Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

HABILITAČNÍ PRÁCE

2003

RNDr. Vít Dolejší, Ph.D.

ADAPTIVE HIGHER ORDER METHODS FOR COMPRESSIBLE FLOW

RNDr. Vít DOLEJŠÍ, Ph.D.

Preface

The field of computational fluid dynamics (CFD) has already had a significant impact on science and engineering of fluid dynamics, ranging from a role of aircraft design to simulation of a spread pollution in the environment. The ultimate goal of the field of CFD is to understand the physical events that occur in the flow of fluids around and within designed objects. These events are related to the action and reaction of phenomena such as dissipation, diffusion, convection, shock waves, boundary layers and turbulence.

In the field of aerodynamics, all of these phenomena are governed by the compressible Navier-Stokes equations. Many of the most important aspects of these relations are nonlinear and, as a consequence, often have no analytical solution. This, of course, motivates the numerical solution of the associated partial differential equations.

The numerical solutions of the Navier-Stokes equations frequently exhibit localised structures, such as propagating discontinuities and sharp transition layers whose reliable numerical approximation presents a challenging computational task. Indeed, in order to resolve such localised phenomena in an accurate and efficient way one has to use locally refined (adapted) computational meshes.

Therefore, an efficient tool for the numerical simulation of compressible flows has to contain a sufficiently robust and accurate solver for the system of governing equations as well as a suitable adaptive method. This work deals with both mentioned aspects of the numerical solution of the Navier-Stokes equations:

1. adaptation of meshes – Chapter 1,
2. solver for the Navier-Stokes equations – Chapter 2.

Ad 1. In Chapter 1 we deal with the use of the *anisotropic mesh adaptation* (AMA) method for the solution of boundary value problems. AMA is a universal problem independent method whose use is not restricted only to fluid dynamics problems. On the basis of the control of the interpolation error, we derive a suitable mesh optimization criterion. The efficiency of the method is demonstrated on the case of a scalar equation and we compare AMA with other adaptive methods. Moreover, we introduce the use of AMA for the simulation of compressible flow. The universality of the method is displayed by an example following from the medicine.

Ad 2. In Chapter 2 we deal with the use of the *discontinuous Galerkin method* (DGM) for the solution of convective–diffusion problems. For the case of scalar equation, we present two versions of DGM and we derive a priori error estimates for both cases. Further, we extend DGM for the system of the compressible Navier-Stokes equation and present several test examples.

The subject matter of this work was included in several projects which were solved at the Charles University Prague, Faculty of Mathematics and Physics, Department of Numerical Mathematics. These projects were supported under the grants No. 201/96/0313, 201/99/0267, 201/02/0684 and 201/00/D116 of the Czech Grant Agency, the grants No. 275/2001/B–MAT/MFF and 3/1998/B MAT/MFF of the Grant Agency of the Charles University Prague and the grant No. MSM 113200007 of the Ministry of Education of the Czech Republic. The presented results are partialy new and partialy were published in scientific journals:

AMA : [Dol98a], [DF], [Dol01], [DF02a],

DGM : [DFS02a], [DFS01], [DFS02b], [Doled], [DFS].

Both chapters are independent, each one has own introduction and conclusion but the bibliography is common. The notation is uniform only within each chapter, some objects are denoted by different symbols (e.g. a triangle or a tetrahedra is denoted by T in Chapter 1 whereas by K in Chapter 2).

Contents

Chapter 1

Anisotropic mesh adaptation method

1.1 Introduction

Numerical solution of complicated problems of computational physics and mathematical engineering gives rise to the necessity of application of adaptive methods, which allow computation of a sufficiently precise solution without enormous requirements for memory and CPU time. The aim of the adaptive methods is to find a suitable discretization of computational domain where sufficiently accurate numerical solution can be computed and moreover the number of unknowns is reasonable. There exists a very large range of different adaptive methods developed during several last decades. Many adaptive methods are based on a priori or a posteriori error analysis. Taking into account the error estimations we define a suitable error indicator. Using the values of numerical solution calculated on the given mesh we compute the values of error indicator for all elements. Then elements, where the value of the indicator is higher than a given tolerance, are refined and we obtain a new grid. This approach is usual for finite element methods. The list of authors dealing with such type of adaptive technique is very large, we cite only the review [Ver96]. Alternative approach is the use of the dual problem, see the review article [Ran98].

Rather different approach is the *anisotropic mesh adaptation* (AMA). It is based on the control of the interpolation error of a piecewise polynomial interpolation of the exact solution on the triangular grid. The mesh adaptation criterion checks the behaviour of the second order derivatives of the solution of the considered problem.

The advantage of AMA is that it can be used without any modification for arbitrary boundary value problem and arbitrary numerical method (finite element method, finite volume method, discontinuous Galerkin method, etc.). The disadvantage of this approach is that we have no control of the discretization error of the numerical solution in fact.

In [DS89], [DS91], [Sim94], it is discussed how these mechanisms arise in the theory of optimal error control, using simple model mesh generation problems. A general recovery technique was developed for determining the derivatives of the finite element solutions in [ZZ92a] and a posteriori error estimation was discussed in [ZZ92b]. The application of AMA for computational fluid dynamics has been performed in INRIA group (see, e.g., [CDBG$^+$96], [DHM95], [BHF97]) and Habashi's group (see, e.g., [FVD$^+$96], [HFaY$^+$96]). In monograph [Ape99], there were presented the anisotropic local interpolation error estimates for several types of two- and three-dimensional finite elements and a priori estimates of the discretization error for model problems with edge singularities or boundary layers.

Our presented mesh adaptation technique for unstructured grids is based on such type of adaptive methods. The original approach of our algorithm is the use of the *edge-based* optimization criterion. In this work we summarize and develop the first formulation of AMA and its use for the numerical solution of compressible flow and heat conduction problem, see [Dol98a], [DF], [Dol01], [BDM] and [DF02a].

Although we try to give a mathematically rigorous explanation, we are not able to avoid several rather heuristic steps. On the other hand the presented method is a very efficient tool for the numerical solution of various practical problems. The presented mathematical treatment of the anisotropic mesh adaptation is almost identical for two and three dimensional problems. Therefore it is carried out for dimension d, $d = 2, 3$. The space discretization is called *mesh*, which is a triangulation ($d = 2$) or a tetrahedrization ($d = 3$), and *element* is a triangle ($d = 2$) or tetrahedron ($d = 3$).

The summary of this chapter is the following. At the beginning we consider a *scalar boundary value problem* (BVP) and suppose that its exact solution is known. In Section 1.2 we derive a necessary condition which guarantees that the discretization error is under a given tolerance. As the necessary condition is not suitable for practical computation, we introduce its stronger (but simpler) version. Moreover, using several heuristic considerations, we define the concept of an *optimal mesh*. Finally, we introduce the *quality parameter* which measures the quality of a mesh.

In Section 1.3 we describe the AMA algorithm which creates the optimal mesh for a known exact solution. Section 1.4 shows how to use the AMA algorithm for practical computation, where the exact solution is not known. Several test examples are given in Section 1.5, where the efficiency of AMA is demonstrated. Section 1.6 and 1.7 show the practical application of AMA to the numerical simulation of inviscid as well as viscous flow and heat conduction problems, respectively. For the clarity of the treatment, some (simple but long) proofs of Lemmas are given in Appendix.

1.2 Interpolation error control

1.2.1 Necessary condition

Let us consider a scalar boundary value problem (BVP) in the set Ω, where $\Omega \subset I\!R^d$ is a bounded domain. By $\overline{\Omega}$ and $\partial\Omega$ we denote the closure and boundary of Ω, respectively. We introduce the weak formulation of (BVP): Find the function $u : \Omega \to I\!R$ satisfying

$$u \in U, \tag{1.2.1}$$

$$\mathcal{L}(u, v) = f(v) \quad \forall v \in V, \tag{1.2.2}$$

$$\mathcal{B}(u) \in V. \tag{1.2.3}$$

Here U and V are the trial and test function spaces, respectively, (1.2.2) represents the weak form of the partial differential equation under consideration and (1.2.3) guarantees that the prescribed boundary conditions are fulfilled in the sense of traces. The function u satisfying (1.2.1) – (1.2.3) is called the *exact solution*. Let us suppose that the problem (1.2.1) – (1.2.3) has a unique exact solution.

Very often the numerical solution of (1.2.1) – (1.2.3) is sought in the space of piecewise polynomial functions and computed by a suitable numerical method (e.g. finite element method, finite volume method, discontinuous Galerkin method). To this end the domain Ω is replaced by its polygonal approximation Ω_h and the mesh $\mathcal{T}_h$ is constructed. The properties of $\mathcal{T}_h$ are presented in Definition 1.1. The *approximate solution* of problem (1.2.1) – (1.2.3) is defined as a function $u_h : \overline{\Omega}_h \to I\!R$ satisfying

$$u_h \in U_h, \tag{1.2.4}$$

$$\mathcal{L}_h(u_h, v_h) = f_h(v_h) \quad \forall v_h \in V_h, \tag{1.2.5}$$

$$\mathcal{B}_h(u_h) \in V_h. \tag{1.2.6}$$

Here

$$U_h = \left\{ u_h \in L^1(\Omega_h); u_h|_T \in P_k(T) \ \forall T \in \mathcal{T}_h \right\}, \tag{1.2.7}$$

and $P_k(T)$ denotes the set of all polynomials of degree $\leq k$ ($k \geq 0$) on the element T and $\mathcal{L}_h,\ f_h,\ \mathcal{B}_h,\ V_h$ are the discrete analogue of $\mathcal{L},\ f,\ \mathcal{B},\ V$, respectively. Both possibilities $U_h \subset U$ and $U_h \not\subset U$ are admissible. Let us suppose, that the *discrete problem* (1.2.4) – (1.2.6) has a unique solution u_h.

In order to measure how close the approximate solution u_h is to the exact solution u, we define the *discretization error* as

$$e_h \equiv \|u - u_h\|_X, \tag{1.2.8}$$

where $\|\cdot\|_X$ is a suitable norm such that the expression $\|v\|_X$ has a sense $\forall v \in U \cup U_h$. (If $U_h \subset U$ we can put $\|\cdot\|_X \equiv \|\cdot\|_U$, where $\|\cdot\|_U$ is the norm in the space U.)

Our aim is, for a given tolerance $\omega > 0$, to compute the approximate solution satisfying

$$e_h \leq \omega. \tag{1.2.9}$$

In what follows, we present the necessary condition for the relation (1.2.9) and show with numerical examples how its satisfaction can be used for the construction of adaptive meshes.

We denote by $\mathcal{O}(U, U_h)$ the set of all operators from U to U_h, i.e.

$$\mathcal{O}(U, U_h) = \{\Pi, \Pi : w \in U \mapsto \Pi w \in U_h\}. \tag{1.2.10}$$

Let us consider the minimization problem: Find $\Pi_h \in \mathcal{O}(U, U_h)$ such that

$$\|w - \Pi_h w\|_X = \min_{w_h \in U_h} \|w - w_h\|_X \quad \forall w \in U. \tag{1.2.11}$$

The existence of Π_h follows from the fact that U_h is a finite dimensional space. The operator Π_h may not be uniquely determined, its uniqueness depends on the choice of $\|\cdot\|_X$.

It is evident that for the approximate solution u_h of (1.2.4) – (1.2.6)

$$\|u - \Pi_h u\|_X \leq e_h. \tag{1.2.12}$$

This is the crucial point of the proposed adaptation strategy: Any numerical method for the solution of (1.2.1) – (1.2.3) computes the approximate solution u_h with the discretization error which is bounded from below by $\|u - \Pi_h u\|_X$, where u is the exact solution of (1.2.1) – (1.2.3). Then the *necessary condition* to fulfill (1.2.9) is

$$\|u - \Pi_h u\|_X \leq \omega. \tag{1.2.13}$$

For a given ω, the condition (1.2.13) is satisfied if the mesh is sufficiently fine. This can be achieved

(i) by a global mesh refinement, which is rather inefficient from the point of view of computer memory and CPU time in further computations,

(ii) by a local grid refinement/coarsening/alignment strategy, which is the subject–matter of Chapter 1.

It is evident that the magnitude of $\|u - \Pi_h u\|_X$ strongly depends on the choice of U_h and therefore on the choice of $\mathcal{T}_h$. In the first part of this chapter, for a given $\omega > 0$ and given function $u \in C^2(\Omega)$, we describe the construction of the mesh $\mathcal{T}_h$ for which the function $\Pi_h u$ satisfies (1.2.13) and the mesh has as small a number of elements as possible. In the second part, we apply these results to the numerical solution of physically relevant problems.

The basic principle of AMA can be illustrated on the following 1D example. Let $\langle A, B\rangle \in I\!R^1$ be an interval and $u \in C^2(\langle A, B\rangle)$ the given function, see Figure 1.1. We introduce two partitions $\mathcal{T}_{h_1}$ and $\mathcal{T}_{h_2}$ of the interval $\langle A, B\rangle$. The former partition (having 5 intervals) is uniform whereas the latter (having 3 intervals) is anisotropic. Let us consider a discontinuous piecewice linear approximation. Let $\Pi_{h_1} u$ and $\Pi_{h_2} u$ be the functions given by (1.2.11) on the partitions $\mathcal{T}_{h_1}$ and $\mathcal{T}_{h_2}$, respectively, where U_h is given by (1.2.15) and we set $\|\cdot\|_X = \|\cdot\|_{L^\infty(\langle A,B\rangle)}$. Moreover we put

$$\|u - \Pi_{h_1} u\|_X = \omega_1, \qquad \|u - \Pi_{h_2} u\|_X = \omega_2, \tag{1.2.14}$$

where the values $\max_{x\in\langle A,B\rangle} |u(x) - \Pi_{h_i} u(x)|,\ i = 1, 2$ are plotted on the Figure 1.1.

We observe that $\omega_2 < \omega_1$ and moreover $\#\mathcal{T}_{h_2} < \#\mathcal{T}_{h_1}$, where $\#\mathcal{T}_h$ denotes the number of subintervals of partition $\mathcal{T}_h$. Therefore, the anisotropic partition $\mathcal{T}_{h_2}$ is more suitable for the piecewise linear capturing of u.

In practical problems, of course, the exact solution is not a priori known. In that case we compute the approximate solution u_h on the starting mesh $\mathcal{T}_h$, we smooth u_h and denote this smoothing as $\tilde{u}$ and apply the AMA algorithm for $\tilde{u}$, see Section 1.4.

1.2.2 ω-minimal mesh

From the practical point of view, we shall use in (1.2.7) the discontinuos piecewise linear polynomials, i.e.

$$U_h = \left\{ u_h \in L^1(\Omega_h); u_h|_T \in P_1(T)\ \forall T \in \mathcal{T}_h \right\}, \tag{1.2.15}$$

and we set

$$\|\cdot\|_X = \|\cdot\|_{L^\infty(\Omega_h)}. \tag{1.2.16}$$

The example of a function from U_h is viewed in Figure 1.2. This choice of U_h covers wide range of numerical methods, piecewise constant approximation used in finite volume method (e.g. [Fei93], [Krö97], [KR94]) and conforming (e.g. [Cia79]), nonconforming (e.g. [ADFF98], [DA96], [Tem77]) and discontinuous Galerkin finite element method (e.g. [Coc]). On the other hand, the use of AMA does not require the piecewise linear approximation and a higher order approximation can be easily adopted for the efficient use of AMA.

Definition 1.1 *Let $\Omega \subset I\!R^d$. We say that $\mathcal{T}_h = \{T_i\}_{i\in I}$ (I = an index set) is a* mesh *of Ω if:*

1. *$T_i \in \mathcal{T}_h$ is a closed simplex (triangle ($d = 2$) or tetrahedron ($d = 3$)).*

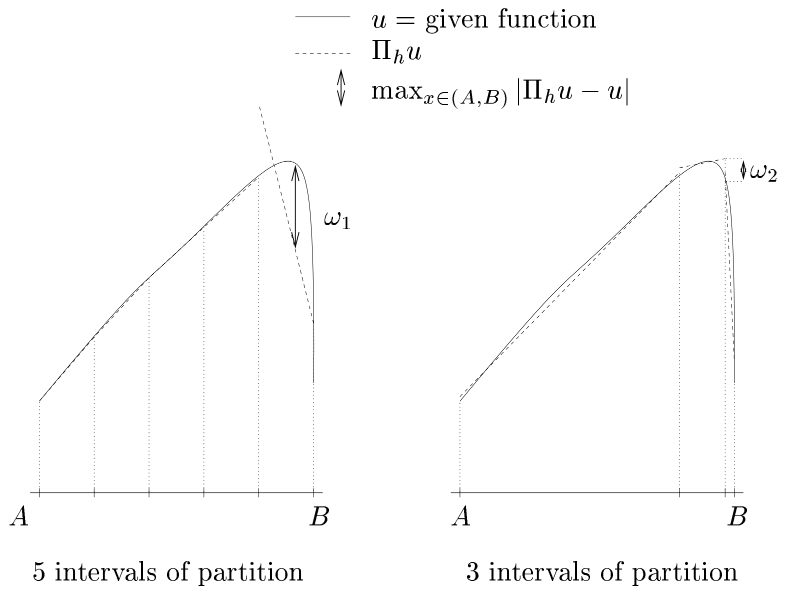


Figure 1.1: Motivation of AMA: partition of the interval AB, u is the given function, $\Pi_{h_1}u$ (left) and $\Pi_{h_2}u$ (right) are discontinuous piecewise linear functions given by (1.2.11), $\max_{x\in(A,B)} |u - \Pi_h u|$ is denoted by $\updownarrow$.

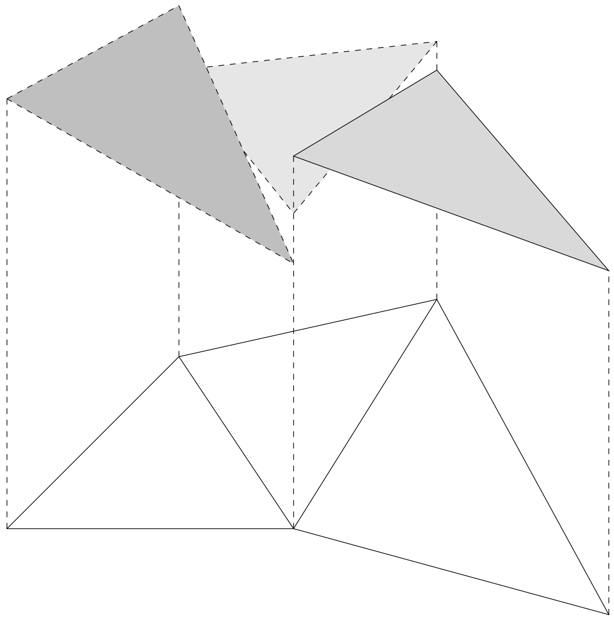


Figure 1.2: Example of a discontinuous piecewise linear function.

2. $\bigcup_{i \in I} T_i = \overline{\Omega_h}$, *where* Ω_h *is a polygonal approximation of* Ω.

3. *Let* x_i *be the barycentre of* T_i, *then* $x_i \in \Omega \ \forall T_i \in \mathcal{T}_h$.

4. *We denote by* $\sigma_h = \{P_1, \dots, P_N\}$ *the set of all vertices of all* $T \in \mathcal{T}_h$. *We suppose that* $\sigma_h \cap \partial\Omega_h \subset \partial\Omega$, $\sigma_h \subset \overline{\Omega}$.

5. *For any* $T_i, T_j \in \mathcal{T}_h$, $T_i \neq T_j$, $T_i \cap T_j$ *are either disjoint or their intersection* $T_i \cap T_j$ *is a common vertex or* $T_i \cap T_j$ *is a common edge or* $T_i \cap T_j$ *is a common face (only for* $d = 3$*).*

The step size h *of the mesh is given by*

$$h = \max_{i \in I} \operatorname{diam}(T_i), \tag{1.2.17}$$

where $\operatorname{diam}(T_i)$ *denotes the diameter of* T_i.

Without proof we present the following Lemma:

Lemma 1.1 *Let* $\Omega \in I\!R^d$ *be a computational domain and* $\bar{h} > 0$. *Then there exists a mesh* $\mathcal{T}_h$ *of* Ω *with the step size* h, $h \leq \bar{h}$.

Definition 1.2 *Let* $w \in C^2(\overline{\Omega} \cup \Omega_h)$ *and* $\mathcal{T}_h = \{T_i\}_{i \in I}$ *be a mesh of* Ω. *Let* $x_i = (x_i^{(1)}, \dots, x_i^{(d)})$ *be the centre of gravity of* $T_i \in \mathcal{T}_h$. *We say that the operator* $r_h \in \mathcal{O}(C^2(\overline{\Omega} \cup \Omega_h), U_h)$ *is the* ∇–interpolation *if*

$$\begin{aligned} r_h w(x_i) &= w(x_i) \qquad \forall T_i \in \mathcal{T}_h, \\ \nabla r_h w(x_i) &= \nabla w(x_i) \qquad \forall T_i \in \mathcal{T}_h. \end{aligned} \tag{1.2.18}$$

Lemma 1.2 *Let* $w \in C^2(\overline{\Omega} \cup \Omega_h)$ *and* $\mathcal{T}_h = \{T_i\}_{i \in I}$ *be a mesh of* Ω. *Let* U_h, $\|\cdot\|_X$, Π_h *and* r_h *are given by (1.2.15), (1.2.16), (1.2.11) and (1.2.18), respectively. Then*

$$\|w - \Pi_h w\|_X \leq \|w - r_h w\|_X. \tag{1.2.19}$$

Proof. As $r_h w \in U_h$, the assertion follows immediately from (1.2.11). □

Remark 1.1 *Lemma 1.2 is valid also for more general* $\|\cdot\|_X$.

Definition 1.3 *Let* u *be the exact solution of the considered problem (1.2.1) – (1.2.3) in the computational domain* Ω *and* r_h *be the operator (1.2.18), where* U_h *is given by (1.2.15). The* interpolation error function $E_{\mathrm{I}} : \Omega \cap \Omega_h \to I\!R^1$ *is defined by*

$$E_{\mathrm{I}}(x) = |u(x) - r_h u(x)|, \ \textit{for almost every } x = (x^{(1)}, \dots, x^{(d)}) \in \Omega \cap \Omega_h. \tag{1.2.20}$$

Lemma 1.2 allows us to control the necessary condition (1.2.13) by the interpolation error function E_{I}. The condition

$$\|u - r_h u\|_X \leq \omega \tag{1.2.21}$$

is stronger than the necessary condition (1.2.13) but as it is easily computable, we use it for further consideration. Therefore our aim is to construct a mesh $\mathcal{T}_h$ of Ω so that the interpolation error function $E_{\mathrm{I}}(x)$ is bounded from above by a given tolerance. On the other hand, we require that the number of elements of the mesh is as small as possible (in order to save a memory and CPU time). The demonstration of the efficiency of (1.2.21) is given in Section 1.5.1

Definition 1.4 *Let $u \in C^2(\overline{\Omega} \cup \Omega_h)$ and $\mathcal{T}_h = \{T_i\}_{i \in I}$ be a mesh of Ω. Let E_I be given by (1.2.20). We say that $\mathcal{T}_h$ is a ω–*mesh

$$E_\mathrm{I}(x) \le \omega \qquad \forall x \in \Omega \cap \Omega_h. \tag{1.2.22}$$

Theorem 1.1 *Let $u \in C^2(\overline{\Omega} \cup \Omega_h)$ and $\omega > 0$. Then there exists an ω-mesh.*

Proof. For $u \in C^2(\overline{\Omega} \cup \Omega_h)$ there exists a constant $\hat{c} > 0$ such that

$$\left| \frac{\partial^2 u(x)}{\partial x^{(i)} \partial x^{(j)}} \right| \le \hat{c}, \quad \forall i, j = 1, \dots, d \;\forall x \in \overline{\Omega}. \tag{1.2.23}$$

Lemma 1.1 gives that there exists a mesh $\mathcal{T}_h$ with step size

$$h \le \sqrt{\frac{2\omega}{\hat{c} d^2}}. \tag{1.2.24}$$

We prove that $\mathcal{T}_h$ is a ω–mesh. Let x_i be the centre of gravity of $T_i \in \mathcal{T}_h$. There exists the uniquely defined function $r_h u$ given by (1.2.18). Then

$$r_h u(x) = u(x_i) + \sum_{j=1}^{d} \frac{\partial u(x_i)}{\partial x^{(j)}} (x^{(j)} - x_i^{(j)}), \; x \in T_i. \tag{1.2.25}$$

Let $x \in \Omega \cap \Omega_h$, then there exists element $T_i \in \mathcal{T}_h$ such that $x \in T_i$. Then the Taylor series expansion gives

$$u(x) = r_h u(x) + \frac{1}{2} \sum_{j,k=1}^{d} \frac{\partial^2 u(\xi)}{\partial x^{(j)} \partial x^{(k)}} (x^{(j)} - x_i^{(j)})(x^{(k)} - x_i^{(k)}), \tag{1.2.26}$$

where $\xi \in \Omega$ is a point lying between x and x_i. As $|x^{(j)} - x_i^{(j)}| \le h, \; j = 1, \dots, d$ then using (1.2.23) and (1.2.25), we have from (1.2.26)

$$|u(x) - r_h u(x)| \le \frac{1}{2} d^2 \hat{c} h^2, \qquad x \in \Omega \cap \Omega_h, \tag{1.2.27}$$

which together with (1.2.24) proves the theorem. □

Definition 1.5 *Let $u \in C^2(\overline{\Omega} \cup \Omega_h)$ and $\omega > 0$. We say that $\mathcal{T}_h^{\min}$ is the ω-*minimal mesh *if $\mathcal{T}_h^{\min}$ is an ω-mesh and*

$$\#\mathcal{T}_h^{\min} = \min \#\mathcal{T}_h, \tag{1.2.28}$$

where $\#\mathcal{T}_h$ denotes the number of elements of $\mathcal{T}_h$ and the minimum is taken over all ω-meshes $\mathcal{T}_h$ of Ω.

Theorem 1.2 *Let $u \in C^2(\overline{\Omega} \cup \Omega_h)$ and $\omega > 0$. Then there exists an ω-minimal mesh.*

Proof. It follows from Theorem 1.1, that the set of ω-meshes is not empty. As the operator $\# : \mathcal{T}_h \to I\!N$ ($I\!N$ = natural numbers) is bounded from below, it attains its minimum. □

1.2.3 Optimal element

In order to approach $\mathcal{T}_h^{\min}$, we introduce the concept of an *optimal element.* Let $u \in C^2(\overline{\Omega} \cup \Omega_h)$ and let its Hessian matrix $\boldsymbol{H}$

$$\boldsymbol{H}(x) = \begin{pmatrix} \frac{\partial^2 u(x)}{\partial x^{(1)} \partial x^{(1)}} & \cdots & \frac{\partial^2 u(x)}{\partial x^{(1)} \partial x^{(d)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 u(x)}{\partial x^{(d)} \partial x^{(1)}} & \cdots & \frac{\partial^2 u(x)}{\partial x^{(d)} \partial x^{(d)}} \end{pmatrix}, \qquad x \in \Omega \cup \Omega_h \tag{1.2.29}$$

be symmetric positive definite. Let $\mathcal{T}_h$ be an ω–mesh and $r_h u$ be the ∇–interpolation given by (1.2.18). Let $x_i = (x_i^{(1)}, \ldots, x_i^{(d)})$ be the centre of gravity of $T_i \in \mathcal{T}_h$ and $x = (x^{(1)}, \ldots, x^{(d)}) \in T_i$. Then, using the Taylor series expansion at x_i, we have

$$u(x) = r_h u(x) + \frac{1}{2} \sum_{j,k=1}^{d} \frac{\partial^2 u(x_i)}{\partial x^{(j)} \partial x^{(k)}} (x^{(j)} - x_i^{(j)})(x^{(k)} - x_i^{(k)}) + o(|x - x_i|^2). \tag{1.2.30}$$

In what follows we shall neglect the terms of order $o(|x-x_i|^2)$ in equalities and inequalities. Up to terms of this order we can write

$$\begin{aligned} E_{\mathrm{I}}(x) &= |u(x) - r_h u(x)| = \qquad\qquad (1.2.31) \\ &= \frac{1}{2} \left| \sum_{j,k=1}^{d} \frac{\partial^2 u(x_i)}{\partial x^{(j)} \partial x^{(k)}} (x^{(j)} - x_i^{(j)})(x^{(k)} - x_i^{(k)}) \right|, \quad x \in T_i. \end{aligned}$$

As $\mathcal{T}_h$ is the ω–mesh we have

$$\left| \sum_{j,k=1}^{d} \frac{\partial^2 u(x_i)}{\partial x^{(j)} \partial x^{(k)}} (x^{(j)} - x_i^{(j)})(x^{(k)} - x_i^{(k)}) \right| \le 2\omega, \quad x \in T_i, \tag{1.2.32}$$

which can be rewritten as

$$\left(x^{(1)} - x_i^{(1)}, \ldots, x^{(d)} - x_i^{(d)}\right) \begin{pmatrix} \frac{\partial^2 u(x_i)}{\partial x^{(1)} \partial x^{(1)}} & \cdots & \frac{\partial^2 u(x_i)}{\partial x^{(1)} \partial x^{(d)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 u(x_i)}{\partial x^{(d)} \partial x^{(1)}} & \cdots & \frac{\partial^2 u(x_i)}{\partial x^{(d)} \partial x^{(d)}} \end{pmatrix} \begin{pmatrix} x^{(1)} - x_i^{(1)} \\ \vdots \\ x^{(d)} - x_i^{(d)} \end{pmatrix} \le 2\omega, \ x \in T_i. \tag{1.2.33}$$

The points $x \in I\!R^d$ satisfying the relation (1.2.33) form an ellipse (for $d = 2$) and an ellipsoid (for $d = 3$) $\epsilon_i(\omega) \subset I\!R^d$, see Appendix.

From this it follows that the interpolation error function $E_{\mathrm{I}}(x)$ is bounded from above by ω for all $x \in \epsilon_i(\omega)$. Consequently, if T is such a element that $T \subset \epsilon_i(\omega)$, then $E_{\mathrm{I}}(x) \le \omega$ for all $x \in T$.

Definition 1.6 *Let $u \in C^2(\overline{\Omega} \cup \Omega_h)$ be a function with symmetric positive definite Hessian matrix, $\omega > 0$, $\mathcal{T}_h$ an ω–mesh, $T_i \in \mathcal{T}_h$ and $\epsilon_i(\omega)$ the ellipse ($d = 2$) or the ellipsoid ($d = 3$) given by (1.2.33). We say that T_i is an* optimal element *if*

$$T_i \subset \epsilon_i(\omega) \tag{1.2.34}$$

and

$$\mathrm{meas}(T_i) \ge \mathrm{meas}(T) \ \textit{for all elements}\ T \subset \epsilon_i(\omega). \tag{1.2.35}$$

Lemma 1.3 *Let $u \in C^2(\overline{\Omega} \cup \Omega_h)$ be a function with symmetric positive definite Hessian matrix, $\omega > 0$ be given, $x_i \in \Omega$ and $\epsilon_i(\omega)$ the ellipse defined by (1.2.33). Then there exist elements T_i^j, $j = 1, \ldots, N(d)$ with the centre of gravity x_i satisfying (1.2.34) and (1.2.35), where $N(2) = 4$ and $N(3) = 24$.*

Proof. See Appendix. □

If T_i is the optimal element for every $T_i \in \mathcal{T}_h$ then $E_{\mathrm{I}}(x) \leq \omega$ for every $x \in \Omega \cap \Omega_h$. Therefore $\mathcal{T}_h$ is the ω–mesh and due to (1.2.35) $\mathcal{T}_h$ has the reasonable small number of elements and approximates well $\mathcal{T}_h^{\mathrm{min}}$.

In order to describe geometrically the optimal element, we introduce the following

Definition 1.7 *Let $\boldsymbol{M}$ be a $d \times d$ symmetric positive definite matrix and $\boldsymbol{v} \in I\!R^d$. We define* the norm *of the vector $\boldsymbol{v}$* corresponding *to the matrix $\boldsymbol{M}$ as*

$$\|\boldsymbol{v}\|_{\boldsymbol{M}} \equiv \left(\boldsymbol{v}^{\mathrm{T}} \boldsymbol{M} \boldsymbol{v}\right)^{\frac{1}{2}}. \tag{1.2.36}$$

Remark 1.2 *If $\boldsymbol{M} = \boldsymbol{I}$ (=identity matrix) then $\|\boldsymbol{v}\|_{\boldsymbol{M}} = |\boldsymbol{v}|$, where $|\boldsymbol{v}|$ is the Euclidean norm of $\boldsymbol{v}$.*

Lemma 1.4 *Let $u \in C^2(\overline{\Omega} \cup \Omega_h)$ be a function with symmetric positive definite Hessian matrix $\boldsymbol{H}(u(x))$, $x \in \Omega$, $\omega > 0$, $\mathcal{T}_h$ a mesh of Ω, $T_i \in \mathcal{T}_h$ with centre of gravity x_i and $\epsilon_i(\omega)$ the ellipse given by (1.2.33). Let V_i^j, $j = 1, \ldots, d+1$ be vertices of T_i. Then T_i is the optimal element if and only if*

$$\|V_i^k - V_i^l\|^2_{\boldsymbol{M}} = 2C(d)\omega, \quad k, l = 1, \ldots, d+1, \ k \neq l, \tag{1.2.37}$$

where we put $\boldsymbol{M} = \boldsymbol{H}(u(x_i))$ and $C(2) = 3$ and $C(3) = \frac{8}{3}$.

Proof. See Appendix □

1.2.4 Optimal mesh

Definition 1.7 and Lemma 1.4 give us a suitable tool for the practical construction of the mesh $\mathcal{T}_h^{\mathrm{min}}$. On the other hand the condition (1.2.37) for $i = i_1, i_2$ can not be generally satisfied for two elements $T_{i_1} \neq T_{i_2}$ with the common edge $\partial T_{i_1} \cap \partial T_{i_2} \neq \emptyset$. In order to avoid these difficulties, we pass from the centre of gravities associated Hessian matrixes to the edge associated Hessian matrixes. This edge associated approach is simpler for the numerical implementation.

Definition 1.8 *Let $u \in C^2(\overline{\Omega} \cup \Omega_h)$ be a function with symmetric positive definite Hessian matrix and $\omega > 0$ be given. Let $\mathcal{T}_h$ be a mesh of the computational domain Ω. Let $\boldsymbol{H}_k$ be the Hessian matrix of u evaluated in the centre of the edge $\boldsymbol{e}_k$, $k \in K$(= an index set) of the mesh $\mathcal{T}_h$. We say that the mesh $\mathcal{T}_h$ is* edge-optimal, *if*

$$\|\boldsymbol{e}_k\|_{\boldsymbol{H}_k} = \sqrt{2C(d)\omega} \qquad \forall k \in K. \tag{1.2.38}$$

Remark 1.3 *The matrix $\boldsymbol{H}_k$ can be interpreted as an* interpolation error matrix *and the norm $\|\boldsymbol{e}_k\|_M$ as an* interpolation error over the edge *$\boldsymbol{e}_k$ (see [CDBG+96], [FVD+96]). It means that the interpolation error (considered in terms of $\|\cdot\|_{\boldsymbol{H}_k}$) is uniformly distributed over the edges of the mesh.*

Remark 1.4 *It is easy to see that the optimal mesh exists only for special types of domains Ω and the Hessian matrixes. Let $\boldsymbol{H}_k = c\boldsymbol{I}$ $\forall k \in K$ where $\boldsymbol{I}$ is the unit matrix and $c > 0$. Then the edge-optimal mesh is uniform, i.e., the (Euclidean) lengths of $\boldsymbol{e}_k$, $k \in K$ are the same therefore all elements are equilateral. Let $\Omega = [0,1]^2$ then the optimal triangulation does not exist because the angle of triangles at the corners of the domain cannot be equal to $\pi/3$.*

It lead us to a weaker form of the concept of edge-optimal mesh. We introduce the quality parameter $Q_{\mathcal{T}_h}$, which measures the edge-optimality in the least squares sense.

Definition 1.9 *Let $u \in C^2(\overline{\Omega} \cup \Omega_h)$ be a function with symmetric positive definite Hessian matrix and $\omega > 0$ be given. Let $\mathcal{T}_h$ be a mesh of the computational domain Ω. Let $\boldsymbol{H}_k$ be the Hessian matrix of u evaluated in the centre of the edge $\boldsymbol{e}_k$, $k \in K$(= an index set) of the mesh $\mathcal{T}_h$. We say that the mesh $\mathcal{T}_h^{\mathrm{opt}}$ is* **optimal** *if*

$$Q_{\mathcal{T}_h^{\mathrm{opt}}} = \min Q_{\mathcal{T}_h}, \tag{1.2.39}$$

where the minimum is taken over all meshes of Ω and

$$Q_{\mathcal{T}_h} = \frac{1}{\#\mathcal{T}_h} \sum_{k \in K} \left(\|\boldsymbol{e}_k\|_{\boldsymbol{H}_k} - \sqrt{2C(d)\omega} \right)^2, \tag{1.2.40}$$

where $\#\mathcal{T}_h$ is the number of edges of $\mathcal{T}_h$. We call the quantity $Q_{\mathcal{T}_h}$ the quality parameter *of $\mathcal{T}_h$.*

The quality parameter $Q_{\mathcal{T}_h}$ is a computable quantity for the given function u and the mesh $\mathcal{T}_h$. Definition 1.9 gives us an idea how to adapt the mesh. The mesh adaptation algorithm is described in Section 1.3. In Section 1.2.5, we present another determination of AMA which leads to the same definition of the quality of a mesh.

1.2.5 Alternative determination of AMA

It is possible to introduce the term edge-optimal mesh from Definition 1.8 without the term optimal element from Section 1.2.3. Our aim is to adapt a given mesh $\mathcal{T}_h$ so that the new one satisfies the stronger condition (1.2.21) and $\#\mathcal{T}_h$ is minimal. Let us suppose that the dependence of $\|u - r_h u\|_X$ on $\#\mathcal{T}_h$ is monotone. Then the inverse problem can be formulated: Adapt a given mesh $\mathcal{T}_h$ so that the interpolation error $\|u - r_h u\|_X$ is minimal for a fixed $\#\mathcal{T}_h$.

The basic idea is the following: In order to minimize the interpolation error for a given $\#\mathcal{T}_h$, the interpolation error function E_{I} defined by (1.2.20) should be equidistantly distributed over the whole computational domain Ω, i.e.

$$E_I(x) \approx C \qquad \forall x \in \Omega, \tag{1.2.41}$$

where $C > 0$ is a constant.

In order to minimize the interpolation error function E_I, we introduce the discrete version of the condition (1.2.41). We consider the interpolation error function E_I over edges for practical reasons. Let $\boldsymbol{e}$ be an edge of $\mathcal{T}_h$ (connecting two nodes of $\mathcal{T}_h$), $|\boldsymbol{e}|$ notes the Euclidean length of $\boldsymbol{e}$ and let $x_{\boldsymbol{e}}$ be the centre of $\boldsymbol{e}$. We approximate $E_I|_{\boldsymbol{e}}$ by the mean value of E_I over $\boldsymbol{e}$. Then the omitting the terms of higher order yields

$$\begin{aligned} E_I|_{\boldsymbol{e}} &\approx \frac{1}{|\boldsymbol{e}|} \int_{\boldsymbol{e}} |u(x) - r_h u(x)| \;\mathrm{d}S \;\approx \\ &\approx \frac{1}{2} \frac{1}{|\boldsymbol{e}|} \int_{\boldsymbol{e}} |(x - x_{\boldsymbol{e}}) \boldsymbol{H}(x_{\boldsymbol{e}}) (x - x_{\boldsymbol{e}})| \;\mathrm{d}S = \frac{1}{24} \left| \vec{\boldsymbol{e}}^{\mathrm{T}} \boldsymbol{H}(x_{\boldsymbol{e}}) \vec{\boldsymbol{e}} \right|, \end{aligned} \tag{1.2.42}$$

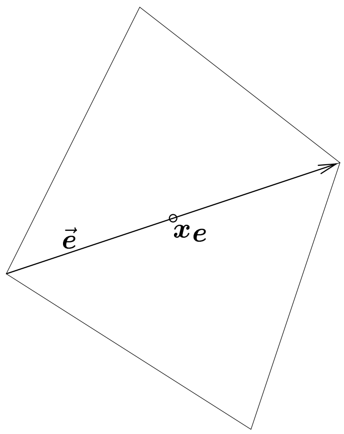


Figure 1.3: The vector $\vec{\boldsymbol{e}}$ corresponding to the edge $\boldsymbol{e} \in \mathcal{T}_h$ and its centre $x_{\boldsymbol{e}}$.

where $\vec{\boldsymbol{e}} \in \mathbb{R}^d$ is the vector parallel with the edge $\boldsymbol{e}$ and with the Euclidean norm equal to $|\boldsymbol{e}|$, see Figure 1.3.

Using (1.2.36), the norm of the edge $\boldsymbol{e}$ corresponding to the matrix $\boldsymbol{H}$ is

$$\|\boldsymbol{e}\|_{\boldsymbol{H}} \equiv \left(\left|\vec{\boldsymbol{e}}^{\mathrm{T}} \boldsymbol{H}(x_{\boldsymbol{e}})\vec{\boldsymbol{e}}\right|\right)^{1/2}. \tag{1.2.43}$$

We have from (1.2.42) and (1.2.43), that the interpolation error function is uniformly distributed over the mesh $\mathcal{T}_h$ if

$$E_I|_{\boldsymbol{e}} \approx \frac{1}{24}\|\boldsymbol{e}\|^2_{\boldsymbol{H}} \approx \omega \qquad \text{for any edge } \boldsymbol{e} \text{ of } \mathcal{T}_h, \tag{1.2.44}$$

where $\omega > 0$ is the given accuracy. With the aid of (1.2.44) we define the following:

Definition 1.10 *The mesh $\mathcal{T}_h$ is* optimal *if and only if*

$$\|\boldsymbol{e}_k\|_{\boldsymbol{H}_k} = \sqrt{24\,\omega} \quad \forall \boldsymbol{e}_k \in \mathcal{T}_h, \qquad \boldsymbol{e}_k \text{ edge of } \mathcal{T}_h, \quad \boldsymbol{H}_k \equiv \boldsymbol{H}(x_{\boldsymbol{e}_k}), \tag{1.2.45}$$

where ω is a given constant which plays a role of accuracy.

Remark 1.5 *Definition 1.10 is identical (modulo a scaling constant) to Definition 1.8. Therefore it leads also to Definition 1.9 introducing the term optimal mesh. The first concept (Sections 1.2.2 – 1.2.4) has better geometrical interpretation. However, the second approach (Section 1.2.5) should be used in order to derive some a priori error estimations.*

1.3 Mesh optimization

The quality parameter $Q_{\mathcal{T}_h}$ given by (1.2.40) is always nonnegative and it is equal to zero if the mesh is edge-optimal in the sense of the Definition 1.8. Therefore we adapt the mesh $\mathcal{T}_h$ in order to decrease $Q_{\mathcal{T}_h}$ and to approach $\mathcal{T}_h^{\mathrm{opt}}$ and so $\mathcal{T}_h^{\min}$. Let $u \in C^2(\overline{\Omega} \cup \Omega_h)$, $\omega > 0$, $\mathcal{T}_{\mathrm{old}}$ be a mesh of Ω and $\boldsymbol{H}_k$, $k \in K$ be the Hessian matrices of u evaluated in the centre of the edge $\boldsymbol{e}_k$. The mesh contains two type of nodes: I (for internal) and B (for boundary). An II edge is an edge that joins two I-nodes, and the other combinations are defined analogously. We want to find a new mesh $\mathcal{T}_{\mathrm{new}}$ such that $Q_{\mathcal{T}_{\mathrm{new}}}$ is minimal from all possible meshes. To obtain $\mathcal{T}_{\mathrm{new}}$, an iterative process is used. Each iteration consists of a certain number of the local mesh operations described in the next paragraphs.

1.3.1 Mesh adaptation – d=2

The local mesh operations for $d = 2$ (viewed in Figure 1.4) are the following

- **Adding a node** ($\mathcal{A}$) - If the norm of some edge of a triangle is greater than $\frac{4}{3}\sqrt{2C(d)\omega}$ then we put a new node.
  - *II or IB-edge*: we put a new node at the centre of this edge.
  - *BB-edge*: a new node is the orthogonal projection of the centre of this edge on $\partial\Omega$.

  The constant $\frac{4}{3}\sqrt{2C(d)\omega}$ is the smallest positive solution of the inequality

  $$\left|\|\boldsymbol{e}\|_{\boldsymbol{H}} - \sqrt{2C(d)\omega}\right| \geq \left|\frac{1}{2}\|\boldsymbol{e}\|_{\boldsymbol{H}} - \sqrt{2C(d)\omega}\right|, \tag{1.3.1}$$

  which means that after adding a node, two edges with the norms $\frac{1}{2}\|\boldsymbol{e}\|_{\boldsymbol{H}}$ arise and the difference $\frac{1}{2}\|\boldsymbol{e}\|_H$ and $\sqrt{2C(d)\omega}$ is smaller than the difference of $\|\boldsymbol{e}\|_H$ and $\sqrt{2C(d)\omega}$.

- **Removing an edge** ($\mathcal{R}$) - If the norm of some edge of a triangle is smaller than $\frac{2}{3}\sqrt{2C(d)\omega}$ we remove this edge. The initial and final nodes of this side pass into a new node, whose position depends on the type of edge.
  - *II-edge*: the new position is at the centre of removed edge.
  - *IB-edge*: the new position is at the position of the B-node
  - *BB-edge*: the new position is at the orthogonal projection of the centre of removed edge on $\partial\Omega$.

  The constant $\frac{2}{3}\sqrt{2C(d)\omega}$ is the highest solution of the inequality

  $$\left|\|\boldsymbol{e}\|_{\boldsymbol{H}} - \sqrt{2C(d)\omega}\right| \geq \left|2\|\boldsymbol{e}\|_{\boldsymbol{H}} - \sqrt{2C(d)\omega}\right|, \tag{1.3.2}$$

  which introduce a heuristical consideration opposite to (1.3.1).

- **Swapping**($\mathcal{S}$) the diagonal of the quadrilateral formed by any pair of adjacent elements. The edge is swapped if the quality parameter $Q_{\mathcal{T}}$ after swapping is smaller than before.

- **Moving a node** ($\mathcal{M}$) is the only non discrete local operation. We seek a new position of a node (positions of other nodes are fixed), which minimize the quality parameter $Q_{\mathcal{T}}$.
  - *B-node*: we move the node along $\partial\Omega$ and the new position minimizes the quality parameter $Q_{\mathcal{T}}$.
  - *I-node*: the moving of node P_i consists of the following steps:

1. Let $\mathcal{P}(i)$ be the set of nodes $P_j \in \sigma_h$ such that $\boldsymbol{e}_{ij} \equiv P_iP_j$ is an edge of the mesh. Then we define the nodes

$$P^*_{i,j} \equiv P_i + \alpha_j(P_j - P_i) \quad \forall j \in \mathcal{P}(i)$$

where $\alpha_j > 0$ is chosen in such a way that $\|P^*_{i,j} - P_i\|_{\boldsymbol{H}_{\boldsymbol{e}_{ij}}} = \sqrt{2C(d)\omega}\ \forall j \in \mathcal{P}(i)$.

2. We define the node P^*_i by

$$P^*_i \equiv \frac{1}{\#\mathcal{P}(i)} \sum_{j \in \mathcal{P}(i)} P^*_{i,j},$$

where $\#\mathcal{P}(i)$ denotes the number of nodes having a common edge with P_i.

3. Let $P'_i(s)$ be a point such that

$$P'_i(s) = P_i + s(P^*_i - P_i), \quad s \in [0,1]$$

If we move P_i into $P'_i(s)$, we obtain a triangulation $\mathcal{T}(s)$ with the quality parameter $Q_\mathcal{T}(s)$. We find $s^* \in [0,1]$ such that $Q_\mathcal{T}(s^*) \le Q_\mathcal{T}(s)$ for all $s \in [0,1]$ and the new position of P is $P_i := P'_i(s^*)$.

Remark 1.6 *The local operation moving a node should seem to be not very effective. There exists another approach ([FOG97]), where the optimal position is sought by a gradient method. On the other hand, the advantage of our proposed algorithm is its simplicity and therefore quickness. As the iterative process consists of many repetitions of all local operations, it is sufficient to find the new position of the node approximately. Numerical experiments show that our algorithm is more effective from CPU-time point of view.*

1.3.2 Mesh adaptation – d=3

The two dimensional operations **adding a node** ($\mathcal{A}$), **removing an edge** ($\mathcal{R}$) and **moving a node** ($\mathcal{M}$) should be easily extended also for the case $d = 3$. Only the operation **swapping** is more complicated and we distinguish the following cases similarly as in [FOG97].

- **Face swapping** ($\mathcal{S}_F$) reconnects the tetrahedra separated by a single interior face. The operation is viewed in Figure 1.5 and Figure 1.6 where two tetrahedra yields three and two new ones for interior and boundary face, respectively. The face is swapped if the quality parameter $Q_\mathcal{T}$ after swapping is smaller than before.

- **Edge swapping** ($\mathcal{S}_E$) is performed for the set of M tetrahedra ($M \ge 3$) having a common edge AB. Moreover, the tetrahedra have the nodes $P_1, \dots, P_M$, see Figure 1.7. So that we have tetrahedra $ABP_jP_{j+1},\ j = 1, \dots, M$ where $P_{M+1} \equiv P_1$. Performing the edge swapping we obtain $2M - 4$ new tetrahedra having the nodes $AP_iP_jP_k$ and $BP_iP_jP_k$, see Figure 1.7. It means that the polygon $P_1 \dots P_M$ is divided into $M-2$ triangles $P_iP_jP_k$. Figure 1.8 shows how we can divide polygon $P_1 \dots P_M$ into $M-2$ triangles for $M = 3, \dots, 7$. The upper indexes of the pictures denote the number of subcases arising by the index shifting. The edge is swapped if the quality parameter $Q_\mathcal{T}$ after swapping is smaller than before.

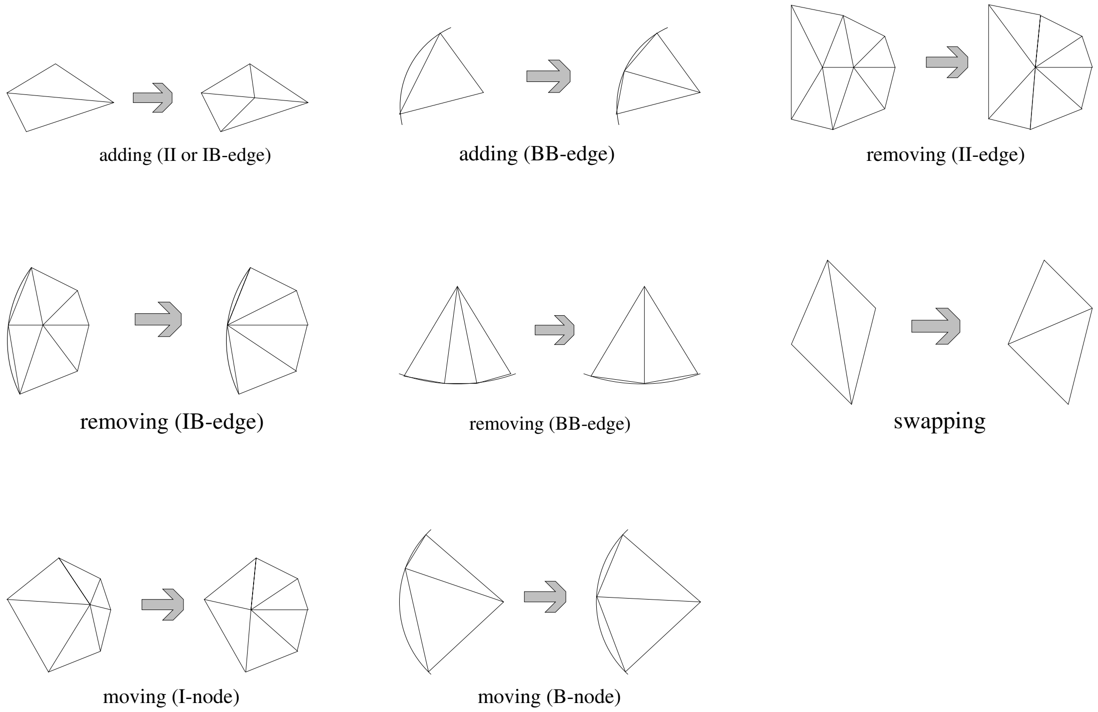


Figure 1.4: Two dimensional local operations.

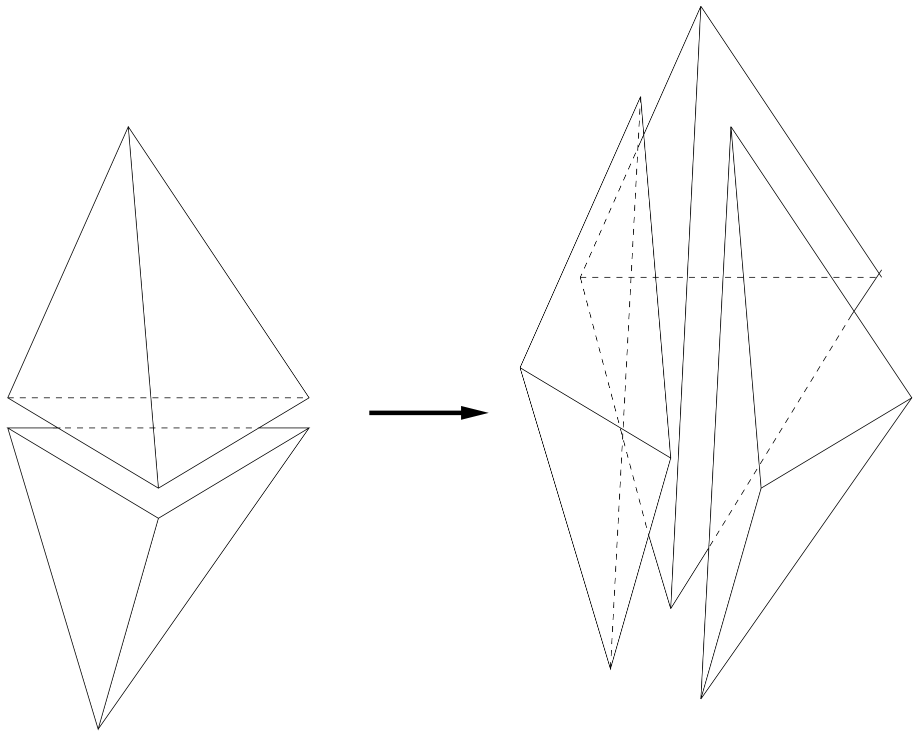


Figure 1.5: Face swapping for an interior face.

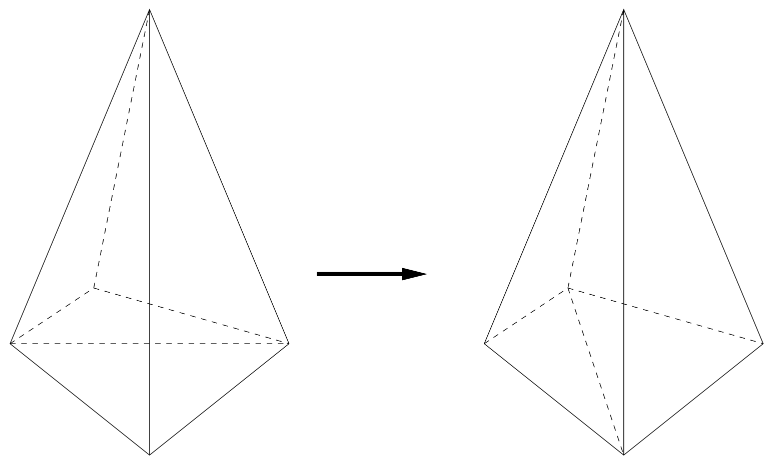


Figure 1.6: Face swapping for a boundary face, the bottom faces belongs to $\partial\Omega_h$.

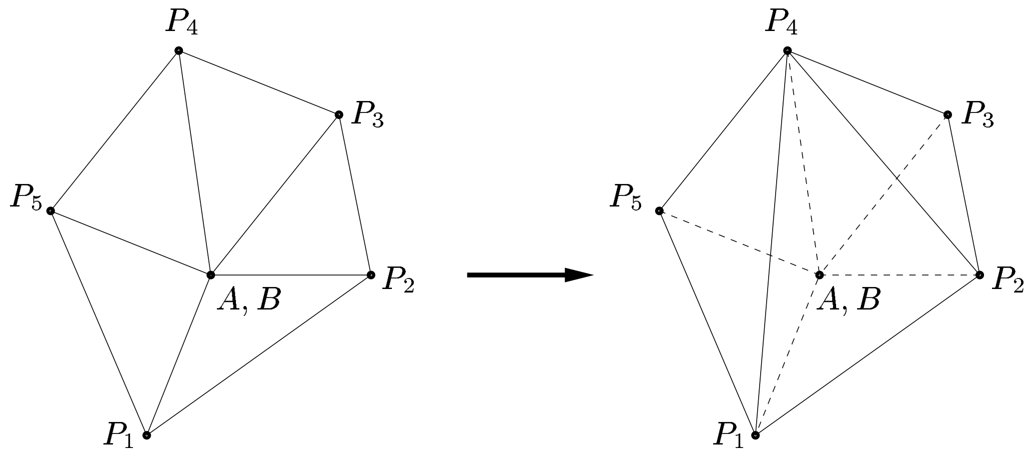


Figure 1.7: Example of edge swapping from 5 to 6 tetrahedra. The view is parallel with the edge AB, therefore the edge AB pass into one point.

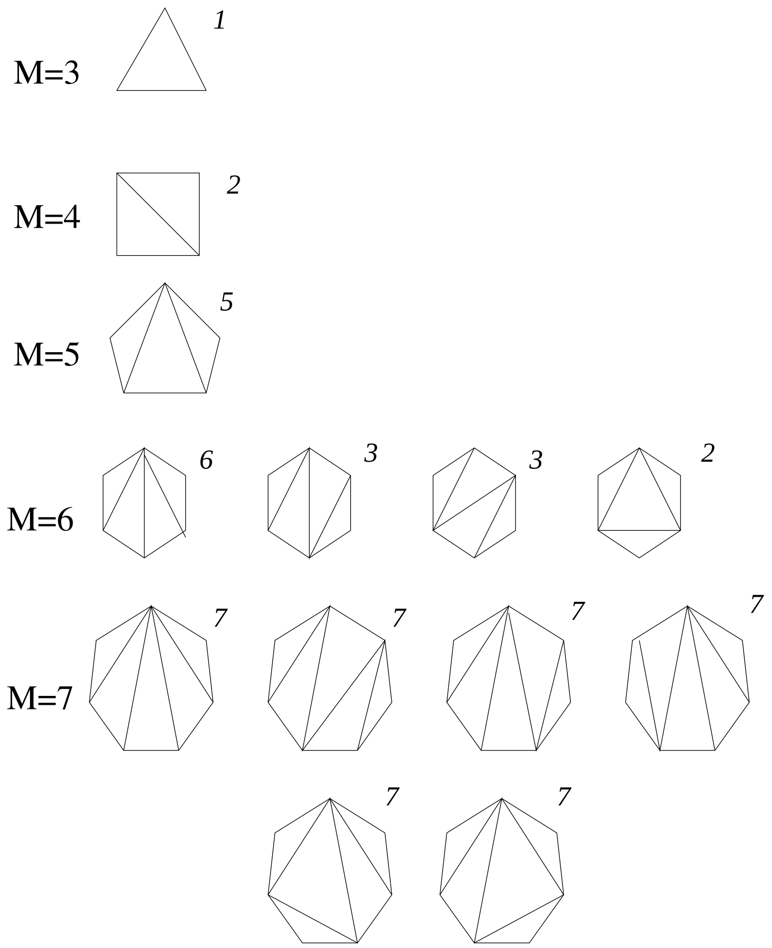


Figure 1.8: Triangles after edge swapping for $3 \leq M \leq 7$, the upper index denotes the number of unique rotations for each case.

1.3.3 Minimal angle condition

From theoretical as well as computational reasons, it is suitable to ensure the shape regularity of the mesh. The mesh is shape regular if there exists a constant β such that

$$\frac{\mathrm{diam}\ (T)}{\rho_T} \leq \beta \qquad \forall T \in \mathcal{T}_h, \tag{1.3.3}$$

where ρ_T denotes the radius of a maximal $d-$dimensional ball inscribed into $T \in \mathcal{T}_h$. To guarantee that the condition (1.3.3) is valid, we define the condition

$$\frac{\mathrm{meas}_d(T)}{\sum_{\boldsymbol{e}_k^T \subset T} |\boldsymbol{e}_k^T|^d} \geq \gamma \qquad \forall T \in \mathcal{T}_h, \tag{1.3.4}$$

where $\mathrm{meas}_d(T)$ is the $d-$dimensional measure of T, $|\boldsymbol{e}_k^T|$ is the (Euclidean) length of the edges $\boldsymbol{e}_k^T$ forming the element $T \in \mathcal{T}_h$. Then any local operation is performed only if the condition (1.3.4) is satisfied for a prescribed number $\gamma > 0$. In examples presented in this work we put $\gamma = 0.02$.

1.3.4 Iterative process

Based on a very large range of numerical experiments we present the AMA algorithm using the combination of the above operations $\mathcal{A}, \mathcal{R}, \mathcal{S}$, and $\mathcal{M}$ as (for $d = 3$ we put $\mathcal{S} = \mathcal{S}_F + \mathcal{S}_E$)

$$m \times \{\mathcal{M} + n_1 \times [\mathcal{R} + \mathcal{S}] + \mathcal{M} + n_2 \times [\mathcal{A} + \mathcal{S}] + M + \mathcal{S}\} + \mathcal{M}, \tag{1.3.5}$$

(the sequence of operations goes from the left to the right), where m, n_1 and n_2 are integers denoting the number of repetitions of the sequence of operations in the square brackets. The sequence in the square brackets is applied until no local discrete operation $\mathcal{S}$, $\mathcal{R}$ and $\mathcal{A}$ are performed. Numerical examples show that the algorithm (1.3.5) converges, i.e. it finish after a finite number of steps.

1.4 Application of AMA in numerical solution process

In the previous section we have shown how to construct the mesh $\mathcal{T}_h^{\mathrm{opt}}$ for the smooth known exact solution u. In practical examples, of course, we do not know the exact solution u and therefore we use a numerical approximation u_h on a mesh $\mathcal{T}_h$, on which we apply a smoothing procedure and use the technique from previous section. What we need is not the smoothed function itself but its Hessian matrices $\boldsymbol{H}$ evaluated in the centres of edges of $\mathcal{T}_h$.

1.4.1 Computation of the second order derivatives

The Hessian matrix $\boldsymbol{H}$ evaluated in the centre of an edge $\boldsymbol{e}_k$ of $\mathcal{T}_h$ is computed as the average of the Hessian matrixes evaluated in the initial and final point forming $\boldsymbol{e}_k$. Let $u_h \in U_h$ (defined in (1.2.15)) be an approximate solution of (1.2.4) – (1.2.6) and

$$\begin{aligned} W_h &= \{w_h; w_h \in C(\overline{\Omega_h}), w_h \text{ is linear on each } T \in \mathcal{T}_h\} \qquad (1.4.1) \\ W_h^0 &= \{w_h; w_h \in W_h, w_h = 0 \text{ on } \partial\Omega_h\}. \end{aligned}$$

We introduce the operator $u_h \in U_h \mapsto \bar{u}_h \in W_h$, such that

$$\bar{u}_h(P_i) = \frac{\sum_{T\in D_i} \int_T u_h(\xi)\, d\xi}{\sum_{T\in D_i} \mathrm{meas}(T)}, \tag{1.4.2}$$

where $P_i \in \sigma_h$ (see Definition 1.1) and D_i is the set of elements from $\mathcal{T}_h$ having P_i in common.

We compute the elements of the Hessian matrix $\boldsymbol{H}$ using the evaluation of the second order derivatives of $\bar{u}_h$ in the following sense. We approximate $\frac{\partial^2 \bar{u}_h}{\partial x^{(i)} \partial x^{(j)}}$, $i,j = 1,\dots,d$ by a function $\boldsymbol{d}^{(i,j)} u_h \in W_h$ such that

$$\left(\frac{\partial^2 \bar{u}_h}{\partial x^{(i)} \partial x^{(j)}}, \varphi\right) \approx \left(\boldsymbol{d}^{(i,j)} u_h, \varphi\right) = -\left(\frac{\partial \bar{u}_h}{\partial x^{(i)}}, \frac{\partial \varphi}{\partial x^{(j)}}\right) \quad \forall \varphi \in W_h^0,\ i,j = 1,\dots,d, \tag{1.4.3}$$

where $(\cdot,\cdot)$ means the scalar product in $L^2(\Omega_h)$. The R.H.S. of (1.4.3) has meaning because $\bar{u}_h, \varphi \in W_h$ and therefore the first derivatives exist and are piecewise constant. The choice of the standard finite element hat test functions from W_h^0 leads to the explicit formula for $\boldsymbol{d}^{(i,j)} u_h$, $i,j = 1,\dots,d$. The relations (1.4.3) determine the values of $\boldsymbol{d}^{(i,j)} u_h$, $i,j = 1,\dots,d$ in inner nodes $P \in \sigma_h \cap \Omega$. For nodes lying on $\partial\Omega$ we use the extrapolation from interior part of the computational domain. The values of $\boldsymbol{d}^{(i,j)} u_h$, $i,j = 1,\dots,d$ in boundary nodes are averages of $\boldsymbol{d}^{(i,j)} u_h$ in nodes having a common edge with a boundary nodes. Numerical examples confirm that such simple extrapolation yields satisfactory results. Finally, we put

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{d}^{(1,1)} u_h & \dots & \boldsymbol{d}^{(1,d)} u_h \\ \vdots & \ddots & \vdots \\ \boldsymbol{d}^{(d,1)} u_h & \dots & \boldsymbol{d}^{(d,d)} u_h \end{pmatrix}. \tag{1.4.4}$$

Lemma 1.5 *Let $u_h \in W_h^0$. Then the approximations of second order derivatives given by (1.4.3) are symmetric, i.e.*

$$\boldsymbol{d}^{(i,j)} u_h = \boldsymbol{d}^{(j,i)} u_h \qquad \forall i,j = 1,\dots,d. \tag{1.4.5}$$

Proof. We perform the proof for the case $d = 3$, the two dimensional case we obtain by an easy simplification. Let $\{\varphi_k,\ \varphi_k(P_l) = \delta_{kl},\ k,l \in J_h^\circ\}$ be a basis of W_h^0, where $\{P_l,\ l \in J_h^\circ\}$ is the set of all inner nodes of $\mathcal{T}_h$. Then using (1.4.3), it is sufficient to prove that

$$\left(\frac{\partial \varphi_k}{\partial x^{(i)}}, \frac{\partial \varphi_l}{\partial x^{(j)}}\right) = \left(\frac{\partial \varphi_k}{\partial x^{(j)}}, \frac{\partial \varphi_l}{\partial x^{(i)}}\right) \qquad \forall i,j = 1,\dots,d,\ \forall k,l \in J_h^\circ. \tag{1.4.6}$$

The relations (1.4.6) are nontrivial if the functions $\partial\varphi_k$ and $\partial\varphi_l$ have a nonempty common support which means that there exists an edge $\boldsymbol{e}_{kl}$ such that P_k and P_l are its final nodes. Then there exists M tetrahedra having the edge $\boldsymbol{e}_{kl}$ in common, we denote them by $T_m \equiv P_k P_l P_m P_{m+1}$, $m = 1,\dots,M$ (we put $P_{M+1} := P_1$), see Figure 1.9.

Let D_m, $m = 1,\dots,M$ be a 4×4 matrix corresponding to the tetrahedron T_m given by

$$\boldsymbol{D}_m \equiv \left\{(\boldsymbol{D}_m)_{r,s}\right\}_{r,s=1}^4 = \begin{pmatrix} x_k^{(1)} & x_k^{(2)} & x_k^{(3)} & 1 \\ x_l^{(1)} & x_l^{(2)} & x_l^{(3)} & 1 \\ x_m^{(1)} & x_m^{(2)} & x_m^{(3)} & 1 \\ x_{m+1}^{(1)} & x_{m+1}^{(2)} & x_{m+1}^{(3)} & 1 \end{pmatrix}, \qquad m = 1,\dots,M, \tag{1.4.7}$$

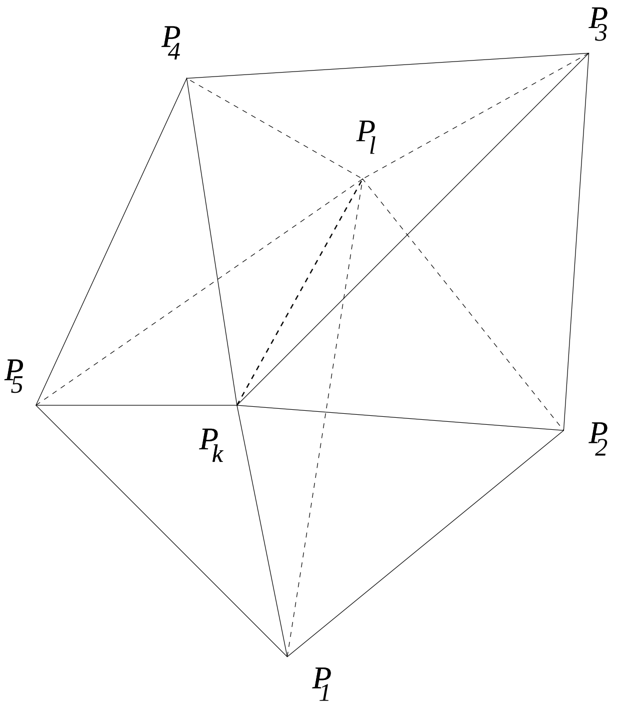


Figure 1.9: Tetrahedra having the edge P_kP_l in common.

where $(x_k^{(1)}, x_k^{(2)}, x_k^{(3)})$ are the coordinates of node P_k and similarly for other nodes. Moreover we define the matrixes $\boldsymbol{D}_m^{p,q}$, $p,q = 1,\dots,4$, $m = 1,\dots,M$ which arise replacing the q-th column of $\boldsymbol{D}_m$ by the column $(\delta_{p1}, \delta_{p2}, \delta_{p3}, \delta_{p4})$, i.e.

$$\boldsymbol{D}_m^{p,q} \equiv \left\{(\boldsymbol{D}_m^{p,q})_{r,s}\right\}_{r,s=1}^{4}, \qquad (\boldsymbol{D}_m^{p,q})_{r,s} = \begin{cases} (\boldsymbol{D}_m)_{r,s} & \text{if } q \neq s \\ \delta_{pr} & \text{if } q = s. \end{cases} \tag{1.4.8}$$

As $\varphi_k(P_l) = \delta_{kl}$, $k,l \in J_h^\circ$ and φ_k, $k \in J_h^\circ$ is linear over each tetrahedron, it is possible observe that

$$\frac{\partial \varphi_p}{\partial x^{(q)}}|_{T_m} = \frac{\det \boldsymbol{D}_m^{p,q}}{\det \boldsymbol{D}_m}, \qquad p = 1,2, \ q = 1,2,3, \ m = 1,\dots,M, \tag{1.4.9}$$

where we put $\varphi_1 = \varphi_k$ and $\varphi_2 = \varphi_l$. From (1.4.6) and (1.4.9) we have

$$\begin{aligned} \left(\frac{\partial \varphi_k}{\partial x^{(i)}}, \frac{\partial \varphi_l}{\partial x^{(j)}}\right) - \left(\frac{\partial \varphi_k}{\partial x^{(j)}}, \frac{\partial \varphi_l}{\partial x^{(i)}}\right) &= \sum_{m=1}^{M} \int_{T_m} \left(\frac{\partial \varphi_k}{\partial x^{(i)}} \frac{\partial \varphi_l}{\partial x^{(j)}} - \frac{\partial \varphi_k}{\partial x^{(j)}} \frac{\partial \varphi_l}{\partial x^{(i)}}\right) \mathrm{d}x \\ = \sum_{m=1}^{M} |T_m| \left(\frac{\det \boldsymbol{D}_m^{k,i}}{\det \boldsymbol{D}_m} \frac{\det \boldsymbol{D}_m^{l,j}}{\det \boldsymbol{D}_m} - \frac{\det \boldsymbol{D}_m^{k,j}}{\det \boldsymbol{D}_m} \frac{\det \boldsymbol{D}_m^{l,i}}{\det \boldsymbol{D}_m}\right) \\ = \frac{1}{6} \sum_{m=1}^{M} \frac{\det \boldsymbol{D}_m^{k,i} \det \boldsymbol{D}_m^{l,j} - \det \boldsymbol{D}_m^{k,j} \det \boldsymbol{D}_m^{l,i}}{\det \boldsymbol{D}_m}, \end{aligned} \tag{1.4.10}$$

where $|T_m| = \det \boldsymbol{D}_m/6$ denotes the volume of T_m. After simple but time consuming computation it is possible to show that

$$\frac{\det \boldsymbol{D}_m^{k,i} \det \boldsymbol{D}_m^{l,j} - \det \boldsymbol{D}_m^{k,j} \det \boldsymbol{D}_m^{l,i}}{\det \boldsymbol{D}_m} = x_m^{(\kappa_{ij})} - x_{m+1}^{(\kappa_{ij})}, \quad m = 1,\dots,M, \tag{1.4.11}$$

where $\kappa_{ij} \in \{1,2,3\} \setminus \{i,j\}$. Now, the equalities (1.4.10) and (1.4.11) with the property

$$\sum_{m=1}^{M} (x_m^{(n)} - x_{m+1}^{(n)}) = 0, \qquad n = 1,2,3 \tag{1.4.12}$$

yield the relation (1.4.6). $\square$

Remark 1.7 *The matrix $\boldsymbol{H}$ given by (1.4.4) is symmetric but not positive definite as is required in previous sections. Moreover if $\bar{u}_h$ is linear in Ω_h then $\boldsymbol{d}^{(i,j)} u_h = 0\ \forall i,j = 1\dots,d$ in Ω. Therefore $\|\boldsymbol{e}_k\|_{\boldsymbol{H}_k} = 0\ \forall k \in K$ which corresponds with the relation $r_h \bar{u}_h(x) = u_h(x)\ \forall x \in \Omega \cup \Omega_h$. Consequently $Q_{\mathcal{T}_h} = 2C(d)\omega$ in (1.2.40) for any mesh $\mathcal{T}_h$ of Ω and the iterative process from the Section 1.3 can not be applied for minimization of $Q_{\mathcal{T}_h}$. From these reasons we use instead of $\boldsymbol{H}$ its modification $\boldsymbol{M}$.*

1.4.2 Generation of the matrices M

Let $\boldsymbol{H}(P_i)$, $P_i \in \sigma_h$ be a Hessian matrix (1.4.4) evaluated in the vertex P_i. As $\boldsymbol{H}(P_i)$ is symmetric, we decompose it in the following way:

$$\boldsymbol{H}(P_i) = R_i \begin{pmatrix} \lambda_{i1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_{id} \end{pmatrix} R_i^{-1}, \quad i \in J_h, \tag{1.4.13}$$

where R_i, $i \in J_h$(= a suitable index set) are orthogonal matrices. As matrices $\boldsymbol{H}(P_i)$ are generally not positive definite, we put

$$\bar{\boldsymbol{H}}(P_i) = R_i \begin{pmatrix} |\lambda_{i1}| & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & |\lambda_{id}| \end{pmatrix} R_i^{-1} \quad i \in J_h. \tag{1.4.14}$$

The matrices $\bar{\boldsymbol{H}}(P_i)$, $P_i \in \sigma_h$ are symmetric positive definite. To avoid the problems mentioned in Remark 1.7 we put

$$\boldsymbol{M}(P_i) = c\left[\boldsymbol{I} + \alpha(\|\bar{\boldsymbol{H}}(P_i)\|)\bar{\boldsymbol{H}}(P_i))\right] \quad P_i \in \sigma_h, \tag{1.4.15}$$

where $\boldsymbol{I}$ is a unit $d \times d$ matrix, $c > 0$ is a constant and $\alpha : \langle 0,\infty) \to \langle 0,\infty)$ is a function. In (1.4.15) we put $\|\bar{\boldsymbol{H}}\| = \max_{i,j=1,\dots,d} |\bar{h}_{ij}|$, where $\bar{h}_{ij}$, $i,j = 1,\dots,d$ are the elements of $\bar{\boldsymbol{H}}$. The first term in the square bracket of (1.4.15) guarantees, that the matrices $\boldsymbol{M}(P_i)$ are always regular.

Setting of parameter c

If $\boldsymbol{H}(P_i) = 0\ \forall P_i \in \sigma_h$ then $\boldsymbol{M}(P_i) = c\boldsymbol{I}\ \forall P_i \in \sigma_h$. If the edge-optimal mesh from Definition 1.8 exists, then it consists of equal equilateral elements. As $\boldsymbol{M}(P_i) = c\boldsymbol{I}\ \forall P_i \in \sigma_h$ then (1.2.38) yields

$$\|\boldsymbol{e}_k\|^2_{\boldsymbol{M}(P_i)} = ch^2 = 2C(d)\omega \qquad \forall k \in K, \tag{1.4.16}$$

where h is the length of edges defined in (1.2.17) and then

$$h = \sqrt{\frac{2C(d)\omega}{c}}. \tag{1.4.17}$$

The d-dimensional measure of equilateral element T with the length of edge is h is

$$\mathrm{meas}_d(T) = C_1(d)h^d, \tag{1.4.18}$$

where $C_1(2) = \sqrt{3}/4$ and $C_1(3) = \sqrt{2}/12$. In this case, the edge-optimal mesh $\mathcal{T}_h$ of Ω consists of $N_{\mathcal{T}_h}$ equal equilateral elements, where

$$N_{\mathcal{T}_h} = \frac{\mathrm{meas}_d(\Omega)}{\mathrm{meas}_d(T)}. \tag{1.4.19}$$

This, (1.4.17) – (1.4.19) imply that

$$c = \left(\frac{N_{\mathcal{T}_h} C_1(d)(2C(d)\omega)^{d/2}}{\mathrm{meas}_d(\Omega)}\right)^{2/d} \tag{1.4.20}$$

and

$$h = \left(\frac{\mathrm{meas}_d(\Omega)}{C_1(d)\, N_{\mathcal{T}_h}}\right)^{1/d}. \tag{1.4.21}$$

It means that for a given $N_{\mathcal{T}_h}$ the relation (1.4.15) leads to the same mesh $\mathcal{T}_h$ independently on $\omega > 0$ provided that $\boldsymbol{H}(P_i) = 0\ \forall P_i \in \sigma_h$.

Remark 1.8 *In order to guarantee the regularity of matrices $\boldsymbol{M}$ we have introduced the term $c\boldsymbol{I}$ in the relation (1.4.15). Then the parameter ω has lost the meaning of the accuracy. It indicates how fine should be a mesh provided that $\boldsymbol{H}(P_i) = 0\ \forall P_i \in \sigma_h$. This can be order also by the value $N_{\mathcal{T}_h}$ in (1.4.20) so that the value ω can be fixed. We set $\omega = 1/2$ in (1.4.20) and then $N_{\mathcal{T}_h}$ is the number of elements prescribed by user for the construction of the initial mesh $\mathcal{T}_{h_0}$.*

Setting of α

The simplest way to define α is to put

$$\alpha(\|\bar{\boldsymbol{H}}(P_i)\|) = \hat{C},\ \ \hat{C} = \mathrm{const},\ \ \hat{C} > 0. \tag{1.4.22}$$

When solving problems with discontinuities (e.g. transonic flow with strong shocks), the norm $\|\bar{\boldsymbol{H}}(P_i)\|$ can tend to infinity for some $P_i \in \sigma_h$. Then the drawback of the choice (1.4.22) is that the condition

$$h^d \leq \bar{c}\,\mathrm{meas}(T),\ \ T \in \mathcal{T}_h,\ \ \bar{c} = \text{constant independent of } h, \tag{1.4.23}$$

is not satisfied as $\|\boldsymbol{H}(P_i)\| \to \infty$ for some $P_i \in \sigma_h$. The condition (1.4.23) is the standard requirement for the finite volume/element methods, see e.g. [Fei93]. To fulfill (1.4.23) we propose the relation

$$\alpha(\|\bar{\boldsymbol{H}}(P_i)\|) = \frac{\varepsilon_1}{\varepsilon_2 + \|\bar{\boldsymbol{H}}(P_i)\|}, \quad P_i \in \sigma_h, \tag{1.4.24}$$

where ε_1 and ε_2 are positive constants. The relation (1.4.15) and (1.4.24) give

$$\lim_{\|\bar{\boldsymbol{H}}(P_i)\| \to \infty} \|\boldsymbol{M}(P_i)\| = c(1 + \varepsilon_1), \quad P_i \in \sigma_h, \tag{1.4.25}$$

from which it follows that the norms of $\boldsymbol{M}(P_i),\ P_i \in \sigma_h$ are bounded and therefore the lengths of edges of $\mathcal{T}$ are bounded from below. The relation (1.4.24) is used in further computation.

The parameters $\varepsilon_1, \varepsilon_2$ are determined in the following way. Let $\bar{\boldsymbol{H}}(P_i) = 0$ for some $P_i \in \sigma_h$. Then $\boldsymbol{M}(P_i) = c\boldsymbol{I}$ and the corresponding optimal element has equal edges with length (in Euclidean norm)

$$l_{\max} = \sqrt{\frac{3}{C(d)}}. \tag{1.4.26}$$

Let $\bar{\boldsymbol{H}}(P_i) = \beta\boldsymbol{I}$ for some $P_i \in \sigma_h$ and let $\beta \gg 1$. Then $\boldsymbol{M}(P_i) \to c(1+\varepsilon_1)\boldsymbol{I}$ for $\beta \to \infty$, as follows from (1.4.25) and the corresponding optimal element has equal edges with limit length (in Euclidean norm)

$$l_{\min} = \sqrt{\frac{C(d)}{c(1+\varepsilon_1)}}. \tag{1.4.27}$$

It follows from (1.4.26) and (1.4.27) that the length l of any edge from the edge-optimal mesh satisfies (provided $\boldsymbol{M}(P_i)$ is defined by (1.4.15) and (1.4.24))

$$l_{\min} = \sqrt{\frac{C(d)}{c(1+\varepsilon_1)}} \leq l \leq \sqrt{\frac{C(d)}{c}} = l_{\max}. \tag{1.4.28}$$

It follows from (1.4.26) and (1.4.27) that

$$\frac{l_{\max}}{l_{\min}} = \sqrt{1+\varepsilon_1} \tag{1.4.29}$$

and then

$$\varepsilon_1 = \left(\frac{l_{\max}}{l_{\min}}\right)^2 - 1. \tag{1.4.30}$$

Relation (1.4.30) explains the role of ε_1. Setting ε_1 we control the ratio between the longest and the shortest edge admissible in the edge-optimal mesh.

Based on series of practical computations we use for ε_2 the following relation

$$\varepsilon_2 = \frac{\varepsilon_1}{p} \tag{1.4.31}$$

where p is some positive number. The role of p is viewed in Figure 1.10, where the function

$$\|\bar{\boldsymbol{H}}\| \mapsto \|\boldsymbol{M}\| = c\left(1 + \frac{\varepsilon_1}{\frac{\varepsilon_1}{p} + \|\bar{\boldsymbol{H}}\|}\|\bar{\boldsymbol{H}}\|\right) \tag{1.4.32}$$

is plotted for different p. The number p controls the transition of the coarsest part of the mesh (with edge length l_{max} given by (1.4.26)) to the finest one (with edge length l_{min} given by (1.4.27)) as it is demonstrated on Figure 1.11.

1.4.3 Multilevel computation

The technique described above is applied in the following way, see Figure 1.12: We start from an initial mesh $\mathcal{T}_{h_0}$ of computational domain Ω. We compute the solution u_{h_0} on the mesh $\mathcal{T}_{h_0}$ using a suitable numerical method (finite volume method, finite element method, ...) for approximate solution of the problem (1.2.1) – (1.2.3). We compute the Hessian matrices

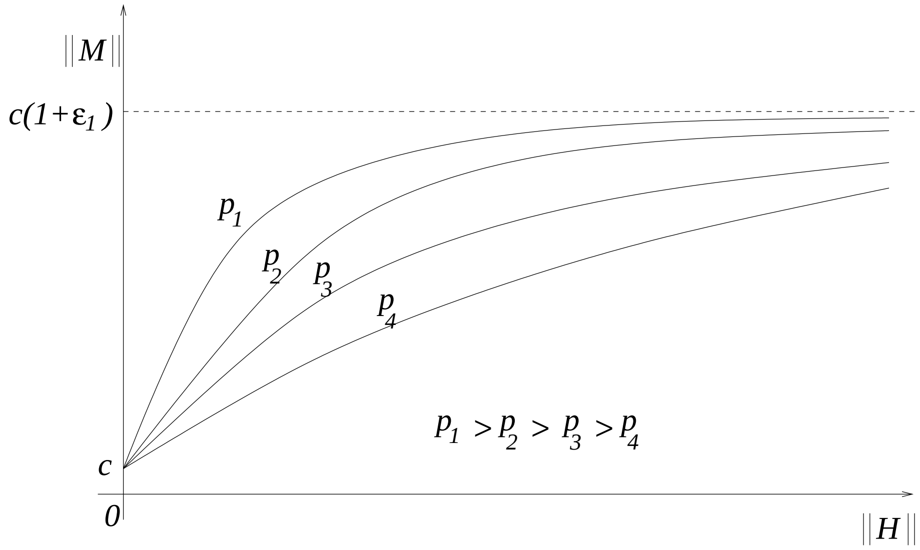


Figure 1.10: Dependence of $\|\boldsymbol{M}\|$ on $\|\bar{\boldsymbol{H}}\|$ for different p.

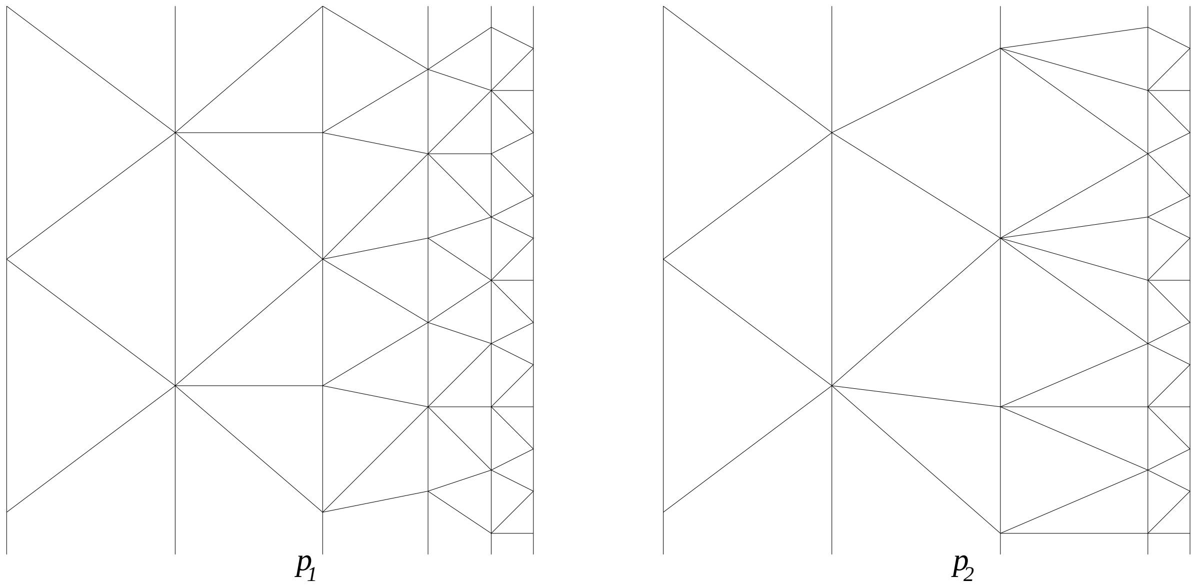


Figure 1.11: Transition of the coarsest part to the finest one for two different $p_1 > p_2$.

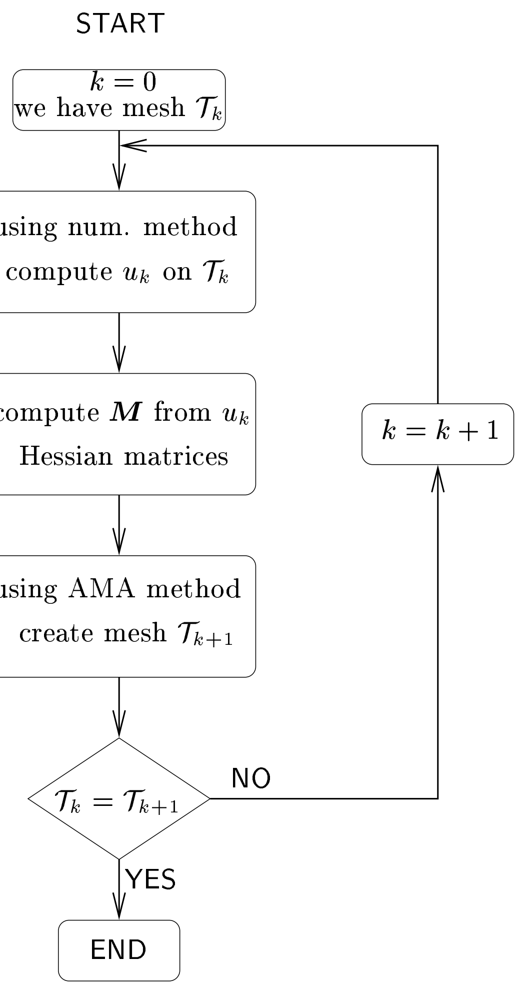


Figure 1.12: Multilevel adaptation AMA algorithm.

$\boldsymbol{M}$ by the technique described above at each vertex of $\mathcal{T}_{h_0}$ and we construct the mesh $\mathcal{T}_{h_1}$ using the iterative process (1.3.5). The results from $\mathcal{T}_{h_0}$ are interpolated onto $\mathcal{T}_{h_1}$ and the numerical solution procedure is applied on $\mathcal{T}_{h_1}$ producing u_{h_1}. In the similar way we construct $\mathcal{T}_{h_2}, u_{h_2}; \mathcal{T}_{h_3}, u_{h_3}; , \ldots$. The multilevel computational process is terminated on the mesh $\mathcal{T}_{h_k}$ if two successive mesh $\mathcal{T}_{h_k}$ and $\mathcal{T}_{h_{k+1}}$ are (approximately) identical. Numerical examples show that for a fixed values of $N_{\mathcal{T}_h}$, ε_1 and p the multilevel computational process terminates, see Section 1.6.3.

1.5 Demonstration of efficiency of AMA

In previous sections, we describe the use of the AMA method in combination with a suitable solver for the numerical solution of BVP. Before presenting practical results, we show several examples demonstrating the efficiency of the proposed adaptive algorithm.

1.5.1 Importance of the stronger necessary condition

At first we show the importance of the stronger necessary condition (1.2.21). Let u be a solution of the given problem. Then the function $\Pi_h u$ introduced in (1.2.11) is the best possible solution from the space U_h in the sense of $\|\cdot\|_X$ norm. Moreover, the value

$$\|u - \Pi_h u\|_X \tag{1.5.1}$$

is the minimal discretization error. Our aim is to show that using anisotropic meshes we can essentially decrease the value of (1.5.1). As the construction of $\Pi_h u$ is difficult, we estimate (1.5.1) by the interpolation error

$$\|u - r_h u\|_X, \tag{1.5.2}$$

where r_h is the ∇-interpolation defined by (1.2.18).

In two examples we investigate (for a given function u) the interpolation error $\|u - r_h u\|_X$ in the norms

$$\|u - r_h u\|_{L^2(\Omega)}, \qquad |u - r_h u|_{H^1(\Omega,\mathcal{T}_h)}, \qquad \|u - r_h u\|_{L^\infty(\Omega)}, \tag{1.5.3}$$

where

$$|u|_{H^1(\Omega,\mathcal{T}_h)} \equiv \left(\sum_{T\in\mathcal{T}_h} \int_T |\nabla u|^2 \,\mathrm{d}x \right)^{1/2} \tag{1.5.4}$$

is the broken H^1 seminorm known from the discontinuous Galerkin method, see e.g. [CKe00]. In (1.5.3), u is a given function and $r_h u$ is defined by (1.2.18) where x_i is a barycentre of $T_i, \mathcal{T}_i \in \mathcal{T}_h$. We compare the values of the interpolation errors (1.5.3) for two types of triangulations: the former uniform triangulation meshes consisting of (almost) equilateral triangles and the latter anisotropic ones obtained by AMA method presented in Section 1.4.

Example 1: Let $\Omega = [-2; 2] \times [-2; 2]$ and

$$u(x_1, x_2) = \frac{1}{2}(100x_1^2 + x_2^2). \tag{1.5.5}$$

The eigenvalues of the Hessian matrix of u are $\lambda_1 = 100,\ \lambda_2 = 1$. Let as consider two sets of triangulations,

$$\left\{\mathcal{T}_{h_1}^{\mathrm{I}},\ \mathcal{T}_{h_2}^{\mathrm{I}},\ \mathcal{T}_{h_3}^{\mathrm{I}},\ \mathcal{T}_{h_4}^{\mathrm{I}}\right\} \quad \text{and} \quad \left\{\mathcal{T}_{h_1}^{\mathrm{A}},\ \mathcal{T}_{h_2}^{\mathrm{A}},\ \mathcal{T}_{h_3}^{\mathrm{A}},\ \mathcal{T}_{h_4}^{\mathrm{A}}\right\}. \tag{1.5.6}$$

The former and the latter sets contain uniform (isotropic) and anisotropic meshes, respectively and

$$\#\mathcal{T}_{h_j}^{\mathrm{I}} \approx \#\mathcal{T}_{h_j}^{\mathrm{A}}, \quad j = 1, \dots, 4, \tag{1.5.7}$$

where $\#\mathcal{T}_h$ denotes the number of triangles of $\mathcal{T}_h$. Figure 1.13 shows meshes $\mathcal{T}_{h_j}^{\mathrm{I}}, \mathcal{T}_{h_j}^{\mathrm{A}},\ j = 1, \dots, 3$.

The computed results are in the following table:

mesh	$\#\mathcal{T}_h$	$\|u - r_h u\|_{L^2(\Omega)}$	$\|u - r_h u\|_{H^1(\Omega,\mathcal{T}_h)}$	$\|u - r_h u\|_{L^\infty(\Omega)}$
$\mathcal{T}_{h_1}^{\mathrm{I}}$	108	4.7131	48.33	8.9073
$\mathcal{T}_{h_1}^{\mathrm{A}}$	110	0.8496	17.37	1.3152
$\mathcal{T}_{h_2}^{\mathrm{I}}$	478	1.1712	24.05	2.6671
$\mathcal{T}_{h_2}^{\mathrm{A}}$	480	0.1828	7.261	0.3496
$\mathcal{T}_{h_3}^{\mathrm{I}}$	1874	0.2770	12.03	0.6582
$\mathcal{T}_{h_3}^{\mathrm{A}}$	1885	0.0462	3.649	0.0930
$\mathcal{T}_{h_4}^{\mathrm{I}}$	7472	0.0713	5.966	0.1697
$\mathcal{T}_{h_4}^{\mathrm{A}}$	7499	0.0116	1.848	0.0259

Example 2: Let $\Omega = [-2;2] \times [-2;2]$ and

$$u(x_1, x_2) = 10(x_1^2 + x_2^2) \exp\left[-10\left((x_1^2 + x_2^2)^{1/2} - 1\right)^2\right]. \tag{1.5.8}$$

This is an example with steep gradients in the neighborhood of a unit circle. Let as consider set of triangulations from previous example

$$\left\{\mathcal{T}_{h_1}^{\mathrm{I}},\ \mathcal{T}_{h_2}^{\mathrm{I}},\ \mathcal{T}_{h_3}^{\mathrm{I}},\ \mathcal{T}_{h_4}^{\mathrm{I}}\right\} \tag{1.5.9}$$

and the new set of anisotropic triangulations

$$\left\{\mathcal{T}_{h_1}^{\mathrm{G}},\ \mathcal{T}_{h_2}^{\mathrm{G}},\ \mathcal{T}_{h_3}^{\mathrm{G}},\ \mathcal{T}_{h_4}^{\mathrm{G}}\right\}. \tag{1.5.10}$$

There is again

$$\#\mathcal{T}_{h_j}^{\mathrm{I}} \approx \#\mathcal{T}_{h_j}^{\mathrm{G}}, \quad j = 1, \ldots, 4. \tag{1.5.11}$$

Figure 1.14 shows meshes $\mathcal{T}_{h_j}^{\mathrm{G}}$, $j = 1, \ldots, 4$. We have investigated the same interpolation errors is in previous example. The results are in the following table:

mesh	$\#\mathcal{T}_h$	$\|u - r_h u\|_{L^2(\Omega)}$	$\|u - r_h u\|_{H^1(\Omega,\mathcal{T}_h)}$	$\|u - r_h u\|_{L^\infty(\Omega)}$
$\mathcal{T}_{h_1}^{\mathrm{I}}$	108	4.2131	39.02	17.22
$\mathcal{T}_{h_1}^{\mathrm{G}}$	114	1.6801	24.43	5.433
$\mathcal{T}_{h_2}^{\mathrm{I}}$	478	0.9241	19.54	4.282
$\mathcal{T}_{h_2}^{\mathrm{G}}$	488	0.3791	11.88	1.664
$\mathcal{T}_{h_3}^{\mathrm{I}}$	1874	0.2371	9.961	1.148
$\mathcal{T}_{h_3}^{\mathrm{G}}$	1880	0.1023	5.887	0.567
$\mathcal{T}_{h_4}^{\mathrm{I}}$	7472	0.0581	4.952	0.350
$\mathcal{T}_{h_4}^{\mathrm{G}}$	7467	0.0243	2.862	0.125

The interpolation errors (1.5.3) are several times smaller for anisotropic meshes than for uniform ones. Consequently, taking into account (1.2.12) and (1.2.19), the disretization error (1.2.8) will be smaller with the use of anisotropic meshes.

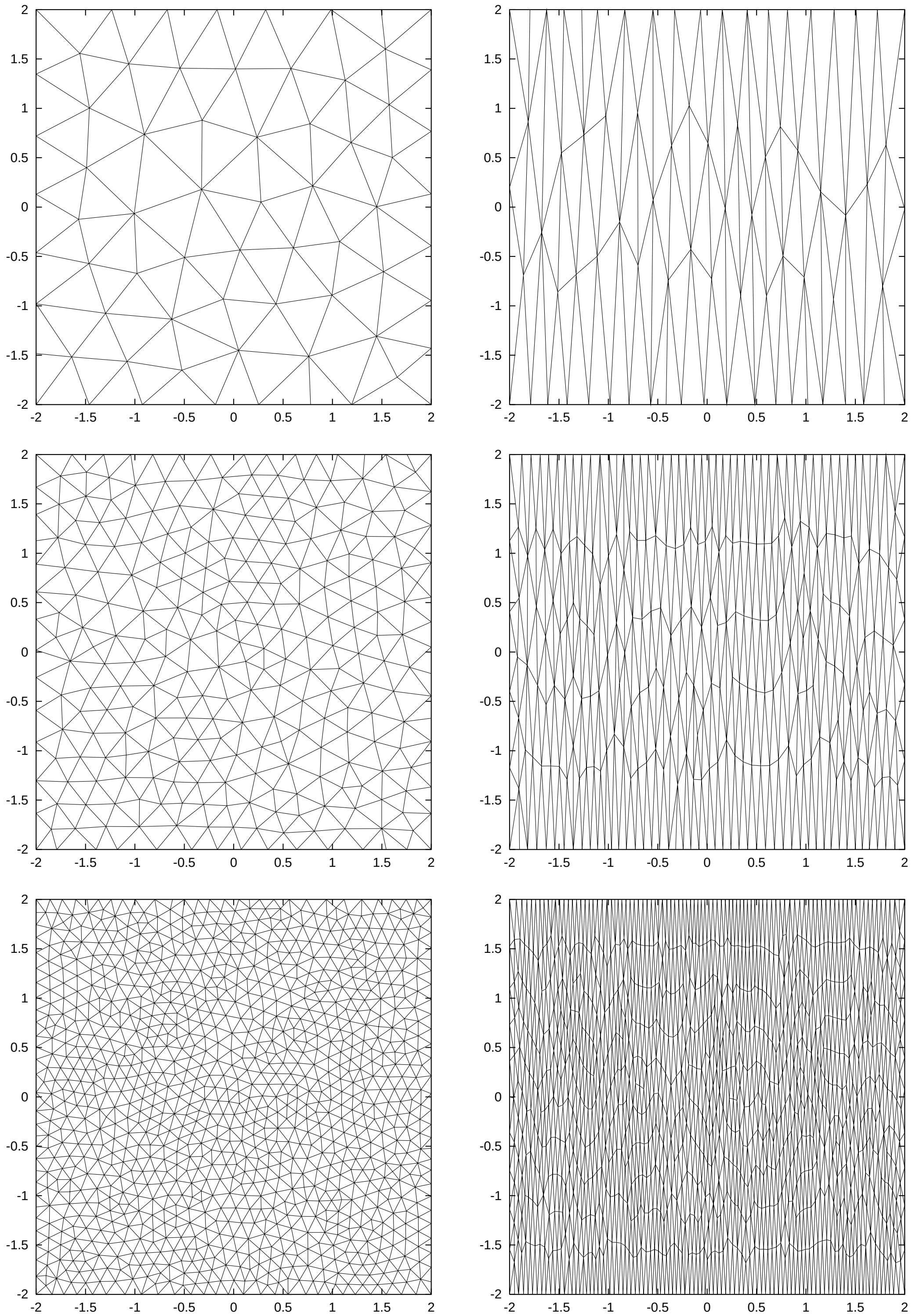


Figure 1.13: Uniform (isotropic) $\mathcal{T}_{h_j}^{\mathrm{I}}$ (left) and anisotropic $\mathcal{T}_{h_j}^{\mathrm{A}}$ (right) meshes for $j = 1, \ldots, 3$.

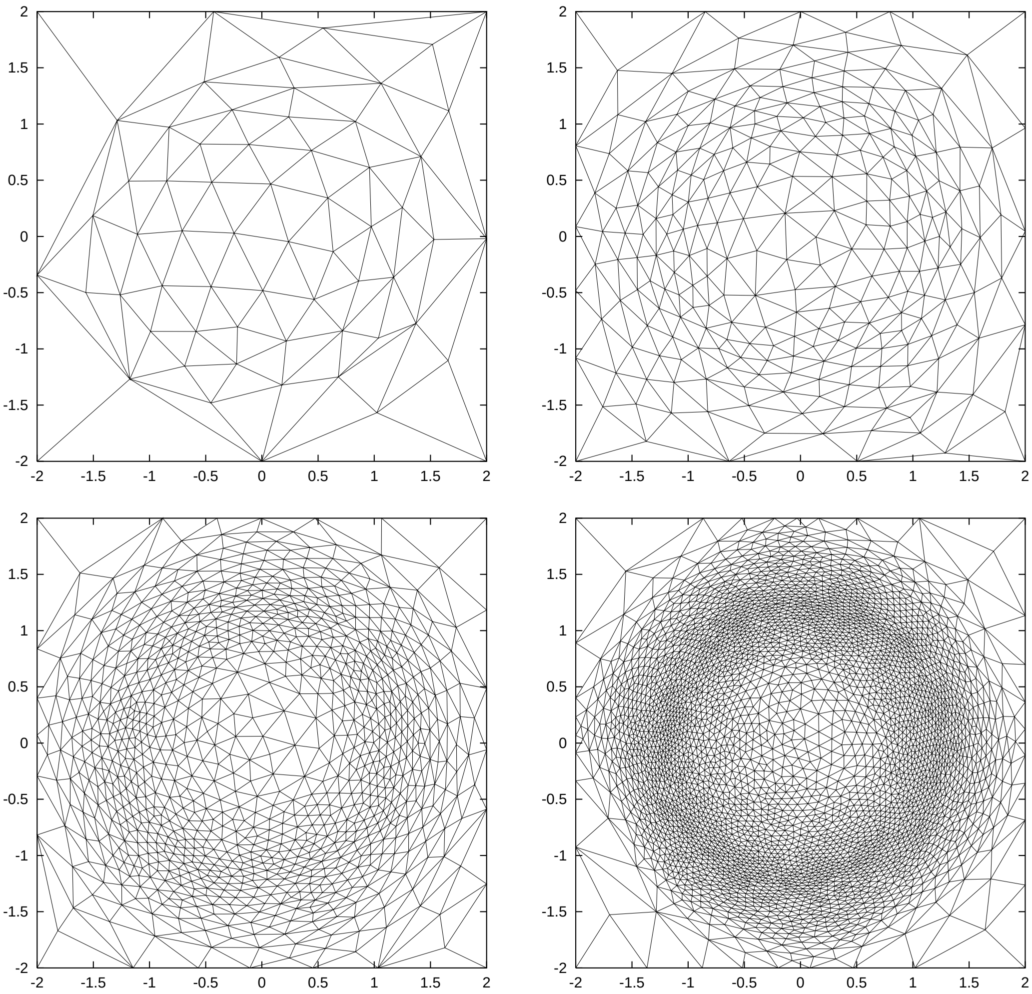


Figure 1.14: Adapted meshes $\mathcal{T}^{\mathrm{G}}_{h_j}$ for $j = 1, \ldots, 4$.

1.5.2 Efficiency of AMA

We apply AMA to the numerical solution of a two dimensional elliptic equation, where the exact solution is known and we demonstrate the practical usefulness of the proposed algorithm. Moreover, in the previous section we described the construction of a new triangulation if the matrices $\boldsymbol{M}(P_i), P_i \in \sigma_h$ are given. To define $\boldsymbol{M}(P_i)$ we use the relations (1.4.13) – (1.4.15), (1.4.20), (1.4.24) and (1.4.31), where the meaning of constants $N_{\mathcal{T}_h}$, ε_1 and p is specified in Section 1.4.2. Their determination will be demonstrated on the two following numerical examples.

In the rest of Section 1.5), following we are interested in the two following items:

- find some suitable algorithm how to choose the constants $N_{\mathcal{T}_h}$, ε_1 and p in order to obtain a suitable meshes for practical computation,
- compare the effectivity of AMA approach with a standard adaptive method.

The value $N_{\mathcal{T}_h}$ represents the number of elements of the initial mesh before the first adaptation. Let $\bar{l}$ be a characteristic length of a given domain. It is sufficient that the step size h of the mesh is z times smaller then $\bar{l}$, where we take z from the interval $\langle 1, 10\rangle$. The d-dimensional measure of the equilateral element (with length of edges h) is equal to $C_1(d)h^d$ where $C_1(2) = \sqrt{3}/4$ and $C_1(3) = \sqrt{2}/12$. Taking into account the relation

$$N_{\mathcal{T}_h} C_1(d) \left(\frac{\bar{l}}{z}\right)^d \approx \mathrm{meas}_d(\Omega), \tag{1.5.12}$$

we put

$$N_{\mathcal{T}_h} := C_1(d)^{-1} \left(\frac{z}{\bar{l}}\right)^d \mathrm{meas}_d(\Omega). \tag{1.5.13}$$

Therefore, we can choose the value $N_{\mathcal{T}_h}$ according to (1.5.13). Numerical experiments show that the choice (1.5.13) is suitable.

To set the constants ε_1 and p we have applied AMA first for scalar elliptic equations (see [BD77]), where the exact solutions are known.

1.5.3 A singular diffusion problem

In [BD77], the following problem is formulated. Find $u : \overline{\Omega} \to I\!R^1$ satisfying

$$\Delta u = 0 \qquad \text{in } \Omega, \tag{1.5.14}$$

where

$$\Omega = \left\{(x_1, x_2) \in I\!R^2, x_1^2 + x_2^2 < 1\right\} \setminus \left\{(x_1, x_2) \in I\!R^2, x_1 \in (0, 1), x_2 = 0\right\} \tag{1.5.15}$$

is the unit circle with a “crack” along the positive x_1 axis. The boundary conditions are $u = \sin(\theta/4)$ along the unit circumference (θ is the angle with respect to the positive x_1 axis), $u = 0$ along the upper surface of the crack, and $\partial u/\partial n = 0$ along the lower surface of the crack. The exact solution of weak problem is

$$u(r, \theta) = r^{1/4} \sin(\theta/4) \tag{1.5.16}$$

(r and θ are the polar coordinates), which belongs to $H^1(\Omega) \setminus H^2(\Omega)$, with a singularity in the origin. If h denotes the longest edge size in the triangulation, it is a classical result ([BD77], p. 4126) that

$$\|u - u_h\| \equiv \|\nabla(u - u_h)\|_{L^2(\Omega)} \leq \beta h^\kappa, \tag{1.5.17}$$

with $\kappa < \frac{1}{4}$ (remember that for regular problems ($u \in H^2(\Omega)$) $\kappa = 1$, see [Fei93], 2.9.41). Here u denotes the exact solution and u_h the piecewise linear finite element numerical solution of (1.5.14).

The fundamental question is the following: *It is possible to increase the convergence rate (1.5.17) with $\kappa < \frac{1}{4}$ if we use the anisotropically adapted grids?* For anisotropic meshes, it is natural to measure the rate of convergence not with respect to the mesh size h but with respect to the number of degrees of freedom, e.g. number of elements of $\mathcal{T}_h$. We define *average mesh size* by

$$h_A \equiv (\#\mathcal{T}_h)^{-1/2}, \tag{1.5.18}$$

where $\#\mathcal{T}_h$ denotes the number of elements of $\mathcal{T}_h$. Obviously, we have for uniform meshes

$$\frac{h}{h_A} \approx \text{constant independent of } h. \tag{1.5.19}$$

Uniform meshes

At first, we have solved (1.5.14) using uniform meshes. In order to generate uniform meshes, we applied the AMA algorithm described in Section 1.2.2 with isotropic matrices, i.e. with $\boldsymbol{M}(P_i) = c\boldsymbol{I},\ P_i \in \sigma_h$, where c is the constant given by (1.4.20), where we prescribe the number of elements $N_{\mathcal{T}_h}$. The equation (1.5.14) is solved by the standard finite element method (FEM) with conforming piecewise linear elements. Then the approximate solution u_h is sought in the space

$$\bar{U}_h = \{u_h \in C(\Omega_h); u_h|_T \in P_1(T)\ \forall T \in \mathcal{T}_h, u_h|_{\Gamma_D} = u_D\}, \tag{1.5.20}$$

where $\Gamma_D \subset \partial\Omega$ is the part of the boundary, where the Dirichlet boundary condition u_D is prescribed.

Let $u_h \in \bar{U}_h$ be the approximate solution of the discrete version of problem (1.5.14) computed on triangulation $\mathcal{T}_h$. As we know the exact solution (1.5.16), we compute the error (1.5.15)

$$e_h \equiv \|u - u_h\| = \left(\sum_{T \in \mathcal{T}_h} \int_T |\nabla(u - u_h)|^2 \,\mathrm{d}x\right)^{1/2}, \tag{1.5.21}$$

where the integral is evaluated by the seven points numerical quadrature exact for the polynomials of degree less or equal to 3.

We define the operator $\bar{\Pi}_h$ using (1.2.11), where we put $\|\cdot\|_X = \|\cdot\|$, i.e. $\bar{\Pi}_h : w \in H^1(\Omega) \mapsto \bar{\Pi}_h w \in \bar{U}_h$ such that

$$\|w - \bar{\Pi}_h w\| = \min_{w_h \in \bar{U}_h} \|w - w_h\|. \tag{1.5.22}$$

Using the standard technique for seeking the minimum of a function of several variables, it is possible to show, that for each $w \in H^1(\Omega)$, there exists a uniquely determined function

$N_{\mathcal{T}_h}$	$\#\mathcal{T}_h$	h	e_h	$e_h^{\min}$	CPU (s)
125	156	0.3079	1.0120	0.8094	7
500	628	0.1540	0.8333	0.6914	9
2000	2442	0.0789	0.7159	0.5958	23
8000	9831	0.0395	0.5837	0.4954	154
32000	39158	0.0199	0.4707	0.4036	2408

Table 1.1: Singular diffusion problem – uniform meshes.

$\bar{\Pi}_h w \in \bar{U}_h$. Then we can compute the *minimal discretization error*

$$e_h^{\min} \equiv \|u - \bar{\Pi}_h u\| = \left(\sum_{T \in \mathcal{T}_h} \int_T |\nabla(u - \bar{\Pi}_h u)|^2 \, \mathrm{d}x \right)^{1/2}. \tag{1.5.23}$$

The function $\bar{\Pi}_h u$ is the *best possible numerical solution* of (1.5.14) – (1.5.15) in the space $\bar{U}_h$ in the sense of the discretization error measured in the $\|\cdot\|$ norm.

Table 1.1 contains the results of computation on five different uniform meshes. The value $N_{\mathcal{T}_h}$ is the number of prescribed elements in (1.4.20) and $\#\mathcal{T}_h$ is the real number of triangles of $\mathcal{T}_h$. The value h in Table 1.1 denotes the step size of a uniform mesh

$$h \equiv \max_{T \in \mathcal{T}_h} \mathrm{diam}\,(T). \tag{1.5.24}$$

The values e_h and $e_h^{\min}$ are given by (1.5.21) and (1.5.23), respectively. The last column introduces the CPU time including the mesh generation process. The computation was performed on a workstation DEC Alpha EV56/500MHz.

Using the data from Table 1.1, we can set the rate of convergence. Let us suppose that

$$e_h = \beta h^{\kappa}, \tag{1.5.25}$$

where e_h is the discretization error given by (1.5.21), h the step size of the mesh given by (1.5.24), β a positive constant and κ the *rate of convergence*. To set κ we apply the *method of the least squares* (for more detail see, e.g., [FD98]). Therefore we obtain $\kappa = 0.28$ and $\beta = 1.49$, which corresponds with the theoretical result (1.5.17).

The smallest value of $e_h = 0.4707$ (computed for $N_{\mathcal{T}_h} = 32000$) is in agreement with low rate of convergence κ. We can easily estimate, how many triangles would be required, so that the discretization error would be significantly small, e.g. $e_h = 0.0567$ (=the smallest error obtained with AMA, see Section 1.5.3, Table 1.2). From the computed values $\kappa = 0.28$ and $\beta = 1.49$ and the relation (1.5.25), we obtain $h = 3.5 \cdot 10^{-15}$, which corresponds approximately to $5.8 \cdot 10^{29}$ equilateral triangles.

Anisotropic meshes

Using the AMA method, we obtain more suitable meshes for the resolution of (1.5.14) in the sense that the minimal discretization error $e_h^{\min}$ is smaller. As we know the exact solution we compute the Hessian matrices exactly, and we compare them with their numerical approximation given by (1.4.3) and (1.4.4). The matrices $\boldsymbol{H}(P_i)$ are given by (1.4.13) – (1.4.14).

$N_{\mathcal{T}_h}$	ε_1	p	$\#\mathcal{T}_h$	e_h	$e_h^{\min}$	CPU (s)
10	10^{10}	50.	759	0.1481	0.1376	122
10	10^{10}	100.	1577	0.1101	0.1022	245
10	10^{10}	150.	2404	0.0992	0.0906	405
10	10^{10}	200.	3194	0.0944	0.0835	610
10	10^{20}	50.	745	0.1378	0.1336	122
10	10^{20}	100.	1610	0.0860	0.0841	247
10	10^{20}	150.	2464	0.0683	0.0673	426
10	10^{20}	200.	3375	0.0567	0.0554	666
10	10^{30}	50.	745	0.1378	0.1336	128
10	10^{30}	100.	1598	0.0942	0.0880	222
10	10^{30}	150.	2452	0.0669	0.0656	400
10	10^{30}	200.	3376	0.0567	0.0554	649

Table 1.2: Singular diffusion problem – anisotropic mesh adaptation.

After simple but time consuming calculation we have

$$\boldsymbol{H}(P_i) = \frac{3}{16} r_i^{-7/4} \boldsymbol{I}, \quad P_i \in \sigma_h, \tag{1.5.26}$$

where $\boldsymbol{I}$ is the unit matrix and r_i is the radius of P_i.

Let $\alpha(\|\bar{\boldsymbol{H}}(P_i)\|)$ be given by (1.4.22). Then the edge-optimal triangulation consists of almost equilateral triangles (provided that $\mathrm{diam}(T) \ll r$), whose areas depend on the radius of their barycentres r_T according

$$\mathrm{meas}(T) \approx \frac{3\sqrt{3}}{4} \left[c \left(1 + \frac{3}{16} \hat{C} r_T^{-7/4} \right) \right]^{-1}. \tag{1.5.27}$$

It follows from (1.5.27) that for $r_T \to 0$ the triangles of the optimal triangulation collapse, i.e. $\mathrm{meas}(T) \to 0$ and the condition (1.4.23) is not satisfied.

On the other hand, let $\alpha(\|\bar{\boldsymbol{H}}(P_i)\|)$ be given by (1.4.24). Then for $\varepsilon_1 \gg 1$ we have

$$\mathrm{meas}(T) \approx \frac{3\sqrt{3}}{4} \left[c \left(1 + \frac{\frac{3}{16} p \varepsilon_1 r_T^{-7/4}}{\varepsilon_1 + \frac{3}{16} p r_T^{-7/4}} \right) \right]^{-1}, \tag{1.5.28}$$

which for $r_T \to 0$ gives

$$\mathrm{meas}(T) \to \frac{3\sqrt{3}}{4} \frac{1}{c(1+\varepsilon_1)}, \tag{1.5.29}$$

and the condition (1.4.23) is satisfied. From this point of view, relation (1.4.24) seems to be better than (1.4.22).

In order to solve (1.5.14), we use the same numerical scheme as for uniform meshes. We performed the mesh adaptation for various combinations of value of ε_1 and p, see Table 1.2. The meaning of values in Table 1.2 is similar as in Table 1.1, $\#\mathcal{T}_h$ denotes the real number of elements after last mesh adaptation and CPU is the time cost for all levels of adaptation including mesh generation.

Figure 1.15 shows the dependence of $h_T \equiv \mathrm{diam}(T)$, $T \in \mathcal{T}_h$ on r_T for the triangulation generated with $\varepsilon_1 = 10^{30}$ and $p = 200$. The separated points present the dependence of the

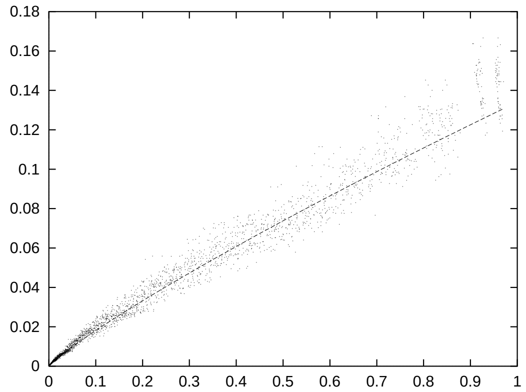


Figure 1.15: Singular diffusion problem – dependence of $h(T)$ on r_T for $\varepsilon_1 = 10^{30}$ and $p = 200$.

ε_1	κ	β
10^{10}	0.637	0.0783
10^{20}	1.172	0.8200
10^{30}	1.221	0.9038

Table 1.3: Singular diffusion problem – anisotropic mesh adaptation, rates of convergence.

value of diameter h_T on corresponding r_T of all $T \in \mathcal{T}_h$, while the dashed line presents the theoretical dependence given by

$$h_T = \frac{4}{\sqrt{3}} \left(\mathrm{meas}(T)\right)^{1/2}, \tag{1.5.30}$$

where meas(T) is given by (1.5.28). The good mutual correspondence is evident.

Using the data from Table 1.2, we set the convergence rate for the anisotropic meshes. From (1.5.19) and (1.5.25), we have

$$e_h = \beta(\#\mathcal{T}_h)^{-\kappa/2}. \tag{1.5.31}$$

In order to set κ, we again apply the method of the least squares. The results are viewed in Table 1.3. We see that the rate of convergence is (for $\varepsilon_1 = 10^{20}$ and $\varepsilon_1 = 10^{30}$) comparable with the regular case ($u \in H^2(\Omega)$), i.e., $\kappa \approx 1$.

Figure 1.16 and Figure 1.17 show two examples of results from Table 1.2. The left figures present the used triangulation and the right ones the corresponding lines of constant values of the numerical solution u_h together with the best possible numerical solution $\bar{\Pi}_h u$. The second and third rows of figures represents the 10 times and 100 times zooms of the neighbourhood of the singularity point, respectively.

Isotropic mesh refinement

We compare AMA technique with a "clasical" adaptive method based on a posteriori error estimation. We define the *error indicator* (compare with (1.5.21))

$$g(T) = \left(\int_T |\nabla(u - u_h)|^2 \, \mathrm{d}x \right)^{1/2}, \quad T \in \mathcal{T}_h, \tag{1.5.32}$$

η	No. of ref.	$\#\mathcal{T}_h$	e_h	$e_h^{\min}$	CPU (s)
$2E-04$	8	1474	0.2175	0.1906	50
$2E-04$	10	1538	0.1658	0.1485	66
$2E-04$	12	1602	0.1330	0.1225	82
$1E-04$	8	2127	0.2121	0.1843	78
$1E-04$	10	2181	0.1585	0.1403	102
$1E-04$	12	2255	0.1237	0.1122	128

Table 1.4: Singular diffusion problem – red-green refinement.

where u and u_h are the exact and numerical solutions, respectively. The error indicator (1.5.32) is the "best possible" a posteriori error indicator as it is based on the direct computation of the error (the exact solution is known).

All triangles $T \in \mathcal{T}_h$ with $g(T)$ greater than a given tolerance η are refined by the red-green technique [BSW83]. The computation was performed for 2 values of the tolerance η and for 8, 10 and 12 level of the mesh refinement. It is necessary to prescribe the maximal number of levels of refinement, as the singularity causes that there exist some triangles with $g(T) > \eta$ for any level of refinement. The achieved results are viewed in Table 1.4 and one of them is visualized in Figure 1.18.

1.5.4 A Poisson problem with steep gradients

The previous test, though essential for checking the convergence properties of the adaptive algorithm, allows no evaluation of the performance of the optimizer in relation to the anisotropy. For problem (1.5.14), both eigenvalues of the Hessian matrix of the solution (1.5.16) differ only on sign (i.e. $\lambda_1 = -\lambda_2$) and thus the optimal triangulation consists of equilateral triangles.

Consider now a diffusion problem exhibiting steep gradients. Let $\Omega = (0,1) \times (0,1)$, and let u be the solution of

$$-\Delta u(x_1, x_2) = 90x_1^8(1 - x_2^{20}) + 380x_2^{18}(1 - x_1^{10}) \tag{1.5.33}$$

with boundary conditions $u = 0$ along the edges $\{(x_1, x_2) \in I\!R^2; x_1 = 1,\ x_2 \in (0,1)\}$ and $\{(x_1, x_2) \in I\!R^2; x_1 \in (0,1),\ x_2 = 1\}$, and $\partial u/\partial n = 0$ along $\{(x_1, x_2) \in I\!R^2; x_1 = 0,\ x_2 \in (0,1)\}$ and $\{(x_1, x_2) \in I\!R^2; x_1 \in (0,1),\ x_2 = 0\}$. The exact solution

$$u = (1 - x_1^{10})(1 - x_2^{20}) \tag{1.5.34}$$

exhibits two steep gradients along the right and top edges, the latter being stronger than the former.

Uniform meshes

The same numerical method as in previous section was applied. Table 1.5 contains the results of computation on four different uniform meshes. In contrast to the problem (1.5.14) – (1.5.15), the discretization error e_h of the problem (1.5.33) (given by (1.5.21)) differs from the minimal discretized error $e_h^{\min}$ (given by (1.5.23)) by the terms of order 10^{-4}. It is caused by the fact, that the problem (1.5.33) has a smooth solution (1.5.34) and the used numerical

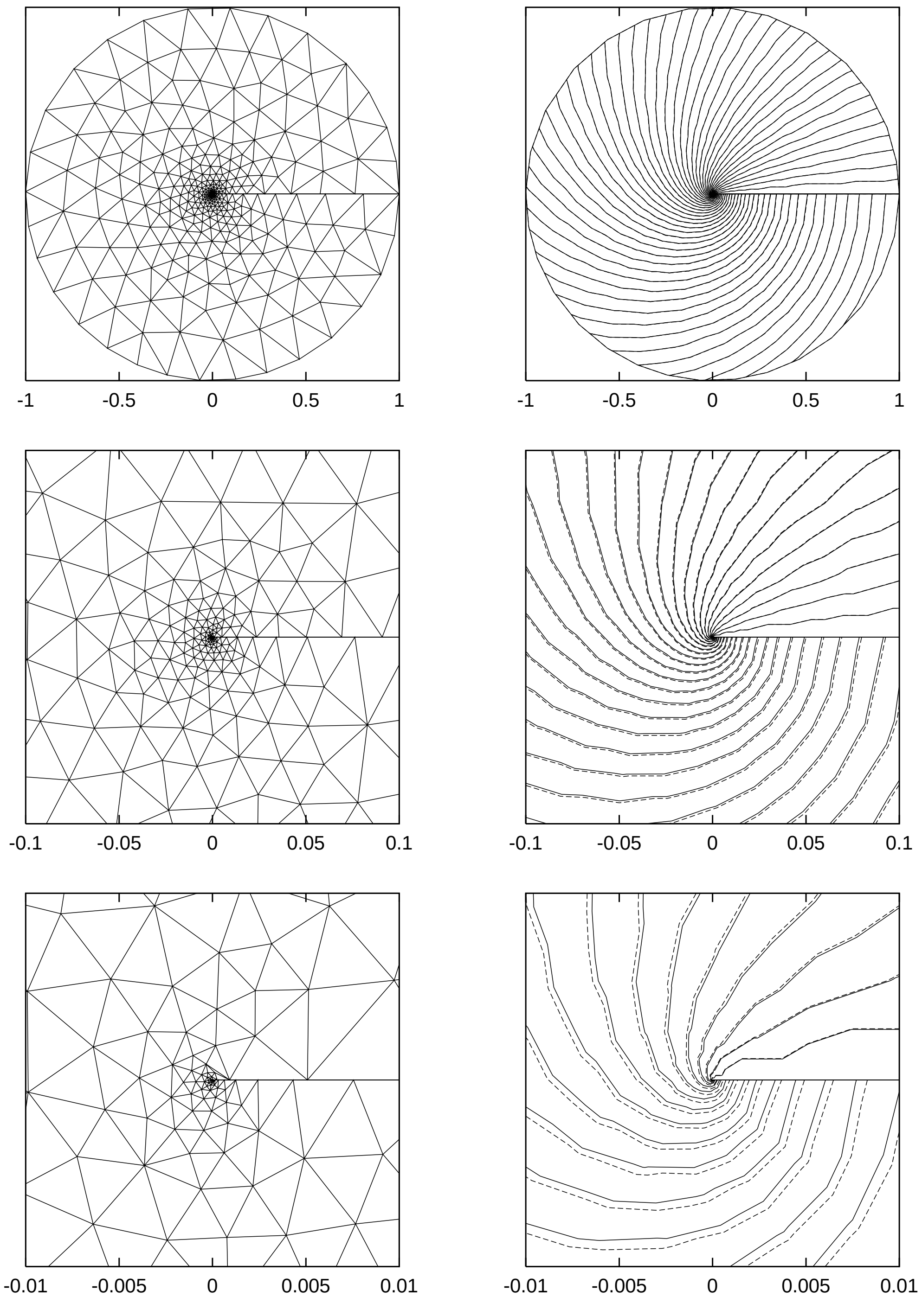


Figure 1.16: Singular diffusion problem, anisotropic mesh adaptation with $N_{\mathcal{T}_h} = 10$, $\varepsilon_1 = 10^{10}$, $p = 50$ ($\#\mathcal{T}_h = 759$) – mesh and isolines of u_h (full line) and $\bar{\Pi}_h u$ (dashed line) with zooms.

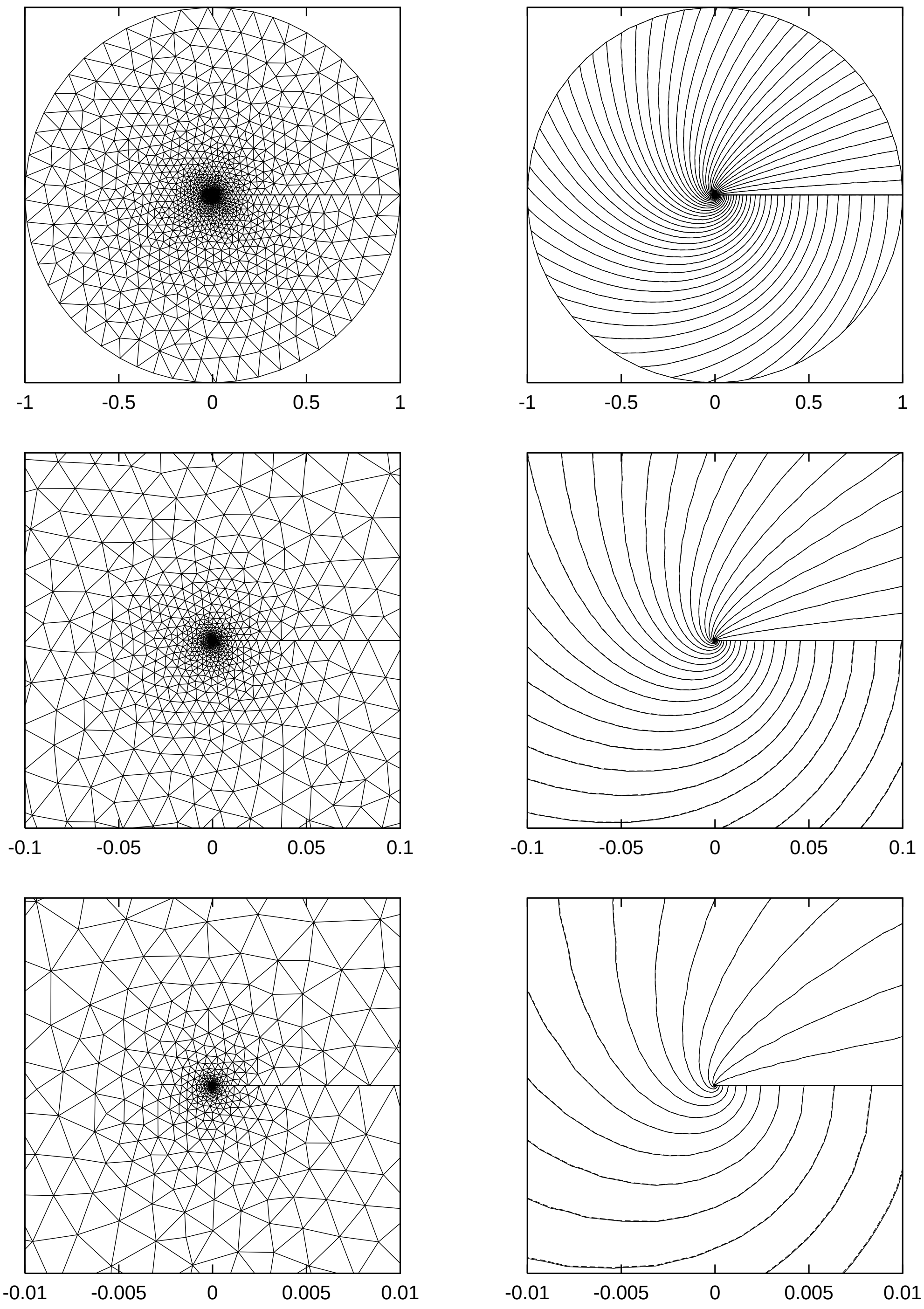


Figure 1.17: Singular diffusion problem, anisotropic mesh adaptation with $N_{\mathcal{T}_h} = 10$, $\varepsilon_1 = 10^{30}$, $p = 200$ ($\#\mathcal{T}_h = 3376$) – mesh and isolines of u_h (full line) and $\bar{\Pi}_h u$ (dashed line) with zooms.

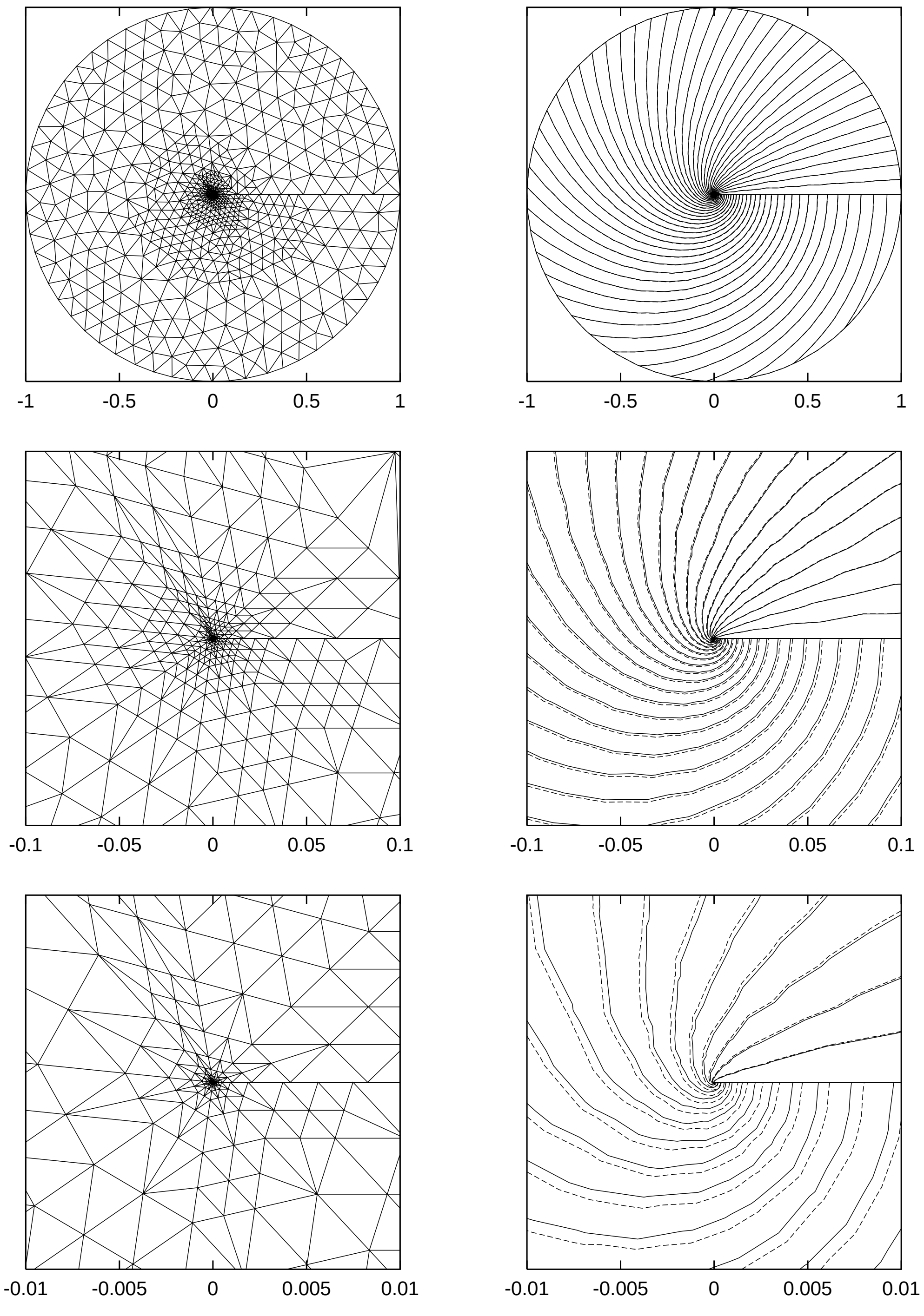


Figure 1.18: Singular diffusion problem, red-green refinement with $\eta = 2 \cdot 10^{-4}$ after 12 level of refinement ($\#\mathcal{T}_h = 1602$) – mesh and isolines of u_h (full line) and $\bar{\Pi}_h u$ (dashed line) with zooms.

$N_{\mathcal{T}_h}$	$\#\mathcal{T}_h$	h	e_h	CPU (s)
125	127	0.1783	1.9281	7
500	546	0.0894	0.9905	9
2000	2314	0.0484	0.5340	29
8000	9332	0.0226	0.2655	252
32000	18950	0.0132	0.1346	87904

Table 1.5: Poisson problem with steep gradients - uniform meshes.

$N_{\mathcal{T}_h}$	ε_1	p	$\#\mathcal{T}_h$	e_h	CPU (s)
1	10^{10}	100	361	0.18392	79
1	10^{20}	100	361	0.18392	79
1	10^{30}	100	361	0.18392	77
1	10^{20}	300	941	0.10783	152
1	10^{10}	300	945	0.10781	153
1	10^{30}	300	941	0.10783	146
1	10^{10}	600	1741	0.07753	287
1	10^{20}	600	1734	0.07666	295
1	10^{30}	600	1734	0.07666	301
1	10^{10}	1000	2883	0.06475	593
1	10^{20}	1000	2908	0.06426	581
1	10^{30}	1000	2908	0.06426	586

Table 1.6: Poisson problem with steep gradients – anisotropic meshes.

scheme gives solution u_h almost identical with $\bar{\Pi}_h u$. Therefore Table 1.5 – Table 1.7 do not contain the value $e_h^{\min}$ and Figure 1.19 - Figure 1.20 do not contains the isolines of $\bar{\Pi}_h u$.

If we again suppose the relation (1.5.25) we determine the rate of convergence as $\kappa = 1.007$ ($\beta = 11.19$), which corresponds to the theoretical results (1.5.17) ($\kappa = 1$ for regular problems).

Anisotropic meshes

Table 1.6 contains results of the solution of (1.5.33) on meshes achieved by AMA (all quantities have the same meaning as in Section 1.5.3). Figure 1.19 shows one example of results from Table 1.6, where the triangulation and the corresponding contours of constant value of the numerical solution u_h are plotted. The expected grid alignment is easily viewed.

Isotropic mesh refinement

We have again applied the red-green refinement method introduced in Section 1.5.3. In this case, it is possible to apply the mesh refinement till the error indicator $g(T)$ is under the tolerance η for each triangle. The computation was performed for 3 values of the tolerance η. The achieved results are viewed in Table 1.7 and one of them visualized Figure 1.20.

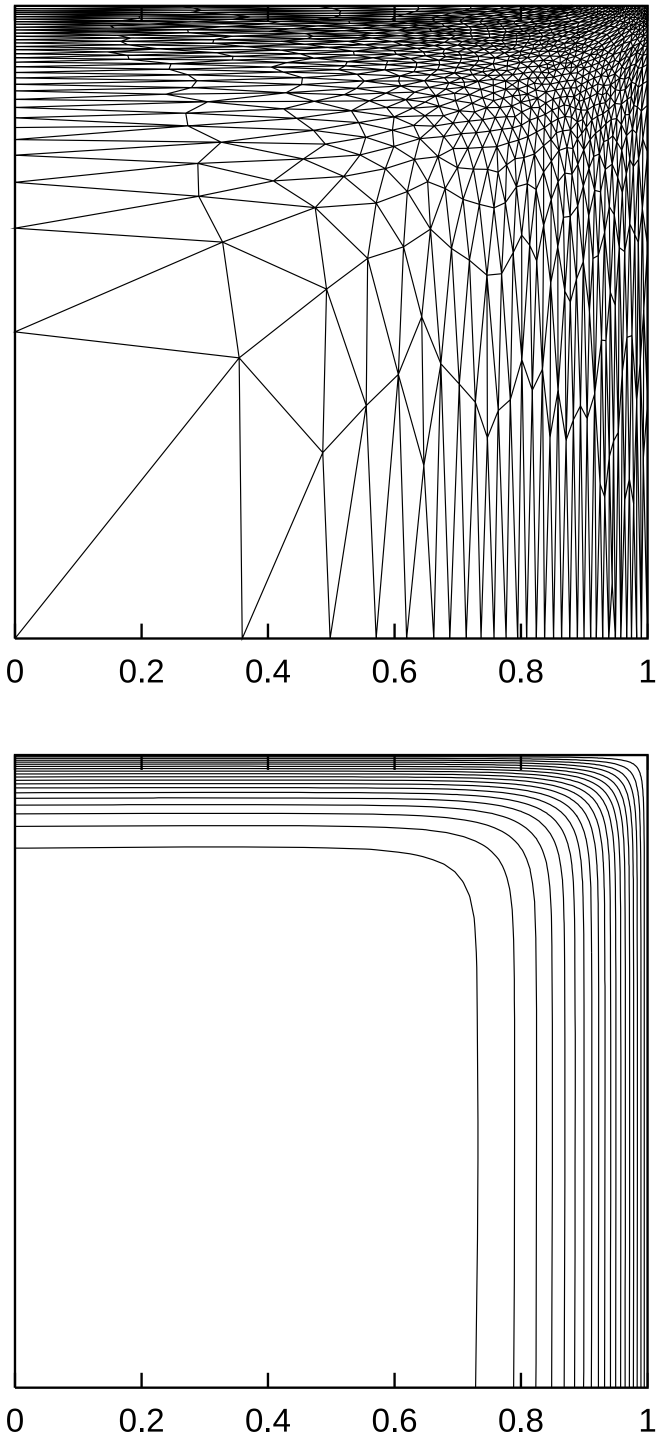
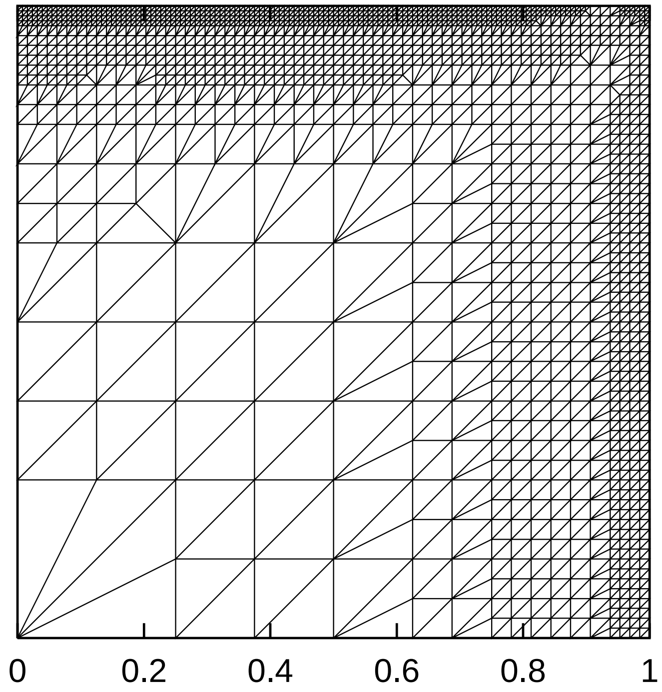


Figure 1.19: Poisson problem with steep gradients - anisotropic mesh adaptation: $N_{\mathcal{T}_h} = 1$, $\varepsilon_1 = 10^{30}$ and $p = 600$ ($\#\mathcal{T}_h = 1734$) - mesh and isolines of u_h.



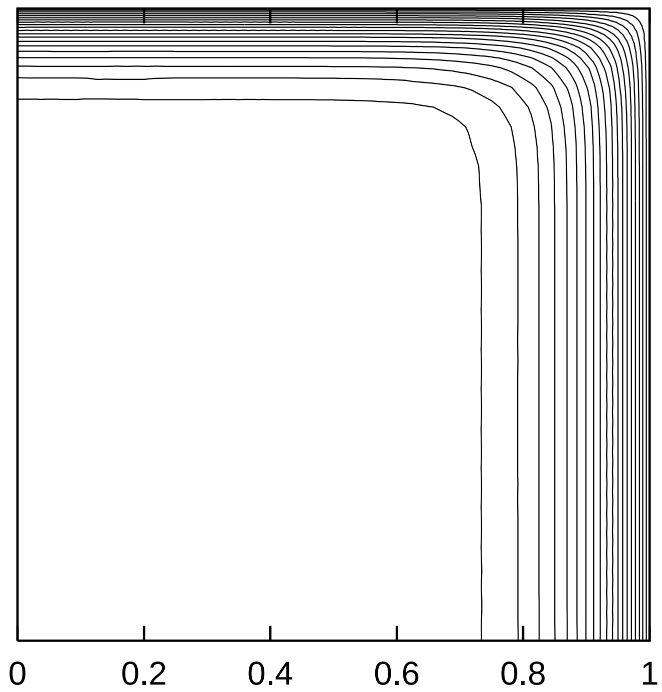


Figure 1.20: Poisson problem with steep gradients – red-green refinement with $\eta = 10^{-5}$ and $p = 1.0$ ($\#\mathcal{T}_h = 8583$) - mesh and isolines of u_h.

η	No. of ref.	$\#\mathcal{T}_h$	e_h	$e_h^{\min}$	CPU (s)
10^{-3}	5	1041	0.3781	0.1906	35
10^{-4}	6	2778	0.2343	0.1485	49
10^{-5}	7	8583	0.1336	0.1225	175

Table 1.7: Poisson problem with steep gradients – red-green refinement.

1.5.5 Discussion of the results

From the results presented in Sections 1.5.3 and 1.5.4, we can conclude the following:

- The computations show that the discretization error e_h (1.5.21) is close to $e_h^{\min}$ (1.5.23) for both uniform and adapted meshes. This is given by the fact, that the used numerical method (conforming linear finite element method) gives sufficiently accurate solution for a given triangulation. This underlines the importance of the necessary condition (1.2.13).
- The global mesh refinement leads to an enormous requirement on the memory and CPU-time. It is absolutely impractical.
- Applying AMA for singular diffusion problem (with nonregular exact solution – $u \in H^1(\Omega) \setminus H^2(\Omega)$), we achieve the rate of convergence $\kappa \approx 1$, which is the theoretical results for regular case ($u \in H^1(\Omega)$).
- In order to achieve the same level of the computational error, the grids adapted by AMA have a smaller number of elements than the grids adapted by the red-green refinement (RGR). The difference is remarkable namely for the second example.
- On the other hand, the CPU-time is higher for AMA than for RGR. It is caused by the simplicity of considered problem – scalar linear equation. Here the mesh adaption algorithm requires more time than the numerical solution of the Poisson equation. The top efficiency of AMA plays a role in practical problems, where the CPU-time rests the same for the mesh adaptation but it is much higher for the numerical solution of the considered problem.
- Observing results from Table 1.2 and Table 1.6, we see that whereas the increase of ε_1 do not cause essential improvement of results, then the increase of p leads to the decrease of the computational error (and naturally the increase of the number of elements). So that we propose the following strategy how to set the parameters $N_{\mathcal{T}_h}$, ε_1 and p:
  1. prescribe $N_{\mathcal{T}_h}$ according to (1.5.13)
  2. set ε_1 sufficiently high, e.g. 10^{20}, see (1.4.30)
  3. start with a small p and increase successively its value till the number of elements is acceptable from the point of view of the computer memory and CPU time.

1.6 Application of AMA for compressible flow simulation

In this section, we present the use of AMA for the solution of two and three dimensional fluid dynamics problems. We deal with the simulation of inviscid and viscous compressible flow, i.e., we seek a numerical solution of the system of the Euler and Navier-Stokes equations.

The application of AMA for inviscid flow simulation is straightforward, see Section 1.6.3. On the other hand, the sufficient accurate viscous flow simulation requires additional mesh adaptation techniques which are determined from the physical properties of the flow, namely the presence of boundary layers and wakes.

At first, we present the system of the Navier-Stokes equations and we briefly describe the used numerical method for its solution, which we developed several years ago. Then the application for inviscid and viscous flow are given.

1.6.1 Governing equations

The complete system describing the motion of compressible viscous flow in a bounded domain $\Omega \subset I\!R^d$ and time interval $(0, \boldsymbol{T})$ is written in the dimensionless form

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{s=1}^{d} \frac{\partial \boldsymbol{f}_s(\boldsymbol{w})}{\partial x_s} = \sum_{s=1}^{d} \frac{\partial \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w})}{\partial x_s} \quad \text{in } Q_{\boldsymbol{T}} = \Omega \times (0, \boldsymbol{T}), \tag{1.6.1}$$

where

$$\begin{aligned}
\boldsymbol{w} &= (\rho, \rho v_1, \dots, \rho v_d, e)^T, \ \boldsymbol{w} = \boldsymbol{w}(x,t), \quad x \in \Omega, \ t \in (0, \boldsymbol{T}), \\
\boldsymbol{f}_s(\boldsymbol{w}) &= (\rho v_s, \rho v_s v_1 + \delta_{s1} p, \dots, \rho v_s v_d + \delta_{sd} p, (e+p)\, v_s)^T, \ s = 1, \dots, d \\
\boldsymbol{R}_s(\boldsymbol{w}) &= (0, \tau_{s1}, \dots, \tau_{sd}, \sum_{r=1}^{d} \tau_{sr} v_r + q_s)^T, \ s = 1, 2 \\
\tau_{sr} &= \frac{1}{Re}\left(\frac{\partial v_s}{\partial x_r} + \frac{\partial v_r}{\partial x_s} - \frac{2}{3} \mathrm{div}\, \boldsymbol{v}\, \delta_{sr}\right), \ q_s = \frac{\gamma}{Re\, Pr} \frac{\partial \theta}{\partial x_s}, \ s, r = 1, \dots, d \\
p &= (\gamma - 1)\,(e - \rho|\boldsymbol{v}|^2/2), \ e = \rho(\theta + |\boldsymbol{v}|^2/2).
\end{aligned} \tag{1.6.2}$$

Notation is the following: t – time coordinate, $x = (x_1, \dots, x_d)$ – Cartesian coordinates, ρ – density, p – pressure, θ – temperature, e – total energy, $\boldsymbol{v} = (v_1, \dots, v_d)$ – velocity, δ_{ij} – Kronecker delta, γ – Poisson adiabatic constant, Pr – Prandtl number, Re – Reynolds number. We assume that $\gamma > 1$, Pr and Re are given constants.

The above system is equipped with the initial condition $\boldsymbol{w}(x, 0) = \boldsymbol{w}_0(x)$, $x \in \Omega$ and boundary conditions, see [FJS]: At inlet we prescribe ρ, $v_1, \dots, v_d$ and use "do-nothing" condition $\sum_{r=1}^{d} \left(\sum_{s=1}^{d} \tau_{sr} n_s\right) v_r + \frac{\gamma}{Re\, Pr} \partial\theta/\partial n = 0$, on fixed walls we assume that $v_1 = \dots = v_d = 0$, $\partial\theta/\partial n = 0$ and at outlet we set $-p n_j + \sum_{i=1}^{d} \tau_{sr} n_s = 0$, $s = 1, \dots, d$, and $\partial\theta/\partial n = 0$. (Here $\partial/\partial n$ is the derivative in the direction of the unit outer normal $\boldsymbol{n} = (n_1, \dots, n_d)$ to the boundary $\partial\Omega$.)

1.6.2 Numerical methods

Our aim is to obtain the steady-state solution of the problem (1.6.1) – (1.6.2) by a time marching method. For a flow with a high Reynolds number, the viscous effects are small, and therefore we can consider the viscous terms $\boldsymbol{R}_s,\ s = 1, \dots, d$ as a perturbation of the inviscid system (the Euler equations). We split (1.6.1) into the inviscid system

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{s=1}^{d} \frac{\partial \boldsymbol{f}_s(\boldsymbol{w})}{\partial x_s} = 0, \tag{1.6.3}$$

and purely viscous system

$$\frac{\partial \boldsymbol{w}}{\partial t} = \sum_{s=1}^{d} \frac{\partial \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w})}{\partial x_s} \tag{1.6.4}$$

and discretize them separately.

The time discretization of (1.6.3) and (1.6.4) is carried out with the use of a partition $0 = t_0 < t_1 < t_2 \dots < t_N = T$ of the time interval $(0, T)$. The inviscid system (1.6.3) is discretized by the cell-centered *finite volume* (FV) *method* on $\mathcal{D}_h = \{D_i\}_{i \in J}$. Here D_i are the finite volumes (which can be triangular, barycentric and dual finite volumes or their 3D analogues), $\mathcal{D}_h$ is a partition of a polygonal approximation Ω_h of the domain Ω and J is a suitable index set. The purely viscous system (1.6.4) is discretized by the piecewise linear conforming or nonconforming *finite element* (FE) *method* on a triangulation ($d = 2$) or a tetrahedrization ($d = 3$) $\mathcal{F}_h$ of Ω_h, compatible with $\mathcal{D}_h$ in some sense. Then the complete system (1.6.1) is discretized via *operator inviscid-viscous splitting* (see e. g. [DA96], [FF96], [FFD96]). One time step $t_k \to t_{k+1}$ consists of two fractional substeps: inviscid FV step on the mesh $\mathcal{D}_h$ and viscous FE step on the mesh $\mathcal{F}_h$.

1. *Inviscid FV step on $\mathcal{D}_h$*:
   Assume that the values $\boldsymbol{w}_i^k,\ i \in J$, approximating the solution on the finite volumes D_i at time t_k are known. Compute the values $w_i^{k+1/2},\ i \in J$, from the FV formula

$$\boldsymbol{w}_i^{k+1/2} = \boldsymbol{w}_i^k - \frac{\tau_k}{|D_i|} \sum_{j \in S(i)} H(\boldsymbol{w}_i^k, \boldsymbol{w}_j^k, \boldsymbol{n}_{ij})\, |\Gamma_{ij}|, \tag{1.6.5}$$

   where $S(i)$ denotes the set of neighbours volumes of D_i and Γ_{ij} denotes the common face between D_i and D_j.

2. *Viscous FE step on $\mathcal{F}_h$*:
   Define the finite element function $\boldsymbol{w}_h^{k+1/2}$ with values $\boldsymbol{w}_h^{k+1/2}(P_i) = \boldsymbol{w}_i^{k+1/2}$ at the vertices $P_i,\ i \in J$ of $\mathcal{F}_h$. At the vertices $P_i \in \partial\Omega_h$, the viscous Dirichlet boundary conditions and suitable extrapolation are used. Compute the finite element function $\boldsymbol{w}_h^{k+1}$ as the solution of the following problem:

   (a) $\boldsymbol{w}_h^{k+1}$ satisfies the viscous Dirichlet boundary conditions,

   (b)

$$(\boldsymbol{w}_h^{k+1}, \varphi_h)_h = (\boldsymbol{w}_h^{k+1/2}, \varphi_h)_h - \tau_k\, a_h(\boldsymbol{w}_h^{k+1/2}, \varphi_h) \tag{1.6.6}$$

   for all test functions $\varphi_h = (\phi_1, ..., \phi_4)$ such that ϕ_j $(j = 1, ..., 4)$ is continuous in Ω, linear on each $T \in \mathcal{F}_h$ and vanishes on the part of $\partial\Omega$, where the j-th component w_j of the state vector $\boldsymbol{w}$ satisfies the Dirichlet boundary condition.

In (1.6.5), H is the numerical flux approximating

$$H(\boldsymbol{w}_i^k, \boldsymbol{w}_j^k, \boldsymbol{n}_{ij}) \approx \sum_{s=1}^{d} f_s(\boldsymbol{w}(x, t_k))(\boldsymbol{n}_{ij})_s, \tag{1.6.7}$$

where $(\boldsymbol{n}_{ij})_1 \ldots, (\boldsymbol{n}_{ij})_d$ are the components of the unit outer normal $\boldsymbol{n}_{ij}$ on Γ_{ij} pointed out of D_i. For the evaluation of the numerical flux, the Osher-Solomon scheme [OS82], [FŠ98] was used.

In (1.6.6), the forms $(\boldsymbol{w}, \varphi)_h$ and $a_h(\boldsymbol{w}, \varphi)$ denote the approximations

$$\begin{aligned} (\boldsymbol{w}, \varphi)_h &\approx \int_\Omega \boldsymbol{w}\, \varphi \, dx, \\ a_h(\boldsymbol{w}, \varphi) &\approx \int_\Omega \sum_{s=1}^{d} \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w})\, \partial\varphi/\partial x_s \, dx, \end{aligned} \tag{1.6.8}$$

which were obtained with the aid of numerical quadrature using the vertices of $T \in \mathcal{F}_h$ as integration points.

We developed three different methods. The first method uses the dual finite volume and conforming FE (see [FFL95]), the second one uses triangular FV and conforming FE (see [FFD96]) and the third one uses the barycentric FV and nonconforming FE (see [DFFK02]).

1.6.3 AMA for inviscid flow simulation

The motion of inviscid compressible flow is described by the system of the Euler equations (1.6.3). The numerical solution is performed by the finite volume method, which is in fact the first step of the method presented in Section 1.6.2, i.e., *Inviscid FV step on the mesh* $\mathcal{D}_h$. All results in Section 1.6.3 was computed by the FV method on triangular or tetrahedral meshes, so that we put $\mathcal{D}_h := \mathcal{T}_h$, where $\mathcal{T}_h$ was obtained by AMA technique.

In order to apply AMA for inviscid flow simulation, we chose the density of the flow ρ as the *significant* quantity. In order to compute the Hessian matrices we simply put $u_h := \rho_h$ in (1.4.4). Then the whole computational process from Section 1.4.3 is applied.

AMA technique is applied for three 2D benchmark cases and two 3D problems. The first example deals with the transonic flow around NACA0012 profile. It shows that the multilevel computational process described in Section 1.4.3 terminates after a finite number of levels of adaptation. The second one compares the AMA technique with two other adaptive methods for the case of the flow through the forward facing step. The third example is a benchmark with complicated geometry of shock waves. The last two examples show the application of AMA for threedimensional problems.

Our aim is to obtain a steady state solution (except 2D and 3D forward facing steps, which are unsteady). It is achieved by a time stabilization technique for $\boldsymbol{T} \to \infty$, i.e. we apply the computational process till the following residuum criterion is satisfied

$$\frac{1}{\tau_k} \left\| \frac{\rho^{k+1} - \rho^k}{\rho^k} \right\|_{L^1(\Omega)} < \mathsf{TOL}, \tag{1.6.9}$$

where ρ^k is the value of the density considered on the time level t_k, τ_k is the time step and $\mathsf{TOL} > 0$ is a given tolerance. All computations were performed on a workstation DEC Alpha EV56/500MHz.

k = level of adaptation	$\#\mathcal{T}_{h_k}$	$\Delta\#\mathcal{T}_{h_k}$	CPU-time (s) adaptation	solver
0	714	–	54	156
1	1483	769	167	289
2	7472	5989	192	1545
3	12254	4785	185	2689
4	14522	2268	231	3352
5	15619	1097	168	3521
6	16270	651	172	3683
7	16525	255	154	3765

Table 1.8: NACA0012, mesh adaptation levels, $\#\mathcal{T}_{h_k}$ = number of elements of $\mathcal{T}_{h_k}$, $\Delta\#\mathcal{T}_{h_k} \equiv \#\mathcal{T}_{h_k} - \#\mathcal{T}_{h_{k-1}}$ and CPU-times for mesh adaptation and computation of the Euler equations itself.

Transonic flow around NACA 0012 profile

We present the simulation of the inviscid flow around NACA 0012 profile with the inlet Mach number ($= |\boldsymbol{v}|/(\gamma p/\rho)^{1/2}$) $M_{\rm in} = 0.8$ and angle of attack $\alpha = 1°25'$. We performed seven levels of AMA algorithm with $c = 200$, $\varepsilon_1 = 10^6$ and $p = 1000$. Table 1.8 contains the number of elements $\#\mathcal{T}_{h_k}$, $k = 0, \dots, 7$ of all levels of adaptation and moreover CPU times needed for the mesh adaptation (left column) and for the solution of the Euler equations itself (right column). Figures 1.21–1.28, show the triangulations and the isolines of the Mach number of all levels of adaptation. Wee see that from very poor results obtained for the initial triangulation, we obtain good results with very sharp shock waves. Figures 1.29 show the distributions of the Mach number along the profile. Finally, the convergence of residuum is plot at Figure 1.30.

The results for the last two levels of refinement (Figure 1.27 and Figure 1.28) are almost identical, which corresponds to the assertion that the AMA algorithm “converges” (see Section 1.4.3). This is also viewed in Table 1.8. The results from “old” mesh are interpolated on the new one in order to save the CPU time. It is vied at Figure 1.30 that for these cases the residuum decreases very fast.

Transonic flow through the forward facing step

We deal with the numerical simulation of inviscid flow through a classical benchmark example – the forward facing step, see [WC84]. The geometry of the problem is viewed in Figure 1.31. The computations of the inviscid Euler equations were performed with the initial conditions: $\rho_0 = 1.4$, $\boldsymbol{v} = (3, 0)$, $p = 1$ and the boundary conditions taken from the initial conditions at inlet and outlet part of $\partial\Omega$. On the impermeable walls, we naturally put $v_n \equiv \boldsymbol{v} \cdot \boldsymbol{n} = 0$. We investigate the solution at the time $\boldsymbol{T} = 2.5$. We compare the efficiency of AMA with two other adaptive methods, a shock indicator [Dol98b], [FDF94] and a residual indicator [Kli00].

The shock indicator (SHOCK) is based on physical properties of shock waves. It checks the density jumps up in the direction of a flow. We define the shock indicator

$$g(i) = \frac{1}{\mathrm{diam}(T_i)} \max_{j \in s(i)} \left[-\left(\rho_i - \rho_j\right) \boldsymbol{v}_i \cdot \boldsymbol{n}_{ij} \right]^+ , \tag{1.6.10}$$

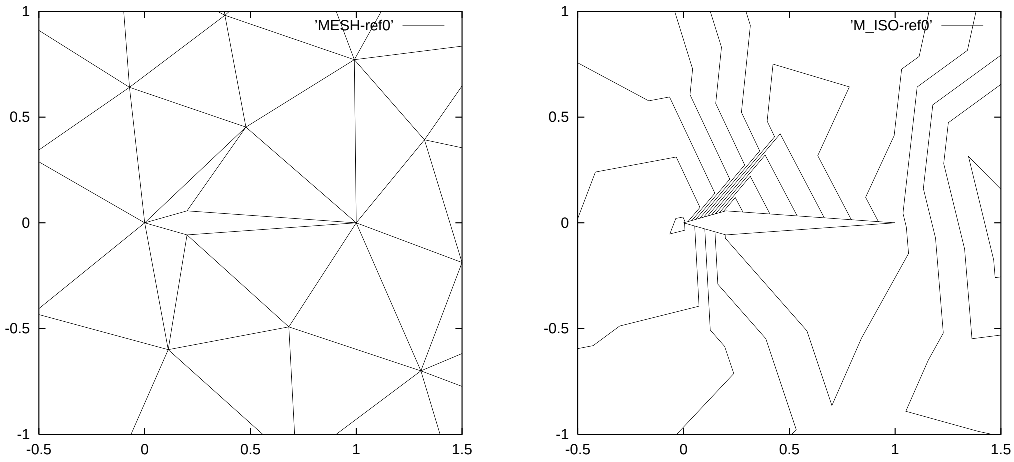


Figure 1.21: Inviscid flow around NACA 0012, triangulation and isolines of Mach number, basic mesh.

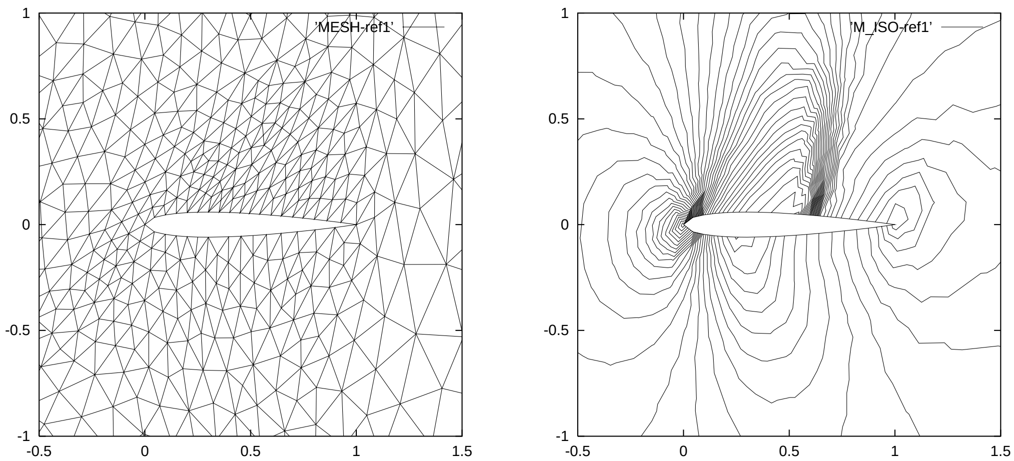


Figure 1.22: Inviscid flow around NACA 0012, triangulation and isolines of Mach number, mesh after first adaptation.

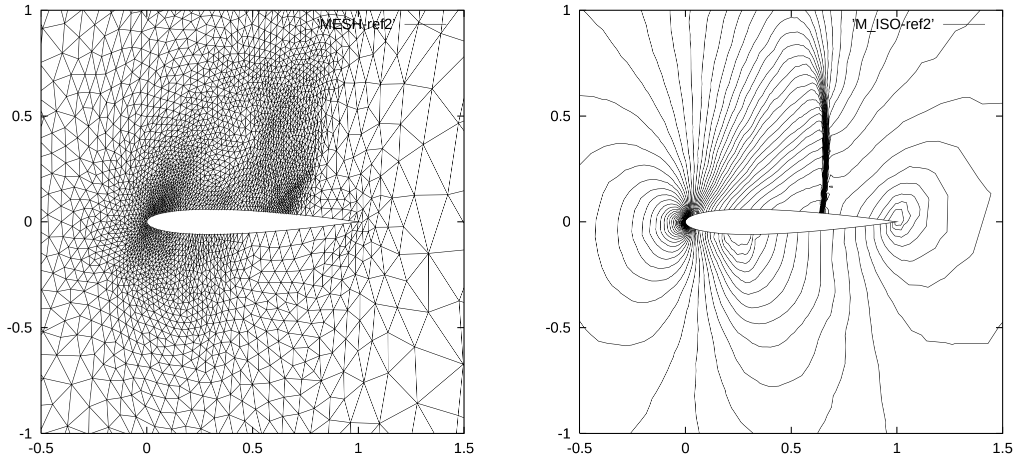


Figure 1.23: Inviscid flow around NACA 0012, triangulation and isolines of Mach number, mesh after second adaptation.

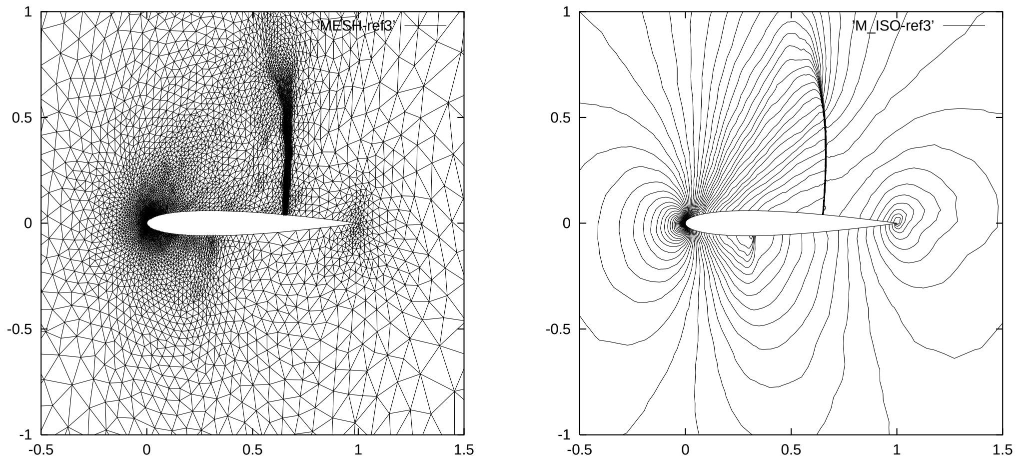


Figure 1.24: Inviscid flow around NACA 0012, triangulation and isolines of Mach number, mesh after third adaptation.

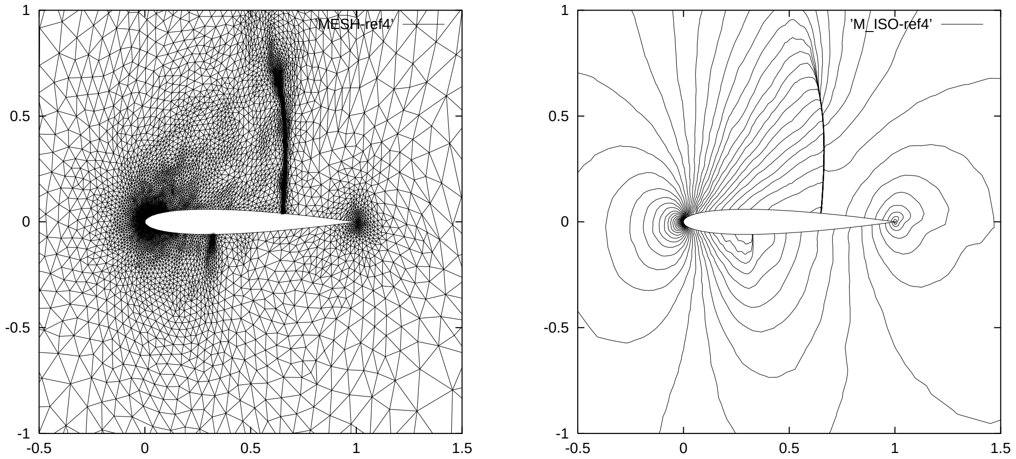


Figure 1.25: Inviscid flow around NACA 0012, triangulation and isolines of Mach number, mesh after fourth adaptation.

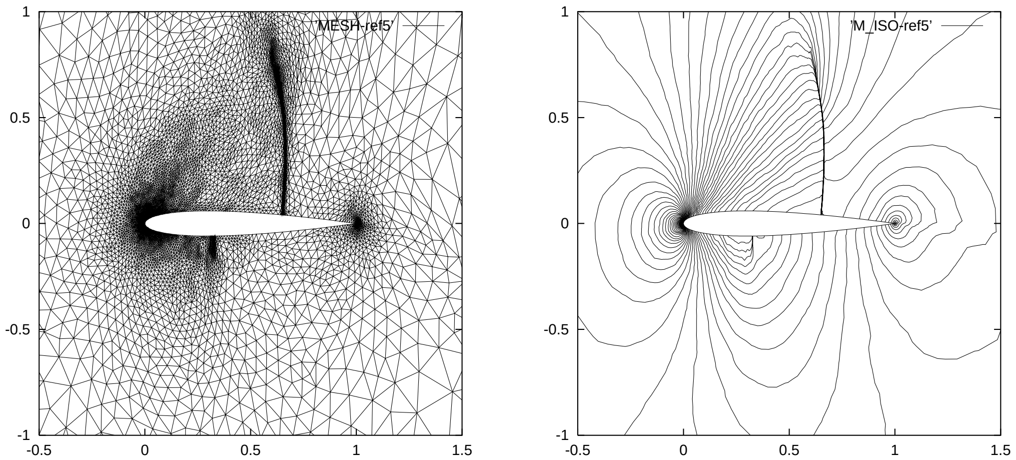


Figure 1.26: Inviscid flow around NACA 0012, triangulation and isolines of Mach number, mesh after fifth adaptation.

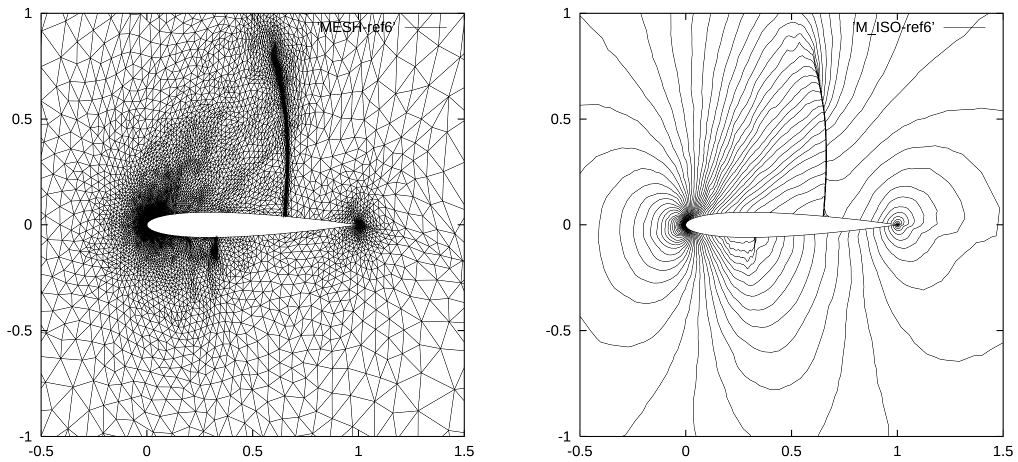


Figure 1.27: Inviscid flow around NACA 0012, triangulation and isolines of Mach number, mesh after sixth adaptation.

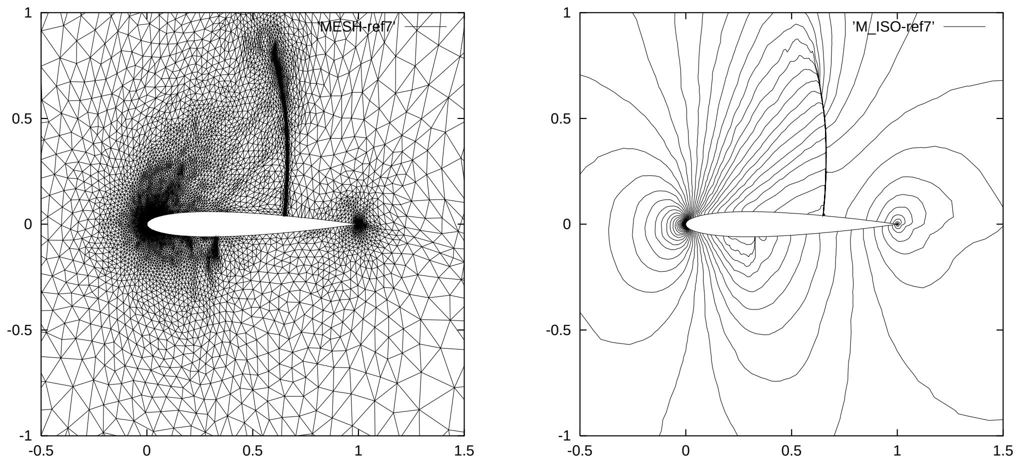


Figure 1.28: Inviscid flow around NACA 0012, triangulation and isolines of Mach number, mesh after seventh adaptation.

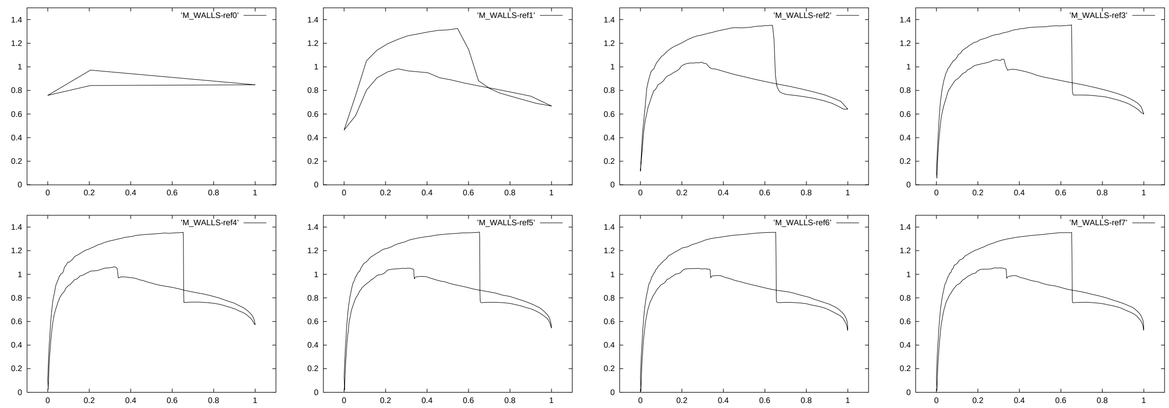


Figure 1.29: Inviscid flow around NACA 0012, Mach number distributions along the profile for all levels of mesh adaptation.

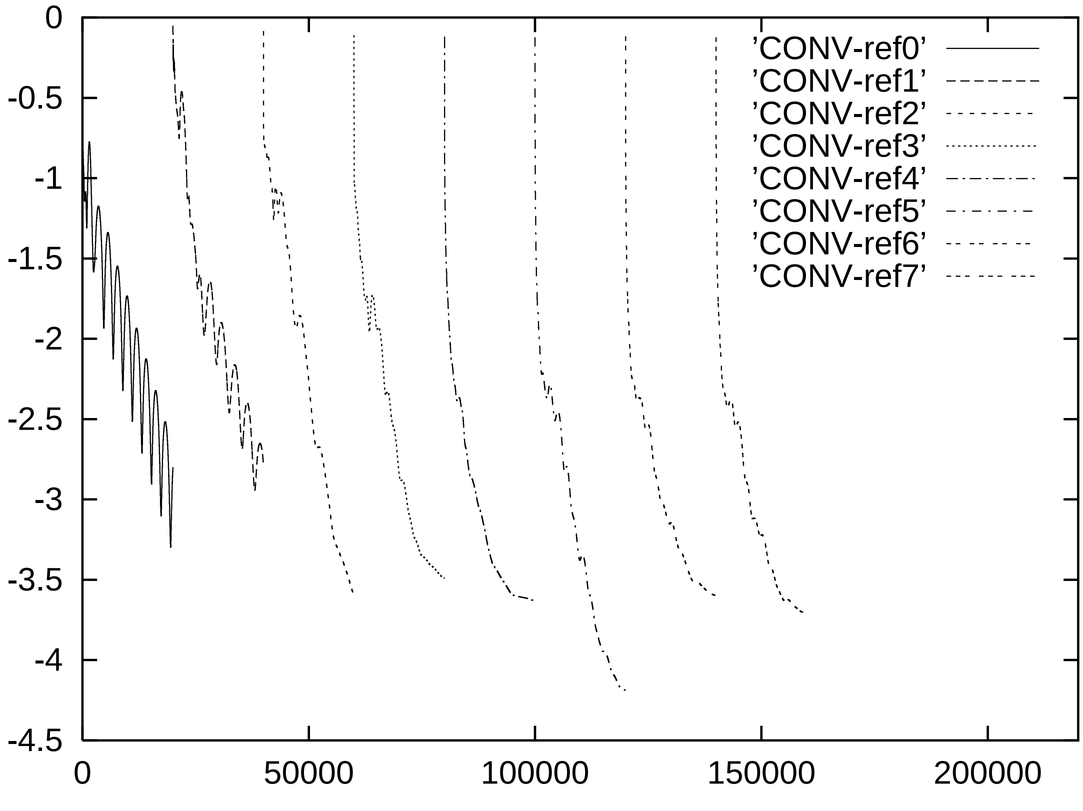


Figure 1.30: Inviscid flow around NACA 0012, history of convergence for all levels of mesh adaptation.

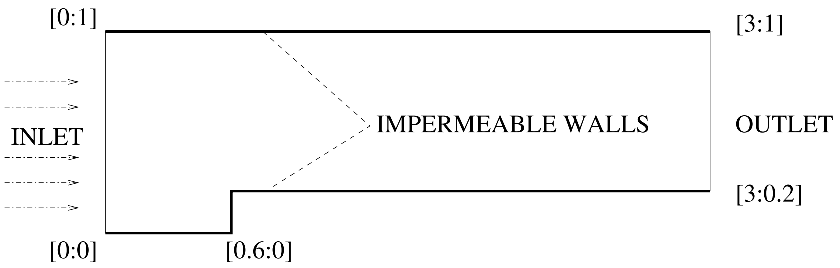


Figure 1.31: Inviscid benchmark: forward facing step.

where $s(i)$ is a set of neighbouring triangles of T_i, ρ_i and $\boldsymbol{v}_i$ is the density and velocity corresponding to T_i and $\boldsymbol{n}_{ij}$ is the unit outer normal to the common edge between triangles T_i and T_j. Moreover we put $a^+ = \max(a, 0)$. If the value of $g(i)$ is higher then a given tolerance η then T_i is refined by the red–greed refinement technique, see [BSW83].

The residual indicator (RESID) is based on the weak formulation of the Euler equations. We define the value

$$r(i) = \sup_{\varphi_h \in V_i} \frac{|(\mathcal{A}(\boldsymbol{w}_h), \varphi_h)|}{\|\varphi_h\|_{H^1_0(T_i)}}, \tag{1.6.11}$$

where $(\mathcal{A}(\boldsymbol{w}_h), \varphi_h)$ is a weak form of the residuum of the Euler equations and supremum is taken over a special set of test functions V_i, see [Kli00]. Again if the value of $r(i)$ is higher then a given tolerance η then T_i is refined by the red–greed refinement technique.

The mesh adaptation for all three adaptive technique was applied several times till the quality of the solution was approximately the same. The quality of the solution is measured only as the sharpness of shock waves in "eye" norm. It is not of course sufficient for practical computation but here we present only a qualitative comparison of three different numerical methods, see Figures 1.32 – 1.34. The following table compare the number of triangles ($\#\mathcal{T}_h$) and CPU-time of all adaptive methods.

	$\#\mathcal{T}_h$			CPU-time			
level	AMA	SHOCK	RESID	AMA adapt	AMA comput	SHOCK	RESID
0	590	590	590	0.9	2.5	2.4	2.5
1	1 680	1 680	2 113	3.1	22.5	12.4	15.5
2	2 813	4 842	6 253	5.8	118.3	67.9	86.9
3	3 480	12 769	10 249	11.3	488.4	366.4	260.8
4	3 727	31 117	14 261	13.1	559.6	1 768.9	712.0
5			17 814				1 719.2
subtotal				34.2	1 191.3		
total	3 727	31 117	17 814		1 225.5	2 218.0	2 796.9

We observe that $\#\mathcal{T}_h$ for AMA is several times smaller than for other adaptive methods and as well as the CPU-time is better for AMA. The CPU-time of AMA is shown in the table separately for the mesh adaptation itself and for the numerical solution of the Euler equations. We see that the anisotropic mesh generation does not cause any essential increasing of CPU-time.

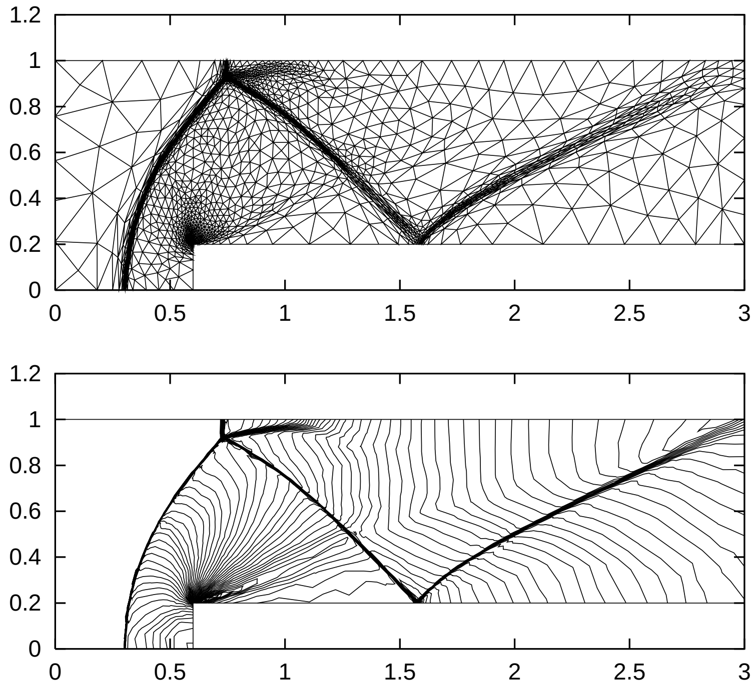


Figure 1.32: Triangulation and isolines of Mach number for AMA.

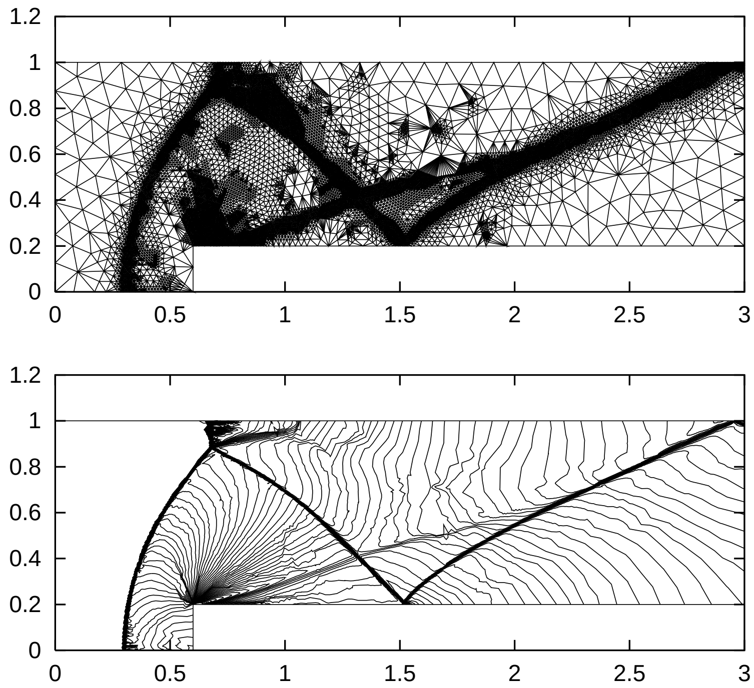


Figure 1.33: Triangulation and isolines of Mach number for shock indicator (1.6.10).

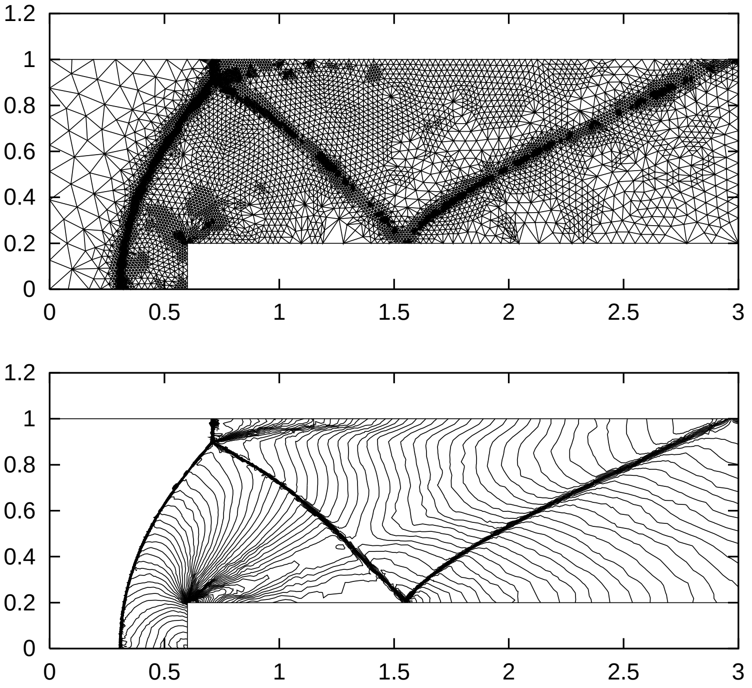


Figure 1.34: Triangulation and isolines of Mach number for residual indicator (1.6.11).

2D supersonic scramjet inlet

This example consists of an internal supersonic flow at Mach number $M_{\text{in}} = 3$ in a scramjet inlet. Figure 1.35 shows the geometry of the problem. The supersonic inlet and several obstacles with sharp angles give the solution with a few shock waves, see [DHM95]. Although the configuration is symmetric, the nonsymmetric mesh of the whole domain has been computed to observe if the solution remains symmetric or not.

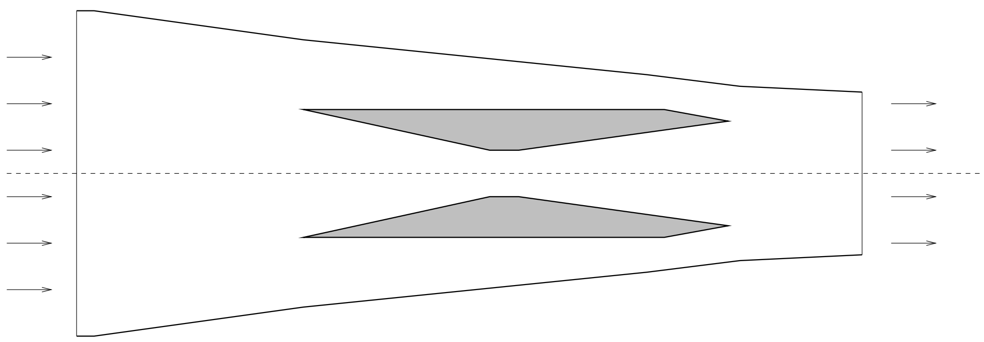


Figure 1.35: Geometry of the 2D supersonic scramjet inlet problem.

Using five automatic mesh adaptations we have obtained the final mesh. Figures 1.36 and 1.37 show the final triangular mesh (52285 elements) and the corresponding isolines of Mach number, respectively. We see rather complicated geometry of shock waves and it is interesting to notice that there are no oscillations in the solution and that the finite volume solver is robust and accurate even for such meshes with very anisotropic triangles. Moreover the numerical solution stays symmetric.

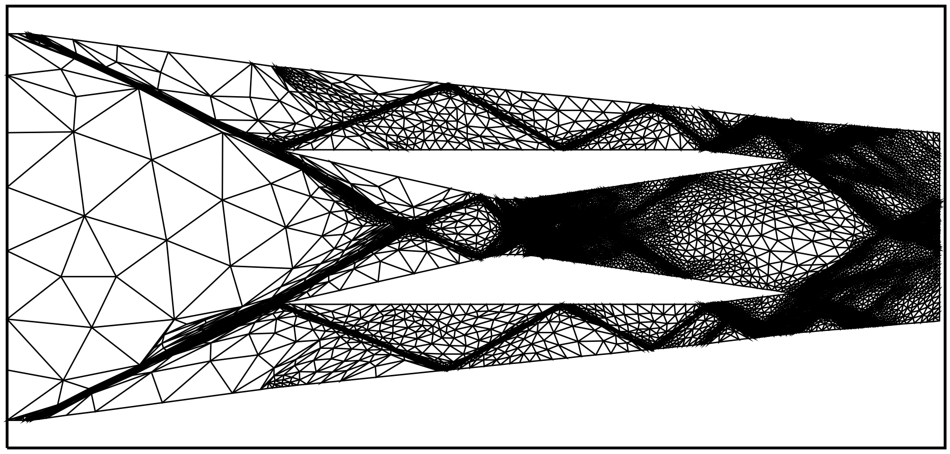


Figure 1.36: Final triangulation for a supersonic scramjet inlet.

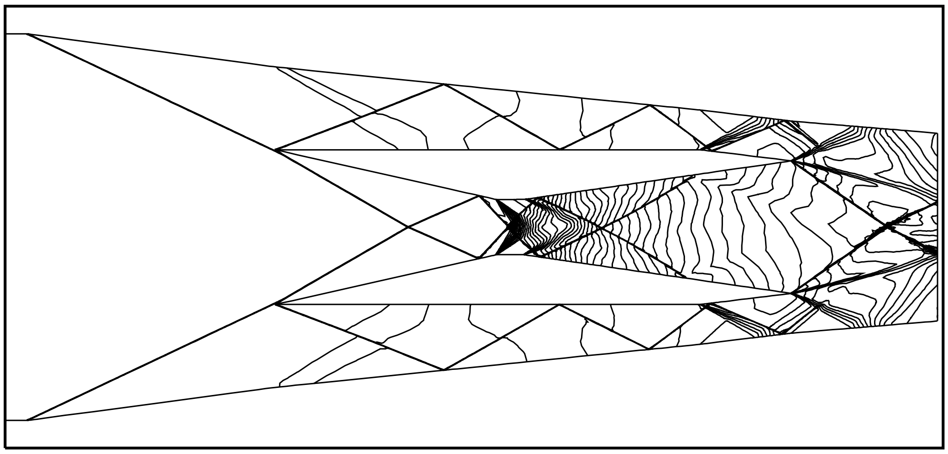


Figure 1.37: The corresponding isolines of Mach number for a supersonic scramjet inlet.

Transonic flow through 3D GAMM channel

The three-dimensional transonic inviscid flow through the channel (with 25 % spherical bump on the lower wall) of air with inlet Mach number $M_{\mathrm{in}} = 0.67$ was solved. The size of the channel is $x_1 \in [0, 2]$, $x_2 \in [0, 1.5]$ and $x_3 \in [0, 1]$. The direction of the flow is parallel with x_1 axis and goes from left to right. Figure 1.38 shows the final tetrahedral mesh (with 8786 elements) achieved after five mesh adaptations and the corresponding isolines of Mach number.

Despite the small number of elements used in the computation, the anisotropic mesh adaptation method leads to satisfactory results.

Supersonic flow through 3D forward facing step

This example is a 3D extension of the plane case, see Figure 1.31. The size in x_3–direction is $[0, 0.25]$ and we consider the same initial and boundary conditions, i.e. $\rho_0 = 1.4$, $\boldsymbol{v} = (3, 0, 0)$, $p = 1$. In fact it is only "kvasi" 3D example, as the solution does not depend on x_3 variable but it is interesting observe meshes generated by AMA.

Figure 1.39 shows the final tetrahedral mesh (with 41709 elements) achieved after three mesh adaptations and the corresponding isolines of Mach number. There are elements having size in x_3–direction equal to the size of the computational domain (=0.25), which corresponds to the fact that the solution is independent of x_3.

1.6.4 AMA for viscous flow simulation

In previous section we present the application of AMA technique for the computation of inviscid flow (i.e. $\boldsymbol{R}_i = 0,\ i = 1, \ldots, d$ in (1.6.1)), where the density ρ was chosen as the significant component of the state vector $\boldsymbol{w}$. This choice seems to be suitable because the triangulation was adapted along discontinuities (shock waves), which they were numerically captured with high accuracy.

Next possibility is to use the local Mach number $M = |\boldsymbol{v}|/\sqrt{\gamma p/\rho}$ as the significant component. The choice of M gives for inviscid flow similar results as for the density, because in subdomains $\Omega^s \subset \Omega$, where the components of $\boldsymbol{w}$ are smooth, the flow is isentropic and the density ρ is the function of M. On the other hand both ρ and M are discontinuous on the shock waves $(\Omega \setminus \Omega^s)$. The choice M as the significant component is more suitable for viscous flow simulation [CDBG+96] because the velocity (and then Mach number) is essentially smaller in boundary layers and wakes than in the external flow and then the high gradient of Mach number will be detected by AMA. The Mach number flow field is viewed in Figure 1.40.

The numerical examples show that the simple use of the Mach number as a significant component do not give yet satisfactory results for high speed flow with a small viscosity, because boundary layers and wakes are not satisfactory captured. Therefore an additional requirement has to be added to the mesh adaptation algorithm.

In the following we describe the viscous modification of AMA for the case $d = 2$ for simplicity. The extension for $d = 3$ is straightforward and it is presented in Section 1.6.5.

Boundary layer

At first we focus on the capturing of boundary layers. We present their physical properties, namely the dependence of the boundary thickness on the Reynolds number. Then we include this property into AMA method.

Let us introduce a local coordinate system $\tilde{x}, \tilde{y}$ at the point lying on the impermeable wall Γ_W, where the condition $(\boldsymbol{v} = 0)$ is prescribed. The axis $\tilde{x}$ is the tangent to Γ_W and $\tilde{y}$ is perpendicular to Γ_W. It is known fact that the velocity does not change very much in the direction $\tilde{x}$ whereas the dependency of the velocity with respect to $\tilde{y}$ is very high. To capture precisely the boundary layer we have to generate a suitable triangulation along Γ_W, Figure 1.41. We denote by h_X and h_Y the size of the triangle in the direction $\tilde{x}$ and $\tilde{y}$, respectively,

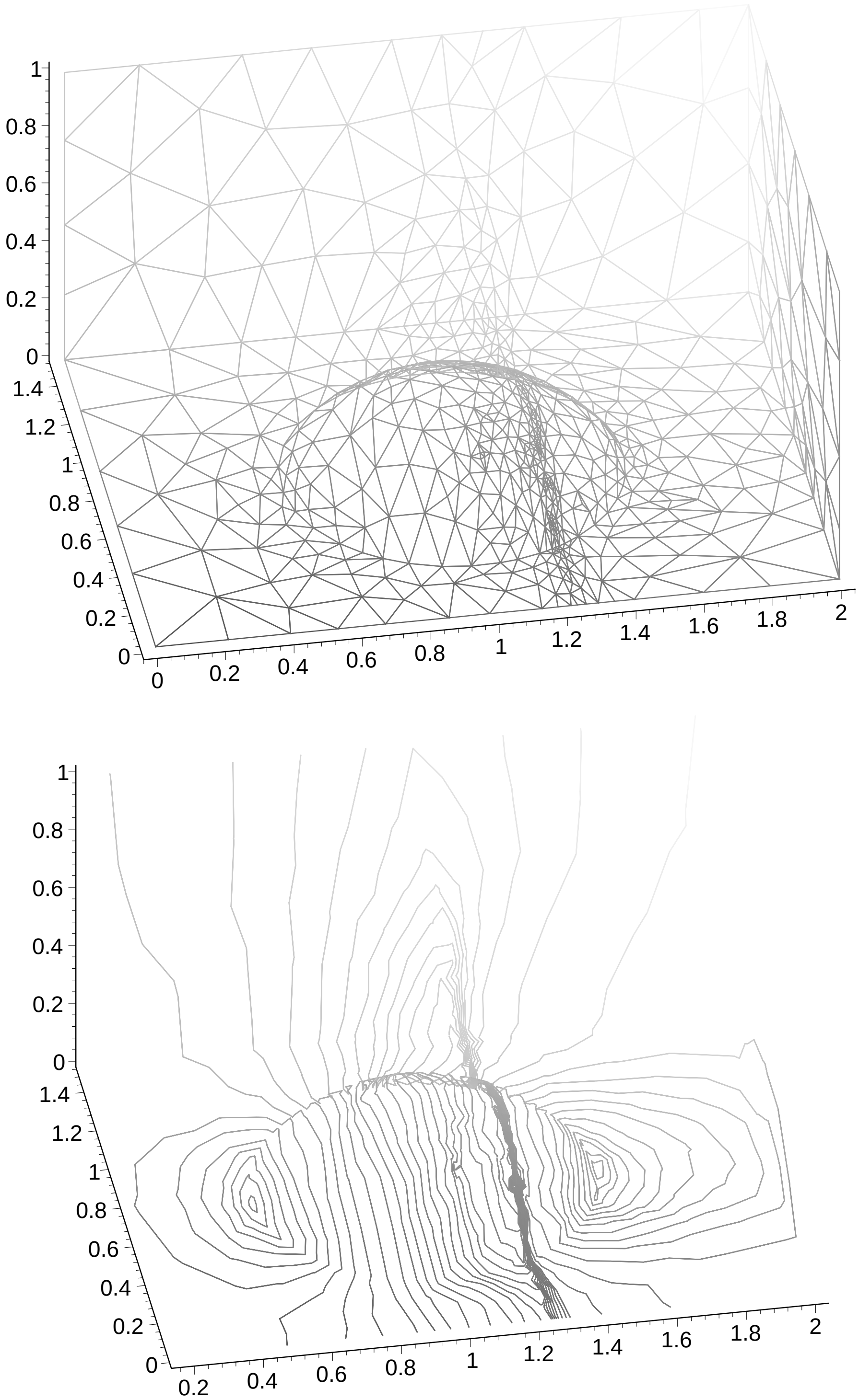


Figure 1.38: Final tetrahedrization (top) and the corresponding isolines of Mach number (bottom) for 3D GAMM channel.

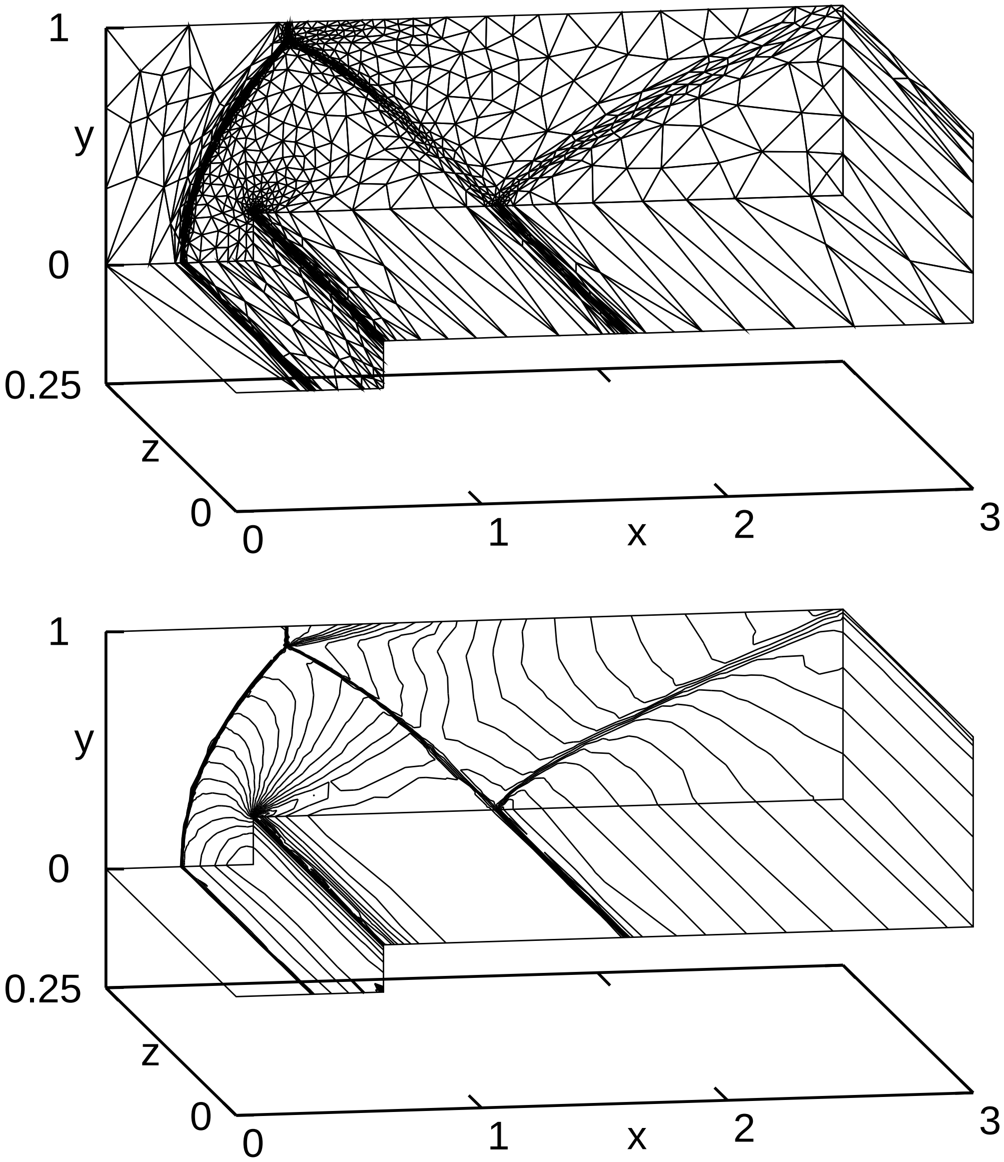


Figure 1.39: Final tetrahedrization (top) and the corresponding isolines of Mach number (bottom) for 3D forward facing step.

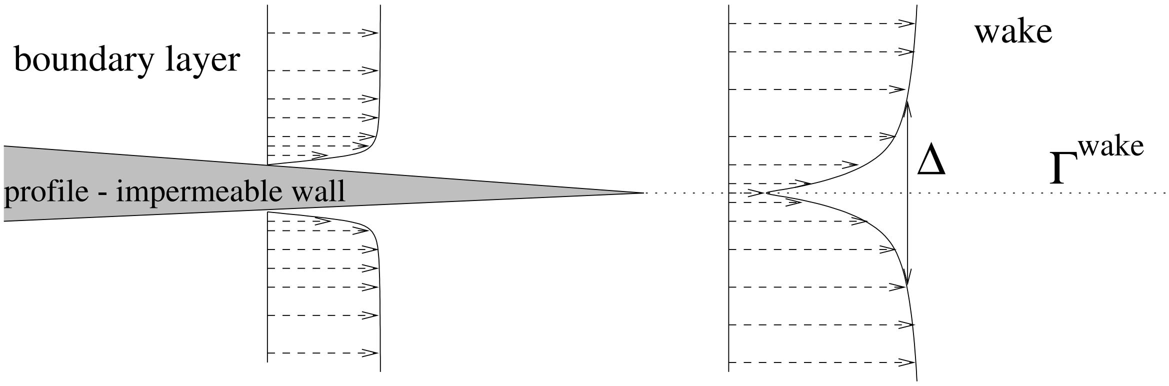


Figure 1.40: Example of Mach number flow field in a boundary layer and in a wake, the wake line Γ^{wake} and the width of the wake Δ.

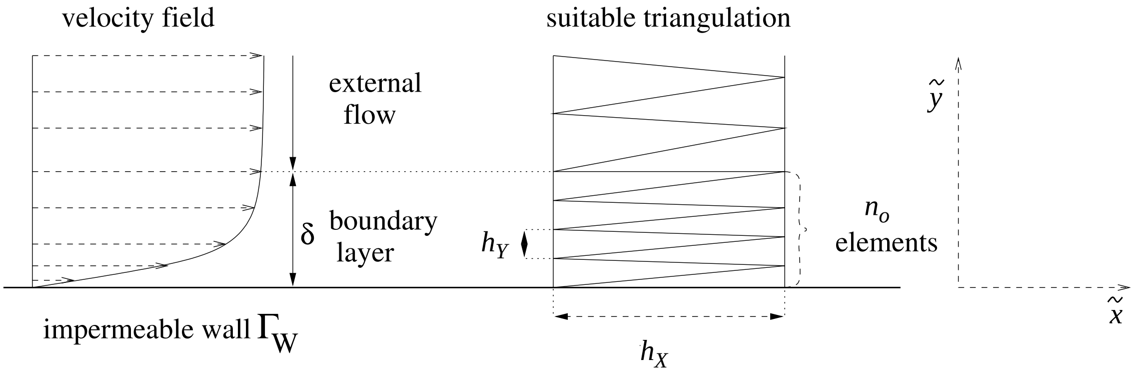


Figure 1.41: Illustration of the boundary layer with thickness δ and the sketch of a suitable triangulation with $n_0 = 4$.

see Figure 1.41. The suitable triangulation for satisfactory capturing of a boundary layer is sketched at the middle part of Figure 1.41. Triangles are thin and long and are alignment along Γ_W. The size h_Y has to be enough small to capture a boundary layer and it should successively increases with the distance from Γ_W. The size h_X has to be as high as possible so that the triangulation has as small as possible number of elements. The thickness of the boundary layer δ depends on the flow viscosity according to the relation

$$\delta \approx \frac{1}{\sqrt{Re}}, \tag{1.6.12}$$

where Re is the Reynolds number. Then to well simulate the boundary layer we put

$$h_Y(0) = \frac{n_0}{\sqrt{Re}}, \tag{1.6.13}$$

where $n_0 > 1$ is a suitable constant and its meaning is viewed in Figure 1.41. Numerical tests lead us to the choice $n_0 \in \langle 3, 10 \rangle$. For triangles laying at the distance r_y from Γ_W we put

$$h_Y(r_y) = \frac{n_0}{\sqrt{Re}} + c_y^B r_y, \tag{1.6.14}$$

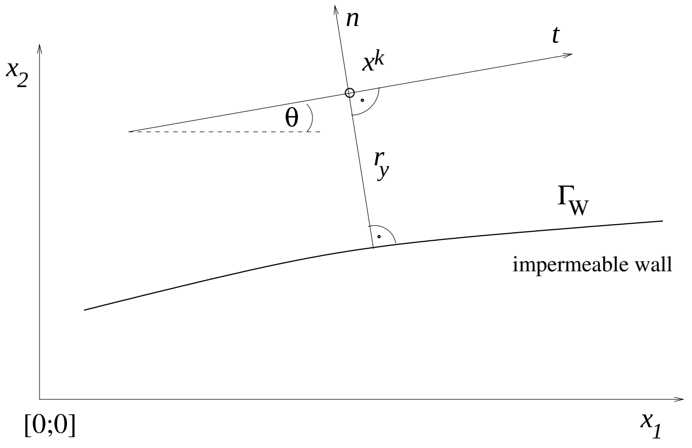


Figure 1.42: Construction of matrix $\boldsymbol{M}$ for boundary layer.

where c_y^B is a suitable constant.

On the other hand the size h_X of the triangle in the direction parallel with Γ_W is restricted only by *the minimal angle condition* required e.q. for FEM. Therefore we put

$$h_X = c_\alpha^B h_Y \tan \alpha_0, \tag{1.6.15}$$

where α_0 is the given minimal angle value and $0 < c_\alpha^B < 1$ is an "security" constant.

The relations (1.6.13) and (1.6.15) guarantee the sufficiently precise capturing of a boundary layer.

Boundary layer in the context of AMA

The previous consideration will be formulated in terms of AMA. Let $x^k \in I\!R^2$ be a point in the vicinity of fixed wall and $\boldsymbol{n}$ denotes the normal from x^k to this wall. We denote by θ the angle between the axis x_1 and the tangent $\boldsymbol{t}$ and by r_y the distance of x^k to Γ_W, see Figure 1.42.

According (1.6.15) and (1.6.14), it is suitable require that the length edge in the direction $\boldsymbol{n}$ and $\boldsymbol{t}$ are h_X and h_Y, respectively. Then using results from Appendix, for the point x^k, we define the matrix

$$\boldsymbol{M}^B(x_k) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \frac{1}{h_X^2} & 0 \\ 0 & \frac{1}{h_Y^2} \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \tag{1.6.16}$$

where θ is the angle between direction parallel with the wall and the axis x (see Figure 1.42) and h_X and h_Y are given by (1.6.15) and (1.6.14), respectively. For the matrix $\boldsymbol{M}^B(x^k)$, there exists an ellipse ϵ_k^B whose lengths and orientation are given by (1.9.3) and in virtue of (1.9.11) the norms (corresponding to $\boldsymbol{M}^B(x^k)$) of all chords of ϵ_k^B going through x^k are equal to $\sqrt{3}$. Moreover, the length of the chord going through x^k parallel with and $\boldsymbol{n}$ is equal to h_X and h_Y, respectively, which corresponds with our imagination of suitable triangulation for capturing a boundary layer.

As it is mentioned in Section 1.2.4, it is suitable for the practical implementation to come to vertex associated matrices. We define $\boldsymbol{M}^B(P_i)$, $P_i \in \sigma_h$ in the same way as $\boldsymbol{M}^B(x^k)$. We

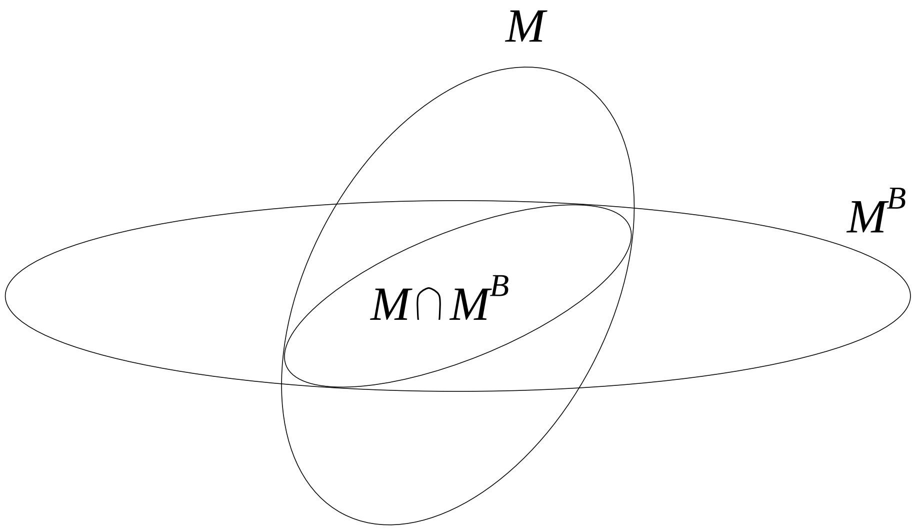


Figure 1.43: Intersection of two metrics.

have for each point $P_i \in \sigma_h$ two matrixes $\boldsymbol{M}(P_i)$ (given by (1.4.15) checking the fulfilling the necessary condition (1.2.13)) and $\boldsymbol{M}^B$ (given by (1.6.16) checking the relations (1.6.14) – (1.6.15)). To satisfy both of this condition we use the so-called *intersection of matrices*, see [DHM95].

Intersection of matrices

We say, that the matrix $\boldsymbol{M} \cap \boldsymbol{M}^B$ is the intersection of matrices $\boldsymbol{M}$ and $\boldsymbol{M}^B$, if the corresponding ellipse is the subset of ellipses corresponding to $\boldsymbol{M}$ and $\boldsymbol{M}^B$ and has the maximal possible area, see Figure 1.43. We apply the transformation of coordinates to the system in which both $\boldsymbol{M}$ and $\boldsymbol{M}^B$ are diagonal, i.e. let $\boldsymbol{P}$ be a matrix such that

$$\tilde{\boldsymbol{M}} = \boldsymbol{P}^T \boldsymbol{M} \boldsymbol{P} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \qquad \tilde{\boldsymbol{M}}^B = \boldsymbol{P}^T \boldsymbol{M}^B \boldsymbol{P} = \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix}, \tag{1.6.17}$$

where $\boldsymbol{P}^T$ denotes the matrix transpose to $\boldsymbol{P}$. Then putting

$$(\boldsymbol{M}, \boldsymbol{M}^B) = (\boldsymbol{P}^{-1})^T \begin{pmatrix} \max(\lambda_1, \mu_1) & 0 \\ 0 & \max(\lambda_2, \mu_2) \end{pmatrix} \boldsymbol{P}^{-1} \tag{1.6.18}$$

we obtain the desired matrix. Here $\boldsymbol{P}^{-1}$ means the inverse matrix to $\boldsymbol{P}$. The matrix $\boldsymbol{P}$ is not defined uniquely and it can be set using the following relations. Let $\boldsymbol{M}$ and $\boldsymbol{M}^B$ are written in the form

$$\boldsymbol{M} = \begin{pmatrix} a_1 & b_1 \\ b_1 & c_1 \end{pmatrix}, \qquad \boldsymbol{M}^B = \begin{pmatrix} a_2 & b_2 \\ b_2 & c_2 \end{pmatrix}, \tag{1.6.19}$$

where $a_i > 0, c_i > 0, a_i c_i > b_i^2,\ i = 1, 2$. Then the matrix $\boldsymbol{P}$ is given by

$$\boldsymbol{P} = \begin{pmatrix} p_1 & p_2 \\ p_3 & p_4 \end{pmatrix} \tag{1.6.20}$$

with components given for example by the following relations

$$
\begin{array}{ll}
\textbf{if} & \exists C > 0 \text{ such that } a_1 = Ca_2,\ b_1 = Cb_2,\ c_1 = Cc_2 \\
\textbf{then} & \\
& a_1 \neq -b_1 \quad \Rightarrow \quad \begin{array}{l} p_1 = -\frac{b_1+c_1}{a_1+b_1} \\ p_2 = 1 \\ p_3 = 1 \\ p_4 = 1 \end{array} \\
& \\
& a_1 = -b_1 \quad \Rightarrow \quad \begin{array}{l} p_1 = 1 \\ p_2 = 1 \\ p_3 = 1 \\ p_4 = -\frac{a_1+b_1}{b_1+c_1} \end{array} \\
\textbf{else} & \\
& \begin{array}{l} d = (a_1c_2 - a_2c_1)^2 - 4(a_1b_2 - b_1a_2)(b_1c_2 - c_1b_2) \\ p_2 = \frac{a_2c_1 - a_1c_2 - \sqrt{d}}{2(a_1b_2 - b_1a_2)} \\ p_1 = -\frac{b_2p_2 + c_2}{p_2a_2 + b_2} \\ p_3 = 1 \\ p_4 = 1 \end{array}
\end{array}
\tag{1.6.21}
$$

It can be proved that if does not exist C such that $a_1 = Ca_2,\ b_1 = Cb_2,\ c_1 = Cc_2$ then d from (1.6.21) is always nonnegative.

The application of the viscous correction is obvious. We use the AMA technique described in Section 1.4.3 but the matrix $\boldsymbol{M}(P_i)$ is replaced by the matrix $\boldsymbol{M}(P_i) \cap \boldsymbol{M}^B(P_i)$ for nodes in vicinity of Γ_W.

Wake

A viscous flow around a profile gives rise to a wake behind the trailing edge of the profile. Its flow field is viewed in Figure 1.40. In order to capture a wake accurately, it is necessary to adapt the triangulation behind the profile. We use a similar approach for nodes in vicinity of a wake as for them in vicinity of boundary layers. We define *the wake line* Γ^{wake} starting at the trailing edge and parallel with the flow field behind the profile, see Figure 1.40. Its position is known a priori (e.q. for symmetric problems) or it is set a posteriori from the results computed on the previous level of adaptation. Similarly as for a boundary layer we introduce a local coordinate system $\tilde{x}, \tilde{y}$ at the point lying on Γ^{wake}. The axis $\tilde{x}$ and $\tilde{y}$ are parallel and perpendicular with Γ^{wake}, respectively. Instead of the relations (1.6.15) and (1.6.14) we define the following relations

$$h_Y(r_y) = \frac{n_0}{\sqrt{Re}} + c_y^W \max(0, r_y - \Delta/2), \tag{1.6.22}$$

$$h_X = c_\alpha^W h_Y \tan \alpha_0, \tag{1.6.23}$$

where Δ means the width of the wake (Figure 1.40) and c_y^W and c_α^W are constants similar as c_y^B and c_α^B, respectively. The magnitude of Δ depends on the considered problem.

Then we define the matrix $\boldsymbol{M}^W$ similarly as in (1.6.16) and use the intersection of matrix $\boldsymbol{M} \cap \boldsymbol{M}^W$ from Section 1.6.4.

1.6.5 AMA for viscous flow – case $d = 3$

As the extension the two dimensional version is straightforward we present here only the relations which have different form. The relation (1.6.14), (1.6.15), (1.6.22), (1.6.23) have the same form. Further the relations (1.6.16) is replaced by

$$\boldsymbol{M}^B(x_k) = \boldsymbol{R}^{\mathrm{T}} \begin{pmatrix} \frac{1}{h_X^2} & 0 & 0 \\ 0 & \frac{1}{h_X^2} & 0 \\ 0 & 0 & \frac{1}{h_Y^2} \end{pmatrix} \boldsymbol{R}, \tag{1.6.24}$$

where $\boldsymbol{R}$ is the rotation matrix

$$\boldsymbol{R} = \begin{pmatrix} \cos\varphi \sin\theta \, \cos 2\varphi + \sin 2\varphi \, \sin\varphi & \sin 2\varphi \, \cos\varphi \, \sin\theta - \sin\varphi \, \cos 2\varphi & \cos\theta \, \cos\varphi \\ \sin\varphi \, \sin\theta \, \cos 2\varphi - \sin 2\varphi \, \cos\varphi & \sin 2\varphi \, \sin\varphi \, \sin\theta - \cos\varphi \, \cos 2\varphi & \cos\theta \, \sin\varphi \\ -\cos\theta \, \cos 2\varphi & -\cos\theta \, \sin 2\varphi & \sin\theta \end{pmatrix}, \tag{1.6.25}$$

where φ and θ are the spherical coordinates of a chosen the direction tangent to the solid wall, see Appendix.

The intersection of matrices is analogue to (1.6.17). We apply the transformation of coordinates to the system in which both $\boldsymbol{M}$ and $\boldsymbol{M}^B$ are diagonal.

1.6.6 Mesh gradation control

The application of AMA technique presented in Sections 1.4.3, 1.6.4 and 1.6.4 can lead to triangulations with high gradation. It means that the sizes of neighbour triangles very differ, see Figure 1.44 left. Too high mesh gradation gives rise some troubles for different numerical solvers (FEM, FVM). The uncontrolled mesh gradation is caused by the fact, that matrixes $\boldsymbol{M}_i$ and $\boldsymbol{M}_j$ very differ for two nearby nodes P_i and P_j. Therefore our aim is to control the mesh gradation of the triangulation generated by AMA. It is illustrated in Figure 1.44, where is shown the passage from the coarse part of the triangulation to the fine one (from left to right). The left and right figures are the triangulation generated without and with mesh gradation control, respectively. The use of mesh gradation control technique gives triangulations with better properties for numerical solution using FEM, but naturally the number of elements is higher.

We have applied two approaches from [BHF97], the H-variation and H-shock. We present only the final relations, for more detail see [BHF97]. Both H-variation and H-shock lead to similar grids.

H-variation

Let P_1P_2 be an edge of the triangulation $\mathcal{T}_h$ and $\boldsymbol{M}_{P_1}$ and $\boldsymbol{M}_{P_2}$ the matrices corresponding to nodes P_1 and P_2, respectively. Then we put

$$\begin{aligned} \boldsymbol{M}_{P_1} &:= \boldsymbol{M}_{P_1} \cap \boldsymbol{M}_{P_2} (1 + \alpha \|P_1P_2\|_{\boldsymbol{M}_{P_1}})^{-2}, \quad \alpha > 0, \\ \boldsymbol{M}_{P_2} &:= \boldsymbol{M}_{P_2} \cap \boldsymbol{M}_{P_1} (1 + \alpha \|P_1P_2\|_{\boldsymbol{M}_{P_2}})^{-2}, \end{aligned} \tag{1.6.26}$$

where $\|P_1P_2\|_{\boldsymbol{M}_{P_1}}$ and $\|P_1P_2\|_{\boldsymbol{M}_{P_2}}$ are the norms of the edge P_1P_2 corresponding to $\boldsymbol{M}_{P_1}$ and $\boldsymbol{M}_{P_2}$, respectively. The operator $\cap$ means the intersection of matrices, Section 1.6.4. We use the value $\alpha = 2$.

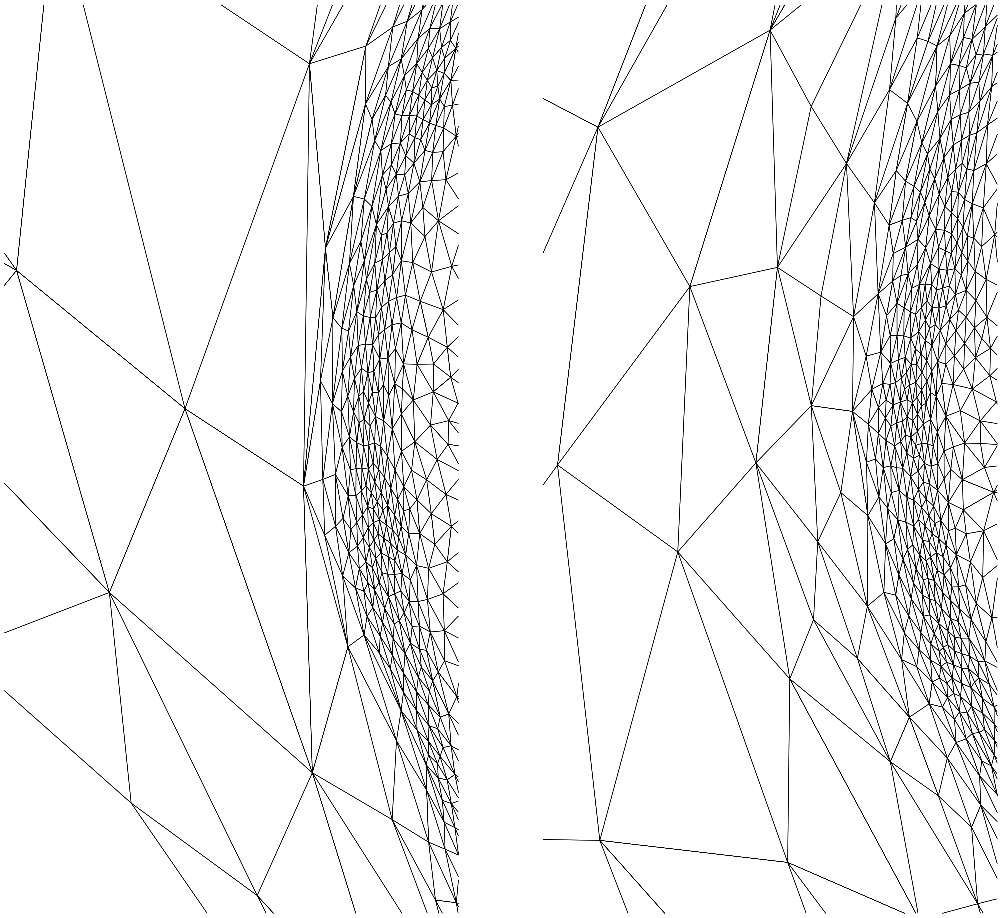


Figure 1.44: Mesh gradation control, triangulation without (left) and with (right) mesh gradation control.

H-shock

Let P_1P_2 be an edge of the triangulation. Let h_{P_1} and h_{P_2} be the norm of the unit vector in the direction P_1P_2 corresponding to matrixes $\boldsymbol{M}_{P_1}$ and $\boldsymbol{M}_{P_2}$, respectively. Let $l_{P_1P_2}$ be the norm of P_1P_2 corresponding to matrix $\bar{\boldsymbol{M}}_{P_1P_2} = (\boldsymbol{M}_{P_1} + \boldsymbol{M}_{P_2})/2$. Then H-shock associated to the edge P_1P_2 is the value

$$c_{P_1P_2} = \left(\frac{h_{P_2}}{h_{P_1}}\right)^{1/l_{P_1P_2}}. \tag{1.6.27}$$

If $c_{P_1P_2} \geq \beta$ then we replace $\boldsymbol{M}_{P_2}$ by $\eta^{-2}\boldsymbol{M}_{P_2}$, where $\eta = \left(\frac{\beta}{c_{P_1P_2}}\right)^{l_{P_1P_2}}$. We use the value $\beta = 1.5$.

1.6.7 Numerical examples of viscous flow simulation

The described method (solver and mesh adaptation) was applied to the numerical simulation of flow past a turbine cascade shown in Figure 1.45. The goal was to obtain the steady state solution with the aid of the time stabilization for $\boldsymbol{T} \to \infty$. The computational results are compared with a wind tunnel experiment (by courtesy of the Institute of Thermodynamics of the Czech Academy of Sciences in Prague, see [ŠŠ90]). The experiment and computations were performed for the following data: angle of attack $= 19^\circ\, 18'$, inlet Mach number $= 0.32$, outlet Mach number $= 1.18$, $\gamma = 1.4$, Reynolds number $Re = 1.5 \cdot 10^6$, Prandtl number $Pr = 0.72$.

Figure 1.46 represents the wind tunnel interferogram showing density isolines. We compare the results achieved by the combined barycentric FV - nonconforming FE method (Section 1.6.2, we set $\mathcal{F}_h := \mathcal{T}_h$) using triangulations $\mathcal{T}_h$ generated by AMA (Section 1.4.3) without and with viscous correction (Section 1.6.4). As the starting triangulation and numerical solution we use the results of inviscid flow simulation [Dol98b]. In Figure 1.47 and Figure 1.48 the final triangular mesh and the corresponding isolines of Mach number with zooms of trailing edge obtained with the aid of anisotropic mesh refinement with and without viscous correction are plotted, respectively. The viscous correction yields to the mesh refinement along profile and then the boundary layers are captured with high accuracy and the shock waves are not smeared as for the example without viscous mesh correction. Figure 1.49 shows the pressure distributions along the profile compared with the measurement. We see that a good agreement of computational results with experiment was achieved, if the viscous mesh correction is applied.

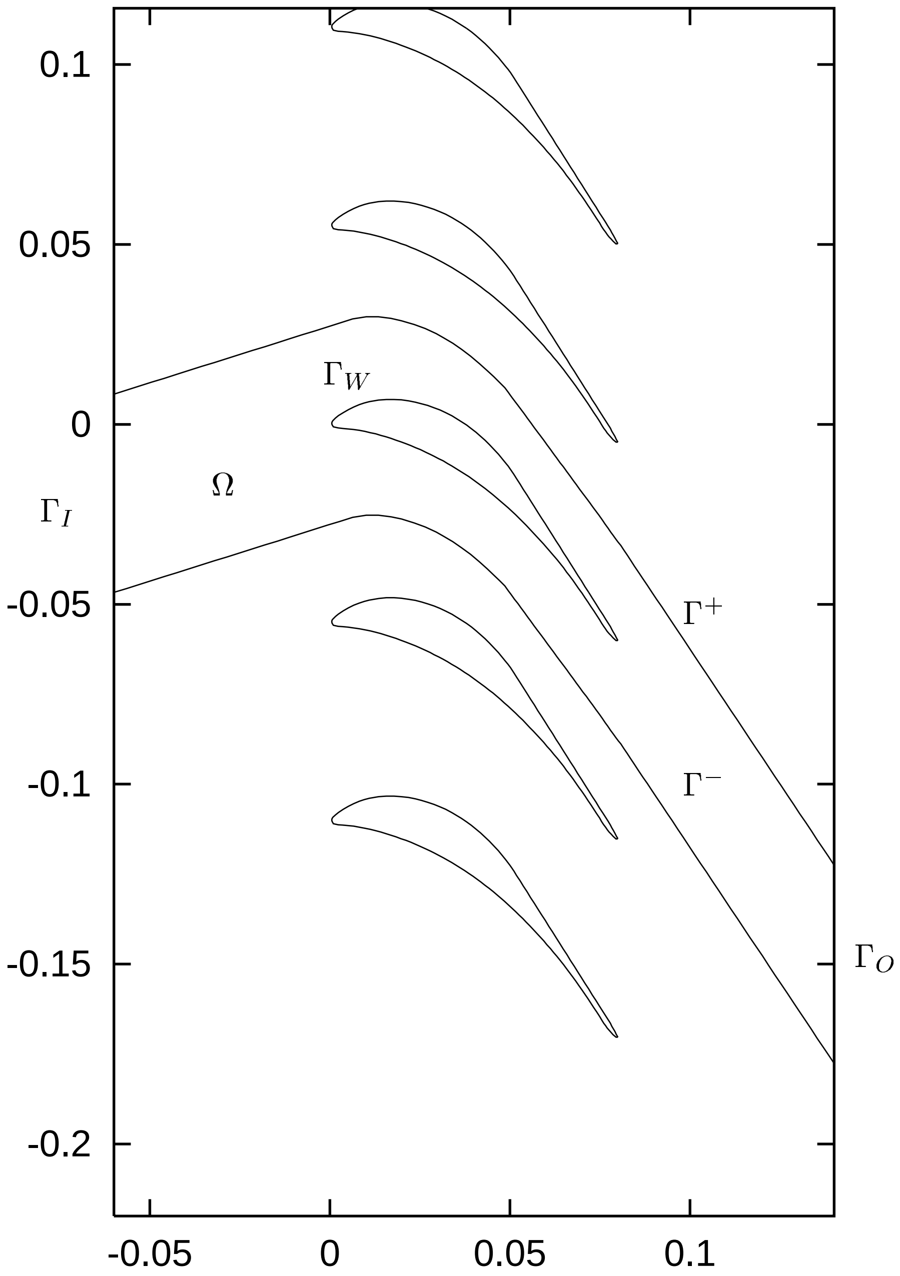


Figure 1.45: Cascade of profiles with the computational domain Ω and the boundary parts Γ_I (inlet), Γ_O (outlet), Γ_W (impearmeable walls) and the artificial periodical cuts Γ^+ and Γ^-, where the periodic boundary conditions are prescribed.

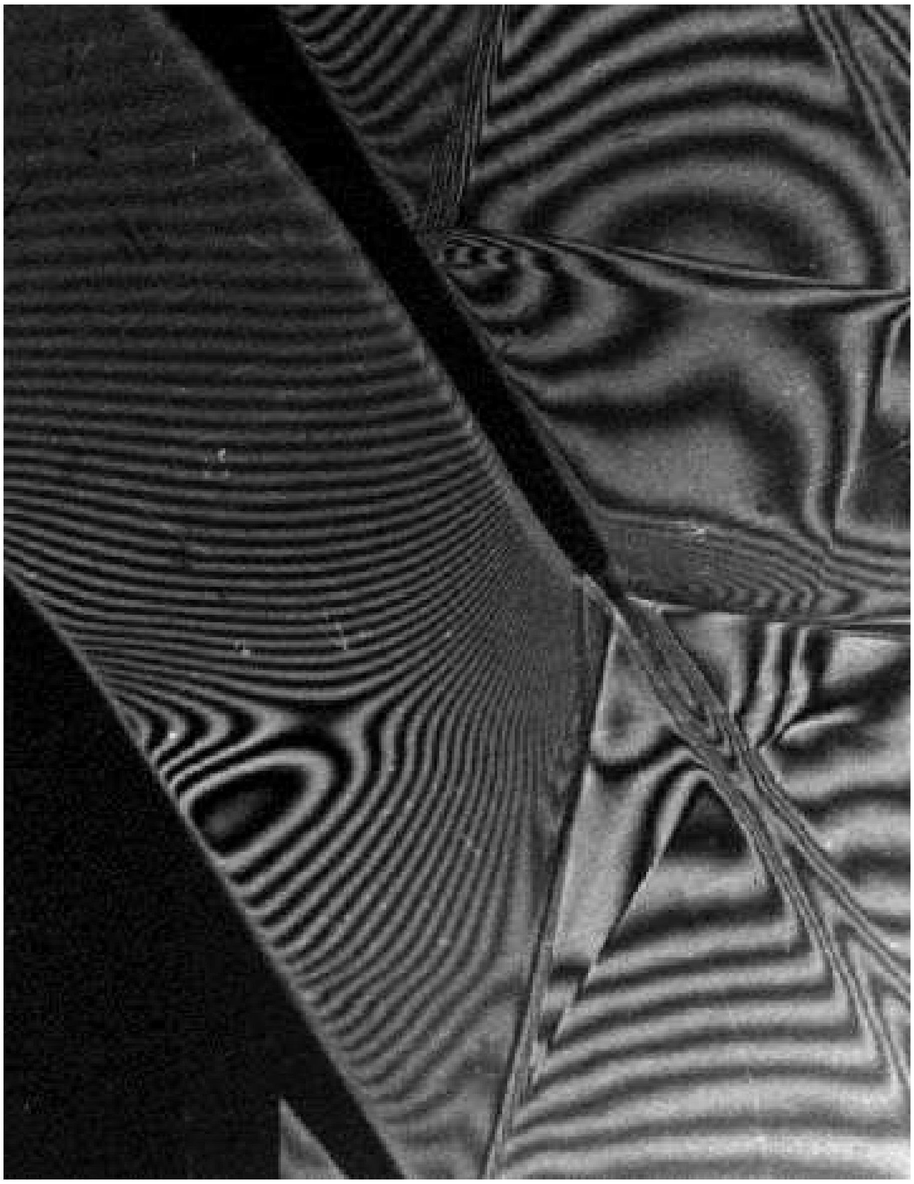


Figure 1.46: The wind tunnel interferogram showing density isolines (Courtesy of the Institute of Thermodynamics, Czech Academy of Science, Prague).

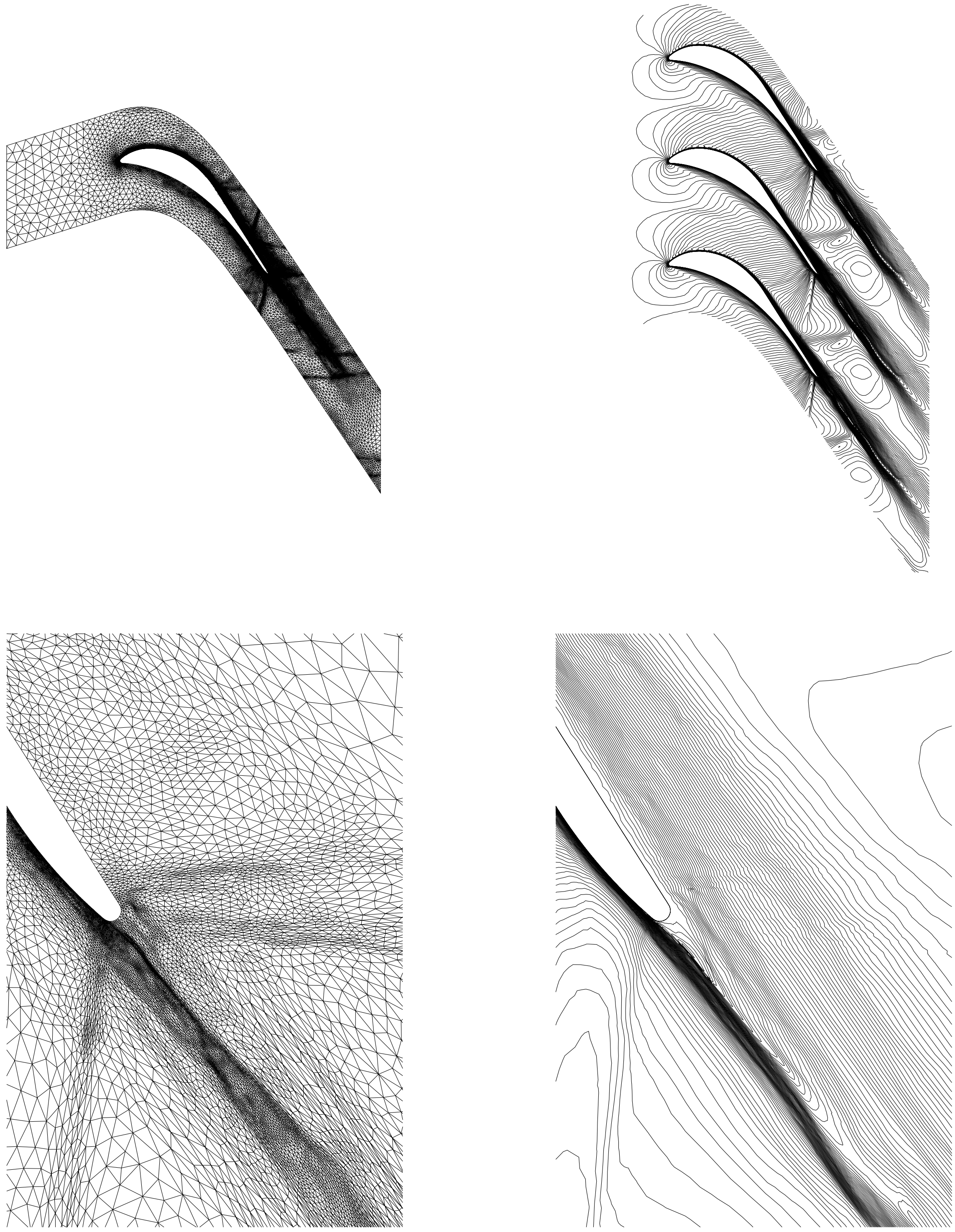


Figure 1.47: Triangulation (64 483 elements) and the corresponding isolines of Mach number with zooms achieved by AMA without viscous correction.

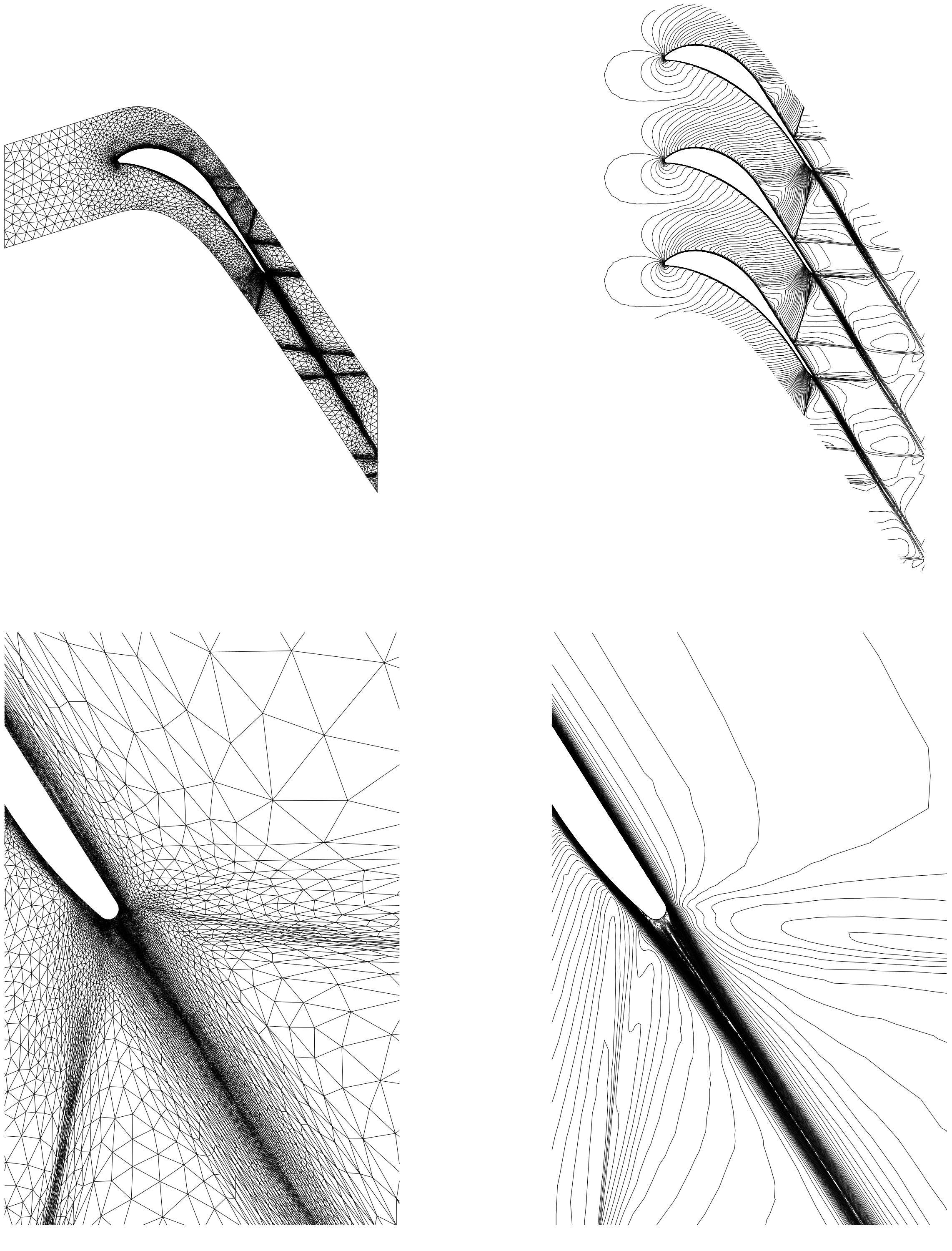


Figure 1.48: Triangulation (38 086 elements) and the corresponding isolines of Mach number with zooms achieved by AMA with viscous correction.

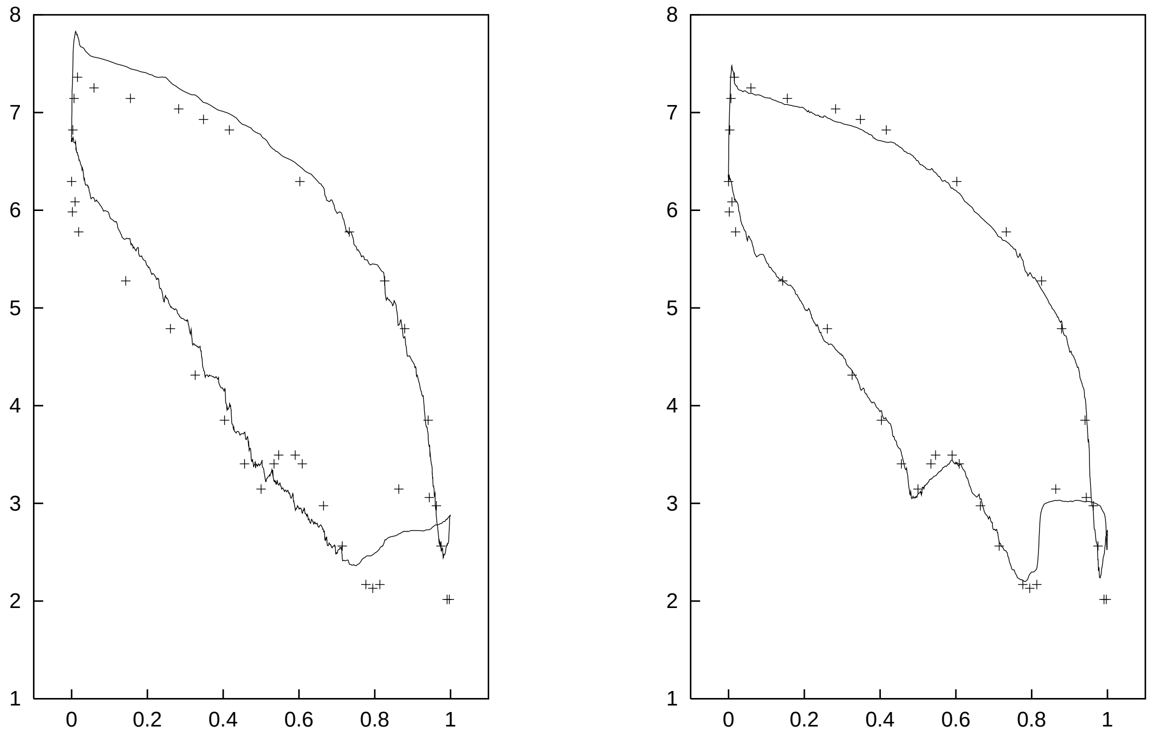


Figure 1.49: Distribution of the pressure (full line) in comparison with measurement (separated nodes) without (left) and with (right) AMA viscous correction.

1.7 Application for heat conduction problems

Although the AMA technique was originally developed for computational fluid dynamics problems, its generality allow us to use it for very different practical problems. We present the results presented in [BDM] where a computer simulation of the thermoregulation of premature infants was presented.

The motivation of this interest is the following. Suddenly occuring lack of oxygen is known to be an important cause of injury to the developing brain of premature infants [MWJG95]. The neuronal loss suffered in such a critical situation evolves over several days [GW92]. One of the factors that influence the degree and distribution of neuronal loss is the cerebral temperature. Clinical studies have shown that lowering the cerebral temperature can prevent much damage [BDGG89]. In order to investigate if the temperature within the brain of a premature infant in an incubator can be lowered by manipulating the environmental parameters, Fischer et al. [FLM00] employed a finite volume technique identical with the numerical method described within this paper but without an adaptive routine to solve a variation of the bio-heat equation which reads as

$$\partial_t \theta(x,t) = \operatorname{div}(\lambda(x)\nabla\theta(x,t)) + q(x,t), \qquad x \in \Omega,\ t \in (0,\boldsymbol{T}) \tag{1.7.1}$$

equipped by the boundary and initial conditions

$$\begin{aligned} \theta(x,t) &= \theta_{\mathrm{D}}(x), \quad x \in \partial\Omega,\ t \in (0,\boldsymbol{T}), \\ \theta(x,0) &= \theta_0(x), \quad x \in \overline{\Omega}, \end{aligned} \tag{1.7.2}$$

and which features strong source terms given by

$$q(x,t) = q_M(x) + q_B(x,t),$$

where θ is temperature and λ is heat conduction coefficient. Thereby, $q_M(x)$ is the metabolical heat production and $q_B(x,t)$ is a model of heat transfer within the human body due to blood flow. For the initial condition we choose $\theta_0(x) = 310.15\,K$. Furthermore, the boundary temperature distribution is given in the form

$$\theta_{\mathrm{D}}(x) = \begin{cases} 299.15\,K, & \text{if } x \in \partial\Omega_H, \\ 309.15\,K, & \text{otherwise}, \end{cases} \tag{1.7.3}$$

where $\partial\Omega_H$ represents the boundary of the head region.

This model is based on the assumption of a steady state situation. Concerning the grid used within the work of [FLM00], it should be noted that structured subgrids were used in the boundary region in order to capture steep heat gradients. For more details and for a theoretical discussion of the method see [FLM00].

In the following, we discuss the simulation of heat transfer with respect to an infant of 1 kg by the use of the described adaptive scheme. It is evident to start from a coarse primary grid and to proceed by computing the corresponding steady state solution. If a steady state solution is reached, the grid is adapted by using the described AMA method and again a steady state calculation is started. We now investigate different levels of refinement computed and used by the method together with the corresponding numerical solutions. Within the Figures 1.50 and 1.51 the evolution the grid undergoes is displayed, and it can be seen that a suitable solution is reached three refinement steps after starting from a quite coarse triangulation. In the following table the numbers of triangles used within the different steps of refinement is given together with the corresponding number of the mesh used in [FLM00] which is denoted as the reference mesh.

Mesh	Number of triangles
First Mesh	1822
Second Mesh	3673
Third Mesh	6731
Fourth Mesh	8956
Reference Mesh	37351

1.8 Conclusion

We present an efficient adaptive technique which can be used for the numerical solution of different problems of the mathematical engineering and computational physics. Although the AMA method is not based on a posteriori error estimation, the presented numerical examples confirm its high efficiency (i.e., the possibility to compute a sufficiently accurate solution without a high increasment of degree of freedom) and its generality (CFD, thermoregulation).

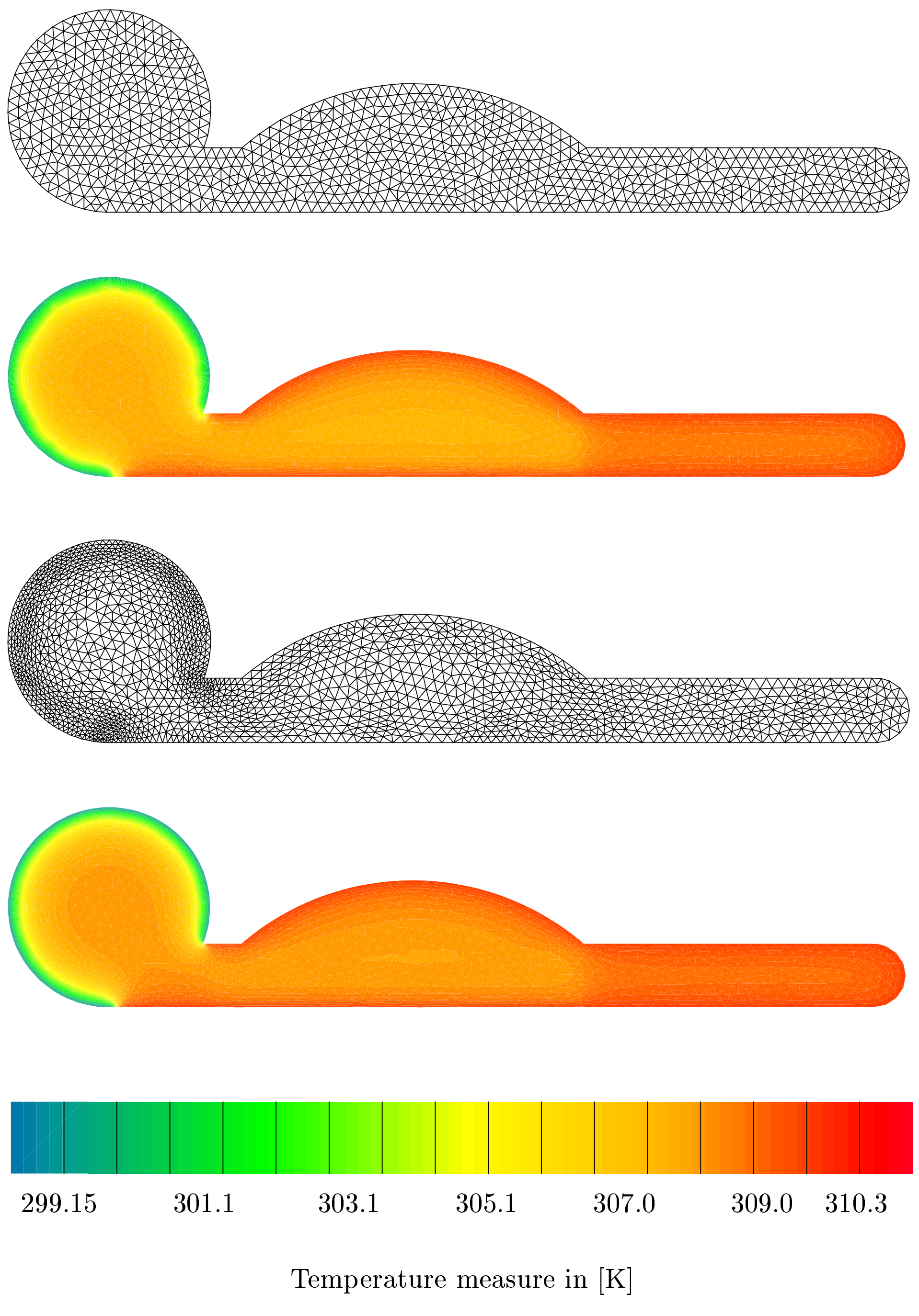


Figure 1.50: Sequence of triangulations and corresponding temperature distributions (part I).

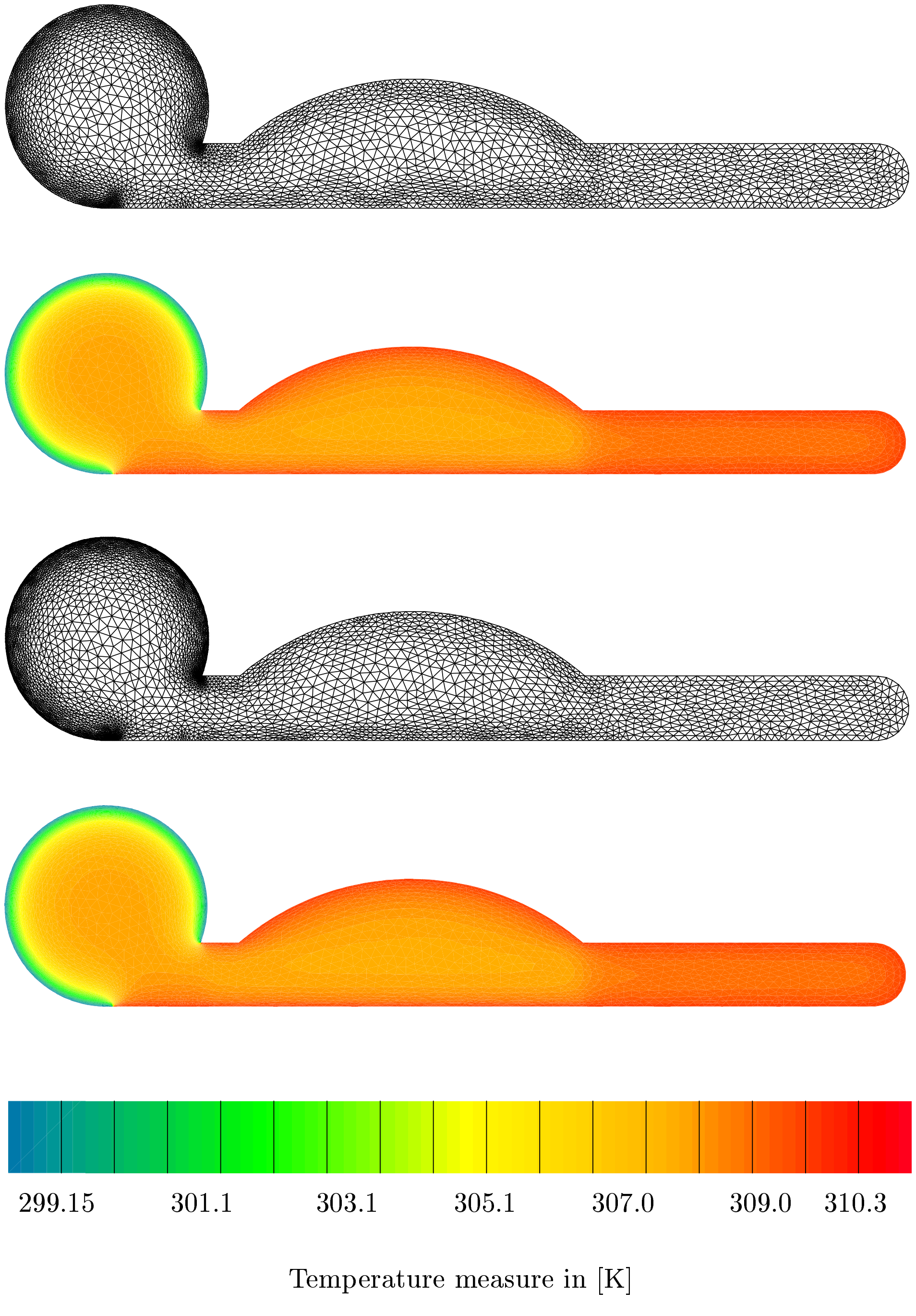


Figure 1.51: Sequence of triangulations and corresponding temperature distributions (part II).

1.9 Appendix

1.9.1 2D case

Let M be a symmetric and positive definite 2×2 matrix

$$M = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \qquad ac - b^2 > 0, \tag{1.9.1}$$

and let $\boldsymbol{v} = (v_1, v_2)$ be a vector in $I\!R^2$. We define *the norm* of the vector $\boldsymbol{v}$ *corresponding* to the matrix M as

$$\|\boldsymbol{v}\|_M \equiv \left(\boldsymbol{v}^{\mathrm{T}} M \boldsymbol{v}\right)^{\frac{1}{2}} = \left[\begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right]^{\frac{1}{2}} = \left(a v_1^2 + 2b v_1 v_2 + c v_2^2\right)^{\frac{1}{2}}. \tag{1.9.2}$$

If $M = I$ (=identical matrix) then $\|\boldsymbol{v}\|_M = |\boldsymbol{v}|$, where $|\boldsymbol{v}| = (v_1^2 + v_2^2)^{1/2}$ is the Euclidean norm of $\boldsymbol{v}$.

We decompose M in the following way:

$$M \equiv \begin{pmatrix} a & b \\ b & c \end{pmatrix} = R^{\mathrm{T}} \Lambda R = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \tag{1.9.3}$$

where λ_1, λ_2 are the eigenvalues of M and $\theta \in [0, 2\pi)$.

Let Ox_1x_2 be a Cartesian coordinate system. The set of all $x = (x_1, x_2) \in I\!R^2$ satisfying the inequality

$$x^{\mathrm{T}} M x = a x_1^2 + 2b x_1 x_2 + c x_2^2 \leq 1 \tag{1.9.4}$$

forms an ellipse ϵ_M with the centre O and with the axes lengths $r_1 = 1/\sqrt{\lambda_1}, r_2 = 1/\sqrt{\lambda_2}$. The angle between the axis x and the major axis of ϵ_M is θ, see Figure 1.52. Let P_0 be the centre of ϵ_M, then for any point P lying on the boundary of ϵ_M the norm of the vector $P_0P = (P - P_0)$ corresponding to M is constant and $\|P_0P\|_M = 1$.

We define *the optimal triangle* T_M *corresponding to* M such that $T_M \subset \epsilon_M$ and $|T_M| \geq |T|$ for any triangle $T \subset \epsilon_M$. (Here $|T|$ means the area of the triangle T). If $M = I$ (= unit matrix) then ϵ_M is an unit circle and T_M is an equilateral triangle with the lengths of edges equal to $\sqrt{3}$.

We express analytically the coordinates of the optimal triangle T_M. Let us define a new Cartesian coordinate system $O\tilde{x}_1\tilde{x}_2$ with the centre at the centre of ϵ_M and with the axes parallel with the axes of the ellipse (Figure 1.53).

The problem of finding a triangle contained in ϵ_M with the maximal area has 4 solutions, see Figure 1.54. These four triangles have the following vertices:

$$\tilde{V}_1^{(1)} = \left[\frac{1}{\sqrt{\lambda_1}}, 0\right], \quad \tilde{V}_2^{(1)} = \left[-\frac{1}{2}\frac{1}{\sqrt{\lambda_1}}, \frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_2}}\right], \quad \tilde{V}_3^{(1)} = \left[-\frac{1}{2}\frac{1}{\sqrt{\lambda_1}}, -\frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_2}}\right], \tag{1.9.5}$$

$$\tilde{V}_1^{(2)} = \left[-\frac{1}{\sqrt{\lambda_1}}, 0\right], \quad \tilde{V}_2^{(2)} = \left[\frac{1}{2}\frac{1}{\sqrt{\lambda_1}}, -\frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_2}}\right], \quad \tilde{V}_3^{(2)} = \left[\frac{1}{2}\frac{1}{\sqrt{\lambda_1}}, \frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_2}}\right], \tag{1.9.6}$$

$$\tilde{V}_1^{(3)} = \left[\frac{1}{\sqrt{\lambda_2}}, 0\right], \quad \tilde{V}_2^{(3)} = \left[-\frac{1}{2}\frac{1}{\sqrt{\lambda_2}}, \frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_1}}\right], \quad \tilde{V}_3^{(3)} = \left[-\frac{1}{2}\frac{1}{\sqrt{\lambda_2}}, -\frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_1}}\right], \tag{1.9.7}$$

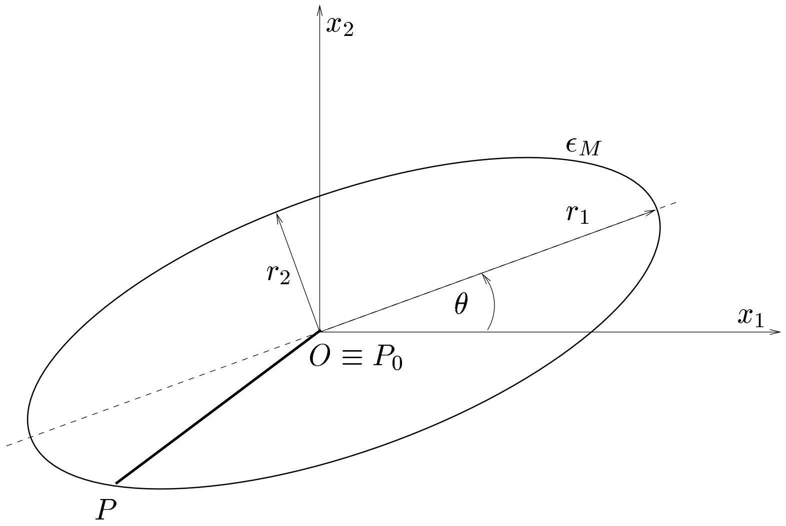


Figure 1.52: An ellipse ϵ_M corresponding to M with semiaxes length r_1, r_2, P_0 centre of ϵ_M, P any point of ϵ_M.

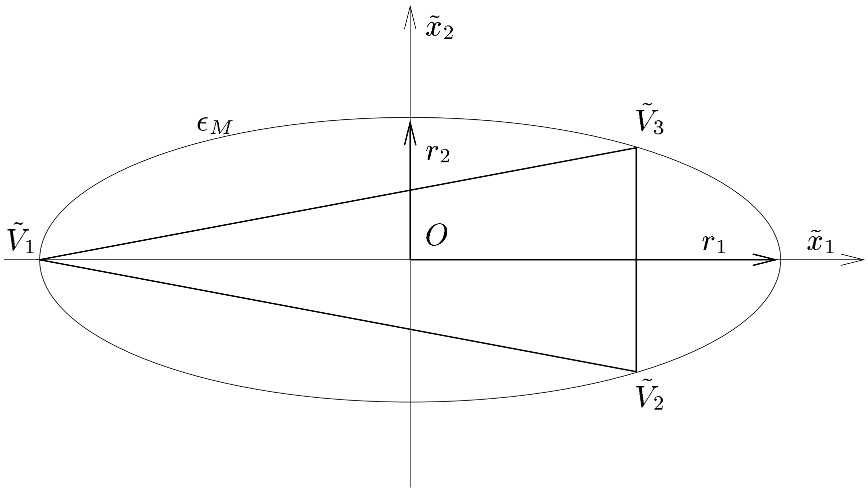


Figure 1.53: An ellipse ϵ_M corresponding to M in the new Cartesian system and the optimal triangle $T_M = \tilde{V}_1\tilde{V}_2\tilde{V}_3$.

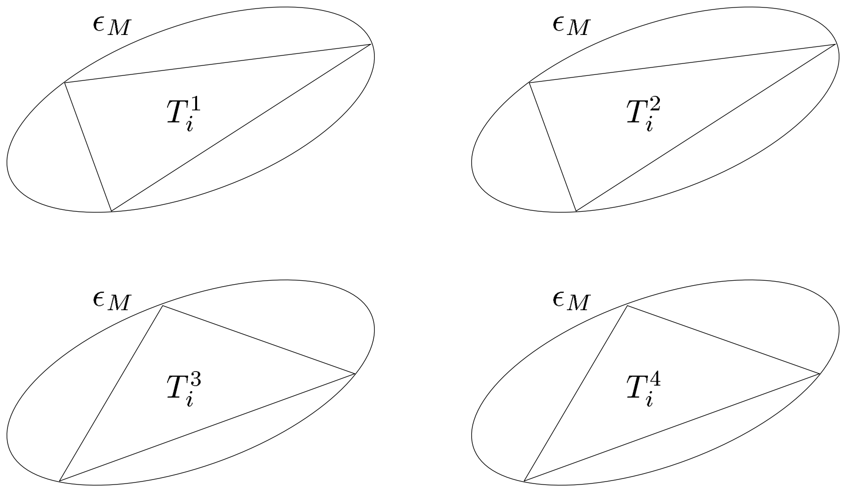


Figure 1.54: Ellipse ϵ_M and the corresponding optimal elements $T_i^1, \dots, T_i^4$.

$$\tilde{V}_1^{(4)} = \left[-\frac{1}{\sqrt{\lambda_2}}, 0\right], \quad \tilde{V}_2^{(4)} = \left[\frac{1}{2}\frac{1}{\sqrt{\lambda_2}}, -\frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_1}}\right], \quad \tilde{V}_3^{(4)} = \left[\frac{1}{2}\frac{1}{\sqrt{\lambda_2}}, \frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_1}}\right]. \tag{1.9.8}$$

Then the area of all these triangles $\tilde{V}_1^{(l)}\tilde{V}_2^{(l)}\tilde{V}_3^{(l)}$, $l = 1, \dots, 4$ is equal to

$$|T_M| = \frac{3\sqrt{3}}{4}\frac{1}{\sqrt{\lambda_1\lambda_2}}. \tag{1.9.9}$$

Figure 1.53 shows the optimal triangle $\tilde{V}_1^{(1)}\tilde{V}_2^{(1)}\tilde{V}_3^{(1)}$. The lengths of edges parallel with the direction $\tilde{x}_2$ are smaller than the lengths of edges parallel with the direction $\tilde{x}_1$.

Now, we can go back to the original coordinate system by the transformation

$$V = R\tilde{V}, \tag{1.9.10}$$

where V is a point in the system Ox_1x_2, $\tilde{V}$ is a point in the system $O\tilde{x}_1\tilde{x}_2$, R is the matrix of the rotation from (1.9.3). Applying (1.9.10) to the points $\tilde{V}_1, \tilde{V}_2, \tilde{V}_3$ (here we use the transformation only for the first triangle given by (1.9.5) and we omit the superscript $^{(1)}$) we obtain the points V_1, V_2, V_3 which are the vertices of the optimal triangle T_M in the original system:

$$\begin{aligned}
V_1 &= \left[\frac{\cos\theta}{\sqrt{\lambda_1}}, \frac{\sin\theta}{\sqrt{\lambda_2}}\right] \\
V_2 &= \left[-\frac{\sqrt{3}}{2}\frac{\sqrt{\lambda_2}\cos\theta + \sqrt{\lambda_1}\sin\theta}{\sqrt{\lambda_1\lambda_2}}, \frac{\sqrt{3}}{2}\frac{\sqrt{\lambda_1}\cos\theta - \sqrt{\lambda_2}\sin\theta}{\sqrt{\lambda_1\lambda_2}}\right], \\
V_3 &= \left[-\frac{\sqrt{3}}{2}\frac{-\sqrt{\lambda_2}\cos\theta + \sqrt{\lambda_1}\sin\theta}{\sqrt{\lambda_1\lambda_2}}, \frac{\sqrt{3}}{2}\frac{\sqrt{\lambda_1}\cos\theta + \sqrt{\lambda_2}\sin\theta}{\sqrt{\lambda_1\lambda_2}}\right].
\end{aligned}$$

Now we compute the norms corresponding to M of the edges of T_M. We put $\boldsymbol{v}_1 = V_1 - V_2$, $\boldsymbol{v}_2 = V_2 - V_3$, $\boldsymbol{v}_3 = V_3 - V_1$. After some elementary but long calculation we find that

$$\|\boldsymbol{v}_i\|_M = \left(\boldsymbol{v}_i^{\mathrm{T}} M \boldsymbol{v}_i\right)^{\frac{1}{2}} = \sqrt{3}, \qquad i = 1, 2, 3, \tag{1.9.11}$$

i.e. the norms of all edges of the optimal triangle T_M are constant and equal to $\sqrt{3}$. The relation (1.9.11) is valid also for triangles which we obtain applying (1.9.10) to (1.9.6) – (1.9.8). This is a very interesting and important property in the context of AMA.

1.9.2 3D case

It is possible to proceed in the similar way as in 2D case. We present here only the main items. Let M be a symmetric and positive definite 3×3 matrix

$$M = \begin{pmatrix} a & b & c \\ b & d & e \\ c & d & f \end{pmatrix}, \tag{1.9.12}$$

and let $\boldsymbol{v} = (v_1, v_2, v_3)$ be a vector in $I\!R^3$. We define analogously as in 2D case *the norm* of the vector $\boldsymbol{v}$ *corresponding* to the matrix M as

$$\|\boldsymbol{v}\|_M \equiv \left(\boldsymbol{v}^{\mathrm{T}} M \boldsymbol{v}\right)^{\frac{1}{2}}. \tag{1.9.13}$$

We decompose M in the following way:

$$\begin{aligned} M &= \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix} = R^T \Lambda R = \qquad\qquad (1.9.14) \\ &= \begin{pmatrix} \cos\varphi \sin\theta \, \cos 2\varphi + \sin 2\varphi \, \sin\varphi & \sin\varphi \, \sin\theta \, \cos 2\varphi - \sin 2\varphi \, \cos\varphi & -\cos\theta \, \cos 2\varphi \\ \sin 2\varphi \, \cos\varphi \, \sin\theta - \sin\varphi \, \cos 2\varphi & \sin 2\varphi \, \sin\varphi \, \sin\theta - \cos\varphi \, \cos 2\varphi & -\cos\theta \, \sin 2\varphi \\ \cos\theta \, \cos\varphi & \cos\theta \, \sin\varphi & \sin\theta \end{pmatrix} \\ &\times \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \\ &\times \begin{pmatrix} \cos\varphi \sin\theta \, \cos 2\varphi + \sin 2\varphi \, \sin\varphi & \sin 2\varphi \, \cos\varphi \, \sin\theta - \sin\varphi \, \cos 2\varphi & \cos\theta \, \cos\varphi \\ \sin\varphi \, \sin\theta \, \cos 2\varphi - \sin 2\varphi \, \cos\varphi & \sin 2\varphi \, \sin\varphi \, \sin\theta - \cos\varphi \, \cos 2\varphi & \cos\theta \, \sin\varphi \\ -\cos\theta \, \cos 2\varphi & -\cos\theta \, \sin 2\varphi & \sin\theta \end{pmatrix}, \end{aligned}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the eigenvalues of M and $\theta \in [\pi/2, \pi/2], \varphi \in [0, 2\pi)$. The inequality

$$x^{\mathrm{T}} M x = a x_1^2 + d x_2^2 + f x_3^2 + 2b x_1 x_2 + 2c x_1 x_3 + 2e x_2 x_3 \leq 1, \qquad x = (x_1, x_2, x_3), \tag{1.9.15}$$

defines an ellipsoid ϵ_M with the axes lengths $r_i = 1/\sqrt{\lambda_i},\ i = 1, 2, 3$ and the values φ and θ are the spherical coordinates of the the direction of major axis, see Figure 1.55. Similarly as in 2D case we define the *optimal tetrahedron* T_M *corresponding to* M such that $T_M \subset \epsilon_M$ and $|T_M| \geq |T|$ for any tetrahedron $T \subset \epsilon_M$. (Here $|T|$ means the volume of the tetrahedron T).

It is possible to prove that there exists 24 optimal tetrahedra for each matrix M. Moreover, if $\boldsymbol{v}_i, i = 1, \ldots, 6$ are the edges of the optimal tetrahedron T_M than

$$\|\boldsymbol{v}_i\|_M = \sqrt{\frac{8}{3}}, \qquad i = 1, \ldots, 6, \tag{1.9.16}$$

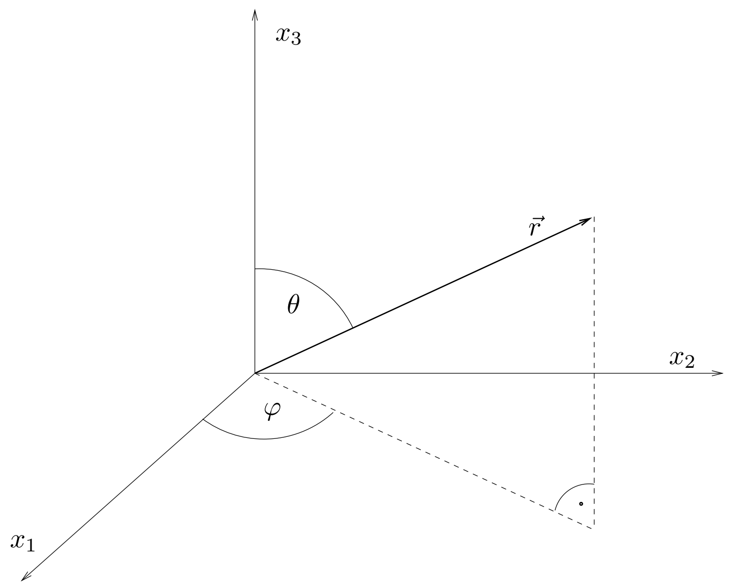


Figure 1.55: Spherical coordinates φ and θ of the direction $\vec{r}$.

Chapter 2

Discontinuous Galerkin method

2.1 Introduction

During the last decade a number of numerical techniques for the solution of nonlinear conservation laws, nonlinear convection-diffusion problems and compressible flow were developed. In the numerical treatment of weak solutions to partial differential equations there are roughly two formulations in common use: a) the energy method, using continuous test functions in the integration leading to the finite element (FE) discretization with piecewise polynomial approximations and applied mainly to elliptic and parabolic, i. e., diffusion problems; b) the finite volume (FV) type methods with piecewise constant approximations corresponding in a natural way to the physical formulation of the basic laws of conservation of mass, momentum and energy, applied mainly to convection (hyperbolic) problems. However, we can also meet with the application of the FV method to elliptic problems (called here a box method, [Hac89]) and the use of conforming finite elements to convection problems, [Joh81].

Often the finite volume method is used for the solution of nonlinear hyperbolic systems of conservation laws as flux vector splitting schemes using the so-called numerical flux (approximate Riemann solver). This approach represents a powerful tool for solving, e. g., the Euler equations describing inviscid flow (see, e. g. [Krö97], [GR91], [Fei93]).

In the solution of convection-diffusion problems it is suitable to use the advantages of both techniques – FE as well as FV. This idea results in the combined (or mixed) finite volume – finite element methods, using the FV approximation of convective terms (in connection with an approximate Riemann solver) and the FE approximation for diffusion terms. The FV and FE discretizations are usually applied on different, but mutually associated types of grids. In [FFL95], [FFLM97], [FFLMW99], the piecewise linear conforming triangular finite elements are combined with the so-called dual finite volumes. Another possibility is the combination of piecewise linear nonconforming Crouzeix–Raviart finite elements (used for the discretization of viscous incompressible flow problems – see [CR73]) with the so-called barycentric finite volumes ([ADFF98], [DFFK02]). For applications, see, e. g., [AM99], [FF96]. Very good computational results were also obtained with the aid of triangular finite volumes combined with conforming finite elements over the so-called adjoint triangular grids ([FF96], [FFD96]). The extension of these methods to 3D problems was carried out in [Fel01].

The combined FV–FE schemes (with adaptive mesh refinement) often give good results (sufficiently accurate resolution of shock waves, boundary layers and wakes), but their disadvantage is the necessity to construct two mutually associated grids, which can be rather complicated, particularly in 3D ([Fel01]). Moreover, these methods have the first order of accuracy only. An attempt to avoid the construction of different FV and FE meshes was proposed, e. g., in [GP99]. Another objection can be the fact that the techniques for the increase of accuracy of FV schemes over unstructured nonuniform meshes is not up to now theoretically based and these schemes do not seem to be sufficiently robust.

The generalization of the FV as well as FE approach allowing an increase of accuracy is represented by the *discontinuous Galerkin finite element* (DG FE) method, based on the idea to approximate the solution of an initial-boundary value problem by piecewise polynomial functions over a FE mesh, without any requirement on interelement continuity. The original DG FE method was introduced by Reed and Hill ([RH73]) for the solution of the neutron transport equation. The first analysis of this method was made by Le Saint and Raviart ([LR74]), later an improvement was achieved by Johnson and Pitkäranta ([JP86]). The DG FE method was applied to nonlinear conservation laws already in 1989 by Cockburn and Shu ([CHS89]). Their approach uses advantages of the FE method and FV schemes with an approximate Riemann solver. During several recent years the DG FE schemes have been extensively developed. For a survey, see e. g. [Coc99] or [CKe00]. An important question is the discretization of diffusion terms in the framework of the DG FE methods. There exist various treatments of this problem. One possibility is to apply a mixed formulation, used, e. g., in [CCSS02], [BR97]. Its disadvantage is a large number of unknowns. Another method is a direct discretization used, e. g., in [BBO99], [OBB98], [BMM$^+$00] and [HSS00]. For the applications to compressible flow, see [BR00] and [BR97].

Our goal is to develop sufficiently accurate schemes for nonlinear conservation laws and convection-diffusion problems which produce numerical solutions not suffering from the so-called spurious oscillations in the vicinity of discontinuities and steep gradients. The main applications are oriented to the simulation of compressible high-speed flow.

We start with a numerical analysis of the discontinuous Galerkin schemes for a scalar convection-diffusion equation (Section 2.2). Assuming that the exact solution is regular, we prove the error estimates for the semidiscretization in space. We deal with two versions of the method.

- Section 2.3, the finite volume discontinuous Galerkin (FV DG) method, which uses a FV approximation of the convective terms applied in the framework of the DG FE method. This method requires only one mesh, but its order of accuracy is one. On the other hand the numerical solution does not suffer from spurious oscillations in the vicinity of discontinuities or steep gradients.

- Section 2.4, the pure DG FE method which allows arbitrary order of accuracy for sufficiently smooth exact solution. On the other hand the numerical solution suffers from spurious oscillations in the vicinity of discontinuities or steep gradients and it is necessary to use a limitation of the order of accuracy.

Further, in Section 2.5 we are interested in problems where the solution contains discontinuities or steep gradients. We present the *jump indicator* which seeks discontinuities and decreases the order of the method to one in a narrow neighbourhood of the discontinuity. Finally, in Section 2.6, we deal with the extension DG FE method for the system of equations, namely we solve the system of the compressible Euler and Navier-Stokes equations. This extension is straightforward but it requires some additional augmentations.

2.2 Discontinuous Galerkin approximation

2.2.1 Continuous problem

Let $\Omega \subset I\!R^d$ ($d = 2$ or 3) be a bounded polyhedral domain and $T > 0$. We set $Q_T = \Omega \times (0, T)$. By $\overline{\Omega}$ and $\partial\Omega$ we denote the closure and boundary of Ω, respectively. Let us consider the following *initial-boundary value problem*: Find $u : Q_T \to I\!R$ such that

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} = \varepsilon\, \Delta u + g \quad \text{in } Q_T, \tag{2.2.1}$$

$$u\Big|_{\partial\Omega\times(0,T)} = u_D, \tag{2.2.2}$$

$$u(x, 0) = u^0(x), \quad x \in \Omega. \tag{2.2.3}$$

We assume that the data satisfy the following conditions:

$$\begin{array}{ll} \text{a)} & f_s \in C^1(I\!R),\ s = 1, \ldots, d, \text{ are Lipschitz-continuous},\ f_s(0) = 0, \\ \text{b)} & \varepsilon > 0, \\ \text{c)} & g \in C([0, T]; L^2(\Omega)), \\ \text{d)} & u_D = \text{trace of some } u^* \in C([0, T]; H^1(\Omega)) \cap L^\infty(Q_T) \text{ on } \partial\Omega \times (0, T), \\ \text{e)} & u^0 \in L^2(\Omega). \end{array} \tag{2.2.4}$$

We use the standard notation for function spaces (see, e. g., [KJk77]): $L^p(\Omega)$, $L^p(Q_T)$ denote the Lebesgue spaces, $W^{k,p}(\Omega)$, $H^k(\Omega) = W^{k,2}(\Omega)$ are the Sobolev spaces, $L^p(0, T; X)$ is the Bochner space of functions p-integrable over the interval $(0, T)$ with values in a Banach space X, $C([0, T]; X)$ $(C^1([0, T]; X))$ is the space of continuous (continuously differential) mappings of the interval $[0, T]$ into X. By $H_0^1(\Omega)$ we denote the subspace of all functions from $H^1(\Omega)$ with zero trace on $\partial\Omega$.

We consider our problem for Ω in a d-dimensional space. From the point of view of applications we have in mind $d = 2$ or $d = 3$. The assumption that $f_s(0) = 0$, $s = 1, \ldots, d$, does not cause any loss of generality, as can be seen from equation (2.2.1). The functions f_s represent convective terms, $\varepsilon > 0$ is the diffusion coefficient. The diffusion term can be, in general, more complicated (in some cases even nonlinear). It is also possible to consider mixed Dirichlet–Neumann boundary conditions, see Section 2.6.1. For simplicity we prescribe the Dirichlet condition on the whole boundary.

A sufficiently regular function satisfying (2.2.1) – (2.2.3) pointwise is called a *classical solution*. It is suitable to introduce the concept of a weak solution. To this end we introduce the following *notation*:

$$(u, v) = \int_\Omega uv \,\mathrm{d}x, \quad u,\, v \in L^2(\Omega) \tag{2.2.5}$$

(scalar product in $L^2(\Omega)$),

$$a(u, v) = \varepsilon \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x, \quad u,\, v \in H^1(\Omega), \tag{2.2.6}$$

$$b(u, v) = \int_\Omega \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} v \,\mathrm{d}x, \quad u \in H^1(\Omega) \cap L^\infty(\Omega),\ v \in L^2(\Omega), \tag{2.2.7}$$

$$\|u\|_{H^1(\Omega)} = \left(\int_\Omega \left(|u|^2 + |\nabla u|^2\right) \mathrm{d}x\right)^{1/2}, \quad u \in H^1(\Omega), \tag{2.2.8}$$

(norm in $H^1(\Omega)$),

$$|u|_{H^1(\Omega)} = \left(\int_\Omega |\nabla u|^2 \,\mathrm{d}x\right)^{1/2}, \quad u \in H^1(\Omega), \tag{2.2.9}$$

(seminorm in $H^1(\Omega)$). It is known that $|\cdot|_{H^1(\Omega)}$ is a norm on $H_0^1(\Omega)$ equivalent to $\|\cdot\|_{H^1(\Omega)}$.

The *weak solution* of (2.2.1) – (2.2.3) is defined as a function u satisfying the conditions

$$\begin{aligned} &\text{a)} \qquad u - u^* \in L^2(0,T;H_0^1(\Omega)), \quad u \in L^\infty(Q_T), \qquad (2.2.10)\\ &\text{b)} \qquad \frac{\mathrm{d}}{\mathrm{d}t}(u(t),v) + b(u(t),v) + a(u(t),v) = (g(t),v)\\ &\qquad\qquad \forall\, v \in H_0^1(\Omega) \text{ in the sense of distributions on } (0,T),\\ &\text{c)} \qquad u(0) = u_0 \quad \text{in } \Omega. \end{aligned}$$

By $u(t)$ we denote the function on Ω such that $u(t)\,(x) = u(x,t)$, $x \in \Omega$.

With the aid of techniques from [Rek82] and [Lio96], it is possible to prove that there exists a unique weak solution. Moreover, it satisfies the condition $\partial u/\partial t \in L^2(Q_T)$. Then (2.2.10), b) can be rewritten as

$$\left(\frac{\partial u(t)}{\partial t}, v\right) + b(u(t),v) + a(u(t),v) = (g(t),v) \tag{2.2.11}$$
$$\forall\, v \in H_0^1(\Omega) \text{ and almost every } t \in (0,T).$$

We say that u satisfying (2.2.10), a) – c) is a *strong solution*, if

$$u \in L^2(0,T;H^2(\Omega)), \quad \frac{\partial u}{\partial t} \in L^2(0,T;H^1(\Omega)). \tag{2.2.12}$$

It is possible to show that the strong solution u satisfies equation (2.2.1) pointwise (almost everywhere) and $u \in C([0,T], H^1(\Omega))$.

2.2.2 Weak formulation in the sense of broken Sobolev space

Triangulations

Let $\mathcal{T}_h$ $(h > 0)$ be a partition of the closure $\overline{\Omega}$ of the domain Ω into a finite number of closed d-dimensional (convex or nonconvex) polyhedra K with mutually disjoint interiors. We call $\mathcal{T}_h$ a triangulation of $\overline{\Omega}$ and *do not* require the conforming properties from the finite element method. In 2D problems we choose usually $K \in \mathcal{T}_h$ as triangles or quadrilaterals. In 3D, $K \in \mathcal{T}_h$ can be, e. g., tetrahedra, pyramids or hexahedra. (We could construct even more general elements K.) In the analysis carried out in Section 2.3.2 the important assumption is that all elements $K \in \mathcal{T}_h$ are uniformly star-shaped.

In our further considerations we shall use the following *notation*. We set $h_K = \mathrm{diam}(K)$ and $h = \max_{K\in\mathcal{T}_h} h_K$. By ρ_K we denote the radius of the largest d-dimensional ball inscribed into K and by $|K|$ we denote the d-dimensional Lebesgue measure of K. All elements of $\mathcal{T}_h$ will be numbered so that $\mathcal{T}_h = \{K_i\}_{i\in I}$, where $I \subset Z^+ = \{0,1,2,\ldots\}$ is a suitable index set. If two elements K_i, $K_j \in \mathcal{T}_h$ contain a nonempty open face, which is a part of a $(d-1)$-dimensional hyperplane, we call them *neighbours*. We set in this case

$$\Gamma_{ij} = \partial K_i \cap \partial K_j \tag{2.2.13}$$

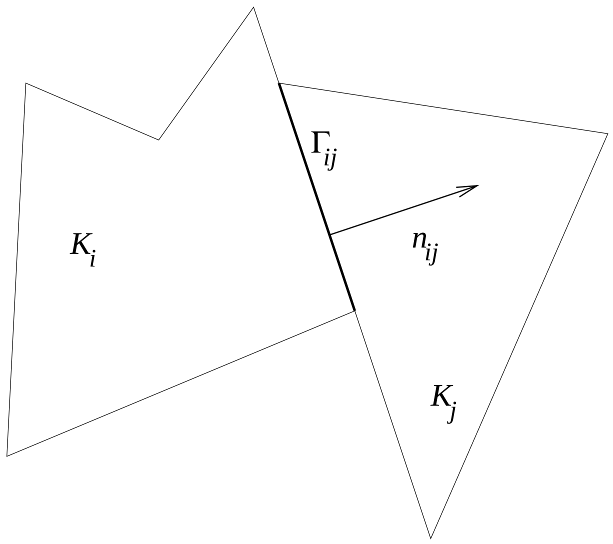


Figure 2.1: Neighbouring elements $K_i,\ K_j$.

and assume that the whole set Γ_{ij} is a part of a $(d-1)$-dimensional hyperplane. (See Figure 2.1, showing a possible 2D situation.) For $i \in I$ we set

$$s(i) = \{j \in I; K_j \text{ is a neighbour of } K_i\}. \tag{2.2.14}$$

The boundary $\partial\Omega$ is formed by a finite number of faces of elements K_i adjacent to $\partial\Omega$. We denote all these boundary faces by S_j, where $j \in I_b \subset Z^- = \{-1, -2, \ldots\}$ and set

$$\begin{aligned} \gamma(i) &= \{j \in I_b; S_j \text{ is a face of } K_i\}, \\ \Gamma_{ij} &= S_j \text{ for } K_i \in \mathcal{T}_h \text{ such that } S_j \subset \partial K_i,\ j \in I_b. \end{aligned} \tag{2.2.15}$$

For K_i not containing any boundary face S_j we set $\gamma(i) = \emptyset$. Obviously, $s(i) \cap \gamma(i) = \emptyset$ for all $i \in I$. Now, if we write $S(i) = s(i) \cup \gamma(i)$, we have

$$\partial K_i = \bigcup_{j \in S(i)} \Gamma_{ij}, \quad \partial K_i \cap \partial\Omega = \bigcup_{j \in \gamma(i)} \Gamma_{ij}. \tag{2.2.16}$$

Furthermore, we use the following notation: $\boldsymbol{n}_{ij}$ = unit outer normal to ∂K_i on the face Γ_{ij} (see Figure 2.1), $|\Gamma_{ij}| = (d-1)$-dimensional measure of Γ_{ij} (i.e., the length of Γ_{ij}, if $d = 2$, and the area of Γ_{ij}, if $d = 3$), $d(\Gamma_{ij}) = \mathrm{diam}(\Gamma_{ij})$, x^i = barycentre of K_i, x^{ij} = barycentre of Γ_{ij}. Obviously, $\boldsymbol{n}_{ij} = -\boldsymbol{n}_{ji}$ and, from the point of view of point sets, we have $\Gamma_{ij} = \Gamma_{ji}$, if $i \in I$, $j \in s(i)$ (and, hence, $j \in I$, $i \in s(j)$).

It is obvious that

$$\begin{aligned} &|K_i| \le h_{K_i}^d \le h^d, \quad |\Gamma_{ij}| \le h_{K_i}^{d-1} \le h^{d-1}, \quad d(\Gamma_{ij}) \le h_{K_i} \le h, \\ &|x^{ij} - x^i| \le h_{K_i} \le h, \quad |x^{ij} - x| \le h_{K_i} \le h \\ &\qquad \text{for } i \in I,\ j \in S(i),\ x \in \Gamma_{ij}. \end{aligned} \tag{2.2.17}$$

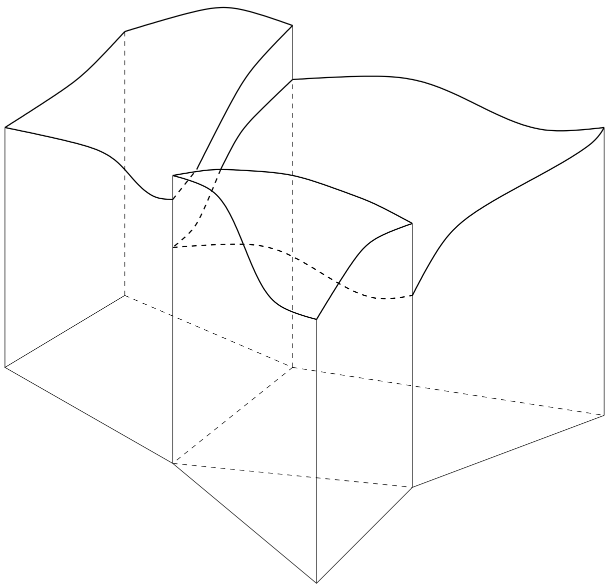


Figure 2.2: Example of a discontinuous piecewise polynomial function.

Broken Sobolev spaces

Over the triangulation $\mathcal{T}_h$ we define the so-called *broken Sobolev space*

$$H^k(\Omega, \mathcal{T}_h) = \{v; v|_K \in H^k(K)\ \forall K \in \mathcal{T}_h\} \tag{2.2.18}$$

and define there the norm

$$\|v\|_{H^k(\Omega,\mathcal{T}_h)} = \left(\sum_{K\in\mathcal{T}_h} \|v\|^2_{H^k(K)} \right)^{1/2} \tag{2.2.19}$$

and the seminorm

$$|v|_{H^k(\Omega,\mathcal{T}_h)} = \left(\sum_{K\in\mathcal{T}_h} |v|^2_{H^k(K)} \right)^{1/2}. \tag{2.2.20}$$

For $v \in H^1(\Omega, \mathcal{T}_h)$ we introduce the following notation

$$\begin{aligned} v|_{\Gamma_{ij}} &= \text{the trace of } v|_{K_i} \quad (\in H^1(K_i)) \text{ on } \Gamma_{ij}, \\ v|_{\Gamma_{ji}} &= \text{the trace of } v|_{K_j} \quad (\in H^1(K_j)) \text{ on } \Gamma_{ji}, \\ \langle v \rangle_{\Gamma_{ij}} &= \frac{1}{2}\left(v|_{\Gamma_{ij}} + v|_{\Gamma_{ji}} \right), \\ [v]_{\Gamma_{ij}} &= v|_{\Gamma_{ij}} - v|_{\Gamma_{ji}}. \end{aligned} \tag{2.2.21}$$

Obviously, $\langle v \rangle_{\Gamma_{ij}} = \langle v \rangle_{\Gamma_{ji}}$, but $[v]_{\Gamma_{ij}} = -[v]_{\Gamma_{ji}}$. On the other hand, $[v]_{\Gamma_{ij}} \boldsymbol{n}_{ij} = [v]_{\Gamma_{ji}} \boldsymbol{n}_{ji}$. Figure 2.2 show an example of a function from $H^k(\Omega, \mathcal{T}_h)$.

Variational formulation of the problem

In order to derive the discrete problem, we start from the strong solution u, multiply equation (2.2.1) by an arbitrary $v \in H^2(\Omega, \mathcal{T}_h)$, integrate over each $K_i \in \mathcal{T}_h$ and apply Green's theorem. We successively obtain the identities

$$\begin{aligned}
\text{a)} \quad & \int_{K_i} \frac{\partial u}{\partial t}(t)\, v \,\mathrm{d}x + \int_{K_i} \sum_{s=1}^{d} \frac{\partial f_s(u(t))}{\partial x_s}\, v \,\mathrm{d}x \\
& + \varepsilon \int_{K_i} \nabla u \cdot \nabla v \,\mathrm{d}x - \varepsilon \int_{\partial K_i} (\nabla u(t) \cdot \boldsymbol{n})\, v \,\mathrm{d}S = \int_{K_i} g(t)\, v \,\mathrm{d}x,
\end{aligned} \tag{2.2.22}$$

$$\begin{aligned}
\text{b)} \quad & \int_{K_i} \frac{\partial u}{\partial t}\, v \,\mathrm{d}x + \int_{\partial K_i} \sum_{s=1}^{d} f_s(u(t))\, n_s\, v \,\mathrm{d}S - \int_{K_i} \sum_{s=1}^{d} f_s(u(t))\, \frac{\partial v}{\partial x_s} \,\mathrm{d}x \\
& + \varepsilon \int_{K_i} \nabla u(t) \cdot \nabla v \,\mathrm{d}x - \varepsilon \int_{\partial K_i} (\nabla u(t) \cdot \boldsymbol{n})\, v \,\mathrm{d}S = \int_{K_i} g(t)\, v \,\mathrm{d}x.
\end{aligned}$$

Here $\boldsymbol{n} = (n_1, \dots, n_d)$ denotes the unit outer normal to ∂K_i. Summing (2.2.22), a) or b) over all $K_i \in \mathcal{T}_h$, using the relations

$$\begin{aligned}
& \int_{\Gamma_{ij}} \langle \nabla v \rangle \cdot \boldsymbol{n}_{ij} [u] \,\mathrm{d}S = 0, \\
& \langle \nabla u \rangle_{\Gamma_{ij}} = \nabla u|_{\Gamma_{ij}} = \nabla u|_{\Gamma_{ji}}
\end{aligned} \tag{2.2.23}$$

valid for the strong solution u and test function $v \in H^2(\Omega, \mathcal{T}_h)$, relations (2.2.16) and the boundary condition (2.2.2), we obtain the identities

$$\begin{aligned}
\text{a)} \quad & \left(\frac{\partial u}{\partial t}(t), v \right) + b(u(t), v) + \varepsilon \sum_{i \in I} \int_{K_i} \nabla u(t) \cdot \nabla v \,\mathrm{d}x \\
& - \varepsilon \sum_{i \in I} \sum_{\substack{j \in s(i) \\ j < i}} \int_{\Gamma_{ij}} \langle \nabla u \rangle \cdot \boldsymbol{n}_{ij} [v] \,\mathrm{d}S + \varepsilon \sum_{i \in I} \sum_{\substack{j \in s(i) \\ j < i}} \int_{\Gamma_{ij}} \langle \nabla v \rangle \cdot \boldsymbol{n}_{ij} [u(t)] \,\mathrm{d}S \\
& - \varepsilon \sum_{j \in I_b} \int_{S_j} (\nabla u(t) \cdot \boldsymbol{n})\, v \,\mathrm{d}S + \varepsilon \sum_{j \in I_b} \int_{S_j} (\nabla v \cdot \boldsymbol{n})\, u(t) \,\mathrm{d}S \\
& = \int_{\Omega} g(t)\, v \,\mathrm{d}x + \varepsilon \sum_{j \in I_b} \int_{S_j} (\nabla v \cdot \boldsymbol{n})\, u_D(t) \,\mathrm{d}S
\end{aligned} \tag{2.2.24}$$

or

$$\begin{aligned}
\text{b)} \quad & \left(\frac{\partial u}{\partial t}(t), v \right) + \sum_{i \in I} \sum_{j \in S(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{d} f_s(u(t))\, n_s\, v|_{\Gamma_{ij}} \,\mathrm{d}S \\
& - \sum_{i \in I} \int_{K_i} \sum_{s=1}^{d} f_s(u(t))\, \frac{\partial v}{\partial x_s} \,\mathrm{d}x + \varepsilon \sum_{i \in I} \int_{K_i} \nabla u(t) \cdot \nabla v \,\mathrm{d}x \\
& - \varepsilon \sum_{i \in I} \sum_{\substack{j \in s(i) \\ j < i}} \int_{\Gamma_{ij}} \langle \nabla u \rangle \cdot \boldsymbol{n}_{ij} [v] \,\mathrm{d}S + \varepsilon \sum_{i \in I} \sum_{\substack{j \in s(i) \\ j < i}} \int_{\Gamma_{ij}} \langle \nabla v \rangle \cdot \boldsymbol{n}_{ij} [u(t)] \,\mathrm{d}S \\
& - \varepsilon \sum_{j \in I_b} \int_{S_j} (\nabla u(t) \cdot \boldsymbol{n})\, v \,\mathrm{d}S + \varepsilon \sum_{j \in I_b} \int_{S_j} (\nabla v \cdot \boldsymbol{n})\, u(t) \,\mathrm{d}S \\
& = \int_{\Omega} g(t)\, v \,\mathrm{d}x + \varepsilon \sum_{j \in I_b} \int_{S_j} (\nabla v \cdot \boldsymbol{n})\, u_D(t) \,\mathrm{d}S.
\end{aligned}$$

(We see that we have added the terms $\int_{S_j}(\nabla v\cdot \boldsymbol{n}\, u(t)\,\mathrm{d}S$ and $\int_{S_j}(\nabla v\cdot \boldsymbol{n})\, u_D(t)\,\mathrm{d}S$, $j\in I_b$, to the left-hand side and the right-hand side, respectively, of (2.2.22), a) or b). Due to the boundary conditions, these terms are equal.)

In the following, we present the numerical analysis of two versions of the discontinuous Galerkin method:

- finite volume discontinuous Galerkin (FV DG) method – Section 2.3,
- discontinuous Galerkin finite element (DG FE) method – Section 2.4,

whose basic difference is in the treatment of the convective terms.

2.3 Finite volume discontinuous Galerkin method

2.3.1 Discretization

Now we shall introduce the discrete problem approximating identity (2.2.24), b) with the aid of the discontinuous Galerkin method combined with a FV approach. We define the space of discontinuous piecewise polynomial functions

$$S^{p,-1}(\Omega,\mathcal{T}_h)=\{v;v|_K\in P^p(K)\ \forall\, K\in\mathcal{T}_h\},\tag{2.3.1}$$

where $P^p(K)$ denotes the space of all polynomials on K of degree $\le p$. Further, we set

$$S_h=S^{1,-1}(\Omega,\mathcal{T}_h).\tag{2.3.2}$$

For arbitrary $v\in L^1_{\mathrm{loc}}(\Omega)$ we define its projection on the space of piecewise constant functions $S^{0,-1}(\Omega,\mathcal{T}_h)$:

$$v\to\pi_0 v:\pi_0 v|_K=\frac{1}{|K|}\int_K v\,\mathrm{d}x\quad\forall\, K\in\mathcal{T}_h.\tag{2.3.3}$$

Obviously, π_0 restricted to $L^2(\Omega)$ is the L^2-projection of $L^2(\Omega)$ into $S^{0,-1}(\Omega,\mathcal{T}_h)$:

$$(\pi_0\, v,q)=(v,q)\quad\forall\, q\in S^{0,-1}(\Omega,\mathcal{T}_h).\tag{2.3.4}$$

Now, we derive the discrete problem in such a way that in (2.2.24), b) $u(t)$ is approximated by $u_h(t)\in S_h$, v is replaced by $v_h\in S_h$, but in the expressions corresponding to the convection terms containing the flux functions f_s we use the approximations $u(t)\approx\pi_0 u_h(t)$, $v\approx\pi_0 v_h$. Hence, the terms $\int_{K_i}\sum_{s=1}^d f_s(u_h)\,\frac{\partial(\pi_0 v_h)}{\partial x_s}\,\mathrm{d}x$ vanish. Similarly as in the FV method the fluxes $\int_{\Gamma_{ij}}\sum_{s=1}^d f_s(u_h)\,n_s\,v\,\mathrm{d}S$ will be approximated with the aid of the so-called numerical flux $H(u,u',\boldsymbol{n})$ by the expression $\int_{\Gamma_{ij}}H(\pi_0 u|_{K_i},\pi_0 u|_{K_j},\boldsymbol{n}_{ij})\,\pi_0 v\,\mathrm{d}S$ for $i\in I$, $j\in S(i)$. Of course, if $j\in\gamma(i)$ and $\Gamma_{ij}\subset\partial\Omega$, it is necessary to specify the meaning of K_j and $\pi_0 u|_{K_j}$. Here we shall use a simple extrapolation putting

$$K_j:=K_i,\quad \pi_0 u|_{K_j}:=\pi_0 u|_{K_i},\quad \text{if } j\in\gamma(i).\tag{2.3.5}$$

In this way we obtain the approximation $\bar b_h(u,v)$ of the convection form $b(u,v)$:

$$\begin{aligned}\bar b_h(u,v)\ &=\ \sum_{i\in I}\pi_0 v|_{K_i}\Big\{\sum_{j\in s(i)}H(\pi_0 u|_{K_i},\pi_0 u|_{K_j},\boldsymbol{n}_{ij})\,|\Gamma_{ij}|\\ &\quad+\sum_{j\in\gamma(i)}H(\pi_0 u|_{K_i},\pi_0 u|_{K_i},\boldsymbol{n}_{ij})\,|\Gamma_{ij}|\Big\},\quad u,\,v\in L^2(\Omega).\end{aligned}\tag{2.3.6}$$

We shall assume that the numerical flux has the following properties:

Assumptions (H):

1. $H(u, v, \boldsymbol{n})$ is defined in $I\!R^d \times \mathcal{S}_1$, where $\mathcal{S}_1 = \{\boldsymbol{n} \in I\!R^d; |\boldsymbol{n}| = 1\}$, and *Lipschitz-continuous* with respect to u, v:

$$|H(u,v,\boldsymbol{n}) - H(u^*,v^*,\boldsymbol{n})| \le C_1(|u-u^*| + |v-v|^*), \qquad (2.3.7)$$
$$u,\ v,\ u^*,\ v^* \in I\!R,\ \boldsymbol{n} \in \mathcal{S}_1.$$

2. $H(u, v, \boldsymbol{n})$ is *consistent*:

$$H(u,u,\boldsymbol{n}) = \sum_{s=1}^{d} f_s(u)\, n_s, \quad u \in I\!R,\ \boldsymbol{n} = (n_1, n_2) \in \mathcal{S}_1. \qquad (2.3.8)$$

3. $H(u, v, \boldsymbol{n})$ is *conservative*:

$$H(u,v,\boldsymbol{n}) = -H(v,u,-\boldsymbol{n}), \quad u,\ v \in I\!R,\ \boldsymbol{n} \in \mathcal{S}_1. \qquad (2.3.9)$$

From (2.2.4), a) and (2.3.8) we see that

$$H(0,0,\boldsymbol{n}) = 0 \quad \forall \boldsymbol{n} \in \mathcal{S}_1. \qquad (2.3.10)$$

Now we introduce for $u, v \in H^2(\Omega, \mathcal{T}_h)$ the forms

$$\begin{aligned}
a_h(u,v) =\ & \varepsilon \sum_{i\in I} \int_{K_i} \nabla u \cdot \nabla v \,\mathrm{d}x \\
& - \varepsilon \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \langle \nabla u \rangle \cdot \boldsymbol{n}_{ij} [v] \,\mathrm{d}S + \varepsilon \sum_{i\in I} \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \langle \nabla v \rangle \cdot \boldsymbol{n}_{ij} [u] \,\mathrm{d}S \qquad (2.3.11)\\
& - \varepsilon \sum_{i\in I} \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} (\nabla u \cdot \boldsymbol{n}_{ij})\, v \,\mathrm{d}S + \varepsilon \sum_{i\in I} \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} (\nabla v \cdot \boldsymbol{n}_{ij})\, u \,\mathrm{d}S, \\
\ell_h(v)\,[t] =\ & (g(t), v) + \varepsilon \sum_{i\in I} \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} u_D(t)\, (\nabla v \cdot \boldsymbol{n}_{ij}) \,\mathrm{d}S \\
& + \varepsilon \sum_{i\in I} \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} \sigma u_D(t)\, v \,\mathrm{d}S, \qquad (2.3.12)\\
J_h^\sigma(u,v) =\ & \sum_{i\in I} \sum_{j\in s(i)} \int_{\Gamma_{ij}} \sigma [u]\,[v] \,\mathrm{d}S + \sum_{i\in I} \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} \sigma u v \,\mathrm{d}S, \qquad (2.3.13)
\end{aligned}$$

where the function $\sigma : \bigcup_{i\in I} \bigcup_{j\in S(i)} \Gamma_{ij} \to I\!R$ is defined by

$$\sigma|_{\Gamma_{ij}} = \frac{1}{d(\Gamma_{ij})}. \qquad (2.3.14)$$

Since $S_h \subset H^2(\Omega, \mathcal{T}_h)$, the above forms have sense for $u := u_h,\ v := v_h \in S_h$. In the discrete problem the form J_h^σ represents a *stabilization term* and replaces the requirement of the continuity of the approximate solution in conforming FE methods.

Now we define the *semidiscrete approximate solution* as a function $u_h : Q_T \to I\!R$ satisfying the conditions

$$\begin{aligned}
&\text{a)} \quad u_h \in C^1([0,T];S_h), && (2.3.15)\\
&\text{b)} \quad \left(\frac{\partial u_h(t)}{\partial t}, v_h\right) + a_h(u_h(t),v_h) + \bar{b}_h(u_h(t),v_h)\\
&\qquad +\varepsilon J_h^\sigma(u_h(t),v_h) = \ell_h(v_h)\,[t] \quad \forall\, v_h \in S_h,\ t \in [0,T],\\
&\text{c)} \quad (u_h(0),v_h) = (u^0,v_h) \quad \forall\, v_h \in S_h.
\end{aligned}$$

The discrete problem (2.3.15), a) – c) is equivalent to a system of ordinary differential equations (ODE). (This approach to the numerical solution of initial-boundary value problems via the space semidiscretization is called the method of lines.) If we apply some ODE solver to problem (2.3.15), a) – c), we obtain the fully discrete problem. In what follows we shall be concerned with the analysis of the semidiscrete problem (2.3.15), a) – c).

Taking into account that the exact strong solution u satisfies (2.2.24), a), $[u]|_{\Gamma_{ij}} = 0$ and $u|_{\partial\Omega\times(0,T)} = u_D$, we can see that for $v_h \in S_h, t \in (0,T)$ it satisfies

$$\left(\frac{\partial u}{\partial t}(t), v_h\right) + a_h(u(t),v_h) + b(u(t),v_h) + \varepsilon J_h^\sigma(u(t),v_h) = \ell(v_h)\,[t]. \qquad (2.3.16)$$

2.3.2 Some auxiliary results

Assumptions (A)

Let us consider a system $\{\mathcal{T}_h\}_{h\in(0,h_0)}$, $h_0 > 0$, of partitions of the domain Ω ($\mathcal{T}_h = \{K_i\}_{i\in I_h}$, $I_h \subset Z^+$) and assume that it has the following properties:

(A1) Each element $K \in \mathcal{T}_h$, $h \in (0,h_0)$ is *star-shaped* with respect to at least one point $x_K \in K$. We assume that

i) there exists a constant $\kappa > 0$ independent of K and h such that

$$\frac{\max_{x\in\partial K}|x-x_K|}{\min_{x\in\partial K}|x-x_K|} \le \kappa \quad \forall K \in \mathcal{T}_h,\ h \in (0,h_0), \qquad (2.3.17)$$

ii) element K has piecewise smooth Lipschitz boundary and K can be divided into a finite number of simplex

$$\overline{K} = \bigcup_{S\in\mathcal{S}(K)} \overline{S}, \qquad (2.3.18)$$

there exists a positive constant C_2 such that

$$\frac{h_S}{\rho_S} \le C_2 \quad \forall S \in \mathcal{S}(K) \qquad \text{(shape regularity)}, \qquad (2.3.19)$$

where h_S is the diameter of S, ρ_S is the radius of the largest d-dimensional ball inscribed into S and moreover

$$0 < \frac{1}{\tilde\kappa} \le \frac{h_K}{h_S} \le \tilde\kappa < \infty \quad \forall S \in \mathcal{S}(K), \qquad (2.3.20)$$

where $\tilde\kappa$ is a constant independent of K and h. The consequence of (2.3.17) is that there exists a constant $C_3 > 0$ such that

$$h_K^d \le C_3|K| \quad \forall\, K \in \mathcal{T}_h,\ \forall\, h \in (0,h_0). \qquad (2.3.21)$$

(A2) There exists a constant C_4 such that

$$\operatorname{card} S(i) \le C_4 \quad \forall\, K_i \in \mathcal{T}_h,\ \forall\, h \in (0, h_0). \tag{2.3.22}$$

(A3) There exists a constant $C_5 > 0$ such that

$$h_{K_i} \le C_5 d(\Gamma_{ij}), \quad i \in I, j \in S(i), h \in (0, h_0). \tag{2.3.23}$$

In what follows, for the sake of simplicity, we shall write I instead of I_h ($h \in (0, h_0)$) and the dependence of index sets $I, I_b, s(i), \gamma(i)$ and $S(i)$ on h will not be emphasized by the notation.

Under the above assumptions, the following results can be established.

Lemma 2.1 *(Multiplicative trace inequality – convex case) Let K be a convex element satisfying (2.3.19), i.e., $\frac{h_K}{\rho_K} \le C_2$. Then there exists a constant $\hat{C}_6 > 0$ independent of h, K such that*

$$\|v\|^2_{L^2(\partial K)} \le \hat{C}_6 \left(\|v\|_{L^2(K)} \, |v|_{H^1(K)} + h_K^{-1} \|v\|^2_{L^2(K)} \right), \tag{2.3.24}$$
$$K \in \mathcal{T}_h,\ v \in H^1(K),\ h \in (0, h_0).$$

Proof. Let us consider an element $K = K_i$, $i \in I$. We denote by $\tilde{x}_i$ the centre of the largest $d-$dimensional ball inscribed into K_i. Without loss of the generality we suppose that $\tilde{x}_i$ is the origin of the coordinate system. We start from the following relation

$$\int_{\partial K_i} v^2 x \cdot \boldsymbol{n} \, \mathrm{d}S = \int_{K_i} \nabla \cdot (v^2 x) \, \mathrm{d}x, \qquad v \in H^1(K_i). \tag{2.3.25}$$

Let $\boldsymbol{n}_{ij}$ be the unit outer normal to K_i on the edge Γ_{ij}, $j \in S(i)$. Then

$$x \cdot \boldsymbol{n}_{ij} = |x||\boldsymbol{n}_{ij}| \cos\alpha = |x| \cos\alpha = t_{ij}, \quad x \in \Gamma_{ij},\ j \in S(i), \tag{2.3.26}$$

where t_{ij} is the distance of $\tilde{x}_i$ from Γ_{ij}, see Figure 2.3. Obviously,

$$t_{ij} \ge \rho_{K_i} \quad \forall j \in S(i). \tag{2.3.27}$$

From (2.3.26) and (2.3.27) we have

$$\int_{\partial K_i} v^2 x \cdot \boldsymbol{n} \, \mathrm{d}S = \sum_{j \in S(i)} \int_{\Gamma_{ij}} v^2 x \cdot \boldsymbol{n}_{ij} \, \mathrm{d}S = \sum_{j \in S(i)} t_{ij} \int_{\Gamma_{ij}} v^2 \, \mathrm{d}S \tag{2.3.28}$$
$$\ge \rho_{K_i} \sum_{j \in S(i)} \int_{\Gamma_{ij}} v^2 \, \mathrm{d}S = \rho_{K_i} \|v\|^2_{L^2(\partial K_i)}.$$

Moreover,

$$\int_{K_i} \nabla \cdot (v^2 x) \, \mathrm{d}x = \int_{K_i} \left(v^2 \nabla \cdot x + x \cdot \nabla v^2 \right) \mathrm{d}x \tag{2.3.29}$$
$$= d \int_{K_i} v^2 \, \mathrm{d}x + 2 \int_{K_i} v x \cdot \nabla v \, \mathrm{d}x \le d \|v\|^2_{L^2(K_i)} + 2 \int_{K_i} |v x \cdot \nabla v| \, \mathrm{d}x.$$

With the aid of the Cauchy inequality the second term of (2.3.29) is estimated as

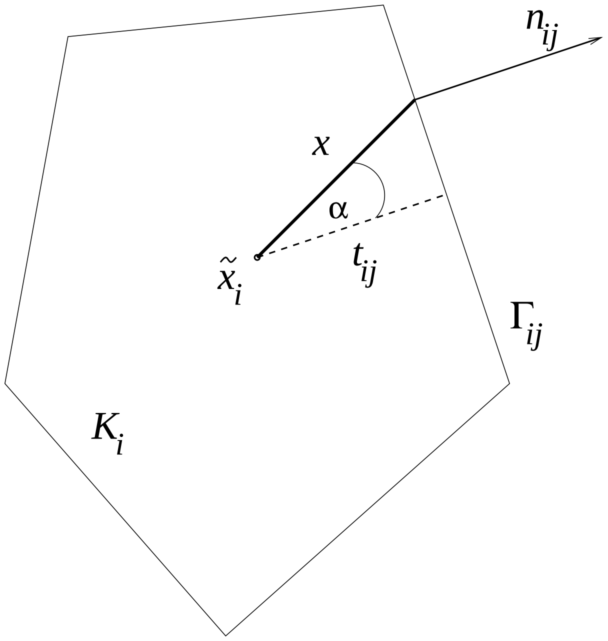


Figure 2.3: Distances t_{ij}.

$$2\int_{K_i} |vx \cdot \nabla v| \, \mathrm{d}x \le 2 \sup_{x \in K_i} |x| \int_{K_i} |v||\nabla v| \, \mathrm{d}x \le 2h_{K_i} \|v\|_{L^2(K_i)} |v|_{H^1(K_i)}. \tag{2.3.30}$$

Then (2.3.19), (2.3.25), (2.3.28), (2.3.29) and (2.3.30) give

$$\begin{aligned} \|v\|^2_{L^2(\partial K_i)} &\le \frac{1}{\rho_{K_i}} \left[2h_{K_i} \|v\|_{L^2(K_i)} |v|_{H^1(K_i)} + d\|v\|^2_{L^2(K_i)} \right] \\ &\le C_2 \left[2\|v\|_{L^2(K_i)} |v|_{H^1(K_i)} + \frac{d}{h_{K_i}} \|v\|^2_{L^2(K_i)} \right], \end{aligned} \tag{2.3.31}$$

which yields (2.3.24) with $\hat{C}_6 = C_2 \max\{2, d\} = dC_2$. □

Lemma 2.2 *(Multiplicative trace inequality – general case) There exists a constant $C_6 > 0$ independent of h, K such that*

$$\|v\|^2_{L^2(\partial K)} \le C_6 \left(\|v\|_{L^2(K)} \, |v|_{H^1(K)} + h_K^{-1} \|v\|^2_{L^2(K)} \right), \tag{2.3.32}$$
$$K \in \mathcal{T}_h, \; v \in H^1(K), \; h \in (0, h_0).$$

Proof. Let $K \in \mathcal{T}_h$ and $S \in \mathcal{S}(K)$. Then (2.3.18), (2.3.19) and Lemma 2.1 yield

$$\|v\|^2_{L^2(\partial S)} \le \hat{C}_6 \left(\|v\|_{L^2(S)} \, |v|_{H^1(S)} + h_S^{-1} \|v\|^2_{L^2(S)} \right), \; S \in \mathcal{S}(K), \; v \in H^1(K), \tag{2.3.33}$$

where $\hat{C}_6 = d\, C_2$ for all $S \in \mathcal{S}(K)$. Using the inclusion $\partial K \subset \bigcup_{S \in \mathcal{S}(K)} \partial S$, relation (2.3.33), the shape regularity (2.3.20) and the Cauchy inequality we find that

$$\|v\|^2_{L^2(\partial K)} \le \sum_{S \in \mathcal{S}(K)} \|v\|^2_{L^2(\partial S)} \tag{2.3.34}$$

$$
\begin{aligned}
&\leq \hat{C}_6 \left[\sum_{S\in\mathcal{S}(K)} \left(\|v\|_{L^2(S)} \, |v|_{H^1(S)} + h_S^{-1} \|v\|^2_{L^2(S)} \right) \right] \\
&\leq \hat{C}_6 \left[\left(\sum_{S\in\mathcal{S}(K)} \|v\|^2_{L^2(S)} \right)^{1/2} \left(\sum_{S\in\mathcal{S}(K)} |v|^2_{H^1(S)} \right)^{1/2} + \tilde{\kappa} h_K^{-1} \|v\|^2_{L^2(K)} \right] \\
&\leq C_6 \left(\|v\|_{L^2(K)} \, |v|_{H^1(K)} + h_K^{-1} \|v\|^2_{L^2(K)} \right),
\end{aligned}
$$

where $C_6 = \hat{C}_6 \max(1, \tilde{\kappa})$. $\square$

Lemma 2.3 *The following assertions hold:*

1. *If $v \in P^1(K_i)$, then $\pi_0 v|_{K_i} = v(x^i)$, $i \in I$.*
2. *If $v \in P^1(K_i)$, $x \in \Gamma_{ij}$, then*

$$\pi_0 v|_{K_i} = v(x) + \nabla v|_{K_i} \cdot (x^i - x), \quad i \in I, \ j \in S(i), \tag{2.3.35}$$

in particular,

$$\pi_0 v|_{K_i} = v(x^{ij}) + \nabla v|_{K_i} \cdot (x^i - x^{ij}). \tag{2.3.36}$$

3. *For $v \in S_h$,*

$$
\begin{aligned}
\pi_0 v|_{K_i} - \pi_0 v|_{K_j} &= [v(x)] - \{\nabla v|_{K_i} \cdot (x - x^i) - \nabla v|_{K_j} \cdot (x - x^j)\}, x \in \Gamma_{ij}, \\
\pi_0 v|_{K_i} - \pi_0 v|_{K_j} &= [v(x^{ij})] - \{\nabla v|_{K_i} \cdot (x^{ij} - x^i) - \nabla v|_{K_j} \cdot (x^{ij} - x^j)\}, \\
& \quad i \in I, \ j \in s(i).
\end{aligned} \tag{2.3.37}
$$

4. *For $v \in S_h$,*

$$[v(x^{ij})] \, |\Gamma_{ij}| = \int_{\Gamma_{ij}} [v] \, \mathrm{d}S, \quad i \in I, \ j \in s(i). \tag{2.3.38}$$

5. *The operator π_0 has the following properties:*

$$
\begin{aligned}
\|\pi_0 v\|_{L^2(K)} &\leq \|v\|_{L^2(K)}, \quad K \in \mathcal{T}_h, \ v \in L^2(K), && (2.3.39) \\
\|\pi_0 v\|_{L^2(\Omega)} &\leq \|v\|_{L^2(\Omega)}, \quad v \in L^2(\Omega), && (2.3.40) \\
\|v - \pi_0 v\|_{L^2(K)} &\leq \frac{h_K}{\pi} |v|_{H^1(K)}, \quad K \in \mathcal{T}_h, \ v \in H^1(K), && (2.3.41) \\
\|v - \pi_0 v\|_{L^2(\partial K)} &\leq C_7 \, h_K^{1/2} |v|_{H^1(K)}, \quad K \in \mathcal{T}_h, \ v \in H^1(K), && (2.3.42)
\end{aligned}
$$

with a constant $C_7 > 0$ independent of K, v, $h \in (0, h_0)$.

6. *There exist a constant $C_8 > 0$ independent of $h \in (0, h_0)$ and a mapping $\Pi : H^2(\Omega, \mathcal{T}_h) \to S_h$ such that*

$$
\begin{aligned}
\|\Pi v - v\|_{L^2(K)} &\leq C_8 h_K |v|_{H^1(K)}, \quad v \in H^1(K), \\
\|\Pi v - v\|_{L^2(K)} &\leq C_8 h_K^2 |v|_{H^2(K)}, \quad v \in H^2(K), \\
|\Pi v - v|_{H^1(K)} &\leq C_8 h_K |v|_{H^2(K)}, \quad v \in H^2(K), \\
& \quad K \in \mathcal{T}_h, \ h \in (0, h_0),
\end{aligned} \tag{2.3.43}
$$

$$
\begin{aligned}
\|\Pi v - v\|_{L^2(\Omega)} &\leq C_8 h |v|_{H^1(\Omega,\mathcal{T}_h)}, \quad v \in H^1(\Omega,\mathcal{T}_h), \\
\|\Pi v - v\|_{L^2(\Omega)} &\leq C_8 h^2 |v|_{H^2(\Omega,\mathcal{T}_h)}, \quad v \in H^2(\Omega,\mathcal{T}_h), \\
|\Pi v - v|_{H^1(\Omega,\mathcal{T}_h)} &\leq C_8 h |v|_{H^2(\Omega,\mathcal{T}_h)}, \quad v \in H^2(\Omega,\mathcal{T}_h), \\
& h \in (0, h_0).
\end{aligned} \tag{2.3.44}
$$

Proof. Assertions 1 – 3 are obtained easily from the definition of the operator π_0 and linearity of v on K. Assertion 4 is a consequence of the fact that the midpoint rule for numerical integration is exact for linear functions.

In order to prove estimate (2.3.39), we use the definition (2.3.3) of π_0 and apply the Cauchy inequality. We get

$$
\|\pi_0 v\|^2_{L^2(K)} = \int_K \frac{1}{|K|^2} \left(\int_K u \,\mathrm{d}x \right)^2 \,\mathrm{d}x \leq \frac{1}{|K^{2|}} |K| \int_K u^2 \,\mathrm{d}x \int_K \mathrm{d}x = \|u\|^2_{L^2(K)}.
$$

Inequality (2.3.40) immediately follows from (2.3.39). Further, the proof of estimate (2.3.41) is a consequence of [PW61], where the inequality

$$
\frac{\int_K |\nabla \eta|^2 \,\mathrm{d}x}{\int_K |\eta|^2 \,\mathrm{d}x} \geq \frac{\pi^2}{h_K^2} \tag{2.3.45}
$$

was established for all $\eta \in H^1(K)$ such that $\int_K \eta \,\mathrm{d}x = 0$. Substituting here $\eta := v - \pi_0 v$ for $v \in H^1(K)$, we obtain (2.3.41). Now, we can prove (2.3.42) with the aid of (2.3.41) and (2.3.32).

Finally, let us prove Assertion 6. The centre of gravity x_K of an element $K \in \mathcal{T}_h$ is expressed

$$
x_K = \frac{1}{|K|} \int_K x \,\mathrm{d}x \in I\!R^d. \tag{2.3.46}
$$

Then

$$
\int_K (x - x_K) \,\mathrm{d}x = 0. \tag{2.3.47}
$$

Let $v \in H^2(K)$. Then $\nabla v \in H^1(K)^d$. By (2.3.3)

$$
\pi_0 v = \frac{1}{|K|} \int_K v \,\mathrm{d}x, \quad \pi_0(\nabla v) = \frac{1}{|K|} \int_K \nabla v \,\mathrm{d}x. \tag{2.3.48}
$$

Then we define the interpolation Πv as

$$
\Pi v = \pi_0 v + (x - x_K) \cdot \pi_0(\nabla v) \in P^1(K). \tag{2.3.49}
$$

Obviously,

$$
\nabla(\Pi v) = \pi_0(\nabla v), \quad \int_K (\nabla(\Pi v) - \pi_0(\nabla v)) \,\mathrm{d}x = 0 \tag{2.3.50}
$$

and, in view of (2.3.48), (2.3.49) and (2.3.47),

$$
\begin{aligned}
\int_K (v - \Pi v) \,\mathrm{d}x &= \int_K (v - \pi_0 v - (x - x_K) \cdot \pi_0(\nabla v)) \,\mathrm{d}x \\
= \pi_0(\nabla v) \cdot \int_K (x - x_K) \,\mathrm{d}x &= 0.
\end{aligned} \tag{2.3.51}
$$

Hence, by (2.3.51) and (2.3.45)

$$\|v - \Pi v\|_{L^2(K)} \le \frac{h_K}{\pi} |v - \Pi v|_{H^1(K)}. \tag{2.3.52}$$

Due to (2.3.50) and (2.3.41),

$$|v - \Pi v|^2_{H^1(K)} = \sum_{j=1}^{d} \left\| \frac{\partial v}{\partial x_j} - \pi_0 \frac{\partial v}{\partial x_j} \right\|^2_{L^2(K)} \le \frac{h_K^2}{\pi^2} \sum_{j=1}^{d} \left| \frac{\partial v}{\partial x_j} \right|^2_{H^1(K)} = \frac{h_K^2}{\pi^2} |v|^2_{H^2(K)}.$$

Hence,

$$|v - \Pi v|_{H^1(K)} \le \frac{h_K}{\pi} |v|_{H^2(K)}, \tag{2.3.53}$$

which yields the third inequality in (2.3.43). Now, (2.3.52) and (2.3.53) imply that

$$\|v - \Pi v\|_{L^2(K)} \le \frac{h_K^2}{\pi^2} |v|_{H^2(K)} \tag{2.3.54}$$

(the second inequality in (2.3.43)). Further in virtue of (2.3.52), (2.3.50) and (2.3.39),

$$\|v - \Pi v\|_{L^2(K)} \le \frac{h_K}{\pi} \left(|v|_{H^1(K)} + |\Pi v|_{H^1(K)} \right) \tag{2.3.55}$$

$$= \frac{h_K}{\pi} \left\{ |v|_{H^1(K)} + \left(\sum_{j=1}^{d} \|\pi_0 \frac{\partial v}{\partial x_j}\|^2_{L^2(K)} \right)^{1/2} \right\}$$

$$\le \frac{h_K}{\pi} \left\{ |v|_{H^1(K)} + \left(\sum_{j=1}^{d} \|\frac{\partial v}{\partial x_j}\|^2_{L^2(K)} \right)^{1/2} \right\} = 2 \frac{h_K}{\pi} |v|_{H^1(K)},$$

which gives the first inequality in (2.3.43). We see that we can set $C_8 = 2/\pi$ in (2.3.43). Finally, the estimates (2.3.44) are obtained from (2.3.43) by squaring and summing. $\square$

Remark 2.1 *The operator Π is not L^2- projector on S_h, which we denote by Π^{L^2}: if $v \in L^2(\Omega)$, then*

$$\left(\Pi^{L^2} v, \varphi \right) = (v, \varphi) \quad \forall \varphi \in S_h. \tag{2.3.56}$$

It is clear that $\Pi v = v$ for $v \in S_h$. Then we have

$$\|v - \Pi^{L^2} v\|_{L^2(\Omega)} \le \|v - \Pi v\|_{L^2(\Omega)} \le C_8 h^p |v|_{H^p(\Omega)}, \ v \in H^p(\Omega), \ p = 1, 2, \tag{2.3.57}$$

as follows from (2.3.44).

2.3.3 Properties of the form $\bar{b}_h$

Let assumptions (2.2.4), a), Assumptions (H) and Assumptions in Section 2.3.2 be satisfied. Now we shall be concerned with properties of the form $\bar{b}_h$. First we investigate its boundedness.

Lemma 2.4 *There exists a constant $C_9 > 0$ such that*

$$|\bar{b}_h(u, v)| \le C_9 \left(J_h^\sigma(v, v)^{1/2} + |v|_{H^1(\Omega, \mathcal{T}_h)} \right) \|u\|_{L^2(\Omega)}, \tag{2.3.58}$$
$$u \in L^2(\Omega), \ v \in S_h, \ h \in (0, h_0).$$

Proof. In view of (2.3.6), for $u \in L^2(\Omega)$ and $v \in S_h$,

$$\bar{b}_h(u,v) = \sum_{i\in I} \pi_0 v|_{K_i} \left\{ \sum_{j\in s(i)} H(\pi_0 u|_{K_i}, \pi_0 u|_{K_j}, \boldsymbol{n}_{ij})|\Gamma_{ij}| + \sum_{j\in\gamma(i)} H(\pi_0 u|_{K_i}, \pi_0 u|_{K_i}, \boldsymbol{n}_{ij})\, |\Gamma_{ij}| \right\}.$$

Due to the conservativity (2.3.9) of the numerical flux, this can be rewritten in the form

$$\begin{aligned} \bar{b}_h(u,v) &= \sum_{i\in I} \left\{ \frac{1}{2} \sum_{j\in s(i)} (\pi_0 v|_{K_i} - \pi_0 v|_{K_j})\, H(\pi_0 u|_{K_i}, \pi_0 u|_{K_j}, \boldsymbol{n}_{ij})\, |\Gamma_{ij}| \right. \\ &\quad \left. + \sum_{j\in\gamma(i)} \pi_0 v\, H(\pi_0 u|_{K_i}, \pi_0 u|_{K_i}, \boldsymbol{n}_{ij})\, |\Gamma_{ij}| \right\}. \end{aligned}$$

Using Assertions 1 – 4 of Lemma 2.3, the Lipschitz-continuity (2.3.7) of H and inequalities (2.2.17) and (2.3.21), we find that

$$\begin{aligned} |\bar{b}_h(u,v)| &= \left| \sum_{i\in I} \left\{ \frac{1}{2} \sum_{j\in s(i)} H(\pi_0 u|_{K_i}, \pi_0 u|_{K_j}, \boldsymbol{n}_{ij}) \right. \right. \\ &\quad \times \left([v(x^{ij})] - (\nabla v|_{K_i} \cdot (x^{ij} - x^i) - \nabla v|_{K_j} \cdot (x^{ij} - x^j)) \right) |\Gamma_{ij}| + \\ &\quad \left. \left. + \sum_{j\in\gamma(i)} H(\pi_0 u|_{K_i}, \pi_0 u|_{K_i}, \boldsymbol{n}_{ij})\, (v(x^{ij}) - \nabla v|_{K_i} \cdot (x^{ij} - x^i))\, |\Gamma_{ij}| \right\} \right| \\ &\le C_1 \sum_{i\in I} \left\{ \frac{1}{2} \sum_{j\in s(i)} \left(|\pi_0 u|_{K_i}| + |\pi_0 u|_{K_j}| \right) \left(\left| \int_{\Gamma_{ij}} [v]\, \mathrm{d}S \right| + \left(|\nabla v|_{K_i}|\, h_{K_i} + |\nabla v|_{K_j}| h_{K_j} \right) |\Gamma_{ij}| \right) \right. \\ &\quad \left. + 2 \sum_{j\in\gamma(i)} |\pi_0 u|_{K_i}| \left(\left| \int_{\Gamma_{ij}} v\, \mathrm{d}S \right| + C_3 |\nabla v|_{K_i}|\, |K_i| \right) \right\}. \end{aligned} \tag{2.3.59}$$

Further, the Cauchy inequality, and similar considerations as above yield

$$\begin{aligned} |\pi_0 u|_{K_i}| \left| \int_{\Gamma_{ij}} [v]\, \mathrm{d}S \right| &\le \int_{\Gamma_{ij}} |\pi_0 u|_{K_i}|\, |[v]|\, \mathrm{d}S \\ &= \int_{\Gamma_{ij}} \left\{ d(\Gamma_{ij})^{1/2}\, |\pi_0 u|_{K_i}| \right\} \left\{ \frac{1}{d(\Gamma_{ij})^{1/2}} |[v]| \right\} \mathrm{d}S \\ &\le \left(\int_{\Gamma_{ij}} d(\Gamma_{ij}) |\pi_0 u|_{K_i}|^2\, \mathrm{d}S \right)^{1/2} \left(\int_{\Gamma_{ij}} \frac{1}{d(\Gamma_{ij})} [v]^2\, \mathrm{d}S \right)^{1/2} \\ &\le \left(h^d_{K_i} |\pi_0 u|_{K_i}|^2 \right)^{1/2} \left(\int_{\Gamma_{ij}} \frac{1}{d(\Gamma_{ij})} [v]^2\, \mathrm{d}S \right)^{1/2} \\ &\le \sqrt{C_3} \|\pi_0 u\|_{L^2(K_i)} \left(\int_{\Gamma_{ij}} \frac{1}{d(\Gamma_{ij})} [v]^2\, \mathrm{d}S \right)^{1/2}. \end{aligned} \tag{2.3.60}$$

We obtain similar estimates, if we write v instead of $[v]$ in the case when $\Gamma_{ij} \subset \partial\Omega$. Further, in virtue of inequalities (2.2.17) and (2.3.21),

$$|\Gamma_{ij}|^2\, h^2_{K_j} = h^2_{K_j}\, |\Gamma_{ij}|^{\frac{d-2}{d-1}}\, |\Gamma_{ij}|^{\frac{d}{d-1}}$$

$$\leq h_{K_j}^d \, h_{K_i}^d \leq C_3^2 |K_j| \, |K_i|.$$

This implies that

$$|\pi_0 u|_{K_i} \left| |\nabla v|_{K_j} \right| \, |\Gamma_{ij}| \, h_{K_j} \leq C_3 \|\pi_0 u\|_{L^2(K_i)} |v|_{H^1(K_j)}. \tag{2.3.61}$$

Now, from (2.3.59) – (2.3.61) we obtain the estimate

$$\begin{aligned}
&|\bar{b}_h(u,v)| \leq \qquad\qquad (2.3.62)\\
\leq\ & C_1 C_3^{1/2} \sum_{i\in I} \left\{ \frac{1}{2} \sum_{j\in(s(i))} \left(\|\pi_0 u\|_{L^2(K_i)} + \|\pi_0 u\|_{L^2(K_j)} \right) \left(\int_{\Gamma_{ij}} \frac{1}{d(\Gamma_{ij})} [v]^2 \, \mathrm{d}S \right)^{1/2} \right.\\
& + C_3^{1/2} \sum_{j\in s(i)} \left(\|\pi_0 u\|_{L^2(K_i)} + \|\pi_0 u\|_{L^2(K_j)} \right) \left(|v|_{H^1(K_i)} + |v|_{H^1(K_j)} \right)\\
& \left. + 2 \sum_{j\in\gamma(i)} \|\pi_0 u\|_{L^2(K_i)} \left(\left(\int_{\Gamma_{ij}} \frac{1}{d(\Gamma_{ij})} v^2 \, \mathrm{d}S \right)^{1/2} + C_3^{1/2} |v|_{H^1(K_i)} \right) \right\}.
\end{aligned}$$

Estimate (2.3.62), assumption (2.3.22) and the discrete Cauchy inequality yield

$$\begin{aligned}
&|\bar{b}_h(u,v)| \leq C^* \left(\sum_{i\in I} \|\pi_0 u\|_{L^2(K_i)}^2 \right)^{1/2} \qquad\qquad (2.3.63)\\
&\times \left\{ \left(\sum_{i\in I} \left(\sum_{j\in s(i)} \int_{\Gamma_{ij}} \frac{1}{d(\Gamma_{ij})} [v]^2 \, \mathrm{d}S + \sum_{j\in\Gamma(i)} \int_{\Gamma_{ij}} \frac{1}{d(\Gamma_{ij})} v^2 \, \mathrm{d}S \right) \right)^{1/2} \right.\\
& \left. + \left(\sum_{i\in I} |v|_{H^1(K_i)}^2 \right)^{1/2} \right\},
\end{aligned}$$

where the constant C^* depends on C_1, C_3 and C_4 and is independent of h, u, v. From this and (2.3.39), the desired estimate (2.3.58) immediately follows. □

The next step is to prove the Lipschitz-continuity of the form $\bar{b}_h$.

Lemma 2.5 *There exists a constant $C_{10} > 0$ such that*

$$\begin{aligned}
&|\bar{b}_h(u_h, v_h) - \bar{b}_h(u, v_h)| \leq \qquad\qquad (2.3.64)\\
\leq\ & C_{10} \left(J_h^\sigma(v_h, v_h)^{1/2} + |v_h|_{H^1(\Omega,\mathcal{T}_h)} \right) \|u - u_h\|_{L^2(\Omega)},\\
& u_h \in S_h, \ u \in L^2(\Omega), \ v_h \in S_h, \ h \in (0, h_0).
\end{aligned}$$

Proof. By (2.3.6), for $u \in L^2(\Omega)$ and $u_h, v_h \in S_h$,

$$\begin{aligned}
&\bar{b}_h(u_h, v_h) - \bar{b}_h(u, v_h) =\\
=\ & \sum_{i\in I} \pi_0 v_h|_{K_i} \left\{ \sum_{j\in s(i)} \left(H(\pi_0 u_h|_{K_i}, \pi_0 u_h|_{K_j}, \boldsymbol{n}_{ij}) - H(\pi_0 u|_{K_i}, \pi_0 u|_{K_j}, \boldsymbol{n}_{ij}) \right) |\Gamma_{ij}| \right.\\
& \left. + \sum_{j\in\gamma(i)} \left(H(\pi_0 u_h|_{K_i}, \pi_0 u_h|_{K_i}, \boldsymbol{n}_{ij}) - H(\pi_0 u|_{K_i}, \pi_0 u|_{K_i}, \boldsymbol{n}_{ij}) \right) |\Gamma_{ij}| \right\}.
\end{aligned}$$

From the Lipschitz-continuity (2.3.7) and conservativity (2.3.9) of H and Lemma 2.3, we find that

$$
\begin{aligned}
&|\bar{b}_h(u_h, v_h) - \bar{b}_h(u, v_h)| \le \\
\le\ & C_1 \sum_{i\in I} \Bigg\{ \frac{1}{2} \sum_{j\in s(i)} \left(|\pi_0(u_h-u)|_{K_i}| + |\pi_0(u_h-u)|_{K_j}| \right) \Bigg(\left| \int_{\Gamma_{ij}} [v_h]\, \mathrm{d}S \right| \\
&+ \left(|\nabla v_h|_{K_i}|\, h_{K_i} + |\nabla v_h|_{K_j}|\, h_{K_j} \right) |\Gamma_{ij}| \Bigg) \\
&+ 2 \sum_{j\in\gamma(i)} |\pi_0(u_h-u)|_{K_i}| \left(\left| \int_{\Gamma_{ij}} v_h \,\mathrm{d}S \right| + C_3 |\nabla v_h|_{K_i}|\, |K_i| \right) \Bigg\}.
\end{aligned}
$$

Now we proceed quite analogously as in the proof of Lemma 2.4, writing $u_h - u$ instead of u and v_h instead of v. □

In what follow we shall be concerned with the *consistency error* of the form $\bar{b}_h$.

Lemma 2.6 *There exists a constant $C_{11} > 0$ such that*

$$
\begin{aligned}
&|\bar{b}_h(u,v) - b(u,v)| \le \\
\le\ & C_{11}\, h|u|_{H^1(\Omega)} \left(J_h^\sigma(v,v)^{1/2} + |v|_{H^1(\Omega,\mathcal{T}_h)} \left(\Phi(\|u\|_{L^\infty(\Omega)}) + 1 \right) \right), \\
& \qquad u \in H^1(\Omega) \cap L^\infty(\Omega),\ v \in S_h,\ h \in (0, h_0),
\end{aligned}
\tag{2.3.65}
$$

where we define, for $M \ge 0$,

$$
\Phi(M) = \max_{\substack{\xi\in[-M,M] \\ s=1,\dots,d}} |f_s'(\xi)|. \tag{2.3.66}
$$

Proof. Let $u \in H^1(\Omega) \cap L^\infty(\Omega)$, $\|u\|_{L^\infty(\Omega)} = M$, $v \in S_h$. We can write

$$
\bar{b}_h(u,v) - b(u,v) = \sigma_1 + \sigma_2, \tag{2.3.67}
$$

where

$$
\sigma_1 := \bar{b}_h(u,v) - b(u, \pi_0 v), \quad \sigma_2 := b(u, \pi_0 v) - b(u,v).
$$

a) First let us deal with the term σ_2. By the definition (2.2.7) of the form b we have

$$
\sigma_2 = \sum_{i\in I} \int_{K_i} \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} (\pi_0 v - v)\, \mathrm{d}x = \sum_{i\in I} \int_{K_i} \sum_{s=1}^{d} f_s'(u) \frac{\partial u}{\partial x_s} (\pi_0 v - v)\, \mathrm{d}x.
$$

Using the continuity of f_s', the boundedness of u and the Cauchy inequality, we immediately get the estimate

$$
|\sigma_2| \le 2 \max_{\substack{\xi\in[-M,M] \\ s=1,\dots,d}} |f_s'(\xi)|\, |u|_{H^1(\Omega)} \|\pi_0 v - v\|_{L^2(\Omega)}.
$$

This and Lemma 2.3, Assertion 5, imply that

$$
|\sigma_2| \le \frac{2}{\pi} \Phi(M)\, h|u|_{H^1(\Omega)} |v|_{H^1(\Omega,\mathcal{T}_h)}. \tag{2.3.68}
$$

b) Furthermore, we shall estimate the term σ_1. Taking into account that $\pi_0 v|_{K_i} = \text{const}$ and using Green's theorem, relations (2.2.16) and the consistency of H, we can write

$$\begin{aligned} b(u, \pi_0 v) &= \sum_{i\in I} \int_{K_i} \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} \pi_0 v \,\mathrm{d}x = \sum_{i\in I} \pi_0 v|_{K_i} \int_{K_i} \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} \,\mathrm{d}x \\ &= \sum_{i\in I} \pi_0 v|_{K_i} \int_{\partial K_i} \sum_{s=1}^{d} f_s(u)\, n_s \,\mathrm{d}S = \sum_{i\in I} \pi_0 v|_{K_i} \sum_{j\in S(i)} \int_{\Gamma_{ij}} H(u, u, \boldsymbol{n}_{ij}) \,\mathrm{d}S. \end{aligned}$$

Hence,

$$\begin{aligned} \sigma_1 &= \bar{b}_h(u, v) - b(u, \pi_0 v) \\ &= \sum_{i\in I} \pi_0 v|_{K_i} \Bigg\{ \sum_{j\in s(i)} \left(H(\pi_0 u|_{K_i}, \pi_0 u|_{K_j}, \boldsymbol{n}_{ij})\, |\Gamma_{ij}| - \int_{\Gamma_{ij}} H(u, u, \boldsymbol{n}_{ij}) \,\mathrm{d}S \right) \\ &\quad + \sum_{j\in\gamma(i)} \left(H(\pi_0 u|_{K_i}, \pi_0 u|_{K_i}, \boldsymbol{n}_{ij})\, |\Gamma_{ij}| - \int_{\Gamma_{ij}} H(u, u, \boldsymbol{n}_{ij}) \,\mathrm{d}S \right) \Bigg\}. \end{aligned}$$

Due to the conservativity of H (see (2.3.9)), the term σ_1 can be expressed in the form

$$\begin{aligned} \sigma_1 &= \frac{1}{2} \sum_{i\in I} \sum_{j\in s(i)} \left\{ H(\pi_0 u|_{K_i}, \pi_0 u|_{K_j}, \boldsymbol{n}_{ij})\, |\Gamma_{ij}| - \int_{\Gamma_{ij}} H(u, u, \boldsymbol{n}_{ij}) \,\mathrm{d}S \right\} \\ &\quad \times (\pi_0 v|_{K_i} - \pi_0 v|_{K_j}) + \sum_{i\in I} \sum_{j\in\gamma(i)} \Big\{ H(\pi_0 u|_{K_i}, \pi_0 u|_{K_i}, \boldsymbol{n}_{ij})\, |\Gamma_{ij}| \\ &\quad - \int_{\Gamma_{ij}} H(u, u, \boldsymbol{n}_{ij}) \,\mathrm{d}S \Big\} \pi_0 v|_{K_i}. \end{aligned}$$

This and the Lipschitz-continuity of the numerical flux yield the estimate

$$\begin{aligned} |\sigma_1| &\le \frac{1}{2} C_1 \sum_{i\in I} \sum_{j\in s(i)} \int_{\Gamma_{ij}} \Big(|\pi_0 u|_{K_i} - u| + |\pi_0 u|_{K_j} - u| \Big)\, |[\pi_0 v]| \,\mathrm{d}S \\ &\quad + 2C_1 \sum_{i\in I} \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} |\pi_0 u|_{K_i} - u|\, |\pi_0 v| \,\mathrm{d}S. \end{aligned}$$

Now, the Cauchy inequality and the inequality $(a+b)^2 \le 2(a^2 + b^2)$ imply that

$$\begin{aligned} |\sigma_1| &\le C_1 \left\{ \sum_{i\in I} \sum_{j\in s(i)} \left(\int_{\Gamma_{ij}} |\pi_0 u|_{K_i} - u|^2 \,\mathrm{d}S + \int_{\Gamma_{ij}} |\pi_0 u|_{K_j} - u|^2 \,\mathrm{d}S \right) \right\}^{1/2} \qquad (2.3.69) \\ &\times \left(\sum_{i\in I} \sum_{j\in s(i)} \int_{\Gamma_{ij}} [\pi_0 v]^2 \mathrm{d}S \right)^{1/2} + 2C_1 \left(\sum_{i\in I} \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} |\pi_0 u|_{K_i} - u|^2 \,\mathrm{d}S \right)^{1/2} \\ &\times \left(\sum_{i\in I} \sum_{j\in s(i)} \int_{\Gamma_{ij}} |\pi_0 v|^2 \,\mathrm{d}S \right)^{1/2} \le \end{aligned}$$

$$
\begin{aligned}
&\leq \sqrt{2}C_1 \left(\sum_{i\in I} \|\pi_0 u - u\|^2_{L^2(\partial K_i)} \right)^{1/2} \left(\sum_{i\in I} \sum_{j\in s(i)} \int_{\Gamma_{ij}} [\pi_0 v]^2 \, \mathrm{d}S \right)^{1/2} \\
&+2C_1 \left(\sum_{i\in I} \|\pi_0 u - u\|^2_{L^2(\partial K_i)} \right)^{1/2} \left(\sum_{i\in I} \sum_{j\in \gamma(i)} \int_{\Gamma_{ij}} |\pi_0 v|^2 \mathrm{d}S \right)^{1/2} .
\end{aligned}
$$

From Lemma 2.3, estimate (2.3.42), we see that

$$
\sum_{i\in I} \|\pi_0 u - u\|^2_{L^2(\partial K_i)} \leq C_7^2 h \sum_{i\in I} |u|^2_{H^1(K_i)} = C_7^2 h |u|^2_{H^1(\Omega)}. \tag{2.3.70}
$$

Moreover, in view of Lemma 2.3, Assertions 2) and 3),

$$
\begin{aligned}
[\pi_0 v]|^2_{\Gamma_{ij}} &\leq 3[v]^2(x) + 3\left(|\nabla v|_{K_i}|^2 \, h^2_{K_i} + |\nabla v|_{K_j}|^2 \, h^2_{K_j} \right), \\
(\pi_0 v|_{K_i})^2 &\leq 2v^2(x) + 2|\nabla v|_{K_i}|^2 h^2_{K_i} \quad \text{for } x \in \Gamma_{ij}.
\end{aligned} \tag{2.3.71}
$$

From (2.3.69) – (2.3.71) and (2.2.17) we find that

$$
\begin{aligned}
|\sigma_1| \leq\; & \sqrt{2}C_1 \, C_7 \, h^{1/2} |u|_{H^1(\Omega)} \times \\
& \times \left\{ 3h \sum_{i\in I} \sum_{j\in s(i)} \left(\int_{\Gamma_{ij}} \frac{[v]}{d(\Gamma_{ij})} \, \mathrm{d}S + \left(|\nabla v|_{K_i}|^2 \, h^d_{K_i} + |\nabla v|_{K_j}|^2 \, h^d_{K_j} \right) \right) \right\}^{1/2} \\
& +2C_1 \, C_7 \, h^{1/2} |u|_{H^1(\Omega)} \left\{ 2h \sum_{i\in I} \sum_{j\in \gamma(i)} \left(\int_{\Gamma_{ij}} \frac{v^2}{d(\Gamma_{ij})} \, \mathrm{d}S + h^d_{K_i} |\nabla v|_{K_i}|^2 \right) \right\}^{1/2} .
\end{aligned}
$$

Using inequality (2.3.21), which implies that $|\nabla v|_{K_i}|^2 h^d_{K_i} \leq C_3 |v|^2_{H^1(K_i)}$, after some manipulation, we obtain the estimate

$$
|\sigma_1| \leq C \, h |u|_{H^1(\Omega)} \left(J_h^\sigma(v,v)^{1/2} + |v|_{H^1(\Omega,\mathcal{T}_h)} \right)
$$

with a constant dependent on C_1, C_3, C_7. This and (2.3.68) already yield (2.3.65). □

2.3.4 Error estimates for the method of lines

Our goal is now to analyze the error estimates of the approximate solution u_h obtained by the method of lines (2.3.15) under the assumption that the exact solution u is a strong one, i.e., it satisfies (2.2.12). We set

$$
M = \|u\|_{L^\infty(Q_T)}. \tag{2.3.72}
$$

Let $\Pi u(t)$ be the S_h-interpolation of $u(t)$ $(t \in [0,T])$ from Lemma 2.3. We set

$$
\xi = u_h - \Pi u \in S_h, \quad \eta = \Pi u - u \in H^2(\Omega, \mathcal{T}_h). \tag{2.3.73}
$$

Then

$$
u_h - u = \xi + \eta. \tag{2.3.74}
$$

From (2.3.15) and (2.3.16), where we substitute $v_h := \xi$, it follows that

$$\begin{aligned} & \left(\frac{\partial \xi}{\partial t}, \xi\right) + a_h(\xi,\xi) + \varepsilon J_h^\sigma(\xi,\xi) \\ = \quad & b(u,\xi) - \bar{b}_h(u_h,\xi) - \left(\frac{\partial \eta}{\partial t}, \xi\right) - a_h(\eta,\xi) - \varepsilon J_h^\sigma(\eta,\xi). \end{aligned} \tag{2.3.75}$$

(Of course, $\xi = \xi(t),\ \eta = \eta(t)$ for $t \in [0,T]$, but we do not emphasize the dependence on t by our notation, if it is not necessary.) In what follows, we estimate the individual terms on the right-hand side of (2.3.75).

In virtue of (2.2.12),

$$\left(\frac{\partial \xi}{\partial t}, \xi\right) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\Omega)}. \tag{2.3.76}$$

The Cauchy inequality implies that

$$\left|\left(\frac{\partial \eta}{\partial t}, \xi\right)\right| \leq \left\|\frac{\partial \eta}{\partial t}\right\|_{L^2(\Omega)} \|\xi\|_{L^2(\Omega)}, \tag{2.3.77}$$

$$J_h^\sigma(\eta,\xi) \leq \left(J_h^\sigma(\eta,\eta)\right)^{1/2}\left(J_h^\sigma(\xi,\xi)\right)^{1/2}. \tag{2.3.78}$$

Lemma 2.7 *We have*

$$\begin{aligned} \|\eta\|_{L^2(\Omega)} &\leq C_8 h|u|_{H^1(\Omega)}, \\ \|\eta\|_{L^2(\Omega)} &\leq C_8 h^2|u|_{H^2(\Omega)}, \\ |\eta|_{H^1(\Omega,\mathcal{T}_h)} &\leq C_8 h|u|_{H^2(\Omega)}, \\ \left\|\frac{\partial \eta}{\partial t}\right\|_{L^2(\Omega)} &\leq C_8 h\left|\frac{\partial u}{\partial t}\right|_{H^1(\Omega)}, \quad h \in (0,h_0). \end{aligned} \tag{2.3.79}$$

Proof is a consequence of (2.2.12), (2.3.44) and the relation $\Pi(\partial u/\partial t) = \partial(\Pi u)/\partial t$. □

Lemma 2.8 *We have*

$$\begin{aligned} & |b(u,\xi) - \bar{b}_h(u_h,\xi)| \\ \leq\ & C_{12}\left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right)\left(\|\xi\|_{L^2(\Omega)} + h^2|u|_{H^2(\Omega)} + h|u|_{H^1(\Omega)}\right), \quad h \in (0,h_0), \end{aligned} \tag{2.3.80}$$

where $C_{12} > 0$ is a constant dependent on $M = \|u\|_{L^\infty(Q_T)}$ but independent of h and ξ.

Proof. We can write

$$\begin{aligned} & b(u,\xi) - \bar{b}_h(u_h,\xi) \\ = \quad & (b(u,\xi) - \bar{b}_h(u,\xi)) + (\bar{b}_h(u,\xi) - \bar{b}_h(\Pi u,\xi)) + (\bar{b}_h(\Pi u,\xi) - \bar{b}_h(u_h,\xi)). \end{aligned} \tag{2.3.81}$$

From Lemma 2.6 and (2.3.72) we find that

$$\begin{aligned} & |b(u,\xi) - \bar{b}_h(u,\xi)| \leq \\ \leq\ & C_{11} h|u|_{H^1(\Omega)}(\Phi(M)+1)\left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right) \\ \leq\ & Ch\left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right)|u|_{H^1(\Omega)} \end{aligned} \tag{2.3.82}$$

with a constant $C > 0$ dependent on u, but independent of h and ξ. Further, by Lemma 2.5 and Assertion 6 of Lemma 2.3,

$$\begin{aligned}
& |\bar{b}_h(u,\xi) - \bar{b}_h(\Pi u,\xi)| \qquad (2.3.83)\\
\le\ & C_{10}\left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right)\|u-\Pi u\|_{L^2(\Omega)}\\
\le\ & C_{10}C_8h^2\left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right)|u|_{H^2(\Omega)}
\end{aligned}$$

and

$$\begin{aligned}
& |\bar{b}_h(\Pi u,\xi) - \bar{b}_h(u_h,\xi)| \qquad (2.3.84)\\
\le\ & C_{10}\left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right)\|\xi\|_{L^2(\Omega)}.
\end{aligned}$$

From (2.3.82) – (2.3.84) we obtain (2.3.80). □

Lemma 2.9 *There exists a constant $C_{13} > 0$ independent of u, h, ξ, ε such that*

$$|a_h(\eta,\xi)| \le C_{13}\varepsilon h|u|_{H^2(\Omega)}\left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right), \quad h\in(0,h_0). \qquad (2.3.85)$$

Proof. In view of (2.3.11), we can write

$$\begin{aligned}
a_h(\eta,\xi) &= \sigma_1+\sigma_2+\sigma_3, \qquad (2.3.86)\\
\sigma_1 &:= \varepsilon\sum_{i\in I}\int_{K_i}\nabla\eta\cdot\nabla\xi\,\mathrm{d}x,\\
\sigma_2 &:= -\varepsilon\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\langle\nabla\eta\rangle\cdot\boldsymbol{n}_{ij}[\xi]\,\mathrm{d}S\\
&\quad +\varepsilon\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\langle\nabla\xi\rangle\cdot\boldsymbol{n}_{ij}[\eta]\,\mathrm{d}S\\
\sigma_3 &:= -\varepsilon\sum_{i\in I}\sum_{j\in\gamma(i)}\int_{\Gamma_{ij}}(\nabla\eta\cdot\boldsymbol{n}_{ij})\,\xi\,\mathrm{d}S\\
&\quad +\varepsilon\sum_{i\in I}\sum_{j\in\gamma(i)}\int_{\Gamma_{ij}}(\nabla\xi\cdot\boldsymbol{n}_{ij})\,\eta\,\mathrm{d}S.
\end{aligned}$$

We estimate the individual terms in (2.3.86) similarly as in [HSS00]. The Cauchy inequality and estimates (2.3.44) imply that

$$|\sigma_1| \le \varepsilon C_8h|u|_{H^2(\Omega)}|\xi|_{H^1(\Omega,\mathcal{T}_h)}. \qquad (2.3.87)$$

In order to estimate σ_2, we set $\gamma_{ij} := d(\Gamma_{ij})$. The Cauchy inequality, (2.3.13)-(2.3.14) and (2.2.17) yield the estimates

$$\begin{aligned}
|\sigma_2| \le\ & \varepsilon\left(\sum_{i\in I}\sum_{j\in s(i)}\gamma_{ij}\|\nabla\eta\|^2_{L^2(\Gamma_{ij})}\right)^{1/2}\left(\sum_{i\in I}\sum_{j\in s(i)}\frac{1}{\gamma_{ij}}\|[\xi]\|^2_{L^2(\Gamma_{ij})}\right)^{1/2} \qquad (2.3.88)\\
&+\varepsilon\left(\sum_{i\in I}\sum_{j\in s(i)}\gamma_{ij}\|\nabla\xi\|^2_{L^2(\Gamma_{ij})}\right)^{1/2}\left(\sum_{i\in I}\sum_{j\in s(i)}\frac{1}{\gamma_{ij}}\|[\eta]\|^2_{L^2(\Gamma_{ij})}\right)^{1/2}
\end{aligned}$$

$$\leq \varepsilon \left(\sum_{i \in I} h_{K_i} \|\nabla \eta\|^2_{L^2(\partial K_i)} \right)^{1/2} J_h^\sigma(\xi,\xi)^{1/2}$$
$$+\varepsilon \left(\sum_{i \in I} h_{K_i} \|\nabla \xi\|^2_{L^2(\partial K_i)} \right)^{1/2} J_h^\sigma(\eta,\eta)^{1/2}.$$

Taking into account that $\eta = \Pi u - u \in H^2(\Omega, \mathcal{T}_h)$ and $\Pi u \in S_h$, we have

$$\|\nabla \eta\|_{L^2(K_i)} = |\eta|_{H^1(K_i)}, \quad |\nabla \eta|_{H^1(K_i)} = |\eta|_{H^2(K_i)} = |u|_{H^2(K_i)}, i \in I. \tag{2.3.89}$$

From (2.3.32) (where we set $v := |\nabla \eta|$) and (2.3.43) we find that

$$\|\nabla \eta\|^2_{L^2(\partial K_i)} \leq 2C_6 C_8^2 |u|^2_{H^2(K_i)} h_{K_i}. \tag{2.3.90}$$

Hence,

$$\left(\sum_{i \in I} h_{K_i} \|\nabla \eta\|^2_{L^2(\partial K_i)} \right)^{1/2} \leq \sqrt{2C_6}\, C_8\, h |u|_{H^2(\Omega)}. \tag{2.3.91}$$

Since $\xi = u_h - \Pi u \in S_h$ and thus, $|\xi|_{H^2(K_i)} = 0$, from (2.3.32) we get

$$\left(\sum_{i \in I} h_{K_i} \|\nabla \xi\|^2_{L^2(\partial K_i)} \right)^{1/2} \leq C_6 |\xi|_{H^1(\Omega, \mathcal{T}_h)}. \tag{2.3.92}$$

Now it remains to estimate the term $J_h^\sigma(\eta,\eta)$. In virtue of (2.3.13), (2.3.23), (2.3.32) and (2.3.43), we have

$$\begin{aligned} J_h^\sigma(\eta,\eta) &\leq \sum_{i \in I} \sum_{j \in s(i)} \int_{\partial \Gamma_{ij}} \frac{1}{\gamma_{ij}} |\eta|^2 \, \mathrm{d}S \\ &\leq C_5 \sum_{i \in I} \frac{1}{h_{K_i}} \int_{\partial K_i} |\eta|^2 \, \mathrm{d}S \\ &\leq C_5\, C_6 \sum_{i \in I} \frac{1}{h_{K_i}} \left(\|\eta\|_{L^2(K_i)} |\eta|_{H^1(K_i)} + h_{K_i}^{-1} \|\eta\|^2_{L^2(K_i)} \right) \\ &\leq C_5\, C_6\, C_8^2\, h^2 |u|^2_{H_2(\Omega)}. \end{aligned} \tag{2.3.93}$$

Finally, from (2.3.88) – (2.3.93) we conclude that

$$|\sigma_2| \leq \sigma(\varepsilon, h, u, \xi) := \varepsilon\, C\, h |u|_{H^2(\Omega)} \left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega, \mathcal{T}_h)} \right), \tag{2.3.94}$$

where $C > 0$ is a constant independent of u, ξ, h, ε.

In a similar way we obtain the estimate of σ_3:

$$|\sigma_3| \leq \sigma(\varepsilon, h, u, \xi). \tag{2.3.95}$$

Estimates (2.3.87), (2.3.94) and (2.3.95) immediately imply (2.3.85). □

In the convergence proof the following version of Gronwall's lemma will be applied:

Lemma 2.10 *Let $y,\, q,\, z,\, r \in C([0,T]),\ r \geq 0$ and let*

$$y(t) + q(t) \leq z(t) + \int_0^t r(s)\, y(s)\, \mathrm{d}S, \quad t \in [0,T]. \tag{2.3.96}$$

Then

$$\begin{aligned} & y(t) + q(t) + \int_0^t r(\vartheta)\, q(\vartheta) \exp\left(\int_\vartheta^t r(s)\, \mathrm{d}s \right) \mathrm{d}\vartheta \\ \leq\ & z(t) + \int_0^t r(\vartheta)\, z(\vartheta) \exp\left(\int_\vartheta^t r(s)\, \mathrm{d}s \right) \mathrm{d}\vartheta, \quad t \in [0,T]. \end{aligned} \tag{2.3.97}$$

Proof can be carried out in a similar way as in [Fei93], Par. 8.2.29. □

Now we will formulate the *main result.*

Theorem 2.1 *Let assumptions (2.2.4), a) – e), (H) and (A1) – (A3) from Section 2.3.2 be satisfied. Let u be the exact strong solution of problem (2.2.10) satisfying (2.2.12) and let u_h be the approximate solution defined by (2.3.15). Then the error $e_h = u_h - u$ satisfies the estimate*

$$\begin{aligned} & \sup_{t\in[0,T]} \|e_h(t)\|^2 + \varepsilon \int_0^T \left(|e_h(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(e_h(\vartheta), e_h(\vartheta)) \right) \mathrm{d}\vartheta \\ & + C_{14}(1+\varepsilon) \int_0^T \exp\left(C_{14} \frac{1+\varepsilon}{\varepsilon} (T - \vartheta) \right) \\ & \times \left\{ \int_0^\vartheta \left(|e_h(\alpha)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(e_h(\alpha), e_h(\alpha)) \right) \mathrm{d}\alpha \right\} \mathrm{d}\vartheta \\ \leq\ & 2h^2 \left\{ \left(\frac{c_1(\varepsilon,h)}{\varepsilon} + c_2 + \varepsilon C_{16} \right) \exp\left(C_{14} \frac{1+\varepsilon}{\varepsilon} T \right) + C_{15} \right\}, \end{aligned} \tag{2.3.98}$$

with constants C_{14}, C_{15}, C_{16}, $c_2 > 0$ independent of h, ε and with a function $c_1(\varepsilon, h)$, specified in the proof.

Proof. As in (2.3.73), we set $\xi = u_h - \Pi u \in S_h$, $\eta = \Pi u - u$. Then (2.3.74) holds: $e_h = u_h - u = \xi + \eta$. From (2.3.75), (2.3.76) and the relation $a_h(\xi,\xi) = \varepsilon |\xi|^2_{H^1(\Omega,\mathcal{T}_h)}$ we get

$$\begin{aligned} & \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|\xi\|^2_{L^2(\Omega)} + \varepsilon |\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + \varepsilon J_h^\sigma(\xi,\xi) \\ =\ & b(u,\xi) - \bar{b}_h(u_h,\xi) - \left(\frac{\partial \eta}{\partial t}, \xi \right) - a_h(\eta,\xi) - \varepsilon J_h^\sigma(\eta,\xi) \quad \text{in } (0,T). \end{aligned} \tag{2.3.99}$$

Now from (2.3.77) – (2.3.80), (2.3.85) and (2.3.93), using Young's inequality, we derive the estimates

$$\begin{aligned} & \frac{\mathrm{d}}{\mathrm{d}t} \|\xi\|^2_{L^2(\Omega)} + 2\varepsilon |\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + 2\varepsilon J_h^\sigma(\xi,\xi) \\ \leq\ & 2C_{12} \left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)} \right) \left(\|\xi\|_{L^2(\Omega)} + h^2 |u|_{H^2(\Omega)} + h|u|_{H^1(\Omega)} \right) \\ & + 2C_8 h \left| \frac{\partial u}{\partial t} \right|_{H^1(\Omega)} \|\xi\|_{L^2(\Omega)} + 2C_{13} \varepsilon\, h |u|_{H^2(\Omega)} \left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)} \right) \end{aligned} \tag{2.3.100}$$

$$
\begin{aligned}
&+2\varepsilon\sqrt{C_5C_6}C_8h|u|_{H^2(\Omega)}J_h^\sigma(\xi,\xi)^{1/2}\\
\le\;& 2C_{14}^*\left(J_h^\sigma(\xi,\xi)^{1/2}+|\xi|_{H^1(\Omega,\mathcal{T}_h)}\right)\left(\|\xi\|_{L^2(\Omega)}+h|u|_{H^1(\Omega)}+(\varepsilon h+h^2)|u|_{H^2(\Omega)}\right)\\
&+2C_8h\|\xi\|_{L^2(\Omega)}\left|\frac{\partial u}{\partial t}\right|_{H^1(\Omega)}\\
\le\;& \varepsilon J_h^\sigma(\xi,\xi)+\varepsilon|\xi|^2_{H^1(\Omega,\mathcal{T}_h)}+C_{14}\left(1+\frac{1}{\varepsilon}\right)\|\xi\|^2_{L^2(\Omega)}\\
&+\frac{C_{14}}{\varepsilon}\left(h^2|u|^2_{H^1(\Omega)}+(\varepsilon^2h^2+h^4)\,|u|^2_{H^2(\Omega)}\right)+C_8^2h^2\left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)}\quad \text{in } (0,T),
\end{aligned}
$$

where $C_{14}^*=\max\{C_{12},C_{13}+\sqrt{C_5C_6}C_8\}$, $C_{14}=\max\{1,8C_{14}^*\}$. Hence,

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\|\xi(t)\|^2_{L^2(\Omega)}+\varepsilon|\xi(t)|^2_{H^1(\Omega,\mathcal{T}_h)}+\varepsilon J_h^\sigma(\xi(t),\xi(t)) \qquad (2.3.101)\\
\le\;& C_{14}\left(1+\frac{1}{\varepsilon}\right)\|\xi(t)\|^2_{L^2(\Omega)}+\frac{C_{14}}{\varepsilon}\left(h^2|u(t)|^2_{H^1(\Omega)}+(\varepsilon^2h^2+h^4)\,|u(t)|^2_{H^2(\Omega)}\right)\\
&+C_8^2h^2\left|\frac{\partial u(t)}{\partial t}\right|^2_{H^1(\Omega)},\quad t\in(0,T).
\end{aligned}
$$

From assumption (2.2.12) it follows that the right-hand side of (2.3.101) is integrable over $(0,T)$. From (2.3.15), c), Remark 3.3 and estimate (2.3.44) it follows that

$$
\|\xi(0)\|^2_{L^2(\Omega)}\le 4C_8^2h^2|u^0|^2_{H^1(\Omega)}. \qquad (2.3.102)
$$

The integration of (2.3.101) from 0 to $t\in[0,T]$ yields

$$
\begin{aligned}
&\|\xi(t)\|^2_{L^2(\Omega)}+\varepsilon\int_0^t\left(|\xi(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)}+J_h^\sigma(\xi(\vartheta),\xi(\vartheta))\right)\mathrm{d}\vartheta \qquad (2.3.103)\\
\le\;& C_{14}\left(1+\frac{1}{\varepsilon}\right)\int_0^t\|\xi(\vartheta)\|^2_{L^2(\Omega)}\,\mathrm{d}\vartheta+\frac{C_{14}}{\varepsilon}h^2\int_0^t\left(|u(\vartheta)|^2_{H^1(\Omega)}+(\varepsilon^2+h^2)\,|u(\vartheta)|^2_{H^2(\Omega)}\right)\mathrm{d}\vartheta\\
&+C_8^2h^2\int_0^t\left|\frac{\partial u(\vartheta)}{\partial t}\right|^2_{H^1(\Omega)}\mathrm{d}\vartheta+4C_8^2h^2|u^0|^2_{H^1(\Omega)},\quad t\in[0,T].
\end{aligned}
$$

Now the application of Gronwall's Lemma 2.10, where we set

$$
\begin{aligned}
y(t) &= \|\xi(t)\|_{L^2(\Omega)}, \qquad (2.3.104)\\
q(t) &= \varepsilon\int_0^t\left(|\xi(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)}+J_h^\sigma(\xi(\vartheta),\xi(\vartheta))\right)\mathrm{d}\vartheta,\\
r(t) &= C_{14}\frac{1+\varepsilon}{\varepsilon},\\
z(t) &= \frac{c_1}{\varepsilon}h^2+c_2h^2,\\
c_1 &= c_1(\varepsilon,h)=C_{14}\left(\|u\|^2_{L^2(0,T;H^1(\Omega))}+(\varepsilon^2+h^2)\,\|u\|^2_{L^2(0,T;H^2(\Omega))}\right),\\
c_2 &= C_8^2\left(\|\partial u/\partial t\|^2_{L^2(0,T;H^1(\Omega))}+4|u^0|^2_{H^1(\Omega)}\right),
\end{aligned}
$$

and the relations

$$
\begin{aligned}
&\int_\vartheta^t r(s)\mathrm{d}s=C_{14}\frac{1+\varepsilon}{\varepsilon}(t-\vartheta), \qquad (2.3.105)\\
&\int_0^t r(\vartheta)\,z(\vartheta)\exp\left(\int_\vartheta^t r(s)\,\mathrm{d}s\right)\mathrm{d}\vartheta=\left(\frac{c_1}{\varepsilon}+c_2\right)h^2\left(\exp\left(C_{14}\frac{1+\varepsilon}{\varepsilon}t\right)-1\right)
\end{aligned}
$$

imply that

$$\begin{aligned}
&\|\xi(t)\|^2_{L^2(\Omega)} + \varepsilon \int_0^t \left(|\xi(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(\xi(\vartheta),\xi(\vartheta)) \right) \mathrm{d}\vartheta \qquad (2.3.106)\\
&+C_{14}(1+\varepsilon) \int_0^t \exp\left(C_{14}\frac{1+\varepsilon}{\varepsilon}(t-\vartheta) \right) \left\{ \int_0^\vartheta \left(|\xi(\alpha)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(\xi(\alpha),\xi(\alpha)) \right) \mathrm{d}\alpha \right\} \mathrm{d}\vartheta\\
\le\ & \left(\frac{c_1}{\varepsilon} + c_2 \right) h^2 \exp\left(C_{14} \frac{1+\varepsilon}{\varepsilon} t \right), \quad t \in [0,T].
\end{aligned}$$

Now, since $e_h = \xi + \eta$, we can write

$$\begin{aligned}
\|e_h\|^2_{L^2(\Omega)} &\le 2\left(\|\xi\|^2_{L^2(\Omega)} + \|\eta\|^2_{L^2} \right), \qquad (2.3.107)\\
|e_h|^2_{H^1(\Omega,\mathcal{T}_h)} &= 2\left(|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + |\eta|^2_{H^1(\Omega,\mathcal{T}_h)} \right),\\
J_h^\sigma(e_h,e_h) &\le 2\left(J_h^\sigma(\xi,\xi) + J_h^\sigma(\eta,\eta) \right).
\end{aligned}$$

Using assumption (2.2.12), from (2.3.93), (2.3.106), (2.3.107) and Lemma 2.7 we deduce that

$$\begin{aligned}
&\|e_h(t)\|^2_{L^2(\Omega)} + \varepsilon \int_0^t \left(|e_h(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(e_h(\vartheta),e_h(\vartheta)) \right) \mathrm{d}\vartheta \qquad (2.3.108)\\
&+C_{14}(1+\varepsilon) \int_0^t \exp\left(C_{14}\frac{1+\varepsilon}{\varepsilon}(t-\vartheta) \right) \left\{ \int_0^\vartheta \left(|e_h(\alpha)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(e_h(\alpha),e_h(\alpha)) \right) \mathrm{d}\alpha \right\} \mathrm{d}\vartheta\\
\le\ & 2h^2 \left\{ \left(\frac{c_1}{\varepsilon} + c_2 \right) \exp\left(C_{14}\frac{1+\varepsilon}{\varepsilon} t \right) + \varepsilon(1+C_5C_6)\, C_8^2 \int_0^t |u(\vartheta)|^2_{H^2(\Omega)} \mathrm{d}\vartheta \right.\\
&\left. +C_{14}(1+\varepsilon)\,(1+C_5C_6)\, C_8^2 \int_0^t \exp\left(C_{14}\frac{1+\varepsilon}{\varepsilon}(t-\vartheta) \right) \int_0^\vartheta |u(\alpha)|^2_{H^2(\Omega)} \mathrm{d}\alpha\, \mathrm{d}\vartheta \right\}\\
&+C_8^2 h^2 |u(t)|^2_{H^1(\Omega)}\\
\le\ & 2h^2 \left\{ \left(\frac{1}{\varepsilon} c_1(\varepsilon,h) + c_2 + \varepsilon C_{16} \right) \exp\left(C_{14}\frac{1+\varepsilon}{\varepsilon} T \right) + C_{15} \right\},
\end{aligned}$$

where

$$\begin{aligned}
C_{15} &= C_8^2 \|u\|^2_{L^\infty(0,T;H^1(\Omega))}, \qquad (2.3.109)\\
C_{16} &= (1+C_5C_6) C_8^2 |u|^2_{L^2(0,T;H^2(\Omega))}.
\end{aligned}$$

This already implies (2.3.98). □

Remark 2.2 *Estimate (2.3.98) is rather pessimistic in the case of a small parameter* ε*. However, if we use the weighted energy seminorm in the form*

$$\left\{ \int_0^T \exp\left(-C_{14}\frac{1+\varepsilon}{\varepsilon}\vartheta \right) \left(\int_0^\vartheta \left(|e_h(\alpha)|^2_{H^1(\Omega),\mathcal{T}_h} + J_h^\sigma(e_h(\alpha),e_h(\alpha)) \right) \mathrm{d}\alpha \right) \mathrm{d}\vartheta \right\}^{1/2}, \qquad (2.3.110)$$

we get the error estimate of order $O(h/\sqrt{\varepsilon})$*. The investigation of the behaviour of the error in dependence on the coefficient* ε *and obtaining uniform estimates with respect to* ε *is particularly important, but rather difficult. The uniform convergence with respect to* ε *has been obtained in a few works analyzing simple problems under rather special assumptions when complete*

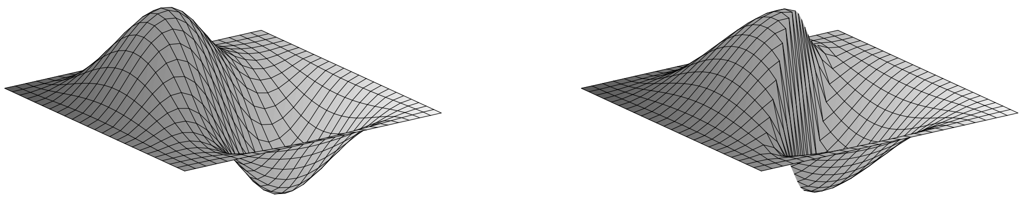


Figure 2.4: The exact solution (5.2) at $t = 1$ for $\nu = 0.5$ (left) and $\nu = 0.01$ (right).

analytic behaviour of solutions is known (cf., e. g. [RST96] and the citations therein, [MS99]). Another possibility is the application of the "hyperbolic machinery" (see, e. g. [Kuz76] and [Krö97] (and the citations therein)). In our work we are interested in a method for fixed positive values of diffusion. This is the case of, e. g., turbulent flow, where the viscosity and diffusion coefficients increase their magnitude by adding the so-called turbulent viscosity and turbulent heat conductivity. Similar situation can be met in financial mathematics ([Duf96]), where the diffusion dominates the convection.

2.3.5 Numerical examples

In this section we present some numerical experiments carried out for the 2D viscous Burgers equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x_1} + u\frac{\partial u}{\partial x_2} = \varepsilon \Delta u + g, \qquad \text{in } \Omega \times (0, T), \tag{2.3.111}$$

where $\Omega = (-1, 1)^2$, $T = 1$, equipped with initial condition (2.2.3) and boundary condition (2.2.2) where $u_D = 0$ on $\partial\Omega$. We set $\varepsilon = 0.002$ and define the function g and the initial condition in such a way that the exact solution has the form

$$u(x_1, x_2, t) = -(1 - x_1^2)^2(1 - x_2^2)^2 \text{arctg}\left(\frac{x_1 + x_2 + 2 - 2t}{\nu}\right), \tag{2.3.112}$$

where $\nu > 0$ is a given constant.

We solve problem (2.3.111) – (2.3.112) for the values of $\nu = 0.5$ and $\nu = 0.01$. The former gives a very smooth solution, whereas the latter leads to a solution with a very steep gradient. Figure 2.4 shows the solution (2.3.112) for both values of ν at time $t = 1$, when the solution has a steep gradient along the line $x_2 = -x_1$.

We solve the initial-boundary value problem by the presented numerical method. In the definition of the form $\bar{b}_h$ the following numerical flux is used

$$H(u_1, u_2, \boldsymbol{n}) = \begin{cases} \sum_{s=1}^{2} f_s(u_1) n_s, & \text{if } A > 0 \\ \sum_{s=1}^{2} f_s(u_2) n_s, & \text{if } A \leq 0 \end{cases}, \tag{2.3.113}$$

mesh	#triangles	h
mesh.125	136	0.351
mesh.500	586	0.214
mesh.2000	2374	0.0938
mesh.8000	9228	0.0452

Table 2.1: Triangulation used for computation.

where

$$A = \sum_{s=1}^{2} f_s'(\bar{u}) n_s, \quad \bar{u} = \frac{1}{2}(u_1 + u_2). \tag{2.3.114}$$

The resulting system of ODE's is solved by the forward Euler method with a very small time step $\tau = 10^{-4}$, which guarantees stability and sufficiently precise resolution with respect to time.

The computational error of the solution is evaluated at time t over a domain $\Omega^c \subset \Omega$ in $L^2(\Omega^c)-$ norm:

$$\|e_h(t)\|_{L^2(\Omega^c)} \equiv \|u_h(\cdot, t) - u(\cdot, t)\|_{L^2(\Omega^c)}. \tag{2.3.115}$$

The steep gradient of the solution for the case $\nu = 0.01$ behaves from the numerical point of view as a discontinuity. Therefore we evaluate the error both for the whole computational domain ($\Omega^c = \Omega$ in (2.3.115)) and only for a part of Ω where the solution is smooth, i .e., $\Omega^c = \Omega \setminus \Omega^{\text{disc}}$, where we put

$$\Omega^{\text{disc}} = \{(x_1, x_2);\ (x_1, x_2) \in \Omega, |x_1 + x_2| \le 0.1\}. \tag{2.3.116}$$

The subdomain Ω^{disc} is a thin region along the "discontinuity" at time $t = 1$. For the case $\nu = 0.5$, it is not necessary to exclude some part of Ω in the error evaluation as the solution does not contain any steep gradient. The integral (2.3.115) is calculated as the sum of integrals over all triangles. To compute the integral over a triangle we use a 7 points quadrature rule which is exact for polynomials of degree 5.

We investigate the error at the final time $t = 1$ and set $e_h = \|e_h(1)\|_{L^2}$. We suppose that the error e_h behaves according to the formula

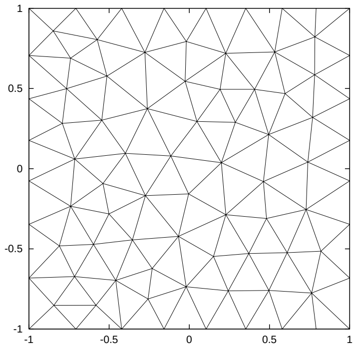
$$e_h \approx Ch^\alpha, \tag{2.3.117}$$

where $C > 0$ is a constant not depending on h and α is the *order of accuracy* of the method. The numerical solution was computed for 4 unstructured meshes with different h, see Figure 2.5 and Table 2.1, and the constants C and α from (2.3.117) were evaluated with the aid of least squares.

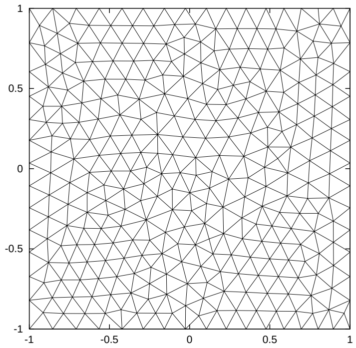
Table 2.2 shows the L^2- error and the constants C, α for the both cases $\nu = 0.5$ and $\nu = 0.01$. The behaviour of the error in dependence on h is plotted in Figure 2.6.

Figures 2.7 and 2.8 show the comparison of the exact solution with the numerical one along the diagonal cut of the domain $\Gamma^{\text{cut}} = \{(x_1, x_2); (x_1, x_2) = (-1, -1) + s(1, 1),\ s \in (0, 2)\}$ for $\nu = 0.5$ and $\nu = 0.01$, respectively, for time instants $t = 0.0$, $t = 0.5$, $t = 0.7$ and $t = 1.0$.

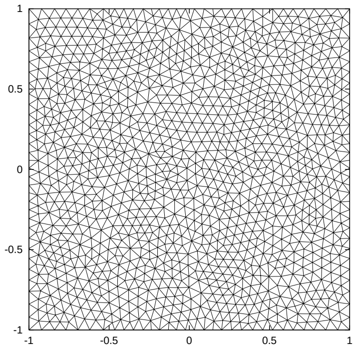
The numerical order of accuracy α is at least equal to 1, so that the numerical examples confirm our theoretical results that the numerical scheme is of order $O(h)$.



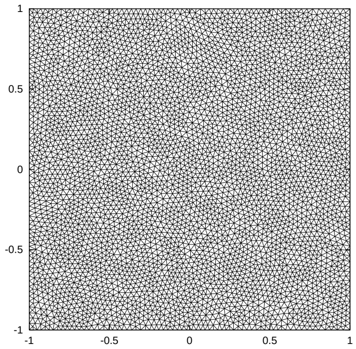
mesh.125



mesh.500



mesh.2000



mesh.8000

Figure 2.5: Triangulations used for the computation.

case	$\nu = 0.5$	$\nu = 0.01$	
e_h	$L^2(\Omega)$	$L^2(\Omega)$	$L^2(\Omega^c)$
mesh.125	0.2455	0.5531	0.3995
mesh.500	0.09450	0.2854	0.1132
mesh.2000	0.05036	0.1800	0.03362
mesh.8000	0.02820	0.1191	0.01575
order α	0.999	0.711	1.540
constant C	0.566	1.011	1.555

Table 2.2: Computational errors in L^2-norm, the order of accuracy α and the constant C.

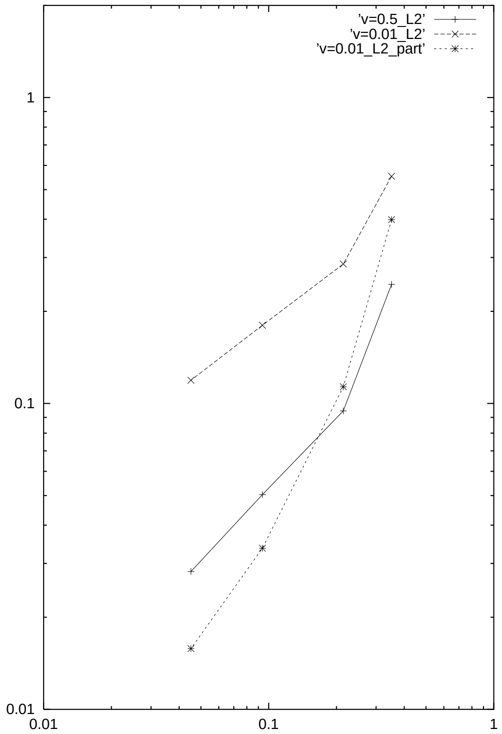


Figure 2.6: Dependence of e_h on h in L^2-norm: case $\nu = 0.5$ – full line, case $\nu = 0.01,\ \Omega^c = \Omega$ – dashed line and case $\nu = 0.01,\ \Omega^c = \Omega \setminus \Omega^{\mathrm{disc}}$ – dotted line.

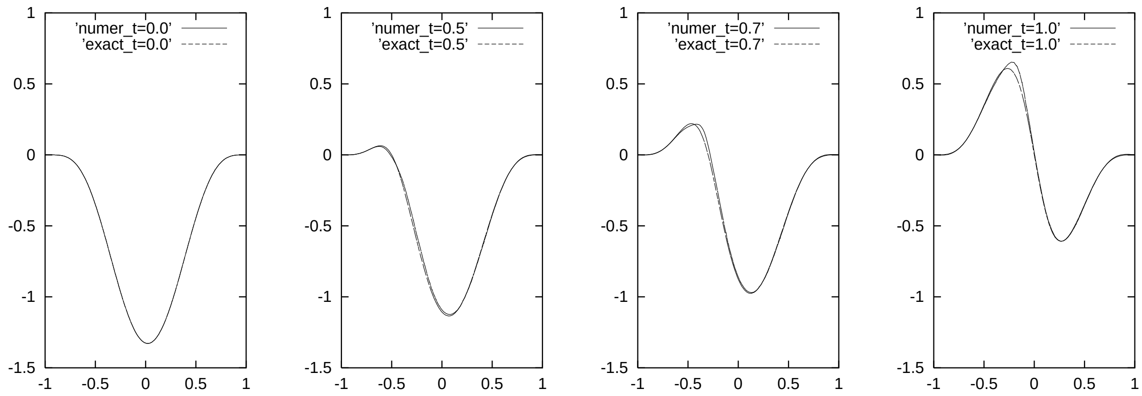


Figure 2.7: Comparison of the numerical solution (full line) with the exact one (dashed line) along the diagonal cut Γ^{cut}. Visualization for $t = 0.0$, $t = 0.5$, $t = 0.7$ and $t = 1.0$ in the case $\nu = 0.5$.

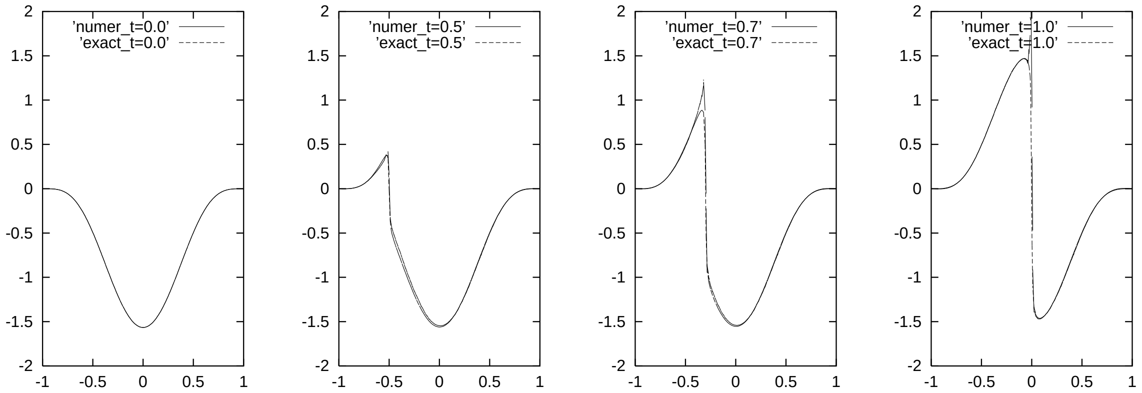


Figure 2.8: Comparison of the numerical solution (full line) with the exact one (dashed line) along the diagonal cut $\Gamma^{\rm cut}$. Visualization for $t = 0.0$, $t = 0.5$, $t = 0.7$ and $t = 1.0$ in the case $\nu = 0.01$.

2.4 Discontinuous Galerkin finite element method

Now we shall introduce the discrete problem approximating identity (2.2.24), b) with the aid of the DG FE method. We shall assume that the weak solution u is sufficiently regular, namely,

$$\frac{\partial u}{\partial t} \in L^2(0,T; H^{p+1}(\Omega)), \tag{2.4.1}$$

where an integer number $p \geq 1$ is a given degree of approximation. It is possible to show that such solution u satisfies equation (2.2.1) pointwise and $u \in C([0,T], H^{p+1}(\Omega))$.

2.4.1 Discretization

We define the space of discontinuous piecewise polynomial functions $S^{p,-1}(\Omega, \mathcal{T}_h)$ by (2.3.1) and we set

$$S_h = S^{p,-1}(\Omega, \mathcal{T}_h). \tag{2.4.2}$$

We remark the different meaning of S_h in (2.3.2) and (2.4.2). Further, the forms $a_h(u,v)$, $\ell_h(v)$ and $J_h^\sigma(u,v)$ are given by (2.3.11), (2.3.12) and (2.3.13), respectively. Moreover, we define the form

$$\tilde{b}_h(u,v) = \sum_{i\in I} \sum_{j \in S(i)} \int_{\Gamma_{ij}} \sum_{s=1}^d f_s(u(t))\, n_s\, v|_{\Gamma_{ij}} \,\mathrm{d}S - \sum_{i \in I} \int_{K_i} \sum_{s=1}^d f_s(u(t))\, \frac{\partial v}{\partial x_s} \,\mathrm{d}x, \tag{2.4.3}$$

which exhibits the convective terms.

Now, we derive the discrete problem in such a way that in (2.2.24), b), $u(t)$ is approximated by $u_h(t) \in S_h$, v is replaced by $v_h \in S_h$. Similarly as in the FV method the fluxes $\int_{\Gamma_{ij}} \sum_{s=1}^d f_s(u_h)\, n_s\, v \,\mathrm{d}S$ will be approximated with the aid of the so-called numerical flux $H(u, u', \boldsymbol{n})$ by the expression $\int_{\Gamma_{ij}} H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ji}}, \boldsymbol{n}_{ij})\, v \,\mathrm{d}S$ for $i \in I$, $j \in S(i)$. Of course, if $j \in \gamma(i)$ and $\Gamma_{ij} \subset \partial\Omega$, it is necessary to specify the meaning of $u|_{\Gamma_{ji}}$. Here we use the extrapolation, i.e. we set $u|_{\Gamma_{ji}} := u|_{\Gamma_{ij}}$.

In this way we obtain the approximation $b_h(u,v)$ of the convection form $\tilde{b}_h(u,v)$:

$$\begin{aligned} b_h(u,v) \;=\; & \sum_{i\in I}\left(\sum_{j\in s(i)}\int_{\Gamma_{ij}} H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ji}}, \boldsymbol{n}_{ij})\, v|_{\Gamma_{ij}}\,\mathrm{d}S\right. \\ +\; & \left.\sum_{j\in\gamma(i)}\int_{\Gamma_{ij}} H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ij}}, \boldsymbol{n}_{ij})\, v|_{\Gamma_{ij}}\,\mathrm{d}S\right) \\ -\; & \sum_{i\in I}\int_{K_i}\sum_{s=1}^{d} f_s(u)\,\frac{\partial v}{\partial x_s}\,\mathrm{d}x, \quad u,\, v\in H^1(\Omega,\mathcal{T}_h). \end{aligned} \tag{2.4.4}$$

We shall assume that the numerical flux has (H) properties (2.3.7) – (2.3.9). In virtue of (2.3.7) and (2.3.8), the functions f_s, $s=1,\dots,d$ are Lipschitz continuous with constant $L_f = 2C_1$.

From (2.2.4), a) and (2.3.8) we see that

$$H(0,0,\boldsymbol{n}) = 0 \quad \forall\, \boldsymbol{n}\in\mathcal{S}_1, \tag{2.4.5}$$

and (2.4.3), (2.4.4) and (2.3.8) yield

$$b_h(u,v) = \tilde{b}_h(u,v) \quad \forall u\in H^2(\Omega)\ \forall v\in H^2(\Omega,\mathcal{T}_h). \tag{2.4.6}$$

Since $S_h\subset H^2(\Omega,\mathcal{T}_h)$, the forms (2.3.11) – (2.3.13) and (2.4.4) have sense for $u := u_h$, $v := v_h \in S_h$. We define the *semidiscrete approximate solution* as a function $u_h : Q_T \to I\!R$ satisfying the conditions

$$\begin{aligned} &\text{a)} && u_h\in C^1([0,T];S_h), \\ &\text{b)} && \left(\frac{\partial u_h(t)}{\partial t}, v_h\right) + a_h(u_h(t),v_h) + b_h(u_h(t),v_h) \\ &&& +\varepsilon J_h^\sigma(u_h(t),v_h) = \ell_h(v_h)\,(t) \quad \forall\, v_h\in S_h,\ t\in[0,T], \\ &\text{c)} && (u_h(0),v_h) = (u^0,v_h) \quad \forall\, v_h\in S_h. \end{aligned} \tag{2.4.7}$$

The discrete problem (2.4.7), a) – c) is equivalent to a system of ordinary differential equations (ODE). (This approach to the numerical solution of initial-boundary value problems via the space semidiscretization is called the method of lines.) If we apply some ODE solver to problem (2.4.7), a) – c), we obtain the fully discrete problem. In what follows we shall be concerned with the analysis of the semidiscrete problem (2.4.7), a) – c).

Taking into account that for the exact solution u satisfying (2.4.1) we have (2.2.24), b), $[u]|_{\Gamma_{ij}} = 0$ and $u|_{\partial\Omega\times(0,T)} = u_D$, we can see that for $v_h\in S_h, t\in(0,T)$, u satisfies the identity

$$\left(\frac{\partial u}{\partial t}(t), v_h\right) + a_h(u(t),v_h) + \tilde{b}_h(u(t),v_h) + \varepsilon J_h^\sigma(u(t),v_h) = \ell_h(v_h)\,(t). \tag{2.4.8}$$

2.4.2 Some auxiliary results

Lemma 2.11 *Let assumption (A1) be valid. Then for each $K\in\mathcal{T}_h$, $h\in(0,h_0)$ there exist axiparallel boxes B_0, B_1 such that*

i) $B_0\subset K\subset B_1$,

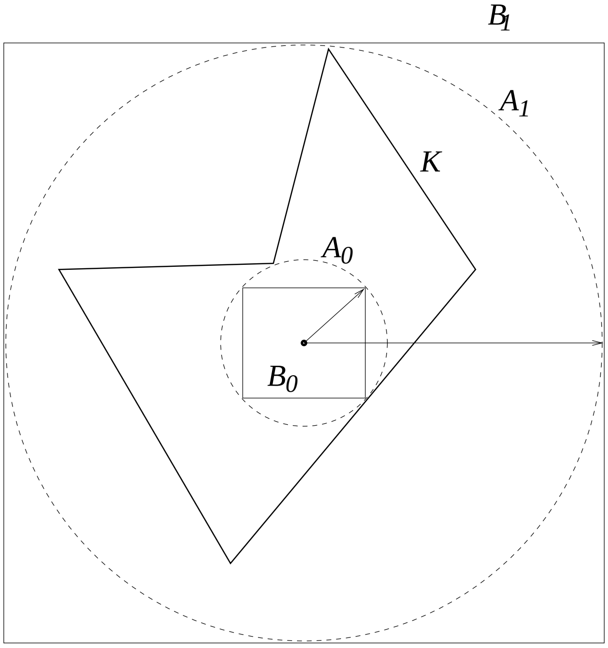


Figure 2.9: The balls A_0, A_1 and the axiparallel boxes B_0, B_1 used in the proof of Lemma 2.11.

ii) centres of B_0 and B_1 are identical,

iii) the ratio of sizes of B_0 and B_1 satisfies the relation

$$0 < \frac{1}{\bar{\kappa}} \leq \frac{h_{B_1}}{h_{B_0}} \leq \bar{\kappa} < \infty, \tag{2.4.9}$$

where $\bar{\kappa}$ is a constant independent of K and h and the numbers h_{B_0} and h_{B_1} are the half of sizes of B_0 and B_1, respectively, i.e. $B_j = (-h_{B_j}, h_{B_j})^d$, $j = 0, 1$.

Proof. Condition (2.3.17) implies that there exists a d-dimensional ball A_0 with the centre x_K inscribed into K and there exists a d-dimensional ball A_1 with the centre x_K containing K. The ratio of diameters of A_0 and A_1 satisfies

$$\frac{r_{A_1}}{r_{A_0}} = \frac{\max_{x\in\partial K} |x - x_K|}{\min_{x\in\partial K} |x - x_K|} \leq \kappa. \tag{2.4.10}$$

Then the existence of the axiparallel boxes B_0, B_1 satisfying conditions i) – iii) follows from Figure 2.9 and (2.4.10) with the constant $\bar{\kappa} = \sqrt{d}\kappa$. $\square$

Lemma 2.12 *(Inverse inequality) There exists a constant $C_{17} > 0$ independent of h, K such that*

$$|v|_{H^1(K)} \leq C_{17} h_K^{-1} \|v\|_{L^2(K)} \quad \forall v \in P_p(K)\ \forall K \in \mathcal{T}_h. \tag{2.4.11}$$

Proof. Let $K \in \mathcal{T}_h$ and let B_0 and B_1 be the boxes from Lemma 2.11 satisfying (2.4.9). Without loss of generality we suppose that the centre of the boxes B_0 and B_1 is the origin of the coordinate system. Then obviously

$$(-1,1)^d \subset h_{B_0}^{-1} K \subset h_{B_0}^{-1} B_1, \tag{2.4.12}$$

where

$$h_{B_0}^{-1} K = \left\{ y = (y_1, \dots, y_d) \in I\!R^d,\ y_l = h_{B_0}^{-1} x_l,\ l = 1, \dots, d,\ x = (x_1, \dots, x_d) \in K \right\}. \tag{2.4.13}$$

We define the reference boxes

$$\hat{B}_0 = (-1,1)^d, \quad \hat{B}_1 = (-L, L)^d, \quad L := \bar{\kappa} \geq 1, \tag{2.4.14}$$

where $\bar{\kappa}$ is the constant from (2.4.9). The equivalence of norms on the finite dimensional space $P_p(\hat{B}_1)$ implies that there exists a constant $C(p, \bar{\kappa})$ such that

$$\|v\|_{L^2(\hat{B}_1)} \leq C(p, \bar{\kappa}) \|v\|_{L^2(\hat{B}_0)} \quad \forall v \in P_p(\hat{B}_1). \tag{2.4.15}$$

The transformation to the reference element and the equivalence of norms $\|\cdot\|_{L^2(\hat{B}_0)}$ and $\|\cdot\|_{H^1(\hat{B}_0)}$ on the finite dimensional space $P_p(\hat{B}_0)$ implies

$$\begin{aligned}
\|\nabla v\|_{L^2(B_0)} &\leq c_1 h_{B_0}^{d/2-1} \|\hat{\nabla} \hat{v}\|_{L^2(\hat{B}_0)} \\
&\leq c_1 h_{B_0}^{d/2-1} \|\hat{v}\|_{H^1(\hat{B}_0)} \\
&\leq c_1 c_2(p) h_{B_0}^{d/2-1} \|\hat{v}\|_{L^2(\hat{B}_0)} \\
&\leq \tilde{C}(p) h_{B_0}^{-1} \|v\|_{L^2(B_0)} \quad \forall v \in P_p(B_0),
\end{aligned} \tag{2.4.16}$$

where $\tilde{C}(p)$ is a constant independent of h.

Let $v \in P_p(B_1))$. Then the transformation to the reference element $\hat{B}_0$ and relations (2.4.12) – (2.4.16) yields

$$\begin{aligned}
\|\nabla v\|_{L^2(K)} &\leq \|\nabla v\|_{L^2(B_1)} \\
&= h_{B_1}^{d/2-1} \|\hat{\nabla} \hat{v}\|_{L^2(\hat{B}_1)} \\
&\leq h_{B_1}^{d/2-1} C(p, \bar{\kappa}) \|\hat{\nabla} \hat{v}\|_{L^2(\hat{B}_0)} \\
&\leq h_{B_1}^{d/2-1} h_{B_0}^{1-d/2} C(p, \bar{\kappa}) \|\nabla v\|_{L^2(B_0)} \\
&= \left(\frac{h_{B_1}}{h_{B_0}} \right)^{d/2-1} C(p, \bar{\kappa}) \|\nabla v\|_{L^2(B_0)} \\
&\leq \left(\frac{h_{B_1}}{h_{B_0}} \right)^{d/2-1} C(p, \bar{\kappa}) \tilde{C}(p) h_{B_0}^{-1} \|v\|_{L^2(B_0)} \\
&\leq \left(\frac{h_{B_1}}{h_{B_0}} \right)^{d/2-1} C(p, \bar{\kappa}) \tilde{C}(p) h_{B_0}^{-1} \|v\|_{L^2(K)} \\
&\leq (\bar{\kappa})^{d/2-1} \bar{\kappa} C(p, \bar{\kappa}) \tilde{C}(p) h_{B_1}^{-1} \|v\|_{L^2(K)} \\
&\leq (\bar{\kappa})^{d/2} C(p, \bar{\kappa}) \tilde{C}(p) h_K^{-1} \|v\|_{L^2(K)} =: C_{17} h_K^{-1} \|v\|_{L^2(K)},
\end{aligned} \tag{2.4.17}$$

which proves the lemma.

□

Lemma 2.13 *There exists a constant* $C_{18} > 0$ *independent of* $h \in (0, h_0)$ *and a mapping* $\Pi : H^1(K) \to P_p(K),\ p \geq 0$ *such that*

$$\begin{aligned} \|\Pi v - v\|_{L^2(K)} &\leq C_{18} h_K^{p+1} |v|_{H^{p+1}(K)},\ v \in H^{p+1}(K), K \in \mathcal{T}_h,\ h \in (0, h_0), \\ |\Pi v - v|_{H^1(K)} &\leq C_{18} h_K^{p} |v|_{H^{p+1}(K)},\ v \in H^{p+1}(K), K \in \mathcal{T}_h,\ h \in (0, h_0). \end{aligned} \tag{2.4.18}$$

Proof. See [Ver99]. □

2.4.3 Properties of the form b_h

Let assumptions (2.2.4), a), Assumptions (H) in Section 2.3.1 and Assumption (A1) in Section 2.3.2 be satisfied. We shall be concerned with the Lipschitz-continuity of the form b_h.

Lemma 2.14 *There exists a constant* $C_{19} > 0$ *such that*

$$|b_h(u, v_h) - b_h(\bar{u}, v_h)| \leq C_{19} \left(J_h^\sigma(v_h, v_h)^{1/2} + |v_h|_{H^1(\Omega, \mathcal{T}_h)} \right) \|u - \bar{u}\|_{L^2(\Omega)},$$
$$u,\ \bar{u},\ v_h \in S_h,\ h \in (0, h_0). \tag{2.4.19}$$

Proof. By (2.4.4), for $u,\ \bar{u} \in H^1(\Omega, \mathcal{T}_h)$ and $v_h \in S_h$,

$$\begin{aligned} b_h(u, v_h) - b_h(\bar{u}, v_h) = & \underbrace{- \sum_{i \in I} \int_{K_i} \sum_{s=1}^{d} (f_s(u) - f_s(\bar{u})) \frac{\partial v_h}{\partial x_s} \, \mathrm{d}x}_{:=\sigma_1} \\ + & \underbrace{\sum_{i \in I} \sum_{j \in s(i)} \int_{\Gamma_{ij}} \left(H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ji}}, \boldsymbol{n}_{ij}) - H(\bar{u}|_{\Gamma_{ij}}, \bar{u}|_{\Gamma_{ji}}, \boldsymbol{n}_{ij}) \right) v_h|_{\Gamma_{ij}} \, \mathrm{d}S}_{:=\sigma_2} \\ + & \underbrace{\sum_{i \in I} \sum_{j \in \gamma(i)} \int_{\Gamma_{ij}} \left(H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ij}}, \boldsymbol{n}_{ij}) - H(\bar{u}|_{\Gamma_{ij}}, \bar{u}|_{\Gamma_{ij}}, \boldsymbol{n}_{ij}) \right) v_h|_{\Gamma_{ij}} \, \mathrm{d}S}_{:=\sigma_3} . \end{aligned} \tag{2.4.20}$$

From the Lipschitz-continuity of the functions $f_s,\ s = 1, \ldots, d$ we have

$$|\sigma_1| \leq L_f \sum_{i \in I} \int_{K_i} \sum_{s=1}^{d} |u - \bar{u}| \left| \frac{\partial v_h}{\partial x_s} \right| \mathrm{d}x \leq \sqrt{d} L_f \|u - \bar{u}\|_{L^2(\Omega)} |v_h|_{H^1(\Omega, \mathcal{T}_h)}. \tag{2.4.21}$$

Using the conservativity (2.3.9) of H we find that

$$\begin{aligned} \sigma_2 + \sigma_3 &= \sum_{i \in I} \left(\sum_{j \in s(i)} \int_{\Gamma_{ij}} v_h|_{\Gamma_{ij}} \left\{ H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ji}}, \boldsymbol{n}_{ij}) - H(\bar{u}|_{\Gamma_{ij}}, \bar{u}|_{\Gamma_{ji}}, \boldsymbol{n}_{ij}) \right\} \mathrm{d}S \right. \\ & \qquad \left. + \sum_{j \in \gamma(i)} \int_{\Gamma_{ij}} v_h|_{\Gamma_{ij}} \left\{ H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ij}}, \boldsymbol{n}_{ij}) - H(\bar{u}|_{\Gamma_{ij}}, \bar{u}|_{\Gamma_{ij}}, \boldsymbol{n}_{ij}) \right\} \mathrm{d}S \right) \\ &= \sum_{i \in I} \left(\frac{1}{2} \sum_{j \in s(i)} \int_{\Gamma_{ij}} \left(v_h|_{\Gamma_{ij}} - v_h|_{\Gamma_{ji}} \right) \left\{ H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ji}}, \boldsymbol{n}_{ij}) - H(\bar{u}|_{\Gamma_{ij}}, \bar{u}|_{\Gamma_{ji}}, \boldsymbol{n}_{ij}) \right\} \mathrm{d}S \right. \\ & \qquad \left. + \sum_{j \in \gamma(i)} \int_{\Gamma_{ij}} v_h|_{\Gamma_{ij}} \left\{ H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ij}}, \boldsymbol{n}_{ij}) - H(\bar{u}|_{\Gamma_{ij}}, \bar{u}|_{\Gamma_{ij}}, \boldsymbol{n}_{ij}) \right\} \mathrm{d}S \right) . \end{aligned}$$

Now, the Lipschitz-continuity (2.3.7) of H, the Cauchy inequality and (2.2.17) imply that

$$
\begin{aligned}
|\sigma_2+\sigma_3| &\le C_1 \sum_{i\in I}\left\{ \frac{1}{2}\sum_{j\in s(i)} \int_{\Gamma_{ij}} \left|[v_h]_{\Gamma_{ij}}\right| \left\{ |u-\bar u|_{\Gamma_{ij}} + |u-\bar u|_{\Gamma_{ji}} \right\} \mathrm{d}S \right. \\
&\qquad \left. + 2\sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} |v_h|_{\Gamma_{ij}}|\,|u-\bar u|_{\Gamma_{ij}} \mathrm{d}S \right\} \qquad (2.4.22) \\
&\le C_1 \sum_{i\in I}\left\{ \sum_{j\in s(i)} \int_{\Gamma_{ij}} \left|[v_h]_{\Gamma_{ij}}\right| |u-\bar u|_{\Gamma_{ij}} \mathrm{d}S + 2\sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} |v_h|_{\Gamma_{ij}}|\,|u-\bar u|_{\Gamma_{ij}} \mathrm{d}S \right\} \\
&\le 2C_1 \sum_{i\in I}\left\{ \left(\sum_{j\in s(i)} \int_{\Gamma_{ij}} \frac{[v_h]^2}{d(\Gamma_{ij})} \mathrm{d}S + \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} \frac{v_h^2}{d(\Gamma_{ij})} \mathrm{d}S \right)^{1/2} \right. \\
&\qquad \left. \times \left(\sum_{j\in s(i)} d(\Gamma_{ij}) \int_{\Gamma_{ij}} (u-\bar u)^2 \mathrm{d}S + \sum_{j\in\gamma(i)} d(\Gamma_{ij}) \int_{\Gamma_{ij}} (u-\bar u)^2 \mathrm{d}S \right)^{1/2} \right\} \\
&\le 2C_1\, J_h^\sigma(v_h,v_h)^{1/2} \left(\sum_{i\in I} h_{K_i} \|u-\bar u\|^2_{L^2(\partial K_i)} \right)^{1/2}.
\end{aligned}
$$

Using the multiplicative trace inequality (2.3.32), the discrete Cauchy inequality and the inverse inequality (2.4.11), for $\varphi \in S_h$ we obtain

$$
\begin{aligned}
\sum_{i\in I} h_{K_i}\|\varphi\|^2_{L^2(\partial K_i)} &\le C_6 \sum_{i\in I}\left(\|\varphi\|^2_{L^2(K_i)} + h_{K_i}\|\varphi\|_{L^2(K_i)}\,|\varphi|_{H^1(K_i)} \right) \qquad (2.4.23) \\
&\le C_6 \left(\|\varphi\|^2_{L^2(\Omega)} + \|\varphi\|_{L^2(\Omega)} \left(\sum_{i\in I} h^2_{K_i} |\varphi|^2_{H^1(K_i)} \right)^{1/2} \right) \\
&\le C_6 \left(\|\varphi\|^2_{L^2(\Omega)} + \|\varphi\|_{L^2(\Omega)} C_{17} \left(\sum_{i\in I} \|\varphi\|^2_{L^2(K_i)} \right)^{1/2} \right) \\
&= C_6(1+C_{17})\|\varphi\|^2_{L^2(\Omega)}.
\end{aligned}
$$

Now, the relations (2.4.22) and (2.4.23), where we put $\varphi := u-\bar u$, give

$$
|\sigma_2+\sigma_3| \le 2C_1(C_6(1+C_{17}))^{1/2}\, J_h^\sigma(v_h,v_h)^{1/2}\|u-\bar u\|_{L^2(\Omega)}. \qquad (2.4.24)
$$

Finally, from (2.4.20), (2.4.21) and (2.4.24) we get the final result

$$
|b_h(u,v_h)-b_h(\bar u,v_h)| \le C_{19}\left(J_h^\sigma(v_h,v_h)^{1/2} + |v_h|_{H^1(\Omega,\mathcal{T}_h)} \right) \|u-\bar u\|_{L^2(\Omega)} \qquad (2.4.25)
$$

with $C_{19} = \max(\sqrt{d}L_f, 2C_1(C_6(1+C_{17}))^{1/2})$. $\square$

Lemma 2.15 *There exists a constant $C_{20}>0$ such that*

$$
\begin{aligned}
&|b_h(u,v_h)-b_h(\Pi\, u,v_h)| \le C_{20}h^{p+1}\left(J_h^\sigma(v_h,v_h)^{1/2} + |v_h|_{H^1(\Omega,\mathcal{T}_h)} \right) |u|_{H^{p+1}(\Omega)}, \\
&\qquad\qquad u\in H^{p+1}(\Omega),\ v_h\in S_h,\ h\in(0,h_0), \qquad (2.4.26)
\end{aligned}
$$

where Πu is the S_h–interpolant of u from Lemma 2.13.

Proof. We proceed analogously as in the proof of Lemma 2.14, writing Πu instead of $\bar{u}$. By (2.4.4), for $u \in H^{p+1}(\Omega)$ and $v_h \in S_h$,

$$b_h(u, v_h) - b_h(\Pi u, v_h) = \underbrace{-\sum_{i\in I}\int_{K_i}\sum_{s=1}^{d}(f_s(u) - f_s(\Pi u))\,\frac{\partial v_h}{\partial x_s}\,\mathrm{d}x}_{:=\sigma_1} \tag{2.4.27}$$
$$+ \underbrace{\sum_{i\in I}\sum_{j\in s(i)}\int_{\Gamma_{ij}}\left(H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ji}}, \boldsymbol{n}_{ij}) - H(\Pi u|_{\Gamma_{ij}}, \Pi u|_{\Gamma_{ji}}, \boldsymbol{n}_{ij})\right) v_h|_{\Gamma_{ij}}\,\mathrm{d}S}_{:=\sigma_2}$$
$$+ \underbrace{\sum_{i\in I}\sum_{j\in \gamma(i)}\int_{\Gamma_{ij}}\left(H(u|_{\Gamma_{ij}}, u|_{\Gamma_{ij}}, \boldsymbol{n}_{ij}) - H(\Pi u|_{\Gamma_{ij}}, \Pi u|_{\Gamma_{ij}}, \boldsymbol{n}_{ij})\right) v_h|_{\Gamma_{ij}}\,\mathrm{d}S}_{:=\sigma_3}.$$

Similarly as in (2.4.21) and from (2.4.18) we have

$$|\sigma_1| \le \sqrt{d}L_f\|u - \Pi u\|_{L^2(\Omega)}|v_h|_{H^1(\Omega,\mathcal{T}_h)} \le \sqrt{d}L_f C_{18} h^{p+1}|u|_{H^{p+1}(\Omega)}|v_h|_{H^1(\Omega,\mathcal{T}_h)}. \tag{2.4.28}$$

Analogously as in (2.4.22) we have

$$|\sigma_2 + \sigma_3| \le 2C_1\, J_h^\sigma(v_h, v_h)^{1/2}\left(\sum_{i\in I} h_{K_i}\|u - \Pi u\|^2_{L^2(\partial K_i)}\right)^{1/2}. \tag{2.4.29}$$

Using the multiplicative trace inequality (2.3.32), the discrete Cauchy inequality and (2.4.18), we obtain

$$\begin{aligned}\sum_{i\in I} h_{K_i}\|u - \Pi u\|^2_{L^2(\partial K_i)} &\le \sum_{i\in I} C_6\left(\|u - \Pi u\|^2_{L^2(K_i)} + h_{K_i}\|u - \Pi u\|_{L^2(K_i)}|u - \Pi u|_{H^1(K_i)}\right)\\ &\le \sum_{i\in I} C_6 C_{18}^2\left(h_{K_i}^{2p+2} + h_{K_i}h_{K_i}^{p+1}h_{K_i}^{p}\right)|u|^2_{H^{p+1}(K_i)}\\ &\le 2C_6C_{18}^2 h^{2p+2}|u|^2_{H^{p+1}(\Omega)}.\end{aligned} \tag{2.4.30}$$

Now, the relations (2.4.29) and (2.4.30) give

$$|\sigma_2 + \sigma_3| \le 2\sqrt{2C_6}C_1C_{18}\, J_h^\sigma(v_h, v_h)^{1/2}h^{p+1}|u|_{H^{p+1}(\Omega)}. \tag{2.4.31}$$

Finally, from (2.4.27), (2.4.28) and (2.4.31) we get the final result

$$|b_h(u, v_h) - b_h(\Pi u, v_h)| \le C_{20}h^{p+1}\left(J_h^\sigma(v_h, v_h)^{1/2} + |v_h|_{H^1(\Omega,\mathcal{T}_h)}\right)|u|_{H^{p+1}(\Omega)} \tag{2.4.32}$$

with $C_{20} = \max(\sqrt{d}L_f, 2\sqrt{2C_6}C_1C_{18})$. □

2.4.4 Error estimates for the method of lines

Our goal is now to analyze the error estimates of the approximate solution u_h obtained by the method of lines (2.4.7) under the assumption that the exact solution u satisfies (2.4.1).

Let $\Pi u(t)$ be the S_h-interpolation of $u(t)$ ($t \in [0, T]$) from Lemma 2.13. We set

$$\xi = u_h - \Pi u \in S_h, \quad \eta = \Pi u - u \in H^{p+1}(\Omega, \mathcal{T}_h). \tag{2.4.33}$$

Then

$$u_h - u = \xi + \eta. \tag{2.4.34}$$

From (2.4.6) – (2.4.8), where we substitute $v_h := \xi$, it follows that

$$\begin{aligned}&\left(\frac{\partial \xi}{\partial t}, \xi\right) + a_h(\xi, \xi) + \varepsilon J_h^\sigma(\xi, \xi) \\ = \quad & b_h(u, \xi) - b_h(u_h, \xi) - \left(\frac{\partial \eta}{\partial t}, \xi\right) - a_h(\eta, \xi) - \varepsilon J_h^\sigma(\eta, \xi).\end{aligned} \tag{2.4.35}$$

(Of course, $\xi = \xi(t)$, $\eta = \eta(t)$ for $t \in [0, T]$, but we do not emphasize the dependence on t by our notation, if it is not necessary.) In what follows, we estimate the individual terms on the right-hand side of (2.4.35).

In virtue of (2.4.1),

$$\left(\frac{\partial \xi}{\partial t}, \xi\right) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\Omega)}. \tag{2.4.36}$$

The Cauchy inequality implies that

$$\left|\left(\frac{\partial \eta}{\partial t}, \xi\right)\right| \leq \left\|\frac{\partial \eta}{\partial t}\right\|_{L^2(\Omega)} \|\xi\|_{L^2(\Omega)}, \tag{2.4.37}$$

$$J_h^\sigma(\eta, \xi) \leq \left(J_h^\sigma(\eta, \eta)\right)^{1/2} \left(J_h^\sigma(\xi, \xi)\right)^{1/2}. \tag{2.4.38}$$

Lemma 2.16 *We have*

$$\begin{aligned}\|\eta\|_{L^2(\Omega)} &\leq C_{18} h^{p+1} |u|_{H^{p+1}(\Omega)}, \\ |\eta|_{H^1(\Omega, \mathcal{T}_h)} &\leq C_{18} h^{p} |u|_{H^{p+1}(\Omega)}, \\ |\eta|_{H^2(\Omega, \mathcal{T}_h)} &\leq C_{18} h^{p-1} |u|_{H^{p+1}(\Omega)}, \\ \left\|\frac{\partial \eta}{\partial t}\right\|_{L^2(\Omega)} &\leq C_{18} h^{p+1} \left|\frac{\partial u}{\partial t}\right|_{H^{p+1}(\Omega)}, \quad h \in (0, h_0).\end{aligned} \tag{2.4.39}$$

Proof is a consequence of (2.4.1), (2.4.18) and the relation $\Pi(\partial u/\partial t) = \partial(\Pi u)/\partial t$. □

Lemma 2.17 *We have*

$$|b_h(u, \xi) - b_h(u_h, \xi)| \leq C_{22}\left(J_h^\sigma(\xi, \xi)^{1/2} + |\xi|_{H^1(\Omega, \mathcal{T}_h)}\right)\left(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\xi\|_{L^2(\Omega)}\right), \quad h \in (0, h_0), \tag{2.4.40}$$

where $C_{22} > 0$ is a constant independent of h and ξ.

Proof. We can write

$$b_h(u, \xi) - b_h(u_h, \xi) = (b_h(u, \xi) - b_h(\Pi u, \xi)) + (b_h(\Pi u, \xi) - b_h(u_h, \xi)). \tag{2.4.41}$$

From Lemma 2.14 and Lemma 2.15 we find that

$$\begin{aligned}|b_h(u, \xi) - b_h(\Pi u, \xi)| &\leq C_{20} h^{p+1}\left(J_h^\sigma(\xi, \xi)^{1/2} + |\xi|_{H^1(\Omega, \mathcal{T}_h)}\right)|u|_{H^{p+1}(\Omega)}, \\ |b_h(\Pi u, \xi) - b_h(u_h, \xi)| &\leq C_{19}\left(J_h^\sigma(\xi, \xi)^{1/2} + |\xi|_{H^1(\Omega, \mathcal{T}_h)}\right)\|\Pi u - u_h\|_{L^2(\Omega)}.\end{aligned} \tag{2.4.42}$$

By (2.4.33), (2.4.41) and (2.4.42),

$$|b_h(u,\xi) - b_h(u_h,\xi)| \le C_{22}\left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right)\left(h^{p+1}|u|_{H^{p+1}(\Omega)} + \|\xi\|_{L^2(\Omega)}\right),$$

which is (2.4.40) with $C_{22} = \max(C_{19}, C_{20})$. □

Lemma 2.18 *There exists a constant $C_{23} > 0$ independent of u, h, ξ, ε such that*

$$|a_h(\eta,\xi)| \le C_{23}\varepsilon h^p |u|_{H^{p+1}(\Omega)}\left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right), \quad h \in (0,h_0). \tag{2.4.43}$$

Proof. In view of (2.3.11), we can write

$$\begin{aligned} a_h(\eta,\xi) &= \sigma_1 + \sigma_2 + \sigma_3, \qquad (2.4.44)\\ \sigma_1 &:= \varepsilon \sum_{i\in I}\int_{K_i} \nabla\eta\cdot\nabla\xi \,\mathrm{d}x,\\ \sigma_2 &:= -\varepsilon\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\Big(\langle\nabla\eta\rangle\cdot\boldsymbol{n}_{ij}[\xi] - \langle\nabla\xi\rangle\cdot\boldsymbol{n}_{ij}[\eta]\Big)\,\mathrm{d}S\\ \sigma_3 &:= -\varepsilon\sum_{i\in I}\sum_{j\in\gamma(i)}\int_{\Gamma_{ij}}\Big((\nabla\eta\cdot\boldsymbol{n}_{ij})\,\xi - (\nabla\xi\cdot\boldsymbol{n}_{ij})\,\eta\Big)\,\mathrm{d}S. \end{aligned}$$

We estimate the individual terms in (2.4.44) similarly as in [HSS00].

The Cauchy inequality and estimates (2.4.39) imply that

$$|\sigma_1| \le \varepsilon C_{18} h^p |u|_{H^{p+1}(\Omega)}|\xi|_{H^1(\Omega,\mathcal{T}_h)}. \tag{2.4.45}$$

In order to estimate σ_2, we set $\gamma_{ij} := d(\Gamma_{ij})$. The Cauchy inequality, (2.3.13), (2.3.14) and (2.2.17) yield the estimates

$$\begin{aligned} |\sigma_2| &\le \varepsilon\left(\sum_{i\in I}\sum_{j\in s(i)}\gamma_{ij}\|\nabla\eta\|^2_{L^2(\Gamma_{ij})}\right)^{1/2}\left(\sum_{i\in I}\sum_{j\in s(i)}\frac{1}{\gamma_{ij}}\|[\xi]\|^2_{L^2(\Gamma_{ij})}\right)^{1/2} \qquad (2.4.46)\\ &\quad +\varepsilon\left(\sum_{i\in I}\sum_{j\in s(i)}\gamma_{ij}\|\nabla\xi\|^2_{L^2(\Gamma_{ij})}\right)^{1/2}\left(\sum_{i\in I}\sum_{j\in s(i)}\frac{1}{\gamma_{ij}}\|[\eta]\|^2_{L^2(\Gamma_{ij})}\right)^{1/2}\\ &\le \varepsilon\left(\sum_{i\in I}h_{K_i}\|\nabla\eta\|^2_{L^2(\partial K_i)}\right)^{1/2} J_h^\sigma(\xi,\xi)^{1/2}\\ &\quad +\varepsilon\left(\sum_{i\in I}h_{K_i}\|\nabla\xi\|^2_{L^2(\partial K_i)}\right)^{1/2} J_h^\sigma(\eta,\eta)^{1/2}. \end{aligned}$$

Taking into account that $\eta = \Pi u - u \in H^2(\Omega,\mathcal{T}_h)$, we have

$$\|\nabla\eta\|_{L^2(K_i)} = |\eta|_{H^1(K_i)}, \quad |\nabla\eta|_{H^1(K_i)} = |\eta|_{H^2(K_i)}, i\in I. \tag{2.4.47}$$

Similarly as in (2.4.30), from (2.3.32) (where we set $v := \nabla\eta$) and (2.4.39) we find that

$$\|\nabla\eta\|^2_{L^2(\partial K_i)} \le 2C_6 C_{18}^2 |u|^2_{H^{p+1}(K_i)} h_{K_i}^{2p-1}. \tag{2.4.48}$$

Hence,

$$\left(\sum_{i\in I} h_{K_i}\|\nabla\eta\|^2_{L^2(\partial K_i)}\right)^{1/2} \le \sqrt{2C_6}\, C_{18}\, h^p|u|_{H^{p+1}(\Omega)}. \tag{2.4.49}$$

From (2.3.32) and (2.4.11), where we put $v := \nabla\xi$, we get

$$\begin{aligned}
\sum_{i\in I} h_{K_i}\|\nabla\xi\|^2_{L^2(\partial K_i)} &\le \sum_{i\in I} C_6\left(\|\nabla\xi\|^2_{L^2(K_i)} + h_{K_i}\|\nabla\xi\|_{L^2(K_i)}|\nabla\xi|_{H^1(K_i)}\right)\\
&\le \sum_{i\in I} C_6\left(\|\nabla\xi\|^2_{L^2(K_i)} + C_{17}\|\nabla\xi\|^2_{L^2(K_i)}\right)\\
&\le C_6(1+C_{17})|\xi|^2_{H^1(\Omega,\mathcal{T}_h)}.
\end{aligned} \tag{2.4.50}$$

Now it remains to estimate the term $J_h^\sigma(\eta,\eta)$. In virtue of (2.3.13), (2.3.23), (2.3.32) and (2.4.39), we have

$$\begin{aligned}
J_h^\sigma(\eta,\eta) &\le 2\sum_{i\in I}\sum_{j\in S(i)}\int_{\partial\Gamma_{ij}} \frac{1}{\gamma_{ij}}|\eta|^2\,\mathrm{d}S\\
&\le 2C_5\sum_{i\in I}\frac{1}{h_{K_i}}\int_{\partial K_i}|\eta|^2\,\mathrm{d}S\\
&\le 2C_5\,C_6\sum_{i\in I}\frac{1}{h_{K_i}}\left(\|\eta\|_{L^2(K_i)}|\eta|_{H^1(K_i)} + h_{K_i}^{-1}\|\eta\|^2_{L^2(K_i)}\right)\\
&\le 4C_5\,C_6\,C_{18}^2\,h^{2p}|u|^2_{H^{p+1}(\Omega)}.
\end{aligned} \tag{2.4.51}$$

Finally, from (2.4.46) – (2.4.51) we conclude that

$$|\sigma_2| \le \sigma(\varepsilon,h,u,\xi) := \varepsilon\, C\, h^p|u|_{H^{p+1}(\Omega)}\left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right), \tag{2.4.52}$$

where $C > 0$ is a constant independent of u, ξ, h, ε.

In a similar way we obtain the estimate of σ_3:

$$|\sigma_3| \le \sigma(\varepsilon,h,u,\xi). \tag{2.4.53}$$

Estimates (2.4.45), (2.4.52) and (2.4.53) immediately imply (2.4.43). □

Now we will formulate the *main result.*

Theorem 2.2 *Let assumptions (2.2.4), a) – e),* (H) *from Section 2.3.1 and* (A1), (A3) *from Section 2.3.2 be satisfied. Let u be the exact strong solution of problem (2.2.10) satisfying (2.4.1) and let u_h be the approximate solution defined by (2.4.7). Then the error $e_h = u_h - u$ satisfies the estimate*

$$\begin{aligned}
&\sup_{t\in[0,T]}\|e_h(t)\|^2_{L^2(\Omega)} + \varepsilon\int_0^T\left(|e_h(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(e_h(\vartheta),e_h(\vartheta))\right)\mathrm{d}\vartheta\\
&+C_{24}(1+\varepsilon)\int_0^T\exp\left(C_{24}\frac{1+\varepsilon}{\varepsilon}(T-\vartheta)\right)\\
&\times\left\{\int_0^\vartheta\left(|e_h(\alpha)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(e_h(\alpha),e_h(\alpha))\right)\mathrm{d}\alpha\right\}\mathrm{d}\vartheta\\
\le\;& 2h^{2p}\left\{\left(\frac{c_1(\varepsilon,h)}{\varepsilon} + c_2 + \varepsilon C_{26}\right)\exp\left(C_{24}\frac{1+\varepsilon}{\varepsilon}T\right) + C_{25}\right\},
\end{aligned} \tag{2.4.54}$$

with constants C_{24}, C_{25}, C_{26}, $c_2 > 0$ independent of h, ε and with a function $c_1(\varepsilon,h)$, specified in the proof.

Proof. As in (2.4.33), we set $\xi = u_h - \Pi u \in S_h$, $\eta = \Pi u - u$. Then (2.4.34) holds: $e_h = u_h - u = \xi + \eta$. From (2.4.35), (2.4.36) and the relation $a_h(\xi,\xi) = \varepsilon|\xi|^2_{H^1(\Omega,\mathcal{T}_h)}$ we get

$$\begin{aligned} &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\Omega)} + \varepsilon|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + \varepsilon J_h^\sigma(\xi,\xi) \\ = \;& b_h(u,\xi) - b_h(u_h,\xi) - \left(\frac{\partial\eta}{\partial t},\xi\right) - a_h(\eta,\xi) - \varepsilon J_h^\sigma(\eta,\xi) \quad \text{in } (0,T). \end{aligned} \tag{2.4.55}$$

Now from (2.4.37) – (2.4.40), (2.4.43) and (2.4.51), using Young's inequality, we derive the estimates

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\Omega)} + 2\varepsilon|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + 2\varepsilon J_h^\sigma(\xi,\xi) \\ \le\;& 2C_{22}\left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right)\left(\|\xi\|_{L^2(\Omega)} + h^{p+1}|u|_{H^{p+1}(\Omega)}\right) \\ &+2C_{18}h^{p+1}\left|\frac{\partial u}{\partial t}\right|_{H^{p+1}(\Omega)}\|\xi\|_{L^2(\Omega)} + 2C_{23}\varepsilon\, h^p|u|_{H^{p+1}(\Omega)}\left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right) \\ &+2\varepsilon\sqrt{4C_5C_6}C_{18}h^p|u|_{H^{p+1}(\Omega)}J_h^\sigma(\xi,\xi)^{1/2} \\ \le\;& 2C_{24}^*\left(J_h^\sigma(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right)\left(\|\xi\|_{L^2(\Omega)} + (\varepsilon h^p + h^{p+1})|u|_{H^{p+1}(\Omega)}\right) \\ &+2C_{18}h^{p+1}\|\xi\|_{L^2(\Omega)}\left|\frac{\partial u}{\partial t}\right|_{H^{p+1}(\Omega)} \\ \le\;& \varepsilon J_h^\sigma(\xi,\xi) + \varepsilon|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + C_{24}\left(1+\frac{1}{\varepsilon}\right)\|\xi\|^2_{L^2(\Omega)} \\ &+\frac{C_{24}}{\varepsilon}\left((\varepsilon^2h^{2p} + h^{2p+2})\,|u|^2_{H^{p+1}(\Omega)}\right) + C_{18}^2h^{2p+2}\left|\frac{\partial u}{\partial t}\right|^2_{H^{p+1}(\Omega)} \quad \text{in } (0,T), \end{aligned} \tag{2.4.56}$$

where $C_{24}^* = \max\{C_{22}, C_{23} + \sqrt{4C_5C_6}C_{18}\}$, $C_{24} = \max\{1, 8(C_{24}^*)^2\}$. Hence,

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\|\xi(t)\|^2_{L^2(\Omega)} + \varepsilon|\xi(t)|^2_{H^1(\Omega,\mathcal{T}_h)} + \varepsilon J_h^\sigma(\xi(t),\xi(t)) \\ \le\;& C_{24}\left(1+\frac{1}{\varepsilon}\right)\|\xi(t)\|^2_{L^2(\Omega)} + \frac{C_{24}}{\varepsilon}\left((\varepsilon^2h^{2p} + h^{2p+2})\,|u(t)|^2_{H^{p+1}(\Omega)}\right) \\ &+C_{18}^2h^{2p+2}\left|\frac{\partial u(t)}{\partial t}\right|^2_{H^{p+1}(\Omega)}, \quad t \in (0,T). \end{aligned} \tag{2.4.57}$$

From assumption (2.4.1) it follows that the right-hand side of (2.4.57) is integrable over $(0,T)$. In virtue of (2.4.7), c), Remark 3.3 and estimate (2.4.39) we have

$$\|xi(0)\|^2_{L^2(\Omega)} \le 4C_{18}^2h^{2p+2}|u^0|^2_{H^{p+1}(\Omega)}. \tag{2.4.58}$$

The integration of (2.4.57) from 0 to $t \in [0,T]$ yields

$$\begin{aligned} &\|\xi(t)\|^2_{L^2(\Omega)} + \varepsilon\int_0^t\left(|\xi(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(\xi(\vartheta),\xi(\vartheta))\right)\mathrm{d}\vartheta \\ \le\;& C_{24}\left(1+\frac{1}{\varepsilon}\right)\int_0^t\|\xi(\vartheta)\|^2_{L^2(\Omega)}\,\mathrm{d}\vartheta + \frac{C_{24}}{\varepsilon}h^{2p}\int_0^t\left((\varepsilon^2+h^2)\,|u(\vartheta)|^2_{H^{p+1}(\Omega)}\right)\mathrm{d}\vartheta \\ &+C_{18}^2h^{2p+2}\int_0^t\left|\frac{\partial u(\vartheta)}{\partial t}\right|^2_{H^{p+1}(\Omega)}\mathrm{d}\vartheta + 4C_{18}^2h^{2p+2}|u^0|^2_{H^{p+1}(\Omega)}, \quad t \in [0,T]. \end{aligned} \tag{2.4.59}$$

Now the application of Gronwall's Lemma 2.10, where we set

$$
\begin{aligned}
y(t) &= |\xi(t)\|_{L^2(\Omega)}, \\
q(t) &= \varepsilon \int_0^t \left(|\xi(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(\xi(\vartheta),\xi(\vartheta)) \right) \mathrm{d}\vartheta, \\
r(t) &= C_{24}\frac{1+\varepsilon}{\varepsilon}, \\
z(t) &= h^{2p}\left(\frac{c_1}{\varepsilon} + c_2 h^2\right), \\
c_1 &= c_1(\varepsilon,h) = C_{24}\left(\left(\varepsilon^2 + h^2\right) \|u\|^2_{L^2(0,T;H^{p+1}(\Omega))} \right), \\
c_2 &= C_{18}^2 \left(\|\partial u/\partial t\|^2_{L^2(0,T;H^{p+1}(\Omega))} + 4|u^0|^2_{H^{p+1}(\Omega)} \right),
\end{aligned}
\tag{2.4.60}
$$

and the relations

$$
\begin{aligned}
&\int_\vartheta^t r(s)\mathrm{d}s = C_{24}\frac{1+\varepsilon}{\varepsilon}(t-\vartheta), \\
&\int_0^t r(\vartheta)\, z(\vartheta) \exp\left(\int_\vartheta^t r(s)\,\mathrm{d}s \right) \mathrm{d}\vartheta = \left(\frac{c_1}{\varepsilon} + c_2 h^2\right) h^{2p} \left(\exp\left(C_{24}\frac{1+\varepsilon}{\varepsilon} t\right) - 1 \right)
\end{aligned}
\tag{2.4.61}
$$

imply that

$$
\begin{aligned}
&|\xi(t)\|^2_{L^2(\Omega)} + \varepsilon \int_0^t \left(|\xi(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(\xi(\vartheta),\xi(\vartheta)) \right) \mathrm{d}\vartheta \\
&+C_{24}(1+\varepsilon)\int_0^t \exp\left(C_{24}\frac{1+\varepsilon}{\varepsilon}(t-\vartheta)\right) \left\{ \int_0^\vartheta \left(|\xi(\alpha)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(\xi(\alpha),\xi(\alpha)) \right) \mathrm{d}\alpha \right\} \mathrm{d}\vartheta \\
\leq\ & \left(\frac{c_1}{\varepsilon} + c_2 h^2\right) h^{2p} \exp\left(C_{24}\frac{1+\varepsilon}{\varepsilon} t\right), \quad t \in [0,T].
\end{aligned}
\tag{2.4.62}
$$

Now, since $e_h = \xi + \eta$, we can write

$$
\begin{aligned}
\|e_h\|^2_{L^2(\Omega)} &\leq 2\left(|\xi\|^2_{L^2(\Omega)} + \|\eta\|^2_{L^2} \right), \\
|e_h|^2_{H^1(\Omega,\mathcal{T}_h)} &\leq 2\left(|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + |\eta|^2_{H^1(\Omega,\mathcal{T}_h)} \right), \\
J_h^\sigma(e_h,e_h) &\leq 2\left(J_h^\sigma(\xi,\xi) + J_h^\sigma(\eta,\eta) \right).
\end{aligned}
\tag{2.4.63}
$$

Using assumption (2.4.1), from (2.4.51), (2.4.62), (2.4.63) and (2.4.39) we deduce that

$$
\begin{aligned}
&\|e_h(t)\|^2_{L^2(\Omega)} + \varepsilon \int_0^t \left(|e_h(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(e_h(\vartheta),e_h(\vartheta)) \right) \mathrm{d}\vartheta \\
&+C_{24}(1+\varepsilon)\int_0^t \exp\left(C_{24}\frac{1+\varepsilon}{\varepsilon}(t-\vartheta)\right) \left\{ \int_0^\vartheta \left(|e_h(\alpha)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(e_h(\alpha),e_h(\alpha)) \right) \mathrm{d}\alpha \right\} \mathrm{d}\vartheta \\
\leq\ & 2h^{2p}\left\{ \left(\frac{c_1}{\varepsilon} + c_2 h^2\right) \exp\left(C_{24}\frac{1+\varepsilon}{\varepsilon} t\right) + \varepsilon(1+4C_5C_6)\,C_{18}^2 \int_0^t |u(\vartheta)|^2_{H^{p+1}(\Omega)}\,\mathrm{d}\vartheta \right. \\
&+C_{24}(1+\varepsilon)\,(1+4C_5C_6)\,C_{18}^2 \int_0^t \exp\left(C_{24}\frac{1+\varepsilon}{\varepsilon}(t-\vartheta)\right) \int_0^\vartheta |u(\alpha)|^2_{H^{p+1}(\Omega)}\mathrm{d}\alpha\,\mathrm{d}\vartheta \\
&\left. + C_{18}^2 h^{2p+2} |u(t)|^2_{H^{p+1}(\Omega)} \right\} \\
\leq\ & 2h^{2p}\left\{ \left(\frac{1}{\varepsilon} c_1(\varepsilon,h) + c_2 h^2 + \varepsilon C_{26} \right) \exp\left(C_{24}\frac{1+\varepsilon}{\varepsilon} T\right) + h^2 C_{25} \right\},
\end{aligned}
\tag{2.4.64}
$$

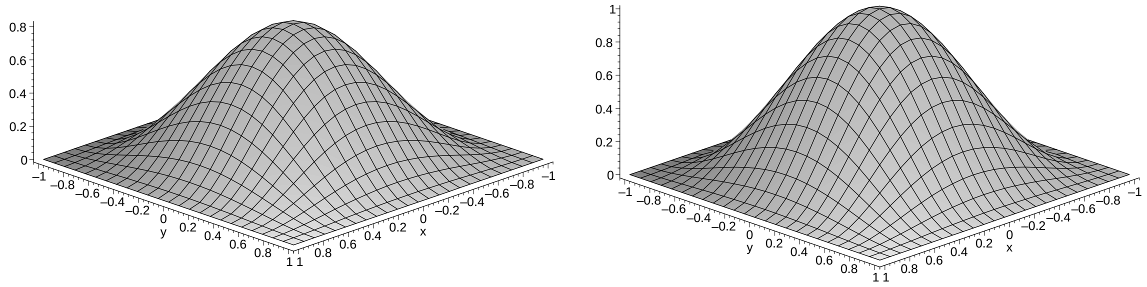


Figure 2.10: The exact solution at $t = 1$ (left) and the steady state solution for $t \to \infty$ (right)

where

$$
\begin{aligned}
C_{25} &= C_{18}^2 \|u\|^2_{L^\infty(0,T;H^{p+1}(\Omega))}, \\
C_{26} &= (1 + 4C_5C_6) C_{18}^2 |u|^2_{L^2(0,T;H^{p+1}(\Omega))}.
\end{aligned}
\tag{2.4.65}
$$

This already implies (2.4.54). □

2.4.5 Numerical examples

In this section we present some numerical experiments carried out for the 2D viscous Burgers equation

$$
\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x_1} + u \frac{\partial u}{\partial x_2} = \varepsilon \Delta u + g, \qquad \text{in } \Omega \times (0,T), \tag{2.4.66}
$$

where $\Omega = (-1,1)^2$, $T = 1$, equipped with initial condition (2.2.3) and boundary condition (2.2.2).

We set $\varepsilon = 0.002$ and define the function g and the initial condition in such a way that the exact solution has the form

$$
u(x_1, x_2, t) = (1 - x_1^2)^2 (1 - x_2^2)^2 \left(1 - \frac{e^{-t}}{2}\right) \tag{2.4.67}
$$

Figure 2.10 shows the exact solution (2.4.67) at $t = 1$ and for $t \to \infty$ when we achieve the steady state solution

$$
u(x_1, x_2) = (1 - x_1^2)^2 (1 - x_2^2)^2. \tag{2.4.68}
$$

We solve the initial-boundary value problem by the presented numerical method. In the definition of the form b_h the following numerical flux is used

$$
H(u_1, u_2, \boldsymbol{n}) = \left\{ \begin{array}{ll} \sum_{s=1}^2 f_s(u_1) n_s, & \text{if } A > 0 \\ \sum_{s=1}^2 f_s(u_2) n_s, & \text{if } A \le 0 \end{array} \right. , \tag{2.4.69}
$$

where

$$
A = \sum_{s=1}^2 f_s'(\bar{u}) n_s, \quad \bar{u} = \frac{1}{2}(u_1 + u_2). \tag{2.4.70}
$$

The resulting system of ODE's is solved by the forward Euler method with a very small time step $\tau = 10^{-4}$, which guarantees stability and sufficiently precise resolution with respect to time.

			$t = 1$		$t \to \infty$	
l	$\#\mathcal{T}_{h_l}$	h_l	e_h	α_l	e_h	α_l
1	136	2.638E-01	1.4778E-02	–	3.5589E-02	–
2	276	1.854E-01	6.6063E-03	2.282	2.7337E-02	0.748
3	586	1.269E-01	3.1208E-03	1.977	7.3604E-03	3.459
4	1046	9.516E-02	1.6331E-03	2.251	2.3878E-03	3.913
5	2374	6.308E-02	7.0742E-04	2.035	1.2455E-03	1.583
6	4150	4.772E-02	4.2615E-04	1.817	6.7826E-04	2.179
global order of accuracy $\bar{\alpha}$				2.078		2.497

Table 2.3: Computational errors in L^2-norm, the values of $\alpha_l,\ l = 2, \dots, 6$ and $\bar{\alpha}$.

The computational error of the solution is evaluated at time t over a domain Ω in $L^2(\Omega)-$norm:

$$\|e_h(t)\|_{L^2(\Omega)} \equiv \|u_h(\cdot, t) - u(\cdot, t)\|_{L^2(\Omega)}. \tag{2.4.71}$$

We investigate the error at the time $t = 1$ and for $t \to \infty$ and set $e_h = \|e_h(1)\|_{L^2(\Omega)}$. We suppose that the error e_h behaves according to the formula

$$e_h \approx Ch^{\alpha}, \tag{2.4.72}$$

where $C > 0$ is a constant not depending on h and α is the *order of accuracy* of the method. The numerical solution was computed for 6 unstructured meshes ($\mathcal{T}_{h_l},\ l = 1, \dots, 6$) with different h_l, see Figure 2.11 and Table 2.3. We set the *local order of accuracy* by

$$\alpha_l = \frac{\log\left(e_{h_l}/e_{h_{l-1}}\right)}{\log\left(h_l/h_{l-1}\right)}, \quad l = 2, \dots, 6. \tag{2.4.73}$$

Moreover we set the *global order of accuracy* $\bar{\alpha}$ by the least square method. Table 2.3 shows the L^2- error and the values of $\alpha_l,\ l = 2, \dots, 6$ and $\bar{\alpha}$. Figures 2.12 and 2.13 show the computed numerical results. We observe in fact smooth numerical approximation (for $l = 5, 6$) even the discontinuous approximation is employed. The behaviour of the error in dependence on h is plotted in Figure 2.14.

The numerical order of accuracy α is approximately equal to 2, so that it indicate that the order of the numerical scheme is higher that our theoretical results: $O(h)$ for $p = 1$.

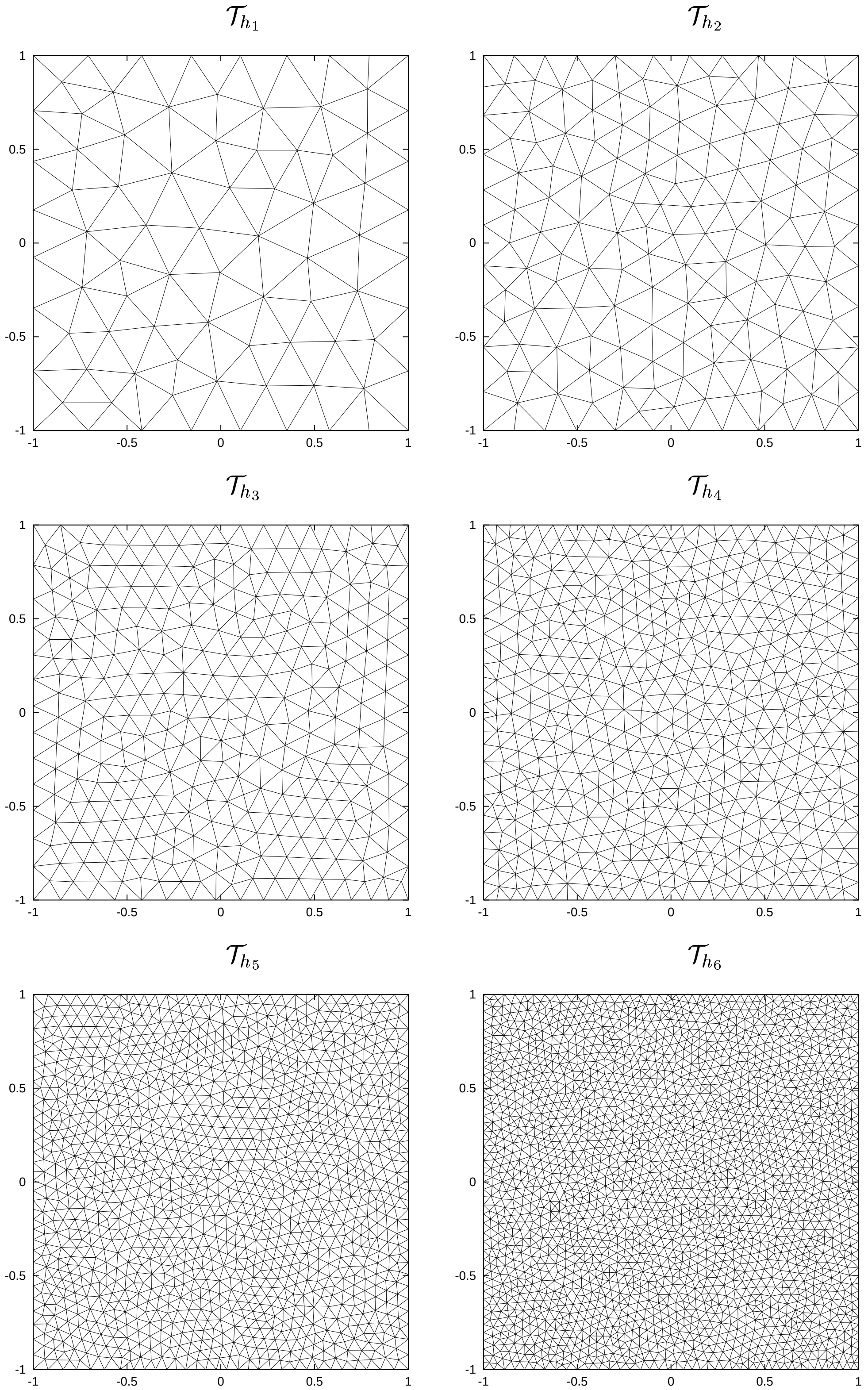


Figure 2.11: Triangulations $\mathcal{T}_{h_l},\ l = 1, \ldots, 6$ used for the computation

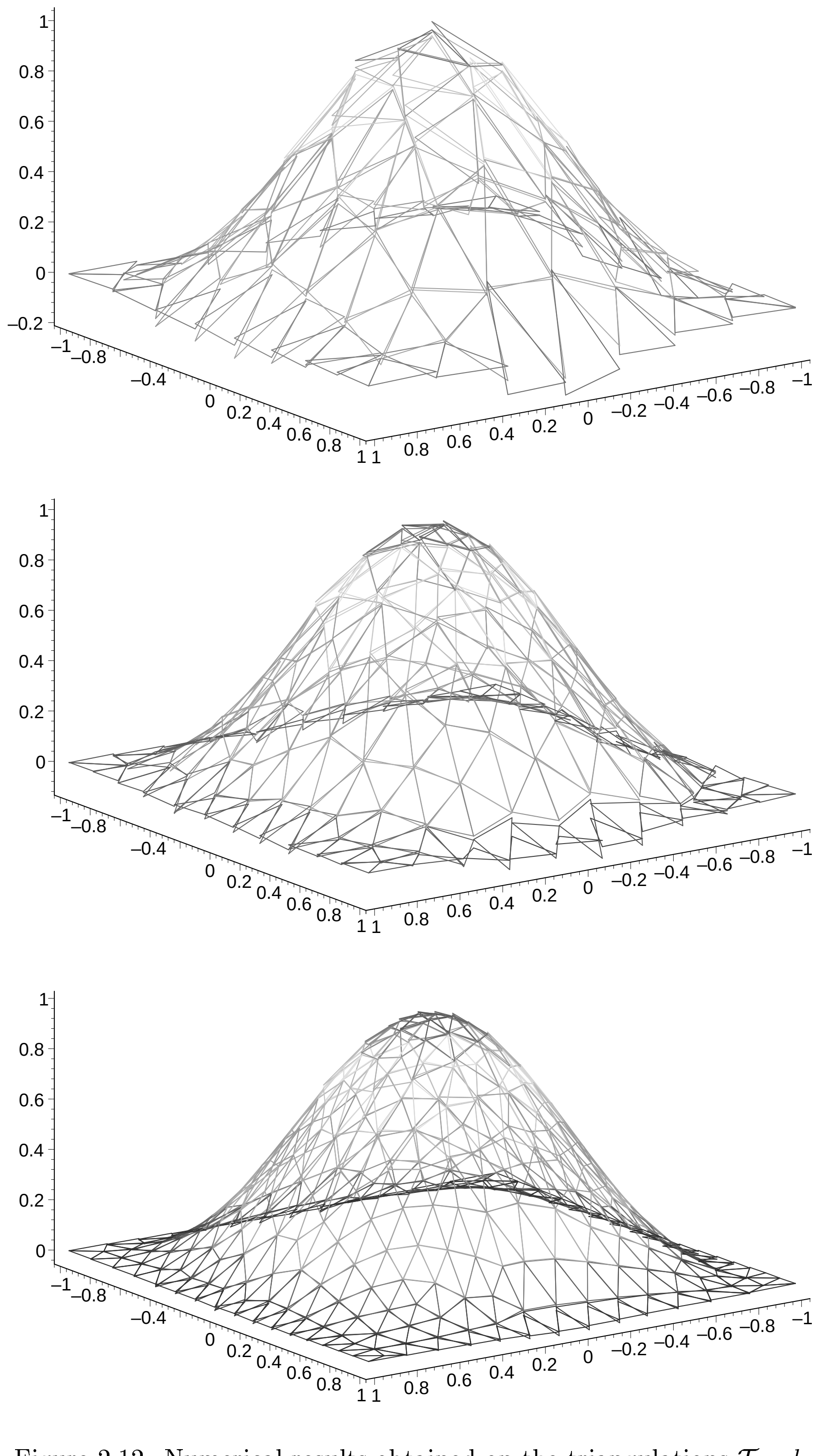


Figure 2.12: Numerical results obtained on the triangulations $\mathcal{T}_{h_l},\ l = 1, \ldots, 3$.

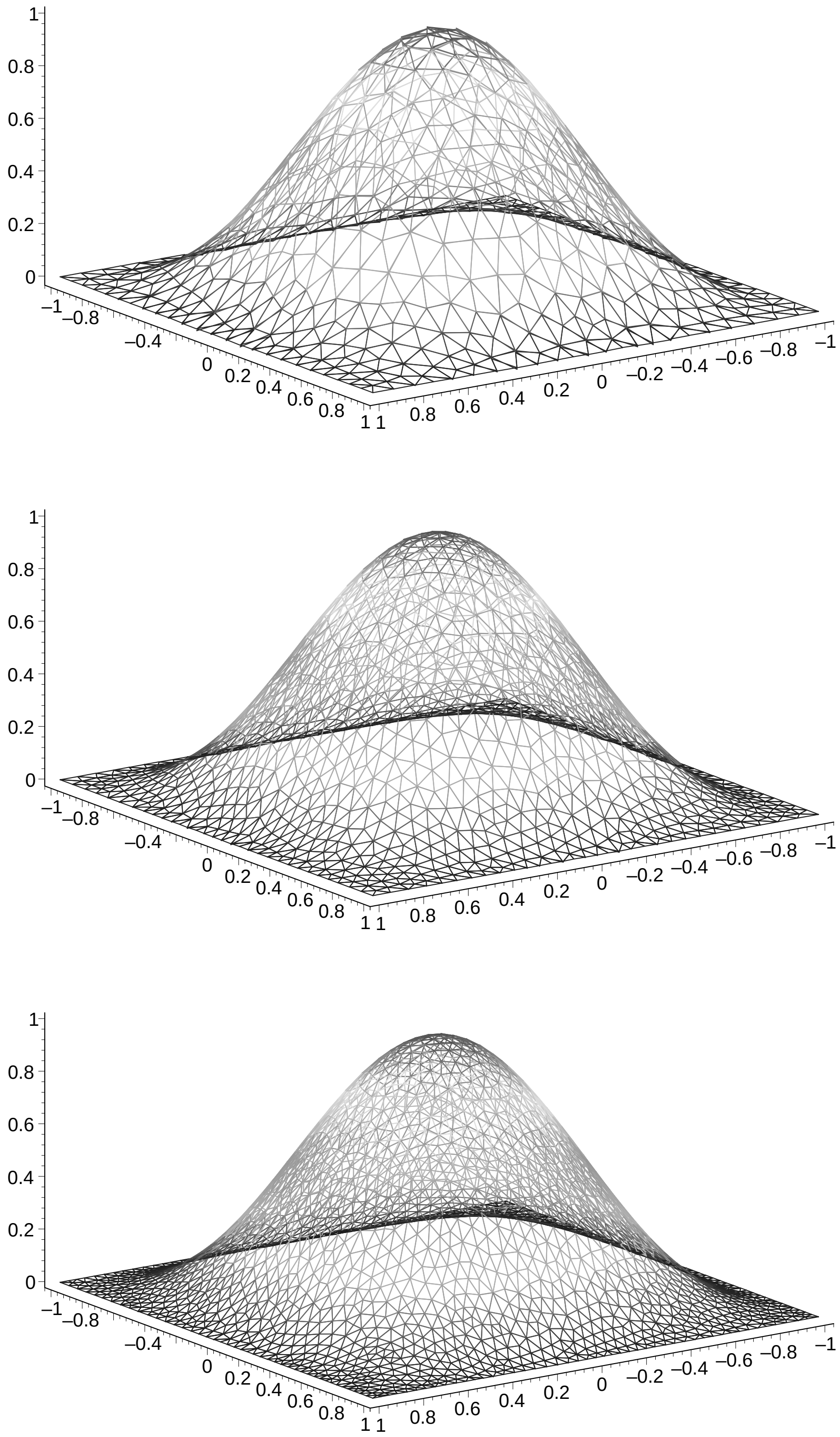


Figure 2.13: Numerical results obtained on the triangulations $\mathcal{T}_{h_l},\ l = 4, \ldots, 6$.

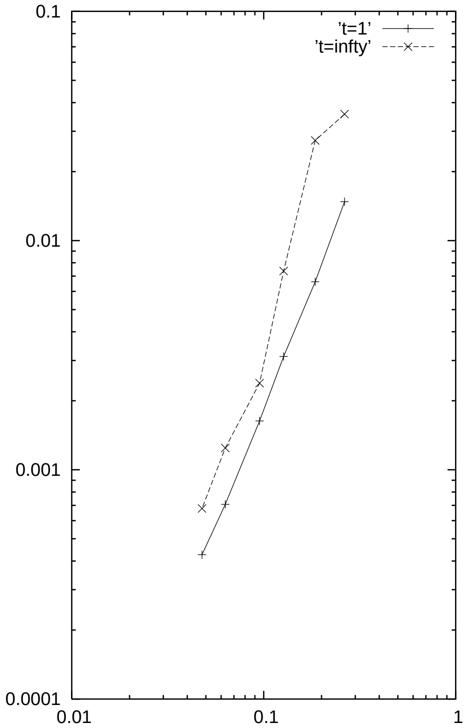


Figure 2.14: Dependence of $e_h(t)$ on h in L^2-norm for $t = 1$ (full line) and $t \to \infty$ (dashed line)

2.5 Scalar convection-diffusion equation – computations

In Section 2.2 we present two method for the numerical solution of a scalar convection-diffusion equation. We have assumed that the exact solution was sufficiently regular. But our goal is to simulate the flow of compressible fluids where the physical quantities describing the flow (density, velocity, pressure) can contain discontinuities. So that the natural question is how the presented method (FV DG and DG FE) solve the problems with discontinuities or steep gradients. The following examples show that the numerical solutions suffer from spurious oscillations in the vicinity of discontinuities or steep gradients. Our goal is to modify the scheme in order to suppress these oscillations but on the other hand to preserve the order of accuracy in regions where the solution is regular.

2.5.1 Numerical solution of a problems with steep gradients

We consider the numerical solution of the viscous Burgers equation

$$\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x_1} + u\,\frac{\partial u}{\partial x_2} = \varepsilon\,\Delta\,u \quad \text{in } Q_T = \Omega \times (0,T), \tag{2.5.1}$$

where $\Omega = (-1,1)^2$, $T = 0.45$, $\varepsilon = 0.002$. The equation (2.5.1) is equipped with periodic boundary conditions and the initial condition

$$u^0(x_1,x_2) = 0.25 + 0.5\sin(\pi(x_1+x_2)), \quad (x_1,x_2) \in \Omega. \tag{2.5.2}$$

Let us note that the periodic boundary conditions are easily treated using a simple modification of schemes described in Sections 2.3 and 2.4.

The inviscid Burgers equation ($\varepsilon = 0$) equipped with the initial condition (2.5.2) and periodic boundary conditions has a discontinuous solution for $t \geq 0.3$. The viscous case, when $\varepsilon = 0.002$, has a similar solution (the discontinuity is replaced by an internal layer due to the diffusion). In our numerical experiments we use again the numerical flux in the form

$$H(u_1,u_2,\boldsymbol{n}) = \left\{ \begin{array}{ll} \sum_{s=1}^2 f_s(u_1)n_s, & \text{if } A > 0 \\ \sum_{s=1}^2 f_s(u_2)n_s, & \text{if } A \leq 0 \end{array} \right., \tag{2.5.3}$$

where $f_s(u) = u^2/2, A = \sum_{s=1}^2 f_s'(\bar{u})n_s,\ \bar{u} = \frac{1}{2}(u_1+u_2)$.

In Figures 2.15 and 2.16 we see the numerical solution at $t = 0.45$ obtained with the aid of the schemes FV DG and DG FE methods, respectively. There the forward Euler time discretization is carried out with a rather small time step $\tau = 10^{-4}$, in order to guarantee the stability and sufficiently accurate resolution in time. (For more details see Section 2.5.2.) The numerical solutions suffer from spurious oscillations in both cases, in the first case the oscillations are noticeable smaller. On the other hand the order of accuracy of the first method is only one in the whole computational domain.

2.5.2 Second-order DG FE method with order limiting

General framework of higher-order schemes for conservation laws

In this section we develop a numerical technique avoiding disadvantages of both schemes discussed above: spurious oscillations in solutions obtained by the pure DG FE method from Section 2.4 and a low order (=first order) of the FV DG scheme from Section 2.3.

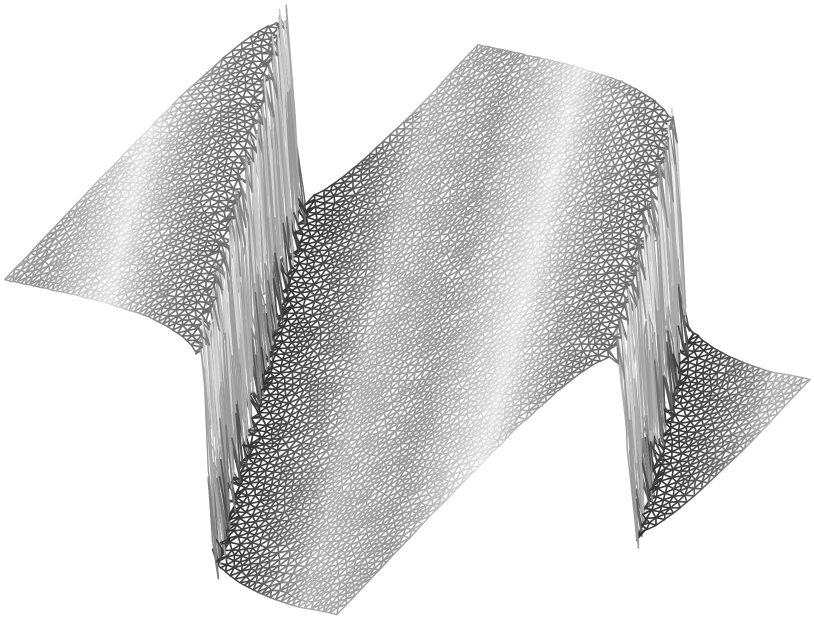


Figure 2.15: FV DG solution of (2.5.1) – (2.5.2) plotted at $t = 0.45$.

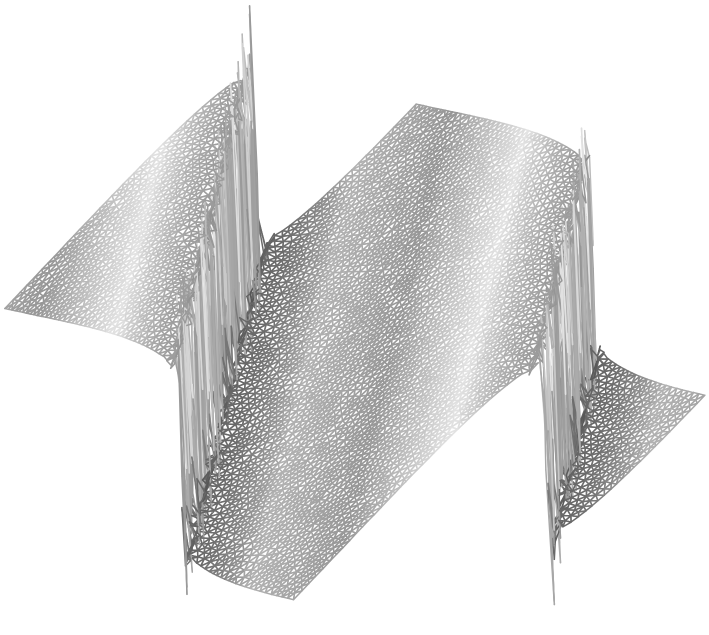


Figure 2.16: DG FE solution of (2.5.1) – (2.5.2) plotted at $t = 0.45$.

All methods for avoiding the Gibbs effect in higher-order schemes are based on a limiting procedure applied in the whole computational domain which should however suppress the accuracy to the first order in the vicinity of steep gradients or discontinuities in the solution. There exists a rigorous method using the concept of a flux limiter for finite difference or finite volume schemes applied to scalar conservation laws with one space variable over uniform meshes. (See, e. g., [Swe84]). However, as numerical experiments show, the heuristic extension of this approach to FV schemes for systems, problems with more space variables and/or nonuniform unstructured meshes can lead to an undesired decrease of accuracy in the whole computational domain. The application of this methodology in the framework of the DG FE method seems to be inadequate.

Here we present a different type of limiting which keeps the order of accuracy in the region of regularity of the solution, because the original higher order scheme remains unchanged in this region and is modified to a first order scheme only in the vicinity of steep gradients or discontinuities.

Fully discrete DG FE scheme with adaptive limiting of order

Let us return to scheme (2.4.7), where *we suppose that* $d = 2$, $\mathcal{T}_h$ is formed by triangles and $p = 1$. We carry out the discretization in time by the forward Euler method. To this end, we consider a partition $0 = t_0 < t_1 < t_2 < \dots$ of the time interval $(0, T)$ and set $\tau_k = t_{k+1} - t_k$.

The *fully discrete problem* of (2.4.7) reads: for each $k \geq 0$ find u_h^{k+1} such that

$$\text{a)} \qquad u_h^{k+1} \in S_h = S_h^{1,-1}(\Omega, \mathcal{T}_h), \tag{2.5.4}$$

$$\begin{aligned} \text{b)} \qquad (u_h^{k+1}, v_h) &= (u_h^k, v_h) - \tau_k\, a_h(u_h^k, v_h) \\ &\quad - \tau_k\, b_h(u_h^k, v_h) - \tau_k\, J_h(u_h^k, v_h) + \tau_k\, \ell_h(v_h)\,(t_k) \quad \forall\, v_h \in S_h. \end{aligned}$$

The formal order of accuracy is $O(\tau + h^2)$. The stability of the scheme requires that $\tau_k \approx O(h^2/\varepsilon + h)$.

Let us go back to problem (2.5.1) – (2.5.2) with periodic boundary conditions. Our first task is to identify areas with very steep gradients (leading to discontinuities for $\varepsilon \to 0+$). Numerical experiments show (see Figures 2.18–2.19) that the interelement jumps in the approximate solution are of the order $O(1)$ on discontinuities, but $O(h^2)$ in the areas where the solution is regular. On general unstructured grids it appears to be suitable to measure the magnitude of interelement jumps in the integral form by

$$\int_{\partial K_i} [u_h^k]^2 \, \mathrm{d}S, \quad K_i \in \mathcal{T}_h. \tag{2.5.5}$$

In areas where the solution is regular, we have

$$g^1(i) = \int_{\partial K_i} [u_h^k]^2 \, \mathrm{d}S/h^5 \approx \int_{\partial K_i} (O(h^2))^2 \, \mathrm{d}S/h^5 \approx O(1), \tag{2.5.6}$$

whereas

$$g^2(i) = \int_{\partial K_i} [u_h^k]^2 \, \mathrm{d}S/h \approx \int_{\partial K_i} (O(1))^2 \, \mathrm{d}S/h \approx O(1) \tag{2.5.7}$$

for very steep gradients (discontinuities). These results lead us to the idea that the switch between the higher order scheme and its first order modification should be tested with the aid of the indicator

$$g(i) = \int_{\partial K_i} [u_h^k]^2 \,\mathrm{d}S/h^\alpha, \quad K_i \in \mathcal{T}_h, \tag{2.5.8}$$

where $\alpha \in (1, 5)$. We choose, e. g., $\alpha = 5/2$. However, in general, using unstructured grids, it is suitable to define the indicator $g(i)$ in terms of h_{K_i} and $|K_i|$. For $\alpha = 5/2$ we set

$$g(i) = \int_{\partial K_i} [u_h^k]^2 \,\mathrm{d}S/(h_{K_i}|K_i|^{3/4}), \quad K_i \in \mathcal{T}_h. \tag{2.5.9}$$

Now we define an *adaptive strategy* for an *automatic limiting* of the order of accuracy of the scheme (2.5.4):

$$\text{a)} \qquad u_h^{k+1} \in S_h = S_h^{1,-1}(\Omega, \mathcal{T}_h), \tag{2.5.10}$$

$$\begin{aligned} \text{b)} \qquad & (u_h^{k+1}, v_h) = (\tilde{u}_h^k, v_h) - \tau_k\, a_h(u_h^k, v_h) \\ & -\tau_k\, \tilde{b}_h(\tilde{u}_h^k, v_h) - \tau_k J_h(u_h^k, v_h) + \tau_k\, \ell_h(v_h)\,(t_k) \quad \forall v_h \in S_h \end{aligned}$$

where $\tilde{u}_h^k$ is the modification of u_h^k defined with the aid of our limiting strategy in the following way:

$$\text{a)} \qquad \text{Set} \quad \tilde{u}_h^k|_{K_i} := u_h^k|_{K_i} \quad \forall i \in I. \tag{2.5.11}$$

$$\text{b)} \qquad \text{If } g(i) > 1 \text{ for some } i \in I, \text{ then set } \tilde{u}_h^k|_{K_i} := \pi_0\, u_h^k|_{K_i}.$$

This means that in (2.5.11) the limiting of the order of the scheme is applied on the elements lying on the discontinuity. In other areas the second order of accuracy is preserved. (The extension to the case $d = 3$ is straightforward.)

Performance of adaptive limiting

At first we show that the interelement jumps in the approximate solution are of the order $O(1)$ on discontinuities, whereas $O(h^2)$ in the regions, where the solution is regular. To demonstrate it, we consider the inviscid case ($\varepsilon = 0$) of (2.5.1) – (2.5.2). This is the limit case of the considered problem with the worst scenario with respect to the behaviour of the solution.

We denote by $\overline{u}$ the exact solution of (2.5.1) – (2.5.2) with $\varepsilon = 0$ at time $t = 0.45$. It was mentioned above that the function $\overline{u}$ contains two line discontinuities expressed as $x_1 + x_2 = -0.78, x_1 + x_2 = 1.22, \quad (x_1, x_2) \in \Omega$, see Figure 2.17. We denote by Ω^d a union of triangles $K_i \in \mathcal{T}_h$ lying in the vicinity of these lines. Further we put $\Omega^c \equiv \Omega \setminus \Omega^d$.

We denote by $\overline{u}_h$ the approximation of $\overline{u}$ computed by the numerical scheme (2.5.10) – (2.5.11) on a mesh $\mathcal{T}_h$. We investigate the magnitude of interelement jumps of $\overline{u}_h$ with the aid of indicators $g^1(i)$, $g^2(i)$ and $g(i)$ given for $K_i \in \mathcal{T}_h$ by (2.5.6), (2.5.7) and (2.5.9), respectively. Figure 2.18 shows the distribution of magnitudes of g^1, g^2 and g computed on the triangulation mesh.500 from Figure 2.20. There is a point of this graph assigned to each $K_i \in \mathcal{T}_h$. The horizontal axis represents the value of $x_1^i + x_2^i$, where (x_1^i, x_2^i) are the coordinates of the barycentre of the element K_i and the vertical axis corresponds to the values $g^1(i)$, $g^2(i)$

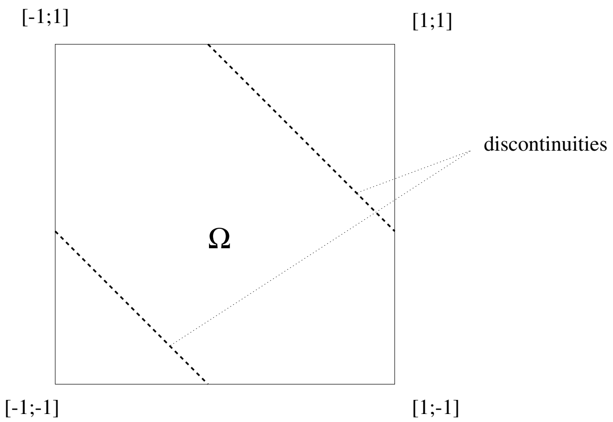


Figure 2.17: Discontinuities in the computational domain.

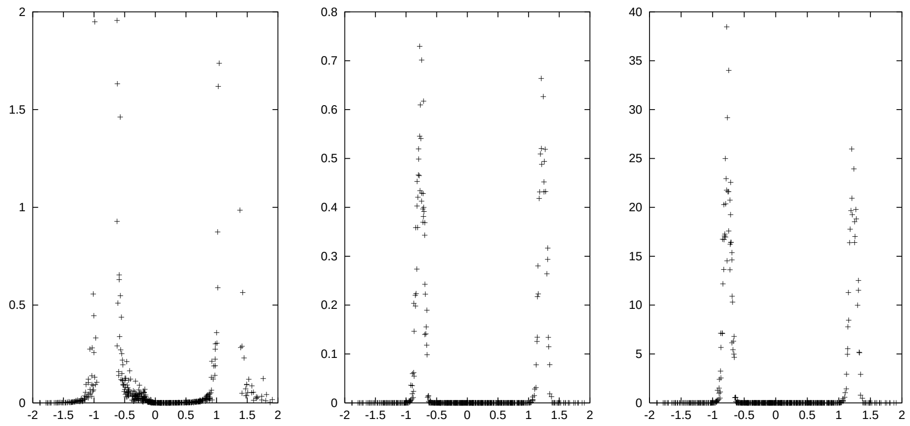


Figure 2.18: Dependence of the quantities $g^1(i)$ (left), $g^2(i)$ (center) and $g(i)$ (right) on the position $x_1 + x_2$ (vertical axes) for triangulation mesh.500.

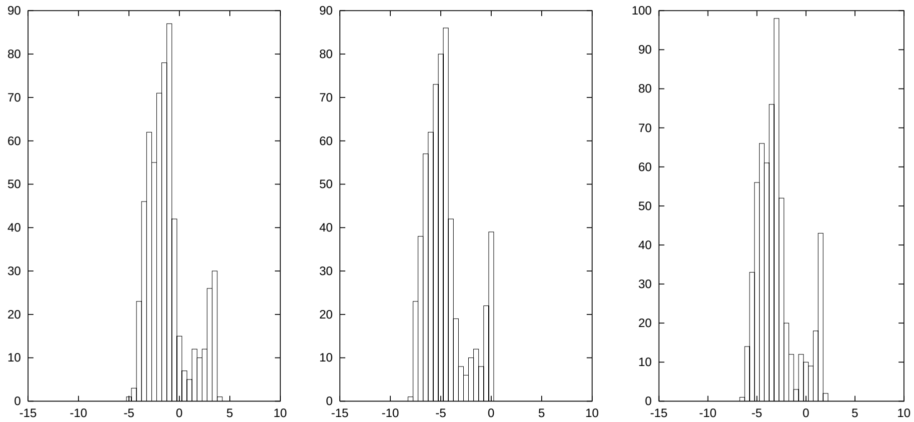


Figure 2.19: Distribution of triangles: horizontal axis (in logarithmic scale) expresses the order of magnitude of $g^1(i)$ (left), $g^2(i)$ (center) and $g(i)$ (right), vertical axis determines the number of triangles, where g^1, g^2 or g have the corresponding order, for triangulation mesh.500.

and $g(i)$ for $K_i \in \mathcal{T}_h$. We see the increments of values of g^1, g^2 and g near the discontinuities. For $K_i \in \Omega^d$, the values of g^2 are from the interval (0.1, 1), which confirms relation (2.5.7). On the other hand, for $K_i \in \Omega^c$ the values of g^1 satisfy the condition $g^1(i) \leq 1$ and this confirms (2.5.6). From the right graph in Figure 2.18 we can see that the limiting given by the condition $g(i) > 1$ indicates the region of steep gradient (discontinuity). Figure 2.19 shows the distribution of triangles. The horizontal axis (in logarithmic scale) expresses the order of magnitude of the values $g^1(i)$, $g^2(i)$ or $g(i)$ and the vertical axis gives the number of triangles, where g^1, g^2 or g have the corresponding order. The Gaussian distribution of elements in Ω^c as well as Ω^d is observed. The graphs with g^1 and g^2 again verify the relations (2.5.6) and (2.5.7), respectively.

Further, Figure 2.20 shows the marked triangles $K_i \in \mathcal{T}_h$ with $g(i) > 1$ where the limiting (2.5.11) should be applied. This proves the correctness of the choice $\alpha = 5/2$ in (2.5.8). Moreover, Figure 2.21 shows the marked triangles ($g(i) > 1$) for another unstructured anisotropic grids, which verifies the form of the denominator in (2.5.9) (i. e. $h_{K_i}|K_i|^{3/4}$).

The efficiency of the limiting is viewed in Figure 2.22, where we present the numerical solution of the inviscid Burgers equation. Figures show the graphs of the approximate solutions along the diagonal connecting the points $(-1, -1)$ and $(1, 1)$. The centre figure is the exact solution and the left and right one are the numerical solutions computed without and with limiting, respectively. The elimination of spurious oscillations and satisfactory correspondence with the exact solution are observed.

The last Figure 2.23 shows the numerical solution of (2.5.1)–(2.5.2) computed by the scheme (2.5.10) - (2.5.11).

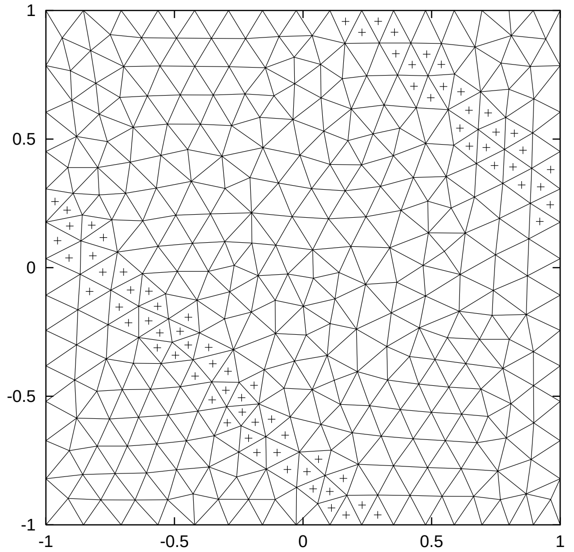


Figure 2.20: Triangulation mesh.500 with marked triangles with $g(i) > 1$.

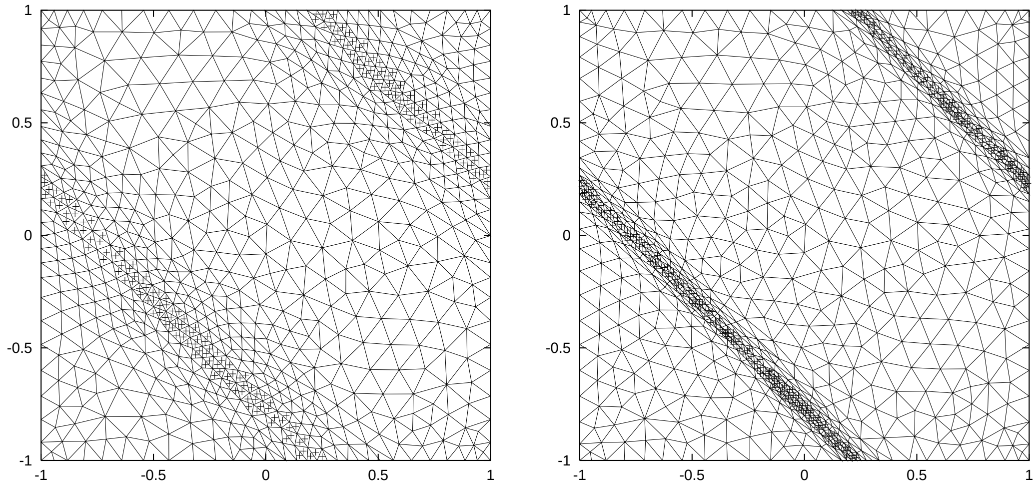


Figure 2.21: Examples of anisotropic triangulations with marked triangles satisfying the condition $g(i) > 1$.

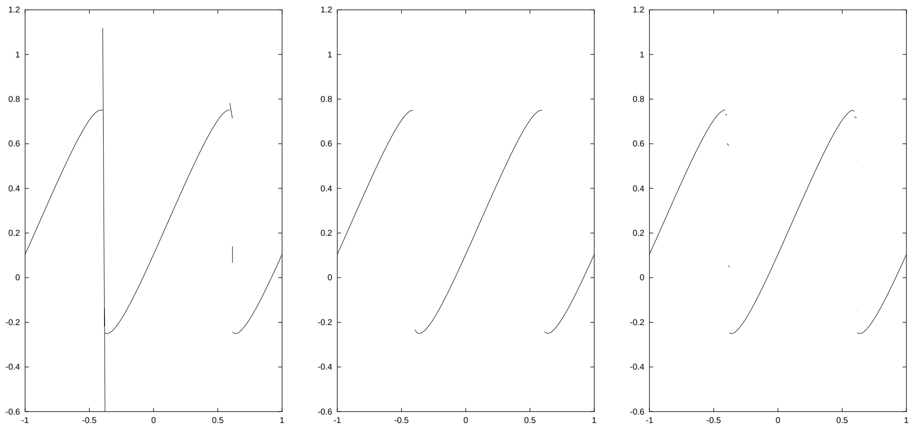


Figure 2.22: Inviscid Burgers equation: graphs along the diagonal, DG FE(left), exact solution (centre), DG FE with limiting (right).

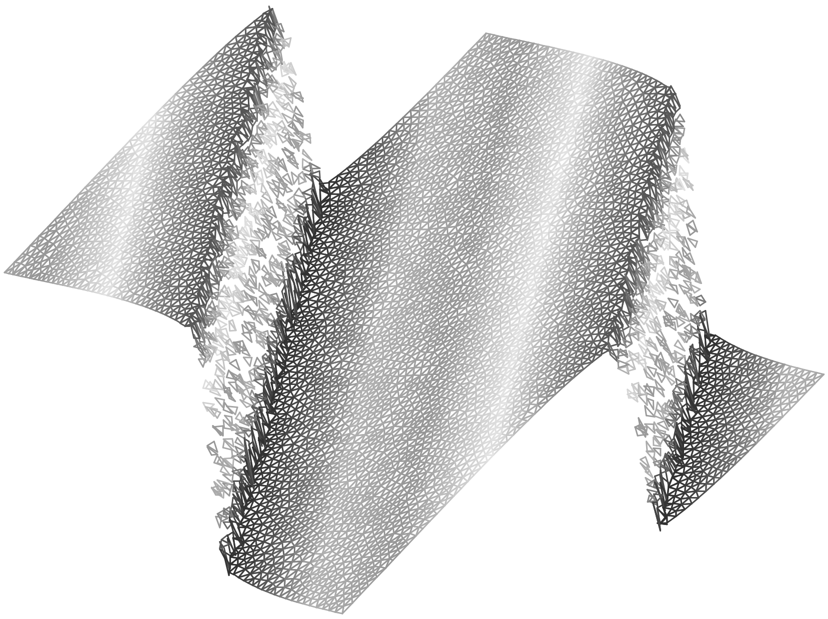


Figure 2.23: Numerical solution of problem (2.5.1) – (2.5.2) computed by DG FE method with limiting, plotted at $t = 0.45$.

2.6 Solution of the Euler and Navier-Stokes equations

2.6.1 Scalar convection-diffusion equation

At first we extend the numerical scheme (2.4.7) from Section 2.4 to a scalar convection-diffusion equation with more general diffusive terms and we also admit Neumann boundary conditions. Then we extent this scheme to the system of equations.

Continuous problem

Let $\Omega \subset I\!R^d$ ($d = 2$ or 3) be a bounded polyhedral domain and $T > 0$. We set $Q_T = \Omega \times (0,T)$ and by $\partial\Omega$ we denote the boundary of Ω which consists of two disjoint parts $\partial\Omega_D$ and $\partial\Omega_N$. We consider the following *initial-boundary value problem*: Find $u : Q_T \to I\!R$ such that

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} = \sum_{s=1}^{d} \frac{\partial R_s(u, \nabla u)}{\partial x_s} + g \quad \text{in } Q_T, \tag{2.6.1}$$

$$u(x,t) = u_D(x) \quad \text{for } x \in \partial\Omega_D, t \in (0,T), \tag{2.6.2}$$

$$\sum_{s=1}^{d} R_s(u, \nabla u) n_s = g_N \quad \text{for } x \in \partial\Omega_N, t \in (0,T), \tag{2.6.3}$$

$$u(x,0) = u^0(x), \quad x \in \Omega. \tag{2.6.4}$$

Here $x = (x_1, \ldots, x_d)$ and t denotes the space and time coordinates and $\boldsymbol{n} = (n_1, \ldots, n_d)$ is a unit outer normal to $\partial\Omega$. We suppose that $f_s \in C^1(I\!R)$, $R_s \in C^1(I\!R^{d+1}), R_s(u, \alpha)$ is linear with respect to $\alpha \in I\!R^d$ and the data are sufficiently regular so that we can assume the existence of a *strong solution* u satisfying (2.6.1) – (2.6.4) pointwise almost everywhere.

Numerical scheme

In order to derive the discrete problem of (2.6.1) – (2.6.4) we use the identical procedure as in Section 2.4. We start from a strong solution u such that $u \in L^2((0,T), H^2(\Omega))$ and $\frac{\partial u}{\partial t} \in L^2((0,T), H^2(\Omega))$. We multiply equation (2.6.1) by any $\varphi \in H^2(\Omega, \mathcal{T}_h)$, integrate over $K \in \mathcal{T}_h$, apply Green's theorem and sum over all $K \in \mathcal{T}_h$. Moreover, we use the relations $\langle \nabla u \rangle_{\Gamma_{ij}} = \nabla u|_{\Gamma_{ij}} = \nabla u|_{\Gamma_{ji}}$ and the identities

$$\int_{\Gamma_{ij}} \sum_{s=1}^{d} \langle R_s(u, \nabla\varphi) \rangle \, (n_{ij})_s \, [u] \, \mathrm{d}S = 0, \; \int_{\Gamma_{ij}} [u][\varphi] \, \mathrm{d}S = 0 \quad \forall j \in s(i) \forall i \in I. \tag{2.6.5}$$

Furthermore, adding some terms which mutually cancel, we find that

$$\begin{aligned}
&\int_{\Omega} \frac{\partial u}{\partial t} \varphi \, \mathrm{d}x + \sum_{i \in I} \sum_{j \in S(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{d} f_s(u) \, (n_{ij})_s \, \varphi|_{\Gamma_{ij}} \, \mathrm{d}S \\
&- \sum_{i \in I} \int_{K_i} \sum_{s=1}^{d} f_s(u) \, \frac{\partial \varphi}{\partial x_s} \, \mathrm{d}x + \sum_{i \in I} \int_{K_i} \sum_{s=1}^{d} R_s(u, \nabla u)) \, \frac{\partial \varphi}{\partial x_s} \, \mathrm{d}x \\
&- \sum_{i \in I} \sum_{\substack{j \in s(i) \\ j < i}} \int_{\Gamma_{ij}} \sum_{s=1}^{d} \langle R_s(u, \nabla u) \rangle \, (n_{ij})_s \, [\varphi] \, \mathrm{d}S
\end{aligned} \tag{2.6.6}$$

$$
\begin{aligned}
& +\sum_{i\in I}\sum_{\substack{j\in s(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{s=1}^{d}\langle R_s(u,\nabla\varphi)\rangle\,(n_{ij})_s\,[u]\,\mathrm{d}S\\
& -\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{d}R_s(u,\nabla u)\,(n_{ij})_s\,\varphi\,\mathrm{d}S\\
& +\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{d}R_s(u,\nabla\varphi)\,(n_{ij})_s\,(u-u_D)\,\mathrm{d}S\\
& +\sum_{i\in I}\sum_{j\in s(i)}\int_{\Gamma_{ij}}\sigma[u]\,[\varphi]\,\mathrm{d}S+\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sigma\,(u-u_D)\,\varphi\,\mathrm{d}S\\
= & \int_{\Omega}g\,\varphi\,\mathrm{d}x+\sum_{i\in I}\sum_{j\in\gamma_N(i)}\int_{\Gamma_{ij}}g_N\,\varphi\,\mathrm{d}S.
\end{aligned}
$$

Here σ is a weight function defined on each Γ_{ij} by $\sigma|_{\Gamma_{ij}}=\varepsilon/\mathrm{d}(\Gamma_{ij})$, where $\varepsilon\approx R_s$.

Similarly as in the FV method the inviscid fluxes will be approximated with the aid of the so-called numerical flux $H(u,u',\boldsymbol{n})$:

$$
\int_{\Gamma_{ij}}\sum_{s=1}^{d}f_s(u_h)\,n_s\,\varphi_h\,\mathrm{d}S\approx\int_{\Gamma_{ij}}H(u_h|_{\Gamma_{ij}},u_h|_{\Gamma_{ji}},\boldsymbol{n}_{ij})\,\varphi_h\,\mathrm{d}S,\ i\in I,\ j\in S(i). \tag{2.6.7}
$$

Of course, if $j\in\gamma(i)$ and $\Gamma_{ij}\subset\partial\Omega$, it is necessary to specify the meaning of $u|_{\Gamma_{ji}}$. Here we use the extrapolation, i.e. we set $u|_{\Gamma_{ji}}:=u|_{\Gamma_{ij}}$. In this way we obtain the convection form $b_h(u,v)$:

$$
\begin{aligned}
b_h(u_h,\varphi_h) = & \sum_{i\in I}\Bigg(\sum_{j\in s(i)}\int_{\Gamma_{ij}}H(u_h|_{\Gamma_{ij}},u_h|_{\Gamma_{ji}},\boldsymbol{n}_{ij})\,\varphi_h|_{\Gamma_{ij}}\,\mathrm{d}S\\
& + \sum_{j\in\gamma(i)}\int_{\Gamma_{ij}}H(u_h|_{\Gamma_{ij}},u_h|_{\Gamma_{ij}},\boldsymbol{n}_{ij})\,\varphi_h|_{\Gamma_{ij}}\,\mathrm{d}S\Bigg)\\
& -\sum_{i\in I}\int_{K_i}\sum_{s=1}^{d}f_s(u_h)\,\frac{\partial\varphi_h}{\partial x_s}\,\mathrm{d}x,\quad u_h,\,\varphi_h\in H^2(\Omega,\mathcal{T}_h).
\end{aligned} \tag{2.6.8}
$$

Now, for $u_h,\varphi_h\in H^2(\Omega,\mathcal{T}_h)$ we set

$$
\begin{aligned}
a_h(u_h,\varphi_h) = & \sum_{i\in I}\int_{K_i}\sum_{s=1}^{d}R_s(u_h,\nabla u_h))\,\frac{\partial\varphi_h}{\partial x_s}\,\mathrm{d}x\\
& -\sum_{i\in I}\sum_{\substack{j\in S(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{s=1}^{d}\langle R_s(u_h,\nabla u_h)\rangle\,(n_{ij})_s\,[\varphi_h]\,\mathrm{d}S\\
& +\sum_{i\in I}\sum_{\substack{j\in S(i)\\ j<i}}\int_{\Gamma_{ij}}\sum_{s=1}^{d}\langle R_s(u_h,\nabla\varphi_h)\rangle\,(n_{ij})_s\,[u_h]\,\mathrm{d}S\\
& -\sum_{i\in I}\sum_{j\in\gamma_D(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{d}R_s(u_h,\nabla u_h)\,(n_{ij})_s\,\varphi_h\,\mathrm{d}S
\end{aligned} \tag{2.6.9}
$$

$$
\begin{aligned}
& + \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{d} R_s(u_h, \nabla\varphi_h)\,(n_{ij})_s\, u_h \,\mathrm{d}S \\
J_h(u_h,\varphi_h) &= \sum_{i\in I} \sum_{j\in s(i)} \int_{\Gamma_{ij}} \sigma[u_h]\,[\varphi_h]\,\mathrm{d}S + \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \sigma\, u_h\, \varphi_h \,\mathrm{d}S \qquad (2.6.10) \\
\ell_h(\varphi_h)\,(t) &= \int_\Omega g(t)\,\varphi_h \,\mathrm{d}x + \sum_{i\in I} \sum_{j\in\gamma_N(i)} \int_{\Gamma_{ij}} g_N\, v \,\mathrm{d}S \qquad (2.6.11) \\
& + \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{d} R_s(u, \nabla\varphi_h)\,(n_{ij})_s\, u_D(t) \,\mathrm{d}S \\
& + \sum_{i\in I} \sum_{j\in\gamma_D(i)} \int_{\Gamma_{ij}} \sigma\, u_D(t)\, \varphi_h \,\mathrm{d}S, \\
(\alpha,\beta) &= \int_\Omega \alpha\beta \,\mathrm{d}x. \qquad (2.6.12)
\end{aligned}
$$

By $u(t)$ we denote the function on Ω such that $u(t)\,(x) = u(x,t)$, $x \in \Omega$. As $S_h \subset H^2(\Omega, \mathcal{T}_h)$ the previous forms have sense as well as for $u_h, \varphi_h \subset S_h$. Then we formulate the *discrete problem*: We define an approximate solution as a function u_h satisfying the conditions

$$
\text{a) } u_h \in C^1([0,T], S_h), \qquad (2.6.13)
$$

$$
\text{b) } \left(\frac{\partial u_h(t)}{\partial t}, \varphi_h\right) + b_h(u_h(t),\varphi_h) + a_h(u_h(t),\varphi_h) + J_h(u_h(t),\varphi_h) = \ell_h(\varphi_h)(t) \quad \forall\, \varphi_h \in S_h \quad \forall\, t \in (0,T),
$$

$$
\text{c) } u_h(0) = u_h^0,
$$

where u_h^0 is an S_h-approximation of u^0 (e. g., L^2-projection).

2.6.2 System of the Navier-Stokes equations

Our aim is to extend the approach explained above to the system of equations, particularly to the compressible Navier-Stokes equations. For simplicity, we shall consider a 2-dimensional problem, i.e., we assume that $d = 2$.

Governing equations

Let $\Omega \subset I\!R^2$ be a bounded domain and $T > 0$. We set $Q_T = \Omega \times (0,T)$ and by $\partial\Omega$ denote the boundary of Ω which consists of several disjoint parts. We distinguish inlet Γ_I, outlet Γ_O and impermeable walls Γ_W on $\partial\Omega$. The system of the compressible Navier-Stokes equations describing 2D viscous flow can be written in the dimensionless form

$$
\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{s=1}^{2} \frac{\partial \boldsymbol{f}_s(\boldsymbol{w})}{\partial x_s} = \frac{\partial \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w})}{\partial x_s} \quad \text{in } Q_T. \qquad (2.6.14)
$$

where

$$
\boldsymbol{w} = (w_1, \dots, w_4)^{\mathrm{T}} = (\rho,\, \rho v_1,\, \rho v_2,\, e)^{\mathrm{T}} \qquad (2.6.15)
$$

is the so-called *state vector*,

$$\begin{aligned} \boldsymbol{f}_s(\boldsymbol{w}) &= (f_s^1(\boldsymbol{w}),\dots,f_s^4(\boldsymbol{w}))^{\mathrm{T}} \\ &= (\rho v_s,\, \rho v_s v_1 + \delta_{s1} p,\, \rho v_s v_2 + \delta_{s2} p,\, (e+p)\, v_s)^{\mathrm{T}},\ \ s = 1,2, \end{aligned} \tag{2.6.16}$$

are the so-called *inviscid (Euler) fluxes* and

$$\begin{aligned} \boldsymbol{R}_s(\boldsymbol{w},\nabla\boldsymbol{w}) &= (R_s^1(\boldsymbol{w},\nabla\boldsymbol{w}),\dots,R_s^4(\boldsymbol{w},\nabla\boldsymbol{w}))^{\mathrm{T}} \\ &= \left(0,\, \tau_{s1}, \tau_{s2}, \sum_{k=1}^{2} \tau_{sk} v_k + \frac{\gamma}{Re\, Pr}\frac{\partial\theta}{\partial x_s}\right)^{\mathrm{T}},\ \ s = 1,2, \end{aligned} \tag{2.6.17}$$

are the so-called *viscous fluxes*. We consider the Newtonian type of fluid, i. e., the viscous part of stress tensor has the form

$$\tau_{sk} = \frac{1}{Re}\left[\left(\frac{\partial v_s}{\partial x_k} + \frac{\partial v_k}{\partial x_s}\right) - \frac{2}{3} \div(\boldsymbol{v})\, \delta_{sk}\right], \quad s,k = 1,2. \tag{2.6.18}$$

We use the following notation: ρ – density, p – pressure, e – total energy, $\boldsymbol{v} = (v_1, v_2)$ – velocity, θ – absolute temperature, γ – Poisson adiabatic constant, Re – Reynolds number, Pr – Prandtl number.

In order to close the system we consider the following thermodynamical relations: the state equation for perfect gas and the relation for total energy

$$p = (\gamma - 1)\,(e - \rho|\boldsymbol{v}|^2/2), \quad e = c_V \rho\theta + \rho|\boldsymbol{v}|^2/2, \tag{2.6.19}$$

where c_V is the specific heat at constant volume which is equal to one in the dimensionless case. The system (2.6.14) – (2.6.19) is *hyperbolic-parabolic* type. It is equipped with the initial condition

$$\boldsymbol{w}(x,0) = \boldsymbol{w}^0(x), \quad x \in \Omega, \tag{2.6.20}$$

and the following set of boundary conditions on appropriate parts of boundary:

$$\begin{aligned} &a) \quad \rho = \rho_D,\ \boldsymbol{v} = \boldsymbol{v}_D,\ \sum_{k=1}^{2}\left(\sum_{l=1}^{2} \tau_{lk} n_l\right) v_k + \frac{\gamma}{Re\, Pr}\frac{\partial\theta}{\partial\boldsymbol{n}} = 0 \quad \text{on } \Gamma_I, \\ &b) \quad \sum_{k=1}^{2} \tau_{sk} n_k = 0,\ s = 1,2,\ \frac{\partial\theta}{\partial\boldsymbol{n}} = 0 \quad \text{on } \Gamma_O, \\ &c) \quad \boldsymbol{v} = 0,\ \frac{\partial\theta}{\partial\boldsymbol{n}} = 0 \quad \text{on } \Gamma_W, \end{aligned} \tag{2.6.21}$$

where ρ_D and $\boldsymbol{v}_D$ are given function and $\boldsymbol{n} = (n_1, n_2)$ is a unit outer normal to $\partial\Omega$. Another possibility is to replace the adiabatic boundary condition (2.6.21), c) by the

$$c')\ \boldsymbol{v} = 0,\ \theta = \theta_D \quad \text{on } \Gamma_W. \tag{2.6.22}$$

The question of a correct choice of the boundary conditions for the compressible Navier-Stokes equations is still open. Our choice is based on some heuristic considerations and various types of numerical experiments.

Putting $\boldsymbol{R}_s = 0,\ s = 1, 2$, in (2.6.14), we obtain the *Euler equations*

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{s=1}^{2} \frac{\partial \boldsymbol{f}_s(\boldsymbol{w})}{\partial x_s} = 0 \quad \text{in } Q_T, \tag{2.6.23}$$

which describe the motion of an inviscid compressible fluid. The system of the Euler equations is *hyperbolic*. For this case we replace the boundary conditions (2.6.21)–(2.6.22) by "inviscid" ones. We put $\boldsymbol{v} \cdot \boldsymbol{n} = 0$ on Γ_W and on Γ_I and Γ_O we consider the boundary conditions, which were derived for a linear system. It means that we prescribe m_n components of $\boldsymbol{w}$, and the other components are extrapolated from interior of Ω. Here m_n is the number of negative eigenvalues of the Jacobi matrix

$$\sum_{s=1}^{2} \frac{D\boldsymbol{f}_s(\boldsymbol{w})}{D\boldsymbol{w}} n_s. \tag{2.6.24}$$

For more detail see [Fei93], [Fei01], [Tor97].

DG FE formulation

The extension of DG FE method from the scalar equation to the system of the Navier-Stokes equations is not quite straightforward. It is caused by the fact that the viscous terms $\boldsymbol{R}_s,\ s = 1, 2$ in (2.6.17) are more complicated than in the scalar case. In order to proceed in a similar way as in Section 2.6.1, we have to modify terms $\boldsymbol{R}_s,\ s = 1, 2$ so that they are linear with respect to $\nabla \boldsymbol{w}$.

Replacing the physical quantities ρ, $\boldsymbol{v}$ and θ by the components of the state vector $\boldsymbol{w}$ we have from (2.6.15) and (2.6.17)

$$\boldsymbol{R}_1(\boldsymbol{w}, \nabla \boldsymbol{w}) = \begin{pmatrix} 0 \\ \frac{2}{3}\frac{1}{Re}\left(2\frac{\partial}{\partial x_1}(\frac{w_2}{w_1}) - \frac{\partial}{\partial x_2}(\frac{w_3}{w_1})\right) \\ \frac{1}{Re}\left(\frac{\partial}{\partial x_1}(\frac{w_3}{w_1}) + \frac{\partial}{\partial x_2}(\frac{w_2}{w_1})\right) \\ \frac{w_2}{w_1}\boldsymbol{R}_1^{(2)} + \frac{w_3}{w_1}\boldsymbol{R}_1^{(3)} + \frac{\gamma}{Re\,Pr}\frac{\partial}{\partial x_1}\left(\frac{w_4}{w_1} - \frac{1}{2w_1^2}(w_2^2 + w_3^2)\right) \end{pmatrix}, \tag{2.6.25}$$

where $\boldsymbol{R}_1^{(2)}$ and $\boldsymbol{R}_1^{(3)}$ are the second and third component of $\boldsymbol{R}_1(\boldsymbol{w}, \nabla \boldsymbol{w})$, respectively, and

$$\boldsymbol{R}_2(\boldsymbol{w}, \nabla \boldsymbol{w}) = \begin{pmatrix} 0 \\ \frac{1}{Re}\left(\frac{\partial}{\partial x_1}(\frac{w_3}{w_1}) + \frac{\partial}{\partial x_2}(\frac{w_2}{w_1})\right) \\ \frac{2}{3}\frac{1}{Re}\left(2\frac{\partial}{\partial x_2}(\frac{w_3}{w_1}) - \frac{\partial}{\partial x_1}(\frac{w_2}{w_1})\right) \\ \frac{w_2}{w_1}\boldsymbol{R}_2^{(2)} + \frac{w_3}{w_1}\boldsymbol{R}_2^{(3)} + \frac{\gamma}{Re\,Pr}\frac{\partial}{\partial x_2}\left(\frac{w_4}{w_1} - \frac{1}{2w_1^2}(w_2^2 + w_3^3)\right) \end{pmatrix}, \tag{2.6.26}$$

where $\boldsymbol{R}_2^{(2)}$ and $\boldsymbol{R}_2^{(3)}$ are the second and third component of $\boldsymbol{R}_2(\boldsymbol{w}, \nabla \boldsymbol{w})$, respectively. Performing the chain rule in the terms $\boldsymbol{R}_s,\ s = 1, 2$ we have

$$\boldsymbol{R}_1(\boldsymbol{w}, \nabla \boldsymbol{w}) = \begin{pmatrix} 0 \\ \frac{2}{3}\frac{1}{Re w_1}\left[2\left(\frac{\partial w_2}{\partial x_1} - \frac{w_2}{w_1}\frac{\partial w_1}{\partial x_1}\right) - \left(\frac{\partial w_3}{\partial x_2} - \frac{w_3}{w_1}\frac{\partial w_1}{\partial x_2}\right)\right] \\ \frac{1}{Re w_1}\left[\left(\frac{\partial w_3}{\partial x_1} - \frac{w_3}{w_1}\frac{\partial w_1}{\partial x_1}\right) + \left(\frac{\partial w_2}{\partial x_2} - \frac{w_2}{w_1}\frac{\partial w_1}{\partial x_2}\right)\right] \\ \frac{w_2}{w_1}\boldsymbol{R}_1^{(2)} + \frac{w_3}{w_1}\boldsymbol{R}_1^{(3)} + \frac{\gamma}{Re\,Pr}\frac{1}{w_1}\times \\ \times\left[\frac{\partial w_4}{\partial x_1} - \frac{w_4}{w_1}\frac{\partial w_1}{\partial x_1} - \frac{1}{w_1}\left(w_2\frac{\partial w_2}{\partial x_1} + w_3\frac{\partial w_3}{\partial x_1}\right) + \frac{1}{w_1^2}(w_2^2 + w_3^2)\frac{\partial w_1}{\partial x_1}\right] \end{pmatrix}, \tag{2.6.27}$$

and

$$
\boldsymbol{R}_2(\boldsymbol{w},\nabla\boldsymbol{w}) = \begin{pmatrix} 0 \\ \frac{1}{Re w_1}\left[\left(\frac{\partial w_3}{\partial x_1} - \frac{w_3}{w_1}\frac{\partial w_1}{\partial x_1}\right) + \left(\frac{\partial w_2}{\partial x_2} - \frac{w_2}{w_1}\frac{\partial w_1}{\partial x_2}\right)\right] \\ \frac{2}{3}\frac{1}{Re\, w_1}\left[2\left(\frac{\partial w_3}{\partial x_2} - \frac{w_3}{w_1}\frac{\partial w_1}{\partial x_2}\right) - \left(\frac{\partial w_2}{\partial x_1} - \frac{w_2}{w_1}\frac{\partial w_1}{\partial x_1}\right)\right] \\ \frac{w_2}{w_1}\boldsymbol{R}_2^{(2)} + \frac{w_3}{w_1}\boldsymbol{R}_2^{(3)} + \frac{\gamma}{Re\, Pr}\frac{1}{w_1}\times \\ \times\left[\frac{\partial w_4}{\partial x_2} - \frac{w_4}{w_1}\frac{\partial w_1}{\partial x_2} - \frac{1}{w_1}\left(w_2\frac{\partial w_2}{\partial x_2} + w_3\frac{\partial w_3}{\partial x_2}\right) + \frac{1}{w_1^2}(w_2^2+w_3^2)\frac{\partial w_1}{\partial x_2}\right] \end{pmatrix}. \qquad (2.6.28)
$$

For $\varphi = (\varphi_1,\varphi_2,\varphi_3,\varphi_4)$ we define the forms

$$
\boldsymbol{D}_1(\boldsymbol{w},\nabla\boldsymbol{w},\varphi,\nabla\varphi) \equiv \qquad (2.6.29)
$$

$$
\begin{pmatrix} 0 \\ \frac{2}{3}\frac{1}{Re w_1}\left[2\left(\frac{\partial\varphi_2}{\partial x_1} - \frac{\varphi_2}{w_1}\frac{\partial w_1}{\partial x_1}\right) - \left(\frac{\partial\varphi_3}{\partial x_2} - \frac{\varphi_3}{w_1}\frac{\partial w_1}{\partial x_2}\right)\right] \\ \frac{1}{Re w_1}\left[\left(\frac{\partial\varphi_3}{\partial x_1} - \frac{\varphi_3}{w_1}\frac{\partial w_1}{\partial x_1}\right) + \left(\frac{\partial\varphi_2}{\partial x_2} - \frac{\varphi_2}{w_1}\frac{\partial w_1}{\partial x_2}\right)\right] \\ \frac{w_2}{w_1}\boldsymbol{D}_1^{(2)} + \frac{w_3}{w_1}\boldsymbol{D}_1^{(3)} + \frac{\gamma}{Re\, Pr}\frac{1}{w_1}\times \\ \times\left[\frac{\partial\varphi_4}{\partial x_1} - \frac{\varphi_4}{w_1}\frac{\partial w_1}{\partial x_1} - \frac{1}{w_1}\left(w_2\frac{\partial\varphi_2}{\partial x_1} + w_3\frac{\partial\varphi_3}{\partial x_1}\right) + \frac{1}{w_1^2}(w_2\varphi_2+w_3\varphi_3)\frac{\partial w_1}{\partial x_1})\right] \end{pmatrix},
$$

and

$$
\boldsymbol{D}_2(\boldsymbol{w},\nabla\boldsymbol{w},\varphi,\nabla\varphi) \equiv \qquad (2.6.30)
$$

$$
\begin{pmatrix} 0 \\ \frac{1}{Re w_1}\left[\left(\frac{\partial\varphi_3}{\partial x_1} - \frac{\varphi_3}{w_1}\frac{\partial w_1}{\partial x_1}\right) + \left(\frac{\partial\varphi_2}{\partial x_2} - \frac{\varphi_2}{w_1}\frac{\partial w_1}{\partial x_2}\right)\right] \\ \frac{2}{3}\frac{1}{Re\, w_1}\left[2\left(\frac{\partial\varphi_3}{\partial x_2} - \frac{\varphi_3}{w_1}\frac{\partial w_1}{\partial x_2}\right) - \left(\frac{\partial\varphi_2}{\partial x_1} - \frac{\varphi_2}{w_1}\frac{\partial w_1}{\partial x_1}\right)\right] \\ \frac{w_2}{w_1}\boldsymbol{D}_2^{(2)} + \frac{w_3}{w_1}\boldsymbol{D}_2^{(3)} + \frac{\gamma}{Re\, Pr}\frac{1}{w_1}\times \\ \times\left[\frac{\partial\varphi_4}{\partial x_2} - \frac{\varphi_4}{w_1}\frac{\partial w_1}{\partial x_2} - \frac{1}{w_1}\left(w_2\frac{\partial\varphi_2}{\partial x_2} + w_3\frac{\partial\varphi_3}{\partial x_2}\right) + \frac{1}{w_1^2}(w_2\varphi_2+w_3\varphi_2)\frac{\partial w_1}{\partial x_2}\right] \end{pmatrix}.
$$

where $\boldsymbol{D}_s^{(2)},\ \boldsymbol{D}_s^{(3)},\ s=1,2$ are the second and third components of $\boldsymbol{D}_s(\boldsymbol{w},\nabla\boldsymbol{w})$, respectively. Obviously

$$
\boldsymbol{D}_s(\boldsymbol{w},\nabla\boldsymbol{w},\boldsymbol{w},\nabla\boldsymbol{w}) = \boldsymbol{R}_s(\boldsymbol{w},\nabla\boldsymbol{w}), \qquad s=1,2 \qquad (2.6.31)
$$

and the forms $\boldsymbol{D}_s(\boldsymbol{w},\nabla\boldsymbol{w},\varphi,\nabla\varphi),\ s=1,2$ are linear with respect φ and $\nabla\varphi$ and therefore they satisfy the assumption of linearity as the forms $R_s(u,\alpha),\ s=1,2$ from the scalar case. There are several possibilities how to define the forms $\boldsymbol{D}_s,\ s=1,2$. Based on a wide range of numerical experiments, the presented possibility (2.6.29)–(2.6.30) seems to us the best one. On the other hand, in the absence of a theoretical background, this question is still open.

In order to derive the weak formulation, we multiply (2.6.14) (where we replace the terms $\boldsymbol{R}_s$ by the terms $\boldsymbol{D}_s,\ s=1,2$) by a function $\varphi\in\left[H^k(\Omega,\mathcal{T}_h)\right]^4$ and integrate over an element $K_i,\ i\in I$. Further we sum over all K_i and add similar stabilization terms as for the scalar equation. Taking into account the boundary condition (2.6.21), we obtain the integral identity (compare with (2.6.6))

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} \sum_{K_i \in \mathcal{T}_h} \int_{K_i} \boldsymbol{w} \cdot \varphi \,\mathrm{d}x = \\
&\quad + \sum_{K_i \in \mathcal{T}_h} \int_{K_i} \sum_{s=1}^{2} \boldsymbol{f}_s(\boldsymbol{w}) \cdot \frac{\partial \varphi}{\partial x_s} \,\mathrm{d}x \\
&\quad - \sum_{K_i \in \mathcal{T}_h} \sum_{j \in S(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \boldsymbol{f}_s(\boldsymbol{w}) \cdot \varphi \,(n_{ij})_s \,\mathrm{d}S \\
&\quad - \sum_{K_i \in \mathcal{T}_h} \int_{K_i} \sum_{s=1}^{2} \boldsymbol{D}_s(\boldsymbol{w}, \nabla \boldsymbol{w}, \boldsymbol{w}, \nabla \boldsymbol{w}) \cdot \frac{\partial \varphi}{\partial x_s} \,\mathrm{d}x \\
&\quad + \sum_{K_i \in \mathcal{T}_h} \sum_{\substack{j \in s(i) \\ j<i}} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \langle \boldsymbol{D}_s(\boldsymbol{w}, \nabla \boldsymbol{w}, \boldsymbol{w}, \nabla \boldsymbol{w}) \rangle \,(n_{ij})_s \cdot [\varphi] \,\mathrm{d}S \\
&\quad - \sum_{K_i \in \mathcal{T}_h} \sum_{\substack{j \in s(i) \\ j<i}} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \langle \boldsymbol{D}_s(\boldsymbol{w}, \nabla \boldsymbol{w}, \varphi, \nabla \varphi) \rangle \,(n_{ij})_s \cdot [\boldsymbol{w}] \,\mathrm{d}S \\
&\quad + \sum_{K_i \in \mathcal{T}_h} \sum_{j \in \gamma_D(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \boldsymbol{D}_s(\boldsymbol{w}, \nabla \boldsymbol{w}, \boldsymbol{w}, \nabla \boldsymbol{w}) \,(n_{ij})_s \cdot \varphi \,\mathrm{d}S \\
&\quad - \sum_{K_i \in \mathcal{T}_h} \sum_{j \in \gamma_D(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{2} \boldsymbol{D}_s(\boldsymbol{w}, \nabla \boldsymbol{w}, \varphi, \nabla \varphi) \,(n_{ij})_s \cdot (\boldsymbol{w} - \boldsymbol{w}_B) \,\mathrm{d}S \\
&\quad - \sum_{K_i \in \mathcal{T}_h} \sum_{\substack{j \in s(i) \\ j<i}} \int_{\Gamma_{ij}} \sigma [\boldsymbol{w}] \cdot [\varphi] \,\mathrm{d}S \\
&\quad - \sum_{K_i \in \mathcal{T}_h} \sum_{j \in \gamma_D(i)} \int_{\Gamma_{ij}} \sigma (\boldsymbol{w} - \boldsymbol{w}_B) \cdot \varphi \,\mathrm{d}S
\end{aligned}
\tag{2.6.32}
$$

where $\sigma = (|\Gamma_{ij}|\, Re)^{-1}$. The set $\gamma_D(i)$ contains edges where the Dirichlet boundary condition is given for at least one component of $\boldsymbol{w} = (w_1, \dots, w_4)$, i.e., edges lying on Γ_I or Γ_W. We prescribe the boundary state $\boldsymbol{w}_B = ((w_B)_1, \dots, (w_B)_4)$ by

$$
(w_B)_r = \begin{cases} (w_D)_r & \text{if } w_r \text{ is prescribed on } \partial\Omega \\ w_r & \text{otherwise,} \end{cases} \quad , \ r = 1, \dots, 4. \tag{2.6.33}
$$

In particular, for the case (2.6.21) a)–c) we have

$$
\begin{aligned}
\boldsymbol{w}_B &= (\rho|_{\Gamma_W}, 0, 0, \rho|_{\Gamma_W} \theta|_{\Gamma_W}) \text{ on } \Gamma_W, \\
\boldsymbol{w}_B &= (\rho_D, \rho_D (\boldsymbol{v}_D)_1, \rho_D (\boldsymbol{v}_D)_2, \rho|_{\Gamma_I} \theta|_{\Gamma_I} + \frac{1}{2} \rho_D |\boldsymbol{v}_D|^2) \text{ on } \Gamma_I,
\end{aligned}
\tag{2.6.34}
$$

and for the case (2.6.21) a)–b), (2.6.22) c')

$$
\begin{aligned}
\boldsymbol{w}_B &= (\rho|_{\Gamma_W}, 0, 0, \rho|_{\Gamma_W} \theta_D) \text{ on } \Gamma_W, \\
\boldsymbol{w}_B &= (\rho_D, \rho_D (\boldsymbol{v}_D)_1, \rho_D (\boldsymbol{v}_D)_2, \rho|_{\Gamma_I} \theta|_{\Gamma_I} + \frac{1}{2} \rho_D |\boldsymbol{v}_D|^2) \text{ on } \Gamma_I,
\end{aligned}
\tag{2.6.35}
$$

where $\rho_D, \boldsymbol{v}_D$ and θ_D are given functions from the boundary conditions (2.6.21)–(2.6.22) and $\rho|_\Gamma$ and $\theta|_\Gamma$ are the values of density and temperature extrapolated from in interior of Ω on the appropriate boundary part Γ, respectively.

The relation (2.6.32) represents a weak form of the Navier-Stokes equations in the sense of the broken Sobolev space $[H^1(\Omega, \mathcal{T}_h)]^4$.

Numerical solution

In order to proceed from (2.6.32) to the numerical scheme for the system of equations, we use the same strategy as for the scalar equation in Section 2.6.1. By $\boldsymbol{w}(t)$ we denote the function on Ω such that $\boldsymbol{w}(t)\,(x) = \boldsymbol{w}(x,t)$, $x \in \Omega$. The relation (2.6.32) has sense for $\boldsymbol{w}(t), \varphi \in [S_h]^4$, $t \in (0,T)$. In the following we present the schemes for the numerical solutions of inviscid or viscous flow described by systems (2.6.23) or (2.6.14), respectively.

Inviscid case

The second term on the right–hand side of (2.6.32) (boundary integral for inviscid fluxes) is approximated by the approach common in the finite volume method. We put

$$\int_{\Gamma_{ij}} \sum_{s=1}^{2} \boldsymbol{f}_s(\boldsymbol{w}(t))\, n_s \varphi \,\mathrm{d}S \approx \int_{\Gamma_{ij}} \boldsymbol{H}(\boldsymbol{w}|_{\Gamma_{ij}}(x,t), \boldsymbol{w}|_{\Gamma_{ji}}(x,t), \boldsymbol{n}_{ij(x)})\varphi|_{\Gamma_{ij}}(x)\mathrm{d}S, \qquad (2.6.36)$$

where $\boldsymbol{H}$ is *the numerical flux*, $\boldsymbol{w}|_{\Gamma_{ij}}(x,\cdot)$ and $\boldsymbol{w}|_{\Gamma_{ji}}(x,\cdot)$ are the values of $\boldsymbol{w}$ at $x \in \Gamma_{ij}$ considered from interior and exterior of K_i, respectively. We use the numerical flux based on the direct solution of the local Riemann problem, see [DF02b, Tor97].

For $\boldsymbol{w}_h(t), \varphi_h \in [S_h]^4$ we introduce the forms

$$\begin{aligned}
(\boldsymbol{w}_h, \varphi_h) &= \int_\Omega \boldsymbol{w}_h(x,t) \cdot \varphi_h(x)\,\mathrm{d}x, \qquad (2.6.37)\\
\boldsymbol{B}_h(\boldsymbol{w}_h, \varphi_h) &= -\sum_{K_i \in \mathcal{T}_h} \int_{K_i} \sum_{s=1}^{2} \boldsymbol{f}_s(\boldsymbol{w}_h(x,t)) \cdot \frac{\partial \varphi_h(x)}{\partial x_s}\,\mathrm{d}x \\
&\quad + \sum_{K_i \in \mathcal{T}_h} \sum_{j \in S(i)} \int_{\Gamma_{ij}} \boldsymbol{H}(\boldsymbol{w}_h|_{\Gamma_{ij}}(x,t), \boldsymbol{w}_h|_{\Gamma_{ji}}(x,t), \boldsymbol{n}_{ij}(x)) \cdot \varphi_h(x)\mathrm{d}S.
\end{aligned}$$

It is necessary define the meaning of $\boldsymbol{w}_h|_{\Gamma_{ji}}$ for $j \in \gamma(i)$. We use the "inviscid" boundary condition mentioned in Section 2.6.2. We say that $\boldsymbol{w}_h$ is the *approximate solution* of the system of the Euler equations (2.6.23), if it satisfies the conditions

$$\begin{aligned}
&\text{a)} \quad \boldsymbol{w}_h \in C^1([0,T],[S_h]^4), \qquad (2.6.38)\\
&\text{b)} \quad \frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{w}_h(t), \varphi_h) + \boldsymbol{B}_h(\boldsymbol{w}_h(t), \varphi_h) = 0 \quad \forall\, \varphi_h \in [S_h]^4\ \forall\, t \in (0,T),\\
&\text{c)} \quad \boldsymbol{w}_h(0) = \boldsymbol{w}_h^0.
\end{aligned}$$

Viscous case

Let $\boldsymbol{B}_h$ be the form introduced in (2.6.37). We denote by $\boldsymbol{A}_h(\boldsymbol{w}, \varphi)$ the diffusive and stabilization part of the right hand side of (2.6.32), i.e., all terms where appears the terms $\boldsymbol{D}_s$

or σ. The form $\boldsymbol{A}_h(\cdot,\cdot)$ has sense for $\boldsymbol{w}_h(t), \varphi_h \in [S_h]^4$. We say that $\boldsymbol{w}_h$ is an *approximate solution* of (2.6.32), if it satisfies the conditions

$$
\begin{aligned}
&\text{a)} \quad \boldsymbol{w}_h \in C^1([0,T],[S_h]^4), \\
&\text{b)} \quad \frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{w}_h(t),\varphi_h) = \boldsymbol{A}_h(\boldsymbol{w}_h(t),\varphi_h) - \boldsymbol{B}_h(\boldsymbol{w}_h(t),\varphi_h), \\
&\qquad\qquad \forall\, \varphi_h \in [S_h]^4 \ \forall\, t \in (0,T), \\
&\text{c)} \quad \boldsymbol{w}_h(0) = \boldsymbol{w}_h^0.
\end{aligned}
\tag{2.6.39}
$$

The relations (2.6.38) and (2.6.39) are the basic numerical schemes for the solution of the Euler and Navier-Stokes equations, respectively, with the aid of DG FE method. They represent a system of ordinary differential equations which can be solved by a suitable numerical method.

Let us remark that the scheme (2.6.38) does not contain any stabilization terms. Moreover, (2.6.38) is a special case of (2.6.39), with $\boldsymbol{A}_h = 0$.

Implementation

The computation of the approximate solution was realized over regular triangular grids with the aid of piecewise linear approximations, i. e., with $p = 1$ in (2.3.1). We use the basis of $S_h := S^{1,-1}$ whose components are piecewise linear functions associated with midpoints of sides of triangles:

$$\{\phi_{ij} \in S_h;\, \phi_{ij}(Q_{kl}) = \delta_{ik}\delta_{jl},\ j \in S(i),\, l \in S(k),\, i,k \in I\}, \tag{2.6.40}$$

where Q_{kl} is the midpoint of Γ_{kl}. This choice leads to a diagonal mass matrix in (2.6.38), b) and (2.6.39), b). For the evaluation of boundary integrals in (2.6.32) we use the two point Gauss quadrature rule

$$\int_{-1}^{1} g(t)\,\mathrm{d}t \approx g\left(-\frac{1}{\sqrt{3}}\right) + g\left(\frac{1}{\sqrt{3}}\right), \tag{2.6.41}$$

which is exact for polynomials of the third degree. The volume integrals in (2.6.32) are evaluated by the three point integration rule

$$\int_{K_i} g(x)\,\mathrm{d}x \approx |K_i|\frac{1}{3}\sum_{j\in S(i)} g(Q_{ij}), \tag{2.6.42}$$

which is exact for second degree polynomials. The systems of ordinary differential equations (2.6.38) and (2.6.39) are solved by the forward Euler method. The use of the explicit scheme requires a restriction on a time step. Based on heuristic consideration and a number of numerical experiments we use the stability condition in the form

$$\Delta t = \frac{1}{3} CFL \min\left(\frac{1}{2}\max_{\substack{K_i\in T_h\\ j\in S(i)}}\left(\frac{|K_i|}{\lambda_{ij}|\Gamma_{ij}|}\right), \frac{1}{8}\max_{K_i\in T_h}\frac{Re}{|K_i|}\right), \tag{2.6.43}$$

where we put usually $CFL = 0.4$ and λ_{ij} is the maximal eigenvalue of the Jacobi matrix (2.6.24) evaluated over Γ_{ij}.

2.6.3 Limiting

The solution of problems with zero or small viscosity (high Reynolds number) contains discontinuities or steep gradients of computed quantities, respectively. Their numerical approximations obtained by (2.6.38) and/or (2.6.39) usually contain spurious oscillations. In order to avoid these unphysical phenomena, we developed an indicator which seeks elements $K \in \mathcal{T}_h$ on which the approximate solution has large interelement jumps, representing discontinuities. On these elements the order limiting is applied, switching from a piecewise linear approximation to a piecewise constant approximation.

We extend the control of jumps from Section 2.5 also to the system of the Euler and Navier-Stokes equations. In order to detect discontinuities of the solutions (shock waves, contact discontinuities), it is suitable to measure the jumps of the density. Based on results from Section 2.5.2 and numerical experiments, we define the *jump indicator*

$$g(i) = \frac{\int_{\partial K_i} [\rho]^2 \,\mathrm{d}S}{h_i |K_i|^{3/4}}. \tag{2.6.44}$$

Let us return to schemes (2.6.38) and (2.6.39). We carry out the discretization in time by the forward Euler method. To this end, we consider a partition $0 = t_0 < t_1 < t_2 < \ldots$ of the time interval $(0, T)$ and set $\tau_k = t_{k+1} - t_k$. Then the *full discrete problems* read: for each $k \geq 0$ find $\boldsymbol{w}_h^{k+1} \in [S_h]^4$ such that

$$(\boldsymbol{w}_h^{k+1}, \varphi_h) = (\boldsymbol{w}_h^k, \varphi_h) - \tau_k \boldsymbol{B}_h(\boldsymbol{w}_h^k, \varphi_h) \quad \forall \varphi_h \in S_h, \tag{2.6.45}$$

Euler equations

$$(\boldsymbol{w}_h^{k+1}, \varphi_h) = (\boldsymbol{w}_h^k, \varphi_h) + \tau_k \left(\boldsymbol{A}_h(\boldsymbol{w}_h^k, \varphi_h) - \boldsymbol{B}_h(\boldsymbol{w}_h^k, \varphi_h) \right) \quad \forall \varphi_h \in S_h. \tag{2.6.46}$$

Navier-Stokes equations

Now, for each time step we apply the limitation procedure in the following way. The schemes (2.6.45) – (2.6.46) are replaced by

$$(\boldsymbol{w}_h^{k+1}, \varphi_h) = (\bar{\boldsymbol{w}}_h^k, \varphi_h) - \tau_k \boldsymbol{B}_h(\bar{\boldsymbol{w}}_h^k, \varphi_h) \quad \forall \varphi_h \in S_h, \tag{2.6.47}$$

Euler equations

$$(\boldsymbol{w}_h^{k+1}, \varphi_h) = (\bar{\boldsymbol{w}}_h^k, \varphi_h) + \tau_k \left(\boldsymbol{A}_h(\boldsymbol{w}_h^k, \varphi_h) - \boldsymbol{B}_h(\bar{\boldsymbol{w}}_h^k, \varphi_h) \right) \quad \forall \varphi_h \in S_h, \tag{2.6.48}$$

Navier-Stokes equations

where

$$\bar{\boldsymbol{w}}_h^k|_{K_i} := \begin{cases} \pi_0 \boldsymbol{w}_h^k|_{K_i} & \text{if } g(i) > 1, \\ \boldsymbol{w}_h^k|_{K_i} & \text{otherwise}, \end{cases} \qquad K_i \in \mathcal{T}_h, \tag{2.6.49}$$

and

$$\pi_0 \boldsymbol{w}_h|_{K_i} \in P_0(K_i), \ \ \pi_0 \boldsymbol{w}_h|_{K_i} = \frac{1}{K_i} \int_{K_i} \boldsymbol{w}_h \,\mathrm{d}x. \tag{2.6.50}$$

Let as recall that the limitation in (2.6.48) is not performed in the diffusive form $\boldsymbol{A}_h(\cdot, \cdot)$. Numerical examples confirm the high efficiency of the limiting procedure (2.6.44), (2.6.47) – (2.6.50). Namely, Figures 2.28, 2.30 and 2.32 in Section 2.6.5 show the capturing of the discontinuity (shock wave) without any unphysical oscillation.

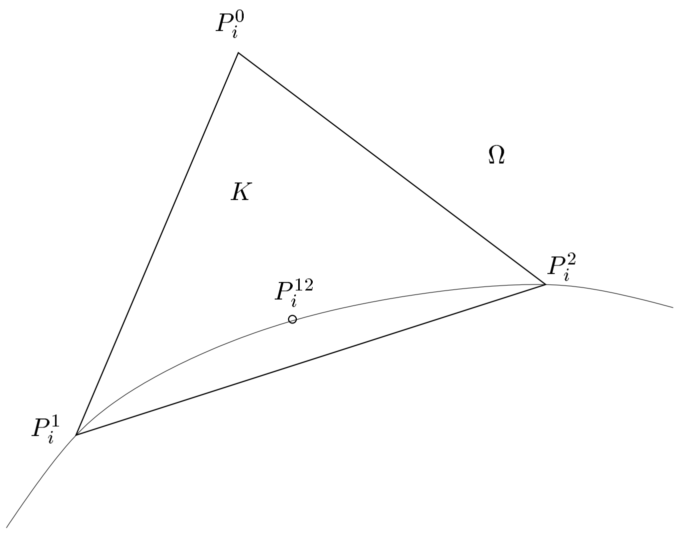


Figure 2.24: Triangle $K_i,\ i \in I_c$ lying on a nonpolygonal part of $\partial\Omega$

2.6.4 Superparametric FE on a nonpolygonal boundary

Numerical experiments show that in some cases the DG FE method does not give good resolution in a neighbourhood of nonpolygonal parts of the boundary $\partial\Omega$, see Example 1 in Section 2.6.5. In order to get a satisfactory solution, it is necessary to use superparametric finite elements or a suitable numerical integration, see [BR00] and [HH02]. We use the first approach.

Let $\{K_i,\ i \in I_c,\ I_c \subset I\}$ be a set of triangles adjacent to a nonpolygonal part of $\partial\Omega$, see Figure 2.24. Let $P_i^k,\ k = 0, 1, 2$ be the vertices of K_i and moreover $P_i^k \in \partial\Omega$ for $k = 1, 2$ and $i \in I_c$. We suppose that the centre $P_i^{12} \in \partial\Omega$ of the curved side with endpoints P_i^1, P_i^2 is close to the centre of the straight segment $P_i^1 P_i^2$, see Figure 2.24.

Then there exists a unique *bilinear* mapping $F_i : \hat{K} \to K_i, F_i = (F_i^1, F_i^2)$, such that

$$\begin{aligned} F_i(\hat{P}^k) &= P_i^k, \quad k = 0, 1, 2, \\ F_i(\hat{P}^{12}) &= P_i^{12}, \end{aligned} \tag{2.6.51}$$

where $\hat{K}$ is a reference triangle with vertices

$$\hat{P}^0 = [0; 0], \quad \hat{P}^1 = [1; 0], \quad \hat{P}^2 = [0; 1] \tag{2.6.52}$$

and

$$\hat{P}^{12} = [1/2; 1/2], \tag{2.6.53}$$

see Figure 2.25.

Then the triangles $K_i,\ i \in I_c$, are replaced by the "curved triangles" defined by

$$\bar{K}_i = F_i(\hat{K}). \tag{2.6.54}$$

The "curved triangle" $\bar{K}_i$ is a plane figure having two straight sides and one curved side, see Figure 2.25. If $i \notin I_c$ then F_i is a linear mapping and, therefore, $\bar{K}_i = K_i$.

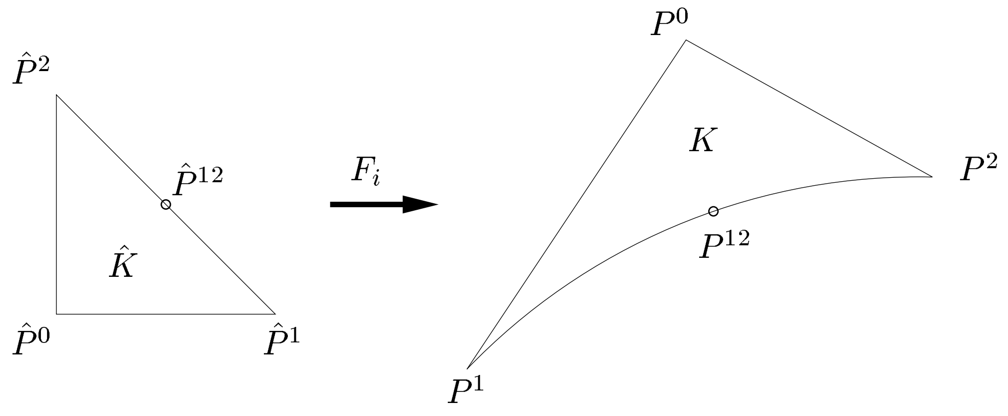


Figure 2.25: Bilinear mapping $F_i : \hat{K} \to K_i$

Now we describe how to evaluate the volume and boundary integrals over the elements $\bar{K}_i,\ i \in I_c$ and their sides Γ_{ij}. We denote by

$$J_{F_i}(\hat{x}) \equiv \frac{D\,F_i}{D\,\hat{x}}(\hat{x}), \quad \hat{x} \in \hat{K}, \tag{2.6.55}$$

the Jacobi matrix of the mapping F_i. Since F_i is bilinear, we easily find that det J_{F_i} is a linear function of the variable $\hat{x}$. The components of the vector valued test functions φ_h from (2.6.38) and (2.6.39) are defined on the boundary elements $\bar{K}_i$ with the aid of the mapping F_i. Denoting by $\hat{\varphi}$ any linear function on the reference element $\hat{K}$, we set

$$\varphi_h(x) = \hat{\varphi}(F_i^{-1}(x)), \quad x \in K_i. \tag{2.6.56}$$

These functions form the components of the test functions from the discrete problem. Then the terms in (2.6.47) and (2.6.47) have modified form. We present it only for inviscid scheme (2.6.47), the modification of integrals of the form $\boldsymbol{A}_h(\cdot,\cdot)$ is very similar as for $\boldsymbol{B}_h(\cdot,\cdot)$.

The first form in (2.6.37) is expressed as

$$\int_{\bar{K}_i} \boldsymbol{w}_h(x,t)\cdot\varphi_h(x)\,\mathrm{d}x = \int_{\hat{K}} \det J_{F_i}(\hat{x})\hat{\boldsymbol{w}}_i(\hat{x},t)\cdot\varphi_h(\hat{x})\,\mathrm{d}\hat{x}, \quad i \in I, \tag{2.6.57}$$

where $\hat{\boldsymbol{w}}_i(\hat{x},t) = \boldsymbol{w}_h(F_i^{-1}(x),t)$.

Moreover, the volume integral in the form $\boldsymbol{B}_h$ in (2.6.37) is computed with the aid of

$$\begin{aligned} &\int_{\bar{K}_i} \sum_{s=1}^{2} \boldsymbol{f}_s(\boldsymbol{w}_h(x,t))\cdot\frac{\partial\varphi_h(x)}{\partial x_s}\,\mathrm{d}x \qquad (2.6.58)\\ =\ &\int_{\hat{K}} \det J_{F_i}(\hat{x}) \sum_{s=1}^{2} \boldsymbol{f}_s(\hat{\boldsymbol{w}}_i(\hat{x},t))\cdot\sum_{j=1}^{2}\frac{\partial\hat{\varphi}}{\partial\hat{x}_j}\frac{\partial(F_i^{-1})^j(F_i(\hat{x}))}{\partial x_s}\,\mathrm{d}\hat{x}, \quad i \in I, \end{aligned}$$

where $(F_i^{-1})^j$ denotes the j-th component of the inverse mapping F_i^{-1}. In order to compute the inverse mapping F_i^{-1} we use the following relation written in matrix form:

$$\frac{D\,F_i^{-1}}{Dx}(F_i(\hat{x})) = \left[\frac{D\,F_i}{D\hat{x}}(\hat{x})\right]^{-1} \tag{2.6.59}$$

following from the identity $x = F_i(F_i^{-1}(x))$. The computation of the inverse matrix in (2.6.59) is simpler than the evaluation of F_i^{-1}.

Further, the boundary integral over a side $\Gamma_{ij} \subset \bar{K}_i$ in the form $\boldsymbol{B}_h$ in (2.6.37) is computed with the aid of

$$\begin{aligned} &\int_{\Gamma_{ij}} \boldsymbol{H}(\boldsymbol{w}_h^+(x,t), \boldsymbol{w}_h^-(x,t), \boldsymbol{n}_{ij}) \cdot \varphi_h(x)\,\mathrm{d}S \\ = \;&\int_{\hat{\Gamma}} \boldsymbol{H}(\boldsymbol{w}_h^+(F_i(\hat{x}),t), \boldsymbol{w}_h^-(F_i(\hat{x}),t), \hat{\boldsymbol{n}}_{ij}(\hat{x})) \cdot \hat{\varphi}(\hat{x})\beta(\hat{x})\,\mathrm{d}\hat{S}, \end{aligned} \tag{2.6.60}$$

where $\hat{\Gamma}$ is the side of $\hat{K}$ corresponding to the curved side of $\bar{K}_i$ in the mapping $F_i = (F_i^1, F_i^2)$ and $\boldsymbol{w}_h^{\pm}(x,t) = \boldsymbol{w}_h^{\pm}(F_i(\hat{x}),t)$. The function $\beta(\hat{x})$ represents a boundary increase for boundary integral in a suitable parametrization $x_1 = x_1(s), x_2 = x_2(s)$:

$$\begin{aligned} &\int_{\Gamma} u(x_1,x_2)\,\mathrm{d}S = \int_{s_1}^{s_2} u(x_1(s),x_2(s))\sqrt{(x_1'(s))^2 + (x_2'(s))^2}\,\mathrm{d}s = \\ = \;&\int_{\hat{s}_1}^{\hat{s}_2} u(F_i^1(\hat{x}_1(\hat{s}),\hat{x}_2(\hat{s})), F_i^2(\hat{x}_1(\hat{s}),\hat{x}_2(\hat{s})))\beta(\hat{x}_1(\hat{s}),\hat{x}_2(\hat{s}))\mathrm{d}\hat{s}. \end{aligned} \tag{2.6.61}$$

Finally, the integrals over the reference triangle $\hat{K}$ and over the reference edge $\hat{\Gamma}$ in (2.6.57), (2.6.58) and (2.6.60) are evaluated with the aid of (2.6.42) and (2.6.41), respectively. Such approach essentially improves the quality of the solution in s neighbourhood of a nonpolygonal part of the boundary as is shown in Section 2.6.5, Figures 2.27 and 2.28.

2.6.5 Numerical examples

We present several examples representing the use of DG FE method for the numerical simulation of compressible flows. We applied the time stabilization method ($t \to \infty$) in order to obtain a steady state solution. The used triangular grid were obtained by the anisotropic mesh adaptation method (AMA), see Chapter 1.

Examples 1 The first set of numerical examples deal with inviscid transonic flow through the GAMM channel. The inlet Mach number $= 0.67$ produces a shock wave. We demonstrate the influence of the use of the superparametric finite elements for triangles lying at the nonpolygonal part of $\partial\Omega$ explained in Section 2.6.4. The computations were performed on a very coarse grid from Figure 2.26 having 784 triangles where such an effect is better observed. Figures 2.27 and 2.28 show the distribution of the density in the computational domain and along the lower wall obtained by scheme (2.6.47) without and with the use of the bilinear mapping, respectively.

Figure 2.30 shows the distribution of the density in Ω and along the lower wall obtained with the aid of the bilinear mapping on the refined mesh shown in Figure 2.29. Here a very sharp capturing of the shock wave and the Zierep singularity are observed.

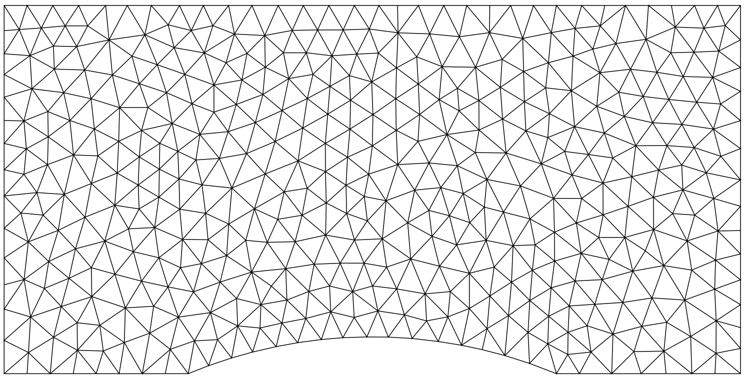


Figure 2.26: Coarse triangular mesh (784 triangles)

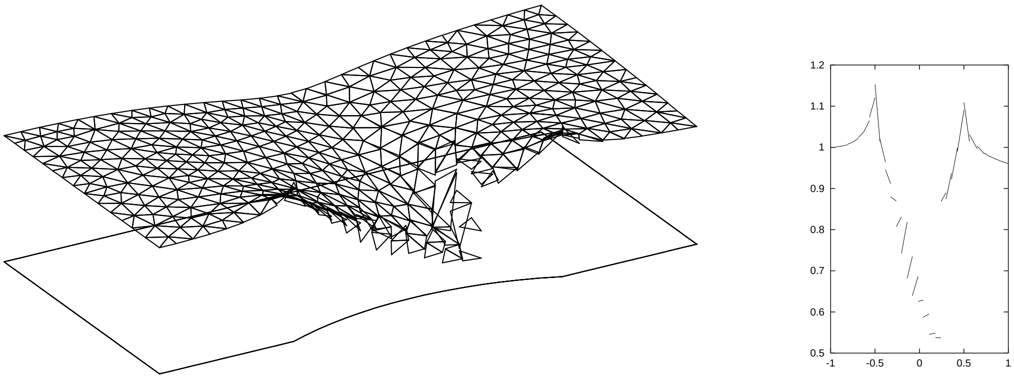


Figure 2.27: Distribution of the density in the computational domain (left) and along the lower wall (right) without the use of the superparametric FE at $\partial\Omega$

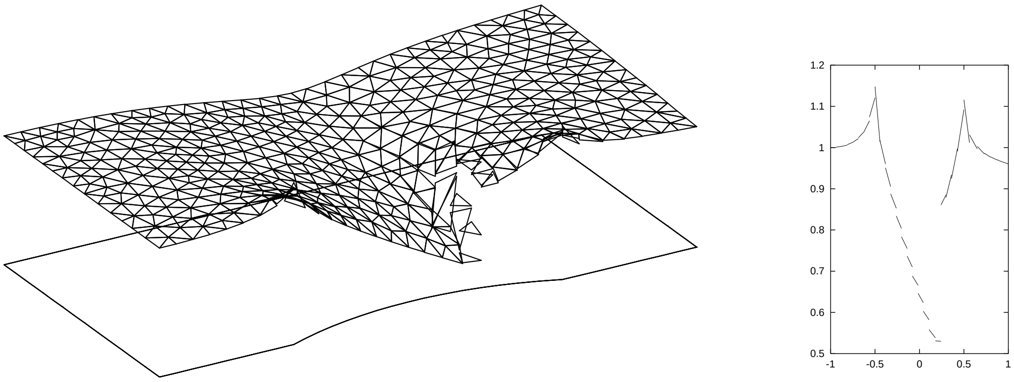


Figure 2.28: Distribution of the density in the computational domain (left) and along the lower wall (right) with the use of the superparametric FE for at $\partial\Omega$

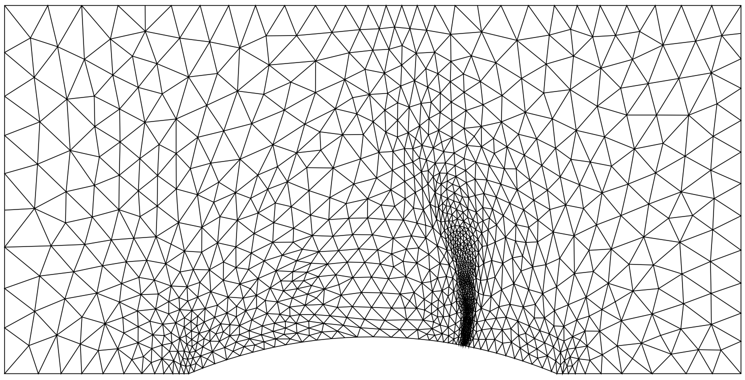


Figure 2.29: Adapted triangular mesh (2131 triangles)

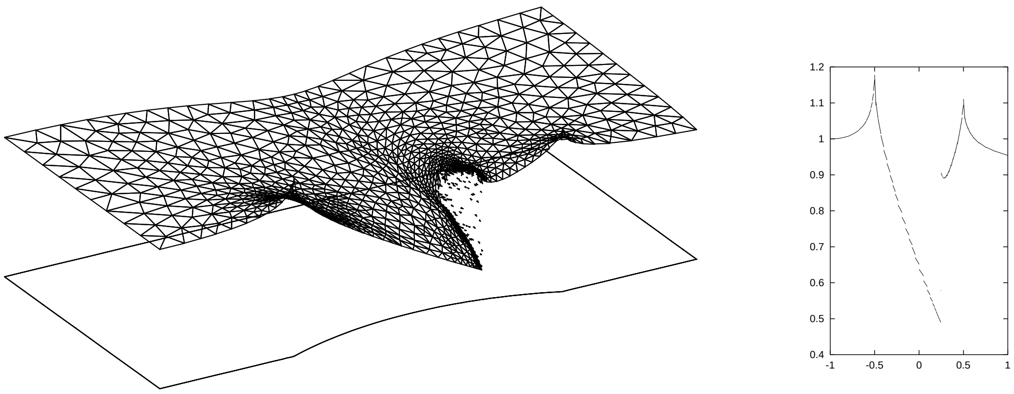


Figure 2.30: Distribution of the density in the computational domain (left) and along the lower wall (right) with superparametric FE on adapted mesh

Example 2 Next example is the the simulation of inviscid around the profile NACA0012 with inlet Mach number $M_{\text{in}} = 0.8$ and zero angle of attack. Figure 2.31 shows the the final mesh (obtained by AMA) and the isolines of the Mach number of this example. As we use a discontinuous approximation then the isolines are discontinuous of course. Figure 2.32 shows the distribution of Mach number and the pressure coefficient c_p around profile. We put

$$c_p = -\frac{p - p_\infty}{\frac{1}{2}\rho_\infty |\boldsymbol{v}_\infty|^2}, \tag{2.6.62}$$

where ρ_∞, $\boldsymbol{v}_\infty$ and p_∞ are the value of density, velocity and pressure at infinity, respectively. Although the used grid is relatively coarse (number of triangles $\#\mathcal{T}_h = 1766$), we obtain qualitatively good results. We observe very sharp capturing of shack wave (thanks adaptivity and discontinuous approximation) without any spurious oscillations. This example is a symmetric problem but the symmetry of the results is slightly smeared. It is caused by the fact, that the used triangulation is completely non symmetric (compare with Example 3, where symmetric mesh is employed).

Examples 3 In the last set of examples, the numerical scheme (2.6.48) was applied for the simulation of viscous flow around the profile NACA0012. The flow computation along NACA0012 profile was performed for the following sets of initial data, see [BGPV87]:

case	M_{in}	α	Re
C1	0.80	10°	500
C2	2.00	10°	106
C3	0.85	0°	500

where M_{in} is the inlet Mach number, α the angle of attack and Re the Reynolds number. We use the set of conditions (2.6.21) a)–b), (2.6.22) c'), where we consider free stream total temperature on the profile, i.e.,

$$\theta_D = \theta_\infty \left(1 + \frac{\gamma - 1}{2} M_\infty^2\right), \tag{2.6.63}$$

where θ_∞ and M_∞ are the value of temperature and Mach number at infinity, respectively. In order to avoid the influence of an artificial external boundary we choose the computational domain Ω with diameter twenty times larger then the diameter of the profile. We compare our results with the numerical results presented in [BGPV87], where ten authors presented their computations using different numerical method. The basic characteristic of the flow is given by the drag (c_D) and lift (c_L) coefficients. The drag and lift coefficient are the first and the second components of the following integral, respectively,

$$\frac{1}{\frac{1}{2}\rho_\infty |u_\infty|^2} \int_{\Gamma_W} \sum_{j=1}^{2} \left[\tau_{ij} + \delta_{ij}(p - p_\infty)\right] n_j \, \mathrm{d}S, \qquad i = 1, 2, \tag{2.6.64}$$

where ρ_∞, $|u_\infty|$ and p_∞ are the values of the density, the magnitude of velocity vector and the pressure at infinity, respectively. The integral is calculated over the NACA0012 profile Γ_W. Further $\tau_{ij},\ i,j = 1,2$ is the viscous part of stress tensor given by (2.6.18), p is the pressure and $\delta_{ij},\ i,j = 1,2$ is the Kronecker delta. The following table contains our computed lift c_L and drag c_D coefficients in comparison with [BGPV87] ($\#\mathcal{T}_h$ denotes the number of elements of $\mathcal{T}_h$)

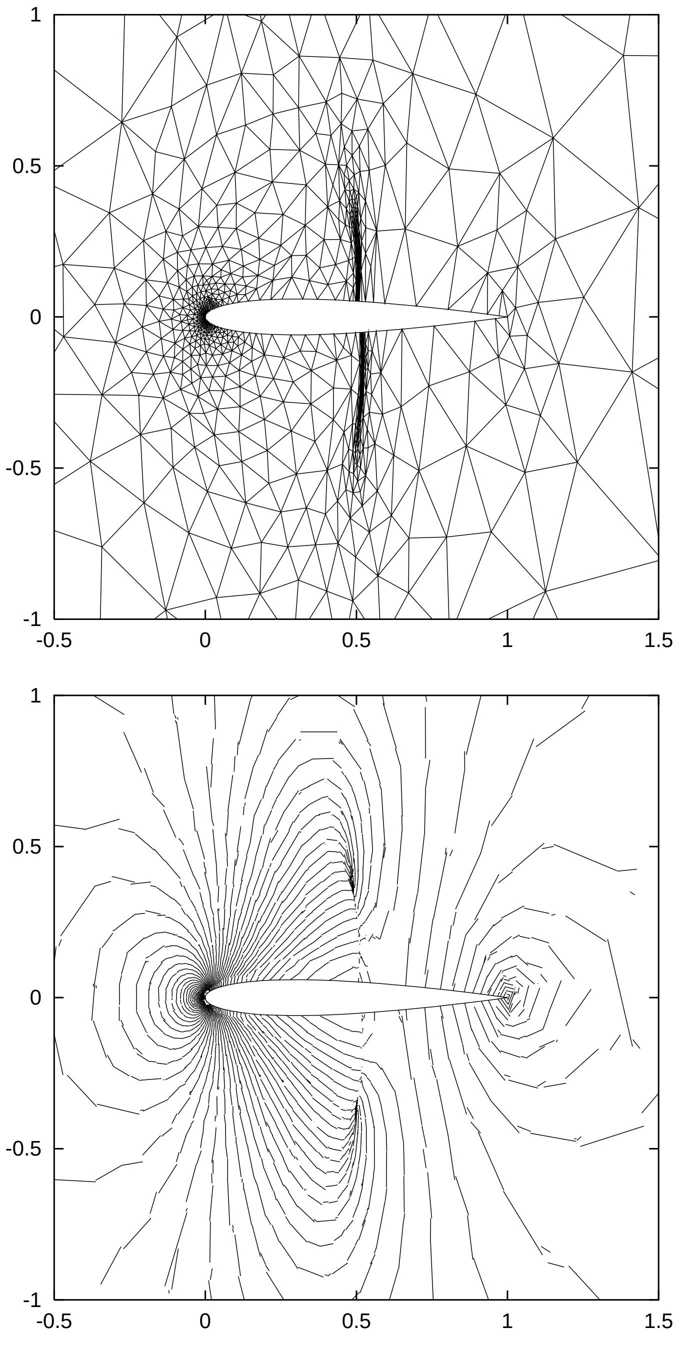
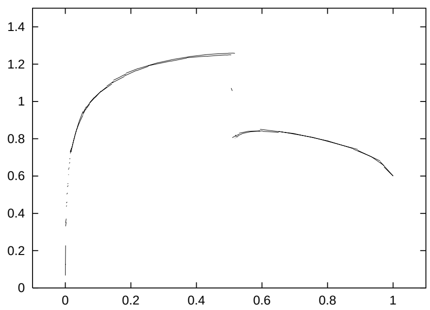


Figure 2.31: Inviscid flow along NACA 0012, $M_{\text{in}} = 0.8$, zero angle of attack, triangulation (top) and isolines of Mach number (bottom)



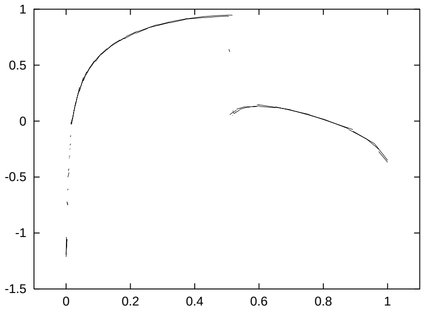


Figure 2.32: Inviscid flow along NACA 0012, $M_{\rm in} = 0.8$, zero angle of attack, distribution of Mach number (left) and pressure coefficient (right) around the profile

case	$\#\mathcal{T}_h$	computed values c_L	computed values c_D	reference values from [BGPV87] c_L range; mean value	reference values from [BGPV87] c_D (range); mean value
C1	4563	0.4985	0.1938	(0.4199 – 0.5170); 0.4526	(0.1597 – 0.2868); 0.2559
C2	5640	0.3969	0.4172	(0.3063 – 0.4059); 0.3443	(0.4120 – 0.4910); 0.4660
C3	4946	0.0003	0.2304	(0.0000 – 0.0007); 0.0001	(0.1790 – 0.2420); 0.2192

The computed values of c_L and c_D correspond with the reference ones. Figures 2.33 – 2.38 show the employed triangulations, computed isolines of the Mach number and density and the distributions of the density and the pressure coefficient along the profile. The isolines are not "smooth" which is caused by the fact that the used meshes are coarse and the visualization is performed without applying any smoothing mechanism. On the other hand, the distributions of ρ and c_p along profile are smooth as the meshes in vicinity of the profile are fine.

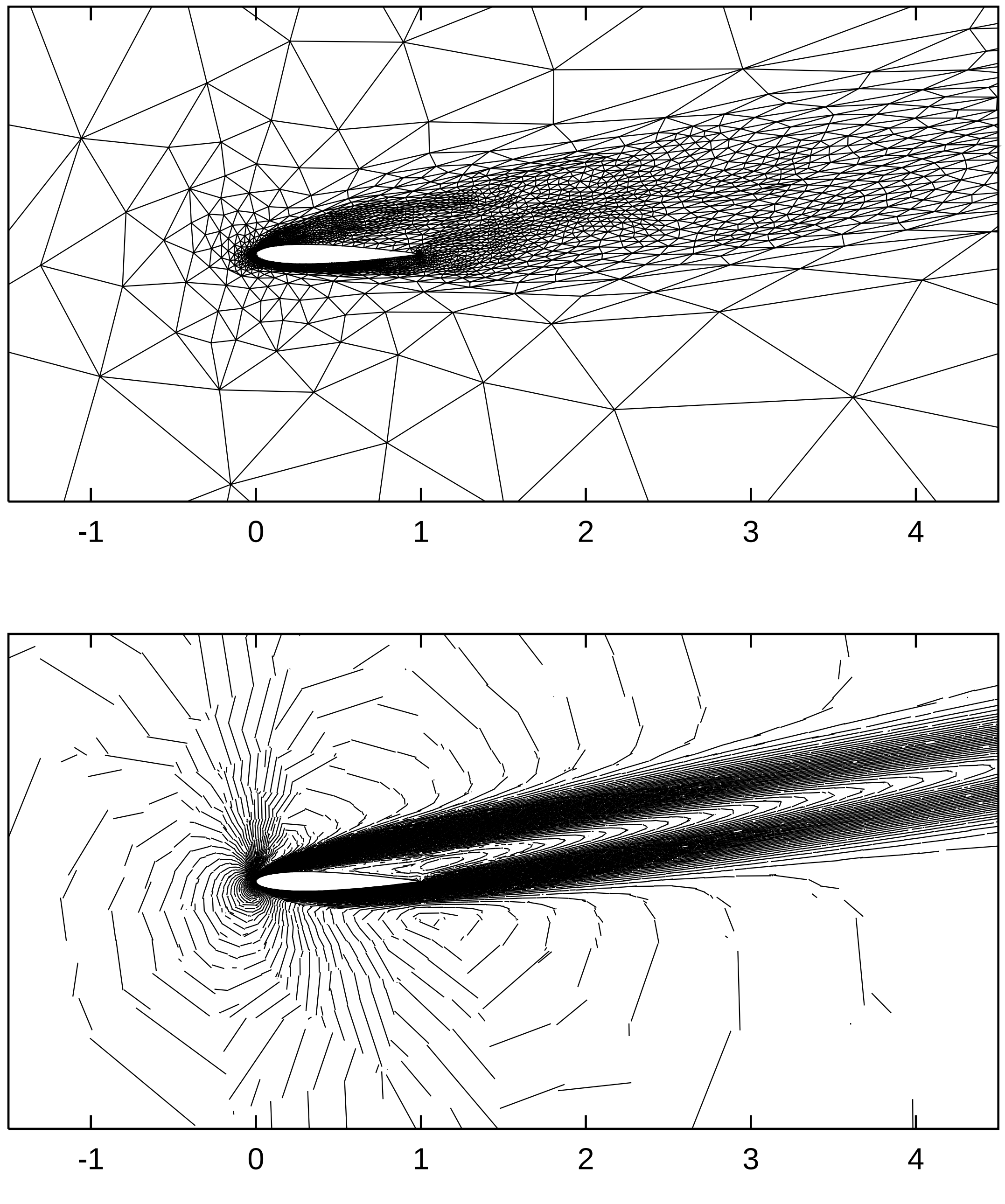


Figure 2.33: Viscous flow along NACA 0012, case C1, triangulation (top), isolines of Mach number (bottom)

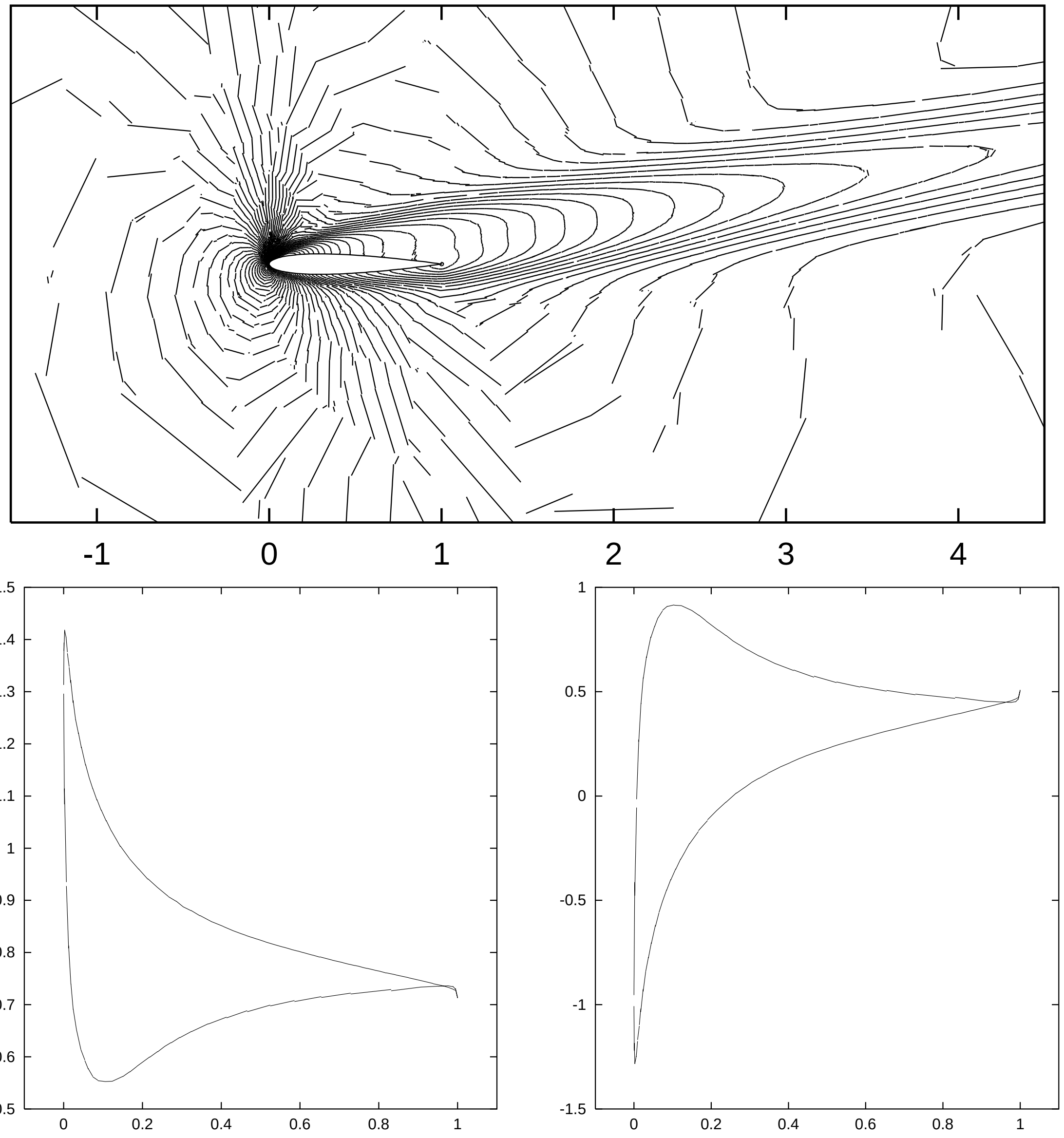


Figure 2.34: Viscous flow along NACA 0012, case C1, isolines of density (top), distribution of density (bottom left) and pressure coefficient (bottom right) along the profile

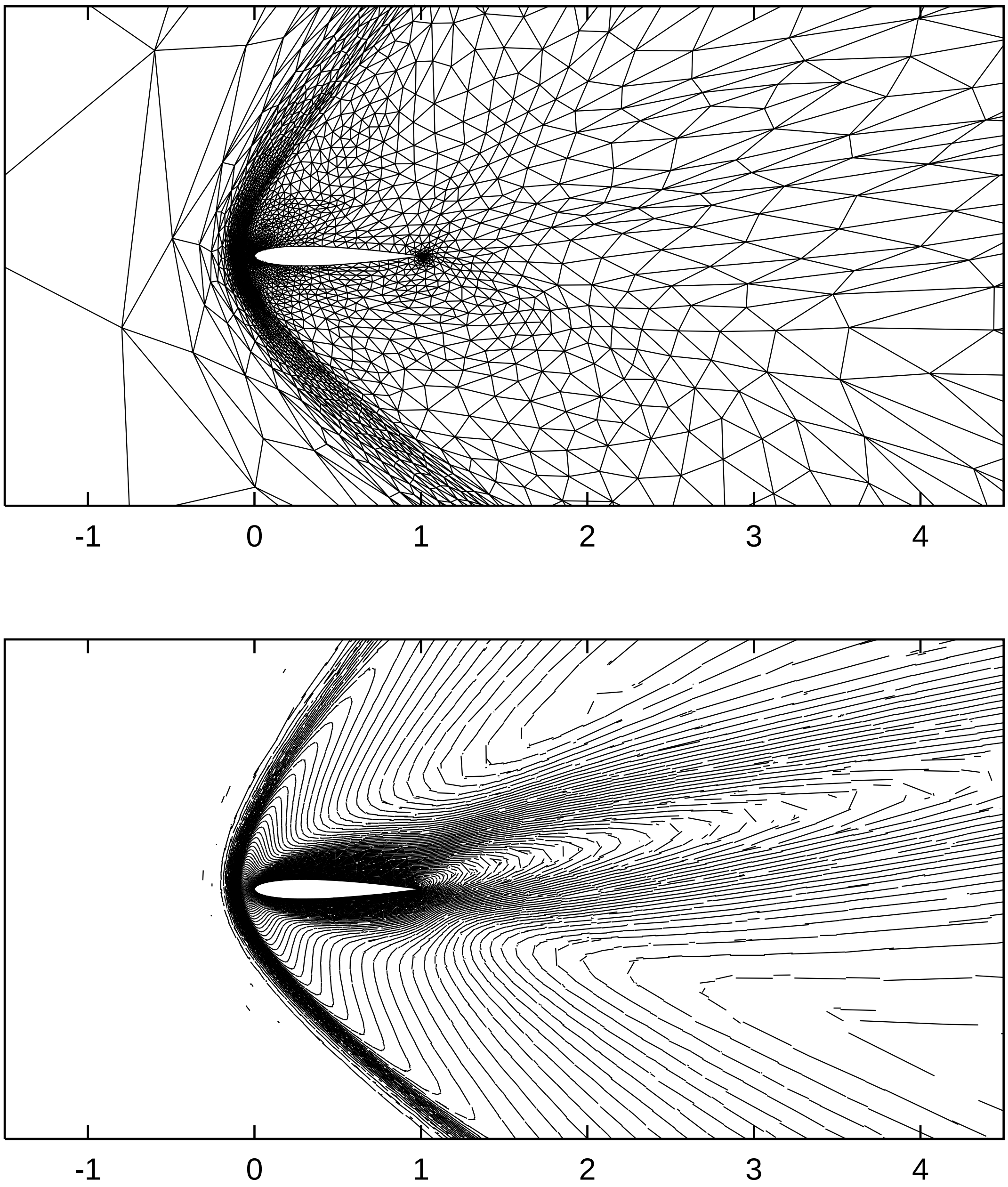


Figure 2.35: Viscous flow along NACA 0012, case C2, triangulation (top), isolines of Mach number (bottom)

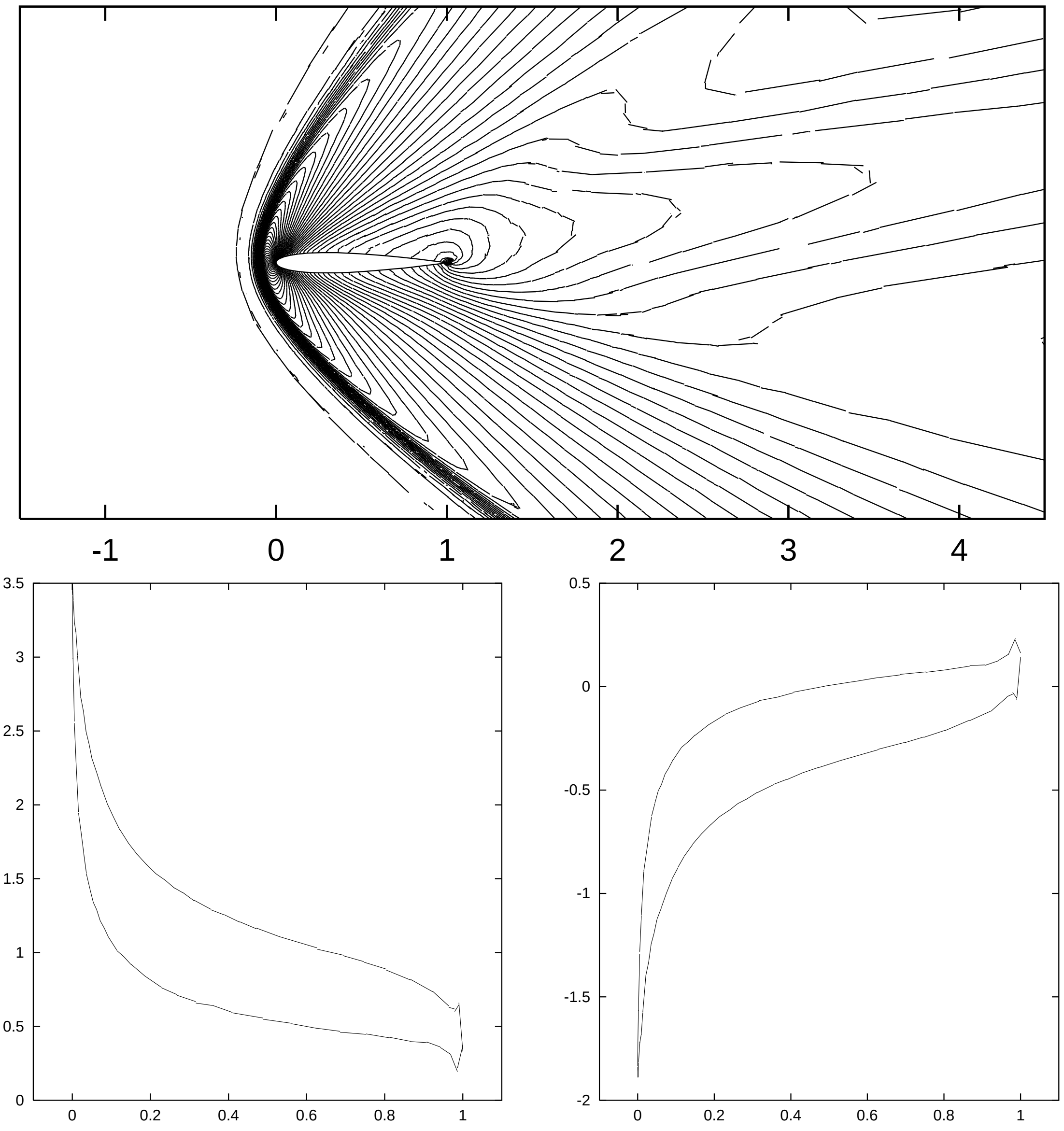


Figure 2.36: Viscous flow along NACA 0012, case C2, isolines of density (top), distribution of density (bottom left) and pressure coefficient (bottom right) along the profile

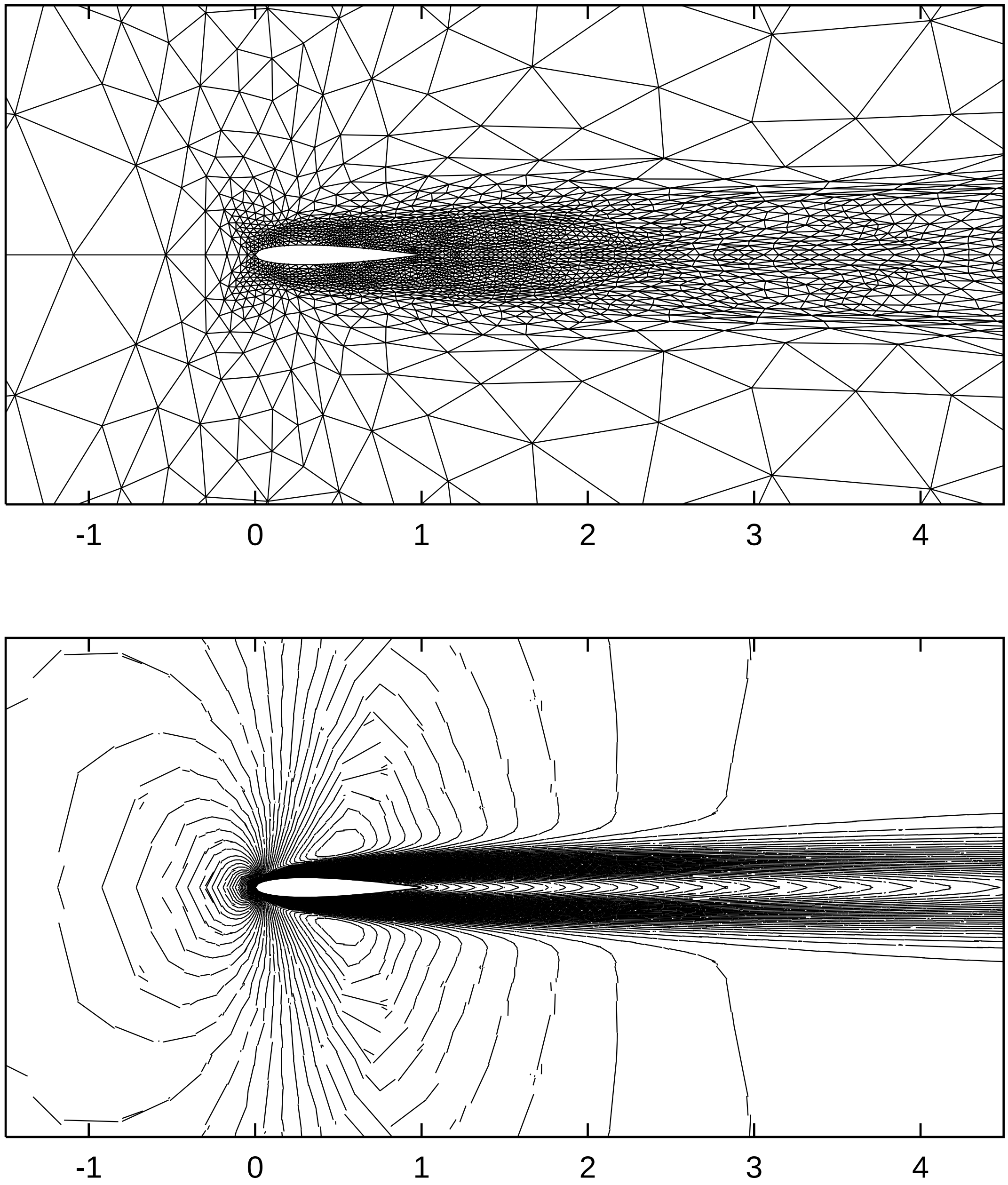


Figure 2.37: Viscous flow along NACA 0012, case C3, triangulation (top), isolines of Mach number (bottom)

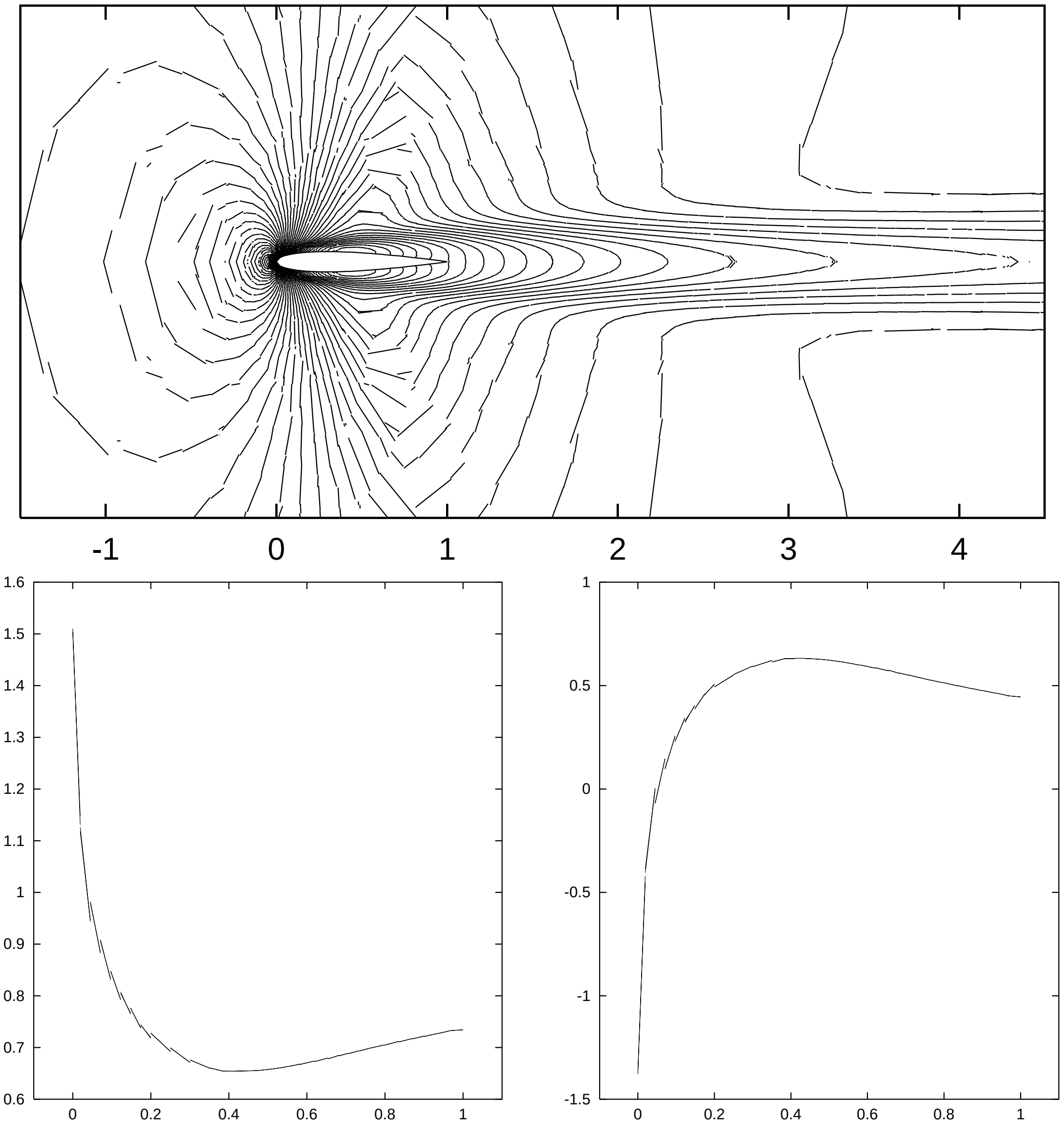


Figure 2.38: Viscous flow along NACA 0012, case C3, isolines of density (top), distribution of density (bottom left) and pressure coefficient (bottom right) along the profile

2.7 Conclusion

From the analysis in Section 2.2 and numerical experiments in Sections 2.5 and 2.6, we see that the discontinuous Galerkin finite element method represents very effective techniques for the solution of nonlinear conservation laws and convection diffusion problems. An interesting observation is the following fact: although a discontinuous approximation is employed, the numerical solutions look continuous (i.e. interelement jumps are not visible) in subdomains where the solution is smooth. The higher order of the scheme and the use of superparametric finite elements for curved parts of boundary produce satisfactory results even on coarse grids.

The important achievement is the construction of adaptive order of accuracy limiting, which avoids spurious oscillations and preserving the higher order numerical scheme unchanged in regions, where the solution is regular. Of course, there is a number of unsolved questions and problems, as, e.g., analysis of error estimates of this method, theoretical justification of the proposed adaptive limiting and extension for higher order of approximation ($p > 1$).

The drawback of the presented methods is a high restriction on the time step given by the stability condition (2.6.43). Then in order to achieve the steady state solution, the high number of iterations has to be performed (in examples from Section 2.6 about $5 \cdot 10^4$). Therefore the future research should be devoted to the use of implicit schemes where the stability condition will be not so restrictive.

Bibliography

[ADFF98] Ph. Angot, V. Dolejší, M. Feistauer, and J. Felcman. Analysis of a combined barycentric finite volume – nonconforming finite element method for nonlinear convection-diffusion problems. *Appl. Math.*, 43(4):263–310, 1998.

[AM99] P. Arminjon and A. Madrane. A mixed finite volume/finite element method for 2-dimensional compressible Navier–Stokes equations on unstructured grids. In M. Fey and R. Jeltsch, editors, *Hyperbolic Problems: Theory, Numerics, Applications*, volume I, pages 11–20. Birkhäuser, Basel, 1999.

[Ape99] T. Apel. *Anisotropic Finite Elements: Local Estimates and Applications.* Teubner, Stuttgart - Leipzig, 1999.

[BBO99] I. Babuška, C. E. Baumann, and J. T. Oden. A discontinuous *hp* finite element method for diffusion problems: 1-d analysis. *Computers and Mathematics with Applications*, 37:103–122, 1999.

[BD77] G. C. Buscaglia and E. A. Dari. Anisotropic mesh optimization and its application in adaptivity. *International Journal for Numerical Methods in Engineering*, 40:4119–4136, 1977.

[BDGG89] R. Busto, W. D. Dietrich, M. Y. Globus, and M. D. Ginsberg. The importance of brain temperature in cerebral ischemic injury. *Stroke*, 20:1114–1134, 1989.

[BDM] M. Breuss, V. Dolejší, and A. Meister. On an adaptive method for heat conduction problems with boundary layers. *Preprint Hamburger Beiträge zur Angenwandten Mathematik.* (submitted to Computing and Visualization in Science).

[BGPV87] M. O. Bristeau, R. Glowinski, J. Periaux, and H. Viviand, editors. *Numerical Simulation of Compressible Navier-Stokes Flows*, volume 18 of *Notes on Numerical Fluid Mechanics.* Vieweg, Braunschweig, 1987.

[BHF97] H. Borouchaki, F. Hecht, and P. J. Frey. Mesh gradation control. In *Proceedings of 6th International Meshing Roundtable*, Park City, Utah, USA, 1997.

[BMM+00] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous Galerkin approximations for elliptic problems. *Numer. Methods Partial Differ. Equations*, 16:35–378, 2000.

[BR97] F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier–Stokes equations. *J. Comput. Phys*, 131:267–279, 1997.

[BR00] F. Bassi and S. Rebay. A high order discontinuous Galerkin method for compressible turbulent flow. In M. Griebel, D. E. Keyes, R. M. Nieminen, D. Rose, and T. Schlick, editors, *Discontinuous Galerkin Method: Theory, Computations and Applications, Lecture Notes in Computation Science and Engineering.* Springer-Verlag, 2000.

[BSW83] R. Bank, A. Sherman, and A. Weiser. Refinement algorithm and data structures for regular local mesh refinement. In R. Stepleman et al., editor, *Scientific Computing*, volume 44, pages 3–17. IMACS North–Holland, Amsterdam, 1983.

[CCSS02] P. Castillo, B. Cockburn, D. Schötzau, and C. Schwab. Optimal a priori estimates for the *hp*–version of the local discontinuous Galerkin method for convection–difussion problems. *Math. Comp.*, 71(238):455–478, 2002.

[CDBG+96] M. J. Castro-Díaz, H. Borouchaki, P. L. George, F. Hecht, and B. Mohammadi. Anisotropic Adaptive Mesch Generation in Two Dimensions for CFD. In J.-A. Désidéri, C. Hirsch, P. Le Tallec, M. Pandolfi, and J. Périaux, editors, *Computational Fluid Dynamics '96*, pages 181–186, Paris, 1996. Wiley, Chichester.

[CHS89] B. Cockburn, S. Hou, and C. W. Shu. TVB Runge–Kutta local projection discontinuous Galerkin finite element for scalar conservation laws II: General framework. *Math. Comp.*, 52:411–435, 1989.

[Cia79] P. G. Ciarlet. *The Finite Elements Method for Elliptic Problems.* North-Holland, Amsterdam, New York, Oxford, 1979.

[CKe00] B. Cockburn, G. E. Karniadakis, and C.-W. Shu (eds.). Discontinuous Galerkin methods. In *Lecture Notes in Computational Science and Engineering 11.* Springer, Berlin, 2000.

[Coc] B. Cockburn. An introduction to the discontinuous Galerkin method for convection-dominated problems.

[Coc99] B. Cockburn. Discontinuous Galerkin methods for convection dominated problems. In T. J. Barth and H. Deconinck, editors, *High–Order Methods for Computational Physics*, Lecture Notes in Computational Science and Engineering 9, pages 69–224. Springer, Berlin, 1999.

[CR73] M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. *RAIRO, Anal. Numér.*, 7:33–76, 1973.

[DA96] V. Dolejší and Ph. Angot. Finite Volume Methods on Unstructured Meshes for Compressible Flows. In F. Benkhaldoun and R. Vilsmeier, editors, *Finite Volumes for Complex Applications (Problems and Perspectives)*, pages 667–674. Hermes, Rouen, 1996.

[DF] V. Dolejší and J. Felcman. Anisotropic mesh adaptation and its application for scalar diffusion equations. (submitted to Numerical Methods for Partial Differential Equations).

[DF02a] 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.

[DF02b] V. Dolejší and J. Felcman. Numerical methods for hyperbolic systems - application in fluid dynamics. *Engineering Mechanics*, 9(3):139–152, 2002.

[DFFK02] V. Dolejší, M. Feistauer, J. Felcman, and A. Kliková. Error estimates for barycentric finite volumes combined with nonconforming finite elements applied to nonlinear convection–diffusion problems. *Applications of Mathematics*, 47(4):301–340, 2002.

[DFS] V. Dolejší, M. Feistauer, and C. Schwab. A discontinuous Galerkin method for nonlinear convection–diffusion problems. (in preparation).

[DFS01] V. Dolejší, M. Feistauer, and C. Schwab. On some aspects of the discontinuous Galerkin finite element method for conservation laws. *The Preprint Series of the School of Mathematics, Charles University Prague, No. MATH–KNM–2001/5*, 2001. (to appear in Mathematics and Computers in Simulation).

[DFS02a] V. Dolejší, M. Feistauer, and C. Schwab. A finite volume discontinuous Galerkin scheme for nonlinear convection–diffusion problems. *Calcolo*, 39:1–40, 2002.

[DFS02b] V. Dolejší, M. Feistauer, and C. Schwab. On discontinuous Galerkin methods for nonlinear convection–diffusion problems and compressible flow. *Mathematica Bohemica*, 127(2):163–179, 2002.

[DHM95] M. J. Castro Díaz, F. Hecht, and B. Mohammadi. New progress in anisotropic grid adaptation for inviscid and viscous flows simulations. In *Proceedings of the 4th Annual International Meshing Roundtable.* Sandia National Laboratories, 1995.

[Dol98a] V. Dolejší. Anisotropic mesh adaptation for finite volume and finite element methods on triangular meshes. *Computing and Visualization in Science*, 1(3):165–178, 1998.

[Dol98b] V. Dolejší. *Sur des méthodes combinant des volumes finis et des éléments finis pour le calcul d'écoulements compressibles sur des maillages non structurés.* PhD thesis, Charles Univeristy Prague and Université Mediterannée Aix–Marseille II, 1998.

[Dol01] V. Dolejší. Anisotropic mesh adaptation technique for viscous flow simulation. *East-West Journal of Numerical Mathematics*, 9(1):1–24, 2001.

[Doled] V. Dolejsi. On the discontinuous galerkin method for the numerical solution of the euler and the Navier–Stokes equations. *Int. J. Numer. Methods Fluids*, (submitted).

[DS89] E. F. D'Azevedo and R. B. Simpson. On optimal interpolation triangle incidences. *SIAM. J. Sci. Statist. Comput.*, 10:1063–1075, 1989.

[DS91] E. F. D'Azevedo and R. B. Simpson. On optimal triangular meshes for minimizing the gradient error. *Numer. Math.*, 59:321–348, 1991.

[Duf96] D. Duffie. *Dynamic Asset Pricing Theory.* Princeton University Press, Princeton New Jersey, 1996.

[FD98] J. Felcman and V. Dolejší. Investigation of accuracy for higher order finite volume schemes. *Engineering Mechanics*, 5(5):327–337, 1998.

[FDF94] J. Felcman, V. Dolejší, and M. Feistauer. Adaptive Finite Volume Method for the Numerical Solution of the Compressible Euler Equations. In J. Périaux S. Wagner, E. H. Hirschel and R. Piva, editors, *Computational Fluid Dynamics '94*, pages 894–901. John Wiley and Sons, Stuttgart, 1994.

[Fei93] M. Feistauer. *Mathematical Methods in Fluid Dynamics.* Longman Scientific & Technical, Harlow, 1993.

[Fei01] M. Feistauer. Numerical methods for compressible flow. In J. Neustupa and P. Penel, editors, *Mathematical Fluid Mechanics. Recent Results and Open Questions.* Birkhäuser, Basel, 2001.

[Fel01] J. Felcman. On a 3D adaptation for compressible flow. In P. Neittaanmäki and M. Křížek, editors, *Finite Element Methods for Three-dimensional Problems*, volume 15 of *GAKUTO Internat. Series, Math. Sci. Appl.*, pages 85–97, Tokyo, 2001. University of Jyväskylä, GAKKOTOSHO CO., LTD.

[FF96] M. Feistauer and J. Felcman. Convection-Diffusion Problems and Compressible Navier–Stokes Equations. In J. R. Whitemen, editor, *The Mathematics of Finite Elements and Applications*, pages 175–194. John Wiley & Sons, 1996.

[FFD96] M. Feistauer, J. Felcman, and V. Dolejší. Numerical simulation of compresssible viscous flow through cascades of profiles. *ZAMM*, 76(S4):297–300, 1996.

[FFL95] M. Feistauer, J. Felcman, and M. Lukáčová. Combined finite element–finite volume solution of compressible flow. *J. Comput. Appl. Math.*, 63:179–199, 1995.

[FFLM97] M. Feistauer, J. Felcman, and M. Lukáčová-Medvidová. On the convergence of a combined finite volume – finite element method for nonlinear convection – diffusion problems. *Numer. Methods Partial Differ. Equations*, 13:163–190, 1997.

[FFLMW99] M. Feistauer, J. Felcman, M. Lukáčová-Medvidová, and G. Warnecke. Error estimates of a combined finite volume – finite element method for nonlinear convection – diffusion problems. *SIAM J. Numer. Anal.*, 36(5):1528–1548, 1999.

[FJS] M. Feistauer, J.Felcman, and I. Straškraba. Mathematical and computational methods for compressible flow.

[FLM00] B. Fischer, M. Ludwig, and A. Meister. An application of the finite volume method to the thermoregulation of infants. Hamburg Beiträge zur Angewandten

Mathematik, Reihe F, Computational Fluid Dynamics and Data Analysis 12, 2000.

[FOG97] L. A. Freitag and C. Ollivier-Gooch. Tetrahedral mesh improvement using swapping and smoothing. *Int. J. Numer. Methods Eng*, 40(21):3979–4002, 1997.

[FŠ98] J. Felcman and P. Šolín. On the construction of the Osher–Solomon scheme for 3D Euler equations. *East–West J. Numer. Math.*, 6(1):43–64, 1998.

[FVD+96] M. Fortin, M.-G. Vallet, J. Dompierre, Y. Bourgault, and W. G. Habashi. Anisotropic Mesh Adaptation: Theory, Validation and Applications. In J.-A. Désidéri, C. Hirsch, P. Le Tallec, M. Pandolfi, and J. Périaux, editors, *Computational Fluid Dynamics '96*, pages 174–180, Paris, 1996. Wiley, Chichester.

[GP99] J. M. Ghidaglia and F. Pascal. Footbridges finite volumes – finite elements. In *C. R. Acad. Sc, Ser. I, Math. 328*, volume 8, pages 711–716. Paris, 1999.

[GR91] E. Godlewski and P. A. Raviart. *Hyperbolic Systems of Conservation Laws*, volume 3–4 of *Mathématiques & Applications.* Paris: Ellipses, 1991.

[GW92] P. D. Gluckman and C. E. Williams. When and why do brain cells die? *Dev. Med. Child Neurol.*, 34:1010–1014, 1992.

[Hac89] W. Hackbusch. On first and second order box schemes. In *Computing 44*, pages 277–296, 1989.

[HFaY+96] W. G. Habashi, M. Fortin, D. Ait ali Yahia, S. Boivin, Y.Bourgault, J. Dompierre, M. P. Robichaud, A. Tam, and M. G. Vallet. Anisotropic mesh optimization: Towards a solver-independent and mesh-independent CFD. In *Compressible Fluid Dynamics*, VKI Lecture Series 1996-06. von Karman Institute for Fluid Dynamics - Concordia University, August 1996.

[HH02] R. Hartmann and P. Houston. Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic conservation laws. *SIAM J. Sci. Comp.*, 24:979–1004, 2002.

[HSS00] 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.

[Joh81] C. Johnson. Finite element methods for convection – diffusion problems. In R. Glowinski and J. L. Lions, editors, *Computing Methods in Engineering and Applied Sciences V.* North–Holland, Amsterdam, 1981.

[JP86] C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, 46:1–26, 1986.

[KJk77] A. Kufner, O. John, and S. Fučík. *Function Spaces.* Academia, Prague, 1977.

[Kli00] A. Kliková. *Finite volume – finite element solution of compressible flow.* PhD thesis, Charles Univeristy, Prague, 2000.

[KR94] D. Kröner and M. Rokyta. Convergence of upwind finite volume schemes for scalar conservation laws in 2D. *SIAM J. Numer. Anal.*, 31:324–343, 1994.

[Krö97] D. Kröner. *Numerical Schemes for Conservation Laws.* Wiley Teubner, Stuttgart, 1997.

[Kuz76] N. N. Kuznetsov. Accuracy of some approximate methods for computing the weak solutions of a first–order quasi–linear equation. *USSR. Comput. Math. Phys.*, 16:105–119, 1976.

[Lio96] P. L. Lions. *Mathematical Topics in Fluid Mechanics.* Oxford Science Publications, 1996.

[LR74] P. Le Saint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In C. de Boor, editor, *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 89–145. Academic Press, 1974.

[MS99] J. M. Melenk and C. Schwab. An *hp* finite element method for convection–diffusion problems. *East–West J. Numer. Math*, 7:31–60, 1999.

[MWJG95] E. C. Mallard, C. E. Williams, B. M. Johnston, and P. D. Gluckman. Neuronal damage in the developing brain following intrauterine asphyxia. *Reprod. Fertil. Dev.*, 7:647–653, 1995.

[OBB98] J. T. Oden, I. Babuška, and C. E. Baumann. A discontinuous *hp* finite element method for diffusion problems. *J. Comput. Phys*, 146:491–519, 1998.

[OS82] S. Osher and F. Solomon. Upwind difference schemes for hyperbolic systems of conservation laws. *Math. Comp.*, 38:339–374, 1982.

[PW61] L. E. Payne and H. F. Weinberger. An optimal poincaré inequality for convex domains. *Archive for Rational Mechanics and Analysis*, 11:286–292, 1961.

[Ran98] R. Rannacher. Adaptive finite element methods. In C. Zenger H. Bulgak, editor, *NATO–Summer School: Error Control and Adaptivity in Scientific Computing*, pages 247–278. Kluwer, Academic Publishers, 1998.

[Rek82] K. Rektorys. *The Method of Discretization in Time and Partial Differential Equations.* Reidel, Dodrecht, 1982.

[RH73] W. H. Reed and T. R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA–UR–73–479, Los Alamos Scientific Laboratory, 1973.

[RST96] H.-G. Roos, M. Stynes, and L. Tobiska. Numerical Methods for Singularly Perturbed Differential Equation. *Springer Series in Computational Mathematics*, 24, 1996. Springer-Verlag, Berlin.

[Sim94] R. B. Simpson. Anisotropic mesh transformations and optimal error control. *Applied Numer. Math.*, 14:183–198, 1994.

[ŠŠ90] M. Štastný and P. Šafařík. Experimental analysis data on the transonic flow past a plane turbine. Technical report, ASME Paper 90-GT-313, New York, 1990.

[Swe84] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 21:995–1011, 1984.

[Tem77] R. Temam. *Navier–Stokes Equations. Theory and numerical analysis.* North–Holland, Amsterdam–New York–Oxford, 1977.

[Tor97] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics.* Springer-Verlag, 1997.

[Ver96] 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.

[Ver99] R. Verfürth. A note on polynomial approximation in Sobolev spaces. M^2AN, 33:715–719, 1999.

[WC84] P. Woodward and P. Colella. The numerical simulation of two-dimensional fluid flow with strong shocks. *J. of Comput. Phys.*, 54:115–173, 1984.

[ZZ92a] 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, 1992.

[ZZ92b] 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, 1992.