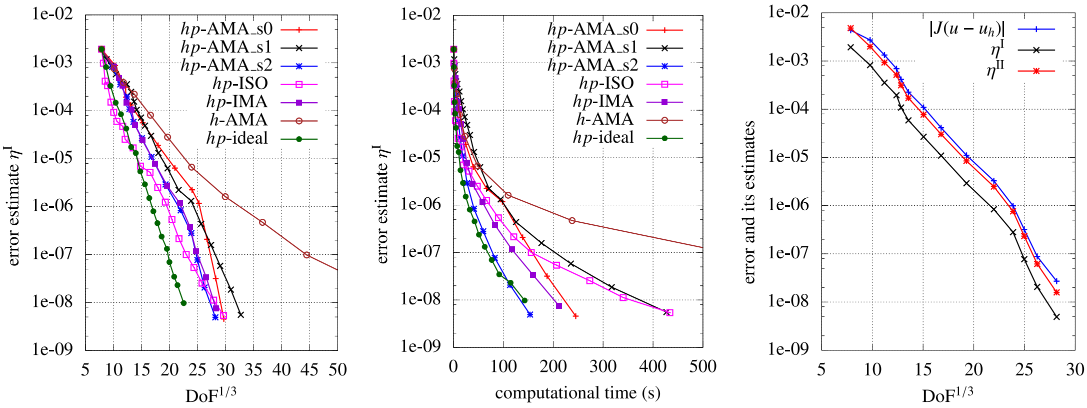

Figure 3.4: Elliptic problem on a "cross" domain, convergence of the error estimate $\eta^{\mathrm{I}} = \eta^{\mathrm{I}}(u_h^{(n)}, z_h^{(n)})$ w. r. t. DoF (left) and w. r. t. the computational time (center) for all tested adapted strategies. Comparison of the error estimates η^{I} and η^{II} with the actual error $J(u-u_h)$ for the *hp*-AMA_s2 method (right).

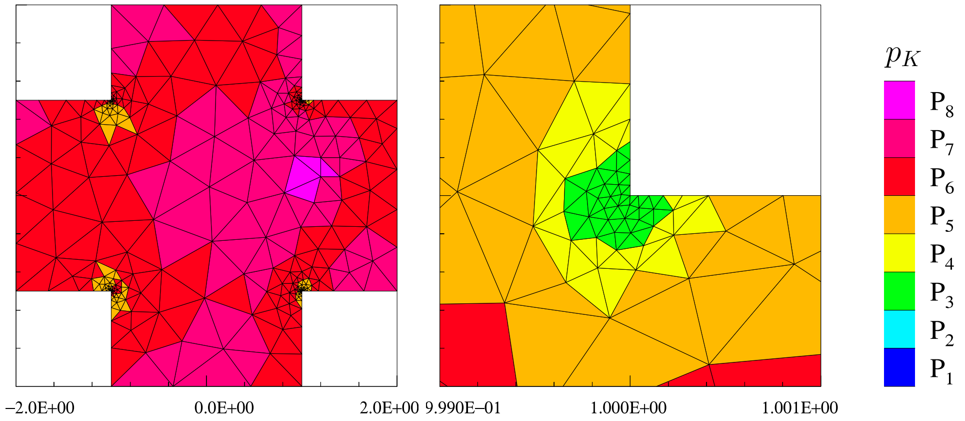

Figure 3.5: Elliptic problem on a "cross" domain – the computational domain Ω with the final *hp*-mesh obtained by the *hp*-AMA_s2 method, total view (left) and a 1000x zoom of the corner singularity at $x = (1,1)$ (right).

3.4.2 Mixed hyperbolic-elliptic problem

In the second example we investigate the performance of the described method for the discretization of elliptic problem (2.1) from [Harriman et al., 2003, Example 2]. We set $\Omega = (0,1)^2$ and $\mathbb{A} = \varepsilon\mathbb{I}$, where

$$\varepsilon = \frac{\delta}{2}\left(1 - \tanh\left(\frac{(r-\frac{1}{4})(r+\frac{1}{4})}{\gamma}\right)\right),$$

$r = \sqrt{(x-1/2)^2 + (y-1/2)^2}$ and $\delta > 0, \gamma > 0$ are constants.

Further, we set $\boldsymbol{b} = (2y^2 - 4x + 1, y + 1)$, $c = -\nabla \cdot \boldsymbol{b} = 3$ and $f = 0$. We choose $\delta = 0.01$ and $\gamma = 0.05$. In this case, the diffusion coefficient ε will be approximately equal to δ in the circle with center $[1/2, 1/2]$ and diameter $1/4$. As r increases over $1/4$, ε quickly decreases reaching values very close to zero ($\approx 10^{-16}$) at the boundary. Therefore, even though the problem is elliptic theoretically, from the computational view the problem behaves like mixed hyperbolic-elliptic problem, since convection is dominating in the region where $r > 1/4$.

The characteristics associated with the convective part of the operator enter the domain Ω through the horizontal edge along $y = 0$ and the vertical edges along $x = 0$ and $x = 1$. We prescribe the Dirichlet boundary condition on this "inflow" part of the boundary $\Gamma_D = \{(x,y) \in \partial\Omega : x = 0 \text{ or } x = 1 \text{ or } y = 0\}$

$$u_D = \begin{cases} 1 & \text{if } x = 0 \text{ and } 0 < y \leq 1, \\ \sin^2(\pi x) & \text{if } 0 \leq x \leq 1 \text{ and } y = 0, \\ e^{-50y^4} & \text{if } x = 1 \text{ and } 0 < y \leq 1, \end{cases} \tag{3.58}$$

which lead to discontinuities in the solution. On the rest of the boundary $\partial\Omega\backslash\Gamma_D$ we prescribe homogeneous Neumann boundary condition. The isocurves of the solution are pictured in the left panel of Figure 3.7, center.

Finally, we choose the target functional as an integral over part of the Neumann boundary

$$J(u) = \int_{0.25}^{0.625} u(x,1)\,\mathrm{d}x \approx 0.324026769433. \tag{3.59}$$

Since the exact solution is unknown we used the reference value $J(u)$ computed by a strong hp-refinement.

This problem has a middle anisotropic feature around the circle $\{(x_1,x_2);\ (x_1 - 0.5)^2 + (x_2 - 0.5)^2 = 1/16\}$. Moreover, the adjoint solution has two line singularity starting at $(x_1,x_2) = (0.25, 1)$ and $(x_1,x_2) = (0.625, 1)$ and going in opposite direction of the convective field $\boldsymbol{b}$, see the sketch in Figure 3.7, right (black lines). These lines are smeared by the diffusion but it is negligible outside of the circle mentioned above.

Remark. We note that due to steep changes of the coefficients $\mathbb{A}(x), \boldsymbol{b}(x)$, the evaluation of the total error (and hence also of the error estimates η_S, η_S^*) is polluted by the errors in numerical integration. The estimate of the quadrature errors are not considered in the presented approach, hence we used an overkill degree of numerical quadrature to suppress these errors.

In Table 3.1 the decrease of the error of the target functional $J(e_h)$ is listed together with the effectivity indices (in brackets) on adaptively refined meshes with N_h elements based on the h-ISO method with fixed $p = 1$. We compare results where

η^{I}_K was replaced by $\eta_{\mathrm{S,K}}$, c.f. (2.33), with z_h^+ being the solution of the adjoint problem (2.15) in S_h^{p+1} (column denoted by η_S^+), estimate based on the reconstruction technique based on solving local problems presented in Section 2.2.2 (η_S^{loc}), and finally weighted least-square reconstruction from Section 2.2.1 (η_S^{LS}) and weighted least-square reconstruction used for the primal problem, i.e., η^{I}_K was replaced by $\eta^*_{\mathrm{S,K}}$, ($\eta_S^{*,LS}$). We observe that even though all of the methods based on reconstruction of the solution underestimate the true error (effectivity indices $i_{\mathrm{eff}} < 1$), the resulting adaptively refined meshes lead to very similar approximation of $J(u_h)$ as the method based on solving the adjoint problem in enriched function space S_h^{p+1} which is computationally much more demanding. We recall that due to the equivalence (2.43) there is no meaning in comparing η_S^{loc} with $\eta_S^{*,loc}$ for the reconstruction based on solving local problems.

Further, we compare computations by all adaptive techniques mentioned at the beginning of Section 3.4 except the *hp*-ideal method since the regularity of the primal and adjoint solutions are not known. Figure 3.6, left and center show the error decay of the error estimate η^{I} with respect to $\mathrm{DoF}^{1/3}$ and with respect to the computational time, respectively. The dominance of the *hp*-AMA_s1 and *hp*-AMA_s2 technique is obvious, namely from the point of view of the computational time. On the other hand, the *hp*-AMA_s0 variant shows poor convergence as was discussed in Section 3.3.1. Additionally, the *h*-AMA technique has a very fast convergence for $\eta^{\mathrm{I}} \geq 5 \cdot 10^{-9}$. This is caused by the presence of discontinuities where the p-adaptation has not a large impact. However, for lower values of η^{I}, the influence of the smooth part of the solution is already non-negligible and the techniques allowing *hp*-adaptation start to dominate. Furthermore, Figure 3.6, right, shows the comparison of the error estimates η^{I} and η^{II} with the actual error $J(e_h)$ for the *hp*-AMA_s2 method. The corresponding final *hp*-mesh and the isolines of the primal and adjoint solutions are plotted in Figure 3.7.

The convergence of the error estimators considerably differs from the elliptic case in Section 3.4.1. The error estimator η^{II} is much larger than η^{I} and, moreover, the estimate η^{II} has no tendency to converge. This is caused by the fact that η^{I} is strongly over-estimated by η^{II} in (3.7), namely the arguments of the scalar product in (3.1) and (2.18) seem to be locally (almost) orthogonal and hence the use of the Cauchy inequality in (3.4) and (3.6) leads to rough bounds.

In order to decrease the error estimate η^{II}, it is possible to modify Algorithm 3 by using η^{II}_K instead of η^{I}_K, $K \in \mathscr{T}_h^{(n)}$ when determining the new element size. Figure 3.8 shows the corresponding results. In contrast with the results shown in Figure 3.7, the mesh is adapted along the left singularity line (coloured in black) of the adjoint solution shown in Figure 3.7, right. Consequently, this singularity line is well resolved (compare the isolines plotted in Figure 3.7, right and Figure 3.8, right.) On the other hand, Figure 3.6, right and Figure 3.8, center, show that using η^{II}_K, $K \in \mathscr{T}_h^{(n)}$ for determining the element sizes instead of η^{I}_K, $K \in \mathscr{T}_h^{(n)}$ requires a little more DoF for achieving the same error level. We conclude that the use of η^{II} and η^{II}_K, $K \in \mathscr{T}_h^{(n)}$ in the mesh adaptation algorithm is possible but may not be always efficient.

3.4.3 Convection-dominated problem

Let us now consider problem (2.1) taken from Formaggia et al. [2004], see also Carpio et al. [2013]. The domain Ω is an L-shaped region given by $[0,4] \times [0,4] \setminus [0,2] \times [0,2]$

η_S^+		η_S^{loc}		η_S^{LS}		$\eta_S^{*,LS}$	
N_h	$J(e_h)$	N_h	$J(e_h)$	N_h	$J(e_h)$	N_h	$J(e_h)$
128	2.02×10^{-3} (0.96)	128	2.02×10^{-3} (0.26)	128	2.02×10^{-3} (0.55)	128	2.02×10^{-3} (0.90)
203	9.12×10^{-4} (1.04)	203	1.12×10^{-3} (0.24)	203	1.37×10^{-3} (0.31)	203	1.48×10^{-3} (0.80)
323	2.99×10^{-4} (1.11)	350	6.67×10^{-4} (0.56)	323	4.40×10^{-4} (0.53)	338	5.75×10^{-4} (0.91)
536	1.89×10^{-4} (1.00)	566	2.45×10^{-4} (0.16)	518	2.20×10^{-4} (0.64)	560	2.99×10^{-4} (0.77)
899	9.77×10^{-5} (1.04)	938	2.14×10^{-4} (0.66)	839	1.53×10^{-4} (0.50)	935	1.13×10^{-4} (0.93)
1460	5.31×10^{-5} (1.08)	1541	9.49×10^{-5} (0.20)	1367	7.96×10^{-5} (0.53)	1541	5.23×10^{-5} (0.87)
2381	2.17×10^{-5} (0.99)	2543	5.67×10^{-5} (0.30)	2198	2.58×10^{-5} (0.94)	2555	2.42×10^{-5} (1.12)
3899	1.42×10^{-5} (1.00)	4157	3.42×10^{-5} (0.47)	3569	1.65×10^{-5} (0.94)	4160	8.56×10^{-6} (1.27)
6305	1.00×10^{-5} (1.00)	6755	1.87×10^{-5} (0.26)	5765	1.18×10^{-5} (0.89)	6758	1.02×10^{-5} (0.88)
10223	4.59×10^{-6} (1.00)	10961	1.03×10^{-5} (0.58)	9272	5.29×10^{-6} (0.80)	10958	4.52×10^{-6} (0.93)
16475	3.07×10^{-6} (1.04)	17723	5.12×10^{-6} (0.39)	14927	3.61×10^{-6} (0.90)	17708	3.22×10^{-6} (0.97)

Table 3.1: Mixed hyperbolic-elliptic problem – decrease of $J(e_h)$ and the corresponding effectivity indices (in brackets) for h-ISO method with fixed $p = 1$.

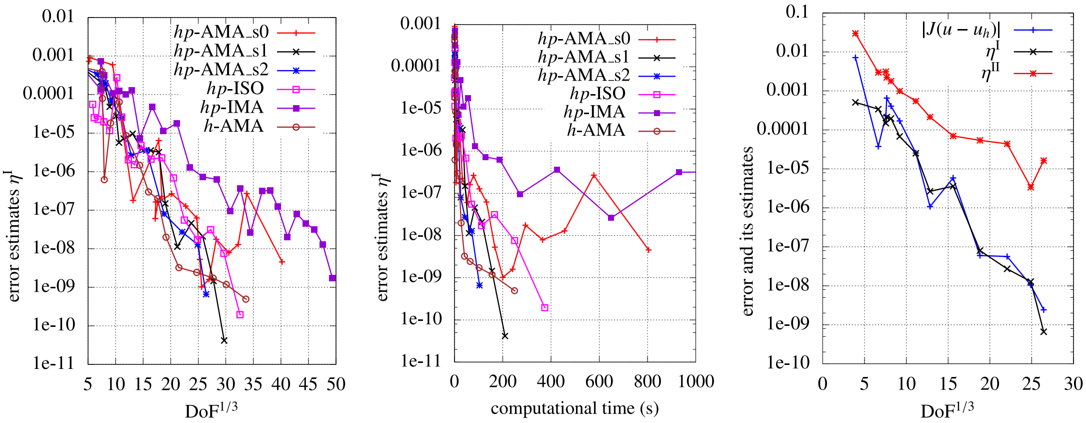

Figure 3.6: Mixed hyperbolic-elliptic problem – convergence of the error estimate $\eta^{\mathrm{I}} = \eta^{\mathrm{I}}(u_h^{(n)}, z_h^{(n)})$ w. r. t. DoF (left) and w. r. t. the computational time (center) for all tested adapted strategies. Comparison of the error estimates η^{I} and η^{II} with the actual error $J(u - u_h)$ for the hp-AMA_s2 method (right).

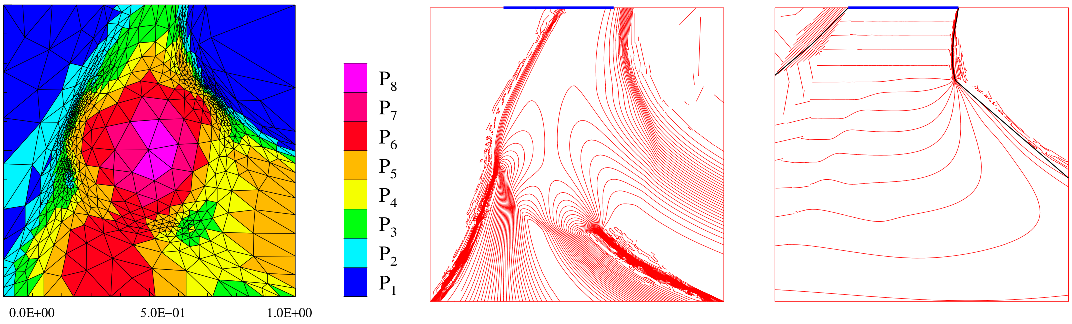

Figure 3.7: Mixed hyperbolic-elliptic problem – the final hp-mesh (left), the isolines of the primal (center) and the adjoint (right) solutions computed in the final grid by the hp-AMA_s2 method. The blue line is the "domain of interest", the black lines are the lines of singularity of the adjoint solution.

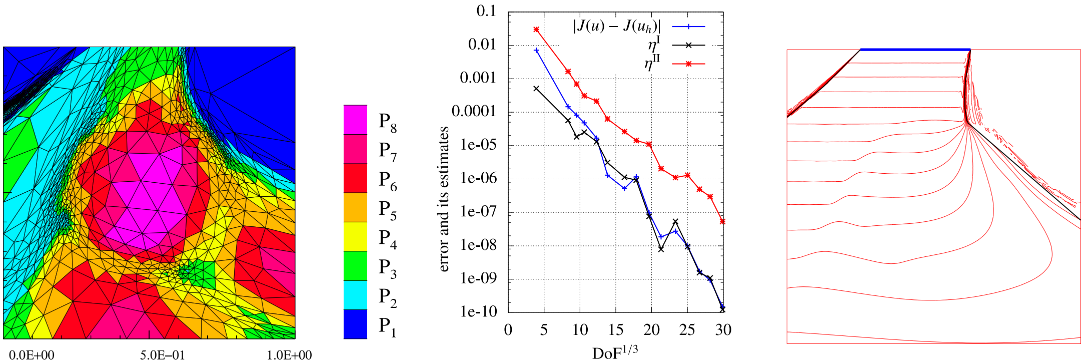

Figure 3.8: Mixed hyperbolic-elliptic problem – computation by the hp-AMA_s2 method with replacing $\eta_K^{\rm I}$ by $\eta_K^{\rm II}$, $K \in \mathscr{T}_h^{(n)}$ in Algorithm 3, the final hp-mesh (left), comparison of the error estimates $\eta^{\rm I}$ and $\eta^{\rm II}$ with the actual error $J(u - u_h)$ (center) and the isolines of the adjoint solution (right).

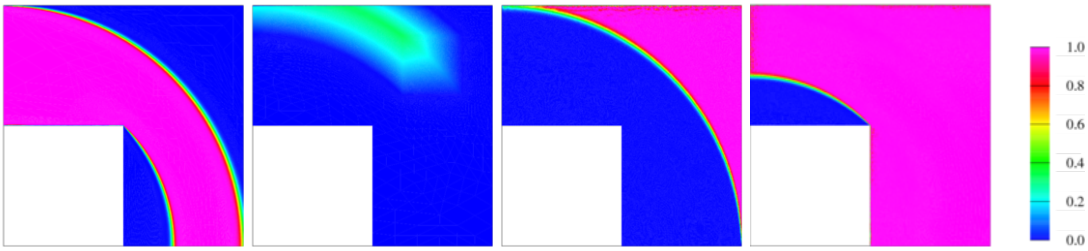

Figure 3.9: Convection-dominated problem, the solutions of primal problem (left) and the adjoint problems with J_V (second), with J_B (third) and with J_D (right).

and we solve

$$-\varepsilon\Delta u + \nabla\cdot(\boldsymbol{b}u) = 0 \qquad \text{in } \Omega, \tag{3.60}$$

where $\varepsilon = 10^{-3}$ and the convection field $\boldsymbol{b} = (x_2, -x_1)$. We prescribe the Dirichlet and Neumann boundary conditions

$$\begin{aligned} u &= 1 && \text{on } \{x_1 = 0\}, \\ \nabla u\cdot\boldsymbol{n} &= 0 && \text{on } \Gamma_1 := \{x\in\partial\Omega; x_1 = 4\}\cup\Gamma_2 := \{x\in\partial\Omega; x_2 = 0\}, \\ u &= 0 && \text{elsewhere.} \end{aligned} \tag{3.61}$$

The solution u exhibits boundary layers as well as two circular-shaped internal layers.

We consider three target functionals $J_V(u)$, $J_B(u)$ and $J_D(u)$ defined hereafter together with the reference values computed by a strong hp-refinement. Let us note that these values slightly differ from those in Carpio et al. [2013].

$$\begin{aligned} J_V(u) &= \int_E u(x)\,\mathrm{d}x \approx 0.20314158 \pm 10^{-8}, && E := (2.5, 3.5)\times(2.5, 3.5), \\ J_B(u) &= \int_{G_B} \boldsymbol{b}\cdot\boldsymbol{n}\,u\,\mathrm{d}S \approx 0.07408122 \pm 10^{-8}, && G_B := \Gamma_1, \\ J_D(u) &= \int_{G_D} \boldsymbol{b}\cdot\boldsymbol{n}\,u\,\mathrm{d}S \approx 3.9670304 \pm 10^{-7}, && G_D := \Gamma_1\cup\Gamma_2, \end{aligned} \tag{3.62}$$

Figure 3.9 shows the solution of the primal problem (3.60) and the solutions of the adjoint problems corresponding to the functionals $J_V(u)$, $J_B(u)$ and $J_D(u)$.

We carried out computations for all three target functionals using all adaptive techniques mentioned at the beginning of Section 3.4 except the hp-ideal method since the regularity of the primal and adjoint solutions are not obvious again. Figure 3.10, left and center show the error decay of the error estimate η^{I} with respect to DoF$^{1/3}$ and with respect to the computational time, respectively. Since primal as well as adjoint solutions have anisotropic features for all three examples (hp)-anisotropic variants are dominant. Again the hp-AMA_s0 methods gives poor convergence for J_V and J_D target functionals. Moreover, the hp-AMA_s2 technique is slightly superior to the hp-AMA_s1 but the difference in the convergence is very small, namely for J_B and J_D. This is caused probably by that fact that for these examples the distribution of η^{I}_K, $K\in\mathscr{T}_h^{(n)}$ is relatively smooth and then the smaller sensitivity of the hp-AMA_s2 technique with respect to $\eta^{\mathrm{I}}_{\max}$ and $\eta^{\mathrm{I}}_{\min}$ does not play an important role.

Furthermore, Figure 3.10, right shows the comparison of the error estimates η^{I} and η^{II} with the actual error $J(e_h)$ for the hp-AMA_s2 method. Both estimators give a reasonable estimate of the magnitude of the error. Here, unlike the problem in Section 3.4.2, η^{II} gives in fact an upper bound of the error.

Finally, the solutions of the primal problem together with the final hp-meshes are presented in Figures 3.11, 3.14. For the case with the functional J_V, we observe a strong refinement along the part of the outer circular-shaped interior layer entering to the domain of interest $E := (2.5, 3.5) \times (2.5, 3.5)$ and also a refinement behind the square E. It can be surprising since the adjoint solution is (almost) constant in front of E in the opposite direction of the convection field. Therefore the adjoint residuals as well as the adjoint weights should be negligible. We suppose that the considered diffusion $\varepsilon = 10^{-3}$ plays also a role which leads to the refinement mentioned above. In order to support this argumentation, we carried out the computation of the same example but with $\varepsilon = 10^{-6}$. The solutions of the primal problem together with the final hp-meshes are presented in Figure 3.12. Obviously, the refinement is just inside of E. Moreover, the interior layer is thinner and lower polynomial approximation degrees are generated.

For the case with the functional J_B, we observe a strong anisotropic refinement along the entire outer circular-shaped interior layer since it attaches the domain of region G_B, cf. (3.62). On the other hand there is almost no refinement along the inner circular-shaped interior since it does no influence the value of the solution on G_B. On the other hand, for the case with the functional J_D, both interior layers act on G_D. However, we do not observe refinement along both interior layers since J_D is the mean value of u over G_D and therefore the smeared layer lead to the same values. A strong refinement is present only in regions where both interior layers begin.

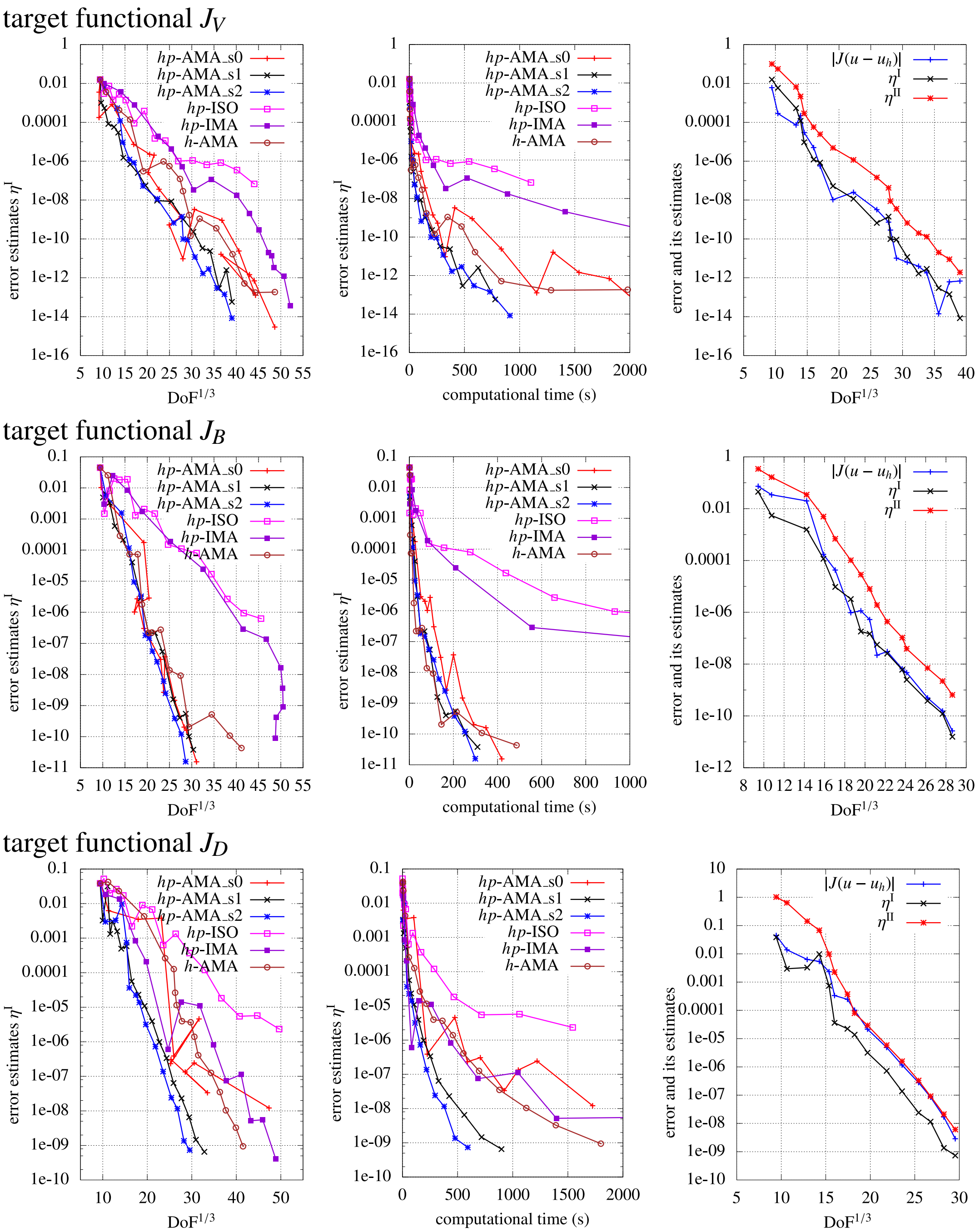

Figure 3.10: Convection-dominated problem, convergence of the error estimate $\eta^{\mathrm{I}} = \eta^{\mathrm{I}}(u_h^{(n)}, z_h^{(n)})$ w. r. t. DoF (left) and w. r. t. the computational time (center) for all tested adapted strategies. Comparison of the error estimates η^{I} and η^{II} with the actual error $J(u-u_h)$ for the hp-AMA_s2 method (right) with the target functionals J_V (top row), J_B (middle row) and J_D (bottom row).

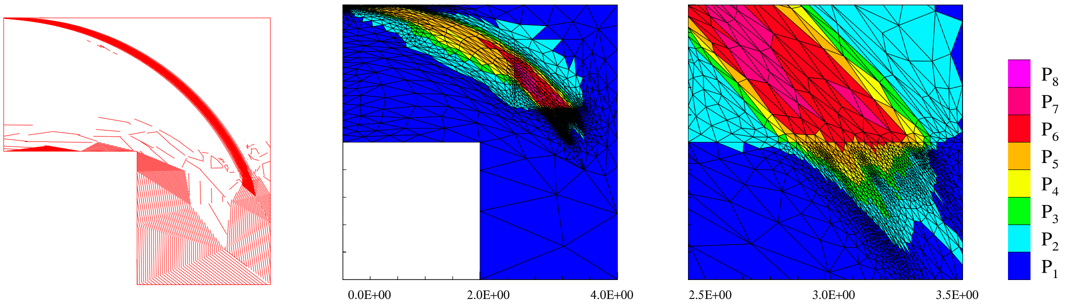

Figure 3.11: Convection-dominated problem, isolines of the primal solution (left), the final hp-mesh (center) and its detail (right) generated by Algorithm 3 with the target functional J_V.

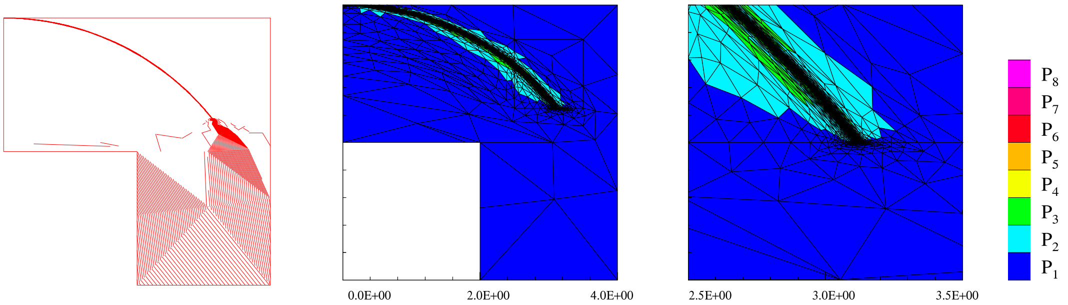

Figure 3.12: Convection-dominated problem with $\varepsilon = 10^{-6}$, isolines of the primal solution (left), the final hp-mesh (center) and its detail (right) generated by Algorithm 3 with the target functional J_V.

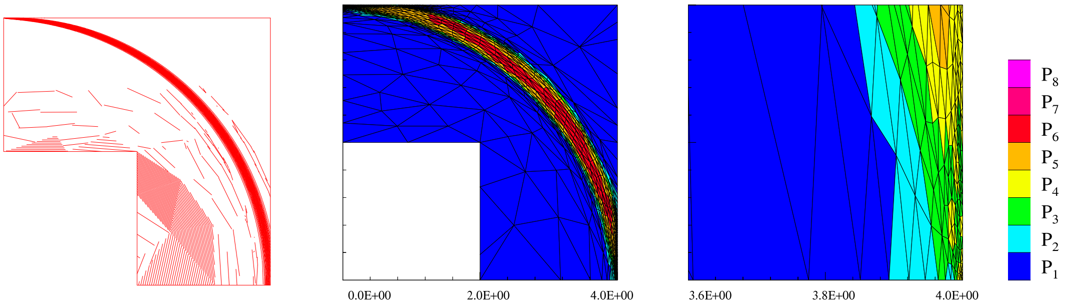

Figure 3.13: Convection-dominated problem, isolines of the primal solution (left), the final hp-mesh (center) and its detail (right) generated by Algorithm 3 with the target functional J_B.

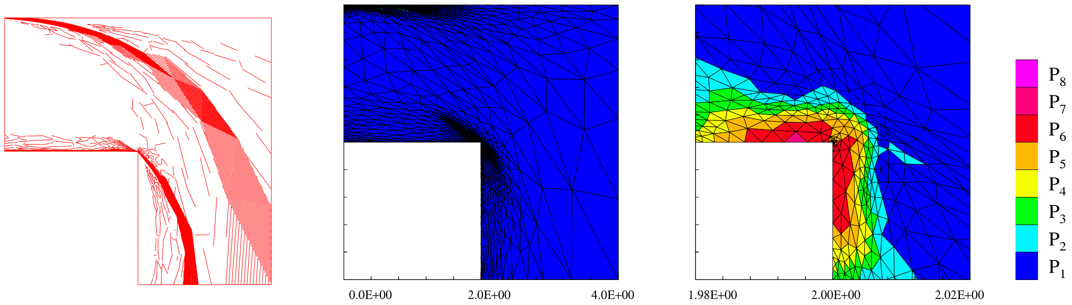

Figure 3.14: Convection-dominated problem, isolines of the primal solution (left), the final *hp*-mesh (center) and its detail (right) generated by Algorithm 3 with the target functional J_D.

4. Inviscid compressible flow

In this chapter we consider steady compressible inviscid adiabatic flow model known as the Euler equations. We focus the steady state solutions, even though generally the Euler equations form a time-dependent problem.

The flow is described by the continuity equations, the Euler equations of motion and the energy equation which we further extend by the thermodynamical relations.

The solution of the Euler equations may often contain discontinuities. For this reason the finite volume method (FV) using piece-wise constant approximation is widely used for solving of compressible flow problems. On the other hand, the conforming finite element method (FEM) is not suitable for numerical solution of such problems due to the assumption, which is usually made, that the exact solution is sufficiently regular. Although there are also conforming finite element techniques applicable to compressible flow, the treatment of discontinuities is rather complicated.

The discontinuous Galerkin method (DG) takes the advantages of both FEM (high order) and FV (discontinuity) which allows to obtain a stable scheme with high-order accuracy. We present the DG method applied to the Euler equations. First we formulate the problem and we describe its DG discretization with emphasis on the definition of the so-called numerical fluxes which describe the flow through the mesh element faces. A special attention is paid to the setting of the boundary conditions, since for hyperbolic problems their treatment is far from obvious and, as it will be shown in Section 4.3.8, it turns up to to be very important for obtaining the adjoint consistency of the numerical scheme.

Utilizing the goal-oriented error estimates for nonlinear problems as presented in Section 1.2 we introduce the goal-oriented error estimation method for Euler equations. We proceed similarly as in the linear case (Chapter 2) – after introducing the goal-oriented error relation, we present the computable estimates based on the reconstructions of the primal and adjoint discrete solutions, and further we derive error indicators enabling the anisotropic *hp*-adaptation of the mesh based on the approach presented in Chapter 3. In all steps we stress the modifications which need to be done compared to the linear problems.

At the end we present several numerical results illustrating the efficiency of the method.

4.1 Euler equations

We consider the steady state compressible inviscid Euler equations in a domain $\Omega \subset \mathbb{R}^d$. We restrict ourselves to the case $d = 2$ in this text, but the the majority of actions bellow can be easily generalized also to $d = 3$.

We use the standard notation: ρ - density, p - pressure (symbol p denotes the degree of polynomial approximation), E - total energy, $v_s,\ s = 1,\dots,d$ - components of the velocity vector $\boldsymbol{v} = (v_1,\dots,v_d)^{\mathrm{T}}$ in the directions x_s, θ - absolute temperature, $c_v > 0$ - specific heat at constant volume, $c_{\mathrm{p}} > 0$ - specific heat at constant pressure, $\gamma = c_{\mathrm{p}}/c_v > 1$ - Poisson adiabatic constant, $R = c_{\mathrm{p}} - c_v > 0$ - gas constant.

The system of governing equations formed by the continuity equation, the Euler equations of motion and the energy equation can be written in the form

$$\sum_{s=1}^{d} \frac{\partial(\rho v_s)}{\partial x_s} = 0, \tag{4.1}$$

$$\sum_{s=1}^{d} \frac{\partial(\rho v_i v_s + \delta_{is}\mathrm{p})}{\partial x_s} = 0, \quad i = 1,\dots,d, \tag{4.2}$$

$$\sum_{s=1}^{d} \frac{\partial((E+\mathrm{p})v_s)}{\partial x_s} = 0. \tag{4.3}$$

To the above system, we add the thermodynamical relations defining the pressure $\mathrm{p} = (\gamma-1)(E-\rho|\boldsymbol{v}|^2/2)$, and the total energy $E = \rho(c_v\theta + |\boldsymbol{v}|^2/2)$. Further, we define the *speed of sound a* and the *Mach number M* by $a = \sqrt{\gamma\mathrm{p}/\rho}$, $M = \frac{|\boldsymbol{v}|}{a}$.

System (4.1)–(4.3) has $m = d+2$ equations and it can be rewritten to

$$\nabla \cdot \boldsymbol{F}(\boldsymbol{w}) = \sum_{s=1}^{d} \frac{\partial \boldsymbol{f}_s(\boldsymbol{w})}{\partial x_s} = 0, \tag{4.4}$$

where $\boldsymbol{w} = (w_1,\dots,w_m)^{\mathrm{T}} = (\rho,\rho v_1,\dots,\rho v_d,E)^{\mathrm{T}} \in \mathbb{R}^m$, is the so-called *state vector* and $\boldsymbol{F}(\boldsymbol{w}) = (\boldsymbol{f}_1(\boldsymbol{w}),\dots,\boldsymbol{f}_d(\boldsymbol{w}))$,

$$\boldsymbol{f}_s(\boldsymbol{w}) = \begin{pmatrix} f_{s,1}(\boldsymbol{w}) \\ f_{s,2}(\boldsymbol{w}) \\ \vdots \\ f_{s,m-1}(\boldsymbol{w}) \\ f_{s,m}(\boldsymbol{w}) \end{pmatrix} = \begin{pmatrix} \rho v_s \\ \rho v_1 v_s + \delta_{1s}\mathrm{p} \\ \vdots \\ \rho v_d v_s + \delta_{ds}\mathrm{p} \\ (E+\mathrm{p})v_s \end{pmatrix} \tag{4.5}$$

$$= \begin{pmatrix} w_{s+1} \\ \frac{w_2 w_{s+1}}{w_1} + \delta_{1s}(\gamma-1)\left(w_m - \frac{1}{2w_1}\sum_{i=2}^{m-1} w_i^2\right) \\ \vdots \\ \frac{w_{m-1}w_{s+1}}{w_1} + \delta_{ds}(\gamma-1)\left(w_m - \frac{1}{2w_1}\sum_{i=2}^{m-1} w_i^2\right) \\ \frac{w_{s+1}}{w_1}\left(\gamma w_m - \frac{\gamma-1}{2w_1}\sum_{i=2}^{m-1} w_i^2\right) \end{pmatrix},$$

is the *flux* of the quantity $\boldsymbol{w}$ in the direction x_s, $s = 1,\dots,d$.

It can be easily shown that

$$\begin{aligned} v_i =& w_{i+1}/w_1,\ i = 1,\dots,d, \\ \mathrm{p} =& (\gamma-1)\left(w_m - \sum_{i=2}^{m-1} w_i^2/(2w_1)\right), \\ \theta =& \left(w_m/w_1 - \frac{1}{2}\sum_{i=2}^{m-1}(w_i/w_1)^2\right)/c_v. \end{aligned} \tag{4.6}$$

The domain of definition of the vector-valued functions $\boldsymbol{f}_s$, $s = 1,\dots,d$, is the open set $\mathscr{D} \subset \mathbb{R}^m$ of vectors $\boldsymbol{w} = (w_1,\dots,w_m)^{\mathrm{T}}$ such that their corresponding density and pressure are positive, i.e.,

$$\mathscr{D} = \left\{\boldsymbol{w} \in \mathbb{R}^m;\ w_1 = \rho > 0,\ w_m - \sum_{i=2}^{m-1} w_i^2/(2w_1) = \mathrm{p}/(\gamma-1) > 0\right\}. \tag{4.7}$$

Obviously, $\boldsymbol{f}_s \in (C^1(\mathscr{D}))^m$.

Using the chain rule in (4.4) leads to a first-order quasilinear system of partial differential equations

$$\sum_{s=1}^{d} \mathbb{A}_s(\boldsymbol{w}) \frac{\partial \boldsymbol{w}}{\partial x_s} = 0, \tag{4.8}$$

where $\mathbb{A}_s(\boldsymbol{w})$ is the $m \times m$ Jacobi matrix of the mapping $\boldsymbol{f}_s$ defined for $\boldsymbol{w} \in \mathscr{D}$:

$$\mathbb{A}_s(\boldsymbol{w}) := \frac{\mathrm{D}\boldsymbol{f}_s(\boldsymbol{w})}{\mathrm{D}\boldsymbol{w}} = \left(\frac{\partial f_{s,i}(\boldsymbol{w})}{\partial w_j}\right)_{i,j=1}^{m}, \ s = 1, \dots, d. \tag{4.9}$$

The Jacobi matrices $\mathbb{A}_s$, $s = 1, 2$, have the form

$$\mathbb{A}_1(\boldsymbol{w}) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2 - v_1^2 & (3-\gamma)v_1 & -\gamma_1 v_2 & \gamma_1 \\ -v_1 v_2 & v_2 & v_1 & 0 \\ v_1\left(\gamma_1|\boldsymbol{v}|^2 - \gamma\frac{E}{\rho}\right) & \gamma\frac{E}{\rho} - \gamma_1 v_1^2 - \frac{\gamma_1}{2}|\boldsymbol{v}|^2 & -\gamma_1 v_1 v_2 & \gamma v_1 \end{pmatrix}, \tag{4.10}$$

$$\mathbb{A}_2(\boldsymbol{w}) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ -v_1 v_2 & v_2 & v_1 & 0 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2 - v_2^2 & -\gamma_1 v_1 & (3-\gamma)v_2 & \gamma_1 \\ v_2\left(\gamma_1|\boldsymbol{v}|^2 - \gamma\frac{E}{\rho}\right) & -\gamma_1 v_1 v_2 & \gamma\frac{E}{\rho} - \gamma_1 v_2^2 - \frac{\gamma_1}{2}|\boldsymbol{v}|^2 & \gamma v_2 \end{pmatrix}, \tag{4.11}$$

where $\gamma_1 = \gamma - 1$.

Let $\mathrm{B}_1 = \{\boldsymbol{n} \in \mathbb{R}^d; |\boldsymbol{n}| = 1\}$ denote the unit sphere in $\mathbb{R}^d$. For $\boldsymbol{w} \in \mathscr{D}$ and $\boldsymbol{n} = (n_1, \dots, n_d)^{\mathrm{T}} \in \mathrm{B}_1$ we denote

$$\boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}) = \sum_{s=1}^{d} \boldsymbol{f}_s(\boldsymbol{w}) n_s = \begin{pmatrix} \rho \boldsymbol{v} \cdot \boldsymbol{n} \\ \rho v_1 \boldsymbol{v} \cdot \boldsymbol{n} + \mathrm{p} n_1 \\ \vdots \\ \rho v_d \boldsymbol{v} \cdot \boldsymbol{n} + \mathrm{p} n_d \\ (E + \mathrm{p}) \boldsymbol{v} \cdot \boldsymbol{n} \end{pmatrix} \tag{4.12}$$

which is the *physical flux* of the quantity $\boldsymbol{w}$ in the direction $\boldsymbol{n}$. Obviously, the Jacobi matrix $\mathrm{D}\boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n})/\mathrm{D}\boldsymbol{w}$ can be expressed in the form

$$\mathbb{P}(\boldsymbol{w}, \boldsymbol{n}) := \frac{\mathrm{D}\boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n})}{\mathrm{D}\boldsymbol{w}} = \sum_{s=1}^{d} n_s \frac{\mathrm{D}\boldsymbol{f}_s(\boldsymbol{w})}{\mathrm{D}\boldsymbol{w}} = \sum_{s=1}^{d} \mathbb{A}_s(\boldsymbol{w}) n_s. \tag{4.13}$$

The matrix $\mathbb{P}(\boldsymbol{w}, \boldsymbol{n})$ has the form

$$\mathbb{P}(\boldsymbol{w}, \boldsymbol{n}) = \left(\begin{array}{c|c|c|c} 0 & n_1 & n_2 & 0 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2 n_1 - v_1 \boldsymbol{v}\cdot\boldsymbol{n} & -\gamma_2 v_1 n_1 + \boldsymbol{v}\cdot\boldsymbol{n} & v_1 n_2 - \gamma_1 v_2 n_1 & \gamma_1 n_1 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2 n_2 - v_2 \boldsymbol{v}\cdot\boldsymbol{n} & v_2 n_1 - \gamma_1 v_1 n_2 & -\gamma_2 v_2 n_2 + \boldsymbol{v}\cdot\boldsymbol{n} & \gamma_1 n_2 \\ \left(\gamma_1|\boldsymbol{v}|^2 - \frac{\gamma E}{\rho}\right) \boldsymbol{v}\cdot\boldsymbol{n} & G n_1 - \gamma_1 v_1 \boldsymbol{v}\cdot\boldsymbol{n} & G n_2 - \gamma_1 v_2 \boldsymbol{v}\cdot\boldsymbol{n} & \gamma \boldsymbol{v}\cdot\boldsymbol{n} \end{array}\right), \tag{4.14}$$

where $\boldsymbol{n} = (n_1, n_2)$, $\gamma_1 = \gamma - 1$, $\gamma_2 = \gamma - 2$ and $G = \gamma\frac{E}{\rho} - \frac{\gamma_1}{2}|\boldsymbol{v}|^2$.

Lemma 4.1. *Let us summarize some important properties of the system of the Euler equations (4.4):*

1. *The vector-valued functions $\boldsymbol{f}_s$ defined by (4.5) are homogeneous mappings of order 1, i.e., $\boldsymbol{f}_s(\alpha \boldsymbol{w}) = \alpha \boldsymbol{f}_s(\boldsymbol{w})$, $\alpha > 0$. Moreover, we have $\boldsymbol{f}_s(\boldsymbol{w}) = \mathbb{A}_s(\boldsymbol{w})\boldsymbol{w}$.*
2. *Similarly, $\boldsymbol{P}(\alpha \boldsymbol{w}, \boldsymbol{n}) = \alpha \boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}), \quad \alpha > 0, \boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}) = \mathbb{P}(\boldsymbol{w}, \boldsymbol{n})\boldsymbol{w}$.*
3. *The system of the Euler equations is diagonally hyperbolic, i.e., the matrix $\mathbb{P}(\boldsymbol{w}, \boldsymbol{n})$ has only real eigenvalues $\lambda_i = \lambda_i(\boldsymbol{w}, \boldsymbol{n})$, $i = 1, \dots, m$, and is diagonalizable: there exists a nonsingular matrix $\mathbb{T} = \mathbb{T}(\boldsymbol{w}, \boldsymbol{n})$ such that*

$$\mathbb{T}^{-1}\mathbb{P}\mathbb{T} = \boldsymbol{\Lambda} = \boldsymbol{\Lambda}(\boldsymbol{w}, \boldsymbol{n}) = \mathrm{diag}(\lambda_1, \dots, \lambda_m) = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & 0 & \lambda_{m-1} & 0 \\ 0 & 0 & \dots & 0 & \lambda_m \end{pmatrix}. \tag{4.15}$$

The columns of the matrix $\mathbb{T}$ are the eigenvectors of the matrix $\mathbb{P}$ and the eigenvalues of the matrix $\mathbb{P}(\boldsymbol{w}, \boldsymbol{n})$, $\boldsymbol{w} \in \mathscr{D}$, $\boldsymbol{n} \in \mathrm{B}_1$ have the form

$$\begin{aligned} \lambda_1(\boldsymbol{w}, \boldsymbol{n}) &= \boldsymbol{v} \cdot \boldsymbol{n} - a, \qquad (4.16) \\ \lambda_2(\boldsymbol{w}, \boldsymbol{n}) &= \dots = \lambda_{d+1}(\boldsymbol{w}, \boldsymbol{n}) \ = \ \boldsymbol{v} \cdot \boldsymbol{n}, \\ \lambda_m(\boldsymbol{w}, \boldsymbol{n}) &= \boldsymbol{v} \cdot \boldsymbol{n} + a, \end{aligned}$$

where $a = \sqrt{\gamma \mathrm{p}/\rho}$ is the speed of sound and $\boldsymbol{v}$ is the velocity vector given by $\boldsymbol{v} = (w_2/w_1, w_3/w_1, \dots, w_{d+1}/w_1)^{\mathrm{T}}$.

4. *Further, we may define the "positive" and "negative" part of $\mathbb{P}$ by*

$$\mathbb{P}^{\pm} = \mathbb{T}\boldsymbol{\Lambda}^{\pm}\mathbb{T}^{-1}, \ \boldsymbol{\Lambda}^{\pm} = diag\ (\lambda_1^{\pm}, \dots, \lambda_m^{\pm}), \tag{4.17}$$

where $\lambda^+ = \max(\lambda, 0)$ and $\lambda^- = \min(\lambda, 0)$ for $\lambda \in \mathbb{R}$.

5. *The system of the Euler equations is rotationally invariant. Namely, for $\boldsymbol{n} = (n_1, \dots, n_d) \in \mathrm{B}_1, \boldsymbol{w} \in \mathscr{D}$ it holds*

$$\boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}) = \sum_{s=1}^{d} \boldsymbol{f}_s(\boldsymbol{w}) n_s = \mathbb{Q}^{-1}(\boldsymbol{n}) \boldsymbol{f}_1(\mathbb{Q}(\boldsymbol{n})\boldsymbol{w}), \tag{4.18}$$

$$\mathbb{P}(\boldsymbol{w}, \boldsymbol{n}) = \sum_{s=1}^{d} \mathbb{A}_s(\boldsymbol{w}) n_s = \mathbb{Q}^{-1}(\boldsymbol{n}) \mathbb{A}_1(\mathbb{Q}(\boldsymbol{n})\boldsymbol{w}) \mathbb{Q}(\boldsymbol{n}). \tag{4.19}$$

Here $\mathbb{Q}(\boldsymbol{n})$ is the $m \times m$ matrix corresponding to $\boldsymbol{n} \in \mathrm{B}_1$ given by

$$\mathbb{Q}(\boldsymbol{n}) = \begin{pmatrix} 1 & \mathbf{0} & 0 \\ \mathbf{0}^{\mathrm{T}} & \mathbb{Q}_0(\boldsymbol{n}) & \mathbf{0}^{\mathrm{T}} \\ 0 & \mathbf{0} & 1 \end{pmatrix}, \tag{4.20}$$

where by $\mathbf{0}$ we denote the vector $(0, 0)$ and the $d \times d$ rotation matrix $\mathbb{Q}_0(\boldsymbol{n})$ is defined by

$$\mathbb{Q}_0(\boldsymbol{n}) = \begin{pmatrix} n_1 & n_2 \\ -n_2 & n_1 \end{pmatrix}. \tag{4.21}$$

Proof. See [Feistauer et al., 2003, Lemma 3.1, Lemma 3.3, Theorem 3.4]. □

4.1.1 Weak formulation of the primal problem

In order to proceed to goal-oriented error estimates, we have to introduce the weak solution of the Euler equations (4.4). However, the weak formulation is delicate since the solution of (4.4) may not be regular, e.g., the presence of the shock waves and contact discontinuities is usual, see Feistauer [1993], Feistauer et al. [2003], LeVeque [1990, 2002]. Further, the solution of (4.4) may be non-unique and thus additional conditions (e.g., entropy condition) have to be added. The typical structure of solutions containing discontinuities is treated using the so-called Rangine-Hugoniot conditions, see [Feistauer et al., 2003, Theorem 2.15]. In order to avoid the difficulties, we introduce the weak formulation only in a formal way and do not deal with the existence of the weak solution. Let $\boldsymbol{V}$ be a suitable space of functions from Ω to $\mathbb{R}^m$, where the weak solution is sought. This space should reflect the physical features of the problem, e.g., the positive density and pressure.

Definition 4.2. *We say that a function $\boldsymbol{w} \in \boldsymbol{V}$ is the weak solution of the steady compressible inviscid Euler equations, if it satisfies*

$$\mathbf{F}(\boldsymbol{w},\boldsymbol{\varphi}) = 0 \qquad \forall \boldsymbol{\varphi} \in \boldsymbol{V}, \tag{4.22}$$

where $\mathbf{F}(\boldsymbol{w},\boldsymbol{\varphi})$ is the weak form corresponding to $\nabla \cdot \boldsymbol{F}(\boldsymbol{w})$, equipped with boundary conditions, cf. (1.62)*, formally given by the boundary operator*

$$\mathscr{B}(\boldsymbol{w}) = 0 \quad \text{on } \partial\Omega. \tag{4.23}$$

We assume that the weak solution exists and that the formulation is consistent, i.e, if $\boldsymbol{w}$ satisfies (4.4) then it fulfills also (4.22). For more details on the solvability of Euler equations and the properties of the solutions see, e.g., LeVeque [1990], Feistauer [1993], Feistauer et al. [2003].

Following the approach presented in Section 1.1.1 for nonconforming discretizations we need to define a space $\boldsymbol{W_h}$ such that the exact solution is contained in it and also $\boldsymbol{S_h^p} \subset \boldsymbol{W_h}$. For the theoretical analysis of the DG method we add an artificial assumption that if the exact solution possesses some discontinuities, these align with the faces of the mesh elements. Then it makes sense to define $\boldsymbol{W_h} := \boldsymbol{H}^1(\Omega,\mathscr{T}_h)$, where denotes the broken Sobolev space $\boldsymbol{H}^1(\Omega,\mathscr{T}_h) := [H^1(\Omega,\mathscr{T}_h)]^m$.

4.1.2 Boundary conditions

A suitable way of employing the boundary conditions is quite delicate problem in numerical simulations of hyperbolic problems, since putting simply $\boldsymbol{w} = \boldsymbol{w}_{\mathrm{BC}}$ for some given data $\boldsymbol{w}_{\mathrm{BC}}$ would make the system overdetermined.

On one hand, setting of the boundary condition for the fluid flows is mainly a physical problem. But on the other hand, it has to correspond to the mathematical formulation of the problem in order to provide a well-posed problem. Basically, we distinguish two part of the boundary $\partial\Omega$ – the *impereable walls* Γ_W and *inlet/outlet* Γ_{IO}.

On $\Gamma_W \subset \partial\Omega$ we prescribe the so-called reflective boundary conditions which simulate fixed walls. This impermeability condition reads $\boldsymbol{v} \cdot \boldsymbol{n} = 0$, where $\boldsymbol{v}$ is the velocity vector and $\boldsymbol{n}$ denotes the outer unit normal to $\partial\Omega$. This gives for $\boldsymbol{w} \in \mathbb{R}^{d+2}$

$$\mathscr{B}(\boldsymbol{w}) = \sum_{i=1}^{d} n_i w_{i+1}. \tag{4.24}$$

On Γ_{IO} we set the transmissive boundaries which enable replacing unbounded domains by some sufficiently large but bounded computational domains if needed. In this case the boundary conditions are designed such that they allow passage of waves through them without any effect.

For the problem (4.4) to be well-posed, the number of prescribed boundary conditions at each point of the boundary has to be equal to the number of the negative eigenvalues of the matrix $\mathbb{P}(\boldsymbol{w}, \boldsymbol{n}_\Gamma)$ defined by (4.13). Here $\boldsymbol{n}_\Gamma$ denotes the unit outer normal on $\partial\Omega$. One way is to distinguish between the four possible cases (supersonic inlet and outlet and subsonic inlet and outlet) and define $\mathscr{B}$ separately for each of them, see, e.g., Hartmann and Leicht [2015], or we may set generally

$$\mathscr{B}(\boldsymbol{w}) = \mathbb{P}^{(-)}(\boldsymbol{w}, \boldsymbol{n})(\boldsymbol{w} - \boldsymbol{w}_{\mathrm{BC}}). \tag{4.25}$$

We note that this approach is related to the treatment of boundary conditions for scalar linear advections problems ($\boldsymbol{b} \cdot \nabla u = f$), where the boundary condition is prescribed only on the part of the boundary where $\boldsymbol{n} \cdot \boldsymbol{n} < 0$.

The numerical treatment of the boundary conditions (4.23) for the DG discretization will be studied in Section 4.3.2.

4.2 Quantity of interest and continuous adjoint problem

The most interesting target quantities in inviscid compressible flows are the aerodynamic coefficients, namely the drag (c_D), lift (c_L) and momentum (c_M). In this section we introduce the target functional J representing either of these coefficients in a unified way.

4.2.1 Quantity of interest

We can define the target functional representing the quantity of interest be the following integral

$$J(\boldsymbol{w}) = \int_{\Gamma_W} j(\boldsymbol{w})\, \mathrm{d}S = \int_{\Gamma_W} \mathrm{p}\boldsymbol{n} \cdot \vartheta\, \mathrm{d}S = \int_{\Gamma_W} \boldsymbol{p}_n \cdot \tilde{\vartheta}\, \mathrm{d}S \tag{4.26}$$

where $\boldsymbol{p}_n = \mathrm{p}(0, n_1, \ldots, n_d, 0)^{\mathrm{T}}$ and $\tilde{\vartheta} = (0, \vartheta_1, \ldots, \vartheta_d, 0)^{\mathrm{T}}$. Hence $j(\boldsymbol{w}) = \boldsymbol{p}_n \cdot \tilde{\vartheta}$ on Γ_W. Here, $\Gamma_W \subset \partial\Omega$ represents a solid profile, $\boldsymbol{n}$ is outer unit normal to the profile pointing into the profile and p is the pressure geven by (4.6).

Further (for $d = 2$) ϑ is given for the drag and lift coefficients by

$$\vartheta_{\mathrm{d}} = \frac{1}{C_\infty}(\cos(\alpha), \sin(\alpha))^{\mathrm{T}} \quad \text{and} \quad \vartheta_{\mathrm{l}} = \frac{1}{C_\infty}(-\sin(\alpha), \cos(\alpha))^{\mathrm{T}}, \tag{4.27}$$

respectively, where α denotes the angle of attack of the flow, $C_\infty = \frac{1}{2}\rho_\infty |\boldsymbol{v}_\infty|^2 L_{\mathrm{ref}}$, ρ_∞ and $\boldsymbol{v}_\infty$ are the far-field density and velocity, respectively and L_{ref} is the reference length.

The coefficient of momentum is usually defined+ as

$$\frac{1}{\frac{1}{2}\rho_\infty |\boldsymbol{v}_\infty|^2 L_{\mathrm{ref}}^2} \int_{\Gamma_W} \mathrm{p}(x - x_{\mathrm{ref}}) \times (\mathbb{Q}(\alpha)\boldsymbol{n})\, \mathrm{d}S, \tag{4.28}$$

where $x_{\text{ref}} \in \Omega$ is the moment reference point and

$$\mathbb{Q}(\alpha) = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix}$$

is the rotation matrix of the angle of attack α. For $x, y \in \mathbb{R}^2$ the notation $\times$ stands for $x \times y = x_1 y_2 - x_2 y_1$.

In order to obtain the standartized shape of target functional, we can rewrite $(x - x_{\text{ref}}) \times (\mathbb{Q}(\alpha)\boldsymbol{n}) = (x - x_{\text{ref}})\mathbb{G}(\mathbb{Q}(\alpha)\boldsymbol{n})$, where $\mathbb{G} = \big((0,1)^{\mathrm{T}}, (-1,0)^{\mathrm{T}}\big)$. Hence, in (4.26) we can choose

$$\vartheta_{\mathrm{m}} := \frac{1}{C_\infty L_{\text{ref}}}((x - x_{\text{ref}})\mathbb{G}\mathbb{Q}(\alpha))^{\mathrm{T}}. \tag{4.29}$$

For further use in the linearized adjoint problem we also need the Fréchet directional derivative $J'[\boldsymbol{w}](\boldsymbol{\varphi}) = \int_{\Gamma_W} j'(\boldsymbol{w})\boldsymbol{\varphi}\, \mathrm{d}S$, c.f. (1.65). Recalling the definition of the pressure (4.6) we may compute its derivative

$$\frac{D\mathrm{p}}{D\boldsymbol{w}} = (\gamma - 1)\begin{pmatrix} \frac{w_2^2 + w_3^2}{2w_1^2} \\ -\frac{w_2}{w_1} \\ -\frac{w_3}{w_1} \\ 1 \end{pmatrix} = (\gamma - 1)\begin{pmatrix} \frac{1}{2}|\boldsymbol{v}|^2 \\ -v_1 \\ -v_2 \\ 1 \end{pmatrix} \tag{4.30}$$

and hence the Jacobi matrix of $\boldsymbol{p}_n$ equals

$$\mathbb{P}_W(\boldsymbol{w}, \boldsymbol{n}) := \frac{D\boldsymbol{p}_n}{D\boldsymbol{w}} = (\gamma - 1)\begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ |\boldsymbol{v}|^2 n_1/2 & -v_1 n_1 & \dots & -v_d n_1 & n_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ |\boldsymbol{v}|^2 n_d/2 & -v_1 n_d & \dots & -v_d n_d & n_d \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}, \tag{4.31}$$

where $\boldsymbol{w} \in \mathscr{D}$, $\boldsymbol{n} = (n_1, \dots, n_d) \in \mathrm{B}_1$, $v_j = w_{j+1}/w_1$, $j = 1, \dots, d$, are the components of the velocity vector and $|\boldsymbol{v}|^2 = v_1^2 + \cdots + v_d^2$.

Altogether we get $j'[\boldsymbol{w}] = \mathbb{P}_W(\boldsymbol{w}, \boldsymbol{n})^{\mathrm{T}}\tilde{\vartheta}$ and

$$J'[\boldsymbol{w}](\boldsymbol{\varphi}) = \int_{\partial\Omega} \tilde{\vartheta}^{\mathrm{T}}\mathbb{P}_W(\boldsymbol{w}, \boldsymbol{n})\boldsymbol{\varphi}\, \mathrm{d}S. \tag{4.32}$$

Let us present several relation between the introduced terms.

Lemma 4.3. *Let $\boldsymbol{w}, \boldsymbol{\varphi} \in \mathscr{D}$ then*

- $\mathbb{P}_W^{\mathrm{T}}(\boldsymbol{w}, \boldsymbol{n})\boldsymbol{\varphi} = \frac{D\mathrm{p}(\boldsymbol{w})}{D\boldsymbol{w}}\left((0, n_1, n_2, 0) \cdot \boldsymbol{\varphi}\right),$
- $\boldsymbol{p}_n = \mathbb{P}_W(\boldsymbol{w}, \boldsymbol{n})\boldsymbol{w}$, *where* $\boldsymbol{p}_n = \boldsymbol{p}_n(\boldsymbol{w})$.
- $J(\boldsymbol{w}) = J'[\boldsymbol{w}](\boldsymbol{w})$.

Proof. The first statement follows directly from the definition of matrix multiplication if we realize that $\mathbb{P}_W^{\mathrm{T}}(\boldsymbol{w}, \boldsymbol{n}) = \frac{D\mathrm{p}(\boldsymbol{w})}{D\boldsymbol{w}} \otimes (0, n_1, \dots, n_d, 0)$, where $\otimes$ denotes the vector outer product.

For the second one, we can write

$$\mathbb{P}_W(\boldsymbol{w},\boldsymbol{n})\boldsymbol{w} = (\gamma-1)r(\boldsymbol{w})(0,n_1,\dots,n_d,0)^{\mathrm{T}},$$

where

$$\begin{aligned} r(\boldsymbol{w}) &= \frac{|\boldsymbol{v}|^2}{2}\boldsymbol{w}_1 - \boldsymbol{v}_1\boldsymbol{w}_2 - \boldsymbol{v}_2\boldsymbol{w}_3 + \boldsymbol{w}_4 \\ &= \frac{|\boldsymbol{v}|^2}{2}\rho - \boldsymbol{v}_1^2\rho - \boldsymbol{v}_2^2\rho + E = E - \frac{|\boldsymbol{v}|^2}{2}\rho = \frac{\mathrm{p}(\boldsymbol{w})}{(\gamma-1)}. \end{aligned} \tag{4.33}$$

The last statement follows from the definition (4.26) of the target functional J, its derivative (4.32) and the previously proven identity, i.e., $\boldsymbol{p}_n(\boldsymbol{w}) = \mathbb{P}_W(\boldsymbol{w},\boldsymbol{n})\boldsymbol{w}$. □

Lemma 4.4. *Let $\tilde{\boldsymbol{w}},\boldsymbol{w} \in \mathscr{D}$ satisfy the boundary condition on Γ_W (4.24), then it holds*

$$\mathbb{P}(\tilde{\boldsymbol{w}},\boldsymbol{n})\boldsymbol{w} = \mathbb{P}_W(\tilde{\boldsymbol{w}},\boldsymbol{n})\boldsymbol{w} \text{ on } \Gamma_W. \tag{4.34}$$

Proof. Since it holds $\boldsymbol{v}\cdot\boldsymbol{n} = 0$ for $\tilde{\boldsymbol{w}}$, some of the entries of the matrix $\mathbb{P}(\tilde{\boldsymbol{w}},\boldsymbol{n})$ disappear and we may write $\mathbb{P}(\tilde{\boldsymbol{w}},\boldsymbol{n}) = \mathbb{P}_W(\tilde{\boldsymbol{w}},\boldsymbol{n}) + \mathbb{R}_W(\tilde{\boldsymbol{w}},\boldsymbol{n})$, where

$$\mathbb{R}_W(\boldsymbol{w},\boldsymbol{n}) := \begin{pmatrix} 0 & n_1 & n_2 & 0 \\ 0 & v_1n_1 & v_1n_2 & 0 \\ 0 & v_2n_1 & v_2n_2 & 0 \\ 0 & Gn_1 & Gn_2 & 0 \end{pmatrix}, \tag{4.35}$$

with $G = \frac{\gamma E}{\rho} - \frac{\gamma-1}{2}|v|^2 = \frac{1}{\rho}(p+E)$. Then we can see that $\mathbb{R}_W(\tilde{\boldsymbol{w}},\boldsymbol{n})\boldsymbol{w} = 0$ due to the property $\boldsymbol{v}\cdot\boldsymbol{n} = 0$ for $\boldsymbol{w}$.

Another way of obtaining (4.34) is to rewrite $\mathbb{P}(\tilde{\boldsymbol{w}},\boldsymbol{n}) = (\gamma-1)(0,n_1,n_2,0)^{\mathrm{T}} \otimes (\frac{D\mathrm{p}}{D\boldsymbol{w}})^{\mathrm{T}} + (1,v_1,v_2,G)^{\mathrm{T}} \otimes (0,n_1,n_2,0)$ and utilize the rules for the $\otimes-$multiplication. □

4.2.2 Formulation of the continuous adjoint problem

For the derivation of the continuous adjoint form to (4.4), we follow the approach presented in Section 1.1.9, which follows the theoretical derivation of the adjoint problem formulation from Giles and Pierce [1997] and Hartmann [2007].

We multiply the equation (4.4) by $\boldsymbol{z} \in \boldsymbol{W_h}$, integrate by parts and finally linearize it around $\boldsymbol{w}$ which leads to

$$\begin{aligned} \int_\Omega \sum_{s=1}^{d} \frac{\partial \boldsymbol{f}_s'[\boldsymbol{w}](\boldsymbol{\varphi})}{\partial x_s}\cdot\boldsymbol{z}\,\mathrm{d}x = &-\int_\Omega \sum_{s=1}^{d} \boldsymbol{f}_s'[\boldsymbol{w}](\boldsymbol{\varphi})\cdot\frac{\partial \boldsymbol{z}}{\partial x_s}\,\mathrm{d}x \\ &+\int_{\partial\Omega}\sum_{s=1}^{d} n_s\boldsymbol{f}_s'[\boldsymbol{w}](\boldsymbol{\varphi})\cdot\boldsymbol{z}\,\mathrm{d}S \qquad \forall\boldsymbol{\varphi}\in\tilde{\boldsymbol{V}}. \end{aligned} \tag{4.36}$$

Here $\boldsymbol{f}_s'[\boldsymbol{w}]$ denotes the Fréchet derivative (see Definition 1.4) of $\boldsymbol{f}_s$ and the direction $\boldsymbol{\varphi}$ belongs to the function space of the permissible variations, see Lu [2005], defined by

$$\tilde{\boldsymbol{V}} := \{\boldsymbol{\varphi}\in\boldsymbol{V} : \boldsymbol{v}\cdot\boldsymbol{n} = 0 \text{ on } \Gamma_W, \mathbb{P}^{(-)}(\boldsymbol{w},\boldsymbol{n})\boldsymbol{\varphi} = 0 \text{ on } \Gamma_{IO}\}. \tag{4.37}$$

Remark. Our definition follows a more general approach from Lu [2005], where

$$\tilde{\boldsymbol{V}} := \{\boldsymbol{\varphi} \in V, \mathscr{B}'[\boldsymbol{w}](\boldsymbol{\varphi}) = 0\}$$

with a small modification that we only approximate $\mathscr{B}'[\boldsymbol{w}](\boldsymbol{\varphi})$ on Γ_{IO} by its linearization $\mathbb{P}^{(-)}(\boldsymbol{w},\boldsymbol{n})\boldsymbol{\varphi}$.

We note that the realization of the Fréchet derivative may be computed by the Jacobi matrix as $\boldsymbol{f}_s'[\boldsymbol{w}](\boldsymbol{\varphi}) = \frac{\mathrm{D}\boldsymbol{f}_s(\boldsymbol{w})}{\mathrm{D}\boldsymbol{w}}\boldsymbol{\varphi} = \mathbb{A}_s(\boldsymbol{w})\boldsymbol{\varphi}$.

Due to (4.9) and (4.13) we have

$$\int_{\partial\Omega} \sum_{s=1}^{d} n_s \boldsymbol{f}_s'[\boldsymbol{w}](\boldsymbol{\varphi}) \cdot \boldsymbol{z} \, \mathrm{d}S = \int_{\partial\Omega} \boldsymbol{z}^{\mathrm{T}} \mathbb{P}(\boldsymbol{w},\boldsymbol{n})\boldsymbol{\varphi} \, \mathrm{d}S \tag{4.38}$$

and hence we may define the variational formulation of the continuous adjoint problem: Find $\boldsymbol{z} \in \boldsymbol{W_h}$ such that

$$-\int_{\Omega} \sum_{s=1}^{d} \frac{\partial \boldsymbol{z}^{\mathrm{T}}}{\partial x_s} \mathbb{A}_s(\boldsymbol{w})\boldsymbol{\varphi} \, \mathrm{d}x + \int_{\partial\Omega} \boldsymbol{z}^{\mathrm{T}} \mathbb{P}(\boldsymbol{w},\boldsymbol{n})\boldsymbol{\varphi} \, \mathrm{d}S = J'[\boldsymbol{w}](\boldsymbol{\varphi}) \qquad \forall \boldsymbol{\varphi} \in \tilde{\boldsymbol{V}}. \tag{4.39}$$

This expression may be further rearranged individually on Γ_W and Γ_{IO}. On Γ_{IO} we exploit that $\boldsymbol{\varphi}$ belongs to $\tilde{\boldsymbol{V}}$ and hence $\mathbb{P}^{(-)}(\boldsymbol{w},\boldsymbol{n})\boldsymbol{\varphi} = 0$. Then due to (4.13) it holds

$$\int_{\Gamma_{IO}} \boldsymbol{z}^{\mathrm{T}} \mathbb{P}(\boldsymbol{w},\boldsymbol{n})\boldsymbol{\varphi} \, \mathrm{d}S = \int_{\Gamma_{IO}} \boldsymbol{z}^{\mathrm{T}} \mathbb{P}^{(+)}(\boldsymbol{w},\boldsymbol{n})\boldsymbol{\varphi} \, \mathrm{d}S. \tag{4.40}$$

On Γ_W we have $n_1\varphi_2 + n_2\varphi_3 = 0$. Functions $\boldsymbol{w}$ and $\boldsymbol{\varphi}$ satisfy the assumption of Lemma 4.4, hence we may use (4.34) which leads to

$$\int_{\Gamma_W} \boldsymbol{z}^{\mathrm{T}} \mathbb{P}(\boldsymbol{w},\boldsymbol{n})\boldsymbol{\varphi} \, \mathrm{d}S = \int_{\Gamma_W} \boldsymbol{z}^{\mathrm{T}} \mathbb{P}_W(\boldsymbol{w},\boldsymbol{n})\boldsymbol{\varphi} \, \mathrm{d}S. \tag{4.41}$$

Recalling (4.32) and noting that

$$\mathbb{P}_W^{\mathrm{T}}(\boldsymbol{w},\boldsymbol{n})\boldsymbol{z} = \frac{D\mathrm{p}(\boldsymbol{w})}{D\boldsymbol{w}}(0,n_1,n_2,0)\cdot\boldsymbol{z}, \quad \mathbb{P}_W^{\mathrm{T}}(\boldsymbol{w},\boldsymbol{n})\tilde{\boldsymbol{\vartheta}} = \frac{D\mathrm{p}(\boldsymbol{w})}{D\boldsymbol{w}}\boldsymbol{n}\cdot\boldsymbol{\vartheta}$$

we obtain a simplified boundary condition

$$n_1 z_2 + n_2 z_3 = \boldsymbol{n}\cdot\boldsymbol{\vartheta} \qquad \text{on } \Gamma_W. \tag{4.42}$$

Therefore, we can write the equation (4.39) in the strong form.

Definition 4.5. *We say that function $\boldsymbol{z}$ is the solution of the adjoint problem to* (4.4) *if it satisfies*

$$-\sum_{s=1}^{d} (\mathbb{A}_s(\boldsymbol{w}))^{\mathrm{T}} \frac{\partial \boldsymbol{z}}{\partial x_s} = 0 \textit{ in } \Omega, \tag{4.43}$$

with boundary conditions

$$\left(\mathbb{P}^{(+)}(\boldsymbol{w},\boldsymbol{n})\right)^{\mathrm{T}} \boldsymbol{z} = 0 \textit{ on } \partial\Omega\backslash\Gamma_W, \qquad n_1 z_2 + n_2 z_3 = \boldsymbol{n}\cdot\boldsymbol{\vartheta} \textit{ on } \Gamma_W. \tag{4.44}$$

4.3 Discontinuous Galerkin discretization

The discontinuous Galerkin (DG) approximate solution of (4.4) is sought in a finite-dimensional subspace of $\boldsymbol{H}^1(\Omega, \mathscr{T}_h)$ which consists of piecewise polynomial functions. We note that the polynomial degrees may differ among mesh elements and we recall the notation used in previous chapters. We denote the local polynomial degree $p_K \in \mathbb{N}$ for each $K \in \mathscr{T}_h$ and we introduce $\boldsymbol{p} := \{p_K; K \in \mathscr{T}_h\}$.

Further, over the triangulation $\mathscr{T}_h$ we define the space of vector-valued discontinuous piecewise polynomial functions

$$\boldsymbol{S}_h^{\boldsymbol{p}} = (S_h^{\boldsymbol{p}})^m, \text{ where } S_h^{\boldsymbol{p}} = \{v \in L^2(\Omega); v|_K \in P^{p_K}(K) \ \forall K \in \mathscr{T}_h\} \tag{4.45}$$

and for further use also

$$\boldsymbol{S}_h^{\boldsymbol{p}+1} = (S_h^{\boldsymbol{p}+1})^m, \text{ where } S_h^{\boldsymbol{p}+1} = \{v \in L^2(\Omega); v|_K \in P^{p_K+1}(K) \ \forall K \in \mathscr{T}_h\}. \tag{4.46}$$

Multiplying (4.4) by $\varphi \in \boldsymbol{H}^1(\Omega, \mathscr{T}_h)$, integrating over Ω and applying the Green theorem separately on each element $K \in \mathscr{T}_h$ we get

$$-\sum_{K\in\mathscr{T}_h} \int_K \sum_{s=1}^d \mathbb{A}(\boldsymbol{w})\boldsymbol{w} \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_s} \,\mathrm{d}x + \sum_{K\in\mathscr{T}_h} \int_{\partial K} \boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}_K) \cdot \boldsymbol{\varphi} \,\mathrm{d}S = 0, \tag{4.47}$$

where we used the statement of Lemma 4.1 which gives $\boldsymbol{f}_s(\boldsymbol{w}) = \mathbb{A}(\boldsymbol{w})\boldsymbol{w}$.

The crucial point of the DG approximation of conservation laws is the evaluation of the boundary terms since the numerical solution may be discontinous on the element interfaces. These integrals are approximated using the *numerical flux* $\mathbf{H} : \mathscr{D} \times \mathscr{D} \times \mathrm{B}_1 \to \mathbb{R}^m$

$$\int_{\partial K} \boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}) \cdot \boldsymbol{\varphi} \,\mathrm{d}S \approx \int_{\partial K} \mathbf{H}(\boldsymbol{w}^{(+)}, \boldsymbol{w}^{(-)}, \boldsymbol{n}_K) \cdot \boldsymbol{\varphi} \,\mathrm{d}S, \tag{4.48}$$

Here the notation $\boldsymbol{w}^{(+)}$ and $\boldsymbol{w}^{(-)}$ denotes the interior and exterior traces of $\boldsymbol{w}$, respectively. More precisely if $\boldsymbol{n}_K$ denotes the unit outer normal for an edge $\Gamma \in \partial K \backslash \partial\Omega$ then $\boldsymbol{w}^{(+)}$ denotes the trace of $\boldsymbol{w}$ in the direction opposite to the direction of $\boldsymbol{n}_K$ (from K) and $\boldsymbol{w}^{(-)}$ denotes the trace in the direction of $\boldsymbol{n}_K$ (from the neigbouring element). The meaning of $\boldsymbol{w}^{(-)}$ on the boundary $\partial\Omega$ will be defined later (Section 4.3.2). Here we only briefly note that its determination is based on the value $\boldsymbol{w}^{(+)}$ and the corresponding boundary condition. Further, we define the mean value of a function $\boldsymbol{\varphi} \in \boldsymbol{H}^1(\Omega, \mathscr{T}_h)$ by $\langle \boldsymbol{\varphi} \rangle = \frac{\boldsymbol{\varphi}^{(+)} + \boldsymbol{\varphi}^{(-)}}{2}$ on interior edges.

The numerical flux has to satisfy some basic conditions:

- *continuity*: $\mathbf{H}(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{n})$ is locally Lipschitz-continuous with respect to the variables $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$,
- *consistency*: $\mathbf{H}(\boldsymbol{w}, \boldsymbol{w}, \boldsymbol{n}) = \boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}), \ \boldsymbol{w} \in \mathscr{D}, \ \boldsymbol{n} = (n_1, \dots, n_d) \in \mathrm{B}_1$,
- *conservativity*: $\mathbf{H}(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{n}) = -\mathbf{H}(\boldsymbol{w}_2, \boldsymbol{w}_1, -\boldsymbol{n}), \quad \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathscr{D}, \ \boldsymbol{n} \in \mathrm{B}_1$.

Definition 4.6. *We say that a function* $\boldsymbol{w}_h \in \boldsymbol{S}_h^{\boldsymbol{p}}$ *is the* discrete DG solution *of the Euler equations* (4.4)*, if*

$$\boldsymbol{a}_h(\boldsymbol{w}_h, \boldsymbol{\varphi}_h) = 0 \qquad \forall \boldsymbol{\varphi}_h \in \boldsymbol{S}_h^{\boldsymbol{p}}, \tag{4.49}$$

where

$$\boldsymbol{a}_h(\boldsymbol{w}_h,\boldsymbol{\varphi}) = -\sum_{K\in\mathscr{T}_h}\int_K \sum_{s=1}^{d}(\mathbb{A}_s(\boldsymbol{w}_h)\boldsymbol{w}_h)\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,\mathrm{d}x + \sum_{K\in\mathscr{T}_h}\int_{\partial K}\mathbf{H}(\boldsymbol{w}_h^{(+)},\boldsymbol{w}_h^{(-)},\boldsymbol{n}_K)\cdot\boldsymbol{\varphi}\,\mathrm{d}S. \tag{4.50}$$

We continue with the specification of the interior and boundary numerical fluxes. The general definition of the numerical flux $\mathbf{H}$ may differ on the inner edges and on the edges laying on the boundary $\partial\Omega = \Gamma_W \cup \Gamma_{IO}$. Further, the boundary numerical fluxes differ for various boundary conditions given on separate parts of the boundary $\partial\Omega$. Each of these cases will be discussed separately, since each one requires a slightly different handling. In the further text we will use the shorter notation $\mathbf{H} = \mathbf{H}(\boldsymbol{w}^{(+)},\boldsymbol{w}^{(-)},\boldsymbol{n}_K)$ when the arguments are evident from context and on the contrary $\mathbf{H}_{\partial\Omega} = \mathbf{H}_{\partial\Omega}(\boldsymbol{w}^{(+)},\boldsymbol{w}^{(-)},\boldsymbol{n}_K)$ when we want emphasize the boundary.

4.3.1 Numerical fluxes on the inner edges

From the possible choices of the numerical fluxes we focus the *Vijayasundaram flux*, see Vijayasundaram [1986]. Other numerical fluxes usable for the Euler equations (such as the Lax-Friedrichs or Van Leer numerical fluxes) can be found in Feistauer et al. [2003]. The definition of the Vijayasundaram numerical flux is based on the spectral decomposition of the matrix $\mathbb{P}$ into $\mathbb{P}^{\pm}$ which was introduced in (4.17).

For any function $\boldsymbol{w} \in \boldsymbol{H}^1(\Omega,\mathscr{T}_h)$ *Vijayasundaram numerical flux* is given by

$$\mathbf{H}_{VS}(\boldsymbol{w}^{(+)},\boldsymbol{w}^{(-)},\boldsymbol{n}) = \mathbb{P}^{+}\left(\langle\boldsymbol{w}\rangle,\boldsymbol{n}\right)\boldsymbol{w}^{(+)} + \mathbb{P}^{-}\left(\langle\boldsymbol{w}\rangle,\boldsymbol{n}\right)\boldsymbol{w}^{(-)}. \tag{4.51}$$

Lemma 4.7. *Vijayasundaram numerical flux* $\mathbf{H}_{VS} = \mathbf{H}(\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{n})$ *given by* (4.51) *is locally Lipschitz-continuous, consistent and conservative.*

Consistency and conservativity follow directly from the definition of the Vijayasundaram flux, a detailed proof of the Lipschitz-continuity can be found, e.g., in Feistauer et al. [2003].

4.3.2 Numerical treatment of the boundary conditions

In the case of first order differential equations, such as the system (4.4), the treatment of boundary conditions is rather complicated, since it is not explicitly clear how much information has to set on the boundary in order to obtain a uniquely solvable problem (not under-determined or over-determined).

The numerical flux $\mathbf{H}$ has to be determined on the boundary edges, but the meaning of $\boldsymbol{w}^{(-)}$ is not so clear there. Since the DG discretization treats the boundary conditions only in a weak sense, the discrete solution $\boldsymbol{w}_h$ generally does not satisfy the boundary condition $\mathscr{B}(\boldsymbol{w}_h) = 0$.

4.3.3 Boundary conditions on impermeable walls

For $\Gamma \in \Gamma_W$ we should interpret in a suitable way the impermeability condition (4.24), i.e., $\boldsymbol{v}\cdot\boldsymbol{n} = 0$, where $\boldsymbol{v}$ is the velocity vector and $\boldsymbol{n}$ the outer unit normal to $\partial\Omega_W$. This condition has to be incorporated in some sense into the expression $\mathbf{H}_{\partial\Omega}(\boldsymbol{w}_h^{(+)},\boldsymbol{w}_h^{(-)},\boldsymbol{n})$ appearing in the definition (4.50) of the form $\boldsymbol{a}_h$.

We describe two possibilities. The first one is based on the direct use of the impermeability condition in the physical flux $\boldsymbol{P}(\boldsymbol{w},\boldsymbol{n})$ and the second applies the so-called *mirror operator* to the state $\boldsymbol{w}$.

Impermeability condition for the boundary value operator

Since the discrete solution $\boldsymbol{w} \in \boldsymbol{S}_h^{\boldsymbol{p}}$ does not satisfy the boundary condition (4.24), we introduce the so-called *boundary value operator*, see Hartmann and Leicht [2015]. $\boldsymbol{u}_\Gamma(\boldsymbol{w})$ on $\partial\Omega$.

This operator has to chosen such that in contrast to $\boldsymbol{w}_h$, $\boldsymbol{u}_\Gamma(\boldsymbol{w}_h)$ satisfies the boundary condition, i.e.$\mathscr{B}(\boldsymbol{u}_\Gamma(\boldsymbol{w}_h)) = 0$, and further that $\boldsymbol{u}_\Gamma(\boldsymbol{w}) = \boldsymbol{w}$ for any $\boldsymbol{w}$ such that $\mathscr{B}(\boldsymbol{w}) = 0$.

In this case, we define $\boldsymbol{u}_\Gamma$ by

$$\boldsymbol{u}_\Gamma(\boldsymbol{w}) := \mathbb{U}_\Gamma \boldsymbol{w} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-n_1^2 & -n_1n_2 & 0 \\ 0 & -n_1n_2 & 1-n_2^2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \boldsymbol{w} \qquad \text{on } \Gamma_W. \tag{4.52}$$

Such choice originates in the substracting of the normal component of the velocity, i.e., $\boldsymbol{v} = (v_1, v_2)$ is replaced by $\boldsymbol{v} - (\boldsymbol{n}\cdot\boldsymbol{v})\boldsymbol{n}$. That also obviously guarantees meeting the boundary condition (4.24).

Then since $\boldsymbol{u}_\Gamma(\boldsymbol{w}_h)\cdot\boldsymbol{n} = 0$, we can define $\boldsymbol{w}_h^{(-)} = \boldsymbol{u}_\Gamma(\boldsymbol{w}_h^{(+)})$. The due to (4.12), the second statement of Lemma 4.1 and (4.34) we obtain

$$\begin{aligned} \mathbf{H}^1_{\partial\Omega,\mathrm{W}}(\boldsymbol{w}^{(+)}, \boldsymbol{w}^{(-)}, \boldsymbol{n}_\Gamma) := \sum_{s=1}^{d} \boldsymbol{f}_s(\boldsymbol{u}_\Gamma(\boldsymbol{w}^{(+)}))n_s = \mathrm{p}(\boldsymbol{u}_\Gamma(\boldsymbol{w}^{(+)}))\,(0, n_1, \ldots, n_d, 0)^{\mathrm{T}} \\ = \mathbb{P}_W(\boldsymbol{u}_\Gamma(\boldsymbol{w}^{(+)}),\boldsymbol{n})\boldsymbol{u}_\Gamma(\boldsymbol{w}^{(+)}) \qquad \text{on } \Gamma_W. \end{aligned} \tag{4.53}$$

Inviscid mirror boundary conditions

This approach is based on the definition of the state vector $\boldsymbol{w}^{(-)}$ on $\Gamma \subset \Gamma_W$ in the form

$$\boldsymbol{w}^{(-)} = \mathscr{M}(\boldsymbol{w}), \tag{4.54}$$

where the boundary operator $\mathscr{M}$, called the *inviscid mirror operator*, is defined in the following way.

For $\boldsymbol{w} \in \mathscr{D}$, $\boldsymbol{w} = (\rho, \rho\boldsymbol{v}, E)^{\mathrm{T}}$ and $\boldsymbol{n} \in \mathrm{B}_1$ is the outer unit normal to $\partial\Omega$ at a point in consideration lying on Γ_W, then we set

$$\boldsymbol{v}^\perp = \boldsymbol{v} - 2(\boldsymbol{v}\cdot\boldsymbol{n})\boldsymbol{n} = (\mathbb{I} - 2\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}})\boldsymbol{v}. \tag{4.55}$$

The vectors $\boldsymbol{v}$ and $\boldsymbol{v}^\perp$ have the same tangential component but opposite normal components.

Then we define $\mathscr{M}(\boldsymbol{w}) := (\rho, \rho\boldsymbol{v}^\perp, E)^{\mathrm{T}} = \mathbb{M}_\Gamma\boldsymbol{w}$, where

$$\mathbb{M}_\Gamma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-2n_1n_1 & -2n_1n_2 & 0 \\ 0 & -2n_1n_2 & 1-2n_2n_2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

We note that employing the boundary value operator (4.52) we can rewrite this as $\mathscr{M}(\boldsymbol{w}) = 2\boldsymbol{u}_\Gamma(\boldsymbol{w}) - \boldsymbol{w}$.

Now we set the mapping $\mathbf{H}^2_{\partial\Omega,\mathrm{W}} : \mathscr{D} \times \mathrm{B}_1 \to \mathbb{R}^m$ by the same formula as the Vijayasundaram numerical fluxes on the interior edge with $\boldsymbol{w}^{(-)} = \mathscr{M}(\boldsymbol{w}^{(+)})$, i.e.,

$$\mathbf{H}^2_{\partial\Omega,\mathrm{W}}(\boldsymbol{w}^{(+)}, \boldsymbol{w}^{(-)}, \boldsymbol{n}) := \mathbf{H}_{VS}(\boldsymbol{w}^{(+)}, \mathscr{M}(\boldsymbol{w}^{(+)}), \boldsymbol{n}), \qquad \text{on } \Gamma_W, \tag{4.56}$$

where the expression $\langle \boldsymbol{w} \rangle$ which appears in the definition of $\mathbf{H}_{VS}$ is artificially defined as $\langle \boldsymbol{w} \rangle = \frac{\boldsymbol{w}^{(+)} + \mathscr{M}(\boldsymbol{w}^{(+)})}{2} = \boldsymbol{u}_\Gamma(\boldsymbol{w}^{(+)})$ on Γ_W.

Remark. For comparison between $\mathbf{H}^1_{\partial\Omega,\mathrm{W}}$ and $\mathbf{H}^2_{\partial\Omega,\mathrm{W}}$ we note that $\mathbf{H}^1_{\partial\Omega,\mathrm{W}}$ can be rewritten to

$$\begin{aligned}\mathbf{H}^1_{\partial\Omega,\mathrm{W}}(\boldsymbol{w}^{(+)}, \boldsymbol{w}^{(-)}, \boldsymbol{n}) &= \mathbb{P}_W(\boldsymbol{u}_\Gamma(\boldsymbol{w}^{(+)}))\mathbb{U}_\Gamma \boldsymbol{w}^{(+)} = \mathbb{P}(\boldsymbol{u}_\Gamma(\boldsymbol{w}^{(+)}))\mathbb{U}_\Gamma \boldsymbol{w}^{(+)} \\ &= (\mathbb{P}^{(+)}(\boldsymbol{u}_\Gamma(\boldsymbol{w}^{(+)})) + \mathbb{P}^{(-)}(\boldsymbol{u}_\Gamma(\boldsymbol{w}^{(+)}))\mathbb{M}_\Gamma)\mathbb{U}_\Gamma \boldsymbol{w}^{(+)} \\ &= \mathbf{H}_{VS}(\boldsymbol{w}^{(+)}, \mathscr{M}(\boldsymbol{w}^{(+)}), \boldsymbol{n})\mathbb{U}_\Gamma \boldsymbol{w}^{(+)} \end{aligned} \tag{4.57}$$

since $\mathbb{M}_\Gamma \mathbb{U}_\Gamma = \mathbb{U}_\Gamma$.

Lemma 4.8. *Both boundary numerical fluxes* $\mathbf{H}^1_{\partial\Omega,\mathrm{W}}$ *and* $\mathbf{H}^2_{\partial\Omega,\mathrm{W}}$ *are consistent.*

Proof. From the consistency of the boundary value operator $\boldsymbol{u}_\Gamma$ we have $\boldsymbol{u}_\Gamma(\boldsymbol{w}) = \boldsymbol{w}$ and hence also $\mathbf{H}^1_{\partial\Omega,\mathrm{W}}$ is consistent. Since $\mathscr{M}(\boldsymbol{w}) = \boldsymbol{w}$ for the exact solution we have $\mathbf{H}^2_{\partial\Omega,\mathrm{W}}(\boldsymbol{w}, \boldsymbol{n}) := \mathbf{H}(\boldsymbol{w}, \boldsymbol{w}, \boldsymbol{n})$ and consistency of $\mathbf{H}^2_{\partial\Omega,\mathrm{W}}$ is a consequence of the consistency of the numerical flux $\mathbf{H}_{VS}$. □

4.3.4 Boundary conditions on the Inlet and Outlet

Setting the boundary value $\boldsymbol{w}_h^{(-)}$ on the inlet and outlet part of the boundary $\Gamma_{IO} \subset \partial\Omega$ is quite delicate task. For instance when dealing with the flow around an isolated profile the state vector $\boldsymbol{w}_{\mathrm{BC}} = \boldsymbol{w}_\infty$ stands for the theoretical free flow in infinite distance from the profile. Yet, we cannot simply set $\boldsymbol{w}_h^{(-)} := \boldsymbol{w}_{\mathrm{BC}}$ on Γ_{IO} since the system (4.4) is hyperbolic. We adopt the approach based on the solution of the nonlinear Riemann problem, described in details in Dolejší and Feistauer [2015], which is derived for the nonstationary Euler equations given by

$$\frac{\partial \boldsymbol{w}}{\partial t} + \nabla \boldsymbol{F}(\boldsymbol{w}) = 0, \tag{4.58}$$

where $\boldsymbol{w}(x,t) : \Omega \times (0,T) \to \mathbb{R}^m$.

For the determination of the boundary condition in any given point x_Γ on the boundary Γ_{IO} we introduce a new coordinate system $(\tilde{x}_1, \dots, \tilde{x}_d)$ such that the coordinate origin lies at the point x_Γ and the axis $\tilde{x}_1$ is parallel to the outer normal $\boldsymbol{n}_\Gamma(x_\Gamma)$, see Figure 4.1. Such transformation of the coordinates is done by the mapping $\tilde{x} = \mathbb{Q}_0(\boldsymbol{n}_\Gamma)(x - x_\Gamma)$, where $\mathbb{Q}_0(\boldsymbol{n}_\Gamma)$ is the rotation matrix by the angle given by the direction of $\boldsymbol{n}_\Gamma$. Then we define

$$\boldsymbol{q}_\Gamma^{(+)} = \mathbb{Q}(\boldsymbol{n}_\Gamma)\boldsymbol{w}_\Gamma^{(+)}, \tag{4.59}$$

where $\mathbb{Q}(\boldsymbol{n}_\Gamma)$ is given by (4.20).

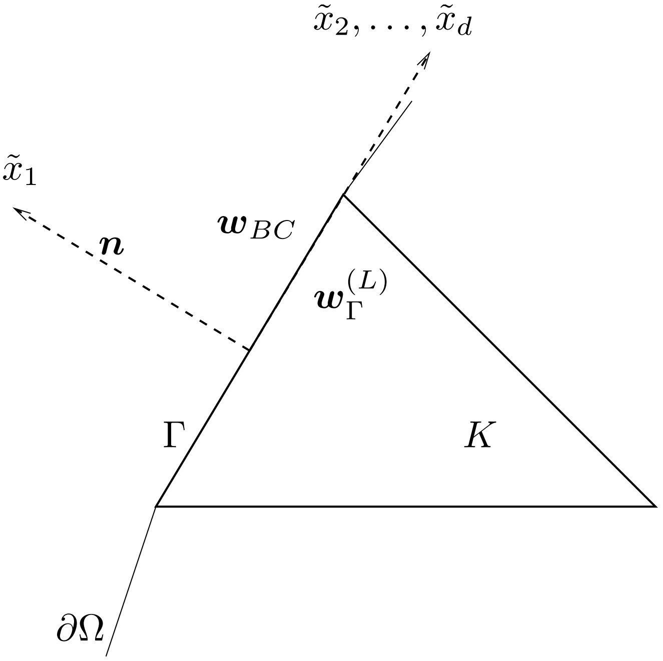

Figure 4.1: New coordinate system $(\tilde{x}_1, \dots, \tilde{x}_d)$.

Now we expoit the rotational invarience of the Euler equation introduced in (4.18) and neglect the derivative with respect to $\tilde{x}_2$ which gives us

$$\frac{\partial \boldsymbol{q}}{\partial t} + \mathbb{A}_1(\boldsymbol{q}) \frac{\partial \boldsymbol{q}}{\partial \tilde{x}_1} = 0, \quad (\tilde{x}_1, t) \in (-\infty, 0) \times [0, \infty), \tag{4.60}$$

for the transformed vector-valued function $\boldsymbol{q} = \mathbb{Q}(\boldsymbol{n}_\Gamma)\boldsymbol{w}$. To this system we add the initial and boundary conditions

$$\begin{aligned} \boldsymbol{q}(\tilde{x}_1, 0) &= \boldsymbol{q}_\Gamma^{(+)}, \quad \tilde{x}_1 < 0, \\ \boldsymbol{q}(0, t) &= \boldsymbol{q}_\Gamma^{(-)}, \quad t > 0. \end{aligned} \tag{4.61}$$

Here, $\boldsymbol{q}_\Gamma^{(+)}$ is given by (4.59) and the unknown state vector $\boldsymbol{q}_\Gamma^{(-)}$ should be determined. In order to define the boundary value $\boldsymbol{q}_\Gamma^{(-)}$ we extend the nonlinear system to

$$\frac{\partial \boldsymbol{q}}{\partial t} + \mathbb{A}_1(\boldsymbol{q}) \frac{\partial \boldsymbol{q}}{\partial \tilde{x}_1} = 0, \quad (\tilde{x}_1, t) \in (-\infty, \infty) \times [0, \infty) \tag{4.62}$$

equipped with the initial condition

$$\boldsymbol{q}(\tilde{x}_1, 0) = \begin{cases} \boldsymbol{q}_\Gamma^{(+)}, & \text{if } \tilde{x}_1 < 0, \\ \boldsymbol{q}_{\mathrm{BC}}, & \text{if } \tilde{x}_1 > 0, \end{cases} \tag{4.63}$$

where $\boldsymbol{q}_{\mathrm{BC}} = \mathbb{Q}(\boldsymbol{n}_\Gamma)\boldsymbol{w}_{\mathrm{BC}}$, see Figure 4.2.

Exact solution of the problem (4.62) and (4.63) can be constructed analytically, see, e.g., Feistauer [1993]. Furthermore, it is a piecewise analytical function which is constant along the lines $\frac{\tilde{x}_1}{t} = \lambda_i$, where λ_i, $i = 1, \dots, m$ are the eigenvalues of the matrix $\mathbb{A}_1(\boldsymbol{q})$. For a detailed derivation see e.g., [Feistauer, 1993, Paragraph 7.1.102].

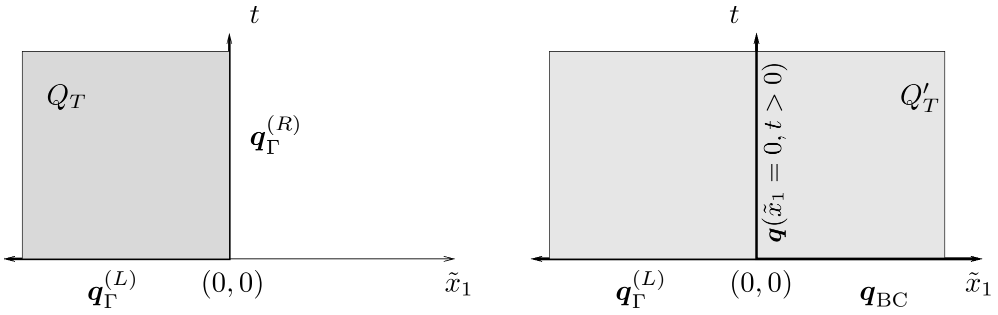
t
Q_T
$\boldsymbol{q}_\Gamma^{(R)}$
$\boldsymbol{q}_\Gamma^{(L)}$
$(0,0)$
$\tilde{x}_1$
t
$\boldsymbol{q}(\tilde{x}_1 = 0, t > 0)$
Q'_T
$\boldsymbol{q}_\Gamma^{(L)}$
$(0,0)$
$\boldsymbol{q}_{\mathrm{BC}}$
$\tilde{x}_1$

Figure 4.2: Initial-boundary value problem (4.60)–(4.61) (left) and the Riemann problem (4.62)–(4.63) (right)

On of those lines is also the ray $\tilde{x}_1 = 0$, $t > 0$ and hence we may define the inlet/outlet boundary operator based on the solution of the nonlinear Riemann problem as

$$\boldsymbol{u}_{\mathrm{RP}}(\boldsymbol{w}_\Gamma^{(+)}, \boldsymbol{w}_{\mathrm{BC}}) := \mathbb{Q}^{-1}(\boldsymbol{n}_\Gamma)\boldsymbol{q}(0, \varepsilon), \tag{4.64}$$

where $\varepsilon > 0$ may be arbitrarily small. Finally, we set $\boldsymbol{w}_h^{(-)} := \boldsymbol{u}_{\mathrm{RP}}(\boldsymbol{w}_h^{(+)}, \boldsymbol{w}_{\mathrm{BC}})$ and

$$\mathbf{H}_{\partial\Omega}(\boldsymbol{w}_h^{(+)}, \boldsymbol{w}_h^{(-)}, \boldsymbol{n}) := \mathbf{H}_{VS}(\boldsymbol{w}_h^{(+)}, \boldsymbol{u}_{\mathrm{RP}}(\boldsymbol{w}_h^{(+)}, \boldsymbol{w}_{\mathrm{BC}}), \boldsymbol{n}), \tag{4.65}$$

where we define $\langle \boldsymbol{w} \rangle = \boldsymbol{w}^{(+)}$ for any $\boldsymbol{w} \in \boldsymbol{H}^1(\Omega, \mathscr{T}_h)$ on Γ_{IO} in (4.51).

4.3.5 Primal consistency

In order to derive the primal residuals we define the primal residual of the problem (4.49) by

$$r_h(\boldsymbol{w}_h)(\boldsymbol{\varphi}) := -\boldsymbol{a}_h(\boldsymbol{w}_h, \boldsymbol{\varphi}). \tag{4.66}$$

Integrating the form (4.50) by parts on each element and using the first statement of Lemma 4.1 we get

$$\begin{aligned} r_h(\boldsymbol{w}_h)(\boldsymbol{\varphi}) := -\sum_{K \in \mathscr{T}_h} \Bigg(& \int_K \sum_{s=1}^d \frac{\partial \boldsymbol{f}_s(\boldsymbol{w}_h)}{\partial x_s} \cdot \boldsymbol{\varphi}_h \,\mathrm{d}x \\ & + \int_{\partial K \setminus \partial\Omega} \left(\boldsymbol{P}(\boldsymbol{w}_h^{(+)}, \boldsymbol{n}) - \mathbf{H}(\boldsymbol{w}_h^{(+)}, \boldsymbol{w}_h^{(-)}, \boldsymbol{n}) \right) \cdot \boldsymbol{\varphi}_h \,\mathrm{d}S \\ & + \int_{\partial K \cap \partial\Omega} \left(\boldsymbol{P}(\boldsymbol{w}_h^{(+)}, \boldsymbol{n}) - \mathbf{H}_{\partial\Omega}(\boldsymbol{w}_h^{(+)}, \boldsymbol{w}_h^{(-)}, \boldsymbol{n}) \right) \cdot \boldsymbol{\varphi}_h \,\mathrm{d}S \Bigg). \end{aligned} \tag{4.67}$$

Now we treat separatelly the terms on the inner edges and boundary edges assuming that the Vijayasundaram numerical flux (4.51) is used in the discretization (4.50). For the inner edges we have

$$\int_{\partial K \setminus \partial\Omega} \left((\boldsymbol{P}(\boldsymbol{w}_h^{(+)}, \boldsymbol{n}) - \mathbb{P}^{(+)}(\langle \boldsymbol{w}_h \rangle, \boldsymbol{n})\boldsymbol{w}_h^{(+)} - \mathbb{P}^{(-)}(\langle \boldsymbol{w}_h \rangle, \boldsymbol{n})\boldsymbol{w}_h^{(-)}) \right) \cdot \boldsymbol{\varphi}_h \,\mathrm{d}S \tag{4.68}$$

On the impermeable walls we distinguish two cases. If the boundary condition is treated by (4.53) then due to the second statement of Lemma 4.1 we have

$$\int_{\partial K\cap\Gamma_W}\left(\mathbb{P}(\boldsymbol{w}_h^{(+)},\boldsymbol{n})\boldsymbol{w}_h^{(+)}-\mathbb{P}_W(\boldsymbol{u}_\Gamma(\boldsymbol{w}_h^{(+)}),\boldsymbol{n})\boldsymbol{u}_\Gamma(\boldsymbol{w}_h^{(+)})\right)\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S. \tag{4.69}$$

If $\mathbf{H}_{\partial\Omega}$ is set according to (4.56) we get

$$\begin{aligned}&\int_{\partial K\cap\Gamma_W}\left(\boldsymbol{P}(\boldsymbol{w}_h^{(+)},\boldsymbol{n})-\mathbf{H}_{VS}(\boldsymbol{w}_h^{(+)},\mathscr{M}(\boldsymbol{w}_h^{(+)}),\boldsymbol{n})\right)\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S \\ =&\int_{\partial K\cap\Gamma_W}\left((\mathbb{P}(\boldsymbol{w}_h^{(+)},\boldsymbol{n})\boldsymbol{w}_h^{(+)}-\mathbb{P}^{(+)}(\boldsymbol{u}_\Gamma(\boldsymbol{w}_h^{(+)}),\boldsymbol{n})\boldsymbol{w}_h^{(+)}\right. \\ &\qquad\qquad\left.-\mathbb{P}^{(-)}(\boldsymbol{u}_\Gamma(\boldsymbol{w}_h^{(+)}),\boldsymbol{n})\mathscr{M}(\boldsymbol{w}_h^{(+)}))\right)\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S.\end{aligned} \tag{4.70}$$

Finally, on Γ_{IO} due to (4.65) we have

$$\begin{aligned}&\int_{\partial K\cap\Gamma_{IO}}\left(\boldsymbol{P}(\boldsymbol{w}_h^{(+)},\boldsymbol{n})-\mathbf{H}_{VS}(\boldsymbol{w}_h^{(+)},v_h{}^{(-)},\boldsymbol{n})\right)\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S \\ =&\int_{\partial K\cap\Gamma_{IO}}\left(\mathbb{P}(\boldsymbol{w}_h^{(+)},\boldsymbol{n})\boldsymbol{w}_h^{(+)}-\mathbb{P}^{(+)}(\boldsymbol{w}_h^{(+)},\boldsymbol{n})\boldsymbol{w}_h^{(+)}\right. \\ &\qquad\qquad\left.-\mathbb{P}^{(-)}(\boldsymbol{w}_h^{(+)},\boldsymbol{n})\boldsymbol{u}_{\mathrm{RP}}(\boldsymbol{w}_h^{(+)},\boldsymbol{w}_{\mathrm{BC}})\right)\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S \\ =&\int_{\partial K\cap\Gamma_{IO}}\mathbb{P}^{(-)}(\boldsymbol{w}_h^{(+)},\boldsymbol{n})(\boldsymbol{w}_h^{(+)}-\boldsymbol{u}_{\mathrm{RP}}(\boldsymbol{w}_h^{(+)},\boldsymbol{w}_{\mathrm{BC}}))\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S.\end{aligned} \tag{4.71}$$

Based on (4.67) we can define for any $K\in\mathscr{T}_h$ the element primal residuals

$$\boldsymbol{R}_K(\boldsymbol{w}_h)=-\sum_{s=1}^{d}\frac{\partial}{\partial x_s}\boldsymbol{f}_s(\boldsymbol{w}_h)=-\sum_{s=1}^{d}\mathbb{A}_s(\boldsymbol{w}_h)\frac{\partial\boldsymbol{w}_h}{\partial x_s}\qquad\text{in } K, \tag{4.72}$$

$$\boldsymbol{r}_K(\boldsymbol{w}_h)=\begin{cases}\boldsymbol{P}(\boldsymbol{w}_h^{(+)},\boldsymbol{n})-\mathbf{H}(\boldsymbol{w}_h^{(+)},\boldsymbol{w}_h^{(-)},\boldsymbol{n}) & \text{on } \partial K\backslash\partial\Omega,\\ \mathbb{P}^{(-)}(\boldsymbol{w}_h^{(+)},\boldsymbol{n})\left(\boldsymbol{w}_h^{(+)}-\boldsymbol{u}_{\mathrm{RP}}(\boldsymbol{w}_h^{(+)},\boldsymbol{w}_{\mathrm{BC}})\right) & \text{on } \partial K\cap\Gamma_{IO}\\ \boldsymbol{P}(\boldsymbol{w}_h^{(+)},\boldsymbol{n})-\mathbf{H}_{\partial\Omega,\mathrm{W}}(\boldsymbol{w}_h^{(+)},\boldsymbol{w}_h^{(-)},\boldsymbol{n}) & \text{on } \partial K\cap\Gamma_W,\end{cases} \tag{4.73}$$

where the term $\mathbf{H}_{\partial\Omega,\mathrm{W}}(\boldsymbol{w}_h^{(+)},\boldsymbol{w}_h^{(-)},\boldsymbol{n})$ stands for either $\mathbb{P}_W(\boldsymbol{u}_\Gamma(\boldsymbol{w}_h^{(+)}),\boldsymbol{n})\boldsymbol{u}_\Gamma(\boldsymbol{w}_h^{(+)})$ or $\mathbf{H}_{VS}(\boldsymbol{w}_h^{(+)},\mathscr{M}(\boldsymbol{w}_h^{(+)}),\boldsymbol{n})$ depending on whether $\mathbf{H}^1_{\partial\Omega,\mathrm{W}}$ or $\mathbf{H}^2_{\partial\Omega,\mathrm{W}}$ is used, respectively.

Employing (4.72), (4.73) in (4.67) we obtain the residual form of the problem (4.49): find $\boldsymbol{w}_h\in\boldsymbol{S}_h^{\boldsymbol{p}}$ such that

$$\sum_{K\in\mathscr{T}_h}\left(\int_K\boldsymbol{R}_K(\boldsymbol{w}_h)\cdot\boldsymbol{\varphi}_h\,\mathrm{d}x+\int_{\partial K}\boldsymbol{r}_K(\boldsymbol{w}_h)\cdot\boldsymbol{\varphi}_h^{(+)}\,\mathrm{d}S\right)=0\qquad\forall\boldsymbol{\varphi}_h\in\boldsymbol{S}_h^{\boldsymbol{p}}. \tag{4.74}$$

We summarize the previous derivation into the following result.

Lemma 4.9. *Let assume that the numerical fluxes* $\mathbf{H}$ *and* $\mathbf{H}_{\partial\Omega}$ *used on inner and boundary edges are consistent, then the discretization* (4.49) *is consistent, i.e., if* $\boldsymbol{w}\in\boldsymbol{H}^1(\Omega,\mathscr{T}_h)$ *is the exact solution of* (4.4) *it also nullifies the discrete formulation* (4.49), *i.e.,*

$$r_h(\boldsymbol{w})(\boldsymbol{\varphi})=0\qquad\forall\boldsymbol{\varphi}\in\boldsymbol{H}^1(\Omega,\mathscr{T}_h). \tag{4.75}$$

4.3.6 Newton-like method for solving the discrete primal problem

The discrete problem (4.49) is solved with the aid of the Newton method, as presented in Section 1.2.4 with several slight changes.

Let $N_{h,\boldsymbol{p}}$ denote the dimension of the space $\boldsymbol{S}_h^{\boldsymbol{p}}$ and $\mathrm{B}_{h,\boldsymbol{p}} = \{\boldsymbol{\varphi}_i(x),\ i = 1,\dots,N_{h,\boldsymbol{p}}\}$ a basis of $\boldsymbol{S}_h^{\boldsymbol{p}}$. We note that it is possible to construct a basis $\mathrm{B}_{h,\boldsymbol{p}}$ as a composition of local bases constructed separately for each $K \in \mathscr{T}_h$ for the DG method, for details see [Dolejší and Feistauer, 2015, Section 8.4.8]

We recall isomorphism between $\boldsymbol{w}_h \in \boldsymbol{S}_h^{\boldsymbol{p}}$ and its algebraic representation $\boldsymbol{\xi} \in \mathbb{R}^{N_{h,\boldsymbol{p}}}$ given by

$$\boldsymbol{w}_h(x) = \sum_{j=1}^{N_{h,\boldsymbol{p}}} \xi^j \boldsymbol{\varphi}_j(x) \in \boldsymbol{S}_h^{\boldsymbol{p}} \longleftrightarrow \boldsymbol{\xi} = (\xi^j)_{j=1}^{N_{h,\boldsymbol{p}}} \in \mathbb{R}^{N_{h,\boldsymbol{p}}}, \tag{4.76}$$

where $\xi^j \in \mathbb{R},\ j = 1,\dots,N_{h,\boldsymbol{p}}$ are its basis coefficients with respect to $\mathrm{B}_{h,\boldsymbol{p}}$. Further the algebraic representation of the systems (4.49) reads: Find $\boldsymbol{\xi} \in \mathbb{R}^{N_{h,\boldsymbol{p}}}$ such that

$$\boldsymbol{F}_h(\boldsymbol{\xi}) = \boldsymbol{0}, \tag{4.77}$$

where

$$\boldsymbol{F}_h(\boldsymbol{\xi}) = (\boldsymbol{a}_h(\boldsymbol{w}_h; \boldsymbol{\varphi}_i))_{i=1}^{N_{h,\boldsymbol{p}}}. \tag{4.78}$$

The exact evaluation of the Jacobi matrix $\frac{D\boldsymbol{F}_h(\bar{\boldsymbol{\xi}})}{D\boldsymbol{\xi}}$ in (4.39) needed for the Newton method, see (1.93), may be costly and moreover the terms corresponding to the numerical Vijayasundaram fluxes are not even continuously differentiable. Therefore we do not compute the derivative $a_h{}'[u_h](\cdot,\cdot)$ precisely, but instead we approximate it by the linearized form

$$a_h{}'[\boldsymbol{w}_h](\cdot,\cdot) \approx \boldsymbol{a}_h^{\mathrm{L}}(\boldsymbol{w}_h; \cdot,\cdot). \tag{4.79}$$

Then we define the $N_{h,\boldsymbol{p}} \times N_{h,\boldsymbol{p}}$ *flux matrix*

$$\mathbb{C}_h\left(\bar{\boldsymbol{\xi}}\right) = \left(\boldsymbol{a}_h^{\mathrm{L}}(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i)\right)_{i,j=1}^{N_{h,\boldsymbol{p}}} \tag{4.80}$$

approximating the Jacobi matrix $\mathbb{C}_h\left(\bar{\boldsymbol{\xi}}\right) \approx \frac{D\boldsymbol{F}_h(\bar{\boldsymbol{\xi}})}{D\boldsymbol{\xi}}$.

Finally, the Newton-like method follows directly algorithm presented in Section 1.2.4, only the Jacobi matrix is replaced by $\mathbb{C}_h\left(\bar{\boldsymbol{\xi}}^l\right)$ in (1.93).

In practical computations, we solve the steady state Euler equations as the time dependent ones where the (pseudo-)time stepping helps to improve the global convergence of the Newton method, similarly as the damping parameter λ^l in (1.92).

Authors of Hartmann and Leicht [2015], Hartmann [2005] suggest to approximate these terms by finite difference method, and in Hartmann [2005] they show that it should significantly speed up the convergence of the nonlinear solver. Our experience is, on the other hand, a bit different. We did not observe any significant change of the convergence when the finite difference members were added or omitted. Furthermore, later in this chapter we present an analysis of a discrete adjoint problem based on the linearization $\boldsymbol{a}_h^{\mathrm{L}}$ which shows that it represents a reasonable discretization of the adjoint problem (4.39).

Remark. Besides solution of the algebraic system (1.93) the most expensive parts within this process is the composition of the matrix $\mathbb{C}_h(\boldsymbol{\xi}^l)$. Hence, in order to accelerate the computation, we do not upgrade the matrix $\mathbb{C}_h(\boldsymbol{\xi}^l)$ in every step of the Newton-like method (1.92)–(1.93).

4.3.7 Linearization of the form a_h

Here, we present this formal linearization of the form $\boldsymbol{a}_h$ which we use as an approximation of the Jacobi matrix. Recalling the form (4.50) from the discrete adjoint problem

$$\begin{aligned}
\boldsymbol{a}_h(\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = &-\sum_{K\in\mathscr{T}_h}\int_K \sum_{s=1}^{d}(\mathbb{A}_s(\boldsymbol{w}_h)\boldsymbol{w}_h)\cdot\frac{\partial\boldsymbol{\varphi}_h}{\partial x_s}\,\mathrm{d}x && (=:\zeta_1)\\
&+\sum_{K\in\mathscr{T}_h}\int_{\partial K\setminus\partial\Omega}\mathbf{H}(\boldsymbol{w}_h^{(+)},\boldsymbol{w}_h^{(-)},\boldsymbol{n})\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S && (=:\zeta_2)\\
&+\sum_{K\in\mathscr{T}_h}\int_{\partial K\cap\Gamma_W}\mathbf{H}^i_{\partial\Omega,\mathrm{W}}(\boldsymbol{w}_h^{(+)},\boldsymbol{w}_h^{(-)},\boldsymbol{n})\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S && (=:\zeta_3)\\
&+\sum_{K\in\mathscr{T}_h}\int_{\partial K\cap\Gamma_{IO}}\mathbf{H}_{\partial\Omega}(\boldsymbol{w}_h^{(+)},\boldsymbol{w}_h^{(-)},\boldsymbol{n})\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S && (=:\zeta_4),
\end{aligned}\tag{4.81}$$

we linearize each of the four terms $\zeta_1,\dots,\zeta_4$.

For the first one we define the linearized form $\zeta_1^{\mathrm{L}}:\boldsymbol{S}_h^{\boldsymbol{p}}\times\boldsymbol{S}_h^{\boldsymbol{p}}\times\boldsymbol{S}_h^{\boldsymbol{p}}\to\mathbb{R}$ by

$$\zeta_1^{\mathrm{L}}(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = -\sum_{K\in\mathscr{T}_h}\int_K\sum_{s=1}^{d}\mathbb{A}_s(\bar{\boldsymbol{w}}_h)\boldsymbol{w}_h\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,\mathrm{d}x. \tag{4.82}$$

Then employing Lemma 4.1 we have $\zeta_1^{\mathrm{L}}(\boldsymbol{w}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h)=\zeta_1$ and obviously ζ_1^{L} is linear with respect to its second and third argument.

For linearization of the term ζ_2 we exploit the definition of the Vijayasundaram numerical fluxes (4.51). Since every inner edge in the triangulation appears twice in the sum we may reorganize the summation and then exploit the fact that

$$\mathbb{P}^-\left(\langle\bar{\boldsymbol{w}}_h\rangle_\Gamma,\boldsymbol{n}_K\right)\boldsymbol{w}_h^{(-)} = -\mathbb{P}^+\left(\langle\bar{\boldsymbol{w}}_h\rangle_\Gamma,-\boldsymbol{n}_K\right)\boldsymbol{w}_h^{(-)}. \tag{4.83}$$

So the linearized form $\zeta_2^{\mathrm{L}}:\boldsymbol{S}_h^{\boldsymbol{p}}\times\boldsymbol{S}_h^{\boldsymbol{p}}\times\boldsymbol{S}_h^{\boldsymbol{p}}\to\mathbb{R}$ reads

$$\begin{aligned}
\zeta_2^{\mathrm{L}}(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) &= \sum_{K\in\mathscr{T}_h}\int_{\partial K\setminus\partial\Omega}\Big[\mathbb{P}^+\big(\langle\bar{\boldsymbol{w}}_h\rangle_\Gamma,\boldsymbol{n}_K\big)\boldsymbol{w}_h^{(+)}\\
&\qquad\qquad+\mathbb{P}^-\big(\langle\bar{\boldsymbol{w}}_h\rangle_\Gamma,\boldsymbol{n}_K\big)\boldsymbol{w}_h^{(-)}\Big]\cdot\boldsymbol{\varphi}_h^{(+)}\,\mathrm{d}S\\
&= \sum_{K\in\mathscr{T}_h}\int_{\partial K\setminus\partial\Omega}\mathbb{P}^+\left(\langle\bar{\boldsymbol{w}}_h\rangle_\Gamma,\boldsymbol{n}_K\right)\boldsymbol{w}_h^{(+)}\cdot[\![\boldsymbol{\varphi}_h]\!]_K\,\mathrm{d}S.
\end{aligned}\tag{4.84}$$

Obviously $\zeta_2^{\mathrm{L}}(\boldsymbol{w}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h)=\zeta_2$ and ζ_2^{L} is linear with respect to its second and third argument.

Regarding the term ζ_3 we have to proceed separately for each of the approaches $\mathbf{H}^i_{\partial\Omega,\mathrm{W}}$, $i=1,2$.

Based on the definition (4.53) of $\mathbf{H}^1_{\partial\Omega,\mathrm{W}}$ we may introduce its in the following form

$$\mathbf{H}^{2,\mathrm{L}}_{\partial\Omega,\mathrm{W}}(\bar{\boldsymbol{w}},\boldsymbol{w},\boldsymbol{n}) = \mathbb{P}_W(\boldsymbol{u}_\Gamma(\bar{\boldsymbol{w}}),\boldsymbol{n})\mathbb{U}_\Gamma\boldsymbol{w}, \quad \bar{\boldsymbol{w}},\boldsymbol{w}\in\mathscr{D},\ \boldsymbol{n}\in\mathrm{B}_1, \tag{4.85}$$

The linearization $\mathbf{H}^2_{\partial\Omega,\mathrm{W}}$ may be introduced similarly to (4.84) and since $\frac{\boldsymbol{w}+\mathscr{M}(\boldsymbol{w})}{2} = u_\Gamma(\boldsymbol{w})$ we get

$$\mathbf{H}^{3,\mathrm{L}}_{\partial\Omega,\mathrm{W}}(\bar{\boldsymbol{w}},\boldsymbol{w},\boldsymbol{n}) = \left(\mathbb{P}^+\left(u_\Gamma(\bar{\boldsymbol{w}}),\boldsymbol{n}\right) + \mathbb{P}^-\left(u_\Gamma(\bar{\boldsymbol{w}}),\boldsymbol{n}\right)\mathbb{M}_\Gamma\right)\boldsymbol{w}^{(+)}. \tag{4.86}$$

Employing the linearized forms (4.85) and (4.86) we set

$$\begin{aligned}\zeta_3^{i,\mathrm{L}}(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) &= \sum_{K\in\mathscr{T}_h}\int_{\partial K\cap\Gamma_W}\mathbf{H}^{\mathrm{i,L}}_{\partial\Omega,\mathrm{W}}(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{n})\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S \\ &= \sum_{K\in\mathscr{T}_h}\int_{\partial K\cap\Gamma_W}\boldsymbol{\varphi}_h^{\mathrm{T}}\mathbb{H}^{\mathrm{i,L}}_{\mathrm{W}}(\bar{\boldsymbol{w}}_h,\boldsymbol{n})\boldsymbol{w}_h\,\mathrm{d}S,\end{aligned} \tag{4.87}$$

where $i=1,2$ and the matrix $\mathbb{H}^{\mathrm{i,L}}_{\mathrm{W}}(\bar{\boldsymbol{w}}_h,\boldsymbol{n})$ corresponds to one of the matrices in (4.85) and (4.86), i.e.,

$$\mathbb{H}^{1,\mathrm{L}}_{\mathrm{W}}(\bar{\boldsymbol{w}}_h,\boldsymbol{n}) = \mathbb{P}_W(\boldsymbol{u}_\Gamma(\bar{\boldsymbol{w}}),\boldsymbol{n})\mathbb{U}_\Gamma, \tag{4.88}$$

$$\mathbb{H}^{2,\mathrm{L}}_{\mathrm{W}}(\bar{\boldsymbol{w}}_h,\boldsymbol{n}) = \mathbb{P}^+\left(u_\Gamma(\bar{\boldsymbol{w}}),\boldsymbol{n}\right) + \mathbb{P}^-\left(u_\Gamma(\bar{\boldsymbol{w}}),\boldsymbol{n}\right)\mathbb{M}_\Gamma. \tag{4.89}$$

By exploring the definitions of $\mathbf{H}^{\mathrm{i,L}}_{\partial\Omega,\mathrm{W}}$, $i=1,2$ we get that both $\zeta_3^{i,\mathrm{L}}$ are linear with respect to the second and third argument and they meet the favorable property $\zeta_3^{i,\mathrm{L}}(\boldsymbol{w}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \zeta_3^i$.

At last, ζ_4 is approximated with the aid of the forms

$$\zeta_4^{\mathrm{L}}(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \sum_{K\in\mathscr{T}_h}\int_{\partial K\cap\Gamma_{IO}}\left(\mathbb{P}^+(\bar{\boldsymbol{w}}_h^{(+)},\boldsymbol{n}_K)\boldsymbol{w}_h^{(+)}\right)\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S, \tag{4.90}$$

and

$$\tilde{\boldsymbol{a}}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_h) = -\sum_{K\in\mathscr{T}_h}\int_{\partial K\cap\Gamma_{IO}}\left(\mathbb{P}^-(\bar{\boldsymbol{w}}_h^{(+)},\boldsymbol{n}_K)\bar{\boldsymbol{w}}_h^{(-)}\right)\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S, \tag{4.91}$$

where according to (4.64) function $\bar{\boldsymbol{w}}_h^{(-)}$ is the solution of the Riemann problem (4.62) given by $\boldsymbol{u}_{\mathrm{RP}}(\bar{\boldsymbol{w}}_h^{(+)},\boldsymbol{w}_{\mathrm{BC}})$ on Γ_{IO}. Let us underline that in the arguments of $\mathbb{P}^\pm$ we do not use the mean value of the state vectors from the left and right side as in (4.51). Moreover, if $\operatorname{supp}\boldsymbol{\varphi}_h\cap(\Gamma_{IO})=\emptyset$, then $\tilde{\boldsymbol{a}}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_h)=0$.

Obviously, due to (4.91) and (4.90), we have

$$\zeta_4^{\mathrm{L}}(\boldsymbol{w}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) - \tilde{\boldsymbol{a}}_h(\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \zeta_4. \tag{4.92}$$

Taking together all the previously defined linearizations, we set

$$\boldsymbol{a}_h^{\mathrm{L}}(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \sum_{i=1}^4\zeta_i^{\mathrm{L}}(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) \tag{4.93}$$

and we evidently get $\boldsymbol{a}_h(\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \boldsymbol{a}_h^{\mathrm{L}}(\boldsymbol{w}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) - \tilde{\boldsymbol{a}}_h(\boldsymbol{w}_h,\boldsymbol{\varphi}_h)$.

If we introduce the vector

$$\boldsymbol{d}_h\left(\bar{\boldsymbol{\xi}}\right) := \left(\tilde{\boldsymbol{a}}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_i)\right)_{i=1}^{N_{h,p}}, \tag{4.94}$$

we obtain algebraic representation of the equality (4.93)

$$\boldsymbol{F}_h(\boldsymbol{\xi}) = \mathbb{C}_h(\boldsymbol{\xi})\boldsymbol{\xi} - \boldsymbol{d}_h(\boldsymbol{\xi}). \tag{4.95}$$

4.3.8 Discrete adjoint problem and adjoint consistency

In this section we introduce the discrete adjoint problem based on the linearization of the form $\boldsymbol{a}_h$ given by (4.93). Further, the adjoint consistency of the discretization is studied.

In order to obtain an adjoint consistent scheme, it is necessary to modify the target functional J defined in (4.26) as generally presented in (1.61). For the functional given by (4.26) we set

$$J_h(\boldsymbol{w}_h) = \int_{\Gamma_W} \mathbf{H}^i_{\partial\Omega,\mathrm{W}}(\boldsymbol{w}_h^{(+)}, \boldsymbol{w}_h^{(-)}, \boldsymbol{n}) \cdot \tilde{\vartheta}\, \mathrm{d}S. \tag{4.96}$$

Here $\mathbf{H}^i_{\partial\Omega,\mathrm{W}} \in \{\mathbf{H}^1_{\partial\Omega,\mathrm{W}}, \mathbf{H}^1_{\partial\Omega,\mathrm{W}}\}$ and $\tilde{\vartheta} = (0, \vartheta_1, \vartheta_2, 0)^{\mathrm{T}}$ on Γ_W, where ϑ is chosen either by (4.27) or (4.29).

For the exact solution of (4.4) it holds due to the consistency of the numerical fluxes $\mathbf{H}_{\partial\Omega,\mathrm{W}}$ that

$$\mathbf{H}_{\partial\Omega,\mathrm{W}}(\boldsymbol{w}^{(+)}, \boldsymbol{w}^{(-)}, \boldsymbol{n}) \cdot \tilde{\vartheta} = \mathrm{p}(\boldsymbol{w})\boldsymbol{n} \cdot \vartheta. \tag{4.97}$$

Hence by comparison of the definitions (4.26) and (4.96) it can be seen that

$$J_h(\boldsymbol{w}) = J(\boldsymbol{w}),$$

i.e., this particular modification J_h is so-called consistent, see Section 1.1.9. Further, using the linearization of the numerical fluxes (4.85) and (4.86), we introduce the linearization of the discrete functional

$$J_h^{\mathrm{L}}(\boldsymbol{w}_h; \boldsymbol{\varphi}_h) = \int_{\Gamma_W} \mathbf{H}^{\mathrm{i,L}}_{\partial\Omega,\mathrm{W}}(\boldsymbol{w}_h; \boldsymbol{\varphi}_h, \boldsymbol{n}) \cdot \tilde{\vartheta}\, \mathrm{d}S = \int_{\Gamma_W} \boldsymbol{\varphi}_h{}^{\mathrm{T}} \mathbb{H}^{\mathrm{i,L}}_{\mathrm{W}}(\boldsymbol{w}_h{}^{(+)}, \boldsymbol{n}))^{\mathrm{T}} \tilde{\vartheta}\, \mathrm{d}S. \tag{4.98}$$

Definition 4.10. *Finally we introduce the discrete adjoint problem. We say that $\boldsymbol{z}_h \in \boldsymbol{S}_h^{\boldsymbol{p}}$ is the discrete adjoint solution if it satisfies*

$$\boldsymbol{a}_h^{\mathrm{L}}(\boldsymbol{w}_h; \boldsymbol{\varphi}_h, \boldsymbol{z}_h) = J_h^{\mathrm{L}}(\boldsymbol{w}_h; \boldsymbol{\varphi}_h) \qquad \forall \boldsymbol{\varphi}_h \in \boldsymbol{S}_h^{\boldsymbol{p}}, \tag{4.99}$$

where $\boldsymbol{a}_h^{\mathrm{L}}$ is given by (4.93)*. Further we define the adjoint residual*

$$r_h^*(\boldsymbol{w}_h, \boldsymbol{z}_h)(\boldsymbol{\varphi}_h) := J_h^{\mathrm{L}}(\boldsymbol{w}_h; \boldsymbol{\varphi}_h) - \boldsymbol{a}_h^{\mathrm{L}}(\boldsymbol{w}_h; \boldsymbol{\varphi}_h, \boldsymbol{z}_h). \tag{4.100}$$

Theorem 4.11. *Let $\mathbf{H}_{VS}$ be the Vijayasundaram numerical flux. Let J_h be the modified target functional defined in* (4.96)*. Then the discretization* (4.49) *is adjoint consistent, i.e., the exact solution $\boldsymbol{w}$ of the flow equations* (4.4) *and its adjoint counter-part $\boldsymbol{z}$, solving the continuous adjoint problem* (4.39)*, satisfy*

$$r_h^*(\boldsymbol{w}, \boldsymbol{z})(\boldsymbol{\varphi}) = 0 \qquad \forall \boldsymbol{\varphi} \in \tilde{\boldsymbol{V}}. \tag{4.101}$$

Proof. Similarly to (4.74) we introduce the residual formulation of the discrete problem (4.99)

$$\sum_{K \in \mathscr{T}_h} \int_K \boldsymbol{R}_K^*(\boldsymbol{w}_h, \boldsymbol{z}_h) \cdot \boldsymbol{\varphi}_h \, \mathrm{d}x + \int_{\partial K} \boldsymbol{r}_K^*(\boldsymbol{w}_h, \boldsymbol{z}_h) \cdot \boldsymbol{\varphi}_h^{(+)}\, \mathrm{d}S = 0 \qquad \forall \boldsymbol{\varphi}_h \in \boldsymbol{S}_h^{\boldsymbol{p}}, \tag{4.102}$$

where the volume and edge residual terms are defined by

$$\boldsymbol{R}_K^*(\boldsymbol{w}_h,\boldsymbol{z}_h) = \sum_{s=1}^{d} \mathbb{A}_s^T(\boldsymbol{w}_h)\frac{\partial \boldsymbol{z}_h}{\partial x_s} \qquad \text{in } K, \tag{4.103}$$

$$\boldsymbol{r}_K^*(\boldsymbol{w}_h,\boldsymbol{z}_h) = \begin{cases} -\mathbb{P}^{(+)}(\langle \boldsymbol{w}_h\rangle, \boldsymbol{n})^{\mathrm{T}}[\![\boldsymbol{z}_h]\!] & \text{on } \partial K\backslash\partial\Omega, \\ -\mathbb{P}^{(+)}(\boldsymbol{w}_h{}^{(+)}, \boldsymbol{n})^{\mathrm{T}}\boldsymbol{z}_h & \text{on } \partial K\cap\Gamma_{IO}, \\ (\mathbb{H}_{\mathrm{W}}^{\mathrm{j,L}}(\boldsymbol{w}_h{}^{(+)}, \boldsymbol{n}))^{\mathrm{T}}(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{z}_h),\, i\in\{1,2\} & \text{on } \partial K\cap\Gamma_W. \end{cases} \tag{4.104}$$

which follows from the definitions of $\zeta_i^{\mathrm{L}}, i = 1,\dots,4$ in (4.82), (4.84), (4.87), (4.90) and the definition of the linearization of the modified target functional (4.98).

Employing the adjoint residuals (4.103), (4.104), we can rewrite (4.101) to

$$\sum_{K\in\mathscr{T}_h}\int_K \boldsymbol{R}_K^*(\boldsymbol{w},\boldsymbol{z})\cdot\boldsymbol{\varphi}\,\mathrm{d}x + \int_{\partial K}\boldsymbol{r}_K^*(\boldsymbol{w},\boldsymbol{z})\cdot\boldsymbol{\varphi}^{(+)}\,\mathrm{d}S \qquad \forall\boldsymbol{\varphi}\in\tilde{\boldsymbol{V}}. \tag{4.105}$$

Reminding the strong formulation of the continuous adjoint problem (4.43) we see that $\boldsymbol{R}_K^*(\boldsymbol{w},\boldsymbol{z}) = 0$ for any $K\in\mathscr{T}_h$. Further, due to the assumed smoothness of the adjoint solution $\boldsymbol{z}$ we also have $\boldsymbol{r}_K^*(\boldsymbol{w},\boldsymbol{z}) = 0$ on $\partial K\backslash\partial\Omega$.

On the boundary we examine separately Γ_W and Γ_{IO}. If the numerical flux $\mathbf{H}^1_{\partial\Omega,\mathrm{W}}$ given by (4.53) is used on Γ_W, we exploit that $\boldsymbol{u}_\Gamma(\boldsymbol{w}^{(+)}) = \boldsymbol{w}^{(+)}$ for the exact solution and $\boldsymbol{u}_\Gamma$ given by (4.52). Hence recalling (4.88) and exploiting the first statement of the Lemma 4.3 we get

$$\begin{aligned} \boldsymbol{r}_K^*(\boldsymbol{w},\boldsymbol{z}) &= \left(\mathbb{H}_{\mathrm{W}}^{1,\mathrm{L}}(\boldsymbol{w},\boldsymbol{n})\right)^{\mathrm{T}}(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{z}) = \mathbb{U}_\Gamma^{\mathrm{T}}\mathbb{P}_W^{\mathrm{T}}(\boldsymbol{u}_\Gamma(\boldsymbol{w}),\boldsymbol{n})(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{z}) \\ &= \frac{D\mathrm{p}(\boldsymbol{w})}{D\boldsymbol{w}}(0,n_1,n_2,0)\cdot\left(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{z}\right) \\ &= \frac{D\mathrm{p}(\boldsymbol{w})}{D\boldsymbol{w}}\left(\boldsymbol{n}\cdot\boldsymbol{\vartheta} - (n_1 z_2 + n_2 z_3)\right) = 0, \end{aligned} \tag{4.106}$$

since the adjoint solution $\boldsymbol{z}$ satisfies the boundary condition (4.44). Furthermore, we did not have to use the properties of the space $\tilde{\boldsymbol{V}}$ and hence this choice of the boundary numerical flux nullifies the edge residual even for any $\boldsymbol{\varphi}_h\in\boldsymbol{S}_h^{\boldsymbol{p}}$.

If the numerical flux $\mathbf{H}^2_{\partial\Omega,\mathrm{W}}$ given by (4.56) is used on Γ_W then we have

$$\begin{aligned} J_h^{\mathrm{L}}(\boldsymbol{w};\boldsymbol{\varphi}) - \zeta_3^{3,\mathrm{L}}(\boldsymbol{w},\boldsymbol{\varphi},\boldsymbol{z}) = \int_{\partial K\cap\Gamma_W} &\left(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{z}\right)^{\mathrm{T}}\big(\mathbb{P}^{(+)}(\boldsymbol{u}_\Gamma(\boldsymbol{w}),\boldsymbol{n}) \\ &+ \mathbb{P}^{(-)}(\boldsymbol{u}_\Gamma(\boldsymbol{w}),\boldsymbol{n})\mathbb{M}_\Gamma\big)\boldsymbol{\varphi}\,\mathrm{d}S. \end{aligned} \tag{4.107}$$

Here, we have to exploit the properties of $\boldsymbol{\varphi}$ belonging to $\tilde{\boldsymbol{V}}$. Since $\boldsymbol{\varphi}\in\tilde{\boldsymbol{V}}$, it holds $\mathscr{B}'[\boldsymbol{w}](\varphi) = n_1\boldsymbol{\varphi}_2 + n_2\boldsymbol{\varphi}_3 = 0$ and hence $\mathbb{M}_\Gamma\boldsymbol{\varphi} = \boldsymbol{\varphi}$ and further since the exact solution satisfies $\boldsymbol{v}\cdot\boldsymbol{n} = 0$ due to (4.34) we obtain that

$$\begin{aligned} &\int_{\partial K\cap\Gamma_W}\left(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{z}\right)^{\mathrm{T}}(\mathbb{P}^{(+)}(\boldsymbol{u}_\Gamma(\boldsymbol{w}),\boldsymbol{n}) + \mathbb{P}^{(-)}(\boldsymbol{u}_\Gamma(\boldsymbol{w}),\boldsymbol{n})\mathbb{M}_\Gamma)\boldsymbol{\varphi}\,\mathrm{d}S \\ &= \int_{\partial K\cap\Gamma_W}(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{z})^{\mathrm{T}}\mathbb{P}(\boldsymbol{w},\boldsymbol{n})\boldsymbol{\varphi}\,\mathrm{d}S = \int_{\partial K\cap\Gamma_W}(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{z})^{\mathrm{T}}\mathbb{P}_W(\boldsymbol{w},\boldsymbol{n})\boldsymbol{\varphi}\,\mathrm{d}S = 0. \end{aligned} \tag{4.108}$$

Finally, on Γ_{IO} the function $\tilde{\boldsymbol{\vartheta}} = 0$ and hence $\boldsymbol{r}_K^*(\boldsymbol{w},\boldsymbol{z}) = 0$ since the residual term in (4.104) precisely equals to the boundary condition in (4.44). □

Let us shortly speculate on the pertinence of the discretization (4.99) of the adjoint problem (4.43)–(4.44). The discrete formulation (4.43) is based on linearization rather than on proper differentiation of the nonlinear discrete problem (4.49) like it is usually done, cf. Hartmann and Leicht [2015] or Hartmann [2006]. On the other hand, the omitted terms contain derivatives of the numerical fluxes (4.51) which lack the required smoothness to be differentiated exactly. In Hartmann [2005] these terms are approximated by finite differences for Lax-Friedrichs and Vijayasundaram numerical fluxes. We note that omitting those terms does not cause any source of inconsistency into the discrete problem and from point of view it nicely corresponds to the continuous formulation of the adjoint problem (4.43), and hence the discretization (4.99) seems as a quite reasonable DG discretization of the problem (4.43).

Remark (Relation to scalar advection equation). Our reasoning is also motivated by the article Bezchlebová et al. [2016], where the scalar advective problem

$$\partial_t u + \boldsymbol{b} \cdot \nabla u = 0 \tag{4.109}$$

for some prescribed flow $\boldsymbol{b}$ and solution u is studied. Here the advective term $\boldsymbol{b} \cdot \nabla u$ is discretized as

$$\begin{aligned} b_h(u_h, \varphi_h) = \sum_{K \in \mathscr{T}_h} \Big(\int_K (\boldsymbol{b} \cdot \nabla u_h) \varphi \, \mathrm{d}x - \int_{\partial K^{(-)} \setminus \partial\Omega} (\boldsymbol{b} \cdot \boldsymbol{n}) [\![u_h]\!] \varphi \, \mathrm{d}S \\ - \int_{\partial K^{(-)} \cap \partial\Omega} (\boldsymbol{b} \cdot \boldsymbol{n}) u_h \varphi \, \mathrm{d}S \Big), \end{aligned} \tag{4.110}$$

where $\partial K^{(-)} = \{x \in \partial K; \boldsymbol{b} \cdot \boldsymbol{n} < 0\}$, $\partial K^{(+)} = \{x \in \partial K; \boldsymbol{b} \cdot \boldsymbol{n} >= 0\}$.

Assuming that $\boldsymbol{w}$ is given, we can progress in a similar way to Bezchlebová et al. [2016] for the discretization of the adjoint problem (4.43).

Multiplying (4.43) by a test function $\boldsymbol{\varphi} \in H^1(\Omega, \mathscr{T}_h)$ and integrating by parts over a mesh element K leads to

$$\begin{aligned} -\sum_{K \in \mathscr{T}_h} \int_K \boldsymbol{\varphi}^{\mathrm{T}} \mathbb{A}_s^T(\boldsymbol{w}) \frac{\partial \boldsymbol{z}_h}{\partial x_s} \, \mathrm{d}x = \sum_{K \in \mathscr{T}_h} \Big(\int_K \frac{\partial}{\partial x_s} (\mathbb{A}_s(\boldsymbol{w}) \boldsymbol{\varphi})^{\mathrm{T}} \boldsymbol{z}_h \, \mathrm{d}x \\ - \int_{\partial K} (\mathbb{P}^{(+)}(\boldsymbol{w}, \boldsymbol{n}) + \mathbb{P}^{(-)}(\boldsymbol{w}, \boldsymbol{n}))^{\mathrm{T}} \boldsymbol{z}_h \boldsymbol{\varphi} \, \mathrm{d}S \Big). \end{aligned} \tag{4.111}$$

We replace $-\mathbb{P}^{(+)}(\boldsymbol{w}, \boldsymbol{n})^{\mathrm{T}} \boldsymbol{z}_h$ by $-\mathbb{P}^{(+)}(\boldsymbol{w}, \boldsymbol{n})^{\mathrm{T}} \boldsymbol{z}_h{}^{(-)}$ on inner edges, c.f. (4.83). Then exploiting the boundary condition $\mathbb{P}^{(+)}(\boldsymbol{w}, \boldsymbol{n})^{\mathrm{T}} \boldsymbol{z} = 0$ we omit the corresponding term on Γ_{IO}.

Further, on Γ_W we may write $\mathbb{P}(\boldsymbol{w}, \boldsymbol{n}) = \mathbb{P}_W(\boldsymbol{w}, \boldsymbol{n}) + \mathbb{R}_W(\boldsymbol{w}, \boldsymbol{n})$ as in the proof of Lemma 4.4. Employing the boundary condition $z_2 n_1 + z_3 n_2 = \boldsymbol{\vartheta} \cdot \boldsymbol{n}$ and the definition of the target functional (4.26) and its directional derivative (4.32), we replace $\boldsymbol{\varphi}^{\mathrm{T}} \mathbb{P}_W(\boldsymbol{w}, \boldsymbol{n})^{\mathrm{T}} \boldsymbol{z}_h$ by $-\boldsymbol{\varphi}^{\mathrm{T}} \mathbb{P}_W(\boldsymbol{w}, \boldsymbol{n})^{\mathrm{T}} \tilde{\boldsymbol{\vartheta}}$ in (4.111). Then integrating by parts for the second time we get (similarly to (4.84) it leads to the jump terms on the inner edges)

$$\begin{aligned} -\sum_{K \in \mathscr{T}_h} \int_K \mathbb{A}_s^T(\boldsymbol{w}) \frac{\partial \boldsymbol{z}_h}{\partial x_s} \boldsymbol{\varphi} \, \mathrm{d}x + \int_{\partial K \setminus \partial\Omega} (\mathbb{P}^{(+)}(\boldsymbol{w}, \boldsymbol{n})^{\mathrm{T}} [\![\boldsymbol{z}_h]\!] \boldsymbol{\varphi} \, \mathrm{d}S \\ + \int_{\partial K \cap \Gamma_{IO}} (\mathbb{P}^{(+)}(\boldsymbol{w}, \boldsymbol{n})^{\mathrm{T}} \boldsymbol{z}_h \boldsymbol{\varphi} \, \mathrm{d}S + \int_{\partial K \cap \Gamma_W} (\mathbb{P}_W(\boldsymbol{w}, \boldsymbol{n})^{\mathrm{T}} \boldsymbol{z}_h \boldsymbol{\varphi} \, \mathrm{d}S = J'[\boldsymbol{w}](\boldsymbol{\varphi}). \end{aligned} \tag{4.112}$$

This equals the adjoint formulation (4.99) using the numerical flux $\mathbf{H}^1_{\partial\Omega,\mathrm{W}}$ on Γ_W if we would have omitted the boundary operator $\boldsymbol{u}_\Gamma$.

Comparing (4.110) to (4.112) shows analogies between

$$\boldsymbol{b}\cdot\nabla u_h \sim -\sum_{s=1}^{d}\mathbb{A}_s^T(\boldsymbol{w})\frac{\partial \boldsymbol{z}_h}{\partial x_s}, \qquad (\boldsymbol{b}\cdot\boldsymbol{n})[\![u_h]\!] \sim -\mathbb{P}^{(+)}(\boldsymbol{w},\boldsymbol{n})^{\mathrm{T}}[\![\boldsymbol{z}_h]\!]. \tag{4.113}$$

4.4 Error estimates

The goal-oriented error estimates were introduced in Section 1.2. We recall the error identity (1.73)

$$J(\boldsymbol{w})-J(\boldsymbol{w}_h)\approx\frac{1}{2}r_h(\boldsymbol{w}_h)(\boldsymbol{z}-\boldsymbol{\varphi}_h)+\frac{1}{2}r_h^*(\boldsymbol{w}_h,\boldsymbol{z}_h)(\boldsymbol{w}-\psi_h)+\mathscr{R}_h^{(3)} \qquad \forall\boldsymbol{\varphi}_h,\psi_h\in\boldsymbol{S}_h^{\boldsymbol{p}}. \tag{4.114}$$

In the formulation of the adjoint problem (4.99), both the derivative of the target functional $J'[\boldsymbol{w}_h]$ and the derivative of the discrete form $a_h{}'[\boldsymbol{w}_h]$ were approximated by linearizations given by (4.98) and (4.93), respectively. That may lead to additional errors, but we omit those in the error estimates similarly as the term $\mathscr{R}_h^{(3)}$ is usually omitted even for exactly differentiated schemes. The numerical experiments, presented in Section 4.5, indicate that this source of errors does not notably change the estimates (compared to results published for similar problem, see, e.g., in Hartmann [2007], Hartmann and Leicht [2015]).

The error identity (1.73) must be replaced by some computable quantities. Hence we replace the exact solutions $\boldsymbol{w}$ and $\boldsymbol{z}$ in (1.73) by some higher-order reconstruction denoted here $\boldsymbol{w}_h^+$ and $\boldsymbol{z}_h^+$, respectively. Those can be computed either globally – on a finer mesh and/or using polynomials of higher degree, or with local reconstructions. Then we take the following approximation of the error of the quantity of interest as a starting point when deriving the computable error estimates. In particular, we define

$$\begin{aligned}J(\boldsymbol{w})-J(\boldsymbol{w}_h)&\approx\eta^{\mathrm{I}}(\boldsymbol{w}_h,\boldsymbol{z}_h) \\ &:=\frac{1}{2}\left(r_h(\boldsymbol{w}_h)(\boldsymbol{z}_h^+-\Pi\boldsymbol{z}_h^+)+r_h^*(\boldsymbol{w}_h,\boldsymbol{z}_h)(\boldsymbol{w}_h^+-\Pi\boldsymbol{w}_h^+)\right),\end{aligned} \tag{4.115}$$

where $\Pi:[L^2(\Omega)]^m\to\boldsymbol{S}_h^{\boldsymbol{p}}$ denotes an arbitrary projection on $\boldsymbol{S}_h^{\boldsymbol{p}}$.

Remark. On the other hand, when we employ only the second order estimate (1.77), we may include the replacement of $a_h{}'$ by $\boldsymbol{a}_h^{\mathrm{L}}$ more naturally into the approximation of the unknown adjoint solution $\boldsymbol{z}$ by $\boldsymbol{z}_h^+$ since (1.77) only gives

$$J(\boldsymbol{w})-J(\boldsymbol{w}_h)=r_h(\boldsymbol{w}_h)(\boldsymbol{z}-\boldsymbol{\psi}_h)+\mathscr{R}_h^{(2)}\approx r_h(\boldsymbol{w}_h)(\boldsymbol{z}_h^+-\boldsymbol{\psi}_h) \tag{4.116}$$

and the term $a_h{}'$ is not explicitly included, but since we follow with the derivation of the goal-oriented anisotropic error estimates, c.f. Chapter 3 we base our estimates on (1.73) even though it is more heuristic for (4.99).

Similarly to Section 2.2 we further rewrite the estimate (4.115) element-wise

$$\eta^{\mathrm{I}}(\boldsymbol{w}_h,\boldsymbol{z}_h)=\sum_{K\in\mathscr{T}_h}\eta_K^{\mathrm{I}} \tag{4.117}$$

where

$$\eta_K^{\rm I} = \eta_K^{\rm I}(\boldsymbol{w}_h, \boldsymbol{z}_h) = \frac{1}{2}\left(r_h(\boldsymbol{w}_h)((\boldsymbol{z}_h^+ - \Pi \boldsymbol{z}_h^+)\,\chi_K) + r_h^*(\boldsymbol{w}_h, \boldsymbol{z}_h)((\boldsymbol{w}_h^+ - \Pi \boldsymbol{z}_h^+)\,\chi_K)\right). \tag{4.118}$$

Here, χ_K denotes the characteristic functions of mesh elements.

Further, we note that the local error indicators which can be used for mesh adaptation may be given by $|\eta_K^{\rm I}|$ or alternatively by $|r_h(\boldsymbol{w}_h)((\boldsymbol{z}_h^+ - \Pi \boldsymbol{z}_h^+)\,\chi_K)|$ or $|r_h^*(\boldsymbol{w}_h, \boldsymbol{z}_h)((\boldsymbol{w}_h^+ - \Pi \boldsymbol{z}_h^+)\,\chi_K)|$. We note that the absolute values are necessary only for defining mesh refinement indicators, while we avoid those for estimating of the error $J(\boldsymbol{w}) - J(\boldsymbol{w}_h)$ since that would lead to a needless overestimation of the true error.

Several examples of local reconstruction techniques relevant for the DG method were introduced in Section 2.2 for linear problems. While the weighted least-squares reconstruction (Section 2.2.1) can be used in the same way as it is defined for the linear problems the reconstruction based on the solution of local problems (Section 2.2.2) has to be adapted to use it for nonlinear problems. That will be the goal of the next section.

4.4.1 Reconstruction based on solving local nonlinear problems

The adjoint discrete solution can reconstructed directly by the algorithm presented in Section 2.2.2 for linear problems. The situation is a bit different for the reconstruction of $\boldsymbol{w}_h$ due to the nonlinearity of the problem (4.49).

Similarly to (2.38), for each $K \in \mathscr{T}_h$, we prescribe $\boldsymbol{w}_K^+ : \Omega \to \mathbb{R}^m$ satisfying:

$$\begin{aligned} &\text{(i)} \quad \boldsymbol{w}_K^+|_{K'} := \boldsymbol{w}_h|_{K'} \text{ for all } K' \neq K,\\ &\text{(ii)} \quad \boldsymbol{w}_K^+|_K \in (P^{p_K+1}(K))^m,\\ &\text{(iii)} \quad a_h(\boldsymbol{w}_K^+, \varphi_h) = 0 \quad \forall \varphi_h \in (P^{p_K+1}(K))^m, \end{aligned} \tag{4.119}$$

where a_h is the form given by (4.50). Since evidently $\boldsymbol{w}_K^+ \in \boldsymbol{S}_h^{p+1}$, we can define $\boldsymbol{w}_h^+ \in \boldsymbol{S}_h^{p+1}$ by $\boldsymbol{w}_h^+|_K := \boldsymbol{w}_K^+ \quad \forall K \in \mathscr{T}_h$.

Since the problem (4.119) (iii) is nonlinear we calculate the reconstruction $\boldsymbol{w}_K^+$ iteratively using Newton method. We set $\boldsymbol{w}_{K,(0)}^+ = \boldsymbol{w}_h$ and $\boldsymbol{w}_{K,(j+1)}^+ = \boldsymbol{w}_{K,(j)}^+ + \tilde{\boldsymbol{w}}_{K,(j)}^+$ for $j = 0, 1, \ldots$

We denote $N_K^+ = \dim(P^{p_K+1}(K))^m = ((p_K+2)(p_K+3)/2)^m$ and we choose a basis $\boldsymbol{\varphi}_K = \varphi_{h,K}^1, \ldots, \varphi_{h,K}^{N_K}, \ldots, \varphi_{h,K}^{N_K^+}$ of $(P^{p_K+1})^m$.

We define the local residual vector $\boldsymbol{R}_K(\boldsymbol{w}_{K,(j)}^+) = \{r_K^i(\boldsymbol{w}_{K,(j)}^+)\}_{i=1}^{N_K^+}$, where

$$r_K^i(\boldsymbol{w}_{K,(j)}^+) = -a_h(\boldsymbol{w}_{K,(j)}^+, \varphi_{h,K}^i) = r_h(\boldsymbol{w}_{K,(j)}^+)(\varphi_{h,K}^i)$$

and $\mathbb{C}_{K,K}^+(\boldsymbol{w}_{K,(j)}^+) \in \mathbb{R}^{N_K^+ \times N_K^+}$ is the diagonal block of matrix $\mathbb{C}_h(\boldsymbol{w}_{K,(j)}^+)$ from (4.80) enlarged by $N_K^+ - N_K$ rows and columns. Let $\tilde{\boldsymbol{W}}_K^{(j)}$ be the vector of the basis coefficents of $\tilde{\boldsymbol{w}}_{K,(j)}^+$, i.e., $\tilde{\boldsymbol{w}}_{K,(j)}^+ = \tilde{\boldsymbol{W}}_K^{(j)} \cdot \boldsymbol{\varphi}_K$. Then the Newton method for the problem (4.119) reads:

```
Algorithm 5: Newton algorithm for (4.119)
Set $\boldsymbol{w}^+_{K,(0)} = \boldsymbol{w}_h$, $j = 0$ and TOL $> 0$ ;
while $\left\|\boldsymbol{R}_K(\boldsymbol{w}^+_{K,(j)})\right\| >$ TOL do
    solve $\mathbb{C}^+_{K,K}(\boldsymbol{w}^+_{K,(j)})\tilde{\boldsymbol{W}}^{(j)}_K = \boldsymbol{R}_K(\boldsymbol{w}^+_{K,(j)})$;
    set $\boldsymbol{w}^+_{K,(j+1)} = \boldsymbol{w}^+_{K,(j)} + \tilde{\boldsymbol{w}}^+_{K,(j)}$ ;
    $j = j+1$ ;
end
```

Lemma 4.12. *When the reconstruction based on solving local problems is used for both primal and adjoint discrete solution and if we perform only one iteration of the Algorithm 5, i.e.*$\boldsymbol{w}^+_K = \boldsymbol{w}^+_{K,(1)}$*, for the primal reconstruction, then it holds that* $\eta_{\mathrm{S,K}} = \eta^*_{\mathrm{S,K}}$ *for any* $K \in \mathscr{T}_h$*.*

Proof. If we carry out only one iteration of the Algorithm 5, then we have $\boldsymbol{w}^+_K = \boldsymbol{w}_h + \tilde{\boldsymbol{w}}^+_{K,(0)}$, where $\tilde{\boldsymbol{w}}^+_{K,(0)}$ solves

$$a_h{}'[\boldsymbol{w}_h](\tilde{\boldsymbol{w}}^+_{K,(0)}, \varphi_{h,K}) = r_h(\boldsymbol{w}_h, \varphi_{h,K}) \quad \forall \varphi_{h,K} \in (P^{p_K+1}(K))^m. \tag{4.120}$$

Further, the adjoint reconstruction $\boldsymbol{z}^+_K = \boldsymbol{z}_h + \tilde{\boldsymbol{z}}^+_K$ satisfies

$$a_h{}'[\boldsymbol{w}_h](\varphi_{h,K}, \tilde{\boldsymbol{z}}^+_K) = r^*_h(\boldsymbol{w}_h, \boldsymbol{z}_h)(\varphi_{h,K}) \quad \forall \varphi_{h,K} \in (P^{p_K+1}(K))^m. \tag{4.121}$$

Hence, employing consecutively $r^*_h(\boldsymbol{w}_h, \boldsymbol{z}_h)(\boldsymbol{w}_h|_K) = 0$, (4.121) and (4.120) we obtain

$$\begin{aligned} \eta^*_{\mathrm{S,K}} &= r^*_h(\boldsymbol{w}_h, \boldsymbol{z}_h)(\boldsymbol{w}^+_K|_K) = r^*_h(\boldsymbol{w}_h, \boldsymbol{z}_h)(\tilde{\boldsymbol{w}}^+_{K,(0)}) \\ &= a_h{}'[\boldsymbol{w}_h](\tilde{\boldsymbol{w}}^+_{K,(0)}, \tilde{\boldsymbol{z}}^+_K) = r_h(\boldsymbol{w}_h, \tilde{\boldsymbol{z}}^+_K) = r_h(\boldsymbol{w}_h, \boldsymbol{z}^+_K|_K) = \eta_{\mathrm{S,K}}. \end{aligned} \tag{4.122}$$

□

4.4.2 Dual weighted residual error estimate

Similarly to (3.4) and (3.6) we further estimate the residuals $r_h(\boldsymbol{w}_h)(\cdot)$ and $r^*_h(\boldsymbol{w}_h, \boldsymbol{z}_h)(\cdot)$ of the primal problem (4.49) and adjoint problem (4.99), respectively.

Employing the integration by parts like in (4.67) and (4.74) the element-wise primal residual can be further estimated by

$$\begin{aligned} r_h(\boldsymbol{w}_h)(\boldsymbol{\varphi}) &= \sum_{K \in \mathscr{T}_h} \left(\int_K \boldsymbol{R}_K(\boldsymbol{w}_h) \cdot \boldsymbol{\varphi} \, \mathrm{d}x + \int_{\partial K} \boldsymbol{r}_K(\boldsymbol{w}_h) \cdot \boldsymbol{\varphi}^{(+)} \, \mathrm{d}S \right) \\ &\leq \sum_{K \in \mathscr{T}_h} \left(\sum_{i=1}^m R^i_{K,V} \left\|\boldsymbol{\varphi}^i\right\|_K + R^i_{K,B} \left\|\boldsymbol{\varphi}^i\right\|_{\partial K} \right) \end{aligned} \tag{4.123}$$

where

$$R^i_{K,V} := \left\|\boldsymbol{R}^i_K(\boldsymbol{w}_h)\right\|_K, \quad R^i_{K,B} := \left\|\boldsymbol{r}^i_K(\boldsymbol{w}_h)\right\|_{\partial K}$$

and the terms $\boldsymbol{R}^i_K(\boldsymbol{w}_h)$ and $\boldsymbol{r}^i_K(\boldsymbol{w}_h)$ denote the i-th component, $i = 1, \dots, m$, of the local residual terms given by (4.72) and (4.73). Similarly $\boldsymbol{\varphi}^i$ denotes the i-th component of the vector function $\boldsymbol{\varphi}$.

Similarly, we may proceed for the adjoint residual

$$r_h^*(\boldsymbol{w}_h,\boldsymbol{z}_h)(\boldsymbol{\varphi}) = \sum_{K\in\mathscr{T}_h}\left(\int_K \boldsymbol{R}_K^*(\boldsymbol{w}_h,\boldsymbol{z}_h)\cdot\boldsymbol{\varphi}\,\mathrm{d}x + \int_{\partial K}\boldsymbol{r}_K^*(\boldsymbol{w}_h,\boldsymbol{z}_h)\cdot\boldsymbol{\varphi}^{(+)}\,\mathrm{d}S\right) \tag{4.124}$$

$$\leq \sum_{K\in\mathscr{T}_h}\left(\sum_{i=1}^m R_{K,V}^{*,i}\,\|\boldsymbol{\varphi}^i\|_K + R_{K,B}^{*,i}\,\|\boldsymbol{\varphi}^i\|_{\partial K}\right) \tag{4.125}$$

where

$$R_{K,V}^{*,i} := \left\|\boldsymbol{R}_K^{*,i}(\boldsymbol{w}_h,\boldsymbol{z}_h)\right\|_K, \quad R_{K,B}^{*,i} := \left\|\boldsymbol{r}_K^{*,i}(\boldsymbol{w}_h,\boldsymbol{z}_h)\right\|_{\partial K}$$

and the terms $\boldsymbol{R}_K^{*,i}(\boldsymbol{w}_h,\boldsymbol{z}_h)$ and $\boldsymbol{r}_K^{*,i}(\boldsymbol{w}_h,\boldsymbol{z}_h)$ are the the i-th components, $i=1,\dots,m$, of the local residual terms given (4.103) and (4.104).

Altogether, we obtain

$$|\eta^{\mathrm{I}}(\boldsymbol{w}_h,\boldsymbol{z}_h)| \leq \eta^{\mathrm{II}}(\boldsymbol{w}_h,\boldsymbol{z}_h), \qquad \eta^{\mathrm{II}}(\boldsymbol{w}_h,\boldsymbol{z}_h) = \sum_{K\in\mathscr{T}_h}\eta_K^{\mathrm{II}}(\boldsymbol{w}_h,\boldsymbol{z}_h), \tag{4.126}$$

where

$$\eta_K^{\mathrm{II}}(\boldsymbol{w}_h,\boldsymbol{z}_h) = \frac{1}{2}\left(\sum_{i=1}^m R_{K,V}^i\,\|(\boldsymbol{z}_h^+ - \Pi\boldsymbol{z}_h^+)^i\|_K + R_{K,B}^i\,\|(\boldsymbol{z}_h^+ - \Pi\boldsymbol{z}_h^+)^i\|_{\partial K}\right. \tag{4.127}$$
$$\left. + R_{K,V}^{*,i}\,\|(\boldsymbol{w}_h^+ - \Pi\boldsymbol{w}_h^+)^i\|_K + R_{K,B}^{*,i}\,\|(\boldsymbol{w}_h^+ - \Pi\boldsymbol{w}_h^+)^i\|_{\partial K}\right).$$

We call the terms including $\boldsymbol{z}_h^+ - \Pi\boldsymbol{z}_h^+$ and $\boldsymbol{w}_h^+ - \Pi\boldsymbol{w}_h^+$ *weights* and for that reason the estimates shaped like (4.126) are usually referred as *dual weighted residual* error estimate, cf. Bangerth and Rannacher [2003].

4.4.3 Goal-oriented anisotropic error estimates

In this section we derive goal-oriented error estimates enabling the anisotropic *hp*-mesh adaptation for the DG discretization of inviscid compressible flow problems. We exploit the approach introduced in Chapter 3, which gives an upper bound to the estimate η_K^{II} on each element $K\in\mathscr{T}_h$.

Let $\boldsymbol{w}_h^+ \in \boldsymbol{S}_h^{p+1}$ and $\boldsymbol{z}_h^+ \in \boldsymbol{S}_h^{p+1}$ be the higher-order reconstructions of the primal and adjoint solutions of (4.49) and (4.99), respectively. Let $K\in\mathscr{T}_h$ be an arbitrary triangle and $\{\lambda_K,\,\sigma_K,\,\phi_K\}$ denotes its anisotropy introduced in Definition 3.1.

We recall the estimates of the interpolation function from Section 3.2.3. For any function $w\in P^{p+1}(K)$ we have

$$w(x) - \Pi_{\bar{x},p}w(x) = w_{\bar{x},p}^{\mathrm{int}}(x), \tag{4.128}$$

where $\Pi_{\bar{x},p}w(x)$ is the $p-$degree polynomial approximation based on the Taylor expansion given by (3.17) and $w_{\bar{x},p}^{\mathrm{int}}$ is the so-called interpolation error function given by (3.18). Function $w_{\bar{x},p}^{\mathrm{int}}(x)$ is a $(p+1)$-function in the sense of Definition 3.2 and hence it can be estimated using Lemma (3.3). There exist $A_w\geq 0$, $\rho_w\geq 1$ and $\varphi_w\in[0,2\pi)$ such that

$$|w_{\bar{x},p}^{\mathrm{int}}(x)| \leq A_w\left((x-\bar{x})^{\mathrm{T}}\mathbb{Q}_{\varphi_w}\mathbb{D}_{\rho_w}^{p+1}\mathbb{Q}_{\varphi_w}^{\mathrm{T}}(x-\bar{x})\right)^{\frac{p+1}{2}},\ x\in\Omega. \tag{4.129}$$

Therefore, in (4.127), we set projection operator $\Pi : S_h^{p+1} \to S_h^p$ by $\Pi\big|_K := \Pi_{x_K,p}$, $K \in \mathscr{T}_h$, where $\Pi_{x_K,p}$ is given by (3.17) and x_K is the barycenter of element K.

Relation (4.128) implies that for each $K \in \mathscr{T}_h$ functions $(\boldsymbol{w}_h^+ - \Pi\boldsymbol{w}_h^+)^i|_K$ and $(\boldsymbol{z}_h^+ - \Pi\boldsymbol{z}_h^+)^i|_K$, $i = 1,\dots,m$ are the $(p_K+1)-$functions and hence we may introduce, c.f. (3.38),

$$\begin{aligned} \{A_{w^i}, \rho_{w^i}, \varphi_{w^i}\} &\text{ the anisotropy of } (\boldsymbol{w}_h^+ - \Pi\boldsymbol{w}_h^+)^i|_K,\, i = 1,\dots,m \\ \{A_{z^i}, \rho_{z^i}, \varphi_{z^i}\} &\text{ the anisotropy of } (\boldsymbol{z}_h^+ - \Pi\boldsymbol{z}_h^+)^i|_K,\, i = 1,\dots,m \end{aligned} \tag{4.130}$$

which are given by the estimate (3.15) and depend only on the $(p+1)^{\text{th}}$-derivatives of $\boldsymbol{w}_h^+$ and $\boldsymbol{z}_h^+$, respectively.

Finally, applying the Lemma 3.5 together with (4.130) gives the anisotropic weighting terms

$$\begin{aligned} \left\|(\boldsymbol{w}_h^+ - \Pi\boldsymbol{w}_h^+)^i\right\|_K &\le \Big(\frac{A_{w^i}^2 \lambda_K^{2(p_K+2)}}{2p_K+4}\boldsymbol{G}_{K,w^i}\Big)^{1/2} &&=: \theta_{K,\mathrm{V}}^i, \\ \left\|(\boldsymbol{z}_h^+ - \Pi\boldsymbol{z}_h^+)^i\right\|_K &\le \Big(\frac{A_{z^i}^2 \lambda_K^{2(p_K+2)}}{2p_K+4}\boldsymbol{G}_{K,z^i}\Big)^{1/2} &&=: \theta_{K,\mathrm{V}}^{*,i}, \\ \left\|(\boldsymbol{w}_h^+ - \Pi\boldsymbol{w}_h^+)^i\right\|_{\partial K} &\le \Big(A_{w^i}^2 \lambda_K^{2p_K+3}\sigma_K \boldsymbol{G}_{K,w^i}\Big)^{1/2} &&=: \theta_{K,\mathrm{B}}^i, \\ \left\|(\boldsymbol{z}_h^+ - \Pi\boldsymbol{z}_h^+)^i\right\|_{\partial K} &\le \Big(A_{z^i}^2 \lambda_K^{2p_K+3}\sigma_K \boldsymbol{G}_{K,z^i}\Big)^{1/2} &&=: \theta_{K,\mathrm{B}}^{*,i}, \end{aligned} \tag{4.131}$$

where

$$\begin{aligned} \boldsymbol{G}_{K,w^i} &= \boldsymbol{G}(p_K+1, p_K+1, \rho_{w^i}, \varphi_{w^i}; \sigma_K, \phi_K), \\ \boldsymbol{G}_{K,z^i} &= \boldsymbol{G}(p_K+1, p_K+1, \rho_{z^i}, \varphi_{z^i}; \sigma_K, \phi_K) \end{aligned} \tag{4.132}$$

are defined by (3.21).

Then applying (4.131) on the weighting terms in (4.127) we get

$$\eta_K^{\mathrm{II}}(\boldsymbol{w}_h, \boldsymbol{z}_h) \le \eta_K^{\mathrm{III}}(\boldsymbol{w}_h, \boldsymbol{z}_h) \tag{4.133}$$

where

$$\eta_K^{\mathrm{III}}(\boldsymbol{w}_h, \boldsymbol{z}_h) := \frac{1}{2}\left(\sum_{i=1}^m R_{K,V}^i \theta_{K,\mathrm{V}}^{*,i} + R_{K,B}^i \theta_{K,\mathrm{B}}^{*,i} + R_{K,V}^{*,i}\theta_{K,\mathrm{V}}^i + R_{K,B}^{*,i}\theta_{K,\mathrm{B}}^i\right). \tag{4.134}$$

Now, the adaptation may be carried out following the Algorithm 3 as presented in Section 3.3.

Remark. The optimal shape of each element K is computed using the optimization technique presented in Section 3.3.1 minimizing (3.51). The error indicator (4.134) contains sixteen weighting terms, which may possibly have different anisotropic features. Therefore, the shape optimization may be simplifies using the error indicator

$$\tilde{\eta}_K^{\mathrm{III}}(\boldsymbol{w}_h, \boldsymbol{z}_h) := \frac{1}{2}\left(R_{K,V}^1 \theta_{K,\mathrm{V}}^{*,1} + R_{K,B}^1 \theta_{K,\mathrm{B}}^{*,1} + R_{K,V}^{*,1}\theta_{K,\mathrm{V}}^1 + R_{K,B}^{*,1}\theta_{K,\mathrm{B}}^1\right) \tag{4.135}$$

which takes into account only the first component (corresponding to density) of the primal and adjoint solutions.

4.5 Numerical experiments

In this section we present several experiments which illustrate performance of the goal-oriented error estimation technique for inviscid compressible flow problems introduced in this chapter.

4.5.1 Adjoint consistency of the DG discretization

There was spend a lot of space in this chapter to analyze the adjoint consistency of the DG discretization of the problem (4.4). A special attention was paid to the impermeability condition on Γ_W ($\boldsymbol{v} \cdot \boldsymbol{n} = 0$) in the discrete scheme since Γ_W is the support of the target functional (4.26). We presented two ways how the numerical flux can be defined on Γ_W. Then target functional (4.26) has to be modified properly in order to provide the adjoint consistency (4.101).

The purpose of this experiment is to investigate how the adjoint consistency of the DG discretization influences the the smoothness of the discrete adjoint solution $\boldsymbol{z}_h$. We consider an inviscid flow with Mach number $M = 0.5$ around the NACA0012 airfoil. The curved shape of the profile, defined by an analytical parametrization, see e.g. [Dolejší and Feistauer, 2015, Section 8.5.3], is approximated by piece-wise cubic polynomial functions. The angle of attack is chosen as $\alpha = 0^\circ$, so the solution should be symmetrical along the axis $y = 0$. Let us note that $\tan\alpha = v_2/v_1$, where $(v_1, v_2)^{\mathrm{T}}$ is the far- field velocity vector.

The experiment was computed on a fixed mesh, initially refined in the vicinity of the profile, see Figure 4.3 (right).

Figure 4.4 shows the first two components of the solution of the discrete adjoint problem (4.99) using the numerical flux $\mathbf{H}^2_{\partial\Omega,\mathrm{W}}$ given by (4.56) accompanied with the modified target functional (4.96) (left) and with the original target functional (4.26) (right). We see that while the adjoint consistent discretization leads to a smooth solution $\boldsymbol{z}_h$, the inconsistent one contains non-physical oscillations. Furthermore, let us denote that discretization using the numerical flux $\mathbf{H}^1_{\partial\Omega,\mathrm{W}}$ given by (4.53) leads to very similar results.

We note that our results are in agreement with [Hartmann and Leicht, 2015, Section 6.1] where a similar example is presented, but with proper differentiation of the form $\boldsymbol{a}_h$.

4.5.2 Subsonic flow

We stay with a laminar flow with a Mach number $M = 0.5$ around the NACA0012 profile and we test the adaptive h and hp-anisotropic refinement based on the Algorithm 3 in Section 3.3 adapted for the indicators (4.134).

Drag coefficient

First, we focus on the decrease of the error measured with respect to the target functional. Therefore, we set $J(\boldsymbol{w})$ expressing the drag coefficient according to (4.26) and the angle of attack $\alpha = 0^\circ$. Then the exact value of the drag is $c_D = 0$ (the flow under consideration is inviscid). Therefore it is easy to plot the decrease of the error of the target quantity compared to its error estimates η^{I} and η^{II}. In Figure 4.5 we see this decrease for the h version of the algorithm with fixed $p = 2$ (top) and for the hp-version

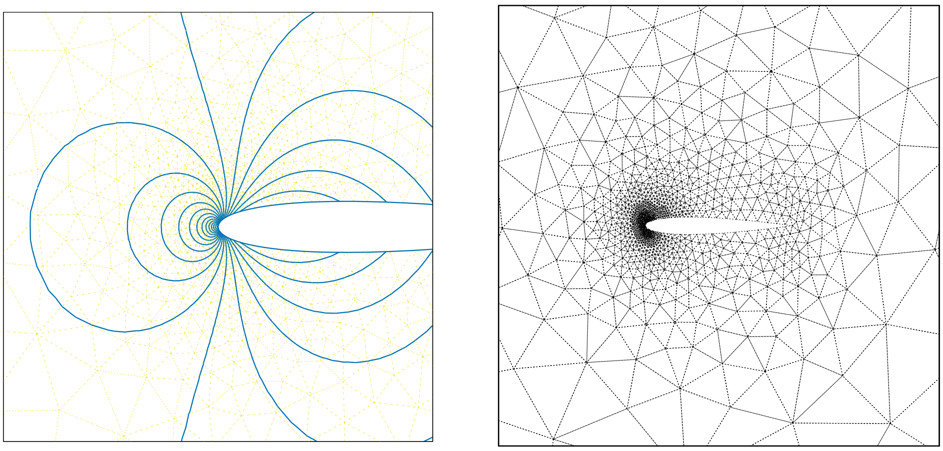

Figure 4.3: Subsonic inviscid flow around the NACA 0012 profile ($M = 0.5$, $\alpha = 0°$): first component of the primal solution $\boldsymbol{w}_h$ (left) and the computational mesh in the vicinity of the profile (right).

(bottom) of the algorithm. It can be observed that both η^{I} and η^{II} approximate the true error quite accurately although η^{I} underestimates the error slightly. We see that the hp-version is superior to the h-version with respect to both number of degrees of freedom (DOF) and computational time.

We note that the computational times depicted on the right side of the figure are only indicative since our software is not fully optimized for fast computations and the experiments were computed on a standard laptop, but we observed at least 30% speedup in the computations when the hp-method was used.

In Figure 4.6 the local refinement indicators η_K^{I} and η_K^{II} for the first, fourth and seventh (final) mesh are depicted and in Figure 4.7 the corresponding hp-meshes are drawn. We see that apart the small region around the trailing edge of the profile strong $p-$ adaptation is performed.

Lift coefficient

For the approximation of the lift coefficient we consider $\alpha = 1.25°$. Since the exact value of $J(\boldsymbol{w})$ is not a priori known for this setting, we use the reference value of $c_L^{\mathrm{ref}} = 1.754 \cdot 10^{-1} \pm 10^{-5}$ computed on the final mesh with the hp-adaptive algorithm.

Therefore, we present the decrease of the error of the target functional with the respective error estimates only for the $h-$ adaptive computation, see Figure 4.8. We see that the error estimates work worse than in the previous case – η^{I} underestimates the error and, quite the other way, η^{II} overestimates it almost ten times.

This may be caused by the weaker regularity of the adjoint solution for the lift coefficient. For comparison, we present the results with the same experiment also for the drag coefficient. We perform several hp-adaptation cycles for both cases and the final hp-meshes are shown in Figure 4.9 (top). In Figure 4.9 (bottom), the first component of the adjoint solutions $\boldsymbol{z}_h$ is drawn in the vicinity of the profile for the drag coefficient (left) and for the lift coefficient (right). According to the hp-adaptation it seems that while the adjoint problem for the drag coefficient is quite smooth (p is high in the surrounding of the profile), for the lift coefficient the polynomial degree p

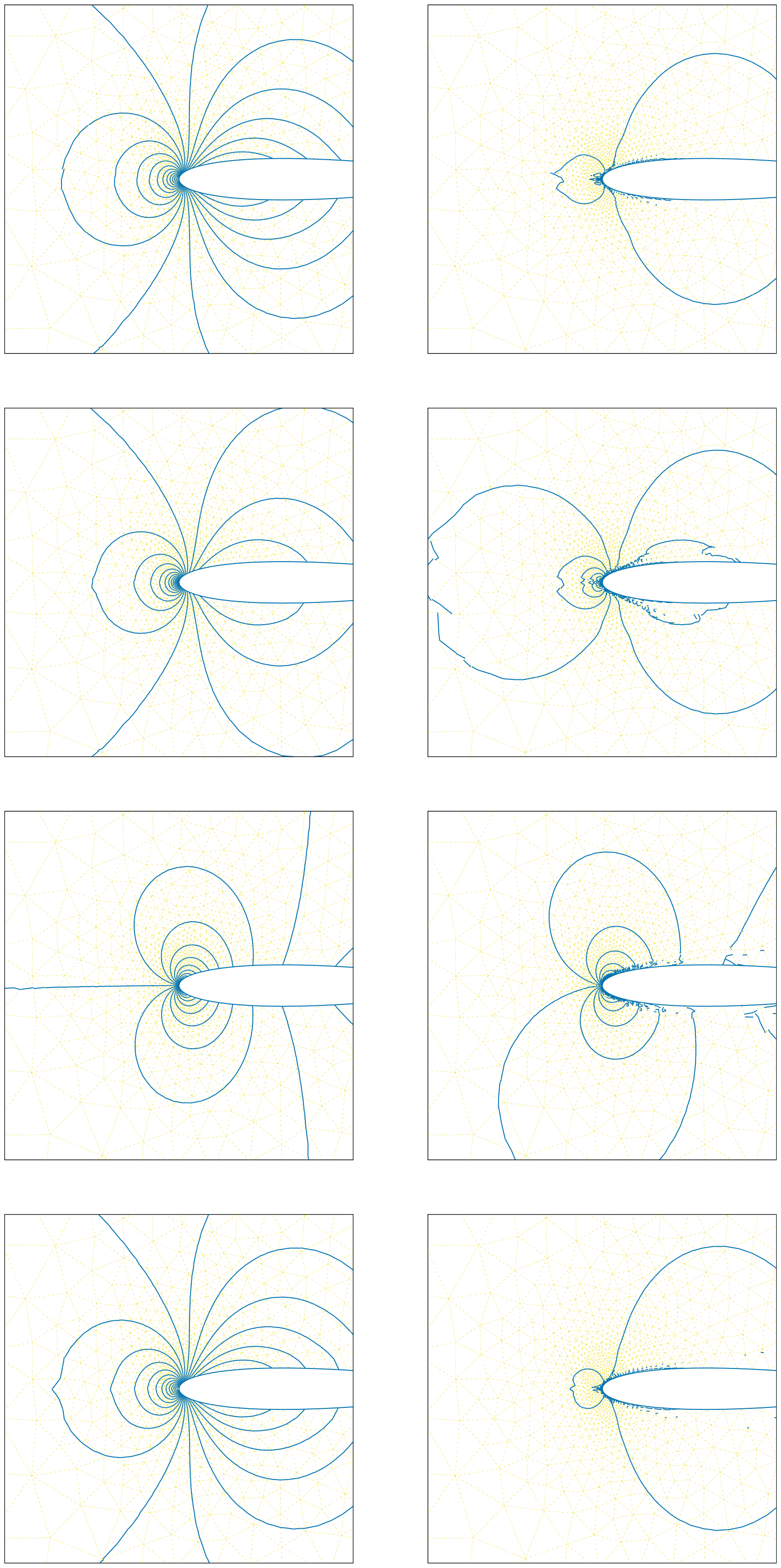

Figure 4.4: Subsonic inviscid flow around the NACA 0012 profile ($M = 0.5$, $\alpha = 0°$): consistent (left) and inconsistent (right) discretization of the target functional of the drag coefficient. The four components of the discrete adjoint solutions $z_h = (z_h^1, z_h^2, z_h^3, z_h^4)$ are consecutively shown.

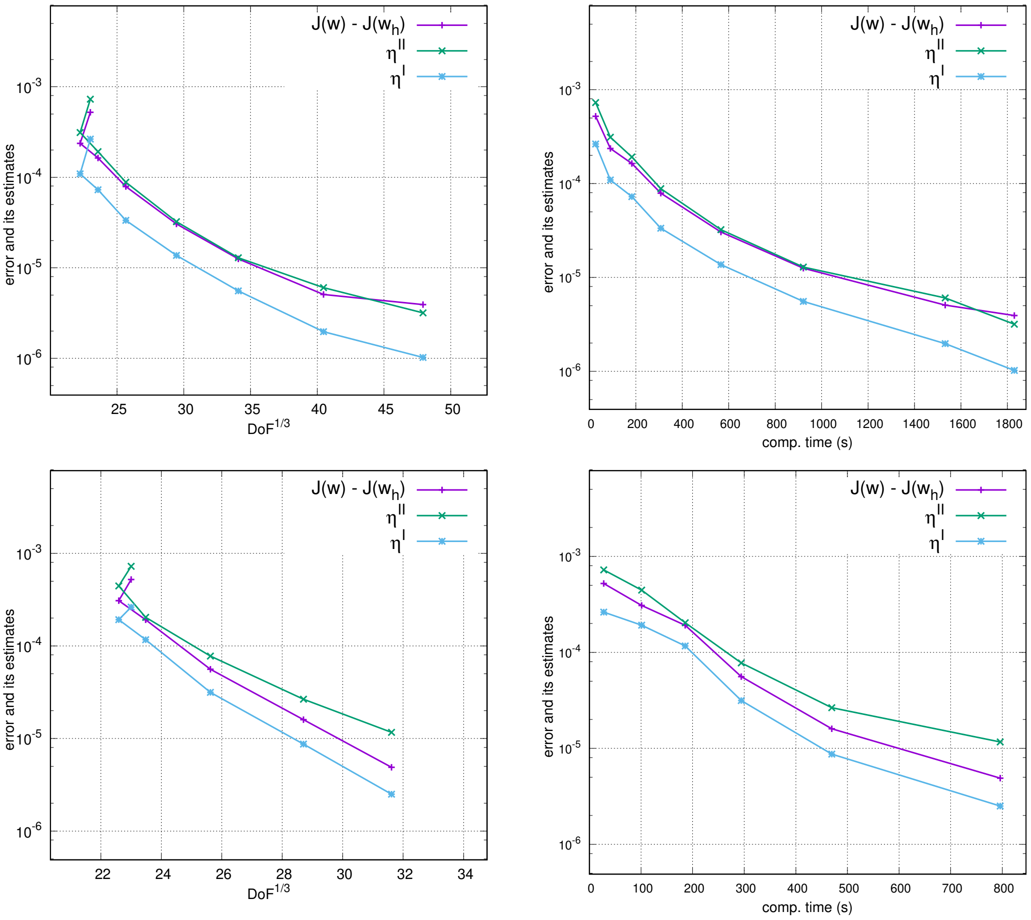

Figure 4.5: Subsonic inviscid flow around the NACA 0012 profile ($M = 0.5$, $\alpha = 0°$): decrease of the error $J(\boldsymbol{w}) - J(\boldsymbol{w}_h)$ and the goal-oriented error estimates η^{I} and η^{II} with respect to the cube root of DOF (left) and the computational time (right) for the h-refinement using $p = 2$ DG approximations (top) and the hp-version (bottom).

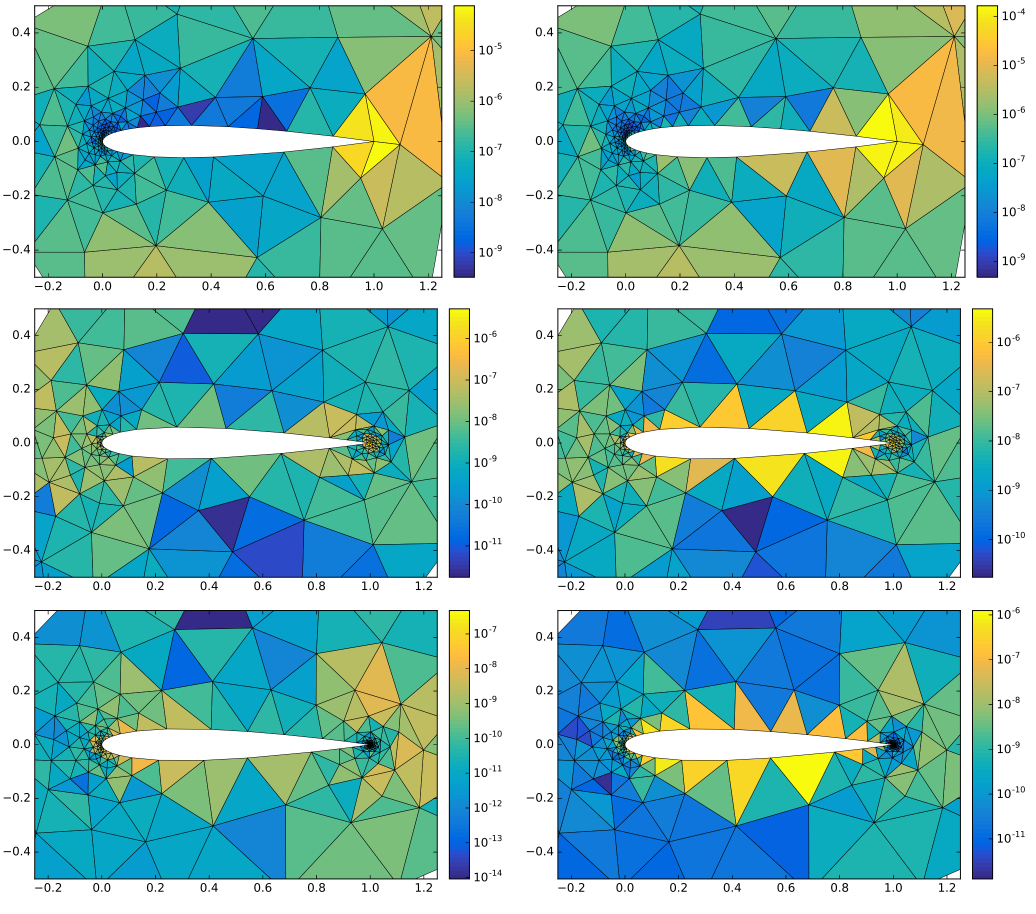

Figure 4.6: Subsonic inviscid flow around the NACA 0012 profile ($M = 0.5$, $\alpha = 0^\circ$): refinement indicators in log-scale for the hp-anisotropic adaptation method on the initial (top), 4th (middle) and 7th (bottom) mesh – η_K^{I} (left) and η_K^{II} (right).

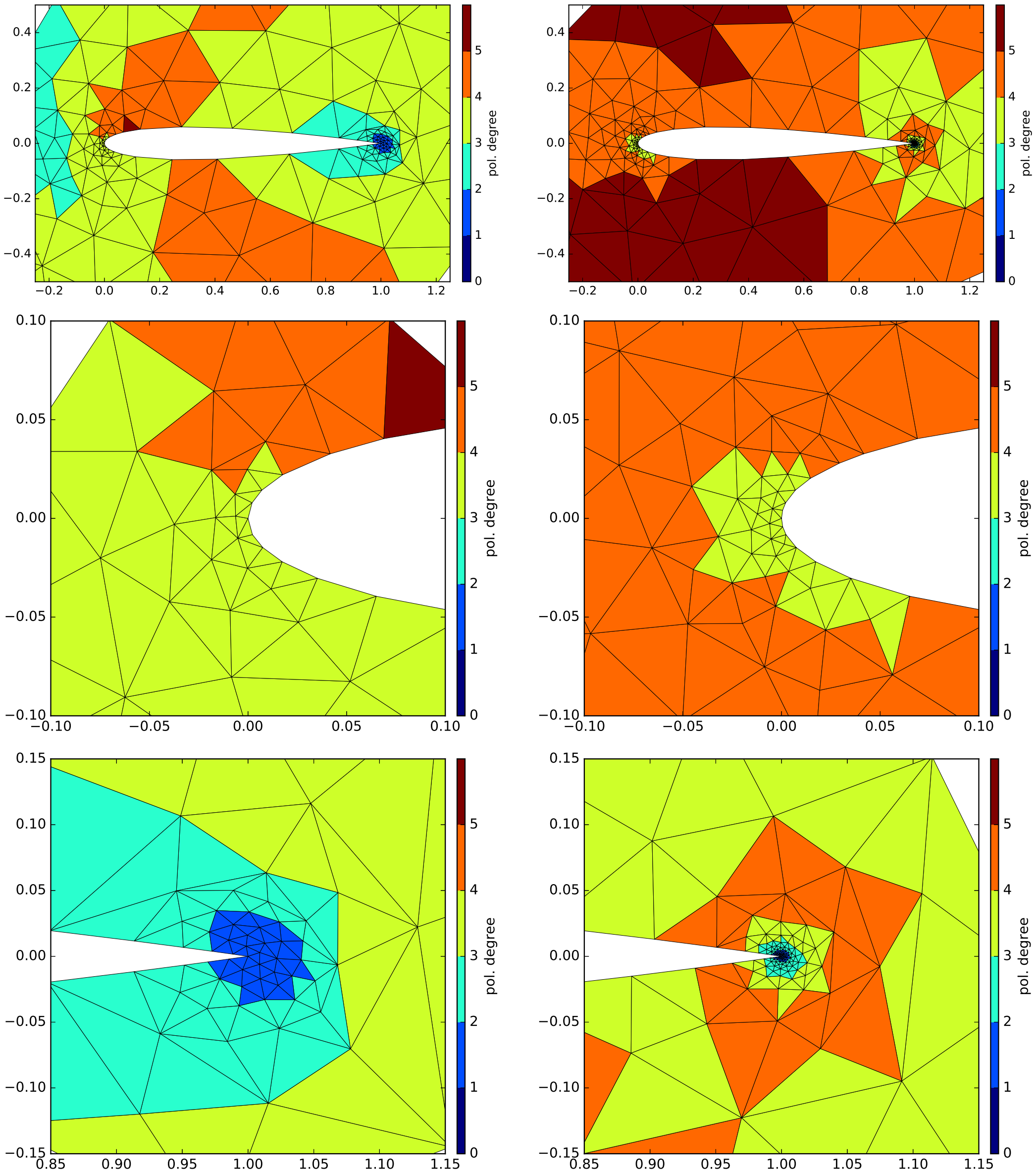

Figure 4.7: Subsonic inviscid flow around the NACA 0012 profile ($M = 0.5$, $\alpha = 0^\circ$): local polynomial degrees on the 4th (left) and 7th (right) mesh for the *hp*-anisotropic adaptation method, the whole profile (top) and zooms to the leading (middle) and trailing (bottom) edge of the profile.

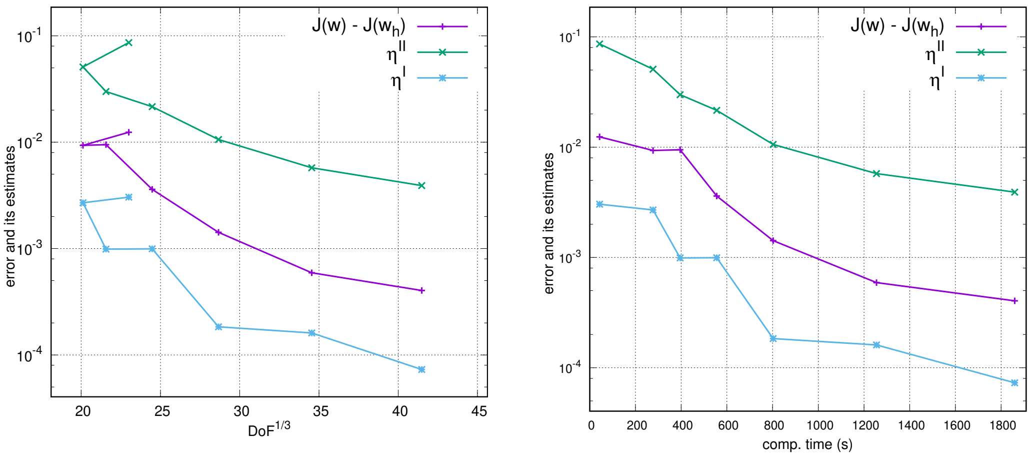

Figure 4.8: Subsonic inviscid flow around the NACA 0012 profile ($M = 0.5$, $\alpha = 1.25°$): decrease of the error $J(\boldsymbol{w}) - J(\boldsymbol{w}_h)$ and the goal-oriented error estimates η^{I} and η^{II} with respect to the cube root of DOF (left) and the computational time (right) for the piecewise quadratic DG approximations.

remains low which may be caused by lower regularity of the adjoint solution.

We remark that in the majority of articles on goal-oriented error estimates for Euler equations, e.g., Hartmann [2005, 2006, 2007], Hartmann and Leicht [2015], the numerical experiments are performed only for the drag coefficient. Actually, we have found only one experiment with the lift coefficient in Sharbatdar and Ollivier-Gooch [2018] for the transonic flow around the NACA 0012 profile.

4.5.3 Shock capturing

In higher-order numerical methods, applied to the problems with high speed flows with shock waves and contact discontinuities, we can observe the so-called Gibbs phenomenon which shapes spurious (nonphysical) oscillations in computed quantities propagating from discontinuities. These phenomena do not occur in low Mach number regimes, when the exact solution is regular (previous experiments), but in the higher-speed flow they cause instabilities in the numerical method.

Even though for DG methods this effect is not as dramatic as for standard FEM, spurious overshoots and undershoots may appear in the vicinity of discontinuities. In order to overcome this undesirable feature, we present a stabilization into the DG scheme based on adding artificial viscosity. Here, we present only the resulting terms needed for the implementation. We refer to Dolejší and Feistauer [2015], where the the stabilization method is derived in detail with notes and citations to other possible methods.

We define the jump indicator

$$g_K(\boldsymbol{w}_h) = \frac{\int_{\partial K\setminus\partial\Omega} [\![\boldsymbol{w}_h^1]\!]^2\,\mathrm{d}S}{|K|\sum_{\Gamma\subset\partial K\setminus\partial\Omega} h_\Gamma}, \qquad K\in\mathscr{T}_h, \tag{4.136}$$

where $\boldsymbol{w}_h^1$ is the first component (corresponding to density) of the vector function $\boldsymbol{w}_h$, $|K|$ denotes the d-dimensional measure of K and h_Γ is the diameter of Γ.

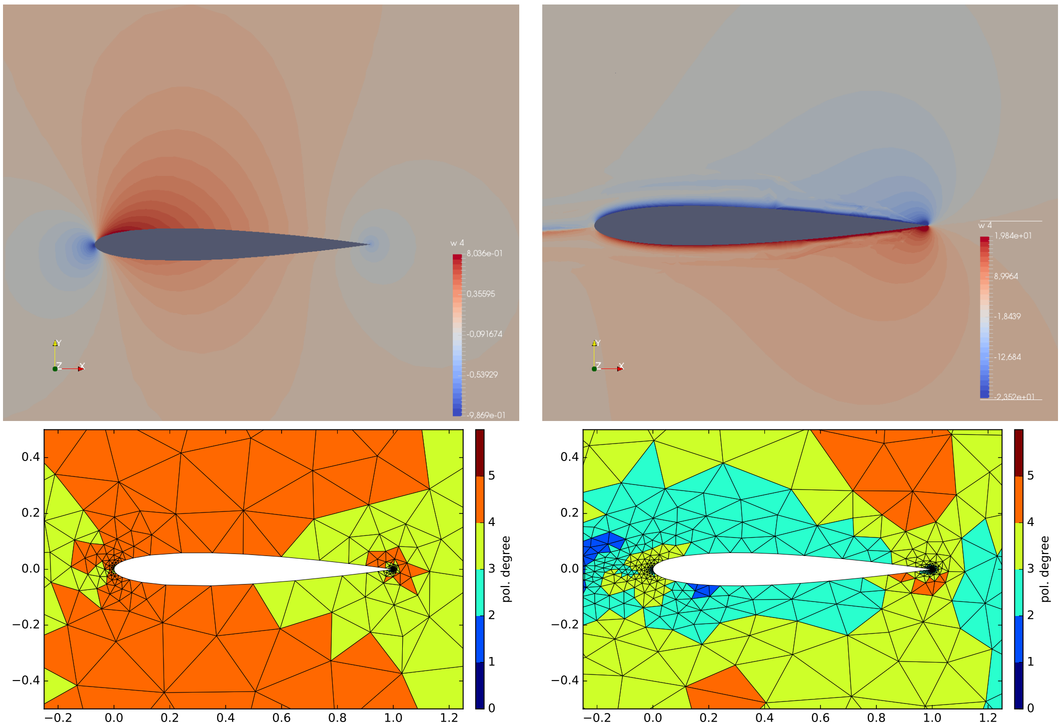

Figure 4.9: Subsonic inviscid flow around the NACA 0012 profile ($M = 0.5$, $\alpha = 1.25°$): the first component of the discrete adjoint solution on the final mesh for the drag target functional (left) and lift (right), bellow the corresponding hp-meshes are depicted.

Further, we set the smoothened discrete jump discontinuity indicator

$$G_K(\boldsymbol{w}_h) = \begin{cases} 0, & \text{if } g_K(\boldsymbol{w}_h) < \xi_{\min}, \\ \frac{1}{2}\sin\left(\pi\frac{g_K(\boldsymbol{w}_h)-(\xi_{\max}-\xi_{\min})}{2(\xi_{\max}-\xi_{\min})}\right)+\frac{1}{2}, & \text{if } g_K(\boldsymbol{w}_h) \in [\xi_{\min},\xi_{\max}), \\ 1, & \text{if } g_K(\boldsymbol{w}_h) \geq \xi_{\max}, \end{cases} \tag{4.137}$$

where we set $\xi_{\min} = 0.5$ and $\xi_{\max} = 1.5$. It is important to note that since we have $g_K(\boldsymbol{w}_h) = O(h^{2p})$ for $K \in \mathscr{T}_h$ where the solution is smooth and $g_K(\boldsymbol{w}_h) = O(h^{-2})$ near discontinuities, $g_K \to 0$ for $h \to 0$ in the case when $K \in \mathscr{T}_h$ is a element where the solution is regular. Hence also the indicators G_K vanish in regions where the solution is regular and the artificial viscosity occurs only locally in the vicinity of discontinuities.

Finally, we introduce the artificial viscosity forms $\boldsymbol{\beta}_h, \boldsymbol{\gamma}_h : \boldsymbol{S}_h^{\boldsymbol{p}} \times \boldsymbol{S}_h^{\boldsymbol{p}} \times \boldsymbol{S}_h^{\boldsymbol{p}} \to \mathbb{R}$, defined as

$$\boldsymbol{\beta}_h(\bar{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) = \nu_1 \sum_{K\in\mathscr{T}_h} G_K(\bar{\boldsymbol{w}}_h)\, h_K^{\alpha_1} \int_K \nabla \boldsymbol{w}_h \cdot \nabla \boldsymbol{\varphi}_h \,\mathrm{d}x, \tag{4.138}$$

$$\boldsymbol{\gamma}_h(\bar{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) = \nu_2 \sum_{\Gamma\in\mathscr{F}_h^I} \frac{1}{2}\left(G_{K_\Gamma^{(L)}}(\bar{\boldsymbol{w}}_h) + G_{K_\Gamma^{(R)}}(\bar{\boldsymbol{w}}_h)\right) h_\Gamma^{\alpha_2} \int_\Gamma [\![\boldsymbol{w}_h]\!] \cdot [\![\boldsymbol{\varphi}_h]\!] \,\mathrm{d}S, \tag{4.139}$$

where $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathscr{T}_h$ are the elements sharing the inner face $\Gamma \in \mathscr{F}_h^I$ and the parameters are set to $\alpha_1 = 2$, $\alpha_2 = 0.5$, $\nu_1 = \nu_2 = 1$. The form $\boldsymbol{\gamma}_h$ allows strengthening the influence of neighboring elements and improves the behavior of the method in the case, when strongly unstructured and/or anisotropic meshes are used.

The artificial viscosity forms $\boldsymbol{\beta}_h$ and $\boldsymbol{\gamma}_h$ are added to the left-hand side of (4.49). Then the DG method with shock capturing reads: find $\boldsymbol{w}_h \in \boldsymbol{S}_h^{\boldsymbol{p}}$: such that

$$a_h(\boldsymbol{w}_h, \boldsymbol{\varphi}_h) + \boldsymbol{\beta}_h(\boldsymbol{w}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) + \boldsymbol{\gamma}_h(\boldsymbol{w}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) = 0 \qquad \forall \boldsymbol{\varphi}_h \in \boldsymbol{S}_h^{\boldsymbol{p}}, \tag{4.140}$$

and it is solved using the Newton-like method as presented in Section 4.3.6 and the discrete adjoint problem is also defined using the linearization of this "augmented" form.

Transonic flow

We consider the far-field Mach number $M = 0.8$ and the angle of attack $\alpha = 1.25°$. This flow regime leads to two shock waves. The first one, lying on the upper side of the profile is stronger than the other one on the lower side. We set target functional $J(\boldsymbol{w})$ as the lift coefficient, see (4.26), (4.27). We consider the reference value $c_L^{\mathrm{ref}} = 3.39576E-01$ which was computed by the hp-anisotropic adaptation method on a mesh with approximately 10^5 degrees of freedom.

In Figure 4.10 the first component of the primal (left) and adjoint (right) discrete solution are depicted in the vicinity of the profile on the final mesh using the hp-anisotropic adaptation. Both shock waves in the primal solution are clearly distinguishable.

We performed several iterations of the adaptation algorithm for the fixed polynomial degree $p = 1$ and the for the hp-version of the method. We note that there is no point in setting globally $p > 1$ - due to the the low regularity of the exact solution.

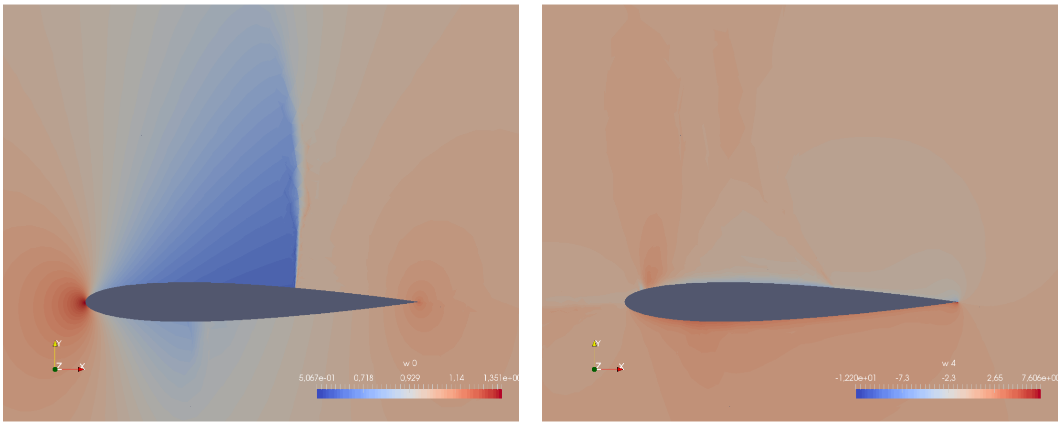

Figure 4.10: Transonic inviscid flow around the NACA 0012 profile ($M = 0.8$, $\alpha = 1.25°$): first component of the primal (left) and adjoint (right) solutions on the final mesh using the hp-anisotropic adaptation.

Indeed, Figure 4.13 shows the local polynomial degrees on the final mesh obtained by the hp-adaptation method and the minimal polynomial degree $p = 1$ is preserved in surrounding of both shock waves.

In Figure 4.11 the decrease of the error of the target quantity compared to the error estimates η^{I} and η^{II} is displayed. The estimate η^{I} follows quite accurate the true error, while the estimate η^{II} overestimates the error by a factor ≈ 10 in both cases. Figure 4.12 pictures the local error indicators η_K^{I} and η_K^{II} on the initial and final mesh. We see the final mesh is strongly refined in the surrounding of the upper shock wave while the refinement around the lower one is only slight. Similar performance can be observed in Figure 4.14 for the hp-adaptation.

We note that the same experiment was carried out in Sharbatdar and Ollivier-Gooch [2018] with similar results.

4.5.4 Turbine cascade

We consider the transonic inviscid compressible flow of air around the 2D turbine cascade SE1050, for cascade data see Šťastný and Šafařík [1992], Halama et al. [2011]. The computational domain Ω is by given a circular cut of a 3D turbine and it is periodically extended in the x_2 direction. We choose the Mach number $M = 0.32$ and the angle of attack $\alpha = 19.3°$. Obviously, the viscous and turbulence effects play an important role for this problem and therefore the use of inviscid model in not appropriate. However, we present this example in order to show abilities (and limits) of our technique to deal with more challenging test problems.

The quantity of interest is chosen as the lift coefficient c_L given by (4.26). We discretized the problem by the discontinuous Galerkin method with piecewise linear approximation ($p = 1$) augmented by the stabilization terms (4.138), (4.139) and we employed the $h-$ version of the goal-oriented anisotropic adaptation algorithm. Unfortunately, the stabilization terms works only partly and after to cycles of the $h-$ adaptation cycle the algebraic solver is not able to converge.

We note that is a problem of the DG discretization and not the error estimates by themselves. If we extend the problem (4.4) by the viscous terms, i.e. for the compress-

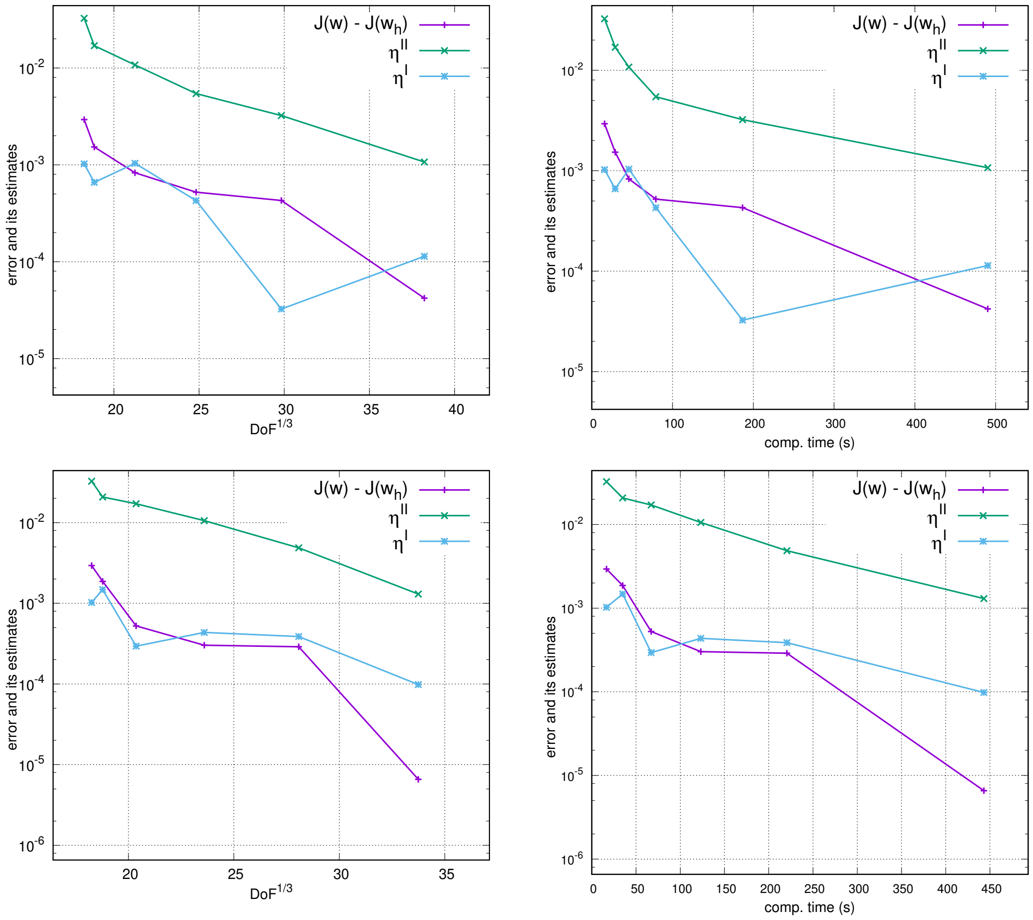

Figure 4.11: Transonic inviscid flow around the NACA 0012 profile ($M = 0.8$, $\alpha = 1.25°$): decrease of the error $J(\boldsymbol{w}) - J(\boldsymbol{w}_h)$ and the goal-oriented error estimates η^{I} and η^{II} for the lift coefficient with respect to the cube root of DOF (left) and the computational time (right) for the piecewise linear (top) and hp (bottom) DG approximations.

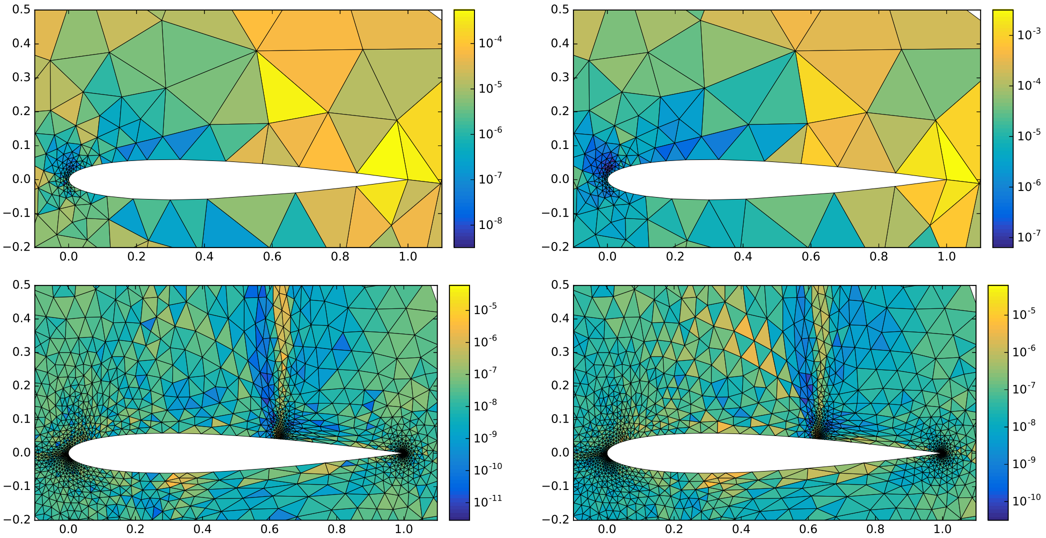

Figure 4.12: Transonic inviscid flow around the NACA 0012 profile ($M = 0.8$, $\alpha = 1.25°$): refinement indicators in log-scale for the $h-$anisotropic adaptation method ($p = 1$) on the initial (top) and final (bottom) mesh – η_K^{I} (left) and η_K^{II} (right).

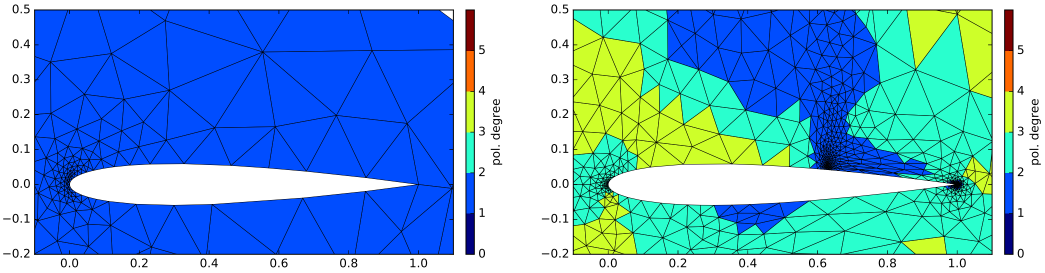

Figure 4.13: Transonic inviscid flow around the NACA 0012 profile ($M = 0.8$, $\alpha = 1.25°$): initial (left) and final (right) mesh using the hp-anisotropic refinement.

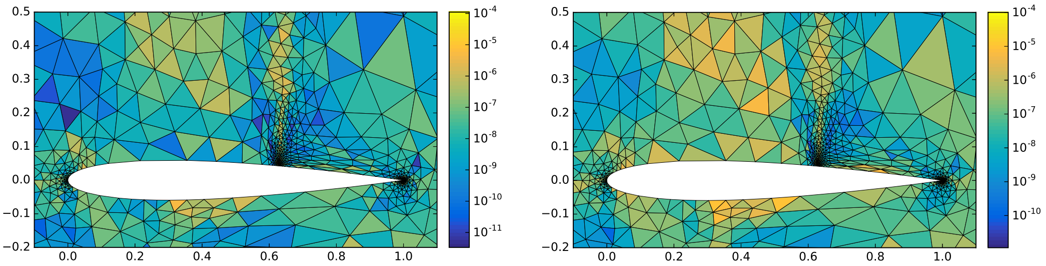

Figure 4.14: Transonic inviscid flow around the NACA 0012 profile ($M = 0.8$, $\alpha = 1.25°$): refinement indicators in log-scale for the hp-anisotropic adaptation method on the final mesh – η_K^{I} (left) and η_K^{II} (right).

ible Navier-Stokes equations, then the problem with the algebraic solver disappears and the DG method works even on finer meshes. The compressible Navier-Stokes equations are a natural target for further research in this fiels. We add that the goal-oriented error estimates were introduced for the compressible Navier-Stokes equations in e.g., Hartmann and Houston [2006b], Hartmann and Leicht [2015].

In Figure 4.15, the first component of the primal (left) and adjoint (right) solutions is depicted. Several shock wave may be observed on the right side of the profile in the primal solution as well as in the adjoint one. In Figure 4.16, the decrease of the error estimates η^{I} and η^{II} is plotted. We omit the true error $J(\boldsymbol{w}) - J(\boldsymbol{w}_h)$, since we were not able to compute the reference value of the drag.

Although the results of our computations are not very convincing, they show the potential of the goal-oriented estimates as the indicators driving the mesh adaptation seem to sensibly detect the areas where errors arise, see Figure 4.17.

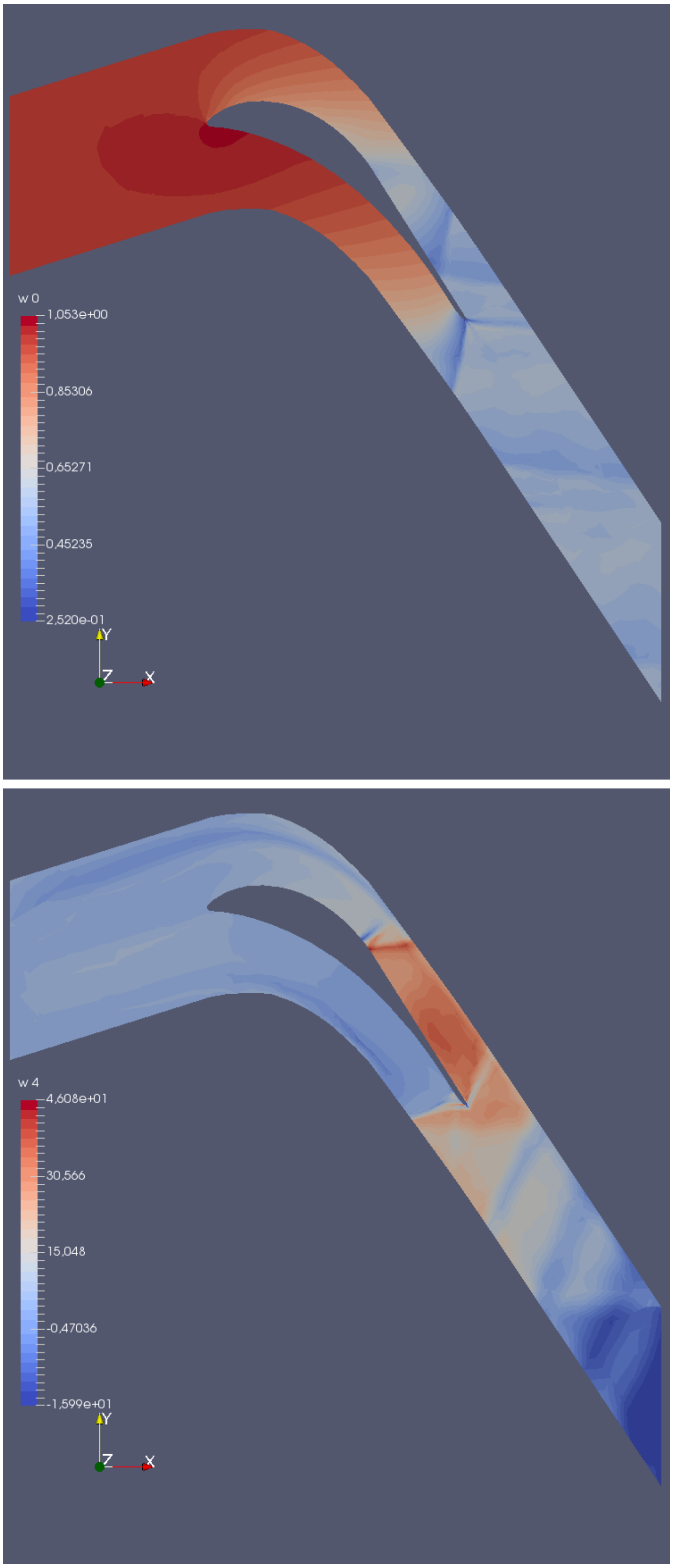

Figure 4.15: Transonic inviscid flow around the SE1050 profile ($M = 0.32$, $\alpha = 19.3°$): first component of the primal (left) and adjoint solution (right) on the final mesh.

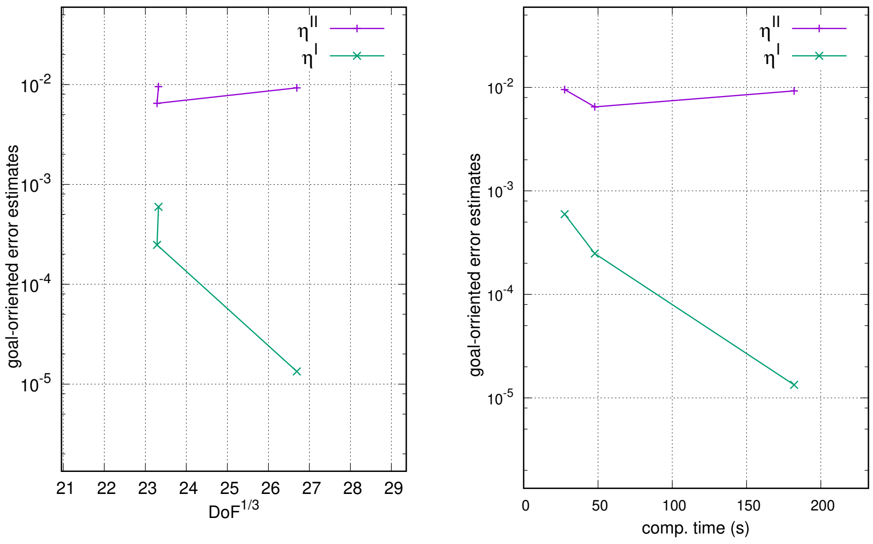

Figure 4.16: Transonic inviscid flow around the SE1050 profile ($M = 0.32$, $\alpha = 19.3°$): decrease of the goal-oriented error estimates η^{I} and η^{II} for the drag coefficient with respect to the cube root of DOF (left) and the computational time (right) for the piecewise linear DG approximations.

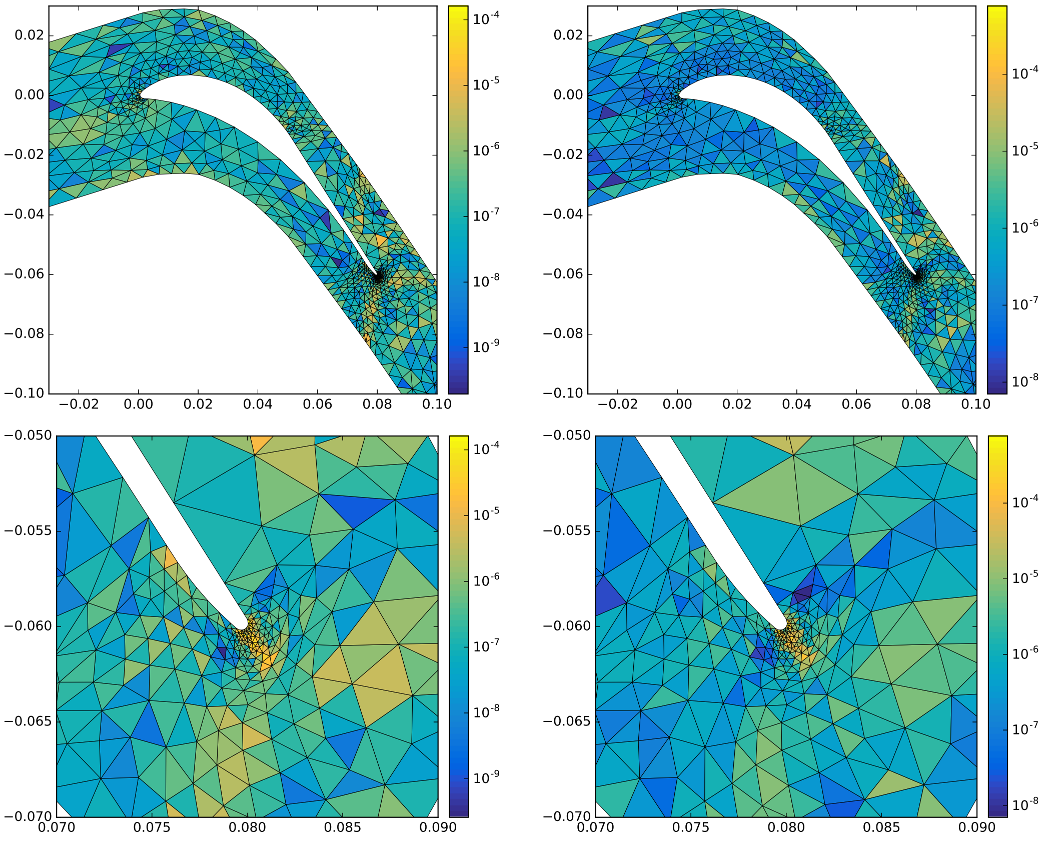

Figure 4.17: Transonic inviscid flow around the SE1050 profile ($M = 0.32$, $\alpha = 19.3°$): refinement indicators in log-scale for the $h-$ anisotropic adaptation method on the final mesh in the vicinity of the profile (top) and zoom on the trailing edge (bottom), η_K^{I} (left) and η_K^{II} (right).

Conclusion

In this thesis we have presented a complex strategy for estimating the computational errors with respect to some given quantity of interest for numerical solutions of partial differential equations.

First, the method was introduced generally and its pros and cons were presented and commented.

Further, the estimates were derived for the scalar linear convection-diffusion equation. We described an adjoint consistent discontinuous Galerkin discretization of the problem and we introduced goal-oriented estimates for both discretization and algebraic errors. Further, we described the influence of the algebraic errors on the estimates based on the primal and adjoint residual, respectively, and we introduced a stopping criterion keeping the algebraic errors controlled by the discretization estimate. In this way the algebraic system may be solved efficiently with satisfactory accuracy with respect to the quantity of interest.

Moreover, two kinds of local reconstructions of the DG solution were proposed. Our method suffers from the common deficiency of DWR approach – due to the approximation of the adjoint solution z we cannot provide guaranteed upper bound for the error of the quantity of interest. On the other hand, it provides results comparable to the approaches based on globally higher order solutions, but due to the local characteristics of the reconstructions it can be computed much faster and straightforwardly in parallel.

The main advantage of the presented strategy is its application for error indicators driving adaptive mesh refinement, where it provides very reliable results. We presented a *hp*-anisotropic adaptive mesh refinement strategy controlled by goal-oriented error indicators, which leads to a very efficient adaptive algorithm.

Finally, a more general approach for the goal-oriented error estimation for nonlinear methods was used for the Euler equations modeling inviscid compressible flows. We introduced the discontinuous Galerkin discretization and the linearized adjoint problem. A solution strategy was presented and finally we derived goal-oriented error estimates for the approximation of the drag, lift and momentum coefficients. These estimates were further modified in order to shape anisotropic error indicators enabling the *hp*-anisotropic adaptive mesh refinement.

Obviously, not all achieved results are satisfactory and they should be further developed. Namely,

- *upper bound of the goal-oriented error estimates* – one possible approach was mentioned in Section 1.1.7. However, application of this technique to discontinuous Galerkin method is not straightforward.

- *control of the algebraic error* – results presented in Section 1.1.8 show that it is difficult to attain an efficient computational process when the primal and adjoint problems are solved alternatively. A possible remedy is to solve the primal and adjoint problems simultaneously, e.g., using the *bi-conjugate gradient* (BiCG) method.

- *control of the algebraic error for nonlinear problems* – the technique mentioned above can be extended to nonlinear problems where the discretized systems are solved by a Newton(-like) method. Then the approximate adjoint solution would be available at each (linear was well as nonlinear) iteration and the linear and nonlinear algebraic errors could be controlled.

- *extension to the compressible Navier-Stokes equations* is a natural step, additionally an application to time-dependent problems is highly challenging.

Bibliography

J. C. Aguilar and J. B. Goodman. Anisotropic mesh refinement for finite element methods based on error reduction. *J. Comput. Appl. Math.*, 193(2):497 – 515, 2006.

M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite element analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2000. ISBN 0-471-29411-X.

M. Ainsworth and R. Rankin. Guaranteed computable bounds on quantities of interest in finite element computations. *International Journal for Numerical Methods in Engineering*, 89(13):1605–1634, 2012.

D. Ait-Ali-Yahia, G. Baruzzi, W. G. Habashi, M. Fortin, J. Dompierre, and M. Vallet. Anisotropic mesh adaptation: towards user-independent, mesh-independent and solver-independent CFD. II. Structured grids. *Internat. J. Numer. Methods Fluids*, 39:657–673, 2002.

F. Alauzet, W. Hassan, and M. Picasso. Goal oriented, anisotropic, a posteriori error estimates for the Laplace equation. In *Proc. of Enumath 2009 in Numerical Mathematics and Advanced Applications*, pages 47–58, Uppsala, Sweden, 2009.

M. Arioli, J. Liesen, A. Midlar, and Z. Strakoš. Interplay between discretization and algebraic computation in adaptive numerical solution of elliptic PDE problems. *GAMM-Mitteilungen*, 36(1):102–129, 2013. ISSN 1522-2608.

I. Babuska and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM Journal on Numerical Analysis*, 15(4):736–754, 1978.

I. Babuška and W. C. Rheinboldt. A posteriori error estimators for the finite element method. *Internat. J. Numer. Methods Engrg.*, 12:1597–1615, 1978.

I. Babuška, T. Strouboulis, and K. Copps. *hp* optimization of finite element approximations: Analysis of the optimal mesh sequences in one dimension. *Comput. Methods Appl. Mech. Engrg.*, 150(1-4):89–108, 1997.

I. Babuška and T. Strouboulis. *The finite element methods and its reliability*. Clarendon Press, Oxford, 2001.

A. Balan, M. Woopen, and G. May. Adjoint-based *hp*-adaptivity on anisotropic meshes for high-order compressible flow simulations. *Computers & Fluids*, 139:47 – 67, 2016. ISSN 0045-7930. 13th USNCCM International Symposium of High-Order Methods for Computational Fluid Dynamics - A special issue dedicated to the 60th birthday of Professor David Kopriva.

W. Bangerth and R. Rannacher. Finite element approximation of the acoustic wave equation: error control and mesh refinement. *East-West Journal of Numerical Mathematics*, 7:262–282, 1999.

W. Bangerth and R. Rannacher. *Adaptive Finite Element Methods for Differential Equations*. Lectures in Mathematics. ETH Zürich. 2003.

R. Bank and A. Weiser. Some a posteriori error estimators for elliptic partial differential equations. *Math. Comp.*, 44:283–301, 1985.

O. Bartoš, V. Dolejší, G. May, A. Rangarajan, and F. Roskovec. A goal-oriented anisotropic *hp*-mesh adaptation method for linear convection–diffusion–reaction problems. *Computers & Mathematics with Applications*, 2019. doi: https://doi.org/10.1016/j.camwa.2019.03.046. published online.

R. Becker and R. Rannacher. Weighted a posteriori error control in FE methods. 1995. Lecture at ENUMATH-95, Paris.

R. Becker and R. Rannacher. A feed-back approach to error control in finite element methods: Basic analysis and examples. *East-West J. Numer. Math.*, 4:237–264, 1996.

R. Becker and R. Rannacher. An optimal control approach to a-posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001.

R. Becker, H. Kapp, and R. Rannacher. Adaptive finite element methods for optimal control of partial differential equations: Basic concept. *SIAM J. Contr. Optim*, 39: 113–132, 1998.

A. Belme, A. Dervieux, and F. Alauzet. Time accurate anisotropic goal-oriented mesh adaptation for unsteady flows. *J. Comp. Phys.*, 231(19):6323–6348, 2012.

E. Bezchlebová, V. Dolejší, and M. Feistauer. Discontinuous Galerkin method for the solution of a transport level-set problem. *Computers & Mathematics with Applications*, 72(3):455–480, 2016.

S. Brenner and R. L. Scott. *The Mathematical Theory of Finite Element Methods*. Spriger, New York, 1994.

M. Breuss, V. Dolejší, and A. Meister. On an adaptive method for heat conduction problems with boundary layers. *ZAMM*, 2(6):450–463, 2006.

K. Bötcher and R. Rannacher. Adaptive error control in solving ordinary differential equations by the discontinuous Galerkin method. Technical report, 1996.

J. Carpio, J. Prieto, and R. Bermejo. Anisotropic "goal-oriented" mesh adaptivity for elliptic problems. *SIAM Journal on Scientific Computing*, 35(2):A861–A885, 2013.

A. Chaillou and M. Suri. A posteriori estimation of the linearization error for strongly monotone nonlinear operators. *J. Comput. Appl. Math.*, 205(1):72–87, 2007.

P. G. Ciarlet. *The Finite Elements Method for Elliptic Problems*. North-Holland, Amsterdam, New York, Oxford, 1979.

L. Demkowicz. *Computing with hp-adaptive finite elements. Vol. 1*. Applied Mathematics and Nonlinear Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2007.

L. Demkowicz, W. Rachowicz, and P. Devloo. A fully automatic *hp*-adaptivity. *J. Sci. Comput.*, 17(1-4):117–142, 2002.

P. Deuflhard. *Newton Methods for Nonlinear Problems*. Springer Series in Computational Mathematics, Vol. 35, 2004.

D. Di Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*. Mathematiques & applications. Springer Berlin Heidelberg, 2012. ISBN 9783642229800.

V. Dolejší. Anisotropic mesh adaptation for finite volume and finite element methods on triangular meshes. *Comput. Vis. Sci.*, 1(3):165–178, 1998.

V. Dolejší. *ANGENER – software package*. Charles University Prague, Faculty of Mathematics and Physics, 2000. www.karlin.mff.cuni.cz/ dolejsi/angen.html.

V. Dolejší. Anisotropic *hp*-adaptive method based on interpolation error estimates in the L^q-norm. *Appl. Numer. Math.*, 82:80–114, 2014.

V. Dolejší and J. Felcman. Anisotropic mesh adaptation for transonic and supersonic flow simulation. In A. Handličová, Z. Krivá, K. Mikola, and D. ševčovič, editors, *Algoritmy 2002,* 16^{th} *Conference on Scientific Computing*, pages 78–85. Slovak University of Technology, Bratislava, 2002.

V. Dolejší, G. May, F. Roskovec, and P. Solin. Anisotropic hp-mesh optimization technique based on the continuous mesh and error models. *Computers & Mathematics with Applications*, 74:45–63, 2017.

V. Dolejší. Anisotropic hp-adaptive method based on interpolation error estimates in the H1-seminorm. *Applications of Mathematics*, 60(6):597–616, Dec 2015.

V. Dolejší and F. Roskovec. Goal oriented a posteriori error estimates for the discontinuos Galerkin method. In *Proceedings of Seminar Programs and Algorithms of Numerical Mathematics 18*, 2016.

V. Dolejší and F. Roskovec. Goal-oriented error estimates including algebraic errors in discontinuous Galerkin discretizations of linear boundary value problems. *Applications of Mathematics*, 62(6):579–605, December 2017.

V. Dolejší and P. Tichý. On efficient numerical solution of linear algebraic systems arising in goal-oriented error estimates. *manuscript*, 2019.

V. Dolejší and M. Feistauer. *Discontinuous Galerkin Method – Analysis and Applications to Compressible Flow*. Springer Verlag, 2015.

V. Dolejší. *hp*-DGFEM for nonlinear convection-diffusion problems. *Mathematics and Computers in Simulation*, 87:87–118, 2013. ISSN 0378-4754.

V. Dolejší and P. Solin. *hp*-discontinuous Galerkin method based on local higher order reconstruction. *Appl. Math. Comput.*, 279:219–235, 2016.

V. Dolejší, G. May, and Rangarajan. A continuous hp-mesh model for adaptive discontinuous galerkin schemes. *Applied Numerical Mathematics*, 124:1–21, 2018.

V. Dolejší, G. May, A. Rangarajan, and F. Roskovec. A goal-oriented high-order anisotropic mesh adaptation using discontinuous galerkin method for linear convection-diffusion-reaction problems. *SIAM Journal on Scientific Computing*, 41 (3):A1899–A1922, 2019.

J. Dompierre, M.-G. Vallet, Y. Bourgault, M. Fortin, and W. G. Habashi. Anisotropic mesh adaptation: towards user-independent, mesh-independent and solver-independent CFD. Part III. unstructured meshes. *Int. J. Numer. Methods Fluids*, 39(8):675–702, 2002.

R. Eymard, T. Gallouët, and R. Herbin. *Solution of equations in R^n (Part 3). Techniques of scientific computing (Part 3).*, chapter Finite volume methods, pages 713–1020. Handbook of numerical analysis. Amsterdam: North-Holland/ Elsevier, 2000.

M. Feistauer. *Mathematical Methods in Fluid Dynamics*. Longman Scientific & Technical, Harlow, 1993.

M. Feistauer, J. Felcman, and I. Straškraba. *Mathematical and Computational Methods for Compressible Flow*. Oxford University Press, Oxford, 2003.

K. Fidkowski and D. Darmofal. Review of output-based error estimation and mesh adaptation in computational fluid dynamics. *AIAA Journal*, 49(4):673–694, 2011.

L. Formaggia and S. Perotto. Anisotropic error estimates for elliptic problems. *Numer. Math.*, 94(1):67–92, 2003.

L. Formaggia, S. Micheletti, and S. Perotto. Anisotropic mesh adaption in computational fluid dynamics: application to the advection-diffusion-reaction and the Stokes problems. *Appl. Numer. Math.*, 51(4):511–533, 2004.

P. J. Frey and F. Alauzet. Anisotropic mesh adaptation for CFD computations. *Comput. Methods Appl. Mech. Engrg.*, 194:5068–5082, 2005.

E. H. Georgoulis, E. Hall, and P. Houston. Discontinuous Galerkin methods on *hp*-anisotropic meshes I: A priori error analysis. *Int. J. Comput. Sci. Math*, 1(2-3): 221–244, 2007.

M. Giles and N. Pierce. Adjoint equations in CFD - Duality, boundary conditions and solution behaviour. In *13th Computational Fluid Dynamics Conference*. American Institute of Aeronautics and Astronautics, 1997.

M. Giles and E. Süli. Adjoint methods for PDEs: a posteriori error analysis and post-processing by duality. *Acta Numerica*, 11:145–236, 2002.

E. Godlewski and P. A. Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, volume 118 of *Applied Mathematical Sciences*. Springer, New York, 1996.

A. Greenbaum, V. Pták, and Z. Strakoš. Any nonincreasing convergence curve is possible for GMRES. *SIAM Journal on Matrix Analysis and Applications*, 17:465–469, 07 1996.

W. Gui and I. Babuška. The *hp* and *h-p* versions of the finite element method in 1 dimension. III. The adaptive *h-p* version. *Numer. Math.*, 49(6):659–683, 1986.

W. G. Habashi, J. Dompierre, Y. Bourgault, D. Ait-Ali-Yahia, M. Fortin, and M.-G. Vallet. Anisotropic mesh adaptation: towards user-independent, mesh-independent and solver-independent CFD. Part I: general principles. *Int. J. Numer. Methods Fluids*, 32(6):725–744, 2000.

J. Halama, F. Benkhaldoun, and J. Fořt. Flux schemes based finite volume method for internal transonic flow with condensation. *International Journal for Numerical Methods in Fluids*, 65:953 – 968, 2011.

K. Harriman, P. Houston, C. Schwab, and E. Süli. *hp*-version discontinuous Galerkin methods with interior penalty for partial differential equations with nonnegative characteristic form. *Recent Advances in Scientific Computing and Partial Differential Equations*, 330:89–119, 2003.

K. Harriman, D. J. Gavaghan, and E. Suli. The importance of adjoint consistency in the approximation of linear functionals using the discontinuous Galerkin finite element method. Technical report, Oxford University Computing Laboratory, 2004.

R. Hartmann and P. Houston. Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic conservation laws. *SIAM J. Sci. Comp.*, 24:979–1004, 2002.

R. Hartmann and P. Houston. Symmetric interior penalty DG methods for the compressible Navier-Stokes equations I: Method formulation. *Int. J. Numer. Anal. Model.*, 1:1–20, 2006a.

R. Hartmann and P. Houston. Symmetric interior penalty DG methods for the compressible Navier-Stokes equations II: Goal-oriented a posteriori error estimation. *Int. J. Numer. Anal. Model.*, 3:141–162, 2006b.

R. Hartmann. *The Role of the Jacobian in the Adaptive Discontinuous Galerkin Method for the Compressible Euler Equations*, pages 301–316. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

R. Hartmann. Derivation of an adjoint consistent discontinuous galerkin discretization of the compressible euler equations. In *G. Lube, G. Papin (Eds): International Conference on Boundary and Interior layers*, 2006.

R. Hartmann. Adjoint Consistency Analysis of Discontinuous Galerkin Discretizations. *SIAM Journal on Numerical Analysis*, 45(6):2671–2696, 2007.

R. Hartmann and T. Leicht. Generalized adjoint consistent treatment of wall boundary conditions for compressible flows. *Journal of Computational Physics*, 300:754–778, 2015.

V. Heuveline and R. Rannacher. A posteriori error control for finite element approximations of elliptic eigenvalue problems. *Advances in Computational Mathematics*, 15(1-4):107–138, 2001.

P. Houston, C. Schwab, and E. Süli. Discontinuous *hp*-finite element method for advection–diffusion problems. Technical Report Research Report No. 2000–07, SAM ETH Zürich, 2000.

P. Houston, C. Schwab, and E. Süli. Discontinuous *hp*-finite element methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 39(6):2133–2163, 2002.

H. T. Huynh. A reconstruction approach to high-order schemes including discontinuous galerkin for diffusion. In *47th AIAA Aerospace Sciences Meeting*, pages AIAA 209–403, 2009.

G. Kanschat and R. Rannacher. Local error analysis of the interior penalty discontinuous galerkin method for second order elliptic problems. *Journal of Numerical Mathematics*, 10(4):249–274, 2002.

P. Ladeveze, F. Pled, and L. Chamoin. New bounding techniques for goal-oriented error estimation applied to linear problems. *International Journal for Numerical Methods in Engineering*, 93:1–36, 2013.

R. LeVeque. *Numerical methods for conservation laws. Lectures in Mathematics ETH Zürich.* Birkhäuser Verlag, Basel, 1990.

R. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.

A. Loseille, A. Dervieux, and F. Alauzet. Fully anisotropic goal-oriented mesh adaptation for 3D steady Euler equations. *J. Comput. Phys.*, 229(8):2866–2897, 2010.

J. Lu. *An a posteriori control framework for adaptive precision optimization using discontinuous Galerkin finite element method*. PhD thesis, M.I.T., 2005.

D. Meidner, R. Rannacher, and J. Vihharev. Goal-oriented error control of the iterative solution of finite element equations. *Journal of Numerical Mathematics*, 17(2):143–172, 2009.

W. F. Mitchell and M. A. McClain. A comparison of hp -adaptive strategies for elliptic partial differential equations. *ACM Transactions on Mathematical Software*, 41: 1–39, 10 2014.

I. Mozolevski and S. Prudhomme. Goal-oriented error estimation based on equilibrated-flux reconstruction for finite element approximations of elliptic problems. *Computer Methods in Applied Mechanics and Engineering*, 288:127–145, 2015.

P. Neittaanmäki and S. Repin. *Reliable methods for computer simulation*, volume 33 of *Studies in Mathematics and its Applications*. Elsevier Science B.V., Amsterdam, 2004. ISBN 0-444-51376-0.

R. H. Nochetto, A. Veeser, and M. Verani. A safeguarded dual weighted residual method. *IMA Journal of Numerical Analysis*, 29(1):126–140, 2009.

M. Park, J. Krakos, T. Michal, A. Loseille, and J. Alonso. Unstructured grid adaptation: Status, potential impacts, and recommended investments toward CFD vision 2030. In *46th AIAA Fluid Dynamics Conference*, 06 2016.

R. Rannacher and F.-T. Suttmeier. A feed-back approach to error control in finite element methods: application to linear elasticity. *Computational Mechanics*, 19(5): 434–446, 1997.

R. Rannacher and J. Vihharev. Adaptive finite element analysis of nonlinear problems: balancing of discretization and iteration errors. *Journal of Numerical Mathematics*, 21:23–47, 2013.

T. Richter and T. Wick. Variational localizations of the dual weighted residual estimator. *Journal of Computational and Applied Mathematics*, 279:192–208, 2015.

C. Schwab. *p- and hp-finite element methods: Theory and applications in solid and fluid mechanics*. Numerical Mathematics and Scientific Computation. Clarendon Press, Oxford, 1998.

M. Sharbatdar and C. Ollivier-Gooch. Mesh adaptation using c1 interpolation of the solution in an unstructured finite volume solver. *International Journal for Numerical Methods in Fluids*, 86(10):637–654, 2018.

R. B. Simpson. Anisotropic mesh transformations and optimal error control. *Applied Numer. Math.*, 14:183–198, 1994.

M. Štastný and P. Šafařík. Boundary layer effects on the transonic flow in a straight turbine cascade. Technical report, ASME Paper 92-GT-155, New York, 1992.

D. Venditti and D. Darmofal. Grid adaptation for functional outputs: Application to two-dimensional inviscid flows. *J. Comput. Phys.*, 176(1):40–69, 2002.

R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques.* Wiley-Teubner Series Advances in Numerical Mathematics. Chichester: John Wiley & Sons. Stuttgart, 1996.

R. Verfürth. *A Posteriori Error Estimation Techniques for Finite Element Methods*. Numerical Mathematics and Scientific Computation. Oxford University Press, 2013.

G. Vijayasundaram. Transonic flow simulation using upstream centered scheme of Godunov type in finite elements. *J. Comput. Phys.*, 63:416–433, 1986.

M. Vohralík. *A posteriori error estimates, stopping criteria and inexpensive implementation*. Habilitation thesis, Université Pierre et Marie Curie – Paris 6, 2010.

P. Šolín and L. Demkowicz. Goal-oriented *hp*-adaptivity for elliptic problems. *Computer Methods in Applied Mechanics and Engineering*, 193(6–8):449–468, 2004.

Z. J. Wang, K. Fidkowski, R. Abgrall, F. Bassi, D. Caraeni, A. Cary, H. Deconinck, R. Hartmann, K. Hillewaert, H. T. Huynh, N. Kroll, G. May, P.-O. Persson, B. van Leer, and M. Visbal. High-order CFD methods: current status and perspective. *Int. J. Numer. Meth. Fluids*, 72:811–845, 2013.

O. C. Zienkiewicz and J. Z. Zhu. The superconvergent patch recovery and a-posteriori error estimates. Part 1: The recovery technique. *Internat. J. Numer. Methods Engrg.*, 33:1331–1364, 1992a.

O. C. Zienkiewicz and J. Z. Zhu. The superconvergent patch recovery and a-posteriori error estimates. Part 2: Error estimated and adaptivity. *Internat. J. Numer. Methods Engrg.*, 33:1365–1382, 1992b.

List of Figures

List of Tables

List of Abbreviations

DG	discontinuous Galerkin (method)
DWR	dual weighted residual (method)
FEM	finite element method
FV	finite volumes (method)
PDE	partial differential equation

List of publications

Journals

- Vít Dolejší, Filip Roskovec, and Miloslav Vlasák. Residual based error estimates for the space–time discontinuous Galerkin method applied to the compressible flows. *Computers & Fluids*, 117:304–324, 2015.
- Miloslav Feistauer, Filip Roskovec, and Anna-Margarete Sändig. Discontinuous Galerkin method for an elliptic problem with nonlinear Newton boundary conditions in a polygon. *IMA Journal of Numerical Analysis*, 39(1):423–453, 2017.
- Vít Dolejší, Georg May, Filip Roskovec, and Pavel Solin. Anisotropic hp-mesh optimization technique based on the continuous mesh and error models. *Computers & Mathematics with Applications*, 74: 45–63, 2017.
- Vít Dolejší and Filip Roskovec. Goal-oriented error estimates including algebraic errors in discontinuous Galerkin discretizations of linear boundary value problems. *Applications of Mathematics*, 62(6):579–605, 2017.
- Vít Dolejší, Georg May, Ajay Rangarajan, and Filip Roskovec. A Goal-oriented high-order anisotropic mesh adaptation discontionuous Galerkin method for linear convection-diffusion-reaction problems. *SIAM Journal on Scientific Computing*, 41(3):A1899–A1922, 2019.
- Ondřej Bartoš, Vít Dolejší, Georg May, Ajay Rangarajan, and Filip Roskovec. A goal-oriented anisotropic *hp*-mesh adaptation method for linear convection-diffusion-reaction problems. *Computers & Mathematics with Applications*, published online, 2019. doi: https://doi.org/10.1016/j.camwa.2019.03.046.
- Ondřej Bartoš, Miloslav Feistauer, and Filip Roskovec. On the effect of numerical integration in the finite element solution of an elliptic problem with a nonlinear Newton boundary condition. *Applications of Mathematics*, 64: 129–167, 2019.

Conference proceedings

- Vít Dolejší, Filip Roskovec, and Miloslav Vlasák. A posteriori error estimates for nonstationary problems. In *Numerical Mathematics and Advanced Applications ENUMATH 2015*, pages 225–233, 2016. Springer International Publishing.
- Vít Dolejší and Filip Roskovec. Goal oriented a posteriori error estimates for the discontinuos Galerkin method. In *Proceedings of Seminar Programs and Algorithms of Numerical Mathematics 18*, 2016.
- Vít Dolejší, Filip Roskovec, and Miloslav Vlasák. On a posteriori error estimates for space–time discontinuous Galerkin method, In *Proceedings of the conference Algoritmy 2016*, pages 125–134, 2016.
- Vít Dolejší, Filip Roskovec. Residual based error estimates for the space-time discontinuous Galerkin method applied to nonlinear hyperbolic equations, In *Proceedings of the conference Algoritmy 2016*, pages 113–124, 2016.