DISCONTINUOUS GALERKIN METHOD

Vít Dolejší, Miloslav Feistauer
Charles University Prague
Faculty of Mathematics and Physics
Czech Republic
dolejsi@karlin.mff.cuni.cz

September 20, 2016

Preface

These lecture notes more or less cover the lecture Discontinuous Galerkin methode given by the author at the master and PhD program at the Charles University in Prague, the Faculty of Mathematics and Physics.

Contents

0.1 Some mathematical concepts

In this section for the reader's convenience, we recall some basic tools of mathematical analysis, which are frequently used in the book. We assume that the reader is familiar with mathematical analysis, including the theory of the Lebesgue integral, and elements of functional analysis, see, for example, [Rud87].

If X is a set or space and $n > 0$ is an integer, then the symbol $X^n = (X)^n$ denotes the Cartesian product $X \times \cdots \times X$ (n-times). This means that

$$X^n = (X)^n = \{(x_1, \dots, x_n);\ x_1, \dots, x_n \in X\}. \tag{1}$$

By $\mathbb{R}$ and $\mathbb{N}$ we denote the set of all real numbers and the set of all positive integers, respectively. In the Euclidean space $\mathbb{R}^d$ ($d \in \mathbb{N}$) we use a Cartesian coordinate system with axes denoted by $x_1, \dots, x_d$. Points from $\mathbb{R}^d$ will usually be denoted by $x = (x_1, \dots, x_d)$, $y = (y_1, \dots, y_d)$, etc. By $|\cdot|$ we denote the Euclidean norm in $\mathbb{R}^d$. Thus, $|x| = \left(\sum_{i=1}^d |x_i|^2\right)^{1/2}$.

Now we introduce some function spaces and their properties, which will be used in the sequel. For deeper results and proofs, we refer the reader to the monographs [AF03], [KJk77], [Žen90].

0.1.1 Spaces of continuous functions

Let us assume that $d \in \mathbb{N}$ and $M \subset \mathbb{R}^d$ is a domain (i.e., an open connected set). By ∂M and $\overline{M}$ we denote its boundary and closure, respectively. By $C(M)$ (or $C^0(M)$) we denote the linear space of all functions continuous in M. For $k \in \mathbb{N}$ and a domain $M \subset \mathbb{R}^d$, $C^k(M)$ denotes the linear space of all functions which have continuous partial derivatives up to the order k in M. The space $C^k(\overline{M})$ is formed by all functions from $C^k(M)$ whose all derivatives up to the order k can be continuously extended onto $\overline{M}$.

Let $M \subset \mathbb{R}^d$. A function $f : M \to \mathbb{R}$ is *μ-Hölder-continuous* with $\mu \in (0, 1]$, if there exists a constant L such that

$$|f(x) - f(y)| \le L|x - y|^\mu \qquad \forall x,\, y \in M. \tag{2}$$

If $\mu = 1$, we speak of a *Lipschitz-continuous* (or simply *Lipschitz*) function. If $M \subset \mathbb{R}^d$ is a domain, then $C^{k,\mu}(\overline{M})$ denotes the set of all functions whose derivatives of order k are μ-Hölder-continuous in $\overline{M}$.

Let us put

$$C^\infty(M) = \bigcap_{k=1}^\infty C^k(M) \quad \text{and} \quad C^\infty(\overline{M}) = \bigcap_{k=1}^\infty C^k(\overline{M}). \tag{3}$$

By $C_0^\infty(M)$ we denote the linear space of all functions $v \in C^\infty(\overline{M})$, whose *support*

$$\operatorname{supp} v = \overline{\{x \in M;\ v(x) \ne 0\}} \tag{4}$$

is a compact (i.e. bounded and closed) subset of the domain M.

If $\alpha_i \ge 0,\ i = 1, \dots, d$, are integers, then we call $\alpha = (\alpha_1, \dots, \alpha_d)$ a multi-index, and define its length as $|\alpha| = \sum_{i=1}^d \alpha_i$. By D^α we denote the multidimensional derivative of order $|\alpha|$:

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}. \tag{5}$$

The linear space $C^k(\overline{M})$, $k = 0, 1, \dots$, equipped with the norm

$$\|u\|_{C^k(\overline{M})} = \sum_{|\alpha| \le k} \sup_{x \in M} |D^\alpha u(x)| \tag{6}$$

is a Banach space. This space is separable but not reflexive.

The linear space $C^{k,\mu}(\overline{M})$, where $k = 0, 1, \dots$, and $\mu \in (0, 1]$, equipped with the norm

$$\|u\|_{C^{k,\mu}(\overline{M})} = \|u\|_{C^k(\overline{M})} + \sum_{|\alpha| = k} \sup_{x, y \in M,\, x \ne y} \frac{|(D^\alpha u)(x) - (D^\alpha u)(y)|}{|x - y|^\mu} \tag{7}$$

is a Banach space. It is called the *Hölder space*. This space is neither separable nor reflexive.

Finally, the symbols ∇ and $\nabla\cdot$ mean the gradient and divergence operators, respectively, i.e.,

$$\nabla u = \operatorname{grad} u = \left(\frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_d}\right)^{\mathrm{T}} \in \mathbb{R}^d \quad \text{for } u : M \to \mathbb{R} \tag{8}$$

and

$$\nabla \cdot \boldsymbol{u} = \operatorname{div} \boldsymbol{u} = \sum_{i=1}^{d} \frac{\partial u_i}{\partial x_i} \in \mathbb{R}, \quad \text{for } \boldsymbol{u} = (u_1, \dots, u_d) : M \to \mathbb{R}^d, \tag{9}$$

where the superscript $^{\mathrm{T}}$ denotes the transposed vector.

The symbols D^α, ∇ and $\nabla\cdot$ are also used for the distributional derivatives; see Section 0.1.3.

0.1.2 Lebesgue spaces

First we recall some standard notation and results from the Lebesgue theory of measure and integral, see, e.g., [Rud87]. Let $M \subset \mathbb{R}^d$, $d = 1, 2, \dots$, be a Lebesgue-measurable set. Its d-dimensional Lebesgue measure will be denoted by $\operatorname{meas}(M)$ or for short $|M|$. We recall that two measurable functions are *equivalent* if they differ at most on a set of zero Lebesgue measure. Then we say that these functions are equal almost everywhere (a.e.) in M.

For $s \in [1, \infty)$ the *Lebesgue space* $L^s(M)$ is the linear space of all functions measurable on M (more precisely, of classes of equivalent measurable functions) such that

$$\int_M |u|^s \, \mathrm{d}x < +\infty. \tag{10}$$

The space $L^s(M)$ is equipped with the norm

$$\|u\|_{L^s(M)} = \left(\int_M |u|^s \, \mathrm{d}x \right)^{1/s}. \tag{11}$$

In case that $s = \infty$, the space $L^\infty(M)$ consists of such measurable functions on M for which the norm

$$\|u\|_{L^\infty(M)} = \operatorname*{ess\,sup}_M |u| = \inf \left\{ \sup_{x \in M \setminus Z} |u(x)|; Z \subset M, \operatorname{meas}(Z) = 0 \right\} \tag{12}$$

is finite. The space $L^s(M)$ is a Banach space for $1 \le s \le \infty$. Moreover, it is separable if and only if $1 \le s < \infty$ and reflexive if and only if $1 < s < \infty$. The space $L^2(M)$ is a Hilbert space with the scalar product

$$(u, v)_{L^2(M)} = \int_M uv \, \mathrm{d}x. \tag{13}$$

The *Cauchy inequality* holds in $L^2(M)$:

$$|(u, v)_{L^2(M)}| \le \|u\|_{L^2(M)} \|v\|_{L^2(M)}, \quad u, v \in L^2(M). \tag{14}$$

0.1.3 Sobolev spaces

Let $M \subset \mathbb{R}^d$, $d = 1, 2, \dots$, be a domain, let $k \ge 0$ be an arbitrary integer and $1 \le s \le \infty$. We define the *Sobolev space* $W^{k,s}(M)$ as the space of all functions from the space $L^s(M)$ whose distributional derivatives $D^\alpha u$, up to the order k, also belong to $L^s(M)$, i.e.,

$$W^{k,s}(M) = \{u \in L^s(M); D^\alpha u \in L^s(M) \ \forall \alpha, \ |\alpha| \le k\}, \tag{15}$$

(See e.g. [KJk77], [Fei93], [Leo09].)

The Sobolev space is equipped with the norm

$$\begin{aligned} \|u\|_{W^{k,s}(M)} &= \left(\sum_{|\alpha| \le k} \|D^\alpha u\|^s_{L^s(M)} \right)^{1/s} && \text{for } 1 \le s < \infty, \\ \|u\|_{W^{k,\infty}(M)} &= \max_{|\alpha| \le k} \left\{ \|D^\alpha u\|_{L^\infty(M)} \right\} && \text{for } s = \infty, \end{aligned} \tag{16}$$

and the seminorm

$$\begin{aligned} |u|_{W^{k,s}(M)} &= \left(\sum_{|\alpha| = k} \|D^\alpha u\|^s_{L^s(M)} \right)^{1/s} && \text{for } 1 \le s < \infty, \\ |u|_{W^{k,\infty}(M)} &= \max_{|\alpha| = k} \left\{ \|D^\alpha u\|_{L^\infty(M)} \right\} && \text{for } s = \infty. \end{aligned} \tag{17}$$

For $1 \le s \le \infty$, the space $W^{k,s}(M)$ is a Banach space; it is separable if and only if $1 \le s < \infty$ and reflexive if and only if $1 < s < \infty$. For $s = 2$, the space $W^{k,2}(M)$ is a Hilbert space and we denote it by $H^k(M)$. Moreover, we put

$$\|u\|_{H^k(M)} = \|u\|_{W^{k,2}(M)} \quad \text{and} \quad |u|_{H^k(M)} = |u|_{W^{k,2}(M)}. \tag{18}$$

If $k = 0$, then we set $W^{0,s}(M) = L^s(M)$, $H^0(M) = L^2(M)$ and

$$|\cdot|_{W^{0,s}(M)} = \|\cdot\|_{W^{0,s}(M)} = \|\cdot\|_{L^s(M)}. \tag{19}$$

For vector-valued functions $\boldsymbol{v} = (v_1, \dots, v_n) \in (H^s(\Omega))^n$, we put

$$\|\boldsymbol{v}\|_{H^k(M)} = \Big(\sum_{i=1}^{n} \|v_i\|^2_{H^k(M)} \Big)^{1/2}. \tag{20}$$

Moreover, with respect to (8), (17), (18) and (20), we write

$$\|\nabla v\|_{L^2(M)} = |v|_{H^1(M)},\ v \in H^1(M), \qquad |\nabla v|_{H^1(M)} = |v|_{H^2(M)},\ v \in H^2(M). \tag{21}$$

0.1.4 Theorems on traces and embeddings

In the modern theory of partial differential equations the concept of a bounded domain $M \subset \mathbb{R}^d$ with Lipschitz boundary ∂M plays an important role. For the definition of a Lipschitz boundary, see, e.g., [KJk77], [Fei93], [Žen90] or Section 2.3.2. It is possible to say that such a boundary ∂M is formed by a finite number of parts expressed as graphs of Lipschitz-continuous functions in local Cartesian coordinate systems. On this boundary, the $(d-1)$-dimensional Lebesgue measure meas_{d-1} and integral are defined and also an outer unit normal vector exists at a.e. point $x \in \partial M$. Moreover, Lebesgue spaces $L^s(\partial M)$ are defined over ∂M.

Theorem 0.1 (Theorem on traces)**.** *Let $1 \le s \le \infty$ and let $M \subset \mathbb{R}^d$ be a domain with Lipschitz boundary. Then there exists a uniquely determined continuous linear mapping $\gamma_0^M : W^{1,s}(M) \to L^s(\partial M)$ such that*

$$\gamma_0^M(u) = u|_{\partial M} \quad \textit{for all } u \in C^\infty(\overline{M}). \tag{22}$$

Moreover, if $1 \ge s \le \infty$, then Green's formula

$$\int_M \left(u \frac{\partial v}{\partial x_i} + v \frac{\partial u}{\partial x_i} \right) \mathrm{d}x = \int_{\partial M} \gamma_0^M(u) \gamma_0^M(v) n_i \, \mathrm{d}S, \tag{23}$$
$$u \in W^{1,s}(M),\ v \in W^{1,s'}(M),\ i = 1, \dots, d,$$

holds, where $s' = s/(s-1)$ and $\boldsymbol{n} = (n_1, \dots, n_d)$ denotes the outer unit normal to ∂M.

The function $\gamma_0^M(u) \in L^s(\partial M)$ is called the *trace* of the function $u \in W^{1,s}(M)$ on the boundary ∂M. For simplicity, when there is no confusion, the notation $u|_{\partial M} = \gamma_0^M(u)$ is used not only for $u \in C^\infty(\overline{M})$ but also for $u \in W^{1,s}(M)$. The continuity of the mapping γ_0^M is equivalent to the existence of a constant $c > 0$ such that

$$\|u|_{\partial M}\|_{L^s(\partial M)} = \|\gamma_0^M(u)\|_{L^s(\partial M)} \le c \|u\|_{W^{1,s}(M)}, \quad u \in W^{1,s}(M). \tag{24}$$

Let $k \ge 1$ be an integer and $1 \le s < \infty$. We define the Sobolev space $W_0^{k,s}(M)$ as the closure of the space $C_0^\infty(M)$ in the topology of the space $W^{k,s}(M)$. If M is a domain with Lipschitz boundary, then $W_0^{1,s}(M) = \{v \in W^{1,s}(M); v|_{\partial M} = 0\}$.

The space of traces on $\partial\Omega$ of all functions $u \in H^1(\Omega)$ is denoted by $H^{1/2}(\partial\Omega)$. Hence, we can write

$$H^{1/2}(\partial\Omega) = \{\gamma_0^\Omega u; u \in H^1(\Omega)\}. \tag{25}$$

If $k \in \mathbb{N}$, we define the space

$$H^{k-1/2}(\partial\Omega) = \{\gamma_0^\Omega u; u \in H^k(\Omega)\}. \tag{26}$$

We speak of Sobolev–Slobodetskii spaces on $\partial\Omega$. (See e. g., [FFS03, Section 1.3.3].)

Note that the symbols c and C will often denote a positive *generic constant*, attaining, in general, different values in different places.

Embedding theorems

Definition 0.2. *Let X, Y be Banach spaces. We say that X is continuously embedded into Y (we write $X \hookrightarrow Y$), if X is a subspace of Y and the identity operator $I : X \to Y$ defined by $Ix = x$ for all $x \in X$ is continuous, i.e., there exists $C > 0$ such that*

$$\|Iv\|_Y \le C\|v\|_X \quad \forall v \in X.$$

We say that X is compactly embedded into Y ($X \hookrightarrow\hookrightarrow Y$) if the embedding operator I is compact.

Theorem 0.3. *The following properties are valid:*
(i) Let $k \ge 0$, $1 \le s \le \infty$ and let $M \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary. Then

$$\begin{aligned}
&W^{k,s}(M) \hookrightarrow L^q(M) \text{ where } \frac{1}{q} = \frac{1}{s} - \frac{k}{d}, \text{ if } k < \frac{d}{s}, \\
&W^{k,s}(M) \hookrightarrow L^q(M) \text{ for all } q \in [1,\infty), \text{ if } k = \frac{d}{s}, \\
&W^{k,s}(M) \hookrightarrow C^{0,k-d/s}(\overline{M}), \text{ if } \frac{d}{s} < k < \frac{d}{s} + 1, \\
&W^{k,s}(M) \hookrightarrow C^{0,\alpha}(\overline{M}) \text{ for all } \alpha \in (0,1), \text{ if } k = \frac{d}{s} + 1, \\
&W^{k,s}(M) \hookrightarrow C^{0,1}(\overline{M}), \text{ if } k > \frac{d}{s} + 1.
\end{aligned} \tag{27}$$

(ii) Let $k > 0$, $1 \le s \le \infty$. Then

$$\begin{aligned}
&W^{k,s}(M) \hookrightarrow\hookrightarrow L^q(M) \text{ for all } q \in [1, s^*) \text{ with } \frac{1}{s^*} = \frac{1}{s} - \frac{k}{d}, \quad \text{if } k < \frac{d}{s}, \\
&W^{k,s}(M) \hookrightarrow\hookrightarrow L^q(M) \text{ for all } q \in [1,\infty), \text{ if } k = \frac{d}{s}, \\
&W^{k,s}(M) \hookrightarrow\hookrightarrow C(\overline{M}), \text{ if } k > \frac{d}{s}.
\end{aligned}$$

(We set $1/\infty := 0$.)
(iii) Let $1 \le s < \infty$. Then $C^\infty(\overline{M})$ is dense in $W^{k,s}(M)$ and $C_0^\infty(\overline{M})$ is dense in $W_0^{k,s}(M)$.
d) By [Leo09, Exercise 1146, page 342], if the domain M is bounded, then the space $W^{1,\infty}(M)$ can be identified with the space $C^{0,1}(\overline{M})$.

Remark 0.4. *In some cases, it is suitable to use the concept of the domain with boundary having the* cone property*. This is more general than the concept of the Lipschitz boundary, but the above definitions and results remain valid. See [AF03].*

0.1.5 Bochner spaces

In the investigation of nonstationary problems we shall work with functions which depend on time and have values in a Banach space. Such functions are elements of the so-called Bochner spaces. If $u(x,t)$ is a function of the space variable x and time t, then it is sometimes suitable to separate these variables and consider u as a function $u(t) = u(\cdot,t)$, which, for each t under consideration, attains a value $u(t)$ that is a function of x and belongs to a suitable space of functions depending on x. This means that $u(t)$ represents the mapping "$x \to (u(t))(x) = u(x,t)$".

Let a, $b \in \mathbb{R}$, $a < b$, and let X be a Banach space with norm $\|\cdot\|$. By a *function defined in the interval $[a,b]$ with its values in the space X* we understand any mapping $u : [a,b] \to X$.

We say that a function $u : [a,b] \to X$ is *continuous at a point $t_0 \in [a,b]$*, if

$$\lim_{\substack{t \to t_0 \\ t \in [a,b]}} \|u(t) - u(t_0)\| = 0. \tag{28}$$

By the symbol $C([a,b];\, X)$ we denote the space of all functions continuous in the interval $[a,b]$ (i.e., continuous at each $t \in [a,b]$) with values in X). The space $C([a,b];\, X)$ equipped with the norm

$$\|u\|_{C([a,b];\, X)} = \max_{t \in [a,b]} \|u(t)\| \tag{29}$$

is a Banach space.

For $s \in [1,\infty]$, we denote by $L^s(a,b;X)$ the space of (classes of equivalent) strongly measurable functions $u : (a,b) \to X$ such that

$$\|u\|_{L^s(a,b;X)} = \left[\int_a^b \|u(t)\|_X^s \,\mathrm{d}t\right]^{1/s} < \infty, \quad \text{if } 1 \le s < \infty, \tag{30}$$

and

$$\begin{aligned}\|u\|_{L^\infty(a,b;X)} &= \operatorname{ess}\sup_{t\in(a,b)} \|u(t)\|_X \\ &= \inf\left\{\sup_{t\in(a,b)\setminus N} \|u(t)\|_X; N \subset (a,b), \operatorname{meas}(N) = 0\right\} < +\infty, \quad \text{if } s = \infty.\end{aligned} \tag{31}$$

We speak of Bochner spaces. It can be proved that $L^s(a,b;X)$ is a Banach space. (The definition of a strongly measurable function $u : (a,b) \to X$ can be found in [KJk77] or [Fei93, Chapter 8].)

If the space X is reflexive, so is $L^s(a,b;X)$ for $s \in (1,\infty)$. Let $1 \le s < \infty$. Then the dual of $L^s(a,b;X)$ is $L^q(a,b;X^*)$, where $1/s + 1/q = 1$ and X^* is the dual of X (for $s = 1$ we set $q = \infty$). The duality between $L^q(a,b;X^*)$ and $L^s(a,b;X)$ becomes

$$\langle f, v\rangle = \int_a^b \langle f(t), v(t)\rangle_{X^*,X} \,\mathrm{d}t, \quad f \in L^q(a,b;X^*),\ v \in L^s(a,b;X). \tag{32}$$

The symbol $\langle f(t), v(t)\rangle_{X^*,X}$ denotes the value of the functional $f(t) \in X^*$ at $v(t) \in X$.

If X is a separable Banach space, then $L^s(a,b;X)$ is also separable, provided $s \in [1,\infty)$. (See, for example, [Edw65, Section 8.18.1].)

Let $|\cdot|_X$ denote a seminorm in the space X. Then a seminorm in $L^s(a,b;X)$ is defined as

$$|f|_{L^s(a,b;X)} = \left(\int_a^b |f(t)|_X^s \,\mathrm{d}t\right)^{1/s} \quad \text{for } 1 \le s < +\infty, \tag{33}$$

and

$$|f|_{L^\infty(a,b;X)} = \operatorname{ess\,sup}_{t\in(a,b)} |f(t)|_X . \tag{34}$$

Similarly we define Sobolev spaces of functions with values in X:

$$W^{k,s}(a,b;X) = \left\{f \in L^s(a,b;X); \frac{\mathrm{d}^j f}{\mathrm{d}t^j} \in L^s(a,b;X),\ j = 1,\dots,k\right\}, \tag{35}$$

where $k \in \mathbb{N}$, $s \in [1,\infty]$ and $\frac{\mathrm{d}^j f}{\mathrm{d}t^j}$ are distributional derivatives. The norm of $f \in W^{k,s}(a,b;X)$ is defined by

$$\|f\|_{W^{k,s}(a,b;X)} = \left(\sum_{j=0}^k \left\|\frac{\mathrm{d}^j f}{\mathrm{d}t^j}\right\|_{L^s(a,b;X)}^s\right)^{1/s} \tag{36}$$

for $s \in [1,\infty)$ and

$$\|f\|_{W^{k,\infty}(a,b;X)} = \max_{j=0,\dots k}\left\|\frac{\mathrm{d}^j f}{\mathrm{d}t^j}\right\|_{L^\infty(a,b;X)}. \tag{37}$$

If $s = 2$, we often use the notation $H^k(a,b;X) = W^{k,2}(a,b;X)$.

Let $|\cdot|_X$ denote a seminorm in the space X. Then a seminorm in $W^{k,s}(a,b;X)$ is defined as

$$|f|_{W^{k,s}(a,b;X)} = \left(\int_a^b \left|\frac{\mathrm{d}^k f}{\mathrm{d}t^k}(t)\right|_X^s \,\mathrm{d}t\right)^{1/s} \quad \text{for } 1 \le s < +\infty, \tag{38}$$

and

$$|f|_{W^{k,\infty}(a,b;X)} = \operatorname{ess\,sup}_{t\in(a,b)} \left|\frac{\mathrm{d}^k f}{\mathrm{d}t^k}(t)\right|_X . \tag{39}$$

For example,

$$|f|_{H^k(a,b;H^1(M))} = \left(\int_a^b \left| \frac{\mathrm{d}^k f}{\mathrm{d}t^k}(t) \right|_{H^1(M)}^2 \mathrm{d}t \right)^{1/2}. \tag{40}$$

We also define spaces of continuously differentiable functions on an interval $I = [a, b]$ with values in X:

$$C^k(I;X) = \left\{ f \in C(I;X);\ \frac{\mathrm{d}^j f}{\mathrm{d}t^j} \in C(I;X) \text{ for all } j = 1, \dots, k \right\}. \tag{41}$$

The norm of $f \in C^k(I;X), k = 0, 1, \dots$, is defined by

$$\|f\|_{C^k(I;X)} = \max\left\{ \left\| \frac{\mathrm{d}^j f}{\mathrm{d}t^j} \right\|_{C(I;X)};\ j = 0, \dots, k \right\}. \tag{42}$$

These spaces are nonreflexive Banach spaces. They are separable if X is separable.

If X is a Banach space with norm $\|\cdot\|_X$, then by X^* we denote its dual space (simply dual), i.e., the space of all continuous linear functionals on X. The space X^* is also a Banach space with norm

$$\|f\|_{X^*} = \sup_{v \in X} \frac{|f(v)|}{\|v\|_X} \quad \forall f \in X^*. \tag{43}$$

Finally, if $p \geq 0$ is an integer and $\omega \subset \mathbb{R}^n$, then by $P_p(\omega)$ we denote the space of the restrictions on ω of all polynomials of degree $\leq p$ depending on $x \in \mathbb{R}^n$. We simply speak of polynomials of degree $\leq p$ on ω.

For nonstationary problems, we shall use spaces of polynomial functions with respect to time. Let $-\infty < a < b < \infty$. If X is a Banach space, then we put

$$P_q(a,b;X) = \left\{ v \in C(a,b;X);\ v(t) = \sum_{i=0}^{q} t^i \varphi_i, \varphi_i \in X, i = 0, \dots, q,\ t \in [a,b] \right\}. \tag{44}$$

0.1.6 Useful theorems and inequalities

Lemma 0.5 (Young's inequality)**.** *If $s, q \in (1, +\infty)$, $1/s + 1/q = 1$ and $a, b \geq 0$, then*

$$ab \leq \frac{a^s}{s} + \frac{b^q}{q}. \tag{45}$$

In particular, if $s = q = 2$ and $\lambda > 0$, then

$$ab \leq \frac{1}{2\lambda} a^2 + \frac{\lambda}{2} b^2. \tag{46}$$

Proof. See, e.g., [FHH+11, Lemma 1.11.] □

Lemma 0.6 (Lax–Milgram)**.** *Let V be a Hilbert space with norm $\|\cdot\|$, let $f : V \to \mathbb{R}$ be a continuous linear functional on V, and let $a : V \times V \to \mathbb{R}$ be a bilinear form on $V \times V$ that is coercive, i.e., there exists a constant $\alpha > 0$ such that*

$$a(u,u) \geq \alpha \|u\|^2 \qquad \forall u \in V, \tag{47}$$

and continuous (also called bounded) and, hence, there exists a constant $C_B > 0$ such that

$$|a(u,v)| \leq C_B \|u\| \|v\| \qquad \forall u, v \in V. \tag{48}$$

Then there exists a unique solution $u_0 \in V$ of the problem

$$a(u_0, v) = f(v) \qquad \forall v \in V. \tag{49}$$

Proof. See [Cia79, Theorem 1.1.3]. □

Corollary 0.7. *Let V_N be a finite-dimensional Hilbert space with norm $\|\cdot\|$, let $f : V_N \to \mathbb{R}$ be a linear functional on V_N, and let $a : V_N \times V_N \to \mathbb{R}$ be a bilinear form on $V_N \times V_N$ which is coercive, i.e., there exists a constant $\alpha > 0$ such that*

$$a(u,u) \geq \alpha \|u\|^2 \qquad \forall u \in V_N. \tag{50}$$

Then there exists a unique solution $u_0 \in V_N$ of the problem

$$a(u_0, v) = f(v) \qquad \forall v \in V_N. \tag{51}$$

Proof. Since the space V_N is finite dimensional, the bilinear form a and the functional f are continuous. Then the application of the Lax–Milgram Lemma 0.6 gives the assertion. Let us note that all norms on the finite-dimensional space are equivalent. □ □

Lemma 0.8 (Discrete Cauchy inequality)**.** *Let $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$ be two sequences of real numbers. Then*

$$\left|\sum_{i=1}^{n} a_i b_i\right| \le \left(\sum_{i=1}^{n} a_i^2\right)^{1/2} \left(\sum_{i=1}^{n} b_i^2\right)^{1/2}. \tag{52}$$

In the analysis of nonstationary problems, the following versions of Gronwall's lemma will be applied.

Lemma 0.9 (Gronwall's lemma)**.** *Let y, q, z, $r \in C([0,T])$, $r \ge 0$, and let*

$$y(t) + q(t) \le z(t) + \int_0^t r(s)\, y(s)\, \mathrm{d}s, \quad t \in [0,T]. \tag{53}$$

Then

$$\begin{aligned} &y(t) + q(t) + \int_0^t r(\vartheta)\, q(\vartheta) \exp\left(\int_\vartheta^t r(s)\, \mathrm{d}s\right) \mathrm{d}\vartheta \\ \le &z(t) + \int_0^t r(\vartheta)\, z(\vartheta) \exp\left(\int_\vartheta^t r(s)\, \mathrm{d}s\right) \mathrm{d}\vartheta, \quad t \in [0,T]. \end{aligned} \tag{54}$$

Proof. Inequality (53) can be written in the form

$$y(t) \le h(t) + \int_0^t r(s)\, y(s)\, \mathrm{d}s, \tag{55}$$

where

$$h(t) = z(t) - q(t). \tag{56}$$

Let us set

$$z_1(t) = \int_0^t r(s)\, y(s)\, \mathrm{d}s. \tag{57}$$

Then $z_1'(t) = r(t)\, y(t)$, $z_1(0) = 0$. Since $r(t) \ge 0$, it follows from (55) that

$$z_1'(t) \le h(t)\, r(t) + r(t)\, z_1(t). \tag{58}$$

If we set

$$w(t) = z_1(t) \exp\left(-\int_0^t r(s)\, \mathrm{d}s\right), \tag{59}$$

then, by (58),

$$\begin{aligned} w'(t) &= z_1'(t) \exp\left(-\int_0^t r(s)\, \mathrm{d}s\right) - z_1(t)\, r(t) \exp\left(-\int_0^t r(s)\, \mathrm{d}s\right) \\ &\le (h(t)\, r(t) + r(t)\, z_1(t)) \exp\left(-\int_0^t r(s)\, \mathrm{d}s\right) - \ r(t)\, z_1(t) \exp\left(-\int_0^t r(s)\, \mathrm{d}s\right) \\ &= h(t)\, r(t) \exp\left(-\int_0^t r(s)\, \mathrm{d}s\right). \end{aligned} \tag{60}$$

Taking into account that $w(0) = 0$ and integrating (60) from 0 to t, we get

$$w(t) \le \int_0^t h(\vartheta)\, r(\vartheta) \exp\left(-\int_0^\vartheta r(s)\, \mathrm{d}s\right) \mathrm{d}\vartheta.$$

This and (59) imply that

$$z_1(t) \le \exp\left(\int_0^t r(s)\,\mathrm{d}s\right)\int_0^t h(\vartheta)\,r(\vartheta)\exp\left(-\int_0^\vartheta r(s)\,\mathrm{d}s\right)\mathrm{d}\vartheta \tag{61}$$
$$= \int_0^t h(\vartheta)\,r(\vartheta)\exp\left(\int_\vartheta^t r(s)\,\mathrm{d}s\right)\mathrm{d}\vartheta.$$

Hence, by (53), (55), (61) and (56), we have

$$y(t)+q(t) \le z(t)+z_1(t) \;\le\; z(t)+\int_0^t h(\vartheta)\,r(\vartheta)\exp\left(\int_\vartheta^t r(s)\,\mathrm{d}s\right)\mathrm{d}\vartheta$$
$$= z(t)+\int_0^t z(\vartheta)\,r(\vartheta)\exp\left(\int_\vartheta^t r(s)\,\mathrm{d}s\right)\mathrm{d}\vartheta$$
$$-\;\int_0^t q(\vartheta)\,r(\vartheta)\exp\left(\int_\vartheta^t r(s)\,\mathrm{d}s\right),$$

which immediately yields inequality (54). □ □

Lemma 0.10 (Gronwall's modified lemma)**.** *Suppose that for all $t \in [0,T]$ we have*

$$\chi^2(t)+R(t) \le A(t)+2\int_0^t B(\vartheta)\chi(\vartheta)\,\mathrm{d}\vartheta, \tag{62}$$

where $R, A, B, \chi \in C([0,T])$ are nonnegative functions. Then for any $t \in [0,T]$

$$\sqrt{\chi^2(t)+R(t)} \le \max_{\vartheta\in[0,t]}\sqrt{A(\vartheta)}+\int_0^t B(\vartheta)\,\mathrm{d}\vartheta. \tag{63}$$

Proof. For any $\vartheta \in [0,T]$ we set

$$\varphi(\vartheta) = 2\int_0^\vartheta B(s)\,\chi(s)\,\mathrm{d}s.$$

Then $\varphi(0)=0$ and

$$\varphi'(\vartheta) = 2B(\vartheta)\,\chi(\vartheta). \tag{64}$$

Let us consider an arbitrary fixed $t \in [0,T]$ and denote

$$S_t = \max_{s\in[0,t]} A(s).$$

It is clear that if $S_t = 0$ for some $t \in [0,T]$, then $S_\tau = 0$ for all $\tau \in [0,t]$. Similarly, the condition $\varphi(\vartheta)=0$ for some $\vartheta \in [0,T]$ implies that $\varphi(\tau)=0$ for all $\tau \in [0,\vartheta]$. Let us set $t_1 = 0$, provided $S_t \ne 0$ for all $t \in [0,T]$, and

$$t_1 = \max\{t\in[0,T]; S_t = 0\},\ t_2 = \max\{\vartheta\in[0,T]; \varphi(\vartheta)=0\},\ t_3 = \min(t_1,t_2).$$

By (64) and (62),

$$\varphi'(\vartheta) \le 2B(\vartheta)\sqrt{S_t+\varphi(\vartheta)}.$$

Then for $t \in (t_3,T]$ we have

$$\int_{t_3}^t \frac{\varphi'(\vartheta)\,\mathrm{d}\vartheta}{2\sqrt{S_t+\varphi(\vartheta)}} \le \int_0^t B(\vartheta)\,\mathrm{d}\vartheta$$

and thus,

$$\sqrt{S_t+\varphi(\vartheta)}\Big|_{\vartheta=t_3}^t = \sqrt{S_t+\varphi(t)}-\sqrt{S_t} \le \int_0^t B(\vartheta)\,\mathrm{d}\vartheta.$$

This implies that

$$\sqrt{S_t + \varphi(t)} \le \sqrt{S_t} + \int_0^t B(\vartheta)\,\mathrm{d}\vartheta. \tag{65}$$

Now, by virtue of (62) and (65),

$$\sqrt{\chi^2(t) + R(t)} \le \sqrt{S_t + \varphi(t)} \le \sqrt{S_t} + \int_0^t B(\vartheta)\,\mathrm{d}\vartheta. \tag{66}$$

Taking into account that

$$\sqrt{S_t} = \sqrt{\max_{s\in[0,t]} A(s)} = \max_{s\in[0,t]} \sqrt{A(s)},$$

from (66) we immediately get (63). Finally, it is obvious that (63) also holds for all $t \in [0, t_3]$. □ □

Lemma 0.11 (Gronwall's discrete lemma)**.** *Let $x_m,\ b_m,\ c_m \ge 0$ and $a_m > 0$ for $m = 0, 1, 2, \ldots$, and let the sequence a_m be nondecreasing. Then, if*

$$\begin{aligned} x_0 + c_0 \le& a_0, \\ x_m + c_m \le& a_m + \sum_{j=0}^{m-1} b_j x_j \quad \text{for } m \ge 1, \end{aligned} \tag{67}$$

we have

$$x_m + c_m \le a_m \prod_{j=0}^{m-1} (1 + b_j) \quad \text{for } m \ge 0. \tag{68}$$

Proof. We start from inequality (67), divided by a_m, and use the assumption that the sequence a_m is nondecreasing. We get

$$\frac{x_m}{a_m} + \frac{c_m}{a_m} \le 1 + \sum_{j=0}^{m-1} b_j \frac{x_j}{a_m} \le 1 + \sum_{j=0}^{m-1} b_j \frac{x_j}{a_j}. \tag{69}$$

Let us set $v_0 = 1$ and $v_m = 1 + \sum_{j=0}^{m-1} b_j \frac{x_j}{a_j}$ for $m \ge 1$. Then by (67) and the inequality $c_{m-1}/a_{m-1} \ge 0$, we have

$$v_m - v_{m-1} = b_{m-1} \frac{x_{m-1}}{a_{m-1}} \le b_{m-1} \left(\frac{x_{m-1}}{a_{m-1}} + \frac{c_{m-1}}{a_{m-1}} \right) \le b_{m-1} v_{m-1}, \quad m \ge 1.$$

This implies that

$$v_m \le (1 + b_{m-1}) v_{m-1} \le v_0 \prod_{j=0}^{m-1} (1 + b_j) = \prod_{j=0}^{m-1} (1 + b_j).$$

Now from (69) we get (68). □ □

Chapter 1

DGM for elliptic problems

This chapter conserns in basic aspects of the discontinuous Galerkin method (DGM), which will be treated in an example of a simple problem for the Poisson equation with mixed Dirichlet–Neumann boundary conditions. We introduce the discretization of this problem with the aid of several variants of the DGM. Further, we prove the existence of the approximate solution and derive error estimates. Finally, several numerical examples are presented.

The book contains a detailed analysis of qualitative properties of DG techniques. It is based on a number of estimates with various constants. We denote by C_A, C_B, C_C, ..., C_a, C_b, C_c, ... positive constants arising in the formulation of results that can be simply named (e.g., $_A$ corresponds to approximation properties, $_B$ - boundedness, $_C$ - coercivity, etc.) Otherwise, we use symbols C, C_1, C_2, ... These constants are always independent of the parameters of the discretization (i.e., the space mesh-size h, time step τ in the case of nonstationary problems, and also the degree p of polynomial approximation in the case of the *hp*-methods), but they may depend on the data in problems. They are often "autonomous" in individual chapters or sections. Some constants are sometimes defined in a complicated way on the basis of a number of constants appearing in previous considerations. For an example, see Remark 2.13.

1.1 Model problem

Let Ω be a bounded domain in $\mathbb{R}^d$, $d = 2, 3$, with Lipschitz boundary $\partial\Omega$. We denote by $\partial\Omega_D$ and $\partial\Omega_N$ parts of the boundary $\partial\Omega$ such that $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, $\partial\Omega_D \cap \partial\Omega_N = \emptyset$ and $\partial\Omega_D \neq \emptyset$.

We consider the following model problem for the Poisson equation: Find a function $u : \Omega \to \mathbb{R}$ such that

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

$$u = u_D \quad \text{on } \partial\Omega_D, \tag{1.1b}$$

$$\boldsymbol{n} \cdot \nabla u = g_N \quad \text{on } \partial\Omega_N, \tag{1.1c}$$

where f, u_D and g_N are given functions. Let us note that $\boldsymbol{n} \cdot \nabla u = \frac{\partial u}{\partial \boldsymbol{n}}$ is the derivative of the function u in the direction $\boldsymbol{n}$, which is the outer unit normal to $\partial\Omega$. A function $u \in C^2(\overline{\Omega})$ satisfying (1.1) pointwise is called a *classical solution*. It is suitable to introduce a weak formulation of the above problem. Let us define the space

$$V = \{v \in H^1(\Omega);\ v|_{\partial\Omega_D} = 0\}.$$

Assuming thet u is a classical solution, we multiply (1.1a) by any function $v \in V$, integrate over Ω and use Green's theorem. Taking into account the boundary condition (1.1c), we obtain the identity

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega f\, v \,\mathrm{d}x + \int_{\partial\Omega_N} g_N\, v \,\mathrm{d}S \quad \forall\, v \in V. \tag{1.2}$$

We can introduce the following definition.

Definition 1.1. *Let us assume the existence of $u^* \in H^1(\Omega)$ such that $u^*|_{\partial\Omega_D} = u_D$ and let $f \in L^2(\Omega)$, $g_N \in L^2(\partial\Omega_N)$. Now we say that a function u is a* weak solution *of problem* (1.1)*, if*
(a) $u - u^* \in V$,
(b) u *satisfies identity (1.2).*

Using the Lax–Milgram Lemma 0.6, we can prove that there exists a unique weak solution of (1.1), see, e.g., [QV99, Section 6.1.2]. In the following, we shall deal with numerical solution of problem (1.1) with the aid of discontinuous piecewise polynomial approximations.

1.2 Abstract numerical method and its theoretical analysis

In order to better understand theoretical foundations of the DGM, we shall describe a possible general approach to deriving error estimates. (Readers familiar with concepts of a priori error estimates in the finite element method can skip this section.)

Let $u \in V$ be a weak solution of a given problem. Let V_h denote a *finite-dimensional space*, where an *approximate solution* u_h is sought. The subscript $h > 0$ (usually chosen as $h \in (0, \overline{h})$ with $\overline{h} > 0$) denotes the parameter of the discretization. Further, we introduce an infinitely dimensional function space W_h such that $V \subset W_h$ and $V_h \subset W_h$. (If $V_h \subset V$, then we usually put $W_h := V$ and thus, W_h is independent of h.) Finally, let $\|\cdot\|_{W_h}$ be a suitable norm in W_h. As we shall see later, the spaces V_h and W_h will be constructed over a suitable mesh in the computational domain, and hence the norm $\|\cdot\|_{W_h}$ may be mesh-dependent.

An *abstract numerical method* reads: Find $u_h \in V_h$ such that

$$A_h(u_h, v_h) = F(v_h) \quad \forall\, v_h \in V_h, \tag{1.3}$$

where $A_h : W_h \times W_h \to \mathbb{R}$ is a bilinear form and $F : W_h \to \mathbb{R}$ is a linear functional.

In the numerical analysis, we want to reach the following goals:

- the approximate solution u_h of (1.3) *exists* and is *unique*,
- the approximate solution u_h *converges* to the exact solution u in the $\|\cdot\|_{W_h}$-norm as $h \to 0$, i.e.,

$$\lim_{h\to 0} \|u - u_h\|_{W_h} = 0, \tag{1.4}$$

- *a priori error estimate*, i.e., we seek $\alpha > 0$ independent of h such that

$$\|u - u_h\|_{W_h} \le Ch^\alpha, \quad h \in (0, \overline{h}), \tag{1.5}$$

  where $C > 0$ is a constant, independent of h (but may depend on u), and α is the *order of convergence*.

Obviously, an a priori error estimate implies the convergence.

The existence and uniqueness of the approximate solution is a consequence of the *coercivity* of A_h, i.e., there exists $C_c > 0$ such that

$$A_h(v_h, v_h) \ge C_c \|v_h\|^2_{W_h} \quad \forall\, v_h \in V_h. \tag{1.6}$$

Then Corollary 0.7 implies the existence and uniqueness of the approximate solution u_h.

In order to derive a priori error estimates, we prove the *consistency* of the method,

$$A_h(u, v_h) = F(v_h) \quad \forall\, v_h \in V_h \tag{1.7}$$

which, together with (1.3), immediately gives the *Galerkin orthogonality* of the error $e_h = u_h - u$ to the space V_h:

$$A_h(e_h, v_h) = 0 \quad \forall\, v_h \in V_h. \tag{1.8}$$

Further, we introduce an *interpolation operator* (usually defined as a suitable *projection*) $\Pi_h : V \to V_h$ and prove its *approximation property*, namely existence of a constant $\alpha > 0$ such that

$$\|v - \Pi_h v\|_{W_h} \le \tilde{C}(v) h^\alpha \quad \forall\, v \in V, \quad h \in (0, \overline{h}), \tag{1.9}$$

where $\tilde{C}(v) > 0$ is a constant independent of h but dependent on v. A further step is the derivation of the inequality

$$A_h(u - \Pi_h u, v_h) \le R(u - \Pi_h u)\|v_h\|_{W_h} \quad \forall\, v_h \in V_h, \tag{1.10}$$

where R depends on suitable norms of the interpolation error $u - \Pi_h u$.

Finally, the *error estimate* is derived in the following way: for each $h \in (0, \overline{h})$ we decompose the error e_h by

$$e_h = u_h - u = \xi + \eta, \tag{1.11}$$

where $\xi := u_h - \Pi_h u \in V_h$ and $\eta := \Pi_h u - u \in W_h$. Putting $v_h := \xi$ in (1.8), we get

$$A_h(e_h, \xi) = A_h(\xi, \xi) + A_h(\eta, \xi) = 0. \tag{1.12}$$

It follows from the coercivity (1.6) and estimate (1.10) that

$$C_c \|\xi\|^2_{W_h} \le A_h(\xi, \xi) = -A_h(\eta, \xi) \le R(\eta)\|\xi\|_{W_h}, \tag{1.13}$$

which immediately implies the inequality

$$\|\xi\|_{W_h} \le \frac{R(\eta)}{C_c}. \tag{1.14}$$

Now, the triangle inequality, relations (1.11) and (1.14) give the error estimate in the form

$$\|e_h\|_{W_h} \le \|\xi\|_{W_h} + \|\eta\|_{W_h} \le \frac{R(\eta)}{C_c} + \|\eta\|_{W_h}. \tag{1.15}$$

This is often called the *abstract error estimate*, which represents an error bound in terms of the interpolation error η.

The last aim is to use the approximation property (1.9) of the operator Π_h and to estimate the expression $R(\eta)$ in terms of the mesh-size h in the form

$$R(\eta) \le \tilde{C}_1(u) h^\alpha, \tag{1.16}$$

which together with (1.15) immediately imply the *error estimate*

$$\|e_h\|_{W_h} \le \left(C_c^{-1} \tilde{C}_1(u) + \tilde{C}(u) \right) h^\alpha, \tag{1.17}$$

valid for all $h \in (0, \overline{h})$. We say that the numerical scheme has the *order of convergence* in the norm $\|\cdot\|_{W_h}$ equal to α.

This concept of numerical analysis is applied in this chapter. (Among other, we specify there the spaces W_h and V_h.) For time dependent problems, treated in Chapters 2 – 4, the analysis is more complicated and the previous technique has to be modified. However, in some parts of the book, error estimates are derived in a different way.

Remark 1.2. *As was mentioned above, we are interested here in deriving of* a priori error estimates *(simply called* error estimates*). We shall not deal with* a posteriori error estimates*, when the error is bounded in a suitable norm in terms of the approximate solution and data of the problem. The subject of a posteriori error estimates plays an important role in practical computations, but is out of the scope of this book. For some results in this direction for the DGM we can refer, e.g., to the papers [AEV11], [DEV13], [EV10], [GHH07], [HH06b], [HSW08], [JSV10] and the references cited therein.*

1.3 Spaces of discontinuous functions

The subject of this section is the construction of DG space partitions of the bounded computational domain Ω and the specification of their properties which are used in the theoretical analysis. Further, function spaces over these meshes are defined.

1.3.1 Partition of the domain

Let $\mathcal{T}_h$ ($h > 0$ is a parameter) be a partition of the closure $\overline{\Omega}$ of the domain Ω into a finite number of closed d-dimensional simplexes K with mutually disjoint interiors such that

$$\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} K. \tag{1.18}$$

This assumption means that the domain Ω is polygonal (if $d = 2$) or polyhedral (if $d = 3$). The case of a 2D nonpolygonal domain is considered, e.g., in [Sob11], where curved elements are used. See also Chapter 6, where curved elements are treated from the implementation point of view. We call $\mathcal{T}_h$ a *triangulation* of Ω and do not require the standard conforming properties from the finite element method, introduced e.g., in [Cia79], [BS94b], [EEHJ96], [Sch00] or [Žen90]. In two-dimensional problems ($d = 2$) we choose $K \in \mathcal{T}_h$ as triangles and in three-dimensional problems ($d = 3$) the elements $K \in \mathcal{T}_h$ are tetrahedra. As we see, we admit that in the finite element mesh the so-called *hanging nodes* (and in 3D also *hanging edges*) appear; see Figure 1.1.

In general, the discontinuous Galerkin method can handle with more general elements as quadrilaterals and convex or even nonconvex star-shaped polygons in 2D and hexahedra, pyramids and convex or nonconvex star-shaped polyhedra in 3D. As an example, we can consider the so-called dual finite volumes constructed over triangular ($d = 2$) or tetrahedral ($d = 3$) meshes (cf., e.g., [FFLMW99]). A use of such elements will be discussed in Section **??**.

In our further considerations we shall use the following notation. By ∂K we denote the boundary of an element $K \in \mathcal{T}_h$ and set $h_K = \mathrm{diam}(K)$ = diameter of K, $h = \max_{K \in \mathcal{T}_h} h_K$. 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.

Let $K, K' \in \mathcal{T}_h$. We say that K and K' are *neighbouring elements* (or simply *neighbours*) if the set $\partial K \cap \partial K'$ has positive $(d-1)$-dimensional measure. We say that $\Gamma \subset K$ is a *face* of K, if it is a maximal connected open subset of either $\partial K \cap \partial K'$, where K' is a neighbour of K, or $\partial K \cap \partial \Omega_D$ or $\partial K \cap \partial \Omega_N$. The symbol $|\Gamma|$ will denote the $(d-1)$-dimensional Lebesgue measure

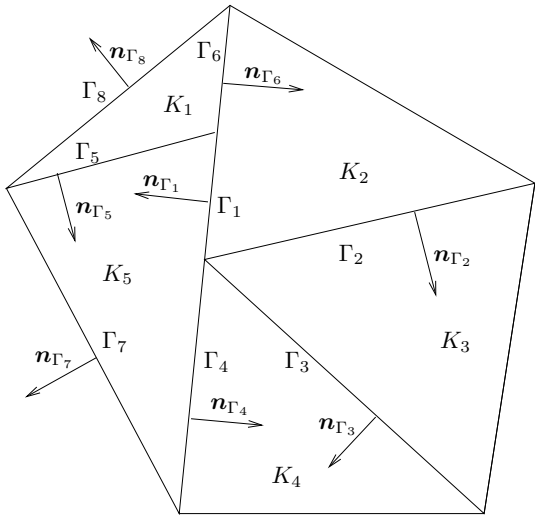

Figure 1.1: Example of elements $K_l,\ l = 1, \ldots, 5$, and faces $\Gamma_l,\ l = 1, \ldots, 8$, with the corresponding normals $\boldsymbol{n}_{\Gamma_l}$. The triangle K_5 has a hanging node. Its boundary is formed by four edges: $\partial K_5 = \Gamma_1 \cup \Gamma_4 \cup \Gamma_7 \cup \Gamma_5$.

of Γ. Hence, if $d = 2$, then $|\Gamma|$ is the length of Γ and for $d = 3$, $|\Gamma|$ denotes the area of Γ. By $\mathcal{F}_h$ we denote the system of all faces of all elements $K \in \mathcal{T}_h$. Further, we define the set of all boundary faces by

$$\mathcal{F}_h^B = \{\Gamma \in \mathcal{F}_h;\ \Gamma \subset \partial\Omega\},$$

the set of all "Dirichlet" boundary faces by

$$\mathcal{F}_h^D = \{\Gamma \in \mathcal{F}_h;\ \Gamma \subset \partial\Omega_D\},$$

the set of all "Neumann" boundary faces by

$$\mathcal{F}_h^N = \{\Gamma \in \mathcal{F}_h,\ \Gamma \subset \partial\Omega_N\}$$

and the set of all inner faces

$$\mathcal{F}_h^I = \mathcal{F}_h \setminus \mathcal{F}_h^B.$$

Obviously, $\mathcal{F}_h = \mathcal{F}_h^I \cup \mathcal{F}_h^D \cup \mathcal{F}_h^N$ and $\mathcal{F}_h^B = \mathcal{F}_h^D \cup \mathcal{F}_h^N$. For a shorter notation we put

$$\mathcal{F}_h^{ID} = \mathcal{F}_h^I \cup \mathcal{F}_h^D.$$

For each $\Gamma \in \mathcal{F}_h$ we define a unit normal vector $\boldsymbol{n}_\Gamma$. We assume that for $\Gamma \in \mathcal{F}_h^B$ the normal $\boldsymbol{n}_\Gamma$ has the same orientation as the outer normal to $\partial\Omega$. For each face $\Gamma \in \mathcal{F}_h^I$ the orientation of $\boldsymbol{n}_\Gamma$ is arbitrary but fixed. See Figure 1.1.

For each $\Gamma \in \mathcal{F}_h^I$ there exist two neighbouring elements $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathcal{T}_h$ such that $\Gamma \subset \partial K_\Gamma^{(L)} \cap \partial K_\Gamma^{(R)}$. (This means that the elements $K_\Gamma^{(L)}, K_\Gamma^{(R)}$ are adjacent to Γ and they share this face.) We use the convention that $\boldsymbol{n}_\Gamma$ is the outer normal to $\partial K_\Gamma^{(L)}$ and the inner normal to $\partial K_\Gamma^{(R)}$; see Figure 1.2.

Moreover, if $\Gamma \in \mathcal{F}_h^B$, then there exists an element $K_\Gamma^{(L)} \in \mathcal{T}_h$ such that $\Gamma \subset K_\Gamma^{(L)} \cap \partial\Omega$.

1.3.2 Assumptions on meshes

Let us consider a system $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$, $\bar{h} > 0$, of triangulations of the domain Ω ($\mathcal{T}_h = \{K\}_{K\in\mathcal{T}_h}$). In our further considerations we shall meet various assumptions on triangulations. The first is usual in the theory of the finite element method:

- The system $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ of triangulations is *shape-regular*: there exists a positive constant C_R such that

$$\frac{h_K}{\rho_K} \le C_R \quad \forall\, K \in \mathcal{T}_h\ \forall\, h \in (0, \bar{h}). \tag{1.19}$$

Moreover, for each face $\Gamma \in \mathcal{F}_h,\ h \in (0, \bar{h})$, we need to introduce a quantity $h_\Gamma > 0$, which represents a "one-dimensional" size of the face Γ. We require that

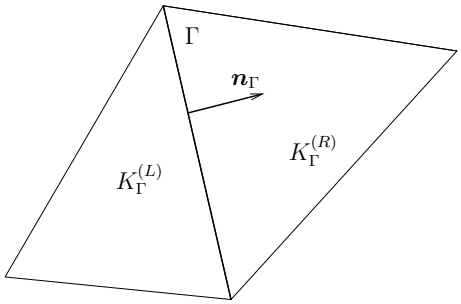

Figure 1.2: Interior face Γ, elements $K_\Gamma^{(L)}$ and $K_\Gamma^{(R)}$ and the orientation of $\boldsymbol{n}_\Gamma$.

- the quantity $h_\Gamma,\ \Gamma \in \mathcal{F}_h,\ h \in (0, \bar{h})$, satisfy the *equivalence condition* with h_K, i.e., there exist constants $C_T, C_G > 0$ independent of h, K and Γ such that

$$C_T h_K \le h_\Gamma \le C_G h_K, \quad \forall\, K \in \mathcal{T}_h,\ \forall\, \Gamma \in \mathcal{F}_h,\ \Gamma \subset \partial K,\ \forall\, h \in (0, \bar{h}). \tag{1.20}$$

The equivalence condition can be fulfilled by additional assumptions on the system of triangulations $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ and by a suitable choice of the quantity $h_\Gamma,\ \Gamma \in \mathcal{F}_h,\ h \in (0, \bar{h})$. We introduce some assumptions on triangulations and several choices of the quantity h_Γ. Then we discuss how the equivalence condition (1.20) is satisfied.

In literature we can find the following assumptions on the system of triangulations:

(MA1) The system $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ is *locally quasi-uniform*: there exists a constant $C_Q > 0$ such that

$$h_K \le C_Q h_{K'} \quad \forall\, K,\, K' \in \mathcal{T}_h,\ K,\, K' \text{ are neighbours},\ \forall\, h \in (0, \bar{h}). \tag{1.21}$$

(MA2) The faces $\Gamma \subset \partial K$ do not degenerate with respect to the diameter of K if $h \to 0$: there exists a constant $C_d > 0$ such that

$$h_K \le C_d \mathrm{diam}(\Gamma) \quad \forall\, K \in \mathcal{T}_h \ \ \forall\, \Gamma \in \mathcal{F}_h,\ \ \Gamma \subset \partial K,\ \ \forall\, h \in (0, \bar{h}). \tag{1.22}$$

(MA3) The system $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ is *quasi-uniform*: there exists a constant $C_U > 0$ such that

$$h \le C_U h_K \quad \forall\, K \in \mathcal{T}_h \ \ \forall\, h \in (0, \bar{h}). \tag{1.23}$$

(MA4) The triangulations $\mathcal{T}_h,\ h \in (0, \bar{h})$, are *conforming*. This means that for two elements $K,\, K' \in \mathcal{T}_h,\ K \ne K'$, either $K \cap K' = \emptyset$ or $K \cap K'$ is a common vertex or $K \cap K'$ is a common face (or for $d = 3$, when $K \cap K'$ is a common edge) of K and K'.

If condition (MA4) is not satisfied, then the triangulations $\mathcal{T}_h$ are called *nonconforming*.

Remark 1.3. *There are some relations among the mesh assumptions (MA1) – (MA4) mentioned above. Obviously, (MA3) $\Rightarrow$ (MA1) . Moreover, if the system of triangulation is shape-regular (i.e.,* (1.19) *is fulfilled) then (MA4) $\Rightarrow$ (MA1) & (MA2) .*

Exercise 1.4. *Prove the implications in Remark 1.3.*

Concerning the choice of the quantity $h_\Gamma,\ \Gamma \in \mathcal{F}_h,\ h \in (0, \bar{h})$, in literature we can find the following basic possibilities:

$$\text{(i)} \qquad h_\Gamma = \mathrm{diam}(\Gamma), \quad \Gamma \in \mathcal{F}_h^{ID}, \tag{1.24}$$

$$\text{(ii)} \qquad h_\Gamma = \begin{cases} \frac{1}{2}\left(h_{K_\Gamma^{(L)}} + h_{K_\Gamma^{(R)}}\right) & \text{for } \Gamma \in \mathcal{F}_h^I \\ h_{K_\Gamma^{(L)}} & \text{for } \Gamma \in \mathcal{F}_h^B, \end{cases} \tag{1.25}$$

$$\text{(iii)} \qquad h_\Gamma = \begin{cases} \max\left(h_{K_\Gamma^{(L)}},\ h_{K_\Gamma^{(R)}}\right) & \text{for } \Gamma \in \mathcal{F}_h^I \\ h_{K_\Gamma^{(L)}} & \text{for } \Gamma \in \mathcal{F}_h^B, \end{cases} \tag{1.26}$$

$$\text{(iv)} \qquad h_\Gamma = \begin{cases} \min\left(h_{K_\Gamma^{(L)}},\ h_{K_\Gamma^{(R)}}\right) & \text{for } \Gamma \in \mathcal{F}_h^I \\ h_{K_\Gamma^{(L)}} & \text{for } \Gamma \in \mathcal{F}_h^B, \end{cases} \tag{1.27}$$

where $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathcal{T}_h$ are the elements adjacent to $\Gamma \in \mathcal{F}_h^I$, see Figure 1.2, and $K_\Gamma^{(L)} \in \mathcal{T}_h$ is the element adjacent to $\Gamma \in \mathcal{F}_h^B$.

The following lemma characterizes assumptions on computational grids and the choice of h_Γ, which guarantee the equivalence condition (1.20).

Lemma 1.5. *Let $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ be a system of triangulations of the domain Ω satisfying the shape-regularity assumption* (1.19)*. Then the equivalence condition* (1.20) *is satisfied in the following cases:*

(i) The triangulations $\mathcal{T}_h$, $h \in (0,\bar{h})$, are conforming (i.e., assumption (MA4) is satisfied) and h_Γ are defined by (1.24) *or* (1.25) *or* (1.26) *or* (1.27) *.*

(ii) The triangulations $\mathcal{T}_h$, $h \in (0,\bar{h})$, are, in general, nonconforming; assumption (MA2) (i.e., (1.22)*) is satisfied and h_Γ are defined by* (1.24)*.*

(iii) The triangulations $\mathcal{T}_h$, $h \in (0,\bar{h})$, are, in general, nonconforming; assumption (MA1) is satisfied (i.e., the system $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ is locally quasi-uniform) and h_Γ are defined by (1.25) *or* (1.26) *or* (1.27)*.*

Exercise 1.6. *Prove the above lemma and find the constants C_T and C_G. For example, in the case (c), when h_Γ is given by (1.25), we have*

$$C_T = (1 + C_Q^{-1})/2, \quad C_G = (1 + C_Q)/2, \tag{1.28}$$

where C_Q is the constant from the local quasi-uniformity condition (1.21).

1.3.3 Broken Sobolev spaces

The discontinuous Galerkin method is based on the use of discontinuous approximations. This is the reason that over a triangulation $\mathcal{T}_h$, for any $k \in \mathbb{N}$, we define the so-called *broken Sobolev space*

$$H^k(\Omega,\mathcal{T}_h) = \{v \in L^2(\Omega); v|_K \in H^k(K) \ \forall\ K \in \mathcal{T}_h\}, \tag{1.29}$$

which consists of functions, whose restrictions on $K \in \mathcal{T}_h$ belong to the Sobolev space $H^k(K)$. On the other hand, functions from $H^k(\Omega,\mathcal{T}_h)$ are, in general, discontinuous on inner faces of elements $K \in \mathcal{T}_h$. For $v \in H^k(\Omega,\mathcal{T}_h)$, we define the norm

$$\|v\|_{H^k(\Omega,\mathcal{T}_h)} = \left(\sum_{K\in\mathcal{T}_h} \|v\|^2_{H^k(K)}\right)^{1/2} \tag{1.30}$$

and the seminorm

$$|v|_{H^k(\Omega,\mathcal{T}_h)} = \left(\sum_{K\in\mathcal{T}_h} |v|^2_{H^k(K)}\right)^{1/2}. \tag{1.31}$$

Let $\Gamma \in \mathcal{F}_h^I$ and let $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathcal{T}_h$ be elements adjacent to Γ. For $v \in H^1(\Omega,\mathcal{T}_h)$ we introduce the following notation:

$$\begin{aligned}
v_\Gamma^{(L)} &= \text{the trace of } v|_{K_\Gamma^{(L)}} \text{ on } \Gamma, \\
v_\Gamma^{(R)} &= \text{the trace of } v|_{K_\Gamma^{(R)}} \text{ on } \Gamma, \\
\langle v\rangle_\Gamma &= \frac{1}{2}\left(v_\Gamma^{(L)} + v_\Gamma^{(R)}\right) \quad \text{(mean value of the traces of } v \text{ on } \Gamma), \\
[v]_\Gamma &= v_\Gamma^{(L)} - v_\Gamma^{(R)} \quad \text{(jump of } v \text{ on } \Gamma).
\end{aligned} \tag{1.32}$$

The value $[v]_\Gamma$ depends on the orientation of $\boldsymbol{n}_\Gamma$, but $[v]_\Gamma \boldsymbol{n}_\Gamma$ is independent of this orientation.

Moreover, let $\Gamma \in \mathcal{F}_h^B$ and $K_\Gamma^{(L)} \in \mathcal{T}_h$ be the element such that $\Gamma \subset \partial K_\Gamma^{(L)} \cap \partial\Omega$. Then for $v \in H^1(\Omega,\mathcal{T}_h)$ we introduce the following notation:

$$\begin{aligned}
v_\Gamma^{(L)} &= \text{the trace of } v|_{K_\Gamma^{(L)}} \text{ on } \Gamma, \\
\langle v\rangle_\Gamma &= [v]_\Gamma = v_\Gamma^{(L)}.
\end{aligned} \tag{1.33}$$

If $\Gamma \in \mathcal{F}_h^B$, then by $v_\Gamma^{(R)}$ we formally denote the exterior trace of v on Γ given either by a boundary condition or by an extrapolation from the interior of Ω.

In case that $\Gamma \in \mathcal{F}_h$ and $[\cdot]_\Gamma$, $\langle\cdot\rangle_\Gamma$ and $\boldsymbol{n}_\Gamma$ appear in integrals $\int_\Gamma \ldots \mathrm{d}S$, then we usually omit the subscript Γ and simply write $[\cdot]$, $\langle\cdot\rangle$ and $\boldsymbol{n}$, respectively.

The discontinuous Galerkin method can be characterized as a finite element technique using piecewise polynomial approximations, in general discontinuous on interfaces between neighbouring elements. Therefore, we introduce a finite-dimensional subspace of $H^k(\Omega,\mathcal{T}_h)$, where the approximate solution will be sought.

Let $\mathcal{T}_h$ be a triangulation of Ω introduced in Section 1.3.1 and let $p \geq 0$ be an integer. We define the space of discontinuous piecewise polynomial functions

$$S_{hp} = \{v \in L^2(\Omega); v|_K \in P_p(K) \ \forall\, K \in \mathcal{T}_h\}, \tag{1.34}$$

where $P_p(K)$ denotes the space of all polynomials of degree $\leq p$ on K. We call the number p the *degree of polynomial approximation*. Obviously, $S_{hp} \subset H^k(\Omega, \mathcal{T}_h)$ for any $k \geq 1$ and its dimension $\dim S_{hp} < \infty$.

1.4 DGM based on a primal formulation

In this section we shall introduce the so-called discontinuous Galerkin method (DGM) based on a *primal formulation* for the solution of problem (1.1). The approximate solution will be sought in the space $S_{hp} \subset H^1(\Omega, \mathcal{T}_h)$. In contrast to the standard (conforming) finite element method, the weak formulation (1.2) given in Section 1.1 is not suitable for the derivation of the DGM, because (1.2) does not make sense for $u \in H^1(\Omega, \mathcal{T}_h) \not\subset H^1(\Omega)$. Therefore, we shall introduce a "weak form of (1.1) in the sense of broken Sobolev spaces".

Let us assume that u is a sufficiently regular solution of (1.1), namely, let $u \in H^2(\Omega)$. Then we speak of a *strong solution*. In deriving the DGM we proceed in the following way. We multiply (1.1a) by a function $v \in H^1(\Omega, \mathcal{T}_h)$, integrate over $K \in \mathcal{T}_h$ and use Green's theorem. Summing over all $K \in \mathcal{T}_h$, we obtain the identity

$$\sum_{K\in\mathcal{T}_h} \int_K \nabla u \cdot \nabla v \,\mathrm{d}x - \sum_{K\in\mathcal{T}_h} \int_{\partial K} (\boldsymbol{n}_K \cdot \nabla u)\, v \,\mathrm{d}S = \int_\Omega f\, v \,\mathrm{d}x, \tag{1.35}$$

where $\boldsymbol{n}_K$ denotes the outer unit normal to ∂K. The surface integrals over ∂K make sense due to the regularity of u. (Since $u \in H^2(K)$, the derivatives $\partial u/\partial x_i$ have the trace on ∂K and $\partial u/\partial x_i|_{\partial K} \in L^2(\partial K)$ for $i = 1, \ldots, d$; see Theorem 0.1 on traces.) We rewrite the surface integrals over ∂K according to the type of faces $\Gamma \in \mathcal{F}_h$ that form the boundary of the element $K \in \mathcal{T}_h$:

$$\begin{aligned}\sum_{K\in\mathcal{T}_h} \int_{\partial K} (\boldsymbol{n}_K \cdot \nabla u)\, v \,\mathrm{d}S = &\sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma (\boldsymbol{n}_\Gamma \cdot \nabla u)\, v \,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^N} \int_\Gamma (\boldsymbol{n}_\Gamma \cdot \nabla u)\, v \,\mathrm{d}S \\ &+ \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \boldsymbol{n}_\Gamma \cdot \left((\nabla u_\Gamma^{(L)})\, v_\Gamma^{(L)} - (\nabla u_\Gamma^{(R)})\, v_\Gamma^{(R)} \right) \mathrm{d}S.\end{aligned} \tag{1.36}$$

(There is the sign "$-$" in the last integral, since $\boldsymbol{n}_\Gamma$ is the outer unit normal to $\partial K_\Gamma^{(L)}$ but the inner unit normal to $\partial K_\Gamma^{(R)}$, see Section 1.3.1 or Figure 1.2.)

Due to the assumption that $u \in H^2(\Omega)$, we have

$$[u]_\Gamma = [\nabla u]_\Gamma = 0, \quad \nabla u_\Gamma^{(L)} = \nabla u_\Gamma^{(R)} = \langle \nabla u \rangle_\Gamma, \quad \Gamma \in \mathcal{F}_h^I. \tag{1.37}$$

Thus, the integrand of the last integral in (1.36) can be written in the form

$$\boldsymbol{n}_\Gamma \cdot (\nabla u)_\Gamma^{(L)}\, v_\Gamma^{(L)} - \boldsymbol{n}_\Gamma \cdot (\nabla u)_\Gamma^{(R)}\, v_\Gamma^{(R)} = \boldsymbol{n}_\Gamma \cdot \langle \nabla u \rangle_\Gamma\, [v]_\Gamma. \tag{1.38}$$

By virtue of the Neumann boundary condition (1.1c),

$$\sum_{\Gamma\in\mathcal{F}_h^N} \int_\Gamma (\boldsymbol{n}_\Gamma \cdot \nabla u) v \,\mathrm{d}S = \int_{\partial\Omega_N} g_N v \,\mathrm{d}S. \tag{1.39}$$

Now, (1.33) and (1.35) – (1.39) imply that

$$\begin{aligned}&\sum_{K\in\mathcal{T}_h} \int_K \nabla u \cdot \nabla v \,\mathrm{d}x - \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \boldsymbol{n} \cdot \langle \nabla u \rangle\, [v] \,\mathrm{d}S - \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \boldsymbol{n} \cdot \nabla u\, v \,\mathrm{d}S \\ &\quad = \sum_{K\in\mathcal{T}_h} \int_K \nabla u \cdot \nabla v \,\mathrm{d}x - \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \boldsymbol{n} \cdot \langle \nabla u \rangle\, [v] \,\mathrm{d}S \\ &\quad = \int_\Omega f\, v \,\mathrm{d}x + \int_{\partial\Omega_N} g_N\, v \,\mathrm{d}S, \qquad v \in H^1(\Omega, \mathcal{T}_h).\end{aligned} \tag{1.40}$$

Here and in what follows, in integrals over Γ the symbol $\boldsymbol{n}$ means $\boldsymbol{n}_\Gamma$.

Relation (1.40) is the basis of the DG discretization of problem (1.1). However, in order to guarantee the existence of the approximate solution and its convergence to the exact one, some additional terms have to be included in the DG formulation.

In order to mimic the continuity of the approximate solution in a weaker sense, we define the *interior and boundary penalty bilinear form*

$$J_h^\sigma(u,v) = \sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma \sigma[u]\,[v]\,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^D}\int_\Gamma \sigma u\,v\,\mathrm{d}S \tag{1.41}$$
$$= \sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma[u]\,[v]\,\mathrm{d}S, \qquad u,v\in H^1(\Omega,\mathcal{T}_h).$$

The boundary penalty is associated with the boundary linear form

$$J_D^\sigma(v) = \sum_{\Gamma\in\mathcal{F}_h^D}\int_\Gamma \sigma u_D\,v\,\mathrm{d}S. \tag{1.42}$$

Here $\sigma>0$ is a penalty weight. Its choice will be discussed in Section 1.6. Obviously, for the exact strong solution $u\in H^2(\Omega)$,

$$J_h^\sigma(u,v) = J_D^\sigma(v) \qquad \forall\, v\in H^1(\Omega,\mathcal{T}_h), \tag{1.43}$$

since $[u]_\Gamma=0$ for $\Gamma\in\mathcal{F}_h^I$ and $[u]_\Gamma=u_\Gamma=u_D$ for $\Gamma\in\mathcal{F}_h^D$.

The interior penalty replaces the continuity of the approximate solution on interior faces, which is required in the standard conforming finite element method. The boundary penalty introduces the Dirichlet boundary condition in the discrete problem.

Moreover, the left-hand side of (1.40) is not symmetric with respect to u and v. In the theoretical analysis, it is advantageous to have some type of symmetry. Hence, it is desirable to include some additional term, which "symmetrizes" the left-hand side of (1.40) and which vanishes for the exact solution. Therefore, let $u\in H^1(\Omega)\cap H^2(\Omega,\mathcal{T}_h)$ be a function which satisfies the Dirichlet boundary condition (1.1b). Then we use the identity

$$\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \boldsymbol{n}\cdot\langle\nabla v\rangle\,[u]\,\mathrm{d}S = \sum_{\Gamma\in\mathcal{F}_h^D}\int_\Gamma \boldsymbol{n}\cdot\nabla v\,u_D\,\mathrm{d}S \quad \forall\, v\in H^2(\Omega,\mathcal{T}_h), \tag{1.44}$$

which is valid since $[u]_\Gamma=0$ for $\Gamma\in\mathcal{F}_h^I$, $[u]_\Gamma=u_\Gamma=u_D$ for $\Gamma\in\mathcal{F}_h^D$ and $\langle\nabla v\rangle_\Gamma=\nabla v_\Gamma$ for $\Gamma\in\mathcal{F}_h^D$ by definition.

Now, without a deeper motivation, we introduce five variants of the *discontinuous Galerkin weak formulation*. Each particular method is commented on in Remark 1.10. Hence, we sum identity (1.40) with -1, 1 or 0-multiple of (1.44) and possibly add equality (1.43). This leads us to the following notation. For $u,v\in H^2(\Omega,\mathcal{T}_h)$ we introduce the bilinear *diffusion forms*

$$a_h^{\mathrm{s}}(u,v) = \sum_{K\in\mathcal{T}_h}\int_K \nabla u\cdot\nabla v\,\mathrm{d}x - \sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \left(\boldsymbol{n}\cdot\langle\nabla u\rangle\,[v] + \boldsymbol{n}\cdot\langle\nabla v\rangle\,[u]\right)\mathrm{d}S, \tag{1.45a}$$

$$a_h^{\mathrm{n}}(u,v) = \sum_{K\in\mathcal{T}_h}\int_K \nabla u\cdot\nabla v\,\mathrm{d}x - \sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \left(\boldsymbol{n}\cdot\langle\nabla u\rangle\,[v] - \boldsymbol{n}\cdot\langle\nabla v\rangle\,[u]\right)\mathrm{d}S, \tag{1.45b}$$

$$a_h^{\mathrm{i}}(u,v) = \sum_{K\in\mathcal{T}_h}\int_K \nabla u\cdot\nabla v\,\mathrm{d}x - \sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \boldsymbol{n}\cdot\langle\nabla u\rangle\,[v]\,\mathrm{d}S, \tag{1.45c}$$

and the right-hand side linear forms

$$F_h^{\mathrm{s}}(v) = \int_\Omega f\,v\,\mathrm{d}x + \sum_{\Gamma\in\mathcal{F}_h^N}\int_\Gamma g_N\,v\,\mathrm{d}S - \sum_{\Gamma\in\mathcal{F}_h^D}\int_\Gamma \boldsymbol{n}\cdot\nabla v\,u_D\,\mathrm{d}S, \tag{1.46a}$$

$$F_h^{\mathrm{n}}(v) = \int_\Omega f\,v\,\mathrm{d}x + \sum_{\Gamma\in\mathcal{F}_h^N}\int_\Gamma g_N\,v\,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^D}\int_\Gamma \boldsymbol{n}\cdot\nabla v\,u_D\,\mathrm{d}S, \tag{1.46b}$$

$$F_h^{\mathrm{i}}(v) = \int_\Omega f\,v\,\mathrm{d}x + \sum_{\Gamma\in\mathcal{F}_h^N}\int_\Gamma g_N\,v\,\mathrm{d}S. \tag{1.46c}$$

Moreover, for $u,v\in H^2(\Omega,\mathcal{T}_h)$ let us define the bilinear forms

$$A_h^{\mathrm{s}}(u,v) = a_h^{\mathrm{s}}(u,v), \tag{1.47a}$$
$$A_h^{\mathrm{n}}(u,v) = a_h^{\mathrm{n}}(u,v), \tag{1.47b}$$
$$A_h^{\mathrm{s},\sigma}(u,v) = a_h^{\mathrm{s}}(u,v) + J_h^\sigma(u,v), \tag{1.47c}$$
$$A_h^{\mathrm{n},\sigma}(u,v) = a_h^{\mathrm{n}}(u,v) + J_h^\sigma(u,v), \tag{1.47d}$$
$$A_h^{\mathrm{i},\sigma}(u,v) = a_h^{\mathrm{i}}(u,v) + J_h^\sigma(u,v), \tag{1.47e}$$

and the linear forms

$$\ell_h^{\mathrm{s}}(v) = F_h^{\mathrm{s}}(v), \tag{1.48a}$$
$$\ell_h^{\mathrm{n}}(v) = F_h^{\mathrm{n}}(v), \tag{1.48b}$$
$$\ell_h^{\mathrm{s},\sigma}(v) = F_h^{\mathrm{s}}(v) + J_D^{\sigma}(v), \tag{1.48c}$$
$$\ell_h^{\mathrm{n},\sigma}(v) = F_h^{\mathrm{n}}(v) + J_D^{\sigma}(v), \tag{1.48d}$$
$$\ell_h^{\mathrm{i},\sigma}(v) = F_h^{\mathrm{i}}(v) + J_D^{\sigma}(v). \tag{1.48e}$$

Since $S_{hp} \subset H^2(\Omega, \mathcal{T}_h)$, the forms (1.47) make sense for $u_h, v_h \in S_{hp}$. Consequently, we define five numerical schemes.

Definition 1.7. *A function $u_h \in S_{hp}$ is called a* DG approximate solution *of problem* (1.1)*, if it satisfies one of the following identities:*

$$\textit{(i)} \quad A_h^{\mathrm{s}}(u_h, v_h) = \ell_h^{\mathrm{s}}(v_h) \quad \forall\, v_h \in S_{hp}, \tag{1.49a}$$
$$\textit{(ii)} \quad A_h^{\mathrm{n}}(u_h, v_h) = \ell_h^{\mathrm{n}}(v_h) \quad \forall\, v_h \in S_{hp}, \tag{1.49b}$$
$$\textit{(iii)} \quad A_h^{\mathrm{s},\sigma}(u_h, v_h) = \ell_h^{\mathrm{s},\sigma}(v_h) \quad \forall\, v_h \in S_{hp}, \tag{1.49c}$$
$$\textit{(iv)} \quad A_h^{\mathrm{n},\sigma}(u_h, v_h) = \ell_h^{\mathrm{n},\sigma}(v_h) \quad \forall\, v_h \in S_{hp}, \tag{1.49d}$$
$$\textit{(v)} \quad A_h^{\mathrm{i},\sigma}(u_h, v_h) = \ell_h^{\mathrm{i},\sigma}(v_h) \quad \forall\, v_h \in S_{hp}, \tag{1.49e}$$

where the forms $A_h^{\mathrm{s}}, A_h^{\mathrm{n}}, \ldots$, and $\ell_h^{\mathrm{s}}, \ell_h^{\mathrm{n}}, \ldots$, are defined by (1.47) and (1.48), respectively.

The diffusion forms $a_h^{\mathrm{s}}, a_h^{\mathrm{n}}, a_h^{\mathrm{i}}$ defined by (1.45) can be simply written in the form

$$a_h(u,v) = \sum_{K\in\mathcal{T}_h} \int_K \nabla u \cdot \nabla v \,\mathrm{d}x - \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \left(\boldsymbol{n}\cdot\langle\nabla u\rangle\,[v] + \Theta\boldsymbol{n}\cdot\langle\nabla v\rangle\,[u]\right)\mathrm{d}S, \tag{1.50}$$

where $\Theta = 1$ in the case of the form a_h^{s}, $\Theta = -1$ for a_h^{n} and $\Theta = 0$ for a_h^{i} and the bilinear forms A_h^{s}, A_h^{n}, $A_h^{\mathrm{s},\sigma}$, $A_h^{\mathrm{n},\sigma}$ and $A_h^{\mathrm{i},\sigma}$ defined by (1.47) can be written in the form

$$A_h(u,v) = a_h(u,v) + \vartheta J_h^{\sigma}(u,v), \tag{1.51}$$

where $\vartheta = 0$ for A_h^{s} and A_h^{n} and $\vartheta = 1$ for $A_h^{\mathrm{s},\sigma}$, $A_h^{\mathrm{n},\sigma}$ and $A_h^{\mathrm{i},\sigma}$.

Similarly we can write

$$F_h(v) = \int_\Omega f\,v\,\mathrm{d}x + \sum_{\Gamma\in\mathcal{F}_h^N}\int_\Gamma g_N\,v\,\mathrm{d}S - \Theta\sum_{\Gamma\in\mathcal{F}_h^D}\int_\Gamma \boldsymbol{n}\cdot\nabla v\,u_D\,\mathrm{d}S, \tag{1.52}$$

with $\Theta = 1$ for F_h^{s}, $\Theta = -1$ for F_h^{n} and $\Theta = 0$ for F_h^{i}, and then the right-hand side form reads

$$\ell_h(v) = F_h(v) + \vartheta J_D^{\sigma}(v), \tag{1.53}$$

where $\vartheta = 0$ for ℓ_h^{s} and ℓ_h^{n} and $\vartheta = 1$ for $\ell_h^{\mathrm{s},\sigma}$, $\ell_h^{\mathrm{n},\sigma}$ and $\ell_h^{\mathrm{i},\sigma}$.

The form a_h^{n} ($\Theta = -1$), a_h^{i} ($\Theta = 0$) and a_h^{s} ($\Theta = 1$) represents the so-called *nonsymmetric*, *incomplete* and *symmetric* variant of the diffusion discretization, respectively.

If we denote by A_h any form defined by (1.47) and by ℓ_h, we denote the form defined by (1.53), i.e., any form given by (1.48), the *discrete problem* (1.49) can be formulated to find $u_h \in S_{hp}$ satisfying the identity

$$A_h(u_h, v_h) = \ell_h(v_h) \quad \forall\, v_h \in S_{hp}. \tag{1.54}$$

The discrete problem (1.54) is equivalent to a system of linear algebraic equations, which can be solved by a suitable direct or iterative method. Namely, let $\{\varphi_i,\ i = 1,\ldots,N_h\}$ be a basis of the space S_{hp}, where $N_h = \dim S_{hp}$ (= dimension of S_{hp}). The approximate solution u_h is sought in the form $u_h(x) = \sum_{j=1}^{N_h} u^j\varphi_j(x)$, where $u^j,\ j = 1,\ldots,N_h$, are unknown real coefficients. Then, due to the linearity of the form A_h, the discrete problem (1.54) is equivalent to the system

$$\sum_{j=1}^{N_h} A_h(\varphi_j,\varphi_i)u^j = \ell_h(\varphi_j), \quad j = 1,\ldots,N_h. \tag{1.55}$$

It can be written in the matrix form

$$\mathbb{A}U = L,$$

where $\mathbb{A} = (a_{ij})_{i,j=1}^{N_h} = (\mathbb{A}_h(\varphi_j, \varphi_i))_{i,j=1}^{N_h}$, $U = (u^j)_{j=1}^{N_h}$ and $L = (\ell_h(\varphi_j))_{j=1}^{N_h}$.

From the construction of the forms A_h and ℓ_h, one can see that the strong solution $u \in H^2(\Omega)$ of problem (1.1) satisfies the identity

$$A_h(u, v) = \ell_h(v) \quad \forall\, v \in H^2(\Omega, \mathcal{T}_h), \tag{1.56}$$

which represents the *consistency* of the method. Relations (1.54) and (1.56) imply the so-called *Galerkin orthogonality* of the error $e_h = u_h - u$ of the method:

$$A_h(e_h, v_h) = 0 \quad \forall\, v_h \in S_{hp}, \tag{1.57}$$

which will be used in analysing error estimates.

Remark 1.8. *Comparing the above process of the derivation of the DG schemes with the abstract numerical method in Section 1.2, we see that we can define the function spaces*

$$V = H^2(\Omega), \quad W_h = H^2(\Omega, \mathcal{T}_h), \quad V_h = S_{hp}. \tag{1.58}$$

However, as we shall see later, the space W_h will not be equipped with the norm $\|\cdot\|_{H^2(\Omega,\mathcal{T}_h)}$ defined by (1.30)*, but by another norm introduced later in* (1.103) *will be used.*

Remark 1.9. *The interior and boundary penalty form J_h^σ together with the form J_D^σ replace the continuity of conforming finite element approximate solutions and represent Dirichlet boundary conditions. Thus, in contrast to standard conforming finite element techniques, both Dirichlet and Neumann boundary conditions are included automatically in the formulation (1.54) of the discrete problem. This is an advantage particularly in the case of nonhomogeneous Dirichlet boundary conditions, because it is not necessary to construct subsets of finite element spaces formed by functions approximating the Dirichlet boundary condition in a suitable way.*

Remark 1.10. *Method (1.49a) was introduced by Delves et al. ([DH79], [DP80], [HD79], [HDP79]), who called it a* global element method*. Its advantage is the symmetry of the discrete problem due to the third term on the right-hand side of* (1.45a)*. On the other hand, a significant disadvantage is that the bilinear form A_h^{s} is indefinite. This causes difficulties when dealing with time-dependent problems, because some eigenvalues of the operator associated with the form A_h can have negative real parts and then the resulting space-time discrete schemes become unconditionally unstable. Therefore, we prove in Lemma 1.36 the continuity of the bilinear form A_h^{s}, but further on we shall not be concerned with this method any more.*

Scheme (1.49b) was introduced by Baumann and Oden in [BBO99], [OBB98] and is usually called the Baumann–Oden method*. It is straightforward to show that the corresponding bilinear form A_h^{n} is positive semidefinite due to the third term on the right-hand side of* (1.45b)*. An interesting property of this method is that it is unstable for piecewise linear approximations, i.e., for $p = 1$.*

Scheme (1.49c) is called the symmetric interior penalty Galerkin *(SIPG) method. It was derived by Arnold ([Arn82]) and Wheeler ([Whe78]) by adding penalty terms to the form A_h^{s}. (In this case a_h and F_h are defined by (1.50) and (1.52) with $\Theta = 1$.) This formulation leads to a symmetric bilinear form, which is coercive, if the penalty parameter σ is sufficiently large. Moreover, the Aubin–Nitsche duality technique (also called Aubin–Nitsche trick) can be used to obtain an optimal error estimate in the $L^2(\Omega)$-norm.*

Method (1.49d), called the nonsymmetric interior penalty Galerkin *(NIPG) method, was proposed by Girault, Rivière and Wheeler in [RWG99]. (Here $\Theta = -1$.) In this case the bilinear form $A_h^{\mathrm{n},\sigma}$ is nonsymmetric and does not allow one to obtain an optimal error estimate in the $L^2(\Omega)$-norm with the aid of the Aubin-Nitsche trick. However, numerical experiments show that in some situations (for example, if uniform grids are used) the odd degrees of the polynomial approximation give the optimal order of convergence. On the other hand, a favorable property of the NIPG method is the coercivity of $A_h^{\mathrm{n},\sigma}(\cdot,\cdot)$ for any penalty parameter $\sigma > 0$.*

Finally, method (1.49e), called the incomplete interior penalty Galerkin *(IIPG) method ($\Theta = 0$), was studied in [DSW04], [Sun03], [SW05]. In this case the bilinear form $A_h^{\mathrm{i},\sigma}$ is nonsymmetric and does not allow one to obtain an optimal error estimate in the $L^2(\Omega)$-norm. The penalty parameter σ has to be chosen sufficiently large in order to guarantee the coercivity of $A_h^{\mathrm{i},\sigma}$. The advantage of the IIPG method is the simplicity of the discrete diffusion operator, because the expressions from* (1.44) *do not appear in* (1.45c)*. This is particularly advantageous in the case when the diffusion operator is nonlinear with respect to ∇u. (See, e.g., [Dol08a] or Chapter 7 of this book.)*

It would also be possible to define the scheme $A_h^{\mathrm{i}}(u, v) = \ell_h^{\mathrm{i}}(v)\ \forall\, v \in S_{hp}$, where $A_h^{\mathrm{i}}(u, v) = a_h^{\mathrm{i}}(u, v)$ and $\ell_h^{\mathrm{i}}(v) = F_h^{\mathrm{i}}(v)$, but this method does not make sense, because it does not contain the Dirichlet boundary data u_D from condition (1.1b)*.*

In the following, we shall deal with the theoretical analysis of the DGM applied to the numerical solution of the model problem (1.1). Namely, we shall pay attention to the existence and uniqueness of the approximate solution defined by (1.54) and derive error estimates.

1.5 Basic tools of the theoretical analysis of DGM

Theoretical analysis of the DG method presented in this book is based on three fundamental tools: the *multiplicative trace inequality*, the *inverse inequality*, and the *approximation properties* of the spaces of piecewise polynomial functions. In this section we introduce and prove these important tools under the assumptions about the meshes in Section 1.3.2.

Our first objective will be to summarize some important concepts and results from finite element theory, treated, e.g., in [Cia79].

Definition 1.11. *Let $n > 0$ be an integer. We say that sets $\omega, \widehat{\omega} \subset \mathbb{R}^n$ are* affine equivalent*, if there exists an invertible affine mapping $F_\omega : \widehat{\omega} \to \omega$ such that $F_\omega(\widehat{\omega}) = \omega$ and*

$$x = F_\omega(\hat{x}) = \mathbb{B}_\omega \hat{x} + \boldsymbol{b}_\omega \in \omega, \quad \hat{x} \in \widehat{\omega}, \tag{1.59}$$

where $\mathbb{B}_\omega$ is an $n \times n$ nonsingular matrix and $\boldsymbol{b}_\omega \in \mathbb{R}^n$.

If $\hat{v} : \widehat{\omega} \to \mathbb{R}$, then the inverse mapping F_ω^{-1} allows us to transform the function $\hat{v}$ to $v : \omega \to \mathbb{R}$ by the relation

$$v(x) = \hat{v}(F_\omega^{-1}(x)), \quad x \in \omega. \tag{1.60}$$

Hence,

$$v = \hat{v} \circ F_\omega^{-1}, \quad \hat{v} = v \circ F_\omega \tag{1.61}$$

and

$$\hat{v}(\hat{x}) = v(x) \text{ for all } \hat{x},\ x \text{ in the correspondence (1.59).}$$

If $\mathbb{B}$ is an $n \times n$ matrix, then its norm associated with the Euclidean norm $|\cdot|$ in $\mathbb{R}^n$ is defined as $\|\mathbb{B}\| = \sup_{x \in \mathbb{R}^n} |\mathbb{B}x|/|x|$.

The following lemmas give us bounds for the norms of matrices $\mathbb{B}_\omega$ and $\mathbb{B}_\omega^{-1}$ and the relations between Sobolev seminorms of functions v and $\hat{v}$ satisfying (1.61). First, we introduce the following notation for bounded domains ω, $\widehat{\omega}$:

$$h_\omega = \text{diam}(\omega),\ h_{\widehat{\omega}} = \text{diam}(\widehat{\omega}), \tag{1.62}$$

$$\rho_\omega = \text{radius of the largest ball inscribed into } \overline{\omega}, \tag{1.63}$$

$$\rho_{\widehat{\omega}} = \text{radius of the largest ball inscribed into } \overline{\widehat{\omega}}.$$

Lemma 1.12. *Let $\omega, \widehat{\omega} \subset \mathbb{R}^n$ be affine-equivalent bounded domains with the invertible mapping $F_\omega(\hat{x}) = \mathbb{B}_\omega \hat{x} + b_\omega \in \omega$ for $\hat{x} \in \widehat{\omega}$. Then*

$$\|\mathbb{B}_\omega\| \le \frac{h_\omega}{2\rho_{\widehat{\omega}}}, \quad \|\mathbb{B}_\omega^{-1}\| \le \frac{h_{\widehat{\omega}}}{2\rho_\omega}. \tag{1.64}$$

Further, the substitution theorem implies that

$$|\det(\mathbb{B}_\omega)| = |\omega|/|\widehat{\omega}|, \tag{1.65}$$

where $|\omega|$ and $|\widehat{\omega}|$ denote the n-dimensional Lebesgue measure of ω and $\widehat{\omega}$, respectively.

For the proof of (1.64) see [Cia79], Theorem 3.1.3. The proof of (1.65) is a consequence of the substitution theorem. Further, we cite here Theorem 3.1.2 from [Cia79].

Lemma 1.13. *Let $\omega, \widehat{\omega} \subset \mathbb{R}^n$ be affine-equivalent bounded domains with the invertible mapping $F_\omega(\hat{x}) = \mathbb{B}_\omega \hat{x} + b_\omega \in \omega$ for $\hat{x} \in \widehat{\omega}$. If $v \in W^{m,\alpha}(\omega)$ for some integer $m \ge 0$ and some $\alpha \in [1, \infty]$, then the function $\hat{v} = v \circ F_\omega \in W^{m,\alpha}(\widehat{\omega})$. Moreover, there exists a constant C depending on m and d only such that*

$$|\hat{v}|_{W^{m,\alpha}(\widehat{\omega})} \le C\, \|\mathbb{B}_\omega\|^m\, |\det(\mathbb{B}_\omega)|^{-1/\alpha}\, |v|_{W^{m,\alpha}(\omega)}, \tag{1.66}$$

$$|v|_{W^{m,\alpha}(\omega)} \le C\, \|\mathbb{B}_\omega^{-1}\|^m\, |\det(\mathbb{B}_\omega)|^{1/\alpha}\, |\hat{v}|_{W^{m,\alpha}(\widehat{\omega})}. \tag{1.67}$$

In our finite element analysis, we have $n = d$ and the set ω represents an element $K \in \mathcal{T}_h$ and $\widehat{\omega}$ is chosen as a reference element $\widehat{K}$, i. e., the simplex with vertices

$$\hat{a}_1 = (0, 0, \dots, 0), \quad \hat{a}_2 = (1, 0, \dots, 0), \quad \hat{a}_3 = (0, 1, 0, \dots, 0), \dots \tag{1.68}$$
$$\dots, \quad \hat{a}_{d+1} = (0, 0, \dots, 1) \in \mathbb{R}^d.$$

The elements K and $\widehat{K}$ are considered as closed sets. The Sobolev spaces over K and $\widehat{K}$ are defined as the spaces over the interiors of these sets. (In Section **??**, we shall also apply the above results to the case with $n = 1$, $\omega = \Gamma \in \mathcal{F}_h$ and $\widehat{\omega} = (0, 1)$.)

As a consequence of the above results we can formulate the following assertions.

Corollary 1.14. *If $K \in \mathcal{T}_h$ and $v \in H^m(K)$, where $m \geq 0$ is an integer, then the function $\hat{v}(\hat{x}) = v(F_K(\hat{x})) \in H^m(\widehat{K})$ and*

$$|v|_{H^m(K)} \leq c_c h_K^{\frac{d}{2}-m} |\hat{v}|_{H^m(\widehat{K})}, \tag{1.69}$$

$$|\hat{v}|_{H^m(\widehat{K})} \leq c_c h_K^{m-\frac{d}{2}} |v|_{H^m(K)}, \tag{1.70}$$

where $c_c > 0$ depends on the shape regularity constant C_R but not on K and v.

Exercise 1.15. *Prove* (1.69) –(1.70) *using the shape-regularity assumption* (1.19) *and the results of Lemmas 1.12 and 1.13.*

In deriving error estimates we shall apply the following important result from [Cia79, Theorem 3.1.4].

Theorem 1.16. *Let $\widehat{\omega} \subset \mathbb{R}^n$ be a bounded domain and for some integers $p \geq 0$ and $m \geq 0$ and some numbers $\alpha, \beta \in [1, \infty]$, let the spaces $W^{p+1,\alpha}(\widehat{\omega})$ and $W^{m,\beta}(\widehat{\omega})$ satisfy the continuous embedding*

$$W^{p+1,\alpha}(\widehat{\omega}) \hookrightarrow W^{m,\beta}(\widehat{\omega}). \tag{1.71}$$

Let $\widehat{\Pi}$ be a continuous linear mapping of $W^{p+1,\alpha}(\widehat{\omega})$ into $W^{m,\beta}(\widehat{\omega})$ such that

$$\widehat{\Pi}\hat{\phi} = \hat{\phi} \quad \forall\, \hat{\phi} \in P_p(\widehat{\omega}). \tag{1.72}$$

Let a set ω be affine-equivalent to the set $\widehat{\omega}$. This means that there exists an affine mapping $x = F_\omega$, $F_\omega(\hat{x}) = \mathbb{B}_\omega \hat{x} + b_\omega \in \omega$ for $\hat{x} \in \widehat{\omega}$, where $\mathbb{B}_\omega$ is a nonsingular $n \times n$ matrix and $b_\omega \in \mathbb{R}^n$. Let the mapping Π_ω be defined by

$$\Pi_\omega v(x) = (\widehat{\Pi}\hat{v})(F_\omega^{-1}(x)), \tag{1.73}$$

for all functions $\hat{v} \in W^{p+1,\alpha}(\widehat{\omega})$ and $v \in W^{p+1,\alpha}(\omega)$ such that $\hat{v}(\hat{x}) = v(F_\omega(\hat{x})) = v(x)$. Then there exists a constant $C(\widehat{\Pi}, \widehat{\omega})$ such that

$$|\widehat{\Pi}\hat{v} - \hat{v}|_{W^{m,\beta}(\widehat{\omega})} \leq C(\widehat{\Pi}, \widehat{\omega}) |\hat{v}|_{W^{p+1,\alpha}(\widehat{\omega})}, \tag{1.74}$$

and

$$|v - \Pi_\omega v|_{W^{m,\beta}(\omega)} \leq C(\widehat{\Pi}, \widehat{\omega})\, |\omega|^{(1/\beta)-(1/\alpha)}\, \frac{h_\omega^{p+1}}{\rho_\omega^m} |v|_{W^{p+1,\alpha}(\omega)} \tag{1.75}$$
$$\forall\, v \in W^{p+1,\alpha}(\omega),$$

with $h_\omega = \operatorname{diam}(\omega)$, ρ_ω defined as the radius of the largest ball inscribed into $\overline{\omega}$ and $|\omega|$ defined as the n-dimensional Lebesgue measure of the set ω. We set $1/\infty := 0$.

Exercise 1.17. *Prove* (1.75) *using* (1.74), (1.66), (1.67), (1.64) *and* (1.65).

Another important result used often in finite element theory is the Bramble–Hilbert lemma (see [Cia79, Theorem 4.1.3] or [Žen90, Theorem 9.3]).

Theorem 1.18 (Bramble–Hilbert lemma)**.** *Let us assume that $\omega \subset \mathbb{R}^n$ is a bounded domain with Lipschitz boundary. Let $p \geq 0$ be an integer and $\alpha \in [1, \infty]$ and let f be a continuous linear functional on the space $W^{p+1,\alpha}(\Omega)$ (i.e., $f \in (W^{p+1,\alpha}(\omega))^*$) satisfying the condition*

$$f(v) = 0 \quad \forall\, v \in P_p(\omega). \tag{1.76}$$

Then there exists a constant $C_{BH} > 0$ depending only on ω such that

$$|f(v)| \leq C_{BH} \|f\|_{(W^{p+1,\alpha}(\omega))^*} |v|_{W^{p+1,\alpha}(\omega)} \quad \forall\, v \in W^{p+1,\alpha}(\omega). \tag{1.77}$$

1.5.1 Multiplicative trace inequality

The forms a_h and J_h^σ given by (1.45) and (1.41), respectively, contain several integrals over faces. Therefore, in the theoretical analysis we need to estimate norms over faces by norms over elements. These estimates are usually obtained using the *multiplicative trace inequality*. In the literature, it is possible to find several variants of the multiplicative trace inequality. Here, we present the variant, which suits our considerations.

Lemma 1.19 (Multiplicative trace inequality)**.** *Let the shape-regularity assumption* (1.19) *be satisfied. Then there exists a constant $C_M > 0$ independent of v, h and K such that*

$$\|v\|^2_{L^2(\partial K)} \leq C_M \left(\|v\|_{L^2(K)} |v|_{H^1(K)} + h_K^{-1} \|v\|^2_{L^2(K)} \right), \tag{1.78}$$
$$K \in \mathcal{T}_h,\ v \in H^1(K),\ h \in (0, \bar{h}).$$

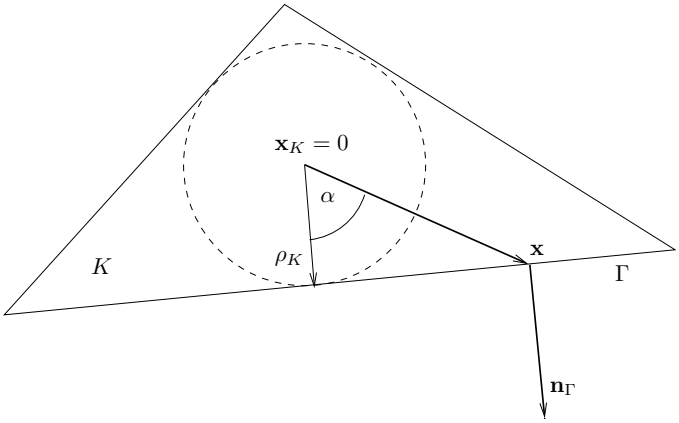

Figure 1.3: Simplex K with its face Γ.

Proof. Let $K \in \mathcal{T}_h$ be arbitrary but fixed. We denote by $\boldsymbol{x}_K$ the center of the largest d-dimensional ball inscribed into the simplex K. Without loss of generality we suppose that $\boldsymbol{x}_K$ is the origin of the coordinate system.

Since the space $C^\infty(K)$ is dense in $H^1(K)$, it is sufficient to prove (1.78) for $v \in C^\infty(K)$. We start from the following relation obtained from Green's identity (23):

$$\int_{\partial K} v^2 \boldsymbol{x} \cdot \boldsymbol{n} \,\mathrm{d}S = \int_K \nabla \cdot (v^2 \boldsymbol{x}) \,\mathrm{d}x, \qquad v \in C^\infty(K), \tag{1.79}$$

where $\boldsymbol{n}$ denotes here the outer unit normal to ∂K. Let $\boldsymbol{n}_\Gamma$ be the outer unit normal to K on a side Γ of K. Then

$$\boldsymbol{x} \cdot \boldsymbol{n}_\Gamma = |\boldsymbol{x}||\boldsymbol{n}_\Gamma| \cos\alpha = |\boldsymbol{x}| \cos\alpha = \rho_K, \quad \boldsymbol{x} \in \Gamma, \tag{1.80}$$

see Figure 1.3. From (1.80) we have

$$\int_{\partial K} v^2 \boldsymbol{x} \cdot \boldsymbol{n} \,\mathrm{d}S = \sum_{\Gamma \subset \partial K} \int_\Gamma v^2 \boldsymbol{x} \cdot \boldsymbol{n}_\Gamma \,\mathrm{d}S = \rho_K \sum_{\Gamma \subset \partial K} \int_\Gamma v^2 \,\mathrm{d}S = \rho_K \|v\|^2_{L^2(\partial K)}. \tag{1.81}$$

Moreover,

$$\begin{aligned} \int_K \nabla \cdot (v^2 \boldsymbol{x}) \,\mathrm{d}x &= \int_K \left(v^2 \nabla \cdot \boldsymbol{x} + \boldsymbol{x} \cdot \nabla v^2 \right) \,\mathrm{d}x \\ &= d \int_K v^2 \,\mathrm{d}x + 2 \int_K v \boldsymbol{x} \cdot \nabla v \,\mathrm{d}x \le d\|v\|^2_{L^2(K)} + 2 \int_K |v \boldsymbol{x} \cdot \nabla v| \,\mathrm{d}x. \end{aligned} \tag{1.82}$$

With the aid of the Cauchy inequality, the second term of (1.82) is estimated as

$$2 \int_K |v \boldsymbol{x} \cdot \nabla v| \,\mathrm{d}x \le 2 \sup_{\boldsymbol{x} \in K} |\boldsymbol{x}| \int_K |v||\nabla v| \,\mathrm{d}x \le 2 h_K \|v\|_{L^2(K)} |v|_{H^1(K)}. \tag{1.83}$$

Then (1.19), (1.79), (1.81), (1.82) and (1.83) give

$$\begin{aligned} \|v\|^2_{L^2(\partial K)} &\le \frac{1}{\rho_K} \left[2 h_K \|v\|_{L^2(K)} |v|_{H^1(K)} + d \|v\|^2_{L^2(K)} \right] \\ &\le C_R \left[2 \|v\|_{L^2(K)} |v|_{H^1(K)} + \frac{d}{h_K} \|v\|^2_{L^2(K)} \right], \end{aligned} \tag{1.84}$$

which proves (1.78) with $C_M = C_R \max\{2, d\}$. □ □

Exercise 1.20. *Prove that the multiplicative trace inequality is valid also for vector-valued functions $\boldsymbol{v} : \Omega \to \mathbb{R}^n$, i.e.,*

$$\|\boldsymbol{v}\|^2_{L^2(\partial K)} \le C_M \left(\|\boldsymbol{v}\|_{L^2(K)} \, |\boldsymbol{v}|_{H^1(K)} + h_K^{-1} \|\boldsymbol{v}\|^2_{L^2(K)} \right), \quad \boldsymbol{v} \in (H^1(K))^n, \ K \in \mathcal{T}_h. \tag{1.85}$$

Hint: Use (1.78) *for each component of $\boldsymbol{v} = (v_1, \ldots, v_n)$, sum these inequalities and apply the discrete Cauchy inequality* (52).

1.5.2 Inverse inequality

In deriving error estimates, we need to estimate the H^1-seminorm of a polynomial function by its L^2-norm, i.e., we apply the so-called *inverse inequality*.

Lemma 1.21 (Inverse inequality)**.** *Let the shape-regularity assumption* (1.19) *be satisfied. Then there exists a constant $C_I > 0$ independent of v, h and K such that*

$$|v|_{H^1(K)} \le C_I h_K^{-1} \|v\|_{L^2(K)} \quad \forall\, v \in P_p(K),\ \forall\, K \in \mathcal{T}_h,\ \forall\, h \in (0, \bar{h}). \tag{1.86}$$

Proof. Let $\widehat{K}$ be a reference triangle and $F_K : \widehat{K} \to K,\ K \in \mathcal{T}_h$ be an affine mapping such that $F_K(\widehat{K}) = K$. By (1.69) (for $m = 1$) and (1.70) (for $m = 0$) we have

$$|v|_{H^1(K)} \le c_c h_K^{\frac{d}{2}-1} |\hat{v}|_{H^1(\widehat{K})}, \qquad \|\hat{v}\|_{L^2(\widehat{K})} \le c_c h_K^{-\frac{d}{2}} \|v\|_{L^2(K)}. \tag{1.87}$$

From [Sch98, Theorem 4.76], we have

$$|\hat{v}|_{H^1(\widehat{K})} \le c_s p^2 \|\hat{v}\|_{L^2(\widehat{K})}, \quad \hat{v} \in P_p(\widehat{K}), \tag{1.88}$$

where $c_s > 0$ depends on d but not on $\hat{v}$ and p. A simple combination of (1.87) – (1.88) proves (1.86) with $C_I = c_s\, c_c^2\, p^2$. Let us note that (1.88) is a consequence of the norm equivalence on finite-dimensional spaces. □ □

Other inverse inequalities will appear in Section **??**, Lemma **??**.

1.5.3 Approximation properties

With respect to the error analysis of the abstract numerical method treated in Section 1.2, a suitable S_{hp}-interpolation has to be introduced. Let $\mathcal{T}_h$ be a given triangulation of the domain Ω. Then for each $K \in \mathcal{T}_h$, we define the mapping $\pi_{K,p} : L^2(K) \to P_p(K)$ such that for every $\varphi \in L^2(K)$

$$\pi_{K,p}\varphi \in P_p(K), \quad \int_K (\pi_{K,p}\varphi) v \,\mathrm{d}x = \int_K \varphi v \,\mathrm{d}x \quad \forall\, v \in P_p(K). \tag{1.89}$$

On the basis of the mappings $\pi_{K,p}$ we introduce the S_{hp}-interpolation Π_{hp}, defined for all $\varphi \in L^2(\Omega)$ by

$$(\Pi_{hp}\varphi)|_K = \pi_{K,p}(\varphi|_K) \quad \forall\, K \in \mathcal{T}_h. \tag{1.90}$$

It can be easily shown that if $\varphi \in L^2(\Omega)$, then

$$\Pi_{hp}\varphi \in S_{hp}, \quad \int_\Omega (\Pi_{hp}\varphi) v \,\mathrm{d}x = \int_\Omega \varphi v \,\mathrm{d}x \quad \forall\, v \in S_{hp}. \tag{1.91}$$

Hence, Π_{hp} is the $L^2(\Omega)$-projection on the space S_{hp}.

The approximation properties of the interpolation operators $\pi_{K,p}$ and Π_{hp} are the consequence of Theorem 1.16.

Lemma 1.22. *Let the shape-regularity assumption* (1.19) *be valid and let p, q, s be integers, $p \ge 0$, $0 \le q \le \mu$, where $\mu = \min(p+1, s)$. Then there exists a constant $C_A > 0$ such that*

$$|\pi_{K,p}v - v|_{H^q(K)} \le C_A h_K^{\mu-q} |v|_{H^\mu(K)} \quad \forall\, v \in H^s(K)\ \forall\, K \in \mathcal{T}_h\ \forall\, h \in (0, \bar{h}). \tag{1.92}$$

Hence, if $p \ge 1$ and $s \ge 2$, then

$$\|\pi_{K,p}v - v\|_{L^2(K)} \le C_A h_K^{\mu} |v|_{H^\mu(K)} \quad \forall\, v \in H^s(K)\ \forall\, K \in \mathcal{T}_h\ \forall\, h \in (0, \bar{h}), \tag{1.93}$$

$$|\pi_{K,p}v - v|_{H^1(K)} \le C_A h_K^{\mu-1} |v|_{H^\mu(K)} \quad \forall\, v \in H^s(K)\ \forall\, K \in \mathcal{T}_h\ \forall\, h \in (0, \bar{h}), \tag{1.94}$$

$$|\pi_{K,p}v - v|_{H^2(K)} \le C_A h_K^{\mu-2} |v|_{H^\mu(K)} \quad \forall\, v \in H^s(K)\ \forall\, K \in \mathcal{T}_h\ \forall\, h \in (0, \bar{h}). \tag{1.95}$$

Moreover, we have

$$\|\pi_{K,1}v - v\|_{L^\infty(K)} \le C_A h_K |v|_{W^{1,\infty}(K)}\ \forall\, v \in W^{1,\infty}(K)\ \forall\, K \in \mathcal{T}_h\ \forall\, h \in (0, \bar{h}). \tag{1.96}$$

Exercise 1.23. *Prove Lemma 1.22 using Theorem 1.16 and assumption* (1.19).

The above results immediately imply the approximation properties of the operator Π_{hp}.

Lemma 1.24. *Let the shape-regularity assumption* (1.19) *be satisfied and let* p, q, s *be integers,* $p \geq 0$, $0 \leq q \leq \mu$, *where* $\mu = \min(p+1, s)$. *Then*

$$|\Pi_{hp} v - v|_{H^q(\Omega,\mathcal{T}_h)} \leq C_A h^{\mu-q} |v|_{H^\mu(\Omega,\mathcal{T}_h)}, \quad v \in H^s(\Omega,\mathcal{T}_h),\ h \in (0,\bar{h}), \tag{1.97}$$

where $\mu = \min(p+1, s)$ *and* C_A *is the constant from* (1.92). *Hence, if* $p \geq 1$ *and* $s \geq 2$, *then*

$$\|\Pi_{hp} v - v\|_{L^2(\Omega)} \leq C_A h^{\mu} |v|_{H^\mu(\Omega,\mathcal{T}_h)}, \quad v \in H^s(\Omega,\mathcal{T}_h),\ h \in (0,\bar{h}), \tag{1.98}$$

$$|\Pi_{hp} v - v|_{H^1(\Omega,\mathcal{T}_h)} \leq C_A h^{\mu-1} |v|_{H^\mu(\Omega,\mathcal{T}_h)}, \quad v \in H^s(\Omega,\mathcal{T}_h),\ h \in (0,\bar{h}), \tag{1.99}$$

$$|\Pi_{hp} v - v|_{H^2(\Omega,\mathcal{T}_h)} \leq C_A h^{\mu-2} |v|_{H^\mu(\Omega,\mathcal{T}_h)}, \quad v \in H^s(\Omega,\mathcal{T}_h),\ h \in (0,\bar{h}). \tag{1.100}$$

Proof. Using (1.90), definition of the seminorm in a broken Sobolev space (1.31) and the approximation properties (1.92), we obtain (1.97). This immediately implies (1.98) – (1.100). □ □

Moreover, using the combination of the multiplicative trace inequality (1.78) and Lemma 1.22, we can prove the approximation properties of the operator Π_{hp} in the norms defined over the boundaries of elements.

Lemma 1.25. *Let the shape-regularity assumption* (1.19) *be satisfied and let* $p \geq 1, s \geq 2$ *be integers and* $\alpha \geq -1$. *Then*

$$\sum_{K\in\mathcal{T}_h} h_K^\alpha \|\Pi_{hp} v - v\|^2_{L^2(\partial K)} \leq 2 C_M C_A^2 h^{2\mu-1+\alpha} |v|^2_{H^\mu(\Omega,\mathcal{T}_h)}, \tag{1.101}$$

$$\sum_{K\in\mathcal{T}_h} h_K^\alpha \|\nabla(\Pi_{hp} v - v)\|^2_{L^2(\partial K)} \leq 2 C_M C_A^2 h^{2\mu-3+\alpha} |v|^2_{H^\mu(\Omega,\mathcal{T}_h)}, \tag{1.102}$$

$$v \in H^s(\Omega,\mathcal{T}_h),\ h \in (0,\bar{h}),$$

where $\mu = \min(p+1, s)$, C_M *is the constant from* (1.78) *and* C_A *is the constant from* (1.92).

Proof. (i) Let $v \in H^s(\Omega,\mathcal{T}_h)$. For simplicity we put $\eta = \Pi_{hp} v - v$. Then relation (1.90) implies that $\eta|_K = \pi_{K,p} v|_K - v|_K$ for $K \in \mathcal{T}_h$. Using the multiplicative trace inequality (1.78), the approximation property (1.92), and the seminorm definition (1.31), we have

$$\begin{aligned}
\sum_{K\in\mathcal{T}_h} h_K^\alpha \|\eta\|^2_{L^2(\partial K)} &\leq C_M \sum_{K\in\mathcal{T}_h} h_K^\alpha \left(\|\eta\|_{L^2(K)} |\eta|_{H^1(K)} + h_K^{-1} \|\eta\|^2_{L^2(K)} \right) \\
&\leq C_M \sum_{K\in\mathcal{T}_h} h_K^\alpha C_A^2 \left(h_K^\mu h_K^{\mu-1} + h_K^{-1} h_K^{2\mu} \right) |v|^2_{H^\mu(K)} \\
&\leq 2 C_M C_A^2 h^{2\mu-1+\alpha} |v|^2_{H^\mu(\Omega,\mathcal{T}_h)}.
\end{aligned}$$

(ii) Similarly as above, using the vector-valued variant of the multiplicative trace inequality (1.85), identities (21) and the approximation property (1.92) we get

$$\begin{aligned}
\sum_{K\in\mathcal{T}_h} h_K^\alpha \|\nabla\eta\|^2_{L^2(\partial K)} &\leq C_M \sum_{K\in\mathcal{T}_h} h_K^\alpha \left(\|\nabla\eta\|_{L^2(K)} |\nabla\eta|_{H^1(K)} + h_K^{-1} \|\nabla\eta\|^2_{L^2(K)} \right) \\
&= C_M \sum_{K\in\mathcal{T}_h} h_K^\alpha \left(|\eta|_{H^1(K)} |\eta|_{H^2(K)} + h_K^{-1} |\eta|^2_{H^1(K)} \right) \\
&\leq C_M \sum_{K\in\mathcal{T}_h} h_K^\alpha C_A^2 \left(h_K^{\mu-1} h_K^{\mu-2} + h_K^{-1} h_K^{2(\mu-1)} \right) |v|^2_{H^\mu(K)} \\
&\leq 2 C_M C_A^2 h^{2\mu-3+\alpha} |v|^2_{H^\mu(\Omega,\mathcal{T}_h)}.
\end{aligned}$$

□ □

1.6 Existence and uniqueness of the approximate solution

We start with the theoretical analysis of the DGM, namely we prove the existence of a numerical solution defined by (1.54). Then, in Section 1.7, we derive error estimates. We follow the formal analysis of the abstract numerical methods in Section 1.2. Therefore, we show the *continuity* and the *coercivity* of the form A_h given by (1.47) in a suitable norm. This norm should reflect the discontinuity of functions from the broken Sobolev spaces $H^1(\Omega,\mathcal{T}_h)$. To this end, we define the following mesh-dependent norm

$$|||u|||_{\mathcal{T}_h} = \left(|u|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(u,u) \right)^{1/2}, \tag{1.103}$$

where $|\cdot|_{H^1(\Omega,\mathcal{T}_h)}$ and J_h^σ are given by (1.31) and (1.41), respectively.

In what follows, because there is no danger of misunderstanding, we shall omit the subscript $\mathcal{T}_h$. This means that we shall simply write $|||\cdot||| = |||\cdot|||_{\mathcal{T}_h}$. We call $|||\cdot|||$ the *DG-norm*.

Exercise 1.26. *Prove that $|||\cdot|||$ is a norm in the spaces $H^1(\Omega,\mathcal{T}_h)$ and S_{hp}.*

1.6.1 The choice of penalty weight σ

In the following considerations we shall assume that the system $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ of triangulations satisfies the shape-regularity assumption (1.19) and the equivalence condition (1.20).

We consider the penalty weight $\sigma : \cup_{\Gamma\in\mathcal{F}_h^{ID}} \to \mathbb{R}$ in the form

$$\sigma|_\Gamma = \sigma_\Gamma = \frac{C_W}{h_\Gamma}, \quad \Gamma \in \mathcal{F}_h^{ID}, \tag{1.104}$$

where $C_W > 0$ is the *penalization constant* and $h_\Gamma(\sim h)$ is the quantity given by one of the possibilities from (1.24) – (1.27) with respect to the considered mesh assumptions (MA1) – (MA4), see Lemma 1.5. Let us note that in some cases it is possible to consider a different form of the penalty parameter σ, as mentioned in Remark 1.51.

Under the introduced notation, in view of (1.41), (1.42) and (1.104), the interior and boundary penalty form and the associated boundary linear form read as

$$J_h^\sigma(u,v) = \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \frac{C_W}{h_\Gamma}[u]\,[v]\,\mathrm{d}S, \qquad J_D^\sigma(v) = \sum_{\Gamma\in\mathcal{F}_h^{D}} \int_\Gamma \frac{C_W}{h_\Gamma} u_D\, v\,\mathrm{d}S. \tag{1.105}$$

In what follows, we shall introduce technical lemmas, which will be useful in the theoretical analysis.

Lemma 1.27. *Let* (1.20) *be valid. Then for each $v \in H^1(\Omega,\mathcal{T}_h)$ we have*

$$\sum_{\Gamma\in\mathcal{F}_h^{ID}} h_\Gamma^{-1} \int_\Gamma [v]^2\,\mathrm{d}S \le \frac{2}{C_T} \sum_{K\in\mathcal{T}_h} h_K^{-1} \int_{\partial K} |v|^2\,\mathrm{d}S, \tag{1.106}$$

$$\sum_{\Gamma\in\mathcal{F}_h^{ID}} h_\Gamma \int_\Gamma \langle v\rangle^2\,\mathrm{d}S \le C_G \sum_{K\in\mathcal{T}_h} h_K \int_{\partial K} |v|^2\,\mathrm{d}S. \tag{1.107}$$

Hence,

$$\sum_{\Gamma\in\mathcal{F}_h^{ID}} \sigma_\Gamma \|[v]\|^2_{L^2(\Gamma)} \le \frac{2C_W}{C_T} \sum_{K\in\mathcal{T}_h} h_K^{-1} \|v\|^2_{L^2(\partial K)}, \tag{1.108}$$

$$\sum_{\Gamma\in\mathcal{F}_h^{ID}} \frac{1}{\sigma_\Gamma} \|\langle v\rangle\|^2_{L^2(\Gamma)} \le \frac{C_G}{C_W} \sum_{K\in\mathcal{T}_h} h_K \|v\|^2_{L^2(\partial K)}. \tag{1.109}$$

Proof. (i) By definition (1.32), the inequality

$$(\gamma+\delta)^2 \le 2(\gamma^2+\delta^2), \quad \gamma,\delta\in\mathbb{R}, \tag{1.110}$$

and (1.20) we have

$$\begin{aligned}
&\sum_{\Gamma\in\mathcal{F}_h^{ID}} h_\Gamma^{-1} \int_\Gamma [v]^2\,\mathrm{d}S \\
&\quad= \sum_{\Gamma\in\mathcal{F}_h^{I}} h_\Gamma^{-1} \int_\Gamma \left|v_\Gamma^{(L)} - v_\Gamma^{(R)}\right|^2 \mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^{D}} h_\Gamma^{-1} \int_\Gamma \left|v_\Gamma^{(L)}\right|^2 \mathrm{d}S \\
&\quad\le 2\sum_{\Gamma\in\mathcal{F}_h^{I}} h_\Gamma^{-1} \int_\Gamma \left(\left|v_\Gamma^{(L)}\right|^2 + \left|v_\Gamma^{(R)}\right|^2\right) \mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^{D}} h_\Gamma^{-1} \int_\Gamma \left|v_\Gamma^{(L)}\right|^2 \mathrm{d}S \\
&\quad\le 2C_T^{-1} \sum_{\Gamma\in\mathcal{F}_h^{ID}} h_{K_\Gamma^{(L)}}^{-1} \int_\Gamma \left|v_\Gamma^{(L)}\right|^2 \mathrm{d}S + 2C_T^{-1} \sum_{\Gamma\in\mathcal{F}_h^{I}} h_{K_\Gamma^{(R)}}^{-1} \int_\Gamma \left|v_\Gamma^{(R)}\right|^2 \mathrm{d}S \\
&\quad\le 2C_T^{-1} \sum_{K\in\mathcal{T}_h} h_K^{-1} \int_{\partial K} |v|^2\,\mathrm{d}S.
\end{aligned}$$

This and (1.104) immediately imply (1.108).

(ii) In the proof of (1.107) we proceed similarly, using (1.32), (1.20) and (1.110). Inequalities (1.108) and (1.109) are obtained from (1.106), (1.107) and (1.104). □ □

1.6.2 Continuity of diffusion bilinear forms

First, we shall prove several auxiliary assertions.

Lemma 1.28. *Any form a_h defined by* (1.45) *satisfies the estimate*

$$|a_h(u,v)| \le \|u\|_{1,\sigma}\,\|v\|_{1,\sigma} \quad \forall\, u,v \in H^2(\Omega,\mathcal{T}_h), \tag{1.111}$$

where

$$\begin{aligned}\|v\|_{1,\sigma}^2 =& |||v|||^2 + \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v\rangle)^2\,\mathrm{d}S \\ =& |v|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(v,v) + \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v\rangle)^2\mathrm{d}S.\end{aligned} \tag{1.112}$$

Proof. It follows from (1.45) that

$$\begin{aligned}|a_h(u,v)| \le & \underbrace{\sum_{K\in\mathcal{T}_h}\int_K |\nabla u\cdot\nabla v|\,\mathrm{d}x}_{\chi_1} \\ & + \underbrace{\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma |\boldsymbol{n}\cdot\langle\nabla u\rangle\,[v]|\,\mathrm{d}S}_{\chi_2} + \underbrace{\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma |\boldsymbol{n}\cdot\langle\nabla v\rangle\,[u]|\,\mathrm{d}S}_{\chi_3}.\end{aligned} \tag{1.113}$$

(For the form a_h^{i} the term χ_3 vanishes, of course.) Obviously, the Cauchy inequality, the discrete Cauchy inequality, and (1.31) imply that

$$\chi_1 \le \sum_{K\in\mathcal{T}_h} |u|_{H^1(K)}|v|_{H^1(K)} \le |u|_{H^1(\Omega,\mathcal{T}_h)}|v|_{H^1(\Omega,\mathcal{T}_h)}. \tag{1.114}$$

Further, by the Cauchy inequality,

$$\begin{aligned}\chi_2 &\le \sum_{\Gamma\in\mathcal{F}_h^{ID}} \left(\int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla u\rangle)^2\,\mathrm{d}S\right)^{1/2}\left(\int_\Gamma \sigma[v]^2\,\mathrm{d}S\right)^{1/2} \\ &\le \left(\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla u\rangle)^2\,\mathrm{d}S\right)^{1/2}\left(\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma[v]^2\,\mathrm{d}S\right)^{1/2},\end{aligned} \tag{1.115}$$

and

$$\chi_3 \le \left(\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v\rangle)^2\,\mathrm{d}S\right)^{1/2}\left(\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma[u]^2\,\mathrm{d}S\right)^{1/2}. \tag{1.116}$$

Using the discrete Cauchy inequality, from (1.114) – (1.116) we derive the bound

$$\begin{aligned}|a_h(u,v)| \le & |u|_{H^1(\Omega,\mathcal{T}_h)}|v|_{H^1(\Omega,\mathcal{T}_h)} \\ & + \left(\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla u\rangle)^2\,\mathrm{d}S\right)^{1/2}\left(\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma[v]^2\,\mathrm{d}S\right)^{1/2} \\ & + \left(\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v\rangle)^2\,\mathrm{d}S\right)^{1/2}\left(\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma[u]^2\,\mathrm{d}S\right)^{1/2} \\ & \le \left(|u|^2_{H^1(\Omega,\mathcal{T}_h)} + \sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla u\rangle)^2\mathrm{d}S + J_h^\sigma(u,u)\right)^{1/2} \\ & \times \left(|v|^2_{H^1(\Omega,\mathcal{T}_h)} + \sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v\rangle)^2\mathrm{d}S + J_h^\sigma(v.v)\right)^{1/2} \\ & = \|u\|_{1,\sigma}\,\|v\|_{1,\sigma}.\end{aligned} \tag{1.117}$$

□ □

Exercise 1.29. *Prove that $\|\cdot\|_{1,\sigma}$ introduced by (1.112) defines a norm in the broken Sobolev space $H^2(\Omega,\mathcal{T}_h)$.*

Corollary 1.30. *By virtue of* (1.47a) – (1.47b)*, Lemma 1.28 and Exercise 1.29, the bilinear forms A_h^{s} and A_h^{n} are bounded with respect to the norm $\|\cdot\|_{1,\sigma}$ in the broken Sobolev space $H^2(\Omega,\mathcal{T}_h)$.*

Exercise 1.31. *Prove Corollary 1.30.*

Further, we shall pay attention on the expression $J_h^\sigma(u,v)$ for $u,v\in H^1(\Omega,\mathcal{T}_h)$.

Lemma 1.32. *Let assumptions* (1.104)*,* (1.19) *and* (1.20) *be satisfied. Then*

$$|J_h^\sigma(u,v)|\le J_h^\sigma(u,u)^{1/2}J_h^\sigma(v,v)^{1/2}\quad\forall\,u,v\in H^1(\Omega,\mathcal{T}_h),\tag{1.118}$$

and

$$\begin{aligned}J_h^\sigma(v,v)&\le\frac{2C_WC_M}{C_T}\sum_{K\in\mathcal{T}_h}\left(h_K^{-2}\|v\|_{L^2(K)}^2+h_K^{-1}\|v\|_{L^2(K)}|v|_{H^1(K)}\right)\\&\le\frac{C_WC_M}{C_T}\sum_{K\in\mathcal{T}_h}\left(3h_K^{-2}\|v\|_{L^2(K)}^2+|v|_{H^1(K)}^2\right)\qquad\forall\,v\in H^1(\Omega,\mathcal{T}_h).\end{aligned}\tag{1.119}$$

Proof. Let $u,v\in H^1(\Omega,\mathcal{T}_h)$. By the definition (1.41) of the form J_h^σ and the Cauchy inequality,

$$\begin{aligned}|J_h^\sigma(u,v)|&\le\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma\sigma|[u]\,[v]|\mathrm{d}S\\&\le\left(\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma\sigma[u]^2\,\mathrm{d}S\right)^{1/2}\left(\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma\sigma[v]^2\,\mathrm{d}S\right)^{1/2}\\&=J_h^\sigma(u,u)^{1/2}J_h^\sigma(v,v)^{1/2}.\end{aligned}\tag{1.120}$$

Further, the definition of the form J_h^σ, (1.104), (1.20) and (1.108) imply that

$$J_h^\sigma(v,v)=\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma\sigma[v]^2\,\mathrm{d}S=\sum_{\Gamma\in\mathcal{F}_h^{ID}}\frac{C_W}{h_\Gamma}\|[v]^2\|_{L^2(\Gamma)}\le\frac{2C_W}{C_T}\sum_{K\in\mathcal{T}_h}h_K^{-1}\|v\|_{L^2(\partial K)}^2.$$

Now, using the multiplicative trace inequality (1.78), we get

$$J_h^\sigma(v,v)\le\frac{2C_WC_M}{C_T}\sum_{K\in\mathcal{T}_h}\left(h_K^{-2}\|v\|_{L^2(K)}^2+h_K^{-1}\|v\|_{L^2(K)}|v|_{H^1(K)}\right).\tag{1.121}$$

The last relation in (1.119) follows from (1.121) and Young's inequality. □ □

Lemmas 1.28 and 1.32 immediately imply the boundedness also of the forms $A_h^{\mathrm{s},\sigma}$, $A_h^{\mathrm{n},\sigma}$ and $A_h^{\mathrm{i},\sigma}$ with respect to the norm $\|\cdot\|_{1,\sigma}$.

Corollary 1.33. *Let assumptions* (1.104)*,* (1.19) *and* (1.20) *be satisfied. Then the forms A_h defined by* (1.47) *satisfy the estimate*

$$|A_h(u,v)|\le2\|u\|_{1,\sigma}\|v\|_{1,\sigma}\quad\forall\,u,v\in H^2(\Omega,\mathcal{T}_h).\tag{1.122}$$

Proof. For the boundedness of $A_h=A_h^{\mathrm{s}}$ and $A_h=A_h^{\mathrm{n}}$, see Corollary (1.30). Let $A_h=A_h^{\mathrm{s},\sigma}$ or $A_h=A_h^{\mathrm{n},\sigma}$ or $A_h=A_h^{\mathrm{i},\sigma}$. Then, by virtue of (1.47c) – (1.47e), Lemmas 1.28 and 1.32 we have

$$\begin{aligned}|A_h(u,v)|&\le|a_h(u,v)|+|J_h^\sigma(u,v)|\le\|u\|_{1,\sigma}\,\|v\|_{1,\sigma}+J_h^\sigma(u,u)^{1/2}J_h^\sigma(v,v)^{1/2}\\&\le\|u\|_{1,\sigma}\,\|v\|_{1,\sigma}+\|u\|_{1,\sigma}\,\|v\|_{1,\sigma}=2\|u\|_{1,\sigma}\,\|v\|_{1,\sigma}.\end{aligned}$$

□ □

The following lemma allows us to estimate the expressions with integrals over $\Gamma\in\mathcal{F}_h$ in terms of norms over elements $K\in\mathcal{T}_h$.

Lemma 1.34. *Let the weight σ be defined by* (1.104)*. Then, under assumptions* (1.19) *and* (1.20)*, for any $v \in H^2(\Omega, \mathcal{T}_h)$ the following estimate holds:*

$$\begin{aligned}\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v\rangle)^2\,\mathrm{d}S &\le \frac{C_G C_M}{C_W}\sum_{K\in\mathcal{T}_h}\Big(h_K\|\nabla v\|_{L^2(K)}\,|\nabla v|_{H^1(K)}+\|\nabla v\|^2_{L^2(K)}\Big)\\ &= \frac{C_G C_M}{C_W}\sum_{K\in\mathcal{T}_h}\Big(h_K|v|_{H^1(K)}\,|v|_{H^2(K)}+|v|^2_{H^1(K)}\Big)\\ &\le \frac{C_G C_M}{2C_W}\sum_{K\in\mathcal{T}_h}\Big(h_K^2|v|^2_{H^2(K)}+3|v|^2_{H^1(K)}\Big).\end{aligned} \tag{1.123}$$

Moreover, if $v \in S_{hp}$, then

$$\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v_h\rangle)^2\,\mathrm{d}S \le \frac{C_G C_M}{C_W}(C_I+1)|v_h|^2_{H^1(\Omega,\mathcal{T}_h)}. \tag{1.124}$$

Proof. Using (1.109) and the multiplicative trace inequality (1.78), we find that

$$\begin{aligned}&\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v\rangle)^2\,\mathrm{d}S\\ &\qquad\le \frac{C_G}{C_W}\sum_{K\in\mathcal{T}_h} h_K\|\nabla v\|^2_{L^2(\partial K)}\\ &\qquad\le \frac{C_G C_M}{C_W}\sum_{K\in\mathcal{T}_h} h_K\Big(\|\nabla v\|_{L^2(K)}\,|\nabla v|_{H^1(K)}+h_K^{-1}\|\nabla v\|^2_{L^2(K)}\Big),\end{aligned}$$

which is the first inequality in (1.123). The second one directly follows from Young's inequality.

If $v \in S_{hp}$, then (1.123) and the inverse inequality (1.86) imply that

$$\begin{aligned}\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v_h\rangle)^2\,\mathrm{d}S &\le \frac{C_G C_M}{C_W}\sum_{K\in\mathcal{T}_h}\Big(C_I\|\nabla v_h\|^2_{L^2(K)}+\|\nabla v_h\|^2_{L^2(K)}\Big)\\ &= \frac{C_G C_M}{C_W}(C_I+1)\sum_{K\in\mathcal{T}_h}\|\nabla v_h\|_{L^2(K)} = \frac{C_G C_M}{C_W}(C_I+1)|v_h|^2_{H^1(\Omega,\mathcal{T}_h)},\end{aligned}$$

which we wanted to prove. □ □

We continue in the derivation of various inequalities based on the estimation of the $\|\cdot\|_{1,\sigma}$-norm.

Lemma 1.35. *Under assumptions of Lemma 1.34, there exist constants C_σ, $\tilde{C}_\sigma > 0$ such that*

$$J_h^\sigma(u,u)^{1/2} \le |||u||| \le \|u\|_{1,\sigma} \le C_\sigma\, R_a(u) \quad \forall\, u \in H^2(\Omega,\mathcal{T}_h),\ h \in (0,\bar{h}), \tag{1.125}$$

$$J_h^\sigma(v_h,v_h)^{1/2} \le |||v_h||| \le \|v_h\|_{1,\sigma} \le \tilde{C}_\sigma|||v_h||| \quad \forall\, v_h \in S_{hp},\ h \in (0,\bar{h}), \tag{1.126}$$

where

$$R_a(u) = \Big(\sum_{K\in\mathcal{T}_h}\Big(|u|^2_{H^1(K)}+h_K^2|u|^2_{H^2(K)}+h_K^{-2}\|u\|^2_{L^2(K)}\Big)\Big)^{1/2}, \quad u \in H^2(\Omega,\mathcal{T}_h). \tag{1.127}$$

Proof. The first two inequalities in (1.125) as well as in (1.126) follow immediately from the definition of the DG-norm (1.103) and the $\|\cdot\|_{1,\sigma}$-norm (1.112). Moreover, in view of (1.123) and (1.119), for $u \in H^2(\Omega,\mathcal{T}_h)$ we have

$$\begin{aligned}\|u\|^2_{1,\sigma} =& |u|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(u,u) + \sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla uv\rangle)^2\mathrm{d}S\\ &\le \sum_{K\in\mathcal{T}_h}|u|^2_{H^1(K)} + \frac{C_W\,C_M}{C_T}\sum_{K\in\mathcal{T}_h}\Big(3h_K^{-2}\|u\|^2_{L^2(K)}+|u|^2_{H^1(K)}\Big)\\ &\quad+\frac{C_G\,C_M}{2C_W}\sum_{K\in\mathcal{T}_h}\Big(h_K^2|u|^2_{H^2(K)}+3|u|^2_{H^1(K)}\Big).\end{aligned}$$

Now, after a simple manipulation, we get

$$\|u\|_{1,\sigma}^2 \le \sum_{K\in\mathcal{T}_h} \Big(|u|_{H^1(K)}^2 \Big(1 + \frac{3C_G\,C_M}{2C_W} + \frac{C_W\,C_M}{C_T}\Big) + |u|_{H^2(K)}^2\, h_K^2\, \frac{C_G\,C_M}{2C_W} + \|u\|_{L^2(K)}^2\, h_K^{-2}\, \frac{3C_W\,C_M}{C_T}\Big).$$

Hence, (1.125) holds with

$$C_\sigma = \left(\max\left(1 + \frac{3C_G\,C_M}{2C_W} + \frac{C_W\,C_M}{C_T},\, \frac{C_G\,C_M}{2C_W},\, \frac{3C_W\,C_M}{C_T}\right)\right)^{1/2}.$$

Further, if $v_h \in S_{hp}$, then (1.112), (1.124) and (1.103) immediately imply (1.126) with $\tilde{C}_\sigma = (1 + C_G\,C_M(C_I+1)/C_W)^{1/2}$. □ □

In what follows, we shall be concerned with properties of the bilinear forms A_h defined by (1.47). First, we prove the continuity of the bilinear forms A_h defined by (1.47) in the space S_{hp} with respect to the norm $|||\cdot|||$.

Lemma 1.36. *Let assumptions* (1.104), (1.19) *and* (1.20) *be satisfied. Then there exists a constant $C_B > 0$ such that the form A_h defined by* (1.47) *satisfies the estimate*

$$|A_h(u_h, v_h)| \le C_B\, |||u_h|||\, |||v_h||| \qquad \forall\, u_h, v_h \in S_{hp}. \tag{1.128}$$

Proof. Estimates (1.122) and (1.126) give (1.128) with $C_B = 2\tilde{C}_\sigma^2$. □ □

Further, we shall prove an inequality similar to (1.128) replacing $u_h \in S_{hp}$ by $u \in H^2(\Omega, \mathcal{T}_h)$.

Lemma 1.37. *Let assumptions* (1.19), (1.20) *and* (1.104) *be satisfied. Then there exists a constant $\tilde{C}_B > 0$ such that*

$$|A_h(u, v_h)| \le \tilde{C}_B\, R_a(u)\, |||v_h||| \quad \forall\, u \in H^2(\Omega, \mathcal{T}_h)\ \forall\, v_h \in S_{hp}\ \forall\, h(0, \bar{h}), \tag{1.129}$$

where R_a is defined by (1.127).

Proof. By (1.122) and (1.125),

$$|A_h(u, v_h)| \le 2\|u\|_{1,\sigma}\|v_h\|_{1,\sigma} \le 2C_\sigma\tilde{C}_\sigma R_a(u)|||v_h|||,$$

which is (1.129) with $\tilde{C}_B = 2C_\sigma\tilde{C}_\sigma$. □ □

1.6.3 Coercivity of diffusion bilinear forms

Lemma 1.38 (NIPG coercivity)**.** *For any $C_W > 0$ the bilinear form $A_h^{\mathrm{n},\sigma}$ defined by* (1.47d) *satisfies the coercivity condition*

$$A_h^{\mathrm{n},\sigma}(v, v) \ge |||v|||^2 \qquad \forall\, v \in H^2(\Omega, \mathcal{T}_h). \tag{1.130}$$

Proof. From (1.45b) and (1.47d) it immediately follows that

$$A_h^{\mathrm{n},\sigma}(v, v) = a_h^{\mathrm{n}}(v, v) + J_h^\sigma(v, v) = |v|_{H^1(\Omega,\mathcal{T}_h)}^2 + J_h^\sigma(v, v) = |||v|||^2, \tag{1.131}$$

which we wanted to prove. □ □

The proof of the coercivity of the symmetric bilinear form $A_h^{\mathrm{s},\sigma}$ is more complicated.

Lemma 1.39 (SIPG coercivity)**.** *Let assumptions* (1.19) *and* (1.20) *be satisfied, let*

$$C_W \ge 4C_G C_M(1 + C_I), \tag{1.132}$$

where C_M, C_I and C_G are the constants from (1.78), (1.86) *and* (1.20)*, respectively, and let the penalty parameter σ be given by* (1.104) *for all $\Gamma \in \mathcal{F}_h^{ID}$. Then*

$$A_h^{\mathrm{s},\sigma}(v_h, v_h) \ge \frac{1}{2}|||v_h|||^2 \quad \forall\, v_h \in S_{hp}\ \forall\, h \in (0, \bar{h}).$$

Proof. Let $\delta > 0$. Then from (1.41), (1.104), (1.45a) and the Cauchy and Young's inequalities it follows that

$$a_h^{\mathrm{s}}(v_h, v_h) = |v_h|^2_{H^1(\Omega,\mathcal{T}_h)} - 2\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \boldsymbol{n}\cdot\langle\nabla v_h\rangle[v_h]\,\mathrm{d}S \tag{1.133}$$

$$\geq |v_h|^2_{H^1(\Omega,\mathcal{T}_h)} - 2\left\{\frac{1}{\delta}\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma h_\Gamma(\boldsymbol{n}\cdot\langle\nabla v_h\rangle)^2\,\mathrm{d}S\right\}^{\frac{1}{2}}\left\{\delta\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \frac{1}{h_\Gamma}[v_h]^2\,\mathrm{d}S\right\}^{\frac{1}{2}}$$

$$\geq |v_h|^2_{H^1(\Omega,\mathcal{T}_h)} - \omega - \frac{\delta}{C_W}J_h^\sigma(v_h, v_h),$$

where

$$\omega = \frac{1}{\delta}\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma h_\Gamma|\langle\nabla v_h\rangle|^2\,\mathrm{d}S. \tag{1.134}$$

Further, from assumption (1.20), inequality (1.107), the multiplicative trace inequality (1.78) and the inverse inequality (1.86) we get

$$\omega \leq \frac{C_G}{\delta}\sum_{K\in\mathcal{T}_h} h_K\|\nabla v_h\|^2_{L^2(\partial K)} \tag{1.135}$$

$$\leq \frac{C_G C_M}{\delta}\sum_{K\in\mathcal{T}_h} h_K\left(|v_h|_{H^1(K)}|\nabla v_h|_{H^1(K)} + h_K^{-1}|v_h|^2_{H^1(K)}\right)$$

$$\leq \frac{C_G C_M(1+C_I)}{\delta}|v_h|^2_{H^1(\Omega,\mathcal{T}_h)}.$$

Now let us choose

$$\delta = 2C_G C_M(1+C_I). \tag{1.136}$$

Then it follows from (1.132) and (1.133) – (1.136) that

$$a_h^{\mathrm{s}}(v_h, v_h) \geq \frac{1}{2}\left(|v_h|^2_{H^1(\Omega,\mathcal{T}_h)} - \frac{4C_G C_M(1+C_I)}{C_W}J_h^\sigma(v_h, v_h)\right) \tag{1.137}$$

$$\geq \frac{1}{2}\left(|v_h|^2_{H^1(\Omega,\mathcal{T}_h)} - J_h^\sigma(v_h, v_h)\right).$$

Finally, definition (1.47c) of the form $A_h^{\mathrm{s},\sigma}$ and (1.137) imply that

$$A_h^{\mathrm{s},\sigma}(v_h, v_h) = a_h^{\mathrm{s}}(v_h, v_h) + J_h^\sigma(v_h, v_h) \tag{1.138}$$

$$\geq \frac{1}{2}\left(|v_h|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(v_h, v_h)\right) = \frac{1}{2}|||v_h|||^2,$$

which we wanted to prove. □ □

Lemma 1.40 (IIPG coercivity)**.** *Let assumptions* (1.19) *and* (1.20) *be satisfied, let*

$$C_W \geq C_G C_M(1+C_I), \tag{1.139}$$

where C_M, C_I and C_G are constants from (1.78), (1.86) *and* (1.20)*, respectively, and let the penalty parameter σ be given by* (1.104) *for all $\Gamma \in \mathcal{F}_h^{ID}$. Then*

$$A_h^{\mathrm{i},\sigma}(v_h, v_h) \geq \frac{1}{2}|||v_h|||^2 \quad \forall\, v_h \in S_{hp}.$$

Proof. The proof is almost identical with the proof of the previous lemma. □ □

Corollary 1.41. *We can summarize the above results in the following way. We have*

$$A_h(v_h, v_h) \geq C_C|||v_h|||^2 \quad \forall\, v_h \in S_{hp}, \tag{1.140}$$

with

$$\begin{array}{lll} C_C = 1 & \text{for } A_h = A_h^{\mathrm{n},\sigma} & \text{if } C_W > 0, \\ C_C = 1/2 & \text{for } A_h = A_h^{\mathrm{s},\sigma} & \text{if } C_W \geq 4C_G C_M(1+C_I), \\ C_C = 1/2 & \text{for } A_h = A_h^{\mathrm{i},\sigma} & \text{if } C_W \geq C_G C_M(1+C_I). \end{array}$$

Corollary 1.42. *By virtue of Corollary 0.7, the coercivity of the forms A_h implies the* existence and uniqueness *of the solution of the discrete problems* (1.49c) – (1.49e) *(SIPG, NIPG and IIPG method).*

1.7 Error estimates

In this section, we derive error estimates of the SIPG, NIPG and IIPG variants of the DGM applied to the numerical solution of the Poisson problem (1.1). Namely, the error $u_h - u$ will be estimated in the DG-norm and the $L^2(\Omega)$-norm.

1.7.1 Estimates in the DG-norm

Let $u \in H^2(\Omega)$ denote the exact strong solution of problem (1.1) and let and $u_h \in S_{hp}$ be the approximate solution obtained by method (1.54), where the forms A_h and ℓ_h are defined by (1.47c) – (1.47e) and (1.48c) – (1.48e), respectively. The error of the method is defined as the function $e_h = u_h - u \in H^2(\Omega, \mathcal{T}_h)$. It can be written in the form

$$e_h = \xi + \eta, \quad \text{with } \xi = u_h - \Pi_{hp} u \in S_{hp},\ \eta = \Pi_{hp} u - u \in H^2(\Omega, \mathcal{T}_h), \tag{1.141}$$

where Π_{hp} is the S_{hp}-interpolation defined by (1.90). Hence, we split the error into two parts ξ and η. The term η represents the error of the S_{hp}-interpolation of the function u. (It is possible to say that η approximates the *distance* of the exact solution from the space S_{hp}, where the approximate solution is sought.) The term η can be simply estimated on the basis of the approximation properties (1.92) and (1.97). On the other hand, the term ξ represents the *distance* between the approximate solution u_h and the projection of the exact solution on the space S_{hp}. The estimation of ξ is sometimes more complicated.

We shall suppose that the system of triangulations $\{\mathcal{T}_h\}_{h \in (0,\bar{h})}$ satisfies the shape-regularity assumptions (1.19) and that the equivalence condition (1.20) holds.

First, we shall prove the so-called *abstract error estimate*, representing a bound of the error in terms of the S_{hp}-interpolation error η.

Theorem 1.43. *Let assumptions* (1.19) *and* (1.20) *be satisfied and let the exact solution of problem* (1.1) *satisfy the condition* $u \in H^2(\Omega)$. *Then there exists a constant* $C_{\mathrm{AE}} > 0$ *such that*

$$|||e_h||| \le C_{\mathrm{AE}}\, R_a(\eta) = C_{\mathrm{AE}}\, R_a(\Pi_{hp} u - u), \quad h \in (0, \bar{h}), \tag{1.142}$$

where $R_a(\eta)$ *is given by* (1.127).

Proof. We express the error by (1.141), i.e., $e_h = u_h - u = \xi + \eta$. The error e_h satisfies the Galerkin orthogonality condition (1.57), which is equivalent to the relation

$$A_h(\xi, v_h) = -A_h(\eta, v_h) \qquad \forall\, v_h \in S_{hp}. \tag{1.143}$$

If we set $v_h := \xi \in S_{hp}$ in (1.143) and use (1.47c) – (1.47e) and the coercivity (1.140), we find that

$$C_C |||\xi|||^2 \le A_h(\xi, \xi) = -A_h(\eta, \xi). \tag{1.144}$$

Now we apply Lemma 1.37 and get

$$|A_h(\eta, \xi)| \le \tilde{C}_B\, R_a(\eta)\, |||\xi|||.$$

The above and (1.144) already imply that

$$|||\xi||| \le \frac{\tilde{C}_B}{C_C}\, R_a(\eta). \tag{1.145}$$

Obviously,

$$|||e_h||| \le |||\xi||| + |||\eta|||. \tag{1.146}$$

Finally, (1.125) gives

$$|||\eta||| \le C_\sigma R_a(\eta). \tag{1.147}$$

Hence, (1.146), (1.145) and (1.147) yield the abstract error estimate (1.142) with $C_{\mathrm{AE}} = C_\sigma + \tilde{C}_B / C_C$. □ □

The abstract error estimate is the basis for estimating the error e_h in terms of the mesh-size h.

Theorem 1.44 (DG-norm error estimate)**.** *Let us assume that* $s \ge 2$, $p \ge 1$, *are integers,* $u \in H^s(\Omega)$ *is the solution of problem* (1.1), $\{\mathcal{T}_h\}_{h \in (0,\bar{h})}$ *is a system of triangulations of the domain* Ω *satisfying the shape-regularity condition* (1.19), *and the equivalence condition* (1.20) *(cf. Lemma 1.5). Moreover, let the penalty constant* C_W *satisfy the conditions from Corollary*

1.41. Let $u_h \in S_{hp}$ be the approximate solution obtained by using of the SIPG, NIPG or IIPG method (1.49c) – (1.49e). *Then the error $e_h = u_h - u$ satisfies the estimate*

$$|||e_h||| \le C_1 h^{\mu-1} |u|_{H^\mu(\Omega)}, \quad h \in (0, \bar{h}), \tag{1.148}$$

where $\mu = \min(p+1, s)$ and C_1 is a constant independent of h and u. Hence, if $s \ge p+1$, we get the error estimate

$$|||e_h||| \le C_1 h^p |u|_{H^{p+1}(\Omega)}.$$

Proof. It is enough to use the abstract error estimate (1.142), where the expressions $|\eta|_{H^1(K)}$, $|\eta|_{H^2(K)}$ and $\|\eta\|_{L^2(K)}$, $K \in \mathcal{T}_h$, are estimated on the basis of the approximation properties (1.93) – (1.95), rewritten for $\eta|_K = (\Pi_{hp} u - u)|_K = \pi_{K,p}(u|_K) - u|_K$ and $K \in \mathcal{T}_h$:

$$\begin{aligned} \|\eta\|_{L^2(K)} &\le C_A\, h_K^{\mu} |u|_{H^\mu(K)}, \\ |\eta|_{H^1(K)} &\le C_A\, h_K^{\mu-1} |u|_{H^\mu(K)}, \\ |\eta|_{H^2(K)} &\le C_A\, h_K^{\mu-2} |u|_{H^\mu(K)}. \end{aligned} \tag{1.149}$$

Thus, the inequality $h_K \le h$ and the relation $\sum_{K\in\mathcal{T}_h} |u|^2_{H^\mu(K)} = |u|^2_{H^\mu(\Omega)}$ imply

$$\begin{aligned} R_a(\eta) &= \left(\sum_{K\in\mathcal{T}_h} \left(|\eta|^2_{H^1(K)} + h_K^2 |\eta|^2_{H^2(K)} + h_K^{-2} \|\eta\|^2_{L^2(K)} \right) \right)^{1/2} \\ &\le \sqrt{3} C_A h^{\mu-1} |u|_{H^\mu(\Omega)}, \end{aligned} \tag{1.150}$$

which together with (1.142) gives (1.148) with the constant $C_1 = \sqrt{3} C_{\mathrm{AE}}\, C_A$. □ □

In order to derive an error estimate in the $L^2(\Omega)$-norm we present the following result.

Lemma 1.45 (Broken Poincaré inequality)**.** *Let the system $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ of triangulations satisfy the shape-regularity assumption* (1.19). *Then there exists a constant $C > 0$ independent of h and v_h such that*

$$\|v_h\|^2_{L^2(\Omega)} \le C \left(\sum_{K\in\mathcal{T}_h} |v_h|^2_{H^1(K)} + \sum_{\Gamma\in\mathcal{F}_h^{ID}} \frac{1}{\mathrm{diam}(\Gamma)} \|[v_h]\|^2_{L^2(\Gamma)} \right) \tag{1.151}$$
$$\forall\, v_h \in S_{hp}\ \forall\, h \in (0, \bar{h}).$$

The proof of the broken Poincaré inequality (1.151) was carried out in [Arn82] in the case where Ω is a convex polygonal domain, $\partial\Omega_D = \partial\Omega$ and the assumption (MA2) in Section 1.3.2 is satisfied. The proof of inequality (1.151) in a general case with the nonempty Neumann part of the boundary can be found in [Bre03].

From Theorem 1.44 and (1.151) we obtain the following result.

Corollary 1.46 ($L^2(\Omega)$-(suboptimal) error estimate)**.** *Let the assumptions of Theorem 1.44 be satisfied. Then*

$$\|e_h\|_{L^2(\Omega)} \le C_2 h^{\mu-1} |u|_{H^\mu(\Omega)}, \quad h \in (0, \bar{h}), \tag{1.152}$$

where C_2 is a constant independent of h. Hence, if $s \ge p+1$, we get the error estimate

$$\|e_h\|_{L^2(\Omega)} \le C_2 h^p |u|_{H^{p+1}(\Omega)}. \tag{1.153}$$

Remark 1.47. *The error estimate* (1.153)*, which is of order $O(h^p)$, is suboptimal with respect to the approximation property* (1.97) *with $q = 0$, $\mu = p+1 \le s$ of the space S_{hp} giving the order $O(h^{p+1})$. In the next section we shall prove an optimal error estimate in the $L^2(\Omega)$-norm for SIPG method using the Aubin–Nitsche technique.*

1.7.2 Optimal $L^2(\Omega)$-error estimate

Our further aim is to derive the optimal error estimate in the $L^2(\Omega)$-norm. It will be based on the *duality technique* sometimes called the *Aubin–Nitsche trick*. Since this approach requires the symmetry of the corresponding bilinear form and the regularity of the exact solution to the dual problem, we shall consider the SIPG method applied to problem (1.1) with $\partial\Omega_D = \partial\Omega$ and $\partial\Omega_N = \emptyset$. This means that we seek u satisfying

$$-\Delta u = f \quad \text{in } \Omega, \tag{1.154a}$$
$$u = u_D \quad \text{on } \partial\Omega. \tag{1.154b}$$

Moreover, for an arbitrary $z \in L^2(\Omega)$, we shall consider the *dual problem*: Given $z \in L^2(\Omega)$, find ψ such that

$$-\Delta\psi = z \quad \text{in } \Omega, \quad \psi = 0 \quad \text{on } \partial\Omega. \tag{1.155}$$

Under the notation

$$V = H_0^1(\Omega) = \left\{v \in H^1(\Omega);\ v = 0 \text{ on } \partial\Omega\right\}, \tag{1.156}$$

the weak formulation of (1.155) reads: Find $\psi \in V$ such that

$$\int_\Omega \nabla\psi \cdot \nabla v \,\mathrm{d}x = \int_\Omega zv \,\mathrm{d}x = (z, v)_{L^2(\Omega)} \quad \forall\ v \in V. \tag{1.157}$$

Let us assume that $\psi \in H^2(\Omega)$ and that there exists a constant $C_D > 0$, independent of z, such that

$$\|\psi\|_{H^2(\Omega)} \le C_D \|z\|_{L^2(\Omega)}. \tag{1.158}$$

This is true provided the polygonal (polyhedral) domain Ω is convex, as follows from [Gri92]. (See Remark 1.50.) Let us note that $H^2(\Omega) \subset C(\overline{\Omega})$, if $d \le 3$.

Let A_h be the symmetric bilinear form given by (1.47c), i.e.,

$$A_h(u, v) = a_h^{\mathrm{s}}(u, v) + J_h^\sigma(u, v), \qquad u, v \in H^2(\Omega, \mathcal{T}_h), \tag{1.159}$$

where a_h^{s} and J_h^σ are defined by (1.45a) and (1.105), respectively.

First, we shall prove the following auxiliary result.

Lemma 1.48. *Let $\psi \in H^2(\Omega)$ be the solution of problem* (1.155)*. Then*

$$A_h(\psi, v) = (v, z)_{L^2(\Omega)} \quad \forall\, v \in H^2(\Omega, \mathcal{T}_h). \tag{1.160}$$

Proof. The function $\psi \in H^2(\Omega)$ satisfies the conditions

$$[\psi]_\Gamma = 0 \qquad \forall\, \Gamma \in \mathcal{F}_h^I, \qquad \psi|_{\partial\Omega} = 0. \tag{1.161}$$

Let $v \in H^2(\Omega, \mathcal{T}_h)$. Using (1.155), (1.161) and Green's theorem, we obtain

$$\begin{aligned}
(v, z)_{L^2(\Omega)} &= \int_\Omega zv \,\mathrm{d}x = -\int_\Omega \Delta\psi v \,\mathrm{d}x \\
&= \sum_{K\in\mathcal{T}_h} \int_K \nabla\psi \cdot \nabla v \,\mathrm{d}x - \sum_{K\in\mathcal{T}_h} \int_{\partial K} \nabla\psi \cdot \boldsymbol{n}\, v \,\mathrm{d}S \\
&= \sum_{K\in\mathcal{T}_h} \int_K \nabla\psi \cdot \nabla v \,\mathrm{d}x \\
&\quad - \left(\sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \langle\nabla\psi\rangle \cdot \boldsymbol{n}\, [v] \,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \langle\nabla v\rangle \cdot \boldsymbol{n}\, [\psi] \,\mathrm{d}S \right) \\
&\quad - \left(\sum_{\Gamma\in\mathcal{F}_h^B} \int_\Gamma \nabla\psi \cdot \boldsymbol{n}\, v \,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^B} \int_\Gamma \nabla v \cdot \boldsymbol{n}\, \psi \,\mathrm{d}S \right) \\
&\quad + \left(\sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \sigma\, [\psi]\, [v] \,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^B} \int_\Gamma \sigma\, \psi\, v \,\mathrm{d}S \right).
\end{aligned}$$

Hence, in view of the definition of the form A_h, we have (1.160). □ □

Theorem 1.49 ($L^2(\Omega)$-optimal error estimate)**.** *Let us assume that $s \ge 2$, $p \ge 1$, are integers, Ω is a bounded convex polyhedral domain, $u \in H^s(\Omega)$ is the solution of problem* (1.1)*, $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ is a system of triangulations of the domain Ω satisfying the shape-regularity condition* (1.19)*, and the equivalence condition* (1.20) *(cf. Lemma 1.5). Moreover, let the penalty constant C_W satisfy the condition from Corollary 1.41. Let $u_h \in S_{hp}$ be the approximate solution obtained using the SIPG method* (1.49c) *(i.e., $\Theta = 1$ and the form $A_h = A_h^{\sigma,s}$ is given by* (1.45a) *and* (1.47c)*. Then*

$$\|e_h\|_{L^2(\Omega)} \le C_3 h^\mu |u|_{H^\mu(\Omega)}, \tag{1.162}$$

where $e_h = u_h - u$, $\mu = \min\{p + 1, s\}$ and C_3 is a constant independent of h and u.

Proof. Let $\psi \in H^2(\Omega)$ be the solution of the dual problem (1.157) with $z := e_h = u_h - u \in L^2(\Omega)$ and let $\Pi_{h1}\psi \in S_{h1}$ be the approximation of ψ defined by (1.90) with $p = 1$. By (1.160), we have

$$A_h(\psi, v) = (e_h, v)_{L^2(\Omega)} \qquad \forall\, v \in H^2(\Omega, \mathcal{T}_h). \tag{1.163}$$

The symmetry of the form A_h, the Galerkin orthogonality (1.57) of the error and (1.163) with $v := e_h$ yield

$$\begin{aligned} \|e_h\|^2_{L^2(\Omega)} &= A_h(\psi, e_h) = A_h(e_h, \psi) \\ &= A_h(e_h, \psi - \Pi_{h1}\psi). \end{aligned} \tag{1.164}$$

Moreover, from (1.122), it follows that

$$A_h(e_h, \psi - \Pi_{h1}\psi) \le 2\|e_h\|_{1,\sigma}\, \|\psi - \Pi_{h1}\psi\|_{1,\sigma}, \tag{1.165}$$

where, by (1.112),

$$\|v\|^2_{1,\sigma} = |||v|||^2 + \sum_{\Gamma \in \mathcal{F}_h^{ID}} \int_\Gamma \sigma^{-1} (\boldsymbol{n} \cdot \langle \nabla v \rangle)^2 \,\mathrm{d}S. \tag{1.166}$$

By (1.125) and (1.150) (with $\mu = 2$), we have

$$\|\psi - \Pi_{h1}\psi\|_{1,\sigma} \le C_\sigma\, R_a(\psi - \Pi_{h1}\psi) \le \sqrt{3} C_\sigma C_A h |\psi|_{H^2(\Omega)}. \tag{1.167}$$

Now, the inverse inequality (1.86) and estimates (1.100), (1.99) imply that

$$\begin{aligned} |\nabla e_h|_{H^1(K)} &= |\nabla(u - u_h)|_{H^1(K)} \\ &\le |\nabla(u - \Pi_{hp}u)|_{H^1(K)} + |\nabla(\Pi_{hp}u - u_h)|_{H^1(K)} \\ &\le |u - \Pi_{hp}u|_{H^2(K)} + C_I h_K^{-1} \|\nabla(\Pi_{hp}u - u_h)\|_{L^2(K)} \\ &\le C_A h_K^{\mu-2} |u|_{H^\mu(K)} + C_I h_K^{-1} \left(\|\nabla(\Pi_{hp}u - u)\|_{L^2(K)} + \|\nabla(u - u_h)\|_{L^2(K)} \right) \\ &\le C_A (1 + C_I) h_K^{\mu-2} |u|_{H^\mu(K)} + C_I h_K^{-1} \|\nabla e_h\|_{L^2(K)}. \end{aligned} \tag{1.168}$$

By (1.123), (1.168) and the discrete Cauchy inequality,

$$\begin{aligned} &\sum_{\Gamma \in \mathcal{F}_h^{ID}} \int_\Gamma \sigma^{-1} (\boldsymbol{n} \cdot \langle \nabla e_h \rangle)^2 \,\mathrm{d}S \\ &\quad\le \frac{C_G C_M}{C_W} \sum_{K \in \mathcal{T}_h} \left(h_K \|\nabla e_h\|_{L^2(K)}\, |\nabla e_h|_{H^1(K)} + \|\nabla e_h\|^2_{L^2(K)} \right) \\ &\quad\le \frac{C_G C_M}{C_W} \left\{ C_A (1 + C_I) h^{\mu-1} |e_h|_{H^1(\Omega, \mathcal{T}_h)} |u|_{H^\mu(\Omega)} + (1 + C_I) |e_h|^2_{H^1(\Omega, \mathcal{T}_h)} \right\}. \end{aligned} \tag{1.169}$$

Since $|e_h|_{H^1(\Omega, \mathcal{T}_h)} \le |||e_h|||$, using (1.148) and (1.169), we have

$$\sum_{\Gamma \in \mathcal{F}_h^{ID}} \int_\Gamma \sigma^{-1} (\boldsymbol{n} \cdot \langle \nabla e_h \rangle)^2 \,\mathrm{d}S \le \frac{C_G C_M}{C_W} C_1 (1 + C_I)(C_1 + C_A) h^{2(\mu-1)} |u|^2_{H^\mu(\Omega)}.$$

Thus, (1.148) and (1.166) yield the estimate

$$\|e_h\|^2_{1,\sigma} \le C_5 h^{2(\mu-1)} |u|^2_{H^\mu(\Omega)} \tag{1.170}$$

with $C_5 = C_1 \left\{ 1 + C_G C_M C_W^{-1} (1 + C_I)(C_1 + C_A) \right\}$. It follows from (1.165), (1.167), and (1.170) that

$$A_h(e_h, \psi - \Pi_{h1}\psi) \le C_6 h^\mu |\psi|_{H^2(\Omega)}\, |u|_{H^\mu(\Omega)}, \tag{1.171}$$

where $C_6 = 2\sqrt{3} C_\sigma C_A \sqrt{C_5}$.

Finally, by (1.164), (1.171), and (1.158) with $z = e_h$,

$$\|e_h\|^2_{L^2(\Omega)} \le C_D C_6 h^\mu |u|_{H^\mu(\Omega)} \|e_h\|_{L^2(\Omega)}, \tag{1.172}$$

which already implies estimate (1.162) with $C_3 = C_D C_6$. □ □

Remark 1.50. *As we see from the above results, if the exact solution $u \in H^{p+1}(\Omega)$ and the finite elements of degree p are used, the error is of the optimal order $O(h^{p+1})$ in the $L^2(\Omega)$-norm. In the case, when the polygonal domain is not convex and/or the Neumann and Dirichlet parts of the boundary $\Omega_N \neq \emptyset$ and $\Omega_D \neq \emptyset$, the exact solution ψ of the dual problem (1.155) is not an element of the space $H^2(\Omega)$. Then it is necessary to work in the Sobolev–Slobodetskii spaces of functions with* noninteger derivatives *and the error in the $L^2(\Omega)$-norm is not of the optimal order $O(h^{p+1})$. The analysis of error estimates for the DG discretization of boundary value problems with boundary singularities is the subject of works [Wih02] and [FS12], where optimal error estimates were obtained with the aid of a suitable graded mesh refinement. The main tools are here the Sobolev–Slobodetskii spaces and weighted Sobolev spaces. For the definitions and properties of these spaces, see [BS94b] and [KS87].*

Remark 1.51. *In [RWG01] the Neumann problem (i.e., $\partial\Omega = \partial\Omega_N$) was solved by the NIPG approach, where the penalty coefficient σ was chosen in the form*

$$\sigma|_\Gamma = \frac{C_W}{h_\Gamma^\beta}, \qquad \Gamma \in \mathcal{F}_h, \tag{1.173}$$

instead of (1.104), where $\beta \geq 1/2$. If triangular grids do not contain any hanging nodes (i.e., the triangulations $\mathcal{T}_h$ are conforming), then an optimal error estimate in the $L^2(\Omega)$-norm of this analogue of the NIPG method was proven provided that $\beta \geq 3$ for $d = 2$ and $\beta \geq 3/2$ for $d = 3$. In this case the interior penalty is so strong that the DG methods behave like the standard conforming (i.e., continuous) finite element schemes. On the other hand, the stronger penalty causes worse computational properties of the corresponding algebraic system, see [Cas02].

1.8 Numerical examples

In this section, we demonstrate by numerical experiments the error estimates (1.148), (1.152) and (1.162). In the first example, we assume that the exact solution is sufficiently regular. We show that the use of a higher degree of polynomial approximation increases the rate of convergence of the method. In the second example, the exact solution has a singularity. Then the order of convergence does not increase with the increasing degree of the polynomial approximation used. The computational results are in agreement with theory and show that the accuracy of the method is determined by the degree of the polynomial approximation as well as the regularity of the solution.

1.8.1 Regular solution

Let us consider the problem of finding a function $u : \Omega = (0,1) \times (0,1) \to \mathbb{R}$ such that

$$\begin{aligned} -\Delta u &= 8\pi^2 \sin(2\pi x_1)\sin(2\pi x_2) \quad \text{in } \Omega, \\ u &= 0 \qquad\qquad\qquad\qquad\quad \text{on } \partial\Omega. \end{aligned} \tag{1.174}$$

It is easy to verify that the exact solution of (1.174) has the form

$$u = \sin(2\pi x_1)\sin(2\pi x_2), \quad (x_1, x_2) \in \Omega. \tag{1.175}$$

Obviously, $u \in C^\infty(\overline{\Omega})$.

We investigate the *experimental order of convergence* (EOC) of the SIPG, NIPG and IIPG methods defined by (1.49c) – (1.49e). We assume that a (semi)norm $\|e_h\|$ of the computational error behaves according to the formula

$$\|e_h\| = Ch^{\mathrm{EOC}}, \tag{1.176}$$

where $C > 0$ is a constant, $h = \max_{K\in\mathcal{T}_h} h_K$, and $\mathrm{EOC} \in \mathbb{R}$ is the experimental order of convergence. Since the exact solution is known and therefore $\|e_h\|$ can be exactly evaluated, it is possible to evaluate EOC in the following way. Let $\|e_{h_1}\|$ and $\|e_{h_2}\|$ be computational errors of the numerical solutions obtained on two different meshes $\mathcal{T}_{h_1}$ and $\mathcal{T}_{h_2}$, respectively. Then from (1.176), eliminating the constant C, we obtain

$$\mathrm{EOC} = \frac{\log(\|e_{h_1}\|/\|e_{h_2}\|)}{\log(h_1/h_2)}. \tag{1.177}$$

Moreover, we evaluate the *global experimental order of convergence* (GEOC) from the approximation of (1.176) with the aid of the least squares method, where all computed pairs $[h, e_h]$ are taken into account simultaneously.

We used a set of four uniform triangular grids having 128, 512, 2048, and 8192 elements, shown in Figure 1.4. The meshes consist of right-angled triangles with the diameter $h = \sqrt{2}/\sqrt{\#\mathcal{T}_h/2}$, where $\#\mathcal{T}_h$ is the number of elements of $\mathcal{T}_h$. EOC is evaluated according to (1.177) for all pairs of "neighbouring" grids. Tables 1.1 – 1.2 show the computational errors in the

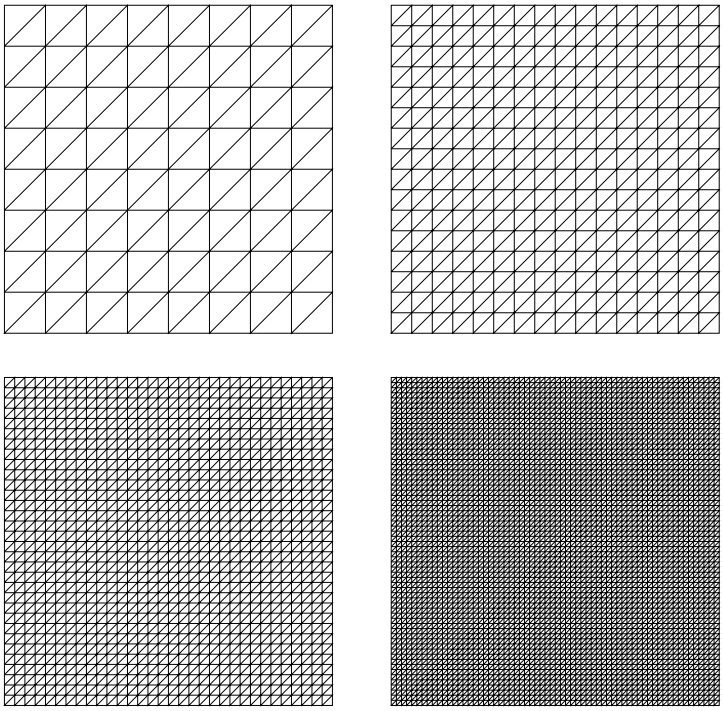

Figure 1.4: Computational grids used for the numerical solution of problems (1.174) and (1.179).

$L^2(\Omega)$-norm and the $H^1(\Omega, \mathcal{T}_h)$-seminorm and EOC obtained by the SIPG, NIPG and IIPG methods using the $P_p,\ p = 1, \ldots, 6$, polynomial approximations. These results are also visualized in Figure 1.5.

We observe that EOC of the SIPG technique are in a good agreement with the theoretical ones, i.e., $O(h^{p+1})$ in the $L^2(\Omega)$-norm (estimate (1.162)) and $O(h^p)$ in the $H^1(\Omega, \mathcal{T}_h)$-seminorm (estimate (1.148)). On the other hand, the experimental order of convergence of the NIPG and IIPG techniques measured in the $L^2(\Omega)$-norm is better than the theoretical estimate (1.152). We deduce that

$$\|e_h\|_{L^2(\Omega)} = O(h^{\bar{p}}), \quad \bar{p} = \begin{cases} p+1 & \text{for } p \text{ odd}, \\ p & \text{for } p \text{ even}. \end{cases} \tag{1.178}$$

This interesting property of the NIPG and IIPG techniques was observed by many authors (cf. [OBB98] and [HSS02]), but up to now a theoretical justification has been missing, see Section 1.8.3 for some comments. The EOC in the $H^1(\Omega, \mathcal{T}_h)$-seminorm of NIPG and IIPG methods is in agreement with (1.148).

1.8.2 Singular case

In the domain $\Omega = (0,1) \times (0,1)$ we consider the Poisson problem

$$\begin{aligned} -\Delta u &= g \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{1.179}$$

with the right-hand side g chosen in such a way that the exact solution has the form

$$u(x_1, x_2) = 2r^\alpha x_1 x_2 (1-x_1)(1-x_2) = r^{\alpha+2} \sin(2\varphi)(1-x_1)(1-x_2), \tag{1.180}$$

where r, φ are the polar coordinates ($r = (x_1^2 + x_2^2)^{1/2}$) and $\alpha \in \mathbb{R}$ is a constant. The function u is equal to zero on $\partial\Omega$ and its regularity depends on the value of α. Namely, by [BS90],

$$u \in H^\beta(\Omega) \quad \forall\, \beta \in (0,\, \alpha + 3), \tag{1.181}$$

where $H^\beta(\Omega)$ denotes the Sobolev–Slobodetskii space of functions with *noninteger derivatives*.

We present numerical results obtained for $\alpha = -3/2$ and $\alpha = 1/2$. If $\alpha = -3/2$, then $u \in H^\beta(\Omega)$ for all $\beta \in (0, 3/2)$, whereas for the value $\alpha = 1/2$, we have $u \in H^\beta(\Omega)$ for all $\beta \in (0, 7/2)$. Figure 1.6 shows the function u for both values of α.

		SIPG		NIPG		IIPG	
p	$h/\sqrt{2}$	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC
1	1/8	6.7452E-02	–	2.9602E-02	–	6.3939E-02	–
1	1/16	1.8745E-02	1.85	7.6200E-03	1.96	1.7383E-02	1.88
1	1/32	4.8463E-03	1.95	1.9292E-03	1.98	4.4579E-03	1.96
1	1/64	1.2252E-03	1.98	4.8536E-04	1.99	1.1239E-03	1.99
	GEOC		1.93		1.98		1.95
2	1/8	3.9160E-03	–	1.0200E-02	–	4.7447E-03	–
2	1/16	4.9164E-04	2.99	2.5723E-03	1.99	8.4877E-04	2.48
2	1/32	6.1644E-05	3.00	6.4259E-04	2.00	1.8081E-04	2.23
2	1/64	7.7184E-06	3.00	1.6032E-04	2.00	4.2670E-05	2.08
	GEOC		3.00		2.00		2.26
3	1/8	3.1751E-04	–	5.5550E-04	–	3.2684E-04	–
3	1/16	1.9150E-05	4.05	3.4481E-05	4.01	2.0077E-05	4.02
3	1/32	1.1775E-06	4.02	2.1333E-06	4.01	1.2414E-06	4.02
3	1/64	7.3124E-08	4.01	1.3250E-07	4.01	7.7176E-08	4.01
	GEOC		4.03		4.01		4.02
4	1/8	2.3496E-05	–	3.7990E-05	–	2.7046E-05	–
4	1/16	7.5584E-07	4.96	2.4304E-06	3.97	1.2929E-06	4.39
4	1/32	2.3824E-08	4.99	1.5512E-07	3.97	7.2190E-08	4.16
4	1/64	7.4627E-10	5.00	9.7626E-09	3.99	4.3310E-09	4.06
	GEOC		4.98		3.97		4.20
5	1/8	1.4133E-06	–	2.3017E-06	–	1.6501E-06	–
5	1/16	2.2193E-08	5.99	3.6590E-08	5.98	2.6160E-08	5.98
5	1/32	3.4686E-10	6.00	5.7147E-10	6.00	4.0753E-10	6.00
5	1/64	5.4139E-12	6.00	8.8468E-12	6.01	6.3670E-12	6.00
	GEOC		6.00		6.00		6.00
6	1/8	7.3313E-08	–	1.1239E-07	–	9.5990E-08	–
6	1/16	5.8381E-10	6.97	1.5138E-09	6.21	1.1620E-09	6.37
6	1/32	4.5855E-12	6.99	2.2864E-11	6.05	1.6380E-11	6.15
6	1/64	3.8771E-14	6.89	3.5354E-13	6.02	2.4417E-13	6.07
	GEOC		6.95		6.09		6.19

Table 1.1: Computational error and EOC in the $L^2(\Omega)$-norm for the regular solution of problem (1.174).

		SIPG		NIPG		IIPG	
p	$h/\sqrt{2}$	$\lvert e_h\rvert_{H^1(\Omega,\mathcal{T}_h)}$	EOC	$\lvert e_h\rvert_{H^1(\Omega,\mathcal{T}_h)}$	EOC	$\lvert e_h\rvert_{H^1(\Omega,\mathcal{T}_h)}$	EOC
1	1/8	1.5018E+00	–	1.2423E+00	–	1.4946E+00	–
1	1/16	7.7679E-01	0.95	6.4615E-01	0.94	7.7519E-01	0.95
1	1/32	3.9214E-01	0.99	3.2741E-01	0.98	3.9181E-01	0.98
1	1/64	1.9666E-01	1.00	1.6450E-01	0.99	1.9658E-01	1.00
	GEOC		0.98		0.97		0.98
2	1/8	2.4259E-01	–	1.9985E-01	–	2.1634E-01	–
2	1/16	6.2760E-02	1.95	5.0217E-02	1.99	5.5693E-02	1.96
2	1/32	1.5849E-02	1.99	1.2536E-02	2.00	1.4053E-02	1.99
2	1/64	3.9743E-03	2.00	3.1305E-03	2.00	3.5244E-03	2.00
	GEOC		1.98		2.00		1.98
3	1/8	2.5610E-02	–	2.4029E-02	–	2.3425E-02	–
3	1/16	3.2202E-03	2.99	3.0531E-03	2.98	2.9699E-03	2.98
3	1/32	4.0238E-04	3.00	3.8298E-04	2.99	3.7253E-04	3.00
3	1/64	5.0260E-05	3.00	4.7890E-05	3.00	4.6607E-05	3.00
	GEOC		3.00		2.99		2.99
4	1/8	2.2049E-03	–	2.2096E-03	–	2.0645E-03	–
4	1/16	1.4023E-04	3.97	1.3801E-04	4.00	1.3039E-04	3.98
4	1/32	8.8035E-06	3.99	8.5962E-06	4.00	8.1650E-06	4.00
4	1/64	5.5077E-07	4.00	5.3601E-07	4.00	5.1038E-07	4.00
	GEOC		3.99		4.00		3.99
5	1/8	1.5680E-04	–	1.6457E-04	–	1.5090E-04	–
5	1/16	4.9305E-06	4.99	5.1666E-06	4.99	4.7527E-06	4.99
5	1/32	1.5413E-07	5.00	1.6126E-07	5.00	1.4865E-07	5.00
5	1/64	4.8146E-09	5.00	5.0316E-09	5.00	4.6439E-09	5.00
	GEOC		5.00		5.00		5.00
6	1/8	9.5245E-06	–	1.0198E-05	–	9.3719E-06	–
6	1/16	1.5092E-07	5.98	1.5951E-07	6.00	1.4762E-07	5.99
6	1/32	2.3666E-09	5.99	2.4862E-09	6.00	2.3083E-09	6.00
6	1/64	3.7008E-11	6.00	3.8770E-11	6.00	3.6051E-11	6.00
	GEOC		5.99		6.00		6.00

Table 1.2: Computational error and EOC in the $H^1(\Omega, \mathcal{T}_h)$-seminorm for the regular solution of problem (1.174).

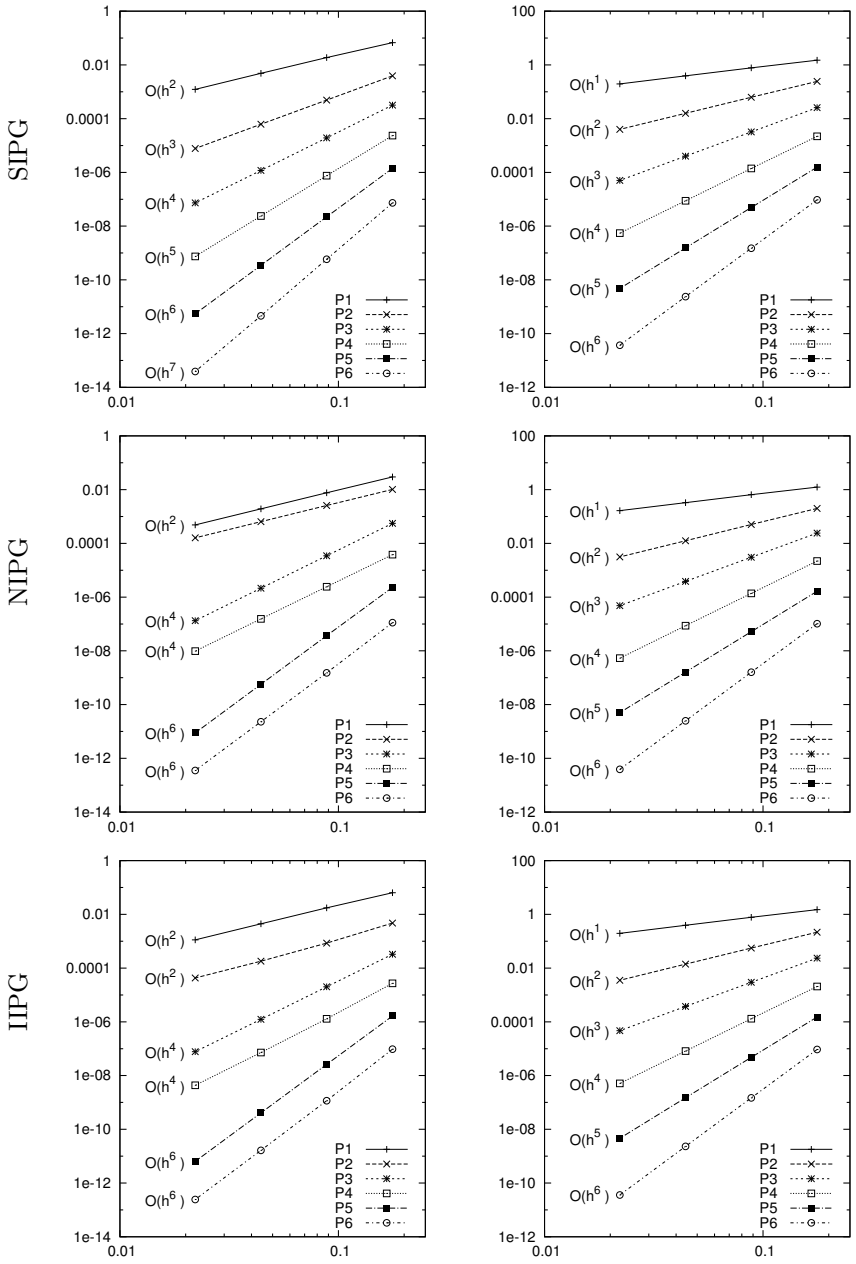

Figure 1.5: Computational error and EOC in the $L^2(\Omega)$-norm (left) and in the $H^1(\Omega, \mathcal{T}_h)$-seminorm (right) for the regular solution of problem (1.174).

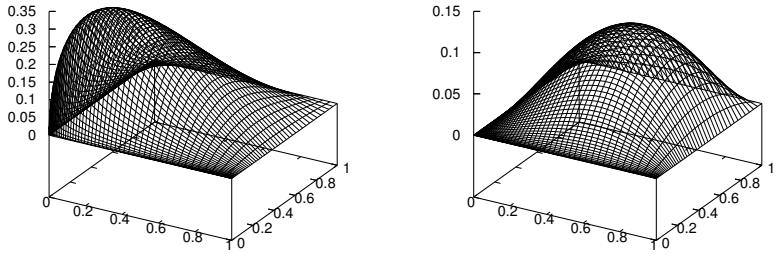

Figure 1.6: Exact solution (1.180) for $\alpha = -3/2$ (left) and $\alpha = 1/2$ (right).

We carried out computations on 4 triangular grids introduced in Section 1.8.1 by the SIPG, NIPG and IIPG technique with the aid of $P_p,\ p = 1, \dots, 6$, polynomial approximations. Tables 1.3 – 1.4 and Tables 1.5 – 1.6 show the computational errors in the $L^2(\Omega)$-norm as well as the $H^1(\Omega, \mathcal{T}_h)$-seminorm, and the corresponding experimental orders of convergence for $\alpha = 1/2$ and $\alpha = -3/2$, respectively. These values are visualized in Figures 1.7 – 1.8 in which the achieved experimental order of convergence is easy to observe.

These results lead us to the proposition that for the SIPG method the error behaves like

$$\begin{aligned} \|u - u_h\|_{L^2(\Omega)} &= O(h^{\mu}), \qquad u \in H^{\beta}(\Omega) \\ |u - u_h|_{H^1(\Omega)} &= O(h^{\mu-1}), \qquad u \in H^{\beta}(\Omega), \end{aligned} \tag{1.182}$$

where $\mu = \min(p+1, \beta)$, and for the IIPG and NIPG methods like

$$\begin{aligned} \|u - u_h\|_{L^2(\Omega)} &= O(h^{\bar{\mu}}), \qquad u \in H^{\beta}(\Omega) \\ |u - u_h|_{H^1(\Omega)} &= O(h^{\mu-1}), \qquad u \in H^{\beta}(\Omega), \end{aligned} \tag{1.183}$$

where $\mu = \min(p+1, \beta)$, $\bar{\mu} = \min(\bar{p}, \beta)$, and $\bar{p}$ is given by (1.178). The statements (1.182) – (1.183) are in agreement with numerical experiments (not presented here) carried out by other authors for additional values of α.

Moreover, the experimental order of convergence of the SIPG technique given by (1.182) corresponds to the result in [Fei89], where for any $\beta \in (1, 3/2)$ we get

$$\begin{aligned} \|v - I_h v\|_{L^2(\Omega)} &\le C(\beta) h^{\mu} \|v\|_{H^{\beta}(\Omega)}, \quad v \in H^{\beta}(\Omega), \\ |v - I_h v|_{H^1(\Omega)} &\le C(\beta) h^{\mu-1} \|v\|_{H^{\beta}(\Omega)}, \quad v \in H^{\beta}(\Omega), \end{aligned} \tag{1.184}$$

where $I_h v$ is a piecewise polynomial Lagrange interpolation to v of degree $\le p$, $\mu = \min(p+1, \beta)$ and $C(\beta)$ is a constant independent of h and v. By [BS01, Section 3.3] and the references therein, where the interpolation in the so-called Besov spaces is used, the precise error estimate of order $O(h^{3/2})$ in the $L^2(\Omega)$-norm and $O(h^{1/2})$ in the $H^1(\Omega, \mathcal{T}_h)$-seminorm can be established, which corresponds to our numerical experiments.

Finally, the experimental order of convergence of the NIPG and IIPG techniques given by (1.183) corresponds to (1.184) and results (1.178).

1.8.3 A note on the $L^2(\Omega)$- optimality of NIPG and IIPG

Numerical experiments from Section 1.8.1 lead us to the observation (1.178), which was presented, e.g., in [BBO99], [Riv08] and the references cited therein. The optimal order of convergence for the odd degrees of approximation was theoretically justified in [LN04], where NIPG and IIPG methods were analyzed for uniform partitions of the one-dimensional domain. See also [Che06], where similar results were obtained.

On the other hand, several examples of 1D special non-uniform (but quasi-uniform) meshes were presented in [GR09], where the NIPG method gives the error in the $L^2(\Omega)$-norm of order $O(h^p)$ even for odd p. A suboptimal EOC can also be obtained for the IIPG method using these meshes, see [Riv08], Section 1.5, Table 1.2.

In [DH10], it was shown that the use of odd degrees of polynomial approximation of IIPG method leads to the optimal order of convergence in the $L^2(\Omega)$-norm on 1D quasi-uniform grids if and only if the penalty parameter (of order $O(h^{-1})$) is chosen in a special way. These results lead us to the hypothesis that the observation (1.178) is not valid in general.

		SIPG		NIPG		IIPG	
p	$h/\sqrt{2}$	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC
1	1/8	2.1789E-03	–	8.1338E-04	–	1.8698E-03	–
1	1/16	5.7581E-04	1.92	2.1069E-04	1.95	4.8403E-04	1.95
1	1/32	1.4740E-04	1.97	5.3806E-05	1.97	1.2267E-04	1.98
1	1/64	3.7248E-05	1.98	1.3609E-05	1.98	3.0848E-05	1.99
	GEOC		1.96		1.97		1.97
2	1/8	5.7796E-05	–	1.0098E-04	–	5.9762E-05	–
2	1/16	7.2545E-06	2.99	2.6758E-05	1.92	1.1004E-05	2.44
2	1/32	9.1150E-07	2.99	6.9525E-06	1.94	2.4341E-06	2.18
2	1/64	1.1434E-07	2.99	1.7734E-06	1.97	5.8760E-07	2.05
	GEOC		2.99		1.94		2.22
3	1/8	2.6233E-06	–	4.0597E-06	–	2.7474E-06	–
3	1/16	1.9366E-07	3.76	3.3583E-07	3.60	2.1985E-07	3.64
3	1/32	1.4898E-08	3.70	2.8012E-08	3.58	1.7889E-08	3.62
3	1/64	1.1930E-09	3.64	2.3717E-09	3.56	1.4838E-09	3.59
	GEOC		3.70		3.58		3.62
4	1/8	2.6498E-07	–	4.1937E-07	–	3.0663E-07	–
4	1/16	2.1097E-08	3.65	3.4292E-08	3.61	2.4522E-08	3.64
4	1/32	1.7819E-09	3.57	2.8705E-09	3.58	2.0460E-09	3.58
4	1/64	1.5429E-10	3.53	2.4482E-10	3.55	1.7516E-10	3.55
	GEOC		3.58		3.58		3.59
5	1/8	5.8491E-08	–	9.3494E-08	–	7.2011E-08	–
5	1/16	4.9611E-09	3.56	8.1022E-09	3.53	6.1832E-09	3.54
5	1/32	4.2999E-10	3.53	7.0989E-10	3.51	5.3944E-10	3.52
5	1/64	3.7656E-11	3.51	6.2465E-11	3.51	4.7387E-11	3.51
	GEOC		3.53		3.52		3.52
6	1/8	1.9318E-08	–	2.9767E-08	–	2.6495E-08	–
6	1/16	1.6677E-09	3.53	2.6000E-09	3.52	2.3079E-09	3.52
6	1/32	1.4570E-10	3.52	2.2856E-10	3.51	2.0259E-10	3.51
6	1/64	1.2809E-11	3.51	2.0149E-11	3.50	1.7847E-11	3.50
	GEOC		3.52		3.51		3.51

Table 1.3: Computational error and EOC in the $L^2(\Omega)$-norm for the solution of problem (1.179) with $\alpha = 1/2$.

		SIPG		NIPG		IIPG	
p	$h/\sqrt{2}$	$\lvert e_h\rvert_{H^1(\Omega,\mathcal{T}_h)}$	EOC	$\lvert e_h\rvert_{H^1(\Omega,\mathcal{T}_h)}$	EOC	$\lvert e_h\rvert_{H^1(\Omega,\mathcal{T}_h)}$	EOC
1	1/8	5.0805E-02	–	4.2283E-02	–	5.0531E-02	–
1	1/16	2.5722E-02	0.98	2.1564E-02	0.97	2.5653E-02	0.98
1	1/32	1.2919E-02	0.99	1.0877E-02	0.99	1.2902E-02	0.99
1	1/64	6.4715E-03	1.00	5.4607E-03	0.99	6.4674E-03	1.00
	GEOC		0.99		0.98		0.99
2	1/8	4.0313E-03	–	3.2281E-03	–	3.5738E-03	–
2	1/16	1.0230E-03	1.98	8.0878E-04	2.00	9.0960E-04	1.97
2	1/32	2.5750E-04	1.99	2.0223E-04	2.00	2.2938E-04	1.99
2	1/64	6.4585E-05	2.00	5.0547E-05	2.00	5.7592E-05	1.99
	GEOC		1.99		2.00		1.99
3	1/8	2.2371E-04	–	2.2267E-04	–	2.0664E-04	–
3	1/16	3.2897E-05	2.77	3.2455E-05	2.78	3.0237E-05	2.77
3	1/32	5.0341E-06	2.71	4.9281E-06	2.72	4.5992E-06	2.72
3	1/64	8.0276E-07	2.65	7.8150E-07	2.66	7.2933E-07	2.66
	GEOC		2.71		2.72		2.72
4	1/8	2.8019E-05	–	2.6863E-05	–	2.3759E-05	–
4	1/16	4.5630E-06	2.62	4.3388E-06	2.63	3.8426E-06	2.63
4	1/32	7.7950E-07	2.55	7.3892E-07	2.55	6.5504E-07	2.55
4	1/64	1.3572E-07	2.52	1.2850E-07	2.52	1.1398E-07	2.52
	GEOC		2.56		2.57		2.57
5	1/8	8.0765E-06	–	8.3686E-06	–	7.0904E-06	–
5	1/16	1.3891E-06	2.54	1.4415E-06	2.54	1.2239E-06	2.53
5	1/32	2.4249E-07	2.52	2.5191E-07	2.52	2.1413E-07	2.51
5	1/64	4.2611E-08	2.51	4.4293E-08	2.51	3.7673E-08	2.51
	GEOC		2.52		2.52		2.52
6	1/8	3.2423E-06	–	3.4916E-06	–	2.9734E-06	–
6	1/16	5.6456E-07	2.52	6.0843E-07	2.52	5.1885E-07	2.52
6	1/32	9.9090E-08	2.51	1.0684E-07	2.51	9.1177E-08	2.51
6	1/64	1.7456E-08	2.50	1.8826E-08	2.50	1.6072E-08	2.50
	GEOC		2.51		2.51		2.51

Table 1.4: Computational error and EOC in the $H^1(\Omega, \mathcal{T}_h)$-seminorm for the solution of problem (1.179) with $\alpha = 1/2$.

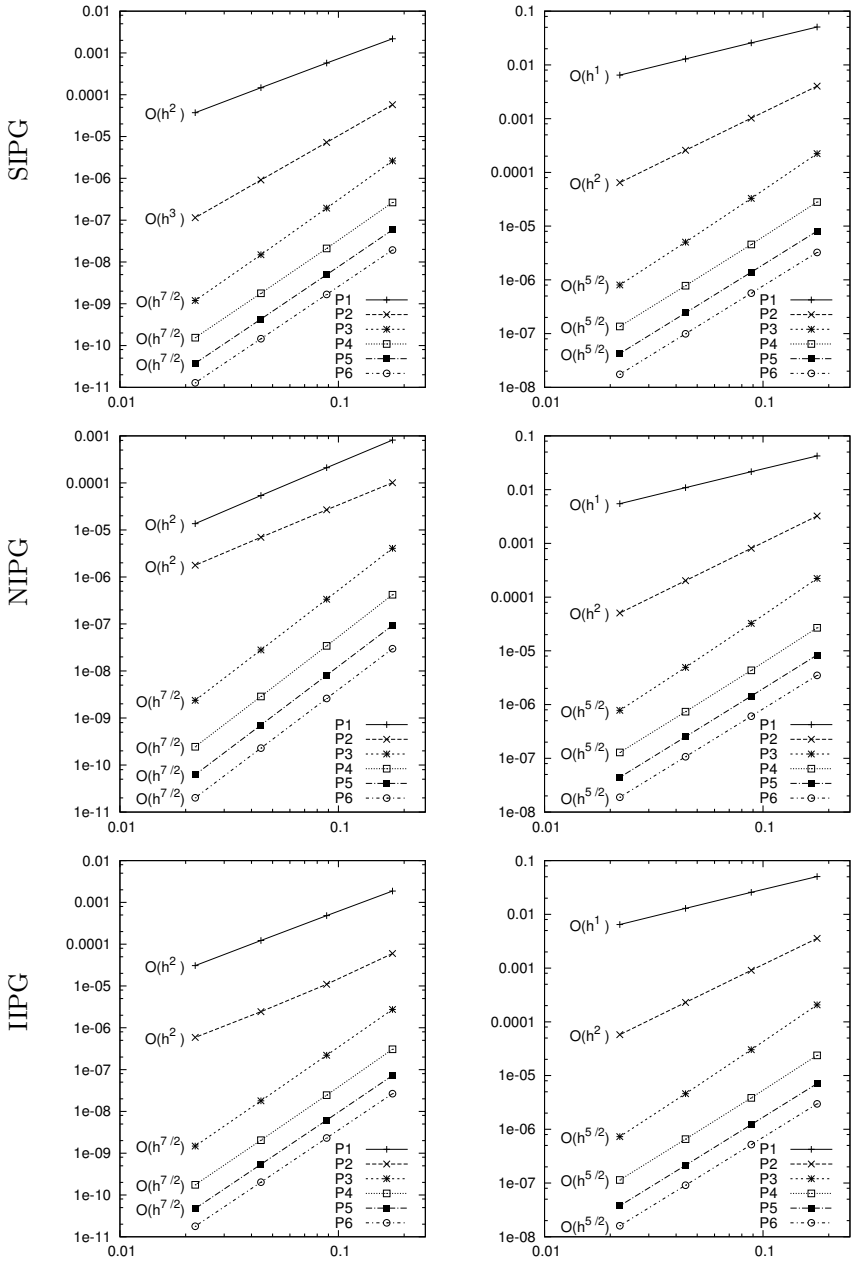

Figure 1.7: Computational error and EOC in the $L^2(\Omega)$-norm (left) and the $H^1(\Omega, \mathcal{T}_h)$-seminorm (right) for the the solution of problem (1.179) with $\alpha = 1/2$.

		SIPG		NIPG		IIPG	
p	$h/\sqrt{2}$	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC
1	1/8	9.2233E-03	–	1.4850E-02	–	7.9896E-03	–
1	1/16	3.2898E-03	1.49	5.3458E-03	1.47	2.8145E-03	1.51
1	1/32	1.1569E-03	1.51	1.8699E-03	1.52	9.8230E-04	1.52
1	1/64	4.0594E-04	1.51	6.5039E-04	1.52	3.4327E-04	1.52
	GEOC		1.50		1.51		1.51
2	1/8	2.3410E-03	–	4.6812E-03	–	1.7779E-03	–
2	1/16	8.1979E-04	1.51	1.6138E-03	1.54	6.0110E-04	1.56
2	1/32	2.8885E-04	1.50	5.6696E-04	1.51	2.0820E-04	1.53
2	1/64	1.0199E-04	1.50	2.0059E-04	1.50	7.2989E-05	1.51
	GEOC		1.51		1.51		1.53
3	1/8	9.7871E-04	–	3.1394E-03	–	1.0279E-03	–
3	1/16	3.4597E-04	1.50	1.1136E-03	1.50	3.6119E-04	1.51
3	1/32	1.2235E-04	1.50	3.9426E-04	1.50	1.2736E-04	1.50
3	1/64	4.3269E-05	1.50	1.3948E-04	1.50	4.4971E-05	1.50
	GEOC		1.50		1.50		1.50
4	1/8	6.4002E-04	–	1.6788E-03	–	7.8547E-04	–
4	1/16	2.2608E-04	1.50	5.9262E-04	1.50	2.7649E-04	1.51
4	1/32	7.9902E-05	1.50	2.0934E-04	1.50	9.7529E-05	1.50
4	1/64	2.8245E-05	1.50	7.3980E-05	1.50	3.4442E-05	1.50
	GEOC		1.50		1.50		1.50
5	1/8	3.8770E-04	–	1.1048E-03	–	6.0190E-04	–
5	1/16	1.3695E-04	1.50	3.9046E-04	1.50	2.1214E-04	1.50
5	1/32	4.8400E-05	1.50	1.3801E-04	1.50	7.4886E-05	1.50
5	1/64	1.7109E-05	1.50	4.8784E-05	1.50	2.6455E-05	1.50
	GEOC		1.50		1.50		1.50
6	1/8	2.7881E-04	–	7.5211E-04	–	5.2298E-04	–
6	1/16	9.8519E-05	1.50	2.6580E-04	1.50	1.8457E-04	1.50
6	1/32	3.4822E-05	1.50	9.3954E-05	1.50	6.5195E-05	1.50
6	1/64	1.2310E-05	1.50	3.3215E-05	1.50	2.3039E-05	1.50
	GEOC		1.50		1.50		1.50

Table 1.5: Computational error and EOC in the $L^2(\Omega)$-norm for the solution of problem (1.179) with $\alpha = -3/2$.

		SIPG		NIPG		IIPG	
p	$h/\sqrt{2}$	$\lvert e_h\rvert_{H^1(\Omega,\mathcal{T}_h)}$	EOC	$\lvert e_h\rvert_{H^1(\Omega,\mathcal{T}_h)}$	EOC	$\lvert e_h\rvert_{H^1(\Omega,\mathcal{T}_h)}$	EOC
1	1/8	4.0604E-01	–	3.9606E-01	–	4.0035E-01	–
1	1/16	2.8999E-01	0.49	2.8508E-01	0.47	2.8631E-01	0.48
1	1/32	2.0555E-01	0.50	2.0312E-01	0.49	2.0309E-01	0.50
1	1/64	1.4539E-01	0.50	1.4413E-01	0.50	1.4370E-01	0.50
	GEOC		0.49		0.49		0.49
2	1/8	1.9294E-01	–	2.3736E-01	–	1.8460E-01	–
2	1/16	1.3627E-01	0.50	1.6750E-01	0.50	1.3052E-01	0.50
2	1/32	9.6419E-02	0.50	1.1842E-01	0.50	9.2389E-02	0.50
2	1/64	6.8224E-02	0.50	8.3741E-02	0.50	6.5385E-02	0.50
	GEOC		0.50		0.50		0.50
3	1/8	1.4304E-01	–	2.3656E-01	–	1.5217E-01	–
3	1/16	1.0145E-01	0.50	1.6731E-01	0.50	1.0794E-01	0.50
3	1/32	7.1853E-02	0.50	1.1833E-01	0.50	7.6459E-02	0.50
3	1/64	5.0852E-02	0.50	8.3679E-02	0.50	5.4113E-02	0.50
	GEOC		0.50		0.50		0.50
4	1/8	9.4937E-02	–	1.7438E-01	–	1.0791E-01	–
4	1/16	6.7297E-02	0.50	1.2334E-01	0.50	7.6474E-02	0.50
4	1/32	4.7649E-02	0.50	8.7229E-02	0.50	5.4139E-02	0.50
4	1/64	3.3715E-02	0.50	6.1686E-02	0.50	3.8306E-02	0.50
	GEOC		0.50		0.50		0.50
5	1/8	7.8490E-02	–	1.4046E-01	–	9.6583E-02	–
5	1/16	5.5605E-02	0.50	9.9348E-02	0.50	6.8396E-02	0.50
5	1/32	3.9357E-02	0.50	7.0261E-02	0.50	4.8400E-02	0.50
5	1/64	2.7843E-02	0.50	4.9686E-02	0.50	3.4238E-02	0.50
	GEOC		0.50		0.50		0.50
6	1/8	6.4288E-02	–	1.2563E-01	–	9.3368E-02	–
6	1/16	4.5518E-02	0.50	8.8855E-02	0.50	6.6077E-02	0.50
6	1/32	3.2208E-02	0.50	6.2836E-02	0.50	4.6744E-02	0.50
6	1/64	2.2782E-02	0.50	4.4434E-02	0.50	3.3060E-02	0.50
	GEOC		0.50		0.50		0.50

Table 1.6: Computational error and EOC in the $H^1(\Omega, \mathcal{T}_h)$-seminorm for the solution of problem (1.179) with $\alpha = -3/2$.

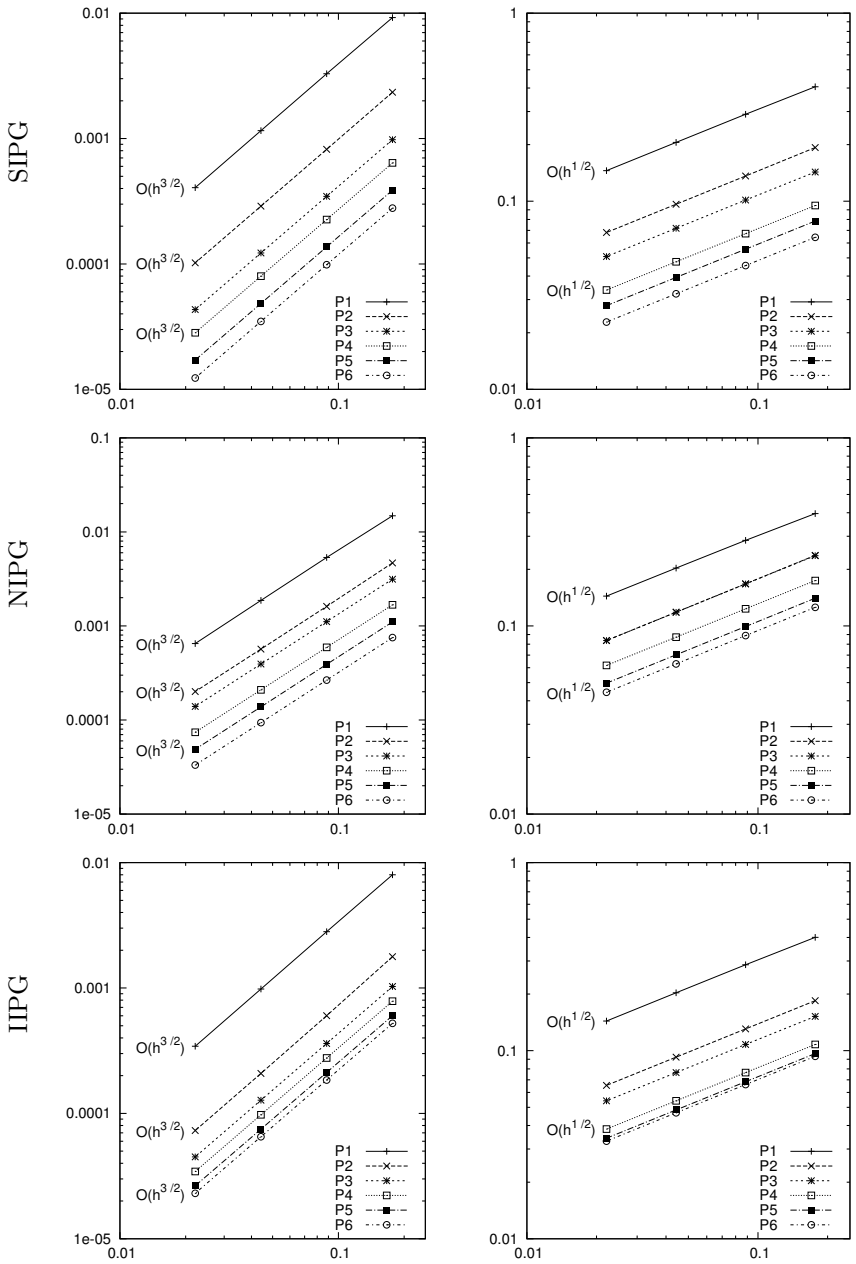

Figure 1.8: Computational error and EOC in the $L^2(\Omega)$-norm (left) and the $H^1(\Omega, \mathcal{T}_h)$-seminorm (right) for the the solution of problem (1.179) with $\alpha = -3/2$.

However, extending theoretical results either to NIPG method or to higher dimensions is problematic. Some attempt was presented in [dDBHM12], where the optimal order of convergence in the $L^2(\Omega)$-norm on equilateral triangular grids was proved for the IIPG method with reduced interior and boundary penalties.

Chapter 2

DGM for convection-diffusion problems

The next chapters 2–4 will be devoted to the DGM for the solution of nonstationary, in general nonlinear, convection-diffusion initial-boundary value problems. Some equations treated here can serve as a simplified model of the Navier–Stokes system describing compressible flow, but the subject of convection-diffusion problems is important for a number of areas in science and technology, as is mentioned in the introduction.

In this chapter we shall be concerned with the analysis of the DGM applied to the space discretization of nonstationary linear and nonlinear convection-diffusion equations. The time variable will be left as continuous. This means that we deal with the so-called *space semidiscretization*, also called the *method of lines*. The full space-time discretization will be the subject of Chapters **??** and 4.

The diffusion terms are discretized by interior penalty Galerkin techniques (SIPG, NIPG and IIPG) introduced in Chapter 1. A special attention is paid to the discretization of convective terms, where the concept of the numerical flux (well-known from the finite volume method) is used. We derive error estimates for a nonlinear equation discretized by all three mentioned techniques. These estimates are suboptimal in the $L^\infty(L^2)$-norm and they are not uniform with respect to the diffusion coefficient. However, for the symmetric SIPG variant, the optimal error estimate in the $L^\infty(L^2)$-norm is derived. Finally, for a linear convection-diffusion equation, we derive error estimates uniform with respect to the diffusion coefficient.

2.1 Scalar nonlinear nonstationary convection-diffusion equation

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded polygonal (if $d = 2$) or polyhedral (if $d = 3$) domain with Lipschitz boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, $\partial\Omega_D \cap \partial\Omega_N = \emptyset$, and $T > 0$. We shall assume that the $(d-1)$-dimensional measure of $\partial\Omega_D$ is positive. Let us denote $Q_T = \Omega \times (0, T)$.

We are concerned with the following nonstationary nonlinear convection-diffusion problem with initial and mixed Dirichlet–Neumann boundary conditions: Find $u : \overline{Q}_T \to \mathbb{R}$ such that

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} = \varepsilon \Delta u + g \quad \text{in } Q_T, \tag{2.1a}$$

$$u\big|_{\partial\Omega_D \times (0,T)} = u_D, \tag{2.1b}$$

$$\varepsilon \frac{\partial u}{\partial n}\Big|_{\partial\Omega_N \times (0,T)} = g_N, \tag{2.1c}$$

$$u(x, 0) = u^0(x), \quad x \in \Omega. \tag{2.1d}$$

We assume that the data satisfy the following conditions:

$$\boldsymbol{f} = (f_1, \dots, f_d),\ f_s \in C^1(\mathbb{R}),\ f_s' \text{ are bounded},\ f_s(0) = 0,\ s = 1, \dots, d, \tag{2.2a}$$

$$\varepsilon > 0, \tag{2.2b}$$

$$g \in C([0, T]; L^2(\Omega)), \tag{2.2c}$$

$$u_D = \text{trace of some } u^* \in C([0, T]; H^1(\Omega)) \cap L^\infty(Q_T) \text{ on } \partial\Omega_D \times (0, T), \tag{2.2d}$$

$$g_N \in C([0, T]; L^2(\partial\Omega_N)), \tag{2.2e}$$

$$u^0 \in L^2(\Omega). \tag{2.2f}$$

The constant ε is a diffusion coefficient, f_s, $s = 1, \dots, d$, are nonlinear convective fluxes and g is a source term. It can be seen

that the assumption that $f_s(0) = 0$ is not limiting. If u satisfies (2.1a), then it also satisfies the equation

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{d} \frac{\partial (f_s(u) - f_s(0))}{\partial x_s} = \varepsilon \Delta u + g,$$

and the new convective fluxes $\tilde{f}_s := f_s(u) - f_s(0),\ s = 1, \dots, d$, satisfy (2.2a). Let us note that in Section 4.2 we shall be concerned with more complicated situation, where both convection and diffusion terms are nonlinear.

It is suitable to introduce the concept of a *weak solution*. To this end, we define the space

$$H^1_{0D}(\Omega) = \{v \in H^1(\Omega);\ v|_{\partial\Omega_D} = 0\},$$

and the following forms:

$$\begin{aligned}
(u, v) &= (u, v)_{L^2(\Omega)} = \int_\Omega uv \,\mathrm{d}x, \quad u,\, v \in L^2(\Omega),\\
a(u, v) &= \varepsilon \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x, \quad u,\, v \in H^1(\Omega),\\
b(u, v) &= \int_\Omega \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s}\, v \,\mathrm{d}x, \quad u \in H^1(\Omega) \cap L^\infty(\Omega),\ v \in L^2(\Omega),\\
(u, v)_N &= \int_{\partial\Omega_N} u\, v \,\mathrm{d}S, \quad u,\, v \in L^2(\partial\Omega_N).
\end{aligned}$$

Definition 2.1. *A function u is called the* weak solution *of problem* (2.1)*, if it satisfies the conditions*

$$u - u^* \in L^2(0, T; H^1_{0D}(\Omega)), \quad u \in L^\infty(Q_T), \tag{2.3a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(u(t), v) + b(u(t), v) + a(u(t), v) = (g(t), v) + (g_N(t), v)_N \ \forall\, v \in H^1_{0D}(\Omega) \tag{2.3b}$$

(in the sense of distributions in $(0, T)$),

$$u(0) = u_0 \quad \text{in } \Omega. \tag{2.3c}$$

Let us recall that by $u(t)$ we denote the function in Ω such that $u(t)\,(x) = u(x, t)$, $x \in \Omega$.

With the aid of techniques from [Rek82], [Lio96] or [Rou05], it is possible to prove that for a function u satisfying (2.3a) – (2.3b) we have $u \in C([0, T]; L^2(\Omega))$, which means that condition (2.3c) makes sense, and that there exists a unique solution of problem (2.3). Moreover, it satisfies the condition $\partial u/\partial t \in L^2(Q_T)$. Then (2.3b) can be rewritten as

$$\left(\frac{\partial u(t)}{\partial t}, v\right) + b(u(t), v) + a(u(t), v) = (g(t), v) + (g_N(t), v)_N \tag{2.4}$$

$$\forall\, v \in H^1_{0D}(\Omega) \text{ and almost every } t \in (0, T).$$

We say that u satisfying (2.3) is a *strong solution*, if

$$u \in L^2(0, T; H^2(\Omega)), \quad \frac{\partial u}{\partial t} \in L^2(0, T; H^1(\Omega)). \tag{2.5}$$

It is possible to show that the strong solution u satisfies equation (2.1) pointwise (almost everywhere) and $u \in C([0, T], H^1(\Omega))$.

2.2 Discretization

In this section we introduce a DG space semidiscretization of problem (2.1). We use the notation and auxiliary results from Sections 1.3–1.5.

By $\mathcal{T}_h$ ($h > 0$) we denote a triangulation of the domain Ω introduced in Section 1.3.1. We start from the strong solution u satisfying (2.5), multiply equation (2.1a) by an arbitrary $v \in H^2(\Omega, \mathcal{T}_h)$, integrate over each $K \in \mathcal{T}_h$, and apply Green's theorem. We obtain the identity

$$\begin{aligned}
\int_K \frac{\partial u(t)}{\partial t}\, v \,\mathrm{d}x + \int_{\partial K} \sum_{s=1}^{d} f_s(u(t)) n_s v \,\mathrm{d}S - \int_K \sum_{s=1}^{d} f_s(u(t)) \frac{\partial v}{\partial x_s} \,\mathrm{d}x \\
+ \varepsilon \int_K \nabla u(t) \cdot \nabla v \,\mathrm{d}x - \varepsilon \int_{\partial K} (\nabla u(t) \cdot \boldsymbol{n})\, v \,\mathrm{d}S = \int_K g(t)\, v \,\mathrm{d}x.
\end{aligned} \tag{2.6}$$

Here $\boldsymbol{n} = (n_1, \dots, n_d)$ denotes the outer unit normal to ∂K. It is possible to write

$$\sum_{s=1}^{d} f_s(u) n_s = \boldsymbol{f}(u) \cdot \boldsymbol{n}, \quad \sum_{s=1}^{d} f_s(u) \frac{\partial v}{\partial x_s} = \boldsymbol{f}(u) \cdot \nabla v. \tag{2.7}$$

Summing (2.6) over all $K \in \mathcal{T}_h$, using the technique introduced in Section 1.4 for the discretization of the diffusion term, we obtain the identity

$$\left(\frac{\partial u(t)}{\partial t}, v \right) + A_h(u(t), v) + \tilde{b}_h(u(t), v) = \ell_h(v)(t), \tag{2.8}$$

where

$$A_h(w, v) = \varepsilon a_h(w, v) + \varepsilon J_h^\sigma(w, v), \tag{2.9}$$

$$a_h(u, v) = \sum_{K \in \mathcal{T}_h} \int_K \nabla u \cdot \nabla v \, \mathrm{d}x - \sum_{\Gamma \in \mathcal{F}_h^{ID}} \int_\Gamma \left(\langle \nabla u \rangle \cdot \boldsymbol{n}[v] + \Theta \langle \nabla v \rangle \cdot \boldsymbol{n}[u] \right) \mathrm{d}S, \tag{2.10}$$

$$J_h^\sigma(u, v) = \sum_{\Gamma \in \mathcal{F}_h^{ID}} \int_\Gamma \sigma [u] \, [v] \, \mathrm{d}S, \tag{2.11}$$

$$\tilde{b}_h(u, v) = \sum_{K \in \mathcal{T}_h} \left\{ \int_{\partial K} \sum_{s=1}^{d} f_s(u(t)) n_s v \, \mathrm{d}S - \int_K \sum_{s=1}^{d} f_s(u(t)) \frac{\partial v}{\partial x_s} \, \mathrm{d}x \right\}, \tag{2.12}$$

$$\ell_h(v)\,(t) = (g(t), v) + (g_N(t), v)_N + \varepsilon \sum_{\Gamma \in \mathcal{F}_h^D} \int_\Gamma \Big(\sigma v - \Theta (\nabla v \cdot \boldsymbol{n}) \Big) u_D(t) \, \mathrm{d}S. \tag{2.13}$$

(The symbols $\langle \cdot \rangle, [\cdot]$ are defined in (1.32) and (1.33).) We call a_h and J_h the diffusion form and the interior and boundary penalty form, respectively. Similarly as in (1.104), the *penalty weight* σ is given by

$$\sigma|_\Gamma = \sigma_\Gamma = \frac{C_W}{h_\Gamma}, \quad \Gamma \in \mathcal{F}_h^{ID}, \tag{2.14}$$

where h_Γ characterizes the "size" of $\Gamma \in \mathcal{F}_h$ defined in Section 1.6 and $C_W > 0$ is a suitable constant. The symbol $\tilde{b}_h$ corresponds to the convection terms. It will be further discretized.

Similarly, as in Section 1.4, for $\Theta = -1$, $\Theta = 0$ and $\Theta = 1$ the form a_h (together with the form J_h^σ) represents the nonsymmetric variant (NIPG), incomplete variant (IIPG) and symmetric variant (SIPG), respectively, of the diffusion form.

Remark 2.2. *Let us note that in contrast to Chapter 1, the form A_h contains the diffusion coefficient ε, compare* (1.45a) – (1.45c) *with* (2.9)*. Therefore, the estimates from Chapter 1, which will be used here, have to be equipped with the multiplication factor $\varepsilon > 0$. We do not emphasize it in the following.*

Now we shall pay a special attention to the approximation of the convective terms represented by the form $\tilde{b}_h$. The integrals $\int_{\partial K} \sum_{s=1}^{d} f_s(u(t)) n_s v \, \mathrm{d}S$ can be expressed in terms of the expressions $\int_\Gamma \sum_{s=1}^{d} f_s(u(t)) n_s v \, \mathrm{d}S$, which will be approximated with the aid of the so-called *numerical flux* $H(u, w, \boldsymbol{n})$:

$$\int_\Gamma \sum_{s=1}^{d} f_s(u(t)) n_s v \, \mathrm{d}S \approx \int_\Gamma H(u_\Gamma^{(L)}, u_\Gamma^{(R)}, \boldsymbol{n}) \, v_\Gamma^{(L)} \, \mathrm{d}S, \qquad \Gamma \in \mathcal{F}_h. \tag{2.15}$$

Here $H : \mathbb{R} \times \mathbb{R} \times \mathrm{B}_1 \to \mathbb{R}$ is a suitably defined function and $\mathrm{B}_1 = \{\boldsymbol{n} \in \mathbb{R}^d; |\boldsymbol{n}| = 1\}$ is the unit sphere in $\mathbb{R}^d$. The simplest are the *central numerical fluxes* given by

$$H(v_1, v_2, \boldsymbol{n}) = \sum_{s=1}^{d} f_s \left(\frac{v_1 + v_2}{2} \right) n_s, \qquad H(v_1, v_2, \boldsymbol{n}) = \sum_{s=1}^{d} \frac{f_s(v_1) + f_s(v_2)}{2} n_s.$$

However, in the most of applications it is suitable to use *upwinding* [1] numerical fluxes as, for example,

$$H(u_1, u_2, \boldsymbol{n}) = \begin{cases} \sum_{s=1}^{d} f_s(u_1) n_s, & \text{if } P > 0 \\ \sum_{s=1}^{d} f_s(u_2) n_s, & \text{if } P \le 0 \end{cases}, \quad \text{where } P = \sum_{s=1}^{d} f_s' \left(\frac{u_1 + u_2}{2} \right) n_s, \tag{2.16}$$

[1] The concept of upwinding is based on the idea that the information on properties of a quantity u is propagated in the flow direction. Therefore, discretization of convective terms is carried out with the aid of data located in the upwind direction from the points in consideration.

or the *Lax–Friedrichs numerical flux*

$$H(v_1, v_2, \boldsymbol{n}) = \sum_{s=1}^{d} \frac{f_s(v_1) + f_s(v_2)}{2} n_s - \lambda |v_1 - v_2|,$$

where $\lambda > 0$ has to be chosen in an appropriate way. For more examples and theoretical background of numerical fluxes we refer to [FFS03].

If $\Gamma \in \mathcal{F}_h^B$, then it is necessary to specify the meaning of $u_\Gamma^{(R)}$ in (2.15). It is possible to use the *extrapolation* from the interior of the computational domain

$$u_\Gamma^{(R)} := u_\Gamma^{(L)}, \qquad \Gamma \in \mathcal{F}_h^B. \tag{2.17}$$

In the theoretical analysis, we shall assume that the numerical flux satisfies the following properties:

1. *continuity*: $H(u, v, \boldsymbol{n})$ is *Lipschitz-continuous* with respect to u, v: there exists a constant $L_H > 0$ such that

$$\begin{aligned} &|H(u, v, \boldsymbol{n}) - H(u^*, v^*, \boldsymbol{n})| \le L_H(|u - u^*| + |v - v^*|), \\ &u,\ v,\ u^*,\ v^* \in \mathbb{R},\ \boldsymbol{n} \in \mathrm{B}_1. \end{aligned} \tag{2.18}$$

2. *consistency*:

$$H(u, u, \boldsymbol{n}) = \sum_{s=1}^{d} f_s(u) n_s, \quad u \in \mathbb{R},\ \boldsymbol{n} = (n_1, \dots, n_d) \in \mathrm{B}_1. \tag{2.19}$$

3. *conservativity*:

$$H(u, v, \boldsymbol{n}) = -H(v, u, -\boldsymbol{n}), \quad u,\ v \in \mathbb{R},\ \boldsymbol{n} \in \mathrm{B}_1. \tag{2.20}$$

By virtue of (2.18) and (2.19), the functions $f_s, s = 1, \dots, d$, are Lipschitz-continuous with constant $L_f = 2L_H$. From (2.2a) and (2.19) we see that

$$H(0, 0, \boldsymbol{n}) = 0 \quad \forall\, \boldsymbol{n} \in \mathrm{B}_1. \tag{2.21}$$

Using the conservativity (2.20) of H and notation (1.32) – (1.33), we find that

$$\begin{aligned} &\sum_{K \in \mathcal{T}_h} \sum_{\Gamma \subset \partial K, \Gamma \in \mathcal{F}_h} \int_\Gamma H(u_\Gamma^{(L)}, u_\Gamma^{(R)}, \boldsymbol{n})\, v_\Gamma^{(L)}\, \mathrm{d}S \\ &= \sum_{\Gamma \in \mathcal{F}_h^I} \int_\Gamma H(u_\Gamma^{(L)}, u_\Gamma^{(R)}, \boldsymbol{n}) \left(v_\Gamma^{(L)} - v_\Gamma^{(R)}\right) \mathrm{d}S + \sum_{\Gamma \in \mathcal{F}_h^B} \int_\Gamma H(u_\Gamma^{(L)}, u_\Gamma^{(R)}, \boldsymbol{n})\, v_\Gamma^{(L)}\, \mathrm{d}S \\ &= \sum_{\Gamma \in \mathcal{F}_h} \int_\Gamma H(u_\Gamma^{(L)}, u_\Gamma^{(R)}, \boldsymbol{n})\, [v]\, \mathrm{d}S \end{aligned} \tag{2.22}$$

Let us recall that in integrals $\int_\Gamma$ the symbol $\boldsymbol{n}$ denotes the normal $\boldsymbol{n}_\Gamma$.

Then, by virtue of (2.15) and (2.22), we define the *convection form* $b_h(u, v)$ approximating $\tilde{b}_h(u, v)$:

$$\begin{aligned} b_h(u, v) = \sum_{\Gamma \in \mathcal{F}_h} \int_\Gamma H(u_\Gamma^{(L)}, u_\Gamma^{(R)}, \boldsymbol{n})\, [v]\, \mathrm{d}S - \sum_{K \in \mathcal{T}_h} \int_K \boldsymbol{f}(u) \cdot \nabla v\, \mathrm{d}x, \\ u,\ v \in H^1(\Omega, \mathcal{T}_h), \quad u \in L^\infty(\Omega). \end{aligned} \tag{2.23}$$

By the definitions (2.12), (2.23) and the consistency (2.19), we have

$$b_h(u, v) = \tilde{b}_h(u, v) \quad \forall u \in H^2(\Omega)\ \forall v \in H^2(\Omega, \mathcal{T}_h). \tag{2.24}$$

Let S_{hp} be the space of discontinuous piecewise polynomial functions (1.34). Since $S_{hp} \subset H^2(\Omega, \mathcal{T}_h) \cap L^\infty(\Omega)$, the forms (2.10), (2.11), (2.13) and (2.23) make sense for $u := u_h,\ v := v_h \in S_{hp}$. Then, we introduce the space DG-discretization of (2.1).

Definition 2.3. *We define the* semidiscrete approximate solution *as a function* $u_h : Q_T \to \mathbb{R}$ *satisfying the conditions*

$$u_h \in C^1([0,T]; S_{hp}), \tag{2.25a}$$

$$\left(\frac{\partial u_h(t)}{\partial t}, v_h\right) + A_h(u_h(t), v_h) + b_h(u_h(t), v_h) = \ell_h(v_h)\,(t) \tag{2.25b}$$

$$\forall\, v_h \in S_{hp},\ \forall\, t \in [0,T],$$

$$(u_h(0), v_h) = (u^0, v_h) \quad \forall\, v_h \in S_{hp}. \tag{2.25c}$$

We see that the initial condition (2.25c) can be written as $u_h(0) = \Pi_{hp} u^0$, where Π_{hp} is the operator of the $L^2(\Omega)$-projection on the space S_{hp} (cf. (1.90)).

The discrete problem (2.25) is equivalent to an initial value problem for a system of ordinary differential equations (ODEs). Namely, let $\{\varphi_i,\ i = 1, \dots, N_h\}$ be a basis of the space S_{hp}, where $N_h = \dim S_{hp}$. The approximate solution u_h is sought in the form

$$u_h(x,t) = \sum_{j=1}^{N_h} u^j(t)\varphi_j(x), \tag{2.26}$$

where $u^j(t) : [0,T] \to \mathbb{R},\ j = 1, \dots, N_h$, are unknown functions. For simplicity, we put

$$B_h(u_h, v_h) = \ell_h(v_h) - A_h(u_h, v_h) - b_h(u_h, v_h), \quad u_h, v_h \in S_{hp}.$$

Now, substituting (2.26) into (2.25b) and putting $v_h := \varphi_i$, we get

$$\sum_{j=1}^{N_h} \frac{\mathrm{d}u^j(t)}{\mathrm{d}t}\,(\varphi_j, \varphi_i) = B_h\left(\sum_{j=1}^{N_h} u^j(t)\varphi_j, \varphi_i\right), \qquad i = 1, \dots, N_h, \tag{2.27}$$

which is the system of the ODEs for the unknown functions $u^j,\ j = 1, \dots, N_h$. 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 (2.27), we obtain a fully discrete problem. In Chapter **??** we shall pay attention to some full space-time discretization techniques. In what follows we shall be concerned with the analysis of the semidiscrete problem (2.25).

Taking into account that the exact solution with property (2.5) satisfies $[u]_\Gamma = 0$ for $\Gamma \in \mathcal{F}_h^I$, $u|_{\partial\Omega_D \times (0,T)} = u_D$ and using (2.8) and (2.24), we find that u satisfies the *consistency* identity

$$\left(\frac{\partial u(t)}{\partial t}, v_h\right) + A_h(u(t), v_h) + b_h(u(t), v_h) = \ell_h(v_h)\,(t) \tag{2.28}$$

for all $v_h \in S_{hp}$ and almost all $t \in (0,T)$. This will be used in the error analysis.

Exercise 2.4. *Verify the relation* (2.28).

2.3 Abstract error estimate

In this section we shall analyze the behaviour of the error in method (2.25). We shall use results derived in Sections 1.6 and 1.7 dealing with the properties of the diffusion form a_h and the penalty form J_h^σ. Similarly as in (1.103), we use the DG-norm

$$|||v||| = \left(|v|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(v,v)\right)^{1/2}, \qquad v \in H^1(\Omega, \mathcal{T}_h). \tag{2.29}$$

In the error analysis we shall suppose that the following basic assumptions are satisfied.

Assumptions 2.5. *Let the following assumptions be satisfied:*

- *assumptions* (2.2) *on data of problem* (2.1),
- *properties (2.18) – (2.20) of the numerical flux* H,
- $\{\mathcal{T}_h\}_{h \in (0,\bar{h})}$ *is a system of triangulations of the domain* Ω *satisfying the shape-regularity assumption (1.19) and the equivalence condition (1.20) of* h_Γ *and* h_K *(cf. Lemma 1.5),*
- *the penalization constant* C_W *satisfies the conditions from Corollary 1.41 for SIPG, NIPG and IIPG versions of the diffusion form* a_h.

We shall again apply the multiplicative trace inequality (1.78), the inverse inequality (1.86) and the approximation properties (1.93) – (1.95) and (1.98) – (1.100).

2.3.1 Consistency of the convection form in the case of the Dirichlet boundary condition

We shall be concerned with Lipschitz-continuity and consistency of the form b_h. The consistency analysis is split in two cases. In this section we consider the case when the Dirichlet boundary condition is considered on the whole boundary $\partial\Omega$, i.e., $\partial\Omega_D = \partial\Omega$ and $\partial\Omega_N = \emptyset$. Analyzing the consistency of the form b_h in the case of mixed boundary conditions is more complicated and is presented in Section 2.3.2.

In what follows we shall assume that $s \geq 2,\ p \geq 1$ are integers.

Lemma 2.6. *Let $\Gamma_N = \emptyset$ (then $\mathcal{F}_h = \mathcal{F}_h^{ID}$). Then there exist constants $C_{b1}, \ldots, C_{b4} > 0$ such that*

$$|b_h(u,v) - b_h(\bar{u},v)| \leq C_{b1}|||v||| \left(\|u-\bar{u}\|^2_{L^2(\Omega)} + \sum_{K\in\mathcal{T}_h} h_K \|u-\bar{u}\|^2_{L^2(\partial K)} \right)^{1/2}, \tag{2.30}$$

$$u,\ \bar{u} \in H^1(\Omega,\mathcal{T}_h) \cap L^\infty(\Omega),\ v \in H^1(\Omega,\mathcal{T}_h),\ h \in (0,\bar{h}),$$

$$|b_h(u_h,v_h) - b_h(\bar{u}_h,v_h)| \leq C_{b2}|||v_h||| \, \|u_h - \bar{u}_h\|_{L^2(\Omega)}, \tag{2.31}$$

$$u_h,\ \bar{u}_h,\ v_h \in S_{hp},\ h \in (0,\bar{h}).$$

If $\Pi_{hp}u$ is the S_{hp}-interpolant of $u \in H^s(\Omega)$ defined by (1.90) and we put $\eta = u - \Pi_{hp}u$, then

$$|b_h(u,v_h) - b_h(\Pi_{hp}u,v_h)| \leq C_{b3}R_b(\eta)|||v_h|||, \quad v_h \in S_{hp},\ h \in (0,\bar{h}), \tag{2.32}$$

where

$$R_b(\eta) = \Big(\sum_{K\in\mathcal{T}_h} \big(\|\eta\|^2_{L^2(K)} + h_K^2 |\eta|^2_{H^1(K)} \big) \Big)^{1/2}. \tag{2.33}$$

Moreover, if $\xi = u_h - \Pi_{hp}u$, then under the above assumptions,

$$|b_h(u,v_h) - b_h(u_h,v_h)| \leq C_{b4}|||v_h||| \left(R_b(\eta) + \|\xi\|_{L^2(\Omega)} \right), \quad v_h \in S_{hp},\ h \in (0,\bar{h}). \tag{2.34}$$

Proof. (i) By (2.23), for $u,\ \bar{u},\ v \in H^1(\Omega,\mathcal{T}_h)$,

$$\begin{aligned} b_h(u,v) - b_h(\bar{u},v) = & \underbrace{- \sum_{K\in\mathcal{T}_h} \int_K (\boldsymbol{f}(u) - \boldsymbol{f}(\bar{u})) \cdot \nabla v \, \mathrm{d}x}_{=:\sigma_1} \\ & \underbrace{+ \sum_{\Gamma\in\mathcal{F}_h} \int_\Gamma \Big(H(u_\Gamma^{(L)}, u_\Gamma^{(R)}, \boldsymbol{n}) - H(\bar{u}_\Gamma^{(L)}, \bar{u}_\Gamma^{(R)}, \boldsymbol{n}) \Big) [v] \, \mathrm{d}S}_{=:\sigma_2}. \end{aligned} \tag{2.35}$$

Let us recall that for $\Gamma \in \mathcal{F}_h^B$ we define the functions $u_\Gamma^{(R)}$ and $\bar{u}_\Gamma^{(R)}$ by extrapolation: $u_\Gamma^{(R)} = u_\Gamma^{(L)}$ and $\bar{u}_\Gamma^{(R)} = \bar{u}_\Gamma^{(L)}$.

From the Lipschitz-continuity of the functions $f_s,\ s = 1, \ldots, d$, and the discrete Cauchy inequality we have

$$|\sigma_1| \leq L_f \sum_{K\in\mathcal{T}_h} \int_K \sum_{s=1}^d |u-\bar{u}| \left| \frac{\partial v}{\partial x_s} \right| \mathrm{d}x \leq \sqrt{d} L_f \|u-\bar{u}\|_{L^2(\Omega)} |v|_{H^1(\Omega,\mathcal{T}_h)}. \tag{2.36}$$

Relation (2.35), the Lipschitz-continuity (2.18) of H, the Cauchy inequality, (1.20), (2.11) and (2.14) imply that

$$\begin{aligned} |\sigma_2| &\leq L_H \sum_{\Gamma\in\mathcal{F}_h} \int_\Gamma \left(\left| u_\Gamma^{(L)} - \bar{u}_\Gamma^{(L)} \right| + \left| u_\Gamma^{(R)} - \bar{u}_\Gamma^{(R)} \right| \right) |[v]| \, \mathrm{d}S \\ &\leq L_H \left(\sum_{\Gamma\in\mathcal{F}_h} \int_\Gamma \frac{[v]^2}{h_\Gamma} \, \mathrm{d}S \right)^{\frac{1}{2}} \left(\sum_{\Gamma\in\mathcal{F}_h} \int_\Gamma h_\Gamma \left(\left| u_\Gamma^{(L)} - \bar{u}_\Gamma^{(L)} \right| + \left| u_\Gamma^{(R)} - \bar{u}_\Gamma^{(R)} \right| \right)^2 \mathrm{d}S \right)^{\frac{1}{2}} \\ &\leq L_H \sqrt{\frac{C_G}{C_W}} J_h^\sigma(v,v)^{1/2} \left(\sum_{K\in\mathcal{T}_h} \int_{\partial K} 2h_K |u-\bar{u}|^2 \, \mathrm{d}S \right)^{1/2} \\ &= L_H \sqrt{\frac{2C_G}{C_W}} J_h^\sigma(v,v)^{1/2} \left(\sum_{K\in\mathcal{T}_h} h_K \|u-\bar{u}\|^2_{L^2(\partial K)} \right)^{1/2}. \end{aligned} \tag{2.37}$$

(Let us note that the third inequality in (2.37) is valid only if $\mathcal{F}_h = \mathcal{F}_h^{ID}$.) Taking into account (2.35) – (2.37) and using the discrete Cauchy inequality, we get

$$
\begin{aligned}
&|b_h(u,v) - b_h(\bar{u},v)| \qquad (2.38)\\
\le&\sqrt{d}L_f\|u-\bar{u}\|_{L^2(\Omega)}|v|_{H^1(\Omega,\mathcal{T}_h)} + L_H\sqrt{\frac{2C_G}{C_W}}J_h^\sigma(v,v)^{1/2}\left(\sum_{K\in\mathcal{T}_h} h_K\|u-\bar{u}\|^2_{L^2(\partial K)}\right)^{\frac{1}{2}}\\
\le&\left(dL_f^2\|u-\bar{u}\|^2_{L^2(\Omega)} + L_H^2\frac{2C_G}{C_W}\sum_{K\in\mathcal{T}_h} h_K\|u-\bar{u}\|^2_{L^2(\partial K)}\right)^{\frac{1}{2}}\left(|v|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(v,v)\right)^{\frac{1}{2}}.
\end{aligned}
$$

This immediately implies (2.30) with $C_{b1} = \left(\max(dL_f^2, 2L_H^2C_G/C_W)\right)^{1/2}$.

(ii) Further, let $u_h,\ \bar{u}_h,\ v_h \in S_{hp}$. Using the multiplicative trace inequality (1.78) and the inverse inequality (1.86), for $\varphi \in S_{hp}$ we obtain

$$
\begin{aligned}
\sum_{K\in\mathcal{T}_h} h_K\|\varphi\|^2_{L^2(\partial K)} &\le C_M\sum_{K\in\mathcal{T}_h}\left(\|\varphi\|^2_{L^2(K)} + h_K\|\varphi\|_{L^2(K)}\,|\varphi|_{H^1(K)}\right) \qquad (2.39)\\
&\le C_M\sum_{K\in\mathcal{T}_h}\left(\|\varphi\|^2_{L^2(K)} + C_I\|\varphi\|^2_{L^2(K)}\right) = C_M(1+C_I)\|\varphi\|^2_{L^2(\Omega)}.
\end{aligned}
$$

Now, if we set $\varphi := u_h - \bar{u}_h$ and use (2.30) with $u := u_h, \bar{u} := \bar{u}_h$ and $v := v_h$, we get (2.31) with $C_{b2} = C_{b1}(1+C_M(1+C_I))^{1/2}$.

(iii) In order to prove (2.32), we start from (2.30) with $u \in H^s(\Omega)$, $\bar{u} := \Pi_{hp}\,u$ and $v := v_h \in S_{hp}$. Using the multiplicative trace inequality (1.78) and Young's inequality, we find that

$$
\begin{aligned}
\sum_{K\in\mathcal{T}_h} h_K\|u-\Pi_{hp}u\|^2_{L^2(\partial K)} &= \sum_{K\in\mathcal{T}_h} h_K\|\eta\|^2_{L^2(\partial K)} \qquad (2.40)\\
&\le C_M\sum_{K\in\mathcal{T}_h}\left(\|\eta\|^2_{L^2(K)} + h_K\|\eta\|_{L^2(K)}|\eta|_{H^1(K)}\right)\\
&\le C_M\sum_{K\in\mathcal{T}_h}\left(\|\eta\|^2_{L^2(K)} + \frac{1}{2}\|\eta\|^2_{L^2(K)} + \frac{1}{2}h_K^2|\eta|^2_{H^1(K)}\right) \le \frac{3}{2}C_M R_b(\eta)^2,
\end{aligned}
$$

where $R_b(\eta)$ is defined in (2.33). Consequently,

$$
\|u-\Pi_{hp}\|^2_{L^2(\Omega)} + \sum_{K\in\mathcal{T}_h} h_K\|u-\Pi_{hp}u\|^2_{L^2(\partial K)} \le (1+\frac{3}{2}C_M)R_b(\eta)^2,
$$

which together with (2.30) immediately yield (2.32) with $C_{b3} = C_{b1}(1+3C_M/2)^{1/2}$.

(iv) The triangle inequality gives

$$
|b_h(u,v_h) - b_h(u_h,v_h)| \le |b_h(u,v_h) - b_h(\Pi_{hp}u,v_h)| + |b_h(\Pi_{hp}u,v_h) - b_h(u_h,v_h)|.
$$

From relations (2.32) and (2.31) with $\bar{u}_h = \Pi_{hp}u$ and $\xi = u_h - \Pi_{hp}u$, we get (2.34) with $C_{b4} = \max(C_{b2}, C_{b3})$. □ □

2.3.2 Consistency of the convective form in the case of mixed boundary conditions

Since Lemma 2.6 is valid only if a Dirichlet boundary condition is prescribed on $\partial\Omega$, we shall be concerned here with the consistency of the form b_h in the case of a nonempty Neumann part $\partial\Omega_N$ of the boundary $\partial\Omega$. We shall start from several auxiliary results.

The first lemma shows the existence of a vector-valued function with suitable properties. Its proof is based on the usual definition of a domain with the Lipschitz boundary.

Lemma 2.7. *There exists a vector-valued function $\boldsymbol{\varphi} \in (W^{1,\infty}(\Omega))^d$ such that*

$$
\boldsymbol{\varphi}\cdot\boldsymbol{n} \ge 1 \quad \text{on } \partial\Omega, \qquad (2.41)
$$

where $\boldsymbol{n}$ is the unit outer normal to $\partial\Omega$.

Proof. By [KJk77] or [Neč67], it follows from the Lipschitz-continuity of $\partial\Omega$ that there exist numbers $\alpha,\ \beta > 0$, Cartesian coordinate systems

$$X_r = (x_{r,1}, \dots, x_{r,d-1}, x_{r,d})^{\mathrm{T}} = (x'_r, x_{r,d})^{\mathrm{T}}, \tag{2.42}$$

Lipschitz-continuous functions

$$a_r : \Delta_r = \left\{x'_r = (x_{r,1}, \dots, x_{r,d-1})^{\mathrm{T}}; |x_{r,i}| < \alpha,\ i = 1, \dots, d-1\right\} \to \mathbb{R} \tag{2.43}$$

with a Lipschitz constant $L > 0$, and orthogonal transformations $A_r : \mathbb{R}^d \to \mathbb{R}^d$, $r = 1, \dots, m$, such that

$$\forall\, x \in \partial\Omega \quad \exists\, r \in \{1, \dots, m\} \quad \exists\, x'_r \in \Delta_r :\ x = A_r^{-1}\left(x'_r, a_r(x'_r)\right). \tag{2.44}$$

Under the notation

$$\begin{aligned} \widehat{V}_r^+ &= \left\{(x'_r, x_{r,d}) \in \mathbb{R}^d; a_r(x'_r) < x_{r,d} < a_r(x'_r) + \beta,\ x'_r \in \Delta_r\right\}, \\ \widehat{V}_r^- &= \left\{(x'_r, x_{r,d}) \in \mathbb{R}^d; a_r(x'_r) - \beta < x_{r,d} < a_r(x'_r),\ x'_r \in \Delta_r\right\}, \\ \widehat{\Lambda}_r &= \left\{(x'_r, x_{r,d}); x_{r,d} = a_r(x'_r) \in \mathbb{R},\ x'_r \in \Delta_r\right\}, \end{aligned} \tag{2.45}$$

we have

$$\widehat{V}_r^+ \subset A_r(\Omega), \quad \widehat{\Lambda}_r \subset A_r(\partial\Omega), \quad \widehat{V}_r^- \subset A_r(\mathbb{R}^d \setminus \overline{\Omega}), \quad \partial\Omega \subset \bigcup_{r=1}^m U_r, \tag{2.46}$$

where the sets U_r are defined by the relations

$$\widehat{U}_r = \widehat{V}_r^+ \cup \widehat{\Lambda}_r \cup \widehat{V}_r^-, \quad U_r = A_r^{-1}(\widehat{U}_r). \tag{2.47}$$

The mappings A_r can be written in the form

$$A_r(x) = \mathbb{Q}_r x + x_r^0, \quad x \in \mathbb{R}^d, \tag{2.48}$$

where $x_r^0 \in \mathbb{R}^d$ and $\mathbb{Q}_r$ are orthogonal $d \times d$ matrices, i. e., $\mathbb{Q}_r \mathbb{Q}_r^{\mathrm{T}} = \mathbb{I}$ = unit matrix. Then the transformation of a d-dimensional vector $y \in \mathbb{R}^d$ reads as

$$y \in \mathbb{R}^d \to \mathbb{Q}_r y \in \mathbb{R}^d. \tag{2.49}$$

The sets U_r are open. There exists an open set U_0 such that

$$\overline{U}_0 \subset \Omega, \quad \overline{\Omega} \subset \bigcup_{r=0}^m U_r. \tag{2.50}$$

By the theorem on partition of unity ([KJk77]), there exist functions $\varphi_r \in C_0^\infty(U_r),\ r = 0, \dots, m$, such that $0 \le \varphi_r \le 1$ and

$$\sum_{r=0}^m \varphi_r(x) = 1 \text{ for } x \in \overline{\Omega} \quad \text{and} \quad \sum_{r=1}^m \varphi_r(x) = 1 \text{ for } x \in \partial\Omega. \tag{2.51}$$

Since the functions a_r are Lipschitz-continuous in Δ_r, they are differentiable almost everywhere in Δ_r. Hence, there exists the gradient

$$\nabla a_r(x'_r) = \left(\frac{\partial a_r}{\partial x_{r,1}}(x'_r), \dots, \frac{\partial a_r}{\partial x_{r,d-1}}(x'_r)\right)^{\mathrm{T}} \quad \text{for a. e. } x'_r \in \Delta_r, \tag{2.52}$$

and

$$|\nabla a_r| \le L \quad \text{a. e. in } \Delta_r,\ r = 1, \dots, m. \tag{2.53}$$

(Here a. e. is meant with respect to $(d-1)$-dimensional measure.) Then there exists an outer unit normal

$$\boldsymbol{n}_r\left(x'_r, a_r(x'_r)\right) = \frac{1}{\sqrt{1 + |\nabla a_r(x'_r)|^2}}\left(\nabla a_r(x'_r),\ -1\right) \tag{2.54}$$

to $\partial\widehat{V}_r^+$ for a. e. $X_r = (x_r', a_r(x_r')) \in \widehat{\Lambda}_r$ (with respect to $(d-1)$-dimensional measure defined on $\widehat{\Lambda}_r$ – cf. [KJk77]) and

$$\boldsymbol{n}(x) = \mathbb{Q}_r^{\mathrm{T}} \boldsymbol{n}_r(A_r(x)), \quad \text{a. e. } x \in \partial\Omega,\ A_r(x) \in \widehat{\Lambda}_r, \tag{2.55}$$

is the outer unit normal to $\partial\Omega$.

If we set $\boldsymbol{e}_d = (0, \dots, 0, -1)^{\mathrm{T}} \in \mathbb{R}^d$, then by (2.52) and (2.53)

$$\boldsymbol{n}_r(X_r) \cdot \boldsymbol{e}_d = \frac{1}{\sqrt{1 + |\nabla a_r(x_r')|^2}} \geq \frac{1}{\sqrt{1+L^2}}, \quad X_r \in \widehat{\Lambda}_r, \quad r = 1, \dots, m. \tag{2.56}$$

By virtue of the orthogonality of $\mathbb{Q}_r$, for a. e. $x \in \partial\Omega$, with $A_r(x) \in \widehat{\Lambda}_r$, we have

$$\begin{aligned} \boldsymbol{n}(x) \cdot (\mathbb{Q}_r^{\mathrm{T}}\, \boldsymbol{e}_d) &= \Big(\mathbb{Q}_r^{\mathrm{T}}\, \boldsymbol{n}_r(A_r(x))\Big) \cdot \Big(\mathbb{Q}_r^{\mathrm{T}}\, \boldsymbol{e}_d\Big) \\ &= \Big(\mathbb{Q}_r^{\mathrm{T}}\, \boldsymbol{n}_r(A_r(x))\Big)^{\mathrm{T}} \cdot \Big(\mathbb{Q}_r^{\mathrm{T}}\, \boldsymbol{e}_d\Big) \\ &= \Big(\boldsymbol{n}_r(A_r(x))^{\mathrm{T}}\, \mathbb{Q}_r\Big) \cdot \Big(\mathbb{Q}_r^{\mathrm{T}}\, \boldsymbol{e}_d\Big) \\ &= \boldsymbol{n}_r(A_r(x)) \cdot \boldsymbol{e}_d \geq \frac{1}{\sqrt{1+L^2}}, \quad r = 1, \dots, m. \end{aligned} \tag{2.57}$$

Now we define the function $\boldsymbol{\varphi}$ by

$$\boldsymbol{\varphi}(x) = \sqrt{1+L^2} \sum_{r=1}^{m} \varphi_r(x)\, \mathbb{Q}_r^T\, \boldsymbol{e}_d, \quad x \in \mathbb{R}^d. \tag{2.58}$$

Obviously, $\boldsymbol{\varphi} \in (C_0^\infty(\mathbb{R}^d))^d$ and thus $\boldsymbol{\varphi} \in W^{1,\infty}(\Omega)^d$. Moreover, by (2.51), (2.57) and (2.58),

$$\boldsymbol{\varphi}(x) \cdot \boldsymbol{n}(x) \geq \sum_{r=1}^{m} \varphi_r(x) = 1, \quad x \in \partial\Omega,$$

what we wanted to prove. □

□

Now we shall prove a "global version" of the multiplicative trace inequality.

Lemma 2.8. *There exists a constant $C_M' > 0$ such that*

$$\|v\|_{L^2(\partial\Omega)}^2 \leq C_M' \left\{ |||v||| \left(\|v\|_{L^2(\Omega)}^2 + \sum_{K \in \mathcal{T}_h} h_K \|v\|_{L^2(\partial K)}^2 \right)^{1/2} + \|v\|_{L^2(\Omega)}^2 \right\}, \quad v \in H^1(\Omega, \mathcal{T}_h),\ h \in (0, \bar{h}). \tag{2.59}$$

Proof. Let $v \in H^1(\Omega, \mathcal{T}_h)$, $h \in (0, \bar{h})$ and $K \in \mathcal{T}_h$. Let $\boldsymbol{\varphi} \in (W^{1,\infty}(\Omega))^d$ be the function from Lemma 2.7. By Green's theorem,

$$\int_{\partial K} v^2 \boldsymbol{\varphi} \cdot \boldsymbol{n}\, \mathrm{d}S = \int_K \nabla \cdot (v^2 \boldsymbol{\varphi})\, \mathrm{d}x = \int_K (v^2 \nabla \cdot \boldsymbol{\varphi} + 2v\boldsymbol{\varphi} \cdot \nabla v)\, \mathrm{d}x.$$

The summation over all $K \in \mathcal{T}_h$ implies that

$$\int_{\partial\Omega} v^2 \boldsymbol{\varphi} \cdot \boldsymbol{n}\, \mathrm{d}S + \sum_{\Gamma \in \mathcal{F}_h^I} \int_\Gamma [v^2]\, \boldsymbol{\varphi} \cdot n\, \mathrm{d}S = \sum_{K \in \mathcal{T}_h} \int_K \big(v^2\, \nabla \cdot \boldsymbol{\varphi} + 2v\boldsymbol{\varphi} \cdot \nabla v\big)\, \mathrm{d}x. \tag{2.60}$$

In view of (2.41) and (2.60),

$$\int_{\partial\Omega} v^2\, \mathrm{d}S \leq \int_{\partial\Omega} v^2 \boldsymbol{\varphi} \cdot \boldsymbol{n}\, \mathrm{d}S \leq \sum_{K \in \mathcal{T}_h} \int_K |v^2 \nabla \cdot \boldsymbol{\varphi} + 2v\boldsymbol{\varphi} \cdot \nabla v| \mathrm{d}x + \sum_{\Gamma \in \mathcal{F}_h^I} \int_\Gamma \big|[v^2]\big|\ |\boldsymbol{\varphi}|\, \mathrm{d}S.$$

Taking into account that $\boldsymbol{\varphi} \in (W^{1,\infty}(\Omega))^d$ and using the Cauchy and Young's inequalities, we find that

$$\|v\|_{L^2(\partial\Omega)}^2 \leq \|\boldsymbol{\varphi}\|_{(W^{1,\infty}(\Omega))^d} \Big(\sum_{\Gamma \in \mathcal{F}_h^I} \int_\Gamma \big|[v^2]\big|\, \mathrm{d}S + \|v\|_{L^2(\Omega)}^2 + 2 \sum_{K \in \mathcal{T}_h} \|v\|_{L^2(K)}\, |v|_{H^1(K)} \Big). \tag{2.61}$$

Further, by the Cauchy inequality, (1.20), (1.107), (2.11) and (2.14), we have

$$\sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma \left|[v^2]\right| \,\mathrm{d}S = 2\sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma |[v]\,\langle v\rangle| \,\mathrm{d}S \tag{2.62}$$
$$\le 2\left(\sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma \sigma[v]^2\mathrm{d}S\right)^{1/2}\left(\sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma \sigma^{-1}\langle v\rangle^2\mathrm{d}S\right)^{1/2}$$
$$\le 2C_W^{-1/2}C_G^{1/2}\,J_h^\sigma(v,v)^{1/2}\left(\sum_{K\in\mathcal{T}_h} h_K\|v\|^2_{L^2(\partial K)}\right)^{1/2}.$$

Now, it follows from (2.61), (2.62) and the discrete Cauchy inequality that

$$\|v\|^2_{L^2(\partial\Omega)} \le \|\boldsymbol{\varphi}\|_{(W^{1,\infty}(\Omega))^d}\Big\{2C_W^{-1/2}C_G^{1/2}J_h^\sigma(v,v)^{1/2}\left(\sum_{K\in\mathcal{T}_h} h_K\|v\|^2_{L^2(\partial K)}\right)^{1/2}$$
$$+\|v\|^2_{L^2(\Omega)} + 2\|v\|_{L^2(\Omega)}\,|v|_{H^1(\Omega,\mathcal{T}_h)}\Big\},$$

which implies (2.59) with $C'_M = \max(2C_W^{-1/2}C_G^{1/2},\,2)\|\boldsymbol{\varphi}\|_{(W^{1,\infty}(\Omega))^d}$. □ □

Now we shall apply the above results to the derivation of the consistency estimate of the form b_h. This form can be expressed as

$$b_h(w,v) = b_h^{ID}(w,v) + b_h^N(w,v), \tag{2.63}$$

where

$$b_h^{ID}(w,v) = -\sum_{K\in\mathcal{T}_h}\int_K \sum_{s=1}^d f_s(w)\,\frac{\partial v}{\partial x_s}\,\mathrm{d}x + \sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma H(w|_\Gamma^{(L)},\,w|_\Gamma^{(R)},\,\boldsymbol{n})[v]_\Gamma\,\mathrm{d}S$$
$$+\sum_{\Gamma\in\mathcal{F}_h^D}\int_\Gamma H(w|_\Gamma^{(L)},\,w|_\Gamma^{(L)},\,\boldsymbol{n})\,v|_\Gamma^{(L)}\,\mathrm{d}S \tag{2.64}$$

and, due to (2.19),

$$b_h^N(w,v) = \sum_{\Gamma\in\mathcal{F}_h^N}\int_\Gamma H(w|_\Gamma^{(L)},\,w|_\Gamma^{(L)},\,\boldsymbol{n})v|_\Gamma^{(L)}\,\mathrm{d}S = \sum_{\Gamma\in\mathcal{F}_h^N}\int_\Gamma\sum_{s=1}^d f_s(w|_\Gamma^{(L)})n_s\,v|_\Gamma^{(L)}\,\mathrm{d}S. \tag{2.65}$$

Let us set $\xi = u_h - \Pi_{hp}u \in S_{hp}$. We are interested estimating the expression

$$b_h(u,\xi) - b_h(u_h,\xi) = \left(b_h^{ID}(u,\xi) - b_h^{ID}(u_h,\xi)\right) + \left(b_h^N(u,\xi) - b_h^N(u_h,\xi)\right). \tag{2.66}$$

Then, by (2.34) with $v_h = \xi$,

$$\left|b_h^{ID}(u,\xi) - b_h^{ID}(u_h,\xi)\right| \le C_{b4}|||\xi|||\left(R_b(\eta) + \|\xi\|_{L^2(\Omega)}\right), \tag{2.67}$$

where $R_b(\eta)$ is defined by (2.33).

It remains to estimate the second term on the right-hand side of (2.66).

Lemma 2.9. *Let $u\in H^s(\Omega)$, $u_h\in S_{hp}$, $\xi = u_h - \Pi_{hp}u$. Then*

$$\left|b_h^N(u,\xi) - b_h^N(u_h,\xi)\right| \le C_N\left(R_c(\eta)^2 + |||\xi|||\|\xi\|_{L^2(\Omega)} + \|\xi\|^2_{L^2(\Omega)}\right), \tag{2.68}$$

where

$$R_c(\eta) = \Big(\sum_{K\in\mathcal{T}_h}\left(h_K^{-1}\|\eta\|^2_{L^2(K)} + h_K|\eta|^2_{H^1(K)}\right)\Big)^{1/2} \tag{2.69}$$

and C_N is a constant independent of u, u_h and h.

Proof. By (2.65), Lipschitz-continuity (2.18), Cauchy and Young's inequalities, and the relation $u_h - u = \eta + \xi$, where $\eta = \Pi_{hp}u - u$, we get

$$\begin{aligned} \left|b_h^N(u,\xi) - b_h^N(u_h,\xi)\right| &\le C_L \|u-u_h\|_{L^2(\partial\Omega_N)} \|\xi\|_{L^2(\partial\Omega_N)} \\ &\le C_L \|u-u_h\|_{L^2(\partial\Omega)} \|\xi\|_{L^2(\partial\Omega)} \le C_L \left(\frac{1}{2}\|\eta\|_{L^2(\partial\Omega)}^2 + \frac{3}{2}\|\xi\|_{L^2(\partial\Omega)}^2\right) \end{aligned} \tag{2.70}$$

with $C_L = 2L_H$. Moreover, using the multiplicative trace inequality (1.78) and Young's inequality, we find that

$$\begin{aligned} \|\eta\|_{L^2(\partial\Omega)}^2 &\le \sum_{K\in\mathcal{T}_h} \|\eta\|_{L^2(\partial K)}^2 \le C_M \sum_{K\in\mathcal{T}_h} \left(h_K^{-1}\|\eta\|_{L^2(K)}^2 + \|\eta\|_{L^2(K)}|\eta|_{H^1(K)}\right) \\ &\le C_M \sum_{K\in\mathcal{T}_h} \left(h_K^{-1}\|\eta\|_{L^2(K)}^2 + \frac{1}{2}h_K^{-1}\|\eta\|_{L^2(K)}^2 + \frac{1}{2}h_K|\eta|_{H^1(K)}^2\right) \\ &\le \frac{3}{2}C_M R_c(\eta)^2, \end{aligned} \tag{2.71}$$

where $R_c(\eta)$ is defined in (2.69).

We estimate $\|\xi\|_{L^2(\partial\Omega)}^2$ according to Lemma 2.8. Taking into account that $\xi \in S_{hp}$ and using the multiplicative trace inequality (1.78) and the inverse inequality (1.86), we find that

$$\begin{aligned} \sum_{K\in\mathcal{T}_h} h_K\|\xi\|_{L^2(\partial K)}^2 &\le C_M \sum_{K\in\mathcal{T}_h} h_K \left(\|\xi\|_{L^2(K)}|\xi|_{H^1(K)} + h_K^{-1}\|\xi\|_{L^2(K)}^2\right) \\ &\le C_M(1+C_I)\,\|\xi\|_{L^2(\Omega)}^2. \end{aligned} \tag{2.72}$$

Hence, in view of (2.59) and (2.72), we have

$$\begin{aligned} \|\xi\|_{L^2(\partial\Omega)}^2 &\le C_M' \left\{(C_M(1+C_I)+1)^{1/2}\ |||\xi|||\,\|\xi\|_{L^2(\Omega)} + \|\xi\|_{L^2(\Omega)}^2\right\} \\ &\le C^* \left(|||\xi|||\|\xi\|_{L^2(\Omega)} + \|\xi\|_{L^2(\Omega)}^2\right), \end{aligned} \tag{2.73}$$

where $C^* = C_M'(C_M(1+C_I)+1)^{1/2}$. Finally, (2.70), (2.71) and (2.73) yield estimate (2.68) with $C_N = \frac{1}{2}C_L\max(2C_M, 3C^*)$, which we wanted to prove. □ □

Let us summarize the above results.

Corollary 2.10. *Let $u \in H^s(\Omega)$, $s \ge 2$, $u_h \in S_{hp}$, $\xi = u_h - \Pi_{hp}u$, $\eta = \Pi_{hp}u - u$. Then*

$$\begin{aligned} &|b_h(u,\xi) - b_h(u_h,\xi)| \\ &\quad\le C_b \left(|||\xi|||\left(R_b(\eta) + \|\xi\|_{L^2(\Omega)}\right) + \delta_N\left(R_c(\eta)^2 + \|\xi\|_{L^2(\Omega)}^2\right)\right), \end{aligned} \tag{2.74}$$

where $\delta_N = 0$, if $\partial\Omega_N = \emptyset$, and $\delta_N = 1$, if $\partial\Omega_N \ne \emptyset$.

Proof. Estimate (2.74) is an immediate consequence of (2.67) and (2.68) with the constant $C_b = C_{b4} + C_N$. □ □

2.3.3 Error estimates for the method of lines

Now we derive the error estimates of the method of lines (2.25) under the assumption that the exact solution u satisfies the condition

$$\frac{\partial u}{\partial t} \in L^2(0,T;H^s(\Omega)), \tag{2.75}$$

where $s \ge 2$ is an integer. Assumption (2.75) implies that $u \in C([0,T];H^s(\Omega))$.

Let $\Pi_{hp}u(t)$ be the S_{hp}-interpolation of $u(t)$ ($t \in [0,T]$) from (1.90). We set

$$\xi = u_h - \Pi_{hp}u \in S_{hp}, \quad \eta = \Pi_{hp}u - u \in H^s(\Omega,\mathcal{T}_h). \tag{2.76}$$

Then the error e_h can be expressed as

$$e_h = u_h - u = \xi + \eta. \tag{2.77}$$

Subtracting (2.28) from (2.25b), where we substitute $v_h := \xi$, we get

$$\left(\frac{\partial \xi}{\partial t}, \xi\right) + A_h(\xi, \xi) = b_h(u, \xi) - b_h(u_h, \xi) - \left(\frac{\partial \eta}{\partial t}, \xi\right) - A_h(\eta, \xi). \tag{2.78}$$

(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 shall estimate the individual terms on the right-hand side of (2.78).

The Cauchy inequality implies that

$$\left|\left(\frac{\partial \eta}{\partial t}, \xi\right)\right| \le \left\|\frac{\partial \eta}{\partial t}\right\|_{L^2(\Omega)} \|\xi\|_{L^2(\Omega)}. \tag{2.79}$$

Moreover, using the result of Lemma 1.37, we have

$$|A_h(\eta, \xi)| \le \varepsilon \widetilde{C}_B R_a(\eta) |||\xi|||, \tag{2.80}$$

where $\widetilde{C}_B$ is the constant from (1.129) and

$$R_a(\eta) = \left(\sum_{K \in \mathcal{T}_h} \left(|\eta|^2_{H^1(K)} + h_K^2 |\eta|^2_{H^2(K)} + h_K^{-2} \|\eta\|^2_{L^2(K)}\right)\right)^{1/2}. \tag{2.81}$$

Finally, we define the term

$$R_Q(\eta) = \frac{2C_1^2}{\varepsilon C_C}\Big(R_b(\eta) + \varepsilon R_a(\eta)\Big)^2 + 2C_1\Big(R_c(\eta)^2 + \|\partial_t \eta\|^2_{L^2(\Omega)}\Big), \tag{2.82}$$

where $R_b(\eta)$ is defined by (2.33), $R_c(\eta)$ is defined by (2.69), and the constant C_1 is defined as $C_1 = \max(C_b + 1, \widetilde{C}_B)$. This notation will be useful in the following.

Now we prove the so-called *abstract error estimate*, representing a bound of the error in terms of the S_{hp}-interpolation error η. Let us recall that in order to increase the readability of the derivation of the error estimate, we number constants appearing in the proofs.

Theorem 2.11. *Let Assumptions 2.5 from Section 2.3 be satisfied. Let u be the exact strong solution of problem* (2.1) *satisfying* (2.75) *and let u_h be the approximate solution obtained by scheme* (2.25)*. Then the error $e_h = u_h - u$ satisfies the estimate*

$$\begin{aligned} &\|e_h(t)\|^2_{L^2(\Omega)} + \varepsilon C_C \varepsilon \int_0^T |||e_h(\vartheta)|||^2 \,\mathrm{d}\vartheta \\ &\quad \le C_2(\varepsilon)\left(\int_0^T R_Q(\eta(t))\,\mathrm{d}t + \|\eta(t)\|^2_{L^2(\Omega)} + C_C \int_0^T |||\eta(\vartheta)|||^2 \,\mathrm{d}\vartheta\right), \\ &\qquad\qquad t \in (0, T), \quad h \in (0, \bar{h}), \end{aligned} \tag{2.83}$$

where C_C is the constant from the coercivity inequality (1.140) *of the form $\frac{1}{\varepsilon} A_h = a_h + J_h^\sigma$, $R_Q(\eta)$ is given by* (2.82) *and $C_2(\varepsilon)$ is a constant independent of h and u, but depending on ε (see* (2.93)*).*

Proof. As in (2.76), we set $\xi = u_h - \Pi_{hp} u \in S_{hp}$, $\eta = \Pi_{hp} u - u$. Then (2.77) holds: $e_h = u_h - u = \xi + \eta$. Due to the coercivity (1.140) of the form A_h,

$$\varepsilon C_C |||\xi|||^2 \le A_h(\xi, \xi). \tag{2.84}$$

It follows from (2.78), (2.84) and the relation

$$\left(\frac{\partial \xi}{\partial t}, \xi\right) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\Omega)}, \tag{2.85}$$

that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\Omega)} + \varepsilon\, C_C |||\xi|||^2 \le b_h(u, \xi) - b_h(u_h, \xi) - \left(\frac{\partial \eta}{\partial t}, \xi\right) - A_h(\eta, \xi). \tag{2.86}$$

Now from (2.74), (2.79), (2.80), using the inequality $(\gamma+\delta)^2 \le 2(\gamma^2+\delta^2)$ and Cauchy and Young's inequalities, we derive the estimates

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\Omega)} + 2\varepsilon C_C |||\xi|||^2 \\
&\qquad \le 2C_b \left(|||\xi||| \left(R_b(\eta) \ + \|\xi\|_{L^2(\Omega)} \right) + R_c(\eta)^2 + \|\xi\|^2_{L^2(\Omega)} \right) \\
&\qquad\quad + 2\|\partial_t \eta\|_{L^2(\Omega)} \|\xi\|_{L^2(\Omega)} + 2\varepsilon \widetilde{C}_B R_a(\eta) |||\xi||| \\
&\qquad \le 2C_1 \Big\{ |||\xi||| \left(\|\xi\|_{L^2(\Omega)} + R_b(\eta) + \varepsilon R_a(\eta) \right) \\
&\qquad\qquad + R_c(\eta)^2 + \|\xi\|^2_{L^2(\Omega)} + \|\partial_t \eta\|^2_{L^2(\Omega)} \Big\} \\
&\qquad \le \varepsilon C_C |||\xi|||^2 + \frac{C_1^2}{\varepsilon C_C} \left(\|\xi\|_{L^2(\Omega)} + R_b(\eta) + \varepsilon R_a(\eta) \right)^2 \\
&\qquad\quad + 2C_1 \Big\{ R_c(\eta)^2 + \|\xi\|^2_{L^2(\Omega)} + \|\partial_t \eta\|^2_{L^2(\Omega)} \Big\} \\
&\qquad \le \varepsilon\, C_C |||\xi|||^2 + C_3 \left(1 + \frac{1}{\varepsilon\, C_C} \right) \|\xi\|^2_{L^2(\Omega)} + R_Q(\eta),
\end{aligned} \tag{2.87}$$

where $C_1 = \max(C_b + 1, \widetilde{C}_B)$, $C_3 = 2\max(C_1, C_1^2)$ and $R_Q(\eta)$ is given by (2.82). Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\xi(t)\|^2_{L^2(\Omega)} + \varepsilon C_C |||\xi(t)|||^2 \le C_3 \left(1 + \frac{1}{\varepsilon\, C_C} \right) \|\xi(t)\|^2_{L^2(\Omega)} + R_Q(\eta(t)). \tag{2.88}$$

Since u, $\frac{\partial u}{\partial t} \in L^2(0,T; H^\mu(\Omega))$, the right-hand side of (2.88) is integrable over $(0,T)$. From (2.76) and (2.25c) we see that $\xi(0) = 0$. The integration of (2.88) from 0 to $t \in [0,T]$ yields

$$\begin{aligned}
&\|\xi(t)\|^2_{L^2(\Omega)} + \varepsilon C_C \int_0^t |||\xi(\vartheta)|||^2 \,\mathrm{d}\vartheta \\
&\qquad \le C_3 \left(1 + \frac{1}{\varepsilon\, C_C} \right) \int_0^t \|\xi(\vartheta)\|^2_{L^2(\Omega)} \,\mathrm{d}\vartheta + \int_0^t R_Q(\eta(\vartheta)) \,\mathrm{d}\vartheta.
\end{aligned} \tag{2.89}$$

Now we shall apply Gronwall's Lemma 0.9 with

$$\begin{aligned}
y(t) &= \|\xi(t)\|^2_{L^2(\Omega)}, & q(t) &= \varepsilon C_C \int_0^t |||\xi(\vartheta)|||^2 \,\mathrm{d}\vartheta, \\
r(t) &= C_3 \frac{\varepsilon C_C + 1}{\varepsilon\, C_C}, & z(t) &= \int_0^t R_Q(\eta(\vartheta)) \,\mathrm{d}\vartheta.
\end{aligned}$$

Further, let us set

$$\begin{aligned}
R(\eta, \varepsilon) &= \int_0^T R_Q(\eta(\vartheta)) \,\mathrm{d}\vartheta, \\
c_1(\varepsilon) &= 1 + C_3 \frac{\varepsilon C_C + 1}{\varepsilon C_C} T \exp\left(C_3 \frac{\varepsilon C_C + 1}{\varepsilon C_C} T \right).
\end{aligned} \tag{2.90}$$

We easily show that

$$\begin{aligned}
&z(t) \le \int_0^T R_Q(\eta(\vartheta)) \,\mathrm{d}\vartheta = R(\eta, \varepsilon), \quad \exp\left(\int_\vartheta^t r(s) \mathrm{d}s \right) \le \exp\left(C_3 \frac{\varepsilon C_C + 1}{\varepsilon C_C} T \right), \\
&z(t) + \int_0^t r(\vartheta) z(\vartheta) \exp\left(\int_\vartheta^t r(s) \mathrm{d}s \right) \mathrm{d}\vartheta \le R(\eta, \varepsilon) c_1(\varepsilon).
\end{aligned}$$

This, (2.89) and Gronwall's lemma 0.9 yield the estimate

$$\|\xi(t)\|^2_{L^2(\Omega)} + \varepsilon C_C \int_0^t |||\xi(\vartheta)|||^2 \mathrm{d}\vartheta \le R(\eta, \varepsilon) c_1(\varepsilon), \quad t \in [0,T],\ h \in (0, \bar{h}). \tag{2.91}$$

By virtue of the relation $e_h = \xi + \eta$ and the inequality $(\gamma+\delta)^2 \le 2(\gamma^2+\delta^2)$, we can write

$$\|e_h\|^2_{L^2(\Omega)} \le 2\left(\|\xi\|^2_{L^2(\Omega)} + \|\eta\|^2_{L^2(\Omega)}\right), \qquad |||e_h|||^2 \le 2\left(|||\xi|||^2 + |||\eta|||^2\right).$$

Using (2.91), we deduce that

$$\begin{aligned} &\|e_h(t)\|^2_{L^2(\Omega)} + \varepsilon C_C \int_0^t |||e_h(\vartheta)|||^2 \,\mathrm{d}\vartheta \\ &\le 2\left(R(\eta,\varepsilon)c_1(\varepsilon) + \|\eta(t)\|^2_{L^2(\Omega)} + \varepsilon C_C \int_0^t |||\eta(\vartheta)|||^2 \,\mathrm{d}\vartheta\right),\ t\in[0,T],\ h\in(0,\bar h), \end{aligned} \tag{2.92}$$

which already implies estimate (2.83) with the constant

$$C_2(\varepsilon) = 2\left(1 + C_3\frac{\varepsilon C_C+1}{\varepsilon C_C} T \exp\left(C_3\frac{\varepsilon C_C+1}{\varepsilon C_C}T\right)\right). \tag{2.93}$$

□ □

2.4 Error estimates in terms of h

Now we derive the first main result of this chapter on the error estimate of the method of lines for the solution of the nonlinear convection-diffusion problem. It will be obtained by estimating the right-hand side of (2.83) in terms of h.

We assume that $s \ge 2$ and the exact solution u satisfies the regularity assumption

$$\frac{\partial u}{\partial t} \in L^2(0,T; H^s(\Omega)). \tag{2.94}$$

Then $u \in C([0,T], H^s(\Omega))$. As usual, we put $\eta(t) = u(u) - \Pi_{hp}u(t)$, $t \in (0,T)$, and $\mu = \min(p+1, s)$. Recalling (1.149), we have

$$\begin{aligned} \|\eta(t)\|_{L^2(K)} &\le C_A\, h_K^{\mu}|u(t)|_{H^\mu(K)}, && K\in\mathcal{T}_h,\ t\in(0,T), \\ |\eta(t)|_{H^1(K)} &\le C_A\, h_K^{\mu-1}|u(t)|_{H^\mu(K)}, && K\in\mathcal{T}_h,\ t\in(0,T), \\ |\eta(t)|_{H^2(K)} &\le C_A\, h_K^{\mu-2}|u(t)|_{H^\mu(K)}, && K\in\mathcal{T}_h,\ t\in(0,T), \end{aligned} \tag{2.95}$$

where C_A is the constant from Lemma 1.22. Then, a simple manipulation gives

$$\sum_{K\in\mathcal{T}_h}\left(|\eta(t)|^2_{H^1(K)} + h_K^2|\eta(t)|^2_{H^2(K)} + h_K^{-2}\|\eta(t)\|_{L^2(K)}\right) \le 3C_A^2 h^{2(\mu-1)}|u(t)|^2_{H^\mu(\Omega)},$$

for any $t \in (0,T)$. This together with (2.81) implies that

$$R_a(\eta(t)) = R_a\left(u(t) - \Pi_{hp}u(t)\right) \le \sqrt{3}C_A h^{\mu-1}|u(t)|_{H^\mu(\Omega)}, \quad t\in(0,T). \tag{2.96}$$

Similarly, from (2.33), we obtain

$$R_b(\eta(t)) = R_b\left(u(t) - \Pi_{hp}u(t)\right) \le \sqrt{2}C_A h^{\mu}|u(t)|_{H^\mu(\Omega)}, \quad t\in(0,T). \tag{2.97}$$

Moreover, (2.69) and (2.95) give

$$R_c(\eta(t)) \le \sqrt{2}C_A h^{\mu-1/2}|u(t)|_{H^\mu(\Omega)}, \quad t\in(0,T). \tag{2.98}$$

Further, we shall use the notation $\partial_t u = \partial u/\partial t$ and $\partial_t(\Pi_{hp}u) = \partial(\Pi_{hp}u)/\partial t$. Then definition (1.90) of the interpolation operator Π_{hp} and the relation

$$\partial_t(\Pi_{hp}u(t)) = \Pi_{hp}(\partial_t u(t)) \in S_{hp} \tag{2.99}$$

imply that

$$\|\partial_t\eta\|_{L^2(\Omega)} = \|\partial_t(\Pi_{hp}u - u)\|_{L^2(\Omega)} = \|\Pi_{hp}(\partial_t u) - \partial_t u\|_{L^2(\Omega)} \le C_A h^{\mu}\,|\partial_t u|_{H^\mu(\Omega)}\,. \tag{2.100}$$

Exercise 2.12. *Using the theorem on differentiating an integral with respect to a parameter, prove (2.99).*

Summarizing (2.82) with (2.96), (2.97), (2.98) and (2.100), we see that for $t \in (0,T)$, we have

$$\begin{aligned} R_Q(\eta(t)) &= \frac{2C_1^2}{\varepsilon C_C}\Big(R_b(\eta(t)) + \varepsilon R_a(\eta(t))\Big)^2 + 2C_1\Big(R_c(\eta(t))^2 + \|\partial_t\eta(t)\|^2_{L^2(\Omega)}\Big) \\ &\le \frac{2C_1^2}{\varepsilon C_C}\Big(\sqrt{2}C_A h^\mu |u(t)|_{H^\mu(\Omega)} + \varepsilon\sqrt{3}C_A h^{\mu-1}|u(t)|_{H^\mu(\Omega)}\Big)^2 \\ &\quad + 4C_1C_A^2h^{2\mu-1}|u(t)|^2_{H^\mu(\Omega)} + 2C_1C_A^2h^{2\mu}\,|\partial_t u(t)|^2_{H^\mu(\Omega)} \\ &\le \frac{2C_1^2C_A^2}{\varepsilon C_C}h^{2(\mu-1)}|u(t)|^2_{H^\mu(\Omega)}(2h^2 + 2\sqrt{6}\varepsilon h + 3\varepsilon^2) \\ &\quad + 4C_1C_A^2h^{2(\mu-1)}\left(|u(t)|^2_{H^\mu(\Omega)} + |\partial_t u(t)|^2_{H^\mu(\Omega)}\right)(h+h^2) \\ &\le C_4 h^{2(\mu-1)}\left(\varepsilon^{-1}h^2 + h + \varepsilon + h^2\right)\left(|u(t)|^2_{H^\mu(\Omega)} + |\partial_t u(t)|^2_{H^\mu(\Omega)}\right), \end{aligned} \tag{2.101}$$

where

$$C_4 = 4C_A^2\max\left(\frac{\sqrt{6}C_1^2}{C_C}, C_1\right). \tag{2.102}$$

The integration of (2.101) over $(0,T)$ yields

$$\begin{aligned} &\int_0^T R_Q(\eta(t))\,\mathrm{d}t \\ &\quad\le C_4 h^{2\mu-2}\left(\varepsilon^{-1}h^2 + h + \varepsilon + h^2\right)\left(|u|^2_{L^2(0,T;H^\mu(\Omega))} + |\partial_t u|^2_{L^2(0,T;H^\mu(\Omega))}\right). \end{aligned} \tag{2.103}$$

Furthermore, using (2.29), (1.119) and (2.95), we get

$$\begin{aligned} |||\eta(t)|||^2 &= |\eta(t)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(\eta(t),\eta(t)) \\ &\le \sum_{K\in\mathcal{T}_h}\left(|\eta(t)|^2_{H^1(K)} + C_W C_M C_T^{-1}\left(3h_K^{-2}\|\eta(t)\|^2_{L^2(K)} + |\eta(t)|^2_{H^1(K)}\right)\right) \\ &\le C_5 h^{2(\mu-1)}|u(t)|^2_{H^\mu(\Omega)}, \quad t\in(0,T), \end{aligned} \tag{2.104}$$

where $C_5 = C_A^2(4C_WC_MC_T^{-1}+1)$. Hence,

$$\varepsilon C_C\int_0^T |||\eta(t)|||^2\,\mathrm{d}t \le \varepsilon C_C C_5 h^{2(\mu-1)}|u|^2_{L^2(0,T;H^\mu(\Omega))}. \tag{2.105}$$

Remark 2.13. *The above estimates illustrate a typical situation in numerical analysis, where a number of constants appear. They are often defined recursively in a complicated way on the basis of constants introduced before. As an example we illustrate this situation by the process leading to the determination of the constant C_4 defined by* (2.102)*. This relation contains the constant C_A appearing in Lemmas 1.22 and 1.24 and the constant C_1, which is defined recursively in the following way:*

$$\begin{aligned} C_1 &= \max(C_b + 1, \widetilde{C}_B), \\ C_b &= C_{b4} + C_N, \\ C_{b4} &= \max(C_{b2}, C_{b3}), \\ C_N &= \frac{1}{2}C_L\max(2C_M, 3C^*), \\ C_L &= 2L_H, \\ C^* &= C'_M(C_M(1+C_I)+1)^{1/2}, \\ C'_M &= \max(2C_W^{-1/2}C_G^{1/2}, 2)\|\boldsymbol{\varphi}\|_{(W^{1,\infty}(\Omega))^d}, \\ C_{b2} &= C_{b1}(1 + C_M(1+C_I))^{1/2}, \\ C_{b3} &= C_{b1}(1 + 3C_M/3)^{1/2}, \\ C_{b1} &= (\max(d\,L_f^2, 2L_H^2\,C_G/C_W))^{1/2}, \end{aligned}$$

where $\widetilde{C}_B$ is the constant from Lemma 1.37, C_M is the constant from Lemma 1.37 (multiplicative trace inequality), L_H is the constant from the Lipschitz continuity (2.18) *of the numerical flux H, C_I is the constant from the inverse inequality* (1.86)*, C_W is the constant from the definition* (1.104) *of the weight in the penalty form J_h^σ, C_G is the constant from the equivalence condition* (1.20)*, $\boldsymbol{\varphi}$ is the function from Lemma 2.7 and L_f is the constant from the Lipschitz continuity of the convective fluxes f_s, $s = 1,\dots,d$.*

Now we are ready to present the final error estimates.

Theorem 2.14. *Let Assumptions 2.5 from Section 2.3 be satisfied. Let u be the exact strong solution of problem* (2.1) *satisfying* (2.75) *and let u_h be the approximate solution obtained by the scheme* (2.25)*. Then the error $e_h = u_h - u$ satisfies the estimate*

$$\max_{t\in[0,T]} \|e_h(t)\|^2_{L^2(\Omega)} + C_C\varepsilon \int_0^T |||e_h(\vartheta)|||^2 \,\mathrm{d}\vartheta \tag{2.106}$$
$$\le \tilde{C}_2(\varepsilon) h^{2(\mu-1)} \left(|u|^2_{L^2(0,T;H^\mu(\Omega))} + |\partial_t u|^2_{L^2(0,T;H^\mu(\Omega))} \right), \quad h \in (0, \bar{h}),$$

where C_C is the constant from the coercivity inequality (1.140) *of the form $\frac{1}{\varepsilon}A_h = a_h + J_h^\sigma$ and $\tilde{C}_2(\varepsilon)$ is a constant independent of h and u, specified in the proof.*

Proof. If $t \in [0,T]$, then the estimation of the right-hand side of (2.83) by (2.103), (2.105) and (2.95) implies that

$$\|e_h(t)\|^2_{L^2(\Omega)} + C_C\varepsilon \int_0^T |||e_h(\vartheta)|||^2 \,\mathrm{d}\vartheta$$
$$\le C_2(\varepsilon) \left(\int_0^T R_Q(\eta(t)) \,\mathrm{d}t + \|\eta(t)\|^2_{L^2(\Omega)} + \varepsilon C_C \int_0^T |||\eta(\vartheta)|||^2 \,\mathrm{d}\vartheta \right),$$
$$\le \tilde{C}_2(\varepsilon) h^{2\mu-2} \left(|u|^2_{L^2(0,T;H^\mu(\Omega))} + |\partial_t u|^2_{L^2(0,T;H^\mu(\Omega))} \right),$$

where $C_2(\varepsilon)$ is the constant from Theorem 2.11 given by (2.93) and

$$\tilde{C}_2(\varepsilon) = C_2(\varepsilon)(C_4 + C_C C_5 + C_A^2) \left(\varepsilon^{-1}\bar{h}^2 + \bar{h} + \varepsilon + \bar{h}^2 \right). \tag{2.107}$$

This proves (2.106). □ □

Remark 2.15. *Estimate (2.106) implies that*

$$\|u - u_h\|_{L^\infty(0,T;L^2(\Omega))} = O(h^{\mu-1}) \quad \text{for } h \to 0+. \tag{2.108}$$

This is in contrast to the approximation properties (1.98) implying that

$$\|u - \Pi_{hp} u\|_{L^\infty(0,T;L^2(\Omega))} = O(h^\mu). \tag{2.109}$$

Numerical experiments presented in the next section demonstrate that the error estimate (2.106) is suboptimal in the $L^\infty(0,T;L^2(\Omega))$-norm. Similarly as in Section 1.7.2 we can derive optimal error estimate in this norm. This is the subject of the next section.

Remark 2.16. *From* (2.107) *and (2.93) we can see that the error estimate* (2.106) *cannot be used for ε very small, because the definition* (2.93) *of the constant $C_2(\varepsilon)$ contains the term of the form $\exp(C/\varepsilon)$, which blows up exponentially for $\varepsilon \to 0+$. This is caused by the technique used in the theoretical analysis (application of Young's inequality and Gronwall's Lemma) in order to overcome the nonlinearity in the convective terms. The nonlinearity of the convective terms represents a serious obstacle for obtaining a uniform error estimate with respect to $\varepsilon \to 0+$. In Section 2.6 we shall be concerned with error estimates of the DGM applied to the numerical solution of a linear convection-diffusion-reaction equation, uniform with respect to the diffusion parameter $\varepsilon \to 0+$.*

2.5 Optimal $L^\infty(0,T;L^2(\Omega))$-error estimate

With respect to Remark 2.15, in this section we derive an optimal error estimate in the $L^\infty(0,T;L^2(\Omega))$-norm. Similarly as in Section 1.7.2, the analysis is based on the *duality technique*. Therefore, we consider only the SIPG variant of the DGM and the Dirichlet boundary condition on the whole boundary $\partial\Omega$.

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a *bounded convex polygonal* (if $d = 2$) or *polyhedral* (if $d = 3$) domain with Lipschitz boundary $\partial\Omega$ and $T > 0$. We are concerned with the nonstationary nonlinear convection-diffusion problem to find $u : Q_T = \Omega \times (0,T) \to \mathbb{R}$ such that

$$\frac{\partial u}{\partial t} + \sum_{s=1}^d \frac{\partial f_s(u)}{\partial x_s} = \varepsilon \Delta u + g \quad \text{in } Q_T, \tag{2.110a}$$
$$u\big|_{\partial\Omega\times(0,T)} = u_D, \tag{2.110b}$$
$$u(x,0) = u^0(x), \quad x \in \Omega. \tag{2.110c}$$

The diffusion coefficient $\varepsilon > 0$ is a given constant, $g : Q_T \to \mathbb{R}$, $u_D : \partial\Omega \times (0,T) \to \mathbb{R}$ and $u^0 : \Omega \to \mathbb{R}$ are given functions satisfying (2.2c), (2.2d) with $\partial\Omega_D = \partial\Omega$, (2.2f), and $f_s \in C^1(\mathbb{R})$, $s = 1, \dots, d$, are fluxes satisfying (2.2a).

Let us recall the definitions of the forms introduced in Section 2.1 by (2.9), (2.10) (with $\Theta = 1$), (2.13), (2.11) and (2.23). Namely, for functions $u, \varphi \in H^2(\Omega, \mathcal{T}_h)$ we write

$$A_h(w,v) = \varepsilon a_h(w,v) + \varepsilon J_h^\sigma(w,v), \tag{2.111}$$

$$a_h(u,\varphi) = \sum_{K\in\mathcal{T}_h} \int_K \nabla u \cdot \nabla\varphi \,\mathrm{d}x - \sum_{\Gamma\in\mathcal{F}_h} \int_\Gamma \left(\langle\nabla u\rangle \cdot \boldsymbol{n}[\varphi] + \langle\nabla\varphi\rangle \cdot \boldsymbol{n}[u]\right) \mathrm{d}S, \tag{2.112}$$

$$J_h^\sigma(u,\varphi) = \sum_{\Gamma\in\mathcal{F}_h} \int_\Gamma \sigma [u]\,[\varphi] \,\mathrm{d}S, \tag{2.113}$$

$$\ell_h(\varphi)(t) = \int_\Omega g(t)\,\varphi \,\mathrm{d}x + \varepsilon \sum_{\Gamma\in\mathcal{F}_h^B} \int_\Gamma \Big(\sigma\varphi - (\nabla\varphi\cdot\boldsymbol{n})\Big) u_D(t) \,\mathrm{d}S, \tag{2.114}$$

$$\begin{aligned} b_h(u,\varphi) = &- \sum_{K\in\mathcal{T}_h} \int_K \sum_{s=1}^d f_s(u) \frac{\partial\varphi}{\partial x_s} \,\mathrm{d}x + \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma H\left(u|_\Gamma^{(L)}, u|_\Gamma^{(R)}, \boldsymbol{n}\right) [\varphi]_\Gamma \,\mathrm{d}S \\ &+ \sum_{\Gamma\in\mathcal{F}_h^B} \int_\Gamma H(u|_\Gamma^{(L)}, u|_\Gamma^{(L)}, \boldsymbol{n}) \varphi_\Gamma^{(L)} \,\mathrm{d}S. \end{aligned}$$

By $(\cdot,\cdot)$ we denote the scalar product in the space $L^2(\Omega)$. The weight σ is again defined by (2.14). We assume that the numerical flux H has properties (2.18) – (2.20) from Section 2.2.

Let the exact solution u of problem (2.110) satisfy the regularity condition (2.94). Moreover, let $u_h \in C^1([0,T]; S_{hp})$ denote the approximate solution defined by (2.25) and let Π_{hp} be the operator of the $L^2(\Omega)$-projection on the space S_{hp} (cf. (1.90)).

In Section 2.3.3, we derived the (sub-optimal) estimate from identity (2.78). The term $A_h(\Pi_{hp}u - u, \xi)$ appearing on the right-hand side of (2.78) cannot be estimated in "an optimal way" (i.e., of order $O(h^\mu)$), because, by virtue of (2.80) and (2.96),

$$|A_h(\Pi_{hp}u - u, \xi)| = |A_h(\eta,\xi)| \le \varepsilon \widetilde{C}_B R_a(\eta) |||\xi|||,$$

and $R_a(\eta) = O(h^{\mu-1})$. Therefore, instead of the $L^2(\Omega)$-projection Π_{hp}, we introduce a new projection P_{hp}, for which the terms mentioned above vanish.

Hence, for every $h \in (0, \bar{h})$ and $t \in [0,T]$, we define the function $\mathrm{P}_{hp}u(t)$ as the *A_h-projection* of $u(t)$ on S_{hp}, i.e., a function satisfying the conditions

$$\mathrm{P}_{hp}u(t) \in S_{hp}, \qquad A_h(\mathrm{P}_{hp}u(t), \varphi_h) = A_h(u(t), \varphi_h) \quad \forall\, \varphi_h \in S_{hp}. \tag{2.115}$$

We are interested in estimates of the functions

$$\chi(t) = u(t) - \mathrm{P}_{hp}u(t) \ \text{ and } \ \partial_t\chi(t) = \frac{\partial}{\partial t}\chi(t) = \frac{\partial}{\partial t}\left(u(t) - \mathrm{P}_{hp}u(t)\right), \quad t \in [0,T],$$

in the DG-norm $|||\cdot|||$ given by (2.29) and in the $L^2(\Omega)$-norm. First, we derive estimates of these functions in the DG-norm.

Lemma 2.17. *There exists a constant $C_{P,\mathrm{e}} > 0$ independent of u, ε and h such that*

$$|||\chi(t)||| \le C_{P,\mathrm{e}}\, h^{\mu-1} |u(t)|_{H^\mu(\Omega)}, \quad t \in [0,T], \tag{2.116}$$

$$|||\partial_t\chi(t)||| \le C_{P,\mathrm{e}}\, h^{\mu-1} |\partial_t u(u)|_{H^\mu(\Omega)}, \quad t \in [0,T], \tag{2.117}$$

for all $h \in (0, \bar{h})$.

Proof. In what follows we usually omit the argument t of the functions $u, \mathrm{P}_{hp}u, \Pi_{hp}u$, etc. By (1.138) and (2.115), we obtain

$$\begin{aligned} \frac{1}{2}\varepsilon |||\Pi_{hp}u - \mathrm{P}_{hp}u|||^2 &\le A_h(\Pi_{hp}u - \mathrm{P}_{hp}u,\, \Pi_{hp}u - \mathrm{P}_{hp}u) \\ &= A_h(\Pi_{hp}u - \mathrm{P}_{hp}u,\, \Pi_{hp}u - \mathrm{P}_{hp}u) + \underbrace{A_h(\mathrm{P}_{hp}u - u,\, \Pi_{hp}u - \mathrm{P}_{hp}u)}_{=0} \\ &= A_h(\Pi_{hp}u - u,\, \Pi_{hp}u - \mathrm{P}_{hp}u). \end{aligned} \tag{2.118}$$

Using the result of Lemma 1.37, we find that

$$A_h(\Pi_{hp}u - u,\, \Pi_{hp}u - \mathrm{P}_{hp}u) \le \varepsilon \widetilde{C}_B\, R_a(\Pi_{hp}u - u)\, |||\Pi_{hp}u - \mathrm{P}_{hp}u|||,$$

where R_a is given by (2.81). This and (2.118) imply that

$$|||\Pi_{hp}u - \mathrm{P}_{hp}u||| \le 2\widetilde{C}_B\, R_a(\Pi_{hp}u - u). \tag{2.119}$$

Further, recalling (1.125), we have

$$|||u - \Pi_{hp}u||| \le C_\sigma R_a(u - \Pi_{hp}u). \tag{2.120}$$

Now it is sufficient to use the triangle inequality

$$|||\chi||| = |||u - \mathrm{P}_{hp}u||| \le |||u - \Pi_{hp}u||| + |||\Pi_{hp}u - \mathrm{P}_{hp}u|||,$$

which implies that

$$|||\chi||| \le (C_\sigma + 2\widetilde{C}_B)\, R_a(\Pi_{hp}u - u). \tag{2.121}$$

Finally, the combination of (2.96) and (2.121) gives

$$|||\chi(t)||| \le \sqrt{3}C_A(C_\sigma + 2\widetilde{C}_B)h^{\mu-1}|u(t)|_{H^\mu(\Omega)}, \quad t \in (0,T),$$

which proves (2.116) with $C_{P,\mathrm{e}} = \sqrt{3}C_A(C_\sigma + 2\widetilde{C}_B)$.

Let us deal now with the norm $|||\partial_t\chi|||$. As

$$A_h\left(u(t) - \mathrm{P}_{hp}u(t),\, \varphi_h\right) = 0 \qquad \forall\, \varphi_h \in S_{hp},\ \forall\, t \in (0,T),$$

from the definitions (2.111) of A_h, for all $\varphi_h \in S_{hp}$, we have

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\left(A_h\left(u(t) - \mathrm{P}_{hp}u(t),\, \varphi_h\right)\right) = A_h\left(\frac{\partial(u(t) - \mathrm{P}_{hp}u(t))}{\partial t},\, \varphi_h\right), \tag{2.122}$$

i.e.,

$$A_h(\partial_t\chi, \varphi_h) = 0 \qquad \forall\, \varphi_h \in S_{hp}. \tag{2.123}$$

Similarly as in (2.118), using the coercivity (1.140) of the form A_h and relation (1.129) from Lemma 1.37, we find that

$$\begin{aligned}
&\frac{\varepsilon}{2}|||\partial_t(\Pi_{hp}u - \mathrm{P}_{hp}u)|||^2 \\
&\quad\le A_h\left(\partial_t(\Pi_{hp}u - \mathrm{P}_{hp}u), \partial_t(\Pi_{hp}u - \mathrm{P}_{hp}u)\right) + \underbrace{A_h\left(\partial_t(\mathrm{P}_{hp}u - u),\, \partial_t(\Pi_{hp}u - \mathrm{P}_{hp}u)\right)}_{=0} \\
&\quad= A_h\left(\partial_t(\Pi_{hp}u - u),\, \partial_t(\Pi_{hp}u - \mathrm{P}_{hp}u)\right) \\
&\quad\le \varepsilon\widetilde{C}_B\, R_a\left(\partial_t(\Pi_{hp}u - u)\right)\, |||\partial_t(\Pi_{hp}u - \mathrm{P}_{hp}u)|||.
\end{aligned}$$

Hence, we have

$$|||\partial_t(\Pi_{hp}u - \mathrm{P}_{hp}u)||| \le 2\widetilde{C}_B\, R_a\left(\partial_t(\Pi_{hp}u - u)\right).$$

Then, similarly as in (2.120), we get

$$|||\partial_t(u - \Pi_{hp}u)||| \le C_\sigma R_a\left(\partial_t(u - \Pi_{hp}u)\right),$$

which together with the triangle inequality gives

$$\begin{aligned}
|||\partial_t(u - \mathrm{P}_{hp}u)(t)||| &\le |||\partial_t(u - \Pi_{hp}u)(t)||| + |||\partial_t(\Pi_{hp}u - \mathrm{P}_{hp}u)(t)||| \\
&\le (2\widetilde{C}_B + C_\sigma)R_a\left(\partial_t(u - \Pi_{hp}u)(t)\right), \quad t \in (0,T).
\end{aligned} \tag{2.124}$$

Finally, we use relation (2.99) and estimate (2.96) rewritten for $\partial_t u(t) - \Pi_{hp}(\partial_t u(t))$:

$$R_a\big(\partial_t u(t) - \Pi_{hp}(\partial_t u(t))\big) \le \sqrt{3}C_A h^{\mu-1}\, |\partial_t u(t)|_{H^\mu(\Omega)}.$$

This and (2.124) already give (2.117). □ □

In what follows, for an arbitrary $z \in L^2(\Omega)$ we shall consider the elliptic *dual problem* (1.155): Given $z \in L^2(\Omega)$, find ψ such that

$$-\Delta\psi = z \qquad \text{in } \Omega, \quad \psi|_{\partial\Omega} = 0. \tag{2.125}$$

Similarly as in (1.157), the weak formulation of problem (2.125) reads: Find $\psi \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla\psi \cdot \nabla v \,\mathrm{d}x = \int_\Omega z\, v \,\mathrm{d}x \qquad \forall\, v \in H_0^1(\Omega). \tag{2.126}$$

As the domain Ω is convex, for every $z \in L^2(\Omega)$ the weak solution ψ is regular, i.e., $\psi \in H^2(\Omega)$, and there exists a constant $C_D > 0$, independent of z such that

$$\|\psi\|_{H^2(\Omega)} \le C_D \|z\|_{L^2(\Omega)}, \tag{2.127}$$

as follows from [Gri92]. Let us note that $H^2(\Omega) \subset C(\overline{\Omega})$.

Further, let $\Pi_{h1}\psi$ be the piecewise linear $L^2(\Omega)$-projection of the function ψ on S_{h1} (cf. (1.91)). Obviously, using (1.125), and (2.96) with $\mu = 2$, we have

$$\|\psi - \Pi_{h1}\psi\|_{1,\sigma} \le C_\sigma R_a(\psi - \Pi_{h1}\psi) \le \sqrt{3} C_A C_\sigma h |\psi|_{H^2(\Omega)}. \tag{2.128}$$

Finally, taking into account that the form A_h is the ε multiple of the form A_h from Chapter 1 and using estimate (1.122), we have

$$|A_h(u,v)| \le 2\,\varepsilon \|u\|_{1,\sigma}\|v\|_{1,\sigma} \quad \forall u, v \in H^2(\Omega, \mathcal{T}_h). \tag{2.129}$$

Now we shall use the dual problem (2.125) to obtain $L^2(\Omega)$-optimal error estimates for $\chi = u - \mathrm{P}_{hp}u$ and $\partial_t\chi = (u - \mathrm{P}_{hp}u)_t$.

Lemma 2.18. *There exists a constant $C_{P,\mathrm{L}} > 0$ independent of ε such that*

$$\|\chi(t)\|_{L^2(\Omega)} \le C_{P,\mathrm{L}} h^\mu |u(t)|_{H^\mu(\Omega)}, \qquad t \in (0,T), \tag{2.130}$$

$$\|\partial_t\chi(t)\|_{L^2(\Omega)} \le C_{P,\mathrm{L}} h^\mu |\partial_t u(t)|_{H^\mu(\Omega)}, \quad t \in (0,T), \tag{2.131}$$

for all $h \in (0, \bar{h})$.

Proof. We have

$$\|\chi\|_{L^2(\Omega)} = \sup_{0 \ne z \in L^2(\Omega)} \frac{|(\chi, z)|}{\|z\|_{L^2(\Omega)}}. \tag{2.132}$$

Taking into account that the form A_h is the ε multiple of the form A_h from Chapter 1, we see that by Lemma 1.48, for $z \in L^2(\Omega)$ and ψ satisfying (2.125), we have

$$(\chi, z) = \frac{1}{\varepsilon} A_h(\psi, \chi). \tag{2.133}$$

Further, the symmetry of A_h and (2.115) give

$$A_h(\Pi_{h1}\psi, \chi) = A_h(\chi, \Pi_{h1}\psi) = A_h(u - \mathrm{P}_{hp}u,\, \Pi_{h1}\psi) = 0, \tag{2.134}$$

and therefore,

$$(\chi, z) = \frac{1}{\varepsilon} A_h(\psi - \Pi_{h1}\psi, \chi). \tag{2.135}$$

Now, using (2.129), we have

$$|(\chi, z)| = \frac{1}{\varepsilon} |A_h(\psi - \Pi_{h1}\psi, \chi)| \le 2 \|\psi - \Pi_{h1}\psi\|_{1,\sigma} \|\chi\|_{1,\sigma}. \tag{2.136}$$

Moreover, by (2.128) and (2.127), we obtain

$$\|\psi - \Pi_{h1}\psi\|_{1,\sigma} \le \sqrt{3} C_A C_\sigma h |\psi|_{H^2(\Omega)} \le \sqrt{3} C_A C_\sigma C_D h \|z\|_{L^2(\Omega)}. \tag{2.137}$$

Triangle inequality, (1.125), (1.126), (2.96) and (2.116) imply the estimate

$$\begin{aligned}\|\chi(t)\|_{1,\sigma} &= \|u - \mathrm{P}_{hp}u\|_{1,\sigma} \le \|u - \Pi_{hp}u\|_{1,\sigma} + \|\Pi_{hp}u - \mathrm{P}_{hp}u\|_{1,\sigma} \\ &\le C_\sigma R_a(u - \Pi_{hp}u) + \widetilde{C}_\sigma |||\Pi_{hp}u - \mathrm{P}_{hp}u||| \\ &\le C_\sigma R_a(u - \Pi_{hp}u) + \widetilde{C}_\sigma |||\Pi_{hp}u - u||| + \widetilde{C}_\sigma |||u - \mathrm{P}_{hp}u||| \\ &\le C_\sigma R_a(u - \Pi_{hp}u) + \widetilde{C}_\sigma C_\sigma R_a(u - \Pi_{hp}u) + \widetilde{C}_\sigma |||\chi||| \\ &\le C_\sigma(1 + \widetilde{C}_\sigma)\sqrt{3}C_A h^{\mu-1}|u(t)|_{H^\mu(\Omega)} + \widetilde{C}_\sigma C_{P,\mathrm{e}}\, h^{\mu-1}|u(t)|_{H^\mu(\Omega)} \\ &= C_6 h^{\mu-1}|u(t)|_{H^\mu(\Omega)}, \quad t \in (0,T),\end{aligned} \tag{2.138}$$

where $C_6 = C_\sigma(1 + \widetilde{C}_\sigma)\sqrt{3}C_A + \widetilde{C}_\sigma C_{P,\mathrm{e}}$. Summarizing (2.136), (2.137) and (2.138), we find that

$$\begin{aligned}(\chi(t), z) \le& 2\sqrt{3}C_A C_\sigma C_D h \|z\|_{L^2(\Omega)} C_6 h^{\mu-1}|u(t)|_{H^\mu(\Omega)} \\ =& C_{P,\mathrm{L}}\, h^\mu |u(t)|_{H^\mu(\Omega)} \|z\|_{L^2(\Omega)}, \quad t \in (0,T),\end{aligned}$$

where $C_{P,\mathrm{L}} = 2\sqrt{3}C_A C_\sigma C_D C_6$. Hence,

$$\|\chi(t)\|_{L^2(\Omega)} = \sup_{0 \ne z \in L^2(\Omega)} \frac{|(\chi(t), z)|}{\|z\|_{L^2(\Omega)}} \le C_{P,\mathrm{L}}\, h^\mu |u(t)|_{H^\mu(\Omega)}, \quad t \in (0,T),$$

which completes the proof of (2.130).

Finally, let us prove estimate (2.131). Differentiating (2.115) with respect to t yields

$$A_h(\partial_t \chi, \varphi_h) = 0 \quad \forall \varphi_h \in S_{hp}. \tag{2.139}$$

We have

$$\|\partial_t \chi\|_{L^2(\Omega)} = \sup_{0 \ne z \in L^2(\Omega)} \frac{|(\partial_t \chi, z)|}{\|z\|_{L^2(\Omega)}}. \tag{2.140}$$

Similarly as in (2.133), we get

$$(\partial_t \chi, z) = \frac{1}{\varepsilon} A_h(\psi, \partial_t \chi). \tag{2.141}$$

The symmetry of A_h and (2.139) imply that

$$A_h(\Pi_{h1}\psi, \partial_t \chi) = A_h(\partial_t \chi, \Pi_{h1}\psi) = A_h\left(\partial_t(u - \mathrm{P}_{hp}u), \Pi_{h1}\psi\right) = 0.$$

These relations, (2.141) and (2.129) yield

$$|(\partial_t \chi, z)| = \frac{1}{\varepsilon}|A_h(\psi - \Pi_{h1}\psi, \partial_t \chi)| \le 2\|\psi - \Pi_{h1}\psi\|_{1,\sigma} \|\partial_t \chi\|_{1,\sigma}. \tag{2.142}$$

The term $\|\psi - \Pi_{h1}\psi\|_{1,\sigma}$ is estimated by (2.137) and similarly as in (2.138), we obtain

$$\|\partial_t \chi(t)\|_{1,\sigma} \le C_6 h^{\mu-1}|\partial_t u(t)|_{H^\mu(\Omega)}, \quad t \in (0,T). \tag{2.143}$$

Finally, from (2.140), (2.142), (2.137) and (2.143), we arrive at estimate (2.131).

□ □

Let us note that assuming the symmetry of the form A_h is crucial in the presented proof. It enables us to exchange arguments in (2.134). This is not possible in the NIPG and IIPG methods, where the analysis of optimal $L^\infty(L^2)$-error estimates still represents an open problem.

Lemma 2.19. *Let us assume that u is the solution of the continuous problem* (2.110) *satisfying condition* (2.75), *u_h is the solution of the discrete problem* (2.25), *$\mathrm{P}_{hp}u$ is defined by* (2.115), *and $\zeta = \mathrm{P}_{hp}u - u_h \in S_{hp}$. Then there exists a constant $C_b > 0$, independent of $h \in (0, \bar{h})$, such that*

$$|b_h(u, \zeta) - b_h(u_h, \zeta)| \le C_b |||\zeta||| \left(h^\mu |u|_{H^\mu(\Omega)} + \|\zeta\|_{L^2(\Omega)}\right). \tag{2.144}$$

Proof. We proceed similarly as in the proof of Lemma 2.6. The triangle inequality gives

$$|b_h(u,\zeta) - b_h(u_h,\zeta)| \le |b_h(u,\zeta) - b_h(\mathrm{P}_{hp}u,\zeta)| + |b_h(\mathrm{P}_{hp}u,\zeta) - b_h(u_h,\zeta)|. \tag{2.145}$$

Applying (2.30) with $\bar{u} := \mathrm{P}_{hp}u$ and $v := \zeta \in S_{hp}$, we get

$$|b_h(u,\zeta) - b_h(\mathrm{P}_{hp}u,\zeta)| \le C_{b1}|||\zeta||| \left(\|\chi\|^2_{L^2(\Omega)} + \sum_{K\in\mathcal{T}_h} h_K \|\chi\|^2_{L^2(\partial K)} \right)^{1/2}. \tag{2.146}$$

(Let us recall that $\chi = u - \mathrm{P}_{hp}u$). The multiplicative trace inequality (1.78) and the Cauchy inequality give

$$\begin{aligned}
\sum_{K\in\mathcal{T}_h} h_K \|\chi\|^2_{L^2(\partial K)} &\le C_M \sum_{K\in\mathcal{T}_h} \left(h_K |\chi|_{H^1(K)} \|\chi\|_{L^2(K)} + \|\chi\|^2_{L^2(K)} \right) \\
&\le C_M \left(h \left(\sum_{K\in\mathcal{T}_h} |\chi|^2_{H^1(K)} \right)^{1/2} \left(\sum_{K\in\mathcal{T}_h} \|\chi\|^2_{L^2(K)} \right)^{1/2} + \sum_{K\in\mathcal{T}_h} \|\chi\|^2_{L^2(K)} \right) \\
&\le C_M \left(h|\chi|_{H^1(\Omega,\mathcal{T}_h)} \|\chi\|_{L^2(\Omega)} + \|\chi\|^2_{L^2(\Omega)} \right).
\end{aligned}$$

The above relations, the inequality $|\chi|_{H^1(\Omega,\mathcal{T}_h)} \le |||\chi|||$ and estimates (2.116) and (2.130) imply that

$$\begin{aligned}
\sum_{K\in\mathcal{T}_h} h_K \|\chi(t)\|^2_{L^2(\partial K)} &\le C_M \left(C_{P,\mathrm{e}} C_{P,\mathrm{L}}\, h\, h^{\mu-1}\, h^{\mu} + C^2_{P,\mathrm{L}} h^{2\mu} \right) |u(t)|^2_{H^\mu(\Omega)} \\
&= C_7 h^{2\mu} |u(t)|^2_{H^\mu(\Omega)}, \quad t \in (0,T),
\end{aligned} \tag{2.147}$$

where $C_7 = C_M(C_{P,\mathrm{e}} C_{P,\mathrm{L}} + C^2_{P,\mathrm{L}})$. Furthermore, (2.146), (2.130) and (2.147) give

$$|b_h(u,\zeta) - b_h(\mathrm{P}_{hp}u,\zeta)| \le C_{b1} \left(C^2_{P,\mathrm{L}} + C_7 \right)^{1/2} h^{\mu} |||\zeta||| |u(t)|_{H^\mu(\Omega)}. \tag{2.148}$$

Furthermore, estimate (2.31) with $\bar{u}_h := \mathrm{P}_{hp}u \in S_{hp}$ and $v_h := \zeta \in S_{hp}$ gives

$$|b_h(\mathrm{P}_{hp}u,\zeta) - b_h(u_h,\zeta)| \le C_{b2}|||\zeta|||\, \|u_h - \mathrm{P}_{hp}u\|_{L^2(\Omega)} = C_{b2}|||\zeta|||\, \|\zeta\|_{L^2(\Omega)}. \tag{2.149}$$

Finally, inserting estimates (2.148) and (2.149) into (2.145), we obtain inequality (2.144) with $C_b = \max\left(C_{b2}, C_{b1}\left(C^2_{P,\mathrm{L}} + C_7 \right)^{1/2} \right)$. □ □

Now we can proceed to the *main result*, which is the optimal error estimate in the norm of the space $L^\infty(0,T;L^2(\Omega))$ of the DG method (2.25) applied on nonconforming meshes.

Theorem 2.20. *Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded convex polygonal (if $d = 2$) or* polyhedral *(if $d = 3$) domain with Lipschitz boundary $\partial\Omega$. Let Assumptions 2.5 in Section 2.3 be satisfied. Let u be the exact solution of problem* (2.1)*, where $\partial\Omega_D = \partial\Omega$ and $\partial\Omega_N = \emptyset$, satisfying the regularity condition* (2.94) *and let u_h be the approximate solution obtained by scheme* (2.25) *with the SIPG version of the diffusion terms and the constant C_W satisfying* (1.132)*. Then the error $e_h = u_h - u$ satisfies the estimate*

$$\|e_h\|_{L^\infty(0,T;L^2(\Omega))} \le C_8 h^{\mu}, \quad h \in (0, \bar{h}), \tag{2.150}$$

with a constant $C_8 > 0$ independent of h.

Proof. Let $\mathrm{P}_{hp}u$ be defined by (2.115) and let χ and ζ be as in Lemmas 2.17, 2.18 and 2.19, i.e., $\chi = u - \mathrm{P}_{hp}u$, $\zeta = \mathrm{P}_{hp}u - u_h$. Then $e_h = u_h - u = -\chi - \zeta$. Let us subtract (2.25b) from (2.28), substitute $\zeta \in S_{hp}$ for v_h, and use the relations

$$\left(\frac{\partial \zeta(t)}{\partial t}, \zeta(t) \right) = \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|\zeta(t)\|^2_{L^2(\Omega)}, \qquad A_h(u(t) - \mathrm{P}_{hp}u(t), \zeta(t)) = 0.$$

Then we get

$$\begin{aligned}
&\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|\zeta(t)\|^2_{L^2(\Omega)} + A_h(\zeta(t), \zeta(t)) \\
&\qquad = (b_h(u_h(t), \zeta(t)) - b_h(u(t), \zeta(t))) - (\partial_t \chi(t), \zeta(t)).
\end{aligned} \tag{2.151}$$

The first term on the right-hand side can be estimated by Lemma 2.19 and Young's inequality. In estimating the second term on the right-hand side we use the Cauchy and Young's inequalities and Lemma 2.18. Finally, the coercivity property (1.140) (where $C_C = 1/2$) of $\frac{1}{\varepsilon}A_h = a_h + J_h^\sigma$ gives the estimate on the left-hand side of (2.151). On the whole, after some manipulation, we get

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\,\|\zeta(t)\|^2_{L^2(\Omega)} + \varepsilon|||\zeta(t)|||^2 \qquad (2.152)\\
&\le 2\,|b_h(u_h(t),\,\zeta(t)) - b_h(u(t),\,\zeta(t))| + 2|(\partial_t\chi(t),\,\zeta(t))|\\
&\le 2C_b|||\zeta|||\,\left(h^\mu|u|_{H^\mu(\Omega)} + \|\zeta\|_{L^2(\Omega)}\right) + 2\|\partial_t\chi(t)\|_{L^2(\Omega)}\|\zeta(t)\|_{L^2(\Omega)}\\
&\le \varepsilon|||\zeta(t)|||^2 + \frac{2C_b^2}{\varepsilon}h^{2\mu}|u|^2_{H^\mu(\Omega)} + \frac{2C_b^2}{\varepsilon}\|\zeta\|^2_{L^2(\Omega)} + C^2_{P,\mathrm{L}}h^{2\mu}|\partial_t u|^2_{H^\mu(\Omega)} + \|\zeta(t)\|^2_{L^2(\Omega)}\\
&\le \varepsilon|||\zeta(t)|||^2 + C_9h^{2\mu}\left(\frac{1}{\varepsilon}|u|^2_{H^\mu(\Omega)} + |\partial_t u|^2_{H^\mu(\Omega)}\right) + C_9\left(1+\frac{1}{\varepsilon}\right)\|\zeta\|^2_{L^2(\Omega)},
\end{aligned}$$

where $C_9 = \max(2C_b^2, C^2_{P,\mathrm{L}}, 1)$. This implies that

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\|\zeta(t)\|^2_{L^2(\Omega)} \le C_9h^{2\mu}\left(\frac{1}{\varepsilon}|u|^2_{H^\mu(\Omega)} + |\partial_t u|^2_{H^\mu(\Omega)}\right) + C_9\left(1+\frac{1}{\varepsilon}\right)\|\zeta\|^2_{L^2(\Omega)}. \qquad (2.153)$$

Using (2.25c), (1.97), (2.130), we have

$$\begin{aligned}
\|\zeta(0)\|^2_{L^2(\Omega)} &= \|\mathrm{P}_{hp}u(0) - u_h(0)\|^2_{L^2(\Omega)} = \|\mathrm{P}_{hp}u(0) - \Pi_{hp}u(0)\|^2_{L^2(\Omega)} \qquad (2.154)\\
&\le 2\|\mathrm{P}_{hp}u(0) - u(0)\|^2_{L^2(\Omega)} + 2\|u(0) - \Pi_{hp}u(0)\|^2_{L^2(\Omega)}\\
&\le 2(C_A^2 + C^2_{P,\mathrm{L}})h^{2\mu}|u^0|^2_{H^\mu(\Omega)} = C_{10}h^{2\mu}|u^0|^2_{H^\mu(\Omega)},
\end{aligned}$$

where $C_{10} = 2(C_A^2 + C^2_{P,\mathrm{L}})$.

Integrating of (2.153) from 0 to $t \in [0,T]$ and (2.154) yield

$$\begin{aligned}
\|\zeta(t)\|^2_{L^2(\Omega)} &\le C_9\,h^{2\mu}\left(\frac{1}{\varepsilon}\int_0^t |u(\vartheta)|^2_{H^\mu(\Omega)}\,\mathrm{d}\vartheta + \int_0^t |\partial_t u(\vartheta)|^2_{H^\mu(\Omega)}\,\mathrm{d}\vartheta\right) \qquad (2.155)\\
&\quad + C_9\left(1+\frac{1}{\varepsilon}\right)\int_0^t \|\zeta(\vartheta)\|^2_{L^2(\Omega)}\,\mathrm{d}\vartheta + C_{10}\,h^{2\mu}|u^0|^2_{H^\mu(\Omega)}.\\
&\le C_9\left(1+\frac{1}{\varepsilon}\right)\int_0^t \|\zeta(\vartheta)\|^2_{L^2(\Omega)}\,\mathrm{d}\vartheta + C_{11}h^{2\mu}N(\varepsilon,u),
\end{aligned}$$

where $C_{11} = \max(C_9, C_{10})$ and

$$N(\varepsilon,u) = \frac{1}{\varepsilon}\int_0^t |u(\vartheta)|^2_{H^\mu(\Omega)}\,\mathrm{d}\vartheta + \int_0^t |\partial_t u(\vartheta)|^2_{H^\mu(\Omega)}\,\mathrm{d}\vartheta + |u^0|^2_{H^\mu(\Omega)}.$$

Now we apply Gronwall's Lemma 0.9, where we put

$$\begin{aligned}
&y(t) = \|\zeta(t)\|^2_{L^2(\Omega)}, && q(t) = 0,\\
&r(t) = C_9\,(1 + 1/\varepsilon)\,, && z(t) = C_{11}h^{2\mu}N(\varepsilon,u).
\end{aligned}$$

Then, after some manipulation, we obtain the estimate

$$\|\zeta(t)\|^2_{L^2(\Omega)} \le C_{11}h^{2\mu}\,N(\varepsilon,u)\,\exp\left(C_9\left(1+\frac{1}{\varepsilon}\right)t\right). \qquad (2.156)$$

Since $e_h = -\chi - \zeta$, to complete the proof, it is sufficient to combine (2.156) with the estimate (2.130) of $\|\chi(t)\|_{L^2(\Omega)}$ in Lemma 2.18. □ □

Exercise 2.21. *Prove estimates (2.155) and (2.156) in detail.*

2.6 Uniform error estimates with respect to the diffusion coefficient

In Sections 2.1–2.5, error estimates for the space DG semidiscretization were derived in the case of nonlinear convection-diffusion problems. From the presented analysis we can see that the constants in these estimates blow up exponentially if the diffusion coefficient $\varepsilon \to 0+$. This means that these estimates are not applicable, if $\varepsilon > 0$ is very small. (See also Remark 2.16.) There is question as to whether it is possible to obtain error estimates that are uniform with respect to the diffusion coefficient $\varepsilon \to 0+$ of convection-diffusion problems.

In this section we are concerned with the error analysis of the DGM of lines applied to a linear convection-diffusion equation, which also contains a reaction term, and its coefficients satisfy some special assumptions used in works analyzing numerical methods for linear convection-diffusion problems (cf. [RST08], Chapter III, or [HSS02]). As a result, we obtain error estimates, uniform with respect to the diffusion coefficient $\varepsilon \to 0+$, and valid even for $\varepsilon = 0$.

2.6.1 Continuous problem

Let $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) be a bounded polygonal (for $d = 2$) or polyhedral (for $d = 3$) domain with Lipschitz boundary $\partial\Omega$ and $T > 0$. We set $Q_T = \Omega \times (0,T)$. Let $\boldsymbol{v} : \overline{Q}_T = \overline{\Omega} \times [0,T] \to \mathbb{R}^d$ be a given *transport flow velocity*. We assume that $\partial\Omega = \partial\Omega^- \cup \partial\Omega^+$, and for all $t \in (0,T)$,

$$\begin{aligned} \boldsymbol{v}(x,t) \cdot \boldsymbol{n}(x) &< 0 \text{ on } \partial\Omega^-, \\ \boldsymbol{v}(x,t) \cdot \boldsymbol{n}(x) &\geq 0 \text{ on } \partial\Omega^+, \end{aligned} \tag{2.157}$$

where $\boldsymbol{n}(x)$ denotes the outer unit normal to the boundary of Ω. We assume that the parts $\partial\Omega^-$ and $\partial\Omega^+$ are independent of time. With respect to our former notation, we can write $\partial\Omega_D = \partial\Omega^-$ and $\partial\Omega_N = \partial\Omega^+$. The part $\partial\Omega^-$ of the boundary $\partial\Omega$ represents the inlet through which the fluid enters the domain Ω. The part of $\partial\Omega^+$, where $\boldsymbol{v} \cdot \boldsymbol{n} > 0$, represents the outlet through which the fluid leaves the domain Ω, and the part on which $\boldsymbol{v} \cdot \boldsymbol{n} = 0$ represents impermeable walls.

We consider the following linear initial-boundary value convection-diffusion-reaction problem: Find $u : Q_T \to \mathbb{R}$ such that

$$\frac{\partial u}{\partial t} + \boldsymbol{v} \cdot \nabla u - \varepsilon \Delta u + cu = g \quad \text{in } Q_T, \tag{2.158a}$$

$$u = u_D \quad \text{on } \partial\Omega^- \times (0,T), \tag{2.158b}$$

$$\varepsilon \frac{\partial u}{\partial \boldsymbol{n}} = g_N \quad \text{on } \partial\Omega^+ \times (0,T), \tag{2.158c}$$

$$u(x,0) = u^0(x), \quad x \in \Omega. \tag{2.158d}$$

In the case $\varepsilon = 0$, we put $g_N = 0$ and ignore the Neumann condition (2.158c).

Equation (2.158a) describes the transport and diffusion in a fluid of a quantity u as, for example, temperature or concentration of some material. The constant $\varepsilon \geq 0$ is the diffusion coefficient, c represents a reaction coefficient, and g defines the source of the quantity u. Such equations appear, for example, in fluid dynamics or heat and mass transfer.

We assume that the data satisfy the following conditions:

$$g \in C([0,T]; L^2(\Omega)), \tag{2.159a}$$

$$u_0 \in L^2(\Omega), \tag{2.159b}$$

$$u_D \text{ is the trace of some } u^* \in C([0,T]; H^1(\Omega)) \cap L^\infty(Q_T) \text{ on } \partial\Omega^- \times (0,T), \tag{2.159c}$$

$$\boldsymbol{v} \in C([0,T]; W^{1,\infty}(\Omega)),\ |\boldsymbol{v}| \leq C_{\boldsymbol{v}} \text{ in } \overline{\Omega} \times [0,T],\ |\nabla \boldsymbol{v}| \leq C_{\boldsymbol{v}} \text{ a.e. in } Q_T, \tag{2.159d}$$

$$c \in C([0,T]; L^\infty(\Omega)),\ |c(x,t)| \leq C_c \text{ a.e.in } Q_T, \tag{2.159e}$$

$$c - \frac{1}{2}\nabla \cdot \boldsymbol{v} \geq \gamma_0 > 0 \text{ in } Q_T \text{ with a constant } \gamma_0, \tag{2.159f}$$

$$g_N \in C([0,T]; L^2(\partial\Omega^+)), \tag{2.159g}$$

$$\varepsilon \geq 0. \tag{2.159h}$$

Assumption (2.159f) is not restrictive, because using the transformation $u = e^{\alpha t} w, \alpha = \text{const}$ substituted into (2.158) leads to the equation for w in the form

$$\frac{\partial w}{\partial t} + \boldsymbol{v} \cdot \nabla w - \varepsilon \Delta w + (c + \alpha) w = g e^{-\alpha t}.$$

Condition (2.159f) now reads $c + \alpha - \frac{1}{2}\nabla \cdot \boldsymbol{v} \geq \gamma_0 > 0$ and is satisfied if we choose $\alpha > 0$ large enough.

The weak formulation is derived in a standard way. Equation (2.158) is multiplied by any $\varphi \in V = \{\varphi \in H^1(\Omega); \varphi|_{\partial\Omega^-} = 0\}$, Green's theorem is applied and condition (2.158c) is used.

Definition 2.22. *We say that a function u is the weak solution to (2.158) if it satisfies the conditions*

$$u - u^* \in L^2(0,T;V), \quad u \in L^\infty(Q_T), \tag{2.160a}$$

$$\frac{d}{dt}\int_\Omega u\varphi \,\mathrm{d}x + \varepsilon\int_\Omega \nabla u\cdot\nabla\varphi\,\mathrm{d}x + \int_{\partial\Omega^+}(\boldsymbol{v}\cdot\boldsymbol{n})u\varphi\,\mathrm{d}S - \int_\Omega u\nabla\cdot(\varphi\boldsymbol{v})\,\mathrm{d}x \tag{2.160b}$$
$$+\int_\Omega cu\varphi\,\mathrm{d}x = \int_\Omega g\varphi\,\mathrm{d}x + \int_{\partial\Omega^+} g_N\varphi\,\mathrm{d}S$$
$$\text{for all } \varphi\in V \text{ in the sense of distributions on } (0,T),$$

$$u(0) = u^0 \text{ in } \Omega. \tag{2.160c}$$

We shall assume that the weak solution u exists and is sufficiently regular, namely,

$$\frac{\partial u}{\partial t} \in L^2(0,T;H^s(\Omega)), \tag{2.161}$$

where $s \geq 2$ is an integer. Then also $u \in C([0,T);H^s(\Omega))$ and it is possible to show that this solution u satisfies equation (2.158a) pointwise (almost everywhere). (If $\varepsilon > 0$, then with the aid of techniques from [Lio96], [Rou05] and [Rek82], it is possible to prove that there exists a unique weak solution. Moreover, it satisfies the condition $\partial u/\partial t \in L^2(Q_T)$.)

2.6.2 Discretization of the problem

Let $\mathcal{T}_h$ be a standard conforming triangulation of the closure of the domain Ω into a finite number of closed triangles ($d = 2$) or tetrahedra ($d = 3$). Hence, the mesh $\mathcal{T}_h$ satisfies assumption (MA4) in Section 1.3.2. This means that we do not consider hanging nodes (or hanging edges) in this case. Otherwise we use the same notation as in Section 1.

We assume that the conforming triangulations satisfy the shape-regularity assumption (1.19). For $K \in \mathcal{T}_h$ we set

$$\partial K^-(t) = \{x \in \partial K;\ \boldsymbol{v}(x,t)\cdot\boldsymbol{n}(x) < 0\}, \tag{2.162}$$

$$\partial K^+(t) = \{x \in \partial K;\ \boldsymbol{v}(x,t)\cdot\boldsymbol{n}(x) \geq 0\}, \tag{2.163}$$

where $\boldsymbol{n}$ denotes the outer unit normal to ∂K. Hence, $\partial K^-(t)$ and $\partial K^+(t)$ denote the inlet and outlet parts of the boundary of K, respectively. In what follows we shall not emphasize the dependence of ∂K^+ and ∂K^- on time by notation.

In order to derive error estimates that are uniform with respect to ε, we discretize the convective terms using the idea of the *upwinding* (see (2.16)). This choice allows us to avoid using Gronwall's Lemma, which causes the non-uniformity of the error estimates in Sections 2.3 and 2.5 (see Remark 2.30). Multiplying the convective term $\boldsymbol{v}\cdot\nabla u$ by any $\varphi \in H^2(\Omega,\mathcal{T}_h)$, integrating over element K and applying Green's theorem, we get

$$\begin{aligned}\int_K (\boldsymbol{v}\cdot\nabla u)\varphi\,\mathrm{d}x &= -\int_K u\,\nabla\cdot(\varphi\boldsymbol{v})\,\mathrm{d}x + \int_{\partial K}(\boldsymbol{v}\cdot\boldsymbol{n})u\varphi\,\mathrm{d}S \\ &= -\int_K u\,\nabla\cdot(\varphi\boldsymbol{v})\,\mathrm{d}x + \int_{\partial K^-}(\boldsymbol{v}\cdot\boldsymbol{n})u\varphi\,\mathrm{d}S + \int_{\partial K^+}(\boldsymbol{v}\cdot\boldsymbol{n})u\varphi\,\mathrm{d}S.\end{aligned} \tag{2.164}$$

On the inflow part of the boundary of K we use information from outside of the element K. Therefore, we write there u^- instead of u. If $x \in \partial\Omega^-$, then we set $u^-(x) := u_D(x)$. The integrals over ∂K^+, where the information "flows out" of the element, remain unchanged. We take into account that $[u] = 0$ on $\Gamma \in \mathcal{F}_h^I$ and $u|_{\partial\Omega^-}$ satisfies the Dirichlet condition (2.158b). We further rearrange the terms in (2.164) and find that

$$\begin{aligned}&\int_K (\boldsymbol{v}\cdot\nabla u)\varphi\,\mathrm{d}x \\ &\quad= -\int_K u\,\nabla\cdot(\varphi\boldsymbol{v})\,\mathrm{d}x + \int_{\partial K^-}(\boldsymbol{v}\cdot\boldsymbol{n})u^-\varphi\,\mathrm{d}S + \int_{\partial K^+}(\boldsymbol{v}\cdot\boldsymbol{n})u\varphi\,\mathrm{d}S \\ &\quad= -\int_K u\,\nabla\cdot(\varphi\boldsymbol{v})\,\mathrm{d}x + \underbrace{\int_{\partial K}(\boldsymbol{v}\cdot\boldsymbol{n})u\varphi\,\mathrm{d}S - \int_{\partial K^+\cup\partial K^-}(\boldsymbol{v}\cdot\boldsymbol{n})u\varphi\,\mathrm{d}S}_{=0} \\ &\qquad\qquad + \int_{\partial K^-}(\boldsymbol{v}\cdot\boldsymbol{n})u^-\varphi\,\mathrm{d}S + \int_{\partial K^+}(\boldsymbol{v}\cdot\boldsymbol{n})u\varphi\,\mathrm{d}S \\ &\quad= \int_K (\boldsymbol{v}\cdot\nabla u)\varphi\,\mathrm{d}x + \int_{\partial K^-}(\boldsymbol{v}\cdot\boldsymbol{n})(u^- - u)\varphi\,\mathrm{d}S \\ &\quad= \int_K (\boldsymbol{v}\cdot\nabla u)\varphi\,\mathrm{d}S - \int_{\partial K^-\setminus\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})[u]\varphi\,\mathrm{d}S - \int_{\partial K^-\cap\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})(u-u_D)\varphi\,\mathrm{d}S,\end{aligned} \tag{2.165}$$

where we set $[u] = u - u^-$ on $\partial K^- \setminus \partial\Omega$.

Remark 2.23. *Let us note that identity* (2.165) *can be derived from the relation*

$$\int_K (\boldsymbol{v}\cdot\nabla u)\varphi\,\mathrm{d}x = -\int_K u\,\nabla\cdot(\varphi\boldsymbol{v})\,\mathrm{d}x + \sum_{\Gamma\subset\partial K}\int_\Gamma H(u_\Gamma^{(L)}, u_\Gamma^{(R)}, \boldsymbol{n}_\Gamma)\varphi\,\mathrm{d}S,$$

where H is the numerical flux given (in analogy to (2.16)*) by*

$$H(u_1,u_2,\boldsymbol{n}) = \begin{cases} \boldsymbol{v}\cdot\boldsymbol{n}\ u_1, & \text{if } \boldsymbol{v}\cdot\boldsymbol{n} > 0 \\ \boldsymbol{v}\cdot\boldsymbol{n}\ u_2, & \text{if } \boldsymbol{v}\cdot\boldsymbol{n} \le 0 \end{cases} \tag{2.166}$$

and $H(u_1,u_2,\boldsymbol{n}) = \boldsymbol{v}\cdot\boldsymbol{n}\,u_D$ on $\partial K^-\cap\partial\Omega$.

Exercise 2.24. *Verify Remark 2.23.*

Now we proceed to the derivation of the discrete problem. We start from equation (2.158a) under assumption (2.161), multiply it by any $\varphi\in H^2(\Omega,\mathcal{T}_h)$, integrate over each element K, apply Green's theorem to the diffusion and convective terms and sum over all elements $K\in\mathcal{T}_h$. Then we use the identity (2.165) for convective terms, add some terms to both sides of the resulting identity or vanishing terms (similarly as in Section 1.4 in the discretization of the diffusion term) and use the boundary conditions (we recall that $\partial\Omega_D = \partial\Omega^- = \cup_{K\in\mathcal{T}_h}\partial K^-\cap\partial\Omega$). After some manipulation we find that the exact solution u satisfies the following identity for $\varphi\in H^2(\Omega,\mathcal{T}_h)$:

$$\left(\frac{\partial u(t)}{\partial t},\varphi\right) + A_h(u(t),\varphi) + b_h(u(t),\varphi) + c_h(u(t),\varphi) = \ell_h(\varphi)(t) \tag{2.167}$$
$$\text{for a.e. } t\in(0,T),$$

where the forms in (2.167) are defined in the following way:

$$(u,\varphi) = \int_\Omega u\varphi\,\mathrm{d}x, \tag{2.168}$$

$$A_h(u,\varphi) = \varepsilon a_h(u,\varphi) + \varepsilon J_h^\sigma(u,\varphi), \tag{2.169}$$

$$\begin{aligned} a_h(u,\varphi) = &\sum_{K\in\mathcal{T}_h}\int_K \nabla u\cdot\nabla\varphi\,\mathrm{d}x - \sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma \left(\langle\nabla u\rangle\cdot\boldsymbol{n}\,[\varphi] + \Theta\langle\nabla\varphi\rangle\cdot\boldsymbol{n}\,[u]\right)\mathrm{d}S \\ &- \sum_{K\in\mathcal{T}_h}\int_{\partial K^-\cap\partial\Omega}\left((\nabla u\cdot\boldsymbol{n})\varphi + \Theta(\nabla\varphi\cdot\boldsymbol{n})u\right)\mathrm{d}S, \end{aligned} \tag{2.170}$$

$$J_h^\sigma(u,\varphi) = \sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma \sigma\,[u]\,[\varphi]\ \mathrm{d}S + \sum_{K\in\mathcal{T}_h}\int_{\partial K^-\cap\partial\Omega}\sigma u\varphi\,\mathrm{d}S, \tag{2.171}$$

$$\begin{aligned} b_h(u,\varphi) = &\sum_{K\in\mathcal{T}_h}\int_K (\boldsymbol{v}\cdot\nabla u)\varphi\,\mathrm{d}x - \sum_{K\in\mathcal{T}_h}\int_{\partial K^-\setminus\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})[u]\varphi\,\mathrm{d}S \\ &- \sum_{K\in\mathcal{T}_h}\int_{\partial K^-\cap\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})u\varphi\,\mathrm{d}S, \end{aligned} \tag{2.172}$$

$$c_h(u,\varphi) = \int_\Omega cu\varphi\,\mathrm{d}x, \tag{2.173}$$

$$\begin{aligned} \ell_h(\varphi)(t) = &\int_\Omega g(t)\varphi\,\mathrm{d}x + \sum_{K\in\mathcal{T}_h}\int_{\partial K^+\cap\partial\Omega} g_N(t)\varphi\,\mathrm{d}S + \varepsilon\sum_{K\in\mathcal{T}_h}\int_{\partial K^-\cap\partial\Omega}\sigma u_D(t)\varphi\,\mathrm{d}S \\ &+ \varepsilon\Theta\sum_{K\in\mathcal{T}_h}\int_{\partial K^-\cap\partial\Omega} u_D(t)(\nabla\varphi\cdot\boldsymbol{n})\,\mathrm{d}S - \sum_{K\in\mathcal{T}_h}\int_{\partial K^-\cap\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})u_D(t)\varphi\ \mathrm{d}S. \end{aligned} \tag{2.174}$$

The weight $\sigma|_\Gamma$ is defined by (1.104), where h_Γ is given by (1.24) or (1.25) or (1.26) and satisfies (1.20). The constant $C_W>0$ from (1.104) is arbitrary for the NIPG version, and it satisfies condition (1.132) or (1.139) for the SIPG or IIPG version, respectively.

In the form representing the discretization of the diffusion term we use the nonsymmetric (NIPG) formulation for $\Theta=-1$,and the incomplete (IIPG) formulation for $\Theta=0$ or symmetric formulation (SIPG) for $\Theta=1$.

The approximate solution will be sought for each $t\in(0,T)$ in the finite dimensional space

$$S_{hp} = \left\{\varphi\in L^2(\Omega);\varphi|_K\in P_p(K)\ \ \forall K\in\mathcal{T}_h\right\}, \tag{2.175}$$

where $p\ge 1$ is an integer and $P_p(K)$ is the space of polynomials on K of degree at most p.

Definition 2.25. *The* DG approximate solution *of problem (2.158) is defined as a function u_h such that*

$$u_h \in C^1([0,T); S_{hp}), \tag{2.176a}$$

$$\left(\frac{\partial u_h(t)}{\partial t}, \varphi_h\right) + A_h(u_h(t), \varphi_h) + b_h(u_h(t), \varphi_h) + c_h(u_h(t), \varphi_h) = \ell_h(\varphi_h)(t)$$

$$\forall \varphi_h \in S_{hp} \ \forall t \in (0,T), \tag{2.176b}$$

$$(u_h(0), \varphi_h) = (u^0, \varphi_h) \qquad \forall \varphi_h \in S_{hp}. \tag{2.176c}$$

If $\varepsilon = 0$, we can also choose $p = 0$. In this case we get the finite volume method using piecewise constant approximations. Thus, the finite volume method is a special case of the DGM.

2.6.3 Error estimates

Let us consider a system $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}, \bar{h} > 0$, of *conforming* triangulations of Ω satisfying the shape-regularity assumption (1.19). By Π_{hp} we again denote the S_{hp}-interpolation defined by (1.90) with approximation properties formulated in Lemma 1.24. Thus, if $\mu = \min(p+1, s),\ s \geq 2$ and $v \in H^s(K)$, then (1.93) – (1.95) hold.

If we denote

$$\xi = u_h - \Pi_{hp} u, \qquad \eta = \Pi_{hp} u - u, \tag{2.177}$$

where u is the exact solution satisfying the regularity conditions (2.161) and u_h is the approximate solution, then the error $e_h = u_h - u = \xi + \eta$. By (1.93) – (1.95) and (2.100), for all $K \in \mathcal{T}_h$ and $h \in (0, \bar{h})$ we have

$$\|\eta\|_{L^2(K)} \leq C_A h^{\mu} |u|_{H^{\mu}(K)}, \tag{2.178}$$

$$|\eta|_{H^1(K)} \leq C_A h^{\mu-1} |u|_{H^{\mu}(K)}, \tag{2.179}$$

$$|\eta|_{H^2(K)} \leq C_A h^{\mu-2} |u|_{H^{\mu}(K)}, \tag{2.180}$$

$$\|\eta\|_{L^2(\Omega)} \leq C_A h^{\mu} |u|_{H^{\mu}(\Omega)}, \tag{2.181}$$

$$\|\partial_t \eta\|_{L^2(\Omega)} \leq C_A h^{\mu} \, |\partial_t u|_{H^{\mu}(\Omega)}, \tag{2.182}$$

almost everywhere in $(0,T)$, where $\partial_t \eta = \partial \eta / \partial t$ and $\partial_t u = \partial u / \partial t$. If $p \geq 0$ and $s \geq 1$, then (2.178), (2.179), (2.181) and (2.182) hold as well, as follows from (1.92).

In the error analysis we use the multiplicative trace inequality (1.78), the inverse inequality (1.86) and the modified variant of Gronwall's lemma 0.10. For simplicity of notation we introduce the following norm over a subset ω of either $\partial\Omega$ or ∂K:

$$\|\varphi\|_{\boldsymbol{v},\omega} = \|\sqrt{|\boldsymbol{v} \cdot \boldsymbol{n}|}\, \varphi\|_{L^2(\omega)}, \tag{2.183}$$

where $\boldsymbol{n}$ is the corresponding outer unit normal.

Now we shall prove the following property of the form b_h given by (2.172).

Lemma 2.26. *There exist positive constants $\overline{C}_b{}'$ and $\overline{C}_b$ independent of $u,\ h,\ \varepsilon$ such that*

$$|b_h(\eta, \xi)| \leq \frac{1}{4} \sum_{K\in\mathcal{T}_h} \left(\|\xi\|^2_{\boldsymbol{v},\partial K^+ \cap \partial\Omega} + \|[\xi]\|^2_{\boldsymbol{v},\partial K^- \setminus \partial\Omega} \right) + \overline{C}_b \sum_{K\in\mathcal{T}_h} \|\eta\|_{L^2(K)} \|\xi\|_{L^2(K)} + R_2(\eta), \tag{2.184}$$

where

$$R_2(\eta) = \overline{C}_b{}' \sum_{K\in\mathcal{T}_h} \left(\|\eta\|_{L^2(K)} |\eta|_{H^1(K)} + h_K^{-1} \|\eta\|^2_{L^2(K)} \right), \tag{2.185}$$

$$\overline{C}_b = C_{\boldsymbol{v}}(1 + C_A C_I), \quad \overline{C}_b{}' = C_{\boldsymbol{v}} C_M \tag{2.186}$$

and $C_{\boldsymbol{v}}$ is the constant in assumption (2.159d).

Proof. Using (2.172) and Green's theorem, we find that

$$
\begin{aligned}
b_h(\eta,\xi) = \sum_{K\in\mathcal{T}_h} \Bigg(& \int_K (\boldsymbol{v}\cdot\nabla\eta)\xi \,\mathrm{d}x \\
& - \int_{\partial K^-\cap\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\xi\eta \,\mathrm{d}S - \int_{\partial K^-\setminus\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\xi(\eta-\eta^-) \,\mathrm{d}S \Bigg) \\
= \sum_{K\in\mathcal{T}_h} \Bigg(& \int_{\partial K} (\boldsymbol{v}\cdot\boldsymbol{n})\xi\eta \,\mathrm{d}S - \int_K \eta(\boldsymbol{v}\cdot\nabla\xi) \,\mathrm{d}x - \int_K \eta\xi\,\nabla\cdot\boldsymbol{v} \,\mathrm{d}x \\
& - \int_{\partial K^-\cap\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\xi\eta \,\mathrm{d}S - \int_{\partial K^-\setminus\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\xi(\eta-\eta^-) \,\mathrm{d}S \Bigg),
\end{aligned}
\tag{2.187}
$$

where the superscript $^-$ denotes the values on ∂K from outside the element K. Hence,

$$
\begin{aligned}
|b_h(\eta,\xi)| \le & \left| \sum_{K\in\mathcal{T}_h} \int_K \eta(\boldsymbol{v}\cdot\nabla\xi) \,\mathrm{d}x \right| + \left| \sum_{K\in\mathcal{T}_h} \int_K \eta\xi\,\nabla\cdot\boldsymbol{v} \,\mathrm{d}x \right| \\
& + \left| \sum_{K\in\mathcal{T}_h} \Bigg(\int_{\partial K} (\boldsymbol{v}\cdot\boldsymbol{n})\xi\eta \,\mathrm{d}S - \int_{\partial K^-\cap\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\xi\eta \,\mathrm{d}S \right. \\
& \left. - \int_{\partial K^-\setminus\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\xi(\eta-\eta^-) \,\mathrm{d}S \Bigg) \right|.
\end{aligned}
\tag{2.188}
$$

The second term on the right-hand side of (2.188) is estimated easily with the aid of the Cauchy inequality and assumption (2.159d):

$$
\left| \sum_{K\in\mathcal{T}_h} \int_K \eta\xi\,\nabla\cdot\boldsymbol{v} \,\mathrm{d}x \right| \le C_{\boldsymbol{v}} \sum_{K\in\mathcal{T}_h} \|\eta\|_{L^2(K)} \|\xi\|_{L^2(K)}.
\tag{2.189}
$$

Since

$$
\sum_{K\in\mathcal{T}_h} \int_{\partial K^+\setminus\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\xi\eta \,\mathrm{d}S = - \sum_{K\in\mathcal{T}_h} \int_{\partial K^-\setminus\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\xi^-\eta^- \,\mathrm{d}S
\tag{2.190}
$$

and $\boldsymbol{v}\cdot\boldsymbol{n} \ge 0$ on ∂K^+, with the aid of Young's inequality, the set decomposition

$$
\partial K = \partial K^+ \cup (\partial K^- \cap \partial\Omega) \cup (\partial K^- \setminus \partial\Omega)
$$

and notation (2.183), the third term on the right-hand side of (2.188) can be rewritten and then estimated in the following way:

$$
\begin{aligned}
& \left| \sum_{K\in\mathcal{T}_h} \left(\int_{\partial K^+} (\boldsymbol{v}\cdot\boldsymbol{n})\xi\eta \,\mathrm{d}S + \int_{\partial K^-\setminus\partial\Omega} \left\{ (\boldsymbol{v}\cdot\boldsymbol{n})\xi\eta - (\boldsymbol{v}\cdot\boldsymbol{n})\xi(\eta-\eta^-) \right\} \mathrm{d}S \right) \right| \\
& = \left| \sum_{K\in\mathcal{T}_h} \left(\int_{\partial K^+\cap\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\xi\eta \,\mathrm{d}S + \int_{\partial K^+\setminus\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\xi\eta \,\mathrm{d}S + \int_{\partial K^-\setminus\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\eta^-\xi \,\mathrm{d}S \right) \right| \\
& = \left| \sum_{K\in\mathcal{T}_h} \left(\int_{\partial K^+\cap\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\xi\eta \,\mathrm{d}S + \int_{\partial K^-\setminus\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\eta^-(\xi-\xi^-) \,\mathrm{d}S \right) \right| \\
& \le \frac{1}{4} \sum_{K\in\mathcal{T}_h} \left(\int_{\partial K^+\cap\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\xi^2 \,\mathrm{d}S + \int_{\partial K^-\setminus\partial\Omega} |\boldsymbol{v}\cdot\boldsymbol{n}|[\xi]^2 \,\mathrm{d}S \right) \\
& \quad + \sum_{K\in\mathcal{T}_h} \left(\int_{\partial K^+\cap\partial\Omega} (\boldsymbol{v}\cdot\boldsymbol{n})\eta^2 \,\mathrm{d}S + \int_{\partial K^-\setminus\partial\Omega} |\boldsymbol{v}\cdot\boldsymbol{n}|(\eta^-)^2 \,\mathrm{d}S \right) \\
& \le \frac{1}{4} \sum_{K\in\mathcal{T}_h} \left(\|\xi\|^2_{\boldsymbol{v},\partial K^+\cap\partial\Omega} + \|[\xi]\|^2_{\boldsymbol{v},\partial K^-\setminus\partial\Omega} \right) \\
& \quad + \sum_{K\in\mathcal{T}_h} \left(\|\eta\|^2_{\boldsymbol{v},\partial K^+\cap\partial\Omega} + \|\eta^-\|^2_{\boldsymbol{v},\partial K^-\setminus\partial\Omega} \right).
\end{aligned}
\tag{2.191}
$$

Using the multiplicative trace inequality, the boundedness of $\boldsymbol{v}$ and estimates (2.178) and (2.179), we get

$$\sum_{K\in\mathcal{T}_h}\left(\|\eta\|^2_{\boldsymbol{v},\partial K^+\cap\partial\Omega}+\|\eta^-\|^2_{\boldsymbol{v},\partial K^-\setminus\partial\Omega}\right) \tag{2.192}$$
$$\le C_{\boldsymbol{v}}\sum_{K\in\mathcal{T}_h}\|\eta\|^2_{L^2(\partial K)}\le C_{\boldsymbol{v}}C_M\sum_{K\in\mathcal{T}_h}\left(\|\eta\|_{L^2(K)}|\eta|_{H^1(K)}+h_K^{-1}\|\eta\|^2_{L^2(K)}\right).$$

By virtue of the definition (2.177) of η and (1.89) – (1.90), the first term on the right-hand side of (2.188) vanishes if the vector $\boldsymbol{v}$ is piecewise linear, because $\boldsymbol{v}\cdot\nabla\xi|_K\in P_p(K)$ in this case. If this is not the case, we have to proceed in a more sophisticated way. For every $t\in[0,T)$ we introduce a function $\Pi_{h1}\boldsymbol{v}(t)$ which is a piecewise linear $L^2(\Omega)$-projection of $\boldsymbol{v}(t)$ on the space S_{hp}. Under assumption (2.159d), by (1.96),

$$\|\boldsymbol{v}-\Pi_{h1}\boldsymbol{v}\|_{L^\infty(K)}\le C_A h_K|\boldsymbol{v}|_{W^{1,\infty}(K)},\quad K\in\mathcal{T}_h,\ h\in(0,\bar h). \tag{2.193}$$

The first term in (2.188) is then estimated with the aid of (1.89), (1.86), (2.193), the Cauchy inequality and assumption (2.159d) in the following way:

$$\left|\sum_{K\in\mathcal{T}_h}\int_K\eta(\boldsymbol{v}\cdot\nabla\xi)\,\mathrm{d}x\right| \tag{2.194}$$
$$\le\sum_{K\in\mathcal{T}_h}\left|\int_K\eta(\Pi_{h1}\boldsymbol{v}\cdot\nabla\xi)\,\mathrm{d}x\right|+\sum_{K\in\mathcal{T}_h}\left|\int_K\eta((\boldsymbol{v}-\Pi_{h1}\boldsymbol{v})\cdot\nabla\xi)\,\mathrm{d}x\right|$$
$$=\sum_{K\in\mathcal{T}_h}\left|\int_K\eta((\boldsymbol{v}-\Pi_{h1}\boldsymbol{v})\cdot\nabla\xi)\,\mathrm{d}x\right|\le\sum_{K\in\mathcal{T}_h}\|\boldsymbol{v}-\Pi_{h1}\boldsymbol{v}\|_{L^\infty(K)}\|\eta\|_{L^2(K)}|\xi|_{H^1(K)}$$
$$\le\sum_{K\in\mathcal{T}_h}C_A h_K|\boldsymbol{v}|_{W^{1,\infty}(K)}\ \|\eta\|_{L^2(K)}\ C_I h_K^{-1}\|\xi\|_{L^2(K)}$$
$$\le C_{\boldsymbol{v}}C_AC_I\sum_{K\in\mathcal{T}_h}\|\eta\|_{L^2(K)}\|\xi\|_{L^2(K)}.$$

Using (2.189), (2.191) and (2.194) in (2.188), we obtain (2.184) with constants defined in (2.186). This finishes the proof of Lemma 2.26. □ □

Further, by (2.80) and Young's inequality,

$$|A_h(\eta,\xi)|\le\varepsilon\widetilde{C}_B R_a(\eta)|||\xi|||\le\frac{\varepsilon}{4}|||\xi|||^2+\varepsilon\widetilde{C}_B^2R_a(\eta)^2=\frac{\varepsilon}{4}|||\xi|||^2+\varepsilon R_1(\eta), \tag{2.195}$$

where

$$R_1(\eta)=\widetilde{C}_B^2\sum_{K\in\mathcal{T}_h}\left(|\eta|^2_{H^1(K)}+h_K^2|\eta|^2_{H^2(K)}+h_K^{-2}\|\eta\|^2_{L^2(K)}\right). \tag{2.196}$$

Finally, the Cauchy inequality gives

$$|c_h(\eta,\xi)|\le C_c\|\eta\|_{L^2(\Omega)}\|\xi\|_{L^2(\Omega)}, \tag{2.197}$$
$$|(\partial_t\eta,\xi)|\le\|\partial_t\eta\|_{L^2(\Omega)}\|\xi\|_{L^2(\Omega)}. \tag{2.198}$$

Now we can formulate the *abstract error estimate*.

Theorem 2.27. *Let us assume that $\{\mathcal{T}_h\}_{h\in(0,\bar h)}$ is a system of conforming shape-regular triangulations (cf.* (1.19)*) of the domain Ω and let assumptions* (2.159) *be satisfied. Let us assume that the constant $C_W>0$ satisfies the conditions in Corollary 1.41 for NIPG, SIPG and IIPG versions of the diffusion form. Let the exact solution u of problem* (2.158) *be regular in the sense of* (2.161) *and let u_h be the approximate solution obtained by the method of lines* (2.176)*. Then the error $e_h=u_h-u$*

satisfies the estimate

$$\begin{aligned}
&\Big(\|e_h(t)\|^2_{L^2(\Omega)} + \frac{\varepsilon}{2}\int_0^t |||e_h(\vartheta)|||^2 \mathrm{d}\vartheta + 2\gamma_0 \int_0^t \|e_h(\vartheta)\|^2_{L^2(\Omega)}\,\mathrm{d}\vartheta \\
&\quad + \frac{1}{2}\int_0^t \sum_{K\in\mathcal{T}_h} \Big(\|e_h(\vartheta)\|^2_{\boldsymbol{v}(\vartheta),\partial K\cap\partial\Omega} + \|[e_h(\vartheta)]\|^2_{\boldsymbol{v}(\vartheta),\partial K^-(\vartheta)\setminus\partial\Omega}\Big)\,\mathrm{d}\vartheta\Big)^{1/2} \\
&\le \sqrt{2}\Big(\int_0^t (\varepsilon R_1(\eta(\vartheta)) + R_2(\eta(\vartheta))\mathrm{d}\vartheta\Big)^{1/2} \\
&\quad + 2\sqrt{2}\int_0^t \Big(\|\eta(\vartheta)\|_{L^2(\Omega)}(C_c + \overline{C}_b) + \Big\|\partial_t\eta(\vartheta)\Big\|_{L^2(\Omega)}\Big)\,\mathrm{d}\vartheta \\
&\quad + \sqrt{2}\left(\|\eta(\vartheta)\|^2_{L^2(\Omega)} + \int_0^t (\frac{\varepsilon}{2}|||\eta(\vartheta)|||^2 + 2\gamma_0\|\eta(\vartheta)\|^2_{L^2(\Omega)} + R_2(\eta(\vartheta)))\,\mathrm{d}\vartheta\right)^{1/2}, \\
&\quad t\in[0,T], \quad h\in(0,\bar{h}),
\end{aligned} \tag{2.199}$$

where R_1 and R_2 are given by (2.196) *and* (2.185)*, respectively.*

Proof. The proof will be carried out in several steps.
We subtract equation (2.167) from (2.176b) and for arbitrary but fixed $t\in[0,T]$, we put $\varphi := \xi(t)$ to get

$$\begin{aligned}
&(\partial_t\xi,\xi) + A_h(\xi,\xi) + b_h(\xi,\xi) + c_h(\xi,\xi) \\
&\qquad = -(\partial_t\eta,\xi) - A_h(\eta,\xi) - b_h(\eta,\xi) - c_h(\eta,\xi).
\end{aligned} \tag{2.200}$$

Obviously,

$$(\partial_t\xi,\xi) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\Omega)}, \tag{2.201}$$

and, in view of Corollary 1.41,

$$A_h(\xi,\xi) \ge \frac{\varepsilon}{2}|||\xi|||^2. \tag{2.202}$$

Further, let us rearrange the terms in the form b_h. We have

$$\begin{aligned}
b_h(\xi,\xi) &= \sum_{K\in\mathcal{T}_h}\left(\int_K (\boldsymbol{v}\cdot\nabla\xi)\xi\,\mathrm{d}x - \int_{\partial K^-\cap\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})\xi^2\,\mathrm{d}S - \int_{\partial K^-\setminus\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})[\xi]\xi\,\mathrm{d}S\right) \\
&= \sum_{K\in\mathcal{T}_h}\left(-\frac{1}{2}\int_K(\nabla\cdot\boldsymbol{v})\,\xi^2\,\mathrm{d}x + \frac{1}{2}\int_{\partial K}(\boldsymbol{v}\cdot\boldsymbol{n})\xi^2\,\mathrm{d}S - \int_{\partial K^-\cap\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})\xi^2\,\mathrm{d}S\right. \\
&\qquad \left. - \int_{\partial K^-\setminus\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})\xi(\xi-\xi^-)\,\mathrm{d}S\right).
\end{aligned}$$

Using the decomposition $\partial K = \partial K^-\cup\partial K^+$, we get

$$\begin{aligned}
b_h(\xi,\xi) = \sum_{K\in\mathcal{T}_h}\frac{1}{2}\Big(&-\int_K \xi^2\,\nabla\cdot\boldsymbol{v}\,\mathrm{d}x - \int_{\partial K^-\cap\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})\xi^2\,\mathrm{d}S \\
&- \int_{\partial K^-\setminus\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})(\xi^2 - 2\xi\xi^-)\,\mathrm{d}S \\
&+ \int_{\partial K^+\cap\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})\xi^2\,\mathrm{d}S + \int_{\partial K^+\setminus\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})\xi^2\,\mathrm{d}S\Big).
\end{aligned}$$

Now, by virtue of the relation

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K^+\setminus\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})\xi^2\,\mathrm{d}S = -\sum_{K\in\mathcal{T}_h}\int_{\partial K^-\setminus\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})(\xi^-)^2\,\mathrm{d}S, \tag{2.203}$$

definition (2.162) and (2.183), we find that

$$
\begin{aligned}
b_h(\xi,\xi) &= \frac{1}{2}\sum_{K\in\mathcal{T}_h}\Big(-\int_K \xi^2\,\nabla\cdot\boldsymbol{v}\;\mathrm{d}x - \int_{\partial K^-\cap\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})\xi^2\,\mathrm{d}S \\
&\quad - \int_{\partial K^-\setminus\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})(\xi^2 - 2\xi\xi^- + (\xi^-)^2)\,\mathrm{d}S + \int_{\partial K^+\cap\partial\Omega}(\boldsymbol{v}\cdot\boldsymbol{n})\xi^2\,\mathrm{d}S\Big) \\
&= \frac{1}{2}\sum_{K\in\mathcal{T}_h}\Big(-\int_K \xi^2\,\nabla\cdot\boldsymbol{v}\;\mathrm{d}x + \int_{\partial K^-\cap\partial\Omega}|(\boldsymbol{v}\cdot\boldsymbol{n})|\xi^2\,\mathrm{d}S \\
&\quad + \int_{\partial K^-\setminus\partial\Omega}|(\boldsymbol{v}\cdot\boldsymbol{n})|(\xi^2 - 2\xi\xi^- + (\xi^-)^2)\,\mathrm{d}S + \int_{\partial K^+\cap\partial\Omega}|(\boldsymbol{v}\cdot\boldsymbol{n})|\xi^2\,\mathrm{d}S\Big) \\
&= \frac{1}{2}\sum_{K\in\mathcal{T}_h}\Big(\|\xi\|^2_{\boldsymbol{v},\partial K^-\cap\partial\Omega} + \|[\xi]\|^2_{\boldsymbol{v},\partial K^-\setminus\partial\Omega} + \|\xi\|^2_{\boldsymbol{v},\partial K^+\cap\partial\Omega}\Big) - \frac{1}{2}\int_\Omega \xi^2\,\nabla\cdot\boldsymbol{v}\,\mathrm{d}x.
\end{aligned}
\tag{2.204}
$$

Finally,

$$
c_h(\xi,\xi) = \int_\Omega c\xi^2\,\mathrm{d}x. \tag{2.205}
$$

On the basis of (2.200) – (2.202), (2.204) and (2.205) we obtain the inequality

$$
\begin{aligned}
\frac{1}{2}\,\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\Omega)} + \frac{\varepsilon}{2}|||\xi|||^2 + \int_\Omega (c - \frac{1}{2}\nabla\cdot\boldsymbol{v})\xi^2\,\mathrm{d}x \\
+ \frac{1}{2}\sum_{K\in\mathcal{T}_h}\Big(\|\xi\|^2_{\boldsymbol{v},\partial K\cap\partial\Omega} + \|[\xi]\|^2_{\boldsymbol{v},\partial K^-\setminus\partial\Omega}\Big) \\
\le |(\partial_t\eta,\xi)| + |A_h(\eta,\xi)| + |b_h(\eta,\xi)| + |c_h(\eta,\xi)|.
\end{aligned}
\tag{2.206}
$$

Now, assumptions (2.159e), (2.159f) and inequalities (2.184), (2.195), (2.197) and (2.198) imply that

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\Omega)} + \frac{\varepsilon}{2}|||\xi|||^2 + 2\gamma_0\|\xi\|^2_{L^2(\Omega)} + \frac{1}{2}\sum_{K\in\mathcal{T}_h}\Big(\|\xi\|^2_{\boldsymbol{v},\partial K\cap\partial\Omega} + \|[\xi]\|^2_{\boldsymbol{v},\partial K^-\setminus\partial\Omega}\Big) \\
&\le 2\|\xi\|_{L^2(\Omega)}\Big(\|\eta\|_{L^2(\Omega)}(C_c + \overline{C}_b) + \|\partial_t\eta\|_{L^2(\Omega)}\Big) + 2\varepsilon R_1(\eta) + 2R_2(\eta).
\end{aligned}
\tag{2.207}
$$

Integrating (2.207) over $(0,t)$ and using the relation $\xi(0)=0$, we get

$$
\begin{aligned}
&\|\xi(t)\|^2_{L^2(\Omega)} + \int_0^t \frac{\varepsilon}{2}|||\xi(\vartheta)|||^2\,\mathrm{d}\vartheta + 2\gamma_0\|\xi\|^2_{L^2(Q_t)} \\
&\quad + \frac{1}{2}\int_0^t\sum_{K\in\mathcal{T}_h}\Big(\|\xi(\vartheta)\|^2_{\boldsymbol{v}(\vartheta),\partial K\cap\partial\Omega} + \|[\xi(\vartheta)]\|^2_{\boldsymbol{v}(\vartheta),\partial K^-(\vartheta)\setminus\partial\Omega}\Big)\,\mathrm{d}\vartheta \\
&\le 2\int_0^t \|\xi(\vartheta)\|_{L^2(\Omega)}\Big(\|\eta(\vartheta)\|_{L^2(\Omega)}(C_c + \overline{C}_b) + \|\partial_t\eta(\vartheta)\|_{L^2(\Omega)}\Big)\,\mathrm{d}\vartheta \\
&\quad + 2\int_0^t (\varepsilon R_1(\eta(\vartheta)) + R_2(\eta(\vartheta)))\mathrm{d}\vartheta.
\end{aligned}
\tag{2.208}
$$

As the last step we make use of the modified Gronwall's Lemma 0.10 with

$$
\begin{aligned}
\chi(t) &= \|\xi(t)\|_{L^2(\Omega)}, \\
R(t) &= \frac{\varepsilon}{2}\int_0^t |||\xi(\vartheta)|||^2\,\mathrm{d}\vartheta + 2\gamma_0\|\xi\|^2_{L^2(Q_t)} \\
&\quad + \frac{1}{2}\int_0^t\sum_{K\in\mathcal{T}_h}\Big(\|\xi(\vartheta)\|^2_{\boldsymbol{v}(\vartheta),\partial K\cap\partial\Omega} + \|[\xi(\vartheta)]\|^2_{\boldsymbol{v}(\vartheta),\partial K^-(\vartheta)\setminus\partial\Omega}\Big)\,\mathrm{d}\vartheta, \\
A(t) &= 2\int_0^t (\varepsilon R_1(\eta(\vartheta)) + R_2(\eta(\vartheta)))\,\mathrm{d}\vartheta, \\
B(t) &= \|\eta(t)\|_{L^2(\Omega)}(C_c + \overline{C}_b) + \|\partial_t\eta(t)\|_{L^2(\Omega)}.
\end{aligned}
\tag{2.209}
$$

For simplicity, we denote the left-hand side of inequality (2.208) as $L(\xi,t)$. Then for $t\in[0,T]$ we get

$$\sqrt{L(\xi,t)}\le\left(2\int_0^t(\varepsilon R_1(\eta(t))+R_2(\eta(t)))\,\mathrm{d}\vartheta\right)^{1/2}+\int_0^t\left(\|\eta(t)\|_{L^2(\Omega)}(C_c+\overline{C}_b)+\|\partial_t\eta(t)\|_{L^2(\Omega)}\right)\mathrm{d}\vartheta. \tag{2.210}$$

To obtain the estimate of $e_h=u_h-u=\xi+\eta$, we note that

$$\begin{aligned}\|e_h\|^2_{L^2(\Omega)}&\le 2\left(\|\xi\|^2_{L^2(\Omega)}+\|\eta\|^2_{L^2(\Omega)}\right),\\ |||e_h|||^2&\le 2\left(|||\xi|||^2+|||\eta|||^2\right),\\ \|e_h\|^2_{\boldsymbol{v},\partial K\cap\partial\Omega}&\le 2\left(\|\xi\|^2_{\boldsymbol{v},\partial K\cap\partial\Omega}+\|\eta\|^2_{\boldsymbol{v},\partial K\cap\partial\Omega}\right),\\ \|[e_h]\|^2_{\boldsymbol{v},\partial K^-\setminus\partial\Omega}&\le 2\left(\|[\xi]\|^2_{\boldsymbol{v},\partial K^-\setminus\partial\Omega}+\|[\eta]\|^2_{\boldsymbol{v},\partial K^-\setminus\partial\Omega}\right).\end{aligned}$$

We can find that

$$\sqrt{L(e_h,t)}\le\sqrt{2}\sqrt{L(\xi,t)+L(\eta,t)}\le\sqrt{2}\left(\sqrt{L(\xi,t)}+\sqrt{L(\eta,t)}\right). \tag{2.211}$$

Similarly as in the proof of (2.192), under the notation (2.185) and (2.186), we find that

$$\sum_{K\in\mathcal{T}_h}\left(\|\eta\|^2_{\boldsymbol{v},\partial K\cap\partial\Omega}+\|[\eta]\|^2_{\boldsymbol{v},\partial K^-\setminus\partial\Omega}\right)\le 2R_2(\eta). \tag{2.212}$$

Now, from (2.210), (2.211) and (2.212) it follows that

$$\begin{aligned}\sqrt{L(e_h,t)}\le{}&2\Big(\int_0^t(\varepsilon R_1(\eta(t))+R_2(\eta(t)))\,\mathrm{d}\vartheta\Big)^{1/2}\\ &+\sqrt{2}\int_0^t\left(\|\eta(t)\|_{L^2(\Omega)}(C_c+\overline{C}_b)+\left\|\partial_t\eta(t)\right\|_{L^2(\Omega)}\right)\mathrm{d}\vartheta\\ &+\sqrt{2}\left(\|\eta(t)\|^2_{L^2(\Omega)}+\int_0^t\left(\frac{\varepsilon}{2}|||\eta(\vartheta)|||^2+2\gamma_0\|\eta\|^2_{L^2(\Omega)}+R_2(\eta(\vartheta))\right)\mathrm{d}\vartheta\right)^{1/2},\end{aligned} \tag{2.213}$$

which is the desired result (2.199). □ □

Now, we formulate the main result of this section, representing the error estimate in terms of the mesh-size h.

Theorem 2.28. *Let us assume that $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ is a system of conforming shape-regular triangulations (cf.* (1.19)*) of the domain Ω and let assumption* (2.159) *be satisfied. Let us assume that the constant $C_W>0$ satisfies the conditions from Corollary 1.41 for NIPG, SIPG and IIPG versions of the diffusion form. Let the exact solution u of problem* (2.158) *be regular in the sense of* (2.161) *and let u_h be the approximate solution obtained by the method of lines* (2.176)*. Then the error $e_h=u_h-u$ satisfies the estimate*

$$\begin{aligned}&\max_{t\in(0,T)}\|e_h(t)\|_{L^2(\Omega)}+\left(\frac{\varepsilon}{2}\int_0^T|||e_h(\vartheta)|||^2\,\mathrm{d}\vartheta+2\gamma_0\|e_h\|^2_{L^2(Q_T)}\right)^{1/2}\\ &\qquad+\left(\frac{1}{2}\sum_{K\in\mathcal{T}_h}\int_0^T\left(\|e_h(t)\|^2_{\boldsymbol{v}(t),\partial K\cap\partial\Omega}+\|[e_h(t)]\|^2_{\boldsymbol{v}(t),\partial K^-(t)\setminus\partial\Omega}\right)\mathrm{d}t\right)^{1/2}\\ &\qquad\le\widetilde{C}h^{\mu-1}(\sqrt{\varepsilon}+\sqrt{h}),\end{aligned} \tag{2.214}$$

where $\widetilde{C}>0$ is a constant independent of ε and h.

Proof. Estimate (2.214) will be derived from the abstract error estimate (2.199) and estimates (2.178) – (2.182) of the term η.

By (2.196), (2.185), (2.178) - (2.180), the inequality $h_K\le h$ and the relation

$$\sum_{K\in\mathcal{T}_h}|u|^2_{H^\mu(K)}=|u|^2_{H^\mu(\Omega)}, \tag{2.215}$$

we have

$$R_1(\eta) \le 3\widetilde{C}_B^2 C_A^2 h^{2(\mu-1)} |u|^2_{H^\mu(\Omega)}, \tag{2.216}$$

$$R_2(\eta) \le 2\overline{C}_b{}' C_A^2 h^{2\mu-1} |u|^2_{H^\mu(\Omega)}. \tag{2.217}$$

From (2.104), we have

$$|||\eta|||^2 \le C_A^2 \left(1 + \frac{4C_W C_M}{C_T}\right) h^{2(\mu-1)} |u|^2_{H^\mu(\Omega)}. \tag{2.218}$$

Now, estimates (2.178), (2.182), (2.199), (2.216) – (2.218) and the inequality $\sqrt{a} + \sqrt{b} + \sqrt{c} \le \sqrt{3}(a+b+c)^{1/2}$ valid for $a, b, c \ge 0$, imply that

$$\begin{aligned}
&\max_{t\in(0,T)} \|e_h(t)\|_{L^2(\Omega)} + \left(\frac{\varepsilon}{2}\int_0^T |||e_h(\vartheta)|||^2 \,\mathrm{d}\vartheta + 2\gamma_0 \|e_h\|^2_{L^2(Q_T)}\right)^{1/2} \\
&\quad + \left(\frac{1}{2}\sum_{K\in\mathcal{T}_h}\int_0^T \left(\|e_h(t)\|^2_{\boldsymbol{v}(t),\partial K\cap\partial\Omega} + \|[e_h(t)]\|^2_{\boldsymbol{v}(t),\partial K^-(t)\setminus\partial\Omega}\right) \mathrm{d}t\right)^{1/2} \\
&\le \left\{\sqrt{6}\left(\left(3\varepsilon\widetilde{C}_B^2 C_A^2 h^{2(\mu-1)} + 2\overline{C}_b{}' C_A^2 h^{2\mu-1}\right)\int_0^T |u(\vartheta)|^2_{H^\mu(\Omega} \,\mathrm{d}\vartheta\right)^{1/2}\right. \\
&\quad + 2\sqrt{6}\left(C_A(C_c + \overline{C}_b)h^\mu \int_0^T |u(\vartheta)|_{H^\mu(\Omega} \,\mathrm{d}\vartheta + C_A h^\mu \int_0^T |\partial_t u(\vartheta)|_{H^\mu(\Omega)} \,\mathrm{d}\vartheta\right) \\
&\quad + \sqrt{6}\Big(C_A^2 h^{2\mu} \max_{t\in[0,T]} |u(t)|^2_{H^\mu(\Omega)} \\
&\quad \left. + C_A^2\Big(2\gamma_0 h^{2\mu} + \frac{\varepsilon}{2}\Big(1 + \frac{4C_W C_M}{C_T}\Big)h^{2(\mu-1)} + 2\overline{C}_b{}' h^{2\mu-1}\Big)\int_0^T |u(\vartheta)|^2_{H^\mu(\Omega)} \,\mathrm{d}\vartheta\Big)^{1/2}\right\}.
\end{aligned} \tag{2.219}$$

The above inequality and the inequality $h < \bar{h}$ already imply estimate (2.214) with a constant $\widetilde{C}$ depending on the constants $\widetilde{C}_B, C_A, \overline{C}_b{}', C_c, \overline{C}_b, \bar{h}, \gamma_0, C_W, C_M, C_T$ and the seminorms

$$|u|_{L^2(0,T;H^\mu(\Omega))},\ |u|_{L^1(0,T;H^\mu(\Omega))},\ |u|_{C([0,T];H^\mu(\Omega))},\ |\partial_t u|_{L^1(0,T;H^\mu(\Omega))}. \qquad \square$$

□

Exercise 2.29. *(i) Prove estimate (2.212) in detail.*

(ii) Verify relations (2.211).

(iii) Express the constant $\widetilde{C}$ from the error estimate (2.214) *in terms of the constants $\widetilde{C}_B$, $C_A, \ldots$, and the norms of u and $\partial_t u$.*

(iv) Prove relations (2.190) and (2.203) in detail.

Remark 2.30. *Let us omit the integrals over $\partial K^- \cap \partial\Omega$ and $\partial K^- \setminus \partial\Omega$ in the form b_h and the corresponding terms on the right-hand side in the definition of the approximate solution u_h (which means that we cancel upwinding). Proceeding in the same way as before, we obtain the estimate of the type*

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\Omega)} + \varepsilon|||\xi|||^2 + 2\int_\Omega (c - \frac{1}{2}\nabla\cdot\boldsymbol{v})\xi^2 \,\mathrm{d}x + \sum_{K\in\mathcal{T}_h}\int_{\partial K}(\boldsymbol{v}\cdot\boldsymbol{n})\xi^2 \,\mathrm{d}S \\
&\qquad \le C\varepsilon h^{2(\mu-1)} + Ch^{2\mu} + \|\xi\|^2_{L^2(\Omega)}.
\end{aligned} \tag{2.220}$$

We can see that it is difficult to handle the terms $\int_\Gamma (\boldsymbol{v}\cdot\boldsymbol{n})\xi^2 \,\mathrm{d}S$ on the left-hand side, as $\boldsymbol{v}\cdot\boldsymbol{n}$ may be both positive and negative. We can make some rearrangements, but then it is necessary to use the standard Gronwall's Lemma 0.9 and we obtain a term like $\exp(cT/\varepsilon)$ on the right-hand side of the final estimate, which is not desirable, especially for small ε. The use of upwinding is therefore important for obtaining the error estimate uniform with respect to the diffusion coefficient ε. Similar result is valid even on an infinite time interval $[0, +\infty)$ as was shown in [FŠ04].

Exercise 2.31. *Prove estimate (2.220) and the error estimate following from (2.220).*

2.7 Numerical examples

In Chapter 1 we presented numerical experiments which demonstrate the high order of convergence of the discontinuous Galerkin method (DGM). However, similar results can be obtained for the standard *conforming finite element method* (FEM) (e.g., [Cia79]). Moreover, in comparison with conforming FEM, DGM requires more degrees of freedom for obtaining the same level of computational error. On the other hand, the numerical solutions obtained by the conforming FEM and DGM are completely different in the case of convection-diffusion problems, particularly for dominating convection.

Let us consider a simple stationary linear convection-diffusion boundary value problem to find such a function u that

$$\begin{aligned} \frac{\partial u}{\partial x_1} - \varepsilon \Delta u &= 1 \quad \text{in} \quad \Omega = (0,1) \times (0,1), \\ u &= 0 \quad \text{on} \quad \partial\Omega, \end{aligned} \tag{2.221}$$

where $\varepsilon > 0$ is a diffusion coefficient. The exact solution possesses an exponential boundary layer along $x_1 = 1$ and two parabolic boundary layers along $x_2 = 0$ and $x_2 = 1$ (cf. [RST96]). In the interior grid points the solution $u(x_1, x_2) \approx x_1$.

We solved this problem with the aid of the conforming FEM and the IIPG variant of DGM on a uniform triangular grid with spacing $h = 1/16$ with the aid of piecewise linear approximation. Figures 2.1 and 2.2 show the approximate solutions for $\varepsilon = 10^{-1},\ 10^{-2},\ 10^{-3},\ 10^{-4},\ 10^{-5}$ and 10^{-6} obtained by FEM and DGM, respectively.

We can see that the conforming finite element solutions suffer from spurious oscillations whose amplitude increases with decreasing diffusion coefficient. On the other hand, for $\varepsilon = 10^{-1},\ 10^{-2}$ and 10^{-3} the discontinuous Galerkin solution contains spurious overshoots and undershoots only in the vicinity of the boundary layers, but inside the domain there are no spurious oscillations. These overshoots and undershoots completely disappear for $\varepsilon \ll 1$. It is caused by the fact that the Dirichlet boundary condition is imposed in a weak sense with the aid of the boundary penalty. From this point of view, the DGM does not require such sophisticated stabilization techniques as the conforming FEM (see [JK07] for an overview).

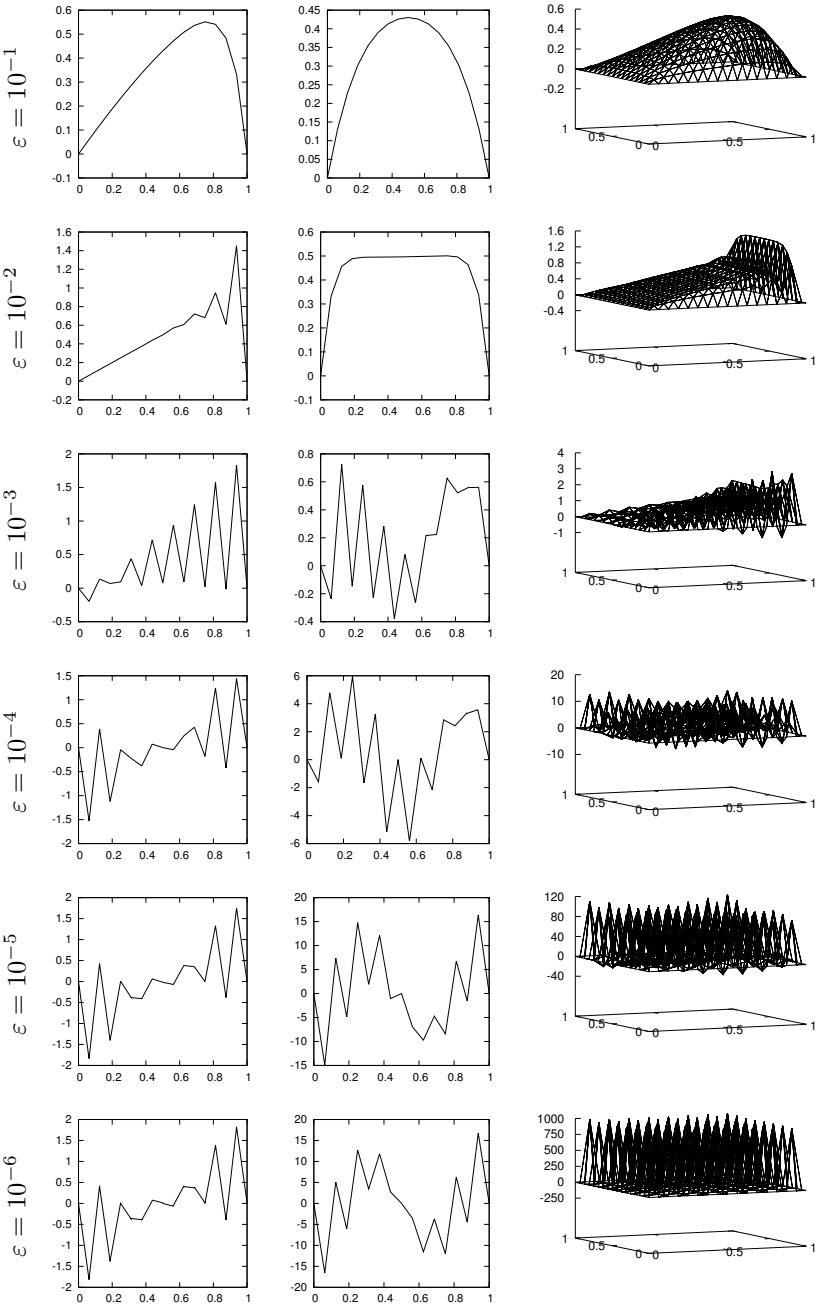

Figure 2.1: Linear convection-diffusion equation, P_1 conforming finite element method, horizontal cut at $x_2 = 0.5$ (left), vertical cut at $x_1 = 0.5$ (center), 3D view (right), for $\varepsilon = 10^{-1},\ 10^{-2},\ 10^{-3},\ 10^{-4},\ 10^{-5}$ and 10^{-6}.

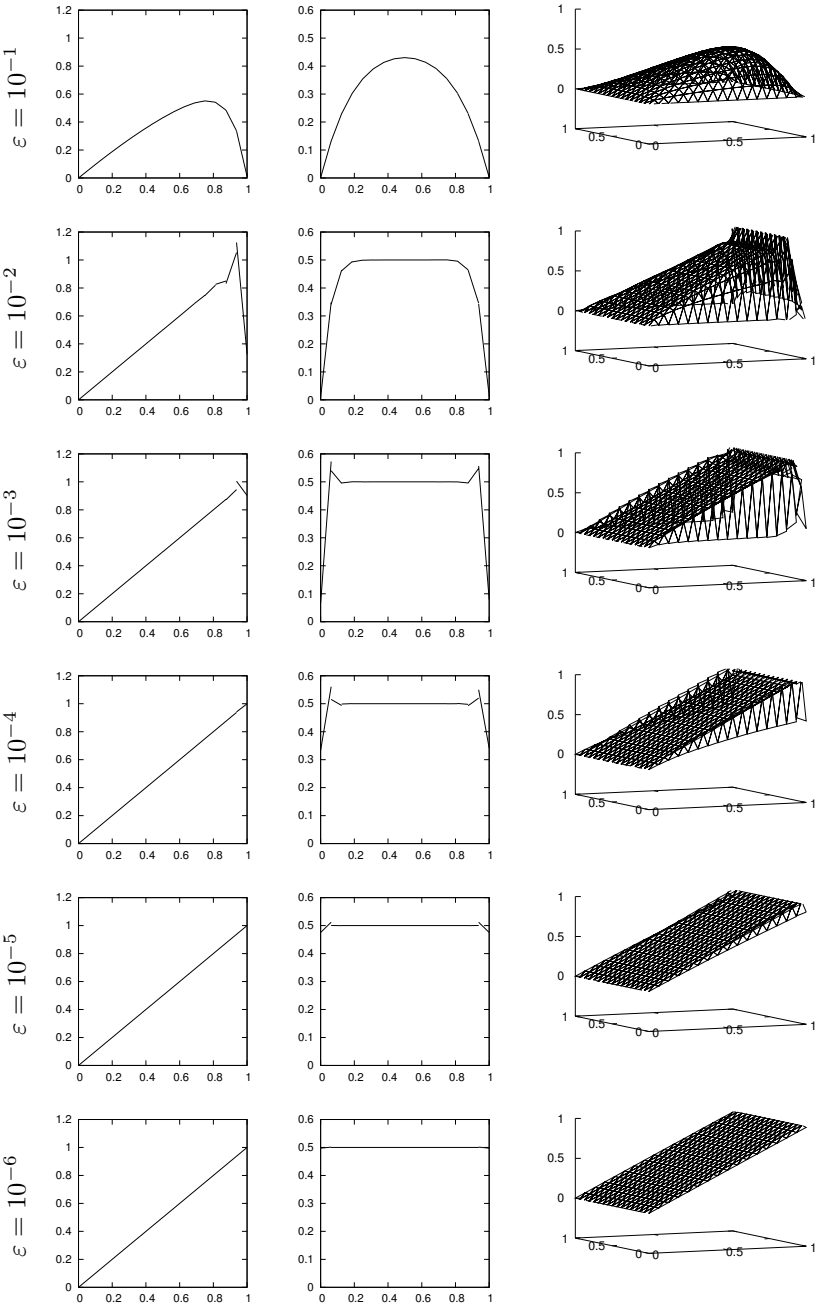

Figure 2.2: Linear convection-diffusion equation, P_1 discontinuous Galerkin method, horizontal cut at $x_2 = 0.5$ (left), vertical cut at $x_1 = 0.5$ (center), 3D view (right), for $\varepsilon = 10^{-1},\ 10^{-2},\ 10^{-3},\ 10^{-4},\ 10^{-5}$ and 10^{-6}.

Chapter 3

Time discretization by the multi-step methods

3.1 Backward difference formula for the time discretization

In Section **??**, we presented the full space-time discretization of the nonstationary initial-boundary value problem (**??**) by the semi-implicit backward Euler time scheme (**??**). This scheme has a high-order of convergence (depending on the degree of polynomial approximation) with respect to the mesh-size h, but only the first order of convergence with respect to the time step τ.

In many applications, computations with a scheme having the first-order of convergence with respect to τ are very inefficient. In this section we introduce a method for solving the nonstationary initial-boundary value problem (2.1) which is based on a combination of the discontinuous Galerkin method for the space semidiscretization and the *k-step backward difference formula* (BDF) for the time discretization. We call this technique as BDF-DGM. The BDF methods are widely used for solving stiff ODEs, see [HNW00], [**?**].

Similarly as in Section **??**, the diffusion, penalty and stabilization terms are treated implicitly, whereas the nonlinear convective terms are treated by a higher-order explicit extrapolation method. This leads to the necessity to solve only a linear algebraic problem at each time step. We analyze this scheme and derive error estimates in the discrete $L^\infty(0,T;L^2(\Omega))$-norm and in the $L^2(0,T;H^1(\Omega,\mathcal{T}_h))$-norm with respect to the mesh-size h and time step τ for $k = 2,\ 3$. Mostly, we follow the strategy from [**?**]. In this section we analyze only the SIPG technique which allows us to obtain h-optimal error estimates in the $L^2(\Omega)$-norm. Concerning NIPG and IIPG approaches, see Remark 3.9.

We consider again the *initial-boundary value problem* (2.1) to find $u : Q_T \to \mathbb{R}$ such that

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} = \varepsilon\,\Delta u + g \quad \text{in } Q_T, \tag{3.1a}$$

$$u\,\big|_{\partial\Omega_D\times(0,T)} = u_D, \tag{3.1b}$$

$$\varepsilon \boldsymbol{n}\cdot\nabla u\,\big|_{\partial\Omega_N\times(0,T)} = g_N, \tag{3.1c}$$

$$u(x,0) = u^0(x), \quad x\in\Omega. \tag{3.1d}$$

We assume that the data satisfy conditions (2.2), i.e.,

$$\begin{aligned}
&\boldsymbol{f} = (f_1,\dots,f_d),\ f_s\in C^1(\mathbb{R}),\ f_s' \text{ are bounded},\ f_s(0)=0,\ s=1,\dots,d,\\
&u_D = \text{trace of some } u^*\in C([0,T];H^1(\Omega))\cap L^\infty(Q_T) \text{ on } \partial\Omega + D\times(0,T),\\
&\varepsilon>0,\quad g\in C([0,T];L^2(\Omega)),\quad g_N\in C([0,T];L^2(\partial\Omega_N)),\quad u^0\in L^2(\Omega).
\end{aligned}$$

We suppose that there exists a weak solution u of (3.1) which is sufficiently regular, namely,

$$u\in W^{1,\infty}(0,T;H^s(\Omega))\cap W^{k,\infty}(0,T;H^1(\Omega))\cap W^{k+1,\infty}(0,T;L^2(\Omega)), \tag{3.2}$$

where $s\ge 2$ is an integer. Such a solution satisfies problem (3.1) pointwise. Under (3.2), we have $u\in C([0,T];H^s(\Omega)),\ u'\in C([0,T];L^2(\Omega))$, where u' means the derivative of $\partial u(t)/\partial t$. (For the definitions of the above function spaces, see Section 0.1.5.)

The symbol $(\cdot,\cdot)$ denotes the scalar product in the space $L^2(\Omega)$.

k	$\alpha_i,\ i=k,k-1,\dots,0$				$\beta_i,\ i=1,\dots,k$		
1	1	-1			1		
2	$\frac{3}{2}$	-2	$\frac{1}{2}$		2	-1	
3	$\frac{11}{6}$	-3	$\frac{3}{2}$	$-\frac{1}{3}$	3	-3	1

Table 3.1: Values of the coefficients $\alpha_i,\ i=0,\dots,k$, and $\beta_i,\ i=1,\dots,k$, for $k=1,2,3$.

3.1.1 Discretization of the problem

We use the same notation and assumptions as in Sections 1.4 and 2.2. This means that we suppose that the domain Ω is polygonal if $d=2$, or polyhedral if $d=3$, with Lipschitz boundary. By $\mathcal{T}_h$ we denote a partition of the domain Ω and use the diffusion, penalty, right-hand side and convection forms $A_h, a_h, \ell_h, J_h^\sigma, b_h$, defined in Section 2.2 by relations (2.9) – (2.13) and (2.23). Let $p\geq 1$ be an integer and let S_{hp} be the space of discontinuous piecewise polynomial functions (1.34). Moreover, we assume that Assumptions 2.5 in Section 2.3 are satisfied. Let us recall that the functions $f_s,\ s=1,\dots,d$, are Lipschitz-continuous with constant $L_f=2L_H$, where the constant L_H is introduced in (2.18).

Furthermore, as was already shown (cf. (2.28)), the exact solution u with property (3.2) satisfies the *consistency* identity

$$\left(\frac{\partial u}{\partial t}(t),v_h\right)+A_h(u(t),v_h)+b_h(u(t),v_h)=\ell_h(v_h)\,(t)\quad\forall v_h\in S_{hp}\ \forall t\in(0,T). \tag{3.3}$$

Now, because of time discretization, we shall consider a uniform partition of the time interval $[0,T]$ formed by the time instants $t_j=j\tau,\ j=0,1,\dots,r$, with a time step $\tau=T/r$, where $r>k$ is an integer. The value $u(t_j)$ of the exact solution will be approximated by an element $u_h^j\in S_{hp},\ j=0,\dots,r$.

Let $k\geq 1$ be an integer. The time derivative in (3.3) will be approximated by a high-order k-step *backward difference formula*

$$\frac{\partial u}{\partial t}(t_{j+k})\approx\frac{1}{\tau}\left(\alpha_k u_h^{j+k}+\alpha_{k-1}u_h^{j+k-1}+\cdots+\alpha_0 u_h^j\right)=\frac{1}{\tau}\sum_{i=0}^{k}\alpha_i u_h^{j+i}, \tag{3.4}$$

where $u_h^{j+l}\approx u(t_{j+l}$ and $\alpha_i,\ i=0,\dots,k$, are the so-called *BDF coefficients* given by

$$\alpha_k=\sum_{i=1}^{k}\frac{1}{i},\quad \alpha_i=(-1)^{k-i}\binom{k}{i}\frac{1}{k-i},\quad i=0,\dots,k-1. \tag{3.5}$$

In order to obtain an accurate, stable, efficient and simple scheme, the forms A_h and ℓ_h will be treated implicitly, whereas the nonlinear terms represented by the form b_h will be treated explicitly. In order to keep the high order of the scheme with respect to the time step, in b_h we employ a *high-order explicit extrapolation*

$$u(t_{j+k})\approx\left(\beta_1 u_h^{j+k-1}+\beta_2 u_h^{j+k-2}+\cdots+\beta_k u_h^j\right)=\sum_{i=1}^{k}\beta_i u_h^{j+k-i}, \tag{3.6}$$

where $\beta_i,\ i=1,\dots,k$, are the coefficients given by

$$\beta_i=(-1)^{i+1}\binom{k}{i}=-\alpha_{k-i}i,\quad i=1,\dots,k. \tag{3.7}$$

Table 3.1 shows the values of $\alpha_i,\ i=0,\dots,k$, and $\beta_i,\ i=1,\dots,k$, for $k=1,2,3$.

Now we are ready to introduce the full space-time BDF-DG discretization of problem (3.1).

Definition 3.1. *Let $k\geq 1$ be an integer and let $u_h^1,\dots,u_h^{k-1}\in S_{hp}$ be given. We define the* approximate solution *of problem* (3.1) *obtained by the* semi-implicit k-step BDF-DG *method as functions u_h^{l+k}, $t_{l+k}\in[0,T]$, satisfying the conditions*

$$u_h^{l+k}\in S_{hp}, \tag{3.8a}$$

$$\frac{1}{\tau}\left(\sum_{i=0}^{k}\alpha_i u_h^{l+i},v_h\right)+A_h(u_h^{l+k},v_h)+b_h(E^{l+k}(u_h),v_h)=\ell_h(v_h)\,(t_{l+k}) \tag{3.8b}$$

$$\forall\, v_h\in S_{hp},\ l=0,1,2,\dots,r-k,$$

where E^m denotes the high-order explicit extrapolation operator at the time level t_m given by

$$E^m(u_h) = \sum_{i=1}^{k} \beta_i u_h^{m-i}, \tag{3.9}$$

and $\alpha_i,\ i = 0, \dots, k$, and $\beta_i,\ i = 1, \dots, k$, are given by (3.5) and (3.7), respectively. The function u_h^l is called the approximate solution *at time $t_l,\ l = 0, \dots, r$.*

Remark 3.2. *(i) We see that the high-order explicit extrapolation $E^{l+k}(u_h)$ depends on $u_h^l, \dots, u_h^{l+k-1}$ and is independent of u_h^{l+k}.*
(ii) Since scheme (3.8) *represents a k-step formula, we have to define the approximate solution u_h^0, $u_h^1, \dots, u_h^{k-1}$ at times $t_0 = 0$, $t_1, \dots, t_{k-1}$. The initial value u_h^0 is defined as the $L^2(\Omega)$ projection of the initial data u^0 on the space S_{hp}. This means that $u_h^0 \in S_{hp}$ and*

$$(u_h^0 - u^0, v_h) = 0 \quad \forall v_h \in S_{hp}.$$

The values $u_h^1, \dots, u_h^{k-1}$ have to be determined, e.g., by a one-step method as, for example, a k^{th}-order Runge–Kutta scheme, see Section **??**.

(iii) The discrete problem (3.8) *is equivalent to a system of linear algebraic equations for each $t_{l+k} \in [0, T]$. The existence and uniqueness of the solution of this linear algebraic problem is proved in Section 3.1.2.*

(iv) The explicit extrapolation can also be applied to $u \in C([0, T]; L^2(\Omega))$ by

$$E^{l+k}(u) = \sum_{i=1}^{k} \beta_i u^{l+k-i}, \qquad t_l,\ t_{l+k} \in [0, T]. \tag{3.10}$$

3.1.2 Theoretical analysis

In what follows we shall be concerned with the analysis of method (3.8) for the SIPG variant of the DGM. Hence, we set $\Theta = 1$ in the definitions (2.10) and (2.13) of the forms A_h and ℓ_h. Moreover, we confine our considerations to the case when $\partial\Omega_N = \emptyset$. This means that we analyze problem (**??**) from Section **??**. Other possibilities will be mentioned in Remark 3.9.

Similarly, as in the analysis of schemes for the numerical solution of ordinary differential equations, we introduce the concept of stability of the BDF method.

Definition 3.3. *The BDF method* (3.8) *is stable (by Dahlquist), if all roots of the polynomial $\rho(\xi) = \sum_{j=0}^{k} \alpha_j \xi^j$ lie in the unit closed circle $\{z \in \mathbb{C}; |z| \le 1\}$ and the roots satisfying the condition $|\xi| = 1$ are simple (the symbol $\mathbb{C}$ denotes the set of complex numbers).*

Theorem 3.4. *Let Assumptions 2.5 from Section 2.3 be satisfied and let $\partial\Omega_N = \emptyset$. Let u be the exact solution of problem* (3.1) *satisfying* (3.2). *Let $t_l = l\tau,\ l = 0, 1, \dots, r,\ \tau = T/r$, be a time partition of $[0, T]$, let $u_h^l,\ l = 0, \dots, r$, be the approximate solution defined by the k-step BDF-DG scheme* (3.8) *with $k = 2$ and let $\tau \le 1$. Then there exists a constant $\widetilde{C}_2 = O\left(\exp(3GT(1 + 2K/\varepsilon))\right)$ independent of h and τ such that*

$$\|e\|^2_{h,\tau,L^\infty(L^2)} \le \widetilde{C}_2 \left((h^{2\mu} + \tau^4)(1 + 1/\varepsilon) + \sum_{j=0}^{1} \|e_h^j\|^2_{L^2(\Omega)} \right), \tag{3.11}$$

where K is defined by (**??**) *and G by* (**??**).

Now, we formulate the $L^\infty(L^2)$-error estimate of the three step method.

Theorem 3.5. *Let Assumptions 2.5 in Section 2.3 be satisfied and let $\partial\Omega_N = \emptyset$. Let u be the exact solution of problem* (3.1) *satisfying* (3.2). *Let $t_l = l\tau,\ l = 0, 1, \dots, r,\ \tau = T/r$, be a partition of the time interval $[0, T]$, let $u_h^l,\ l = 0, \dots, r$ be defined by the k-step BDF-DG scheme* (3.8) *with $k = 3$ and let $\tau \le 1$. Then there exists a constant $\widetilde{C}_3 = O\left(\exp(GT(30 + 117K/4\varepsilon))\right)$ such that*

$$\|e\|^2_{h,\tau,L^\infty(L^2)} \le \widetilde{C}_3 \left((h^{2\mu} + \tau^6)(1 + 1/\varepsilon) + \sum_{l=0}^{2} \|e_h^l\|^2_{L^2(\Omega)} + \tau\varepsilon |||\zeta^2|||^2 \right), \tag{3.12}$$

where K is defined by (**??**) *and $\zeta^2 = u_h^2 - \mathrm{P}_{hp} u^2$ is given by* (**??**).

Theorem 3.6. *Let Assumptions 2.5 in Section 2.3 be satisfied and let $\partial\Omega_N = \emptyset$. Let u be the exact solution of problem* (3.1) *satisfying* (3.2). *Let $t_l = l\tau,\ l = 0, 1, \dots, r,\ \tau = T/r$, be a partition of $[0, T]$ and let $u_h^l,\ l = 0, \dots, r$, be the approximate solution defined by the k-step BDF-DG scheme* (3.8) *with $k = 2, 3$. Then there exists a constant $\widehat{C}$ such that*

$$\begin{aligned} &\|e\|^2_{h,\tau,L^2(H^1)} \\ &\le \widehat{C}\Big(\varepsilon h^{2(\mu-1)} + (1+1/\varepsilon)^2(h^{2\mu} + \tau^{2k}) + (1+1/\varepsilon)\sum_{j=0}^{k-1}\Big(\|e_h^j\|_{L^2(\Omega)} + \tau\varepsilon|||e_h^j|||^2\Big)\Big). \end{aligned} \tag{3.13}$$

Remark 3.7. *We observe that estimates (3.11), (3.12) and (3.13) are optimal with respect to h as well as τ in the discrete $L^\infty(0, T; L^2(\Omega))$-norm and $L^2(0, T; H^1(\Omega, \mathcal{T}_h))$-norm.*

It can be seen that these estimates are not of practical use for $\varepsilon \to 0+$, because they blow up exponentially with respect to $1/\varepsilon$. This is caused by the treatment of nonlinear terms in the error analysis. The nonlinearity of the convective terms represents a serious obstacle for obtaining a uniform error estimate with respect to $\varepsilon \to 0+$.

Remark 3.8. *The proven unconditional stability may seem to be in contradiction with the Dahlquist barrier (see [**?**, Theorem 1.4]) which implies that the 3-step BDF method cannot be unconditionally A-stable. However, in our case, the k-step BDF scheme with $k = 2, 3$ was not applied to a general system of ODEs, but to system* (3.1) *arising from the space semi-discretization of (3.1) under the assumptions of the symmetry of the form A_h and some favourable properties of the form b_h, which cause that all eigenvalues of the Jacobi matrix of the corresponding ODE system lie in the stability region of the k-step BDF method with $k = 2, 3$ for any $\tau \le 1$ and $h \in (0, \bar{h})$.*

Remark 3.9. *The presented numerical analysis can be partly extended also to NIPG and IIPG variants of the DG method. However, the determination of error estimates for the 3-step BDF-DG method employs equality (**??**), which is not valid for NIPG and IIPG variants due to their non-symmetry. It is not clear to us whether it is possible to avoid this obstacle.*

On the other hand, for the 2-step BDF-DG method, a weaker result than (3.11) *can be derived for NIPG and IIPG variants, for example,*

$$\|e\|^2_{h,\tau,L^\infty(L^2)} \le \widetilde{C}\left((h^{2(\mu-1)} + \tau^4)(1+1/\varepsilon) + \sum_{j=0}^{1}\|e_h^j\|^2_{L^2(\Omega)}\right), \tag{3.14}$$

where $\widetilde{C}$ is independent of h and τ. Estimate (3.14) *can also be proved in the case of mixed Dirichlet–Neumann boundary conditions, i.e., for nonempty $\partial\Omega_N$.*

3.1.3 Numerical examples

In this section we demonstrate the theoretical error estimates (3.11), (3.12) and (3.13) derived in the previous section. We try to investigate the dependence of the computational error on h and τ independently. Based on (3.11), (3.12) and (3.13) we expect that the computational error $e_{h,\tau}$ in the $L^2(\Omega)$-norm as well as the $H^1(\Omega, \mathcal{T}_h)$-seminorm depends on h and τ according to the formula

$$e_{h,\tau} \approx c_h h^{p+1} + c_\tau \tau^k, \tag{3.15}$$

where c_h and c_τ are constants independent of h and τ.

In our numerical experiments we solve equation (3.1a) in $\Omega = (0, 1)^2$, $\partial\Omega = \partial\Omega_D$, $f_i(u) = u^2/2,\ i = 1, 2$, equipped with the boundary condition (3.1b) and the initial condition (3.1d).

Convergence with respect to τ

In this case we put $\varepsilon = 0.01$, $T = 1$ and the functions u_D, u_0 and g are chosen in such a way that the exact solution has the form $u(x_1, x_2, t) = 16\,(e^{10t} - 1)/(e^{10} - 1)\, x_1(1 - x_1)x_2(1 - x_2)$.

The computations were carried out on a fine triangular mesh having 4219 elements with a piecewise cubic approximation in space and using 6 different time steps: 1/20, 1/40, 1/80, 1/160, 1/320, 1/640. For such data setting we expect that $c_h h^{p+1} \ll c_\tau \tau^k$ and, therefore the space discretization errors are negligible. Fig. 3.1 shows the computational errors at $t = T$ and the corresponding experimental order of convergence with respect to τ in the $L^2(\Omega)$-norm and the $H^1(\Omega, \mathcal{T}_h)$-seminorm for the k-step BDF scheme (3.8) with $k = 1$, $k = 2$ and $k = 3$. The expected order of convergence $O(\tau^k)$ is observed in each case. A smaller decrease of the order of convergence in the $H^1(\Omega, \mathcal{T}_h)$-seminorm for $k = 3$ and $\tau = 1/640$ is caused by the influence of the spatial discretization since in this case the statement $c_h h^{p+1} \ll c_\tau \tau^k$ is no longer valid.

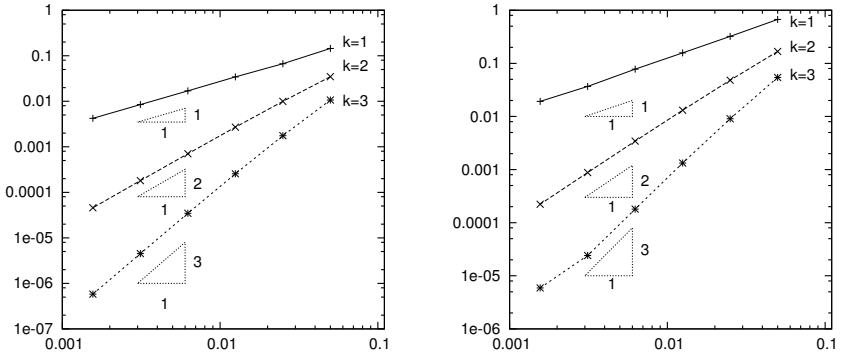

Figure 3.1: Computational errors and orders of convergence with respect to the time step τ in the $L^2(\Omega)$-norm (left) and the $H^1(\Omega, \mathcal{T}_h)$-seminorm (right) for scheme (3.8) with $k = 1$ (full line), $k = 2$ (dashed line) and $k = 3$ (dotted line).

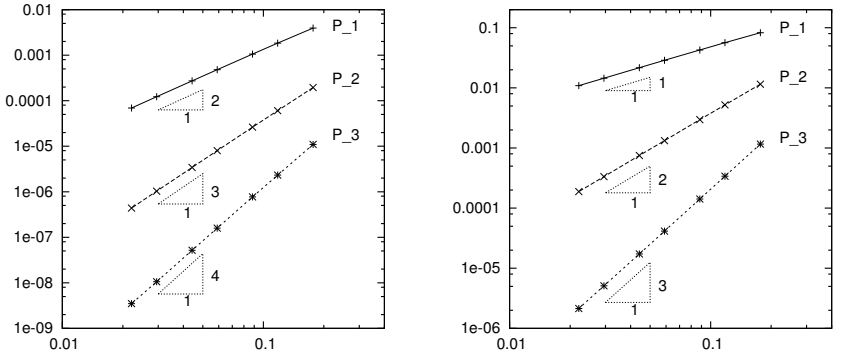

Figure 3.2: Computational errors and orders of convergence with respect to the mesh-size h in the $L^2(\Omega)$-norm (left) and the $H^1(\Omega, \mathcal{T}_h)$-seminorm (right) for scheme (3.8) with P_1 (full line), P_2 (dashed line) and P_3 (dotted line) approximations.

Convergence with respect to h

In this case we put $\varepsilon = 0.1$, $T = 10$ and the functions u_D, u_0 and g are chosen in such a way that the exact solution has the form $u(x_1, x_2, t) = (1 - e^{-10t})(x_1^2 + x_2^2)x_1x_2(1 - x_1)(1 - x_2)$. As we see, we have $\mu = p + 1$.

The computations were carried out with the 3-step BDF scheme (3.8) on 7 triangular meshes having 128, 288, 512, 1152, 2048, 4608 and 8192 elements, using the time step $\tau = 0.01$. For such data setting we expect that $c_h h^{p+1} \gg c_\tau \tau^k$ and the time discretization errors can be neglected. Fig. 3.2 shows the computational errors at $t = T$ and the corresponding experimental order of convergence with respect to h in the $L^2(\Omega)$-norm and the $H^1(\Omega, \mathcal{T}_h)$-seminorm for piecewise linear P_1, quadratic P_2 and cubic P_3 approximations. We observe the order of convergence $O(h^{p+1})$ for $p = 1, 2, 3$ in the $L^2(\Omega)$–norm and $O(h^p)$ in the $H^1(\Omega, \mathcal{T}_h)$-seminorm, which perfectly corresponds to the theoretical results (3.13).

Chapter 4

Time discretization by time discontinuous Galerkin method

In Chapter **??**, we introduced and analyzed methods based on the combination of the DGM space discretization with the backward difference formula in time. Although this approach gives satisfactory results in a number of applications (see Chapter 7), its drawback is a complicated adaptation of the space computational mesh and the time step. From this point of view, more suitable approach is the *space-time discontinuous Galerkin* method (ST-DGM), where the DGM is applied separately in space and in time.

The ST-DGM can use different triangulations on different time levels arising due to a mesh adaptation and, thus, it perfectly suits the numerical solution of nonstationary problems. Moreover, the ST-DGM can (locally) employ different polynomial degrees p and q in space and time discretization, respectively.

Section 4.1 will be concerned with basic ideas and techniques of the ST-DGM applied to a linear model heat equation. In Section 4.2, we extend the analysis to a more general convection-diffusion problem with nonlinear convection and nonlinear diffusion. Sections **??** and **??** will be devoted to some special ST-DG techniques.

4.1 Space-time DGM for a heat equation

In this section, we present and analyze the *space-time discontinuous Galerkin* method applied to a simple model problem represented by the linear heat equation. We explain the main aspects of the ST-DG discretization for this problem and derive the error estimates in the $L^\infty(0,T;L^2(\Omega))$-norm and the DG-norm formed by the $L^2(0,T;H^1(\Omega,\mathcal{T}_h))$-norm and penalty terms.

Let $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3, be a bounded polygonal or polyhedral domain, $T > 0$ and $Q_T := \Omega \times (0,T)$. We consider the problem to find $u : Q_T \to \mathbb{R}$ such that

$$\frac{\partial u}{\partial t} = \varepsilon\,\Delta u + g \quad \text{in } Q_T, \tag{4.1a}$$

$$u\,\big|_{\partial\Omega_D\times(0,T)} = u_D, \tag{4.1b}$$

$$\nabla u\cdot\boldsymbol{n}\,\big|_{\partial\Omega_N\times(0,T)} = g_N, \tag{4.1c}$$

$$u(x,0) = u^0(x), \quad x\in\Omega, \tag{4.1d}$$

Similarly as in Section 2.2 we assume that the boundary $\partial\Omega$ is formed by two disjoint parts $\partial\Omega_D$ and $\partial\Omega_N$ with $\mathrm{meas}_{d-1}(\partial\Omega_D) > 0$, and that the data satisfy the usual conditions (cf. (2.2)): u_D = trace of some $u^* \in C([0,T];H^1(\Omega))$ on $\partial\Omega_D\times(0,T)$, $\varepsilon > 0$, $g \in C([0,T];L^2(\Omega))$, $g_N \in C([0,T];L^2(\partial\Omega_N))$ and $u^0 \in L^2(\Omega)$.

4.1.1 Discretization of the problem

Space-time partition and function spaces

In order to derive the space-time discontinuous Galerkin discretization, we introduce some notation.

Let $r > 1$ be an integer. In the time interval $[0,T]$ we construct a partition $0 = t_0 < \cdots < t_r = T$ and denote

$$I_m = (t_{m-1},t_m), \quad \overline{I}_m = [t_{m-1},t_m], \quad \tau_m = t_m - t_{m-1}, \quad \tau = \max_{m=1,\ldots,r}\tau_m.$$

Then

$$[0,T] = \cup_{m=1}^r \overline{I}_m, \quad I_m\cap I_n = \emptyset \ \text{ for } m\neq n,\ m,n = 1,\ldots,r.$$

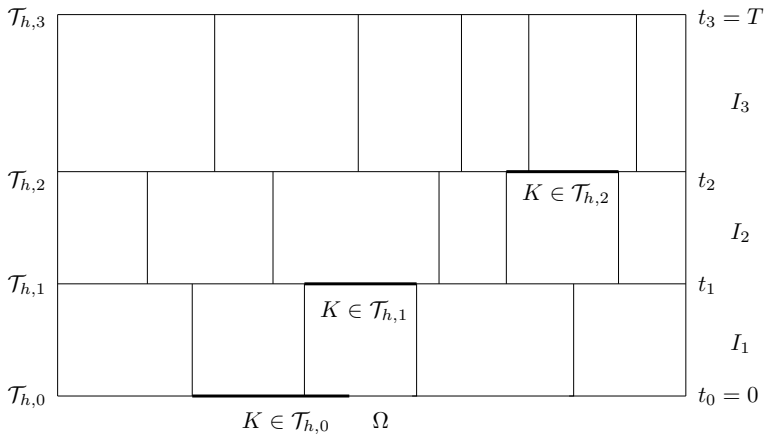

Figure 4.1: Space-time discretization for space dimension $d = 1$.

If φ is a function defined in $\bigcup_{m=1}^{r} I_m$, we introduce the notation

$$\varphi_m^{\pm} = \varphi(t_m \pm) = \lim_{t \to t_m \pm} \varphi(t), \quad \{\varphi\}_m = \varphi_m^+ - \varphi_m^-, \tag{4.2}$$

provided the one-sided limits $\lim_{t \to t_m \pm} \varphi(t)$ exist.

For each time instant $t_m,\ m = 0, \ldots, r$, and interval $I_m,\ m = 1, \ldots, r$, we consider a partition $\mathcal{T}_{h,m}$ (called triangulation) of the closure $\overline{\Omega}$ of the domain Ω into a finite number of closed simplexes (triangles for $d = 2$ and tetrahedra for $d = 3$) with mutually disjoint interiors. The partitions $\mathcal{T}_{h,m}$ may be in general different for different m. Figure 4.1 shows an illustrative example of the space-time partition for $d = 1$.

In what follows, we shall use a similar notation as in Section 1.1, only a subscript $_m$ has to be added to the notation because of different grids $\mathcal{T}_{h,m}$. By $\mathcal{F}_{h,m}$ we denote the system of all faces of all elements $K \in \mathcal{T}_{h,m}$. Further, we denote the set of all inner faces by $\mathcal{F}_{h,m}^I$ and the set of all boundary faces by $\mathcal{F}_{h,m}^B$. Each $\Gamma \in \mathcal{F}_{h,m}$ will be associated with a unit normal vector $\boldsymbol{n}_\Gamma$, which has the same orientation as the outer normal to $\partial\Omega$ for $\Gamma \in \mathcal{F}_{h,m}^B$. In $\Gamma \in \mathcal{F}_{h,m}^B$ we distinguish the subsets the of all "Dirichlet" boundary faces $\mathcal{F}_h^D = \{\Gamma \in \mathcal{F}_h;\ \Gamma \subset \partial\Omega_D\}$ and of all "Neumann" boundary faces $\mathcal{F}_h^N = \{\Gamma \in \mathcal{F}_h,\ \Gamma \subset \partial\Omega_N\}$. We set

$$h_K = \mathrm{diam}(K) \text{ for } K \in \mathcal{T}_{h,m}, \quad h_m = \max_{K \in \mathcal{T}_{h,m}} h_K, \quad h = \max_{m=1,\ldots,r} h_m.$$

By ρ_K we denote the radius of the largest ball inscribed into K.

For any integer $k \geq 1$, over a triangulation $\mathcal{T}_{h,m}$ we define the *broken Sobolev space*

$$H^k(\Omega, \mathcal{T}_{h,m}) = \{v \in L^2(\Omega); v|_K \in H^k(K)\ \forall\, K \in \mathcal{T}_{h,m}\}, \tag{4.3}$$

with seminorm

$$|v|_{H^k(\Omega,\mathcal{T}_{h,m})} = \left(\sum_{K \in \mathcal{T}_{h,m}} |v|^2_{H^k(K)} \right)^{1/2}. \tag{4.4}$$

In the same way as in Chapter 1, we use the symbols $\langle v \rangle_\Gamma$ and $[v]_\Gamma$ for the mean value and the jump of $v \in H^k(\Omega, \mathcal{T}_{h,m})$ on the face $\Gamma \in \mathcal{F}_{h,m}$, see (1.32).

Let $p,\ q \geq 1$ be integers. For each $m = 1, \ldots, r$ we define the finite-dimensional space

$$S_{h,m}^p = \left\{\varphi \in L^2(\Omega); \varphi|_K \in P_p(K)\ \forall\, K \in \mathcal{T}_{h,m}\right\}. \tag{4.5}$$

Over each mesh $\mathcal{T}_{h,m}$ we shall use the L^2-projections analogous to $\pi_{K,p}$ and Π_{hp} defined in (1.89) and (1.90). For simplicity we denote these projections by $\Pi_{h,m}$. Hence, if $K \in \mathcal{T}_{h,m}$, $m = 1, \ldots, r$, and $v \in L^2(K)$, then

$$(\Pi_{h,m} v)|_K \in P_p(K), \quad (\Pi_{h,m} v - v, \varphi)_{L^2(K)} = 0 \quad \forall \varphi \in P_p(K), \tag{4.6}$$

and, if $v \in L^2(\Omega)$, then

$$\Pi_{h,m} v \in S_{h,m}^p, \quad (\Pi_{h,m} v - v, \varphi) = 0 \quad \forall \varphi \in S_{h,m}^p. \tag{4.7}$$

As in previous sections, $(\cdot,\cdot)_{L^2(K)}$ and $(\cdot,\cdot)$ denote the $L^2(K)$-scalar product and the $L^2(\Omega)$-scalar product, respectively, and $P_p(K)$ denotes the space of all polynomials on K of degree $\le p$. Properties of these projections follow from Lemmas 1.22 and 1.24 and they are summarized in (**??**) – (**??**).

The approximate solution will be sought in the space of functions piecewise polynomial in space and time:

$$S_{h,\tau}^{p,q} = \Big\{ \varphi \in L^2(Q_T);\ \varphi(x,t)\big|_{I_m} = \sum_{i=0}^{q} t^i\, \varphi_{m,i}(x) \qquad \text{with } \varphi_{m,i} \in S_{h,m}^{p},\ i = 0,\dots,q,\quad m = 1,\dots,r \Big\}. \tag{4.8}$$

4.1.2 Space-time DG discretization

We derive the full space-time discontinuous Galerkin discretization in the similar way as the space discretization introduced in detail in Chapter 1. We consider a regular exact solution satisfying the conditions

$$u \in L^2(0,T;H^2(\Omega)), \quad \frac{\partial u}{\partial t} \in L^2(0,T;H^1(\Omega)). \tag{4.9}$$

Then $u \in C([0,T];H^1(\Omega))$. Such solution satisfies (4.1) pointwise. Moreover, let $m \in \{1,\dots,r\}$ be arbitrary but fixed. We multiply (4.1a) by $\varphi \in S_{h,\tau}^{p,q}$, integrate over $K \times I_m$ and sum over all elements $K \in \mathcal{T}_{h,m}$. Then

$$\begin{aligned} &\int_{I_m} (u',\varphi)\,\mathrm{d}t + \varepsilon \int_{I_m} \Big(\sum_{K\in\mathcal{T}_{h,m}} \int_K \nabla u \cdot \nabla\varphi \,\mathrm{d}x - \sum_{K\in\mathcal{T}_{h,m}} \int_{\partial K} \nabla u \cdot \boldsymbol{n}\varphi \,\mathrm{d}S \Big)\,\mathrm{d}t \\ &\quad = \int_{I_m} (g,\varphi)\,\mathrm{d}t, \end{aligned} \tag{4.10}$$

where we use the notation $u' = \partial u/\partial t$.

First, we deal with the time derivative term. With the aid of the integration by parts, we have

$$\int_{I_m} (u',\varphi)\,\mathrm{d}t = -\int_{I_m} (u,\varphi')\,\mathrm{d}t + (u_m^-,\varphi_m^-) - (u_{m-1}^+,\varphi_{m-1}^+). \tag{4.11}$$

Since the exact solution u is continuous with respect to t, we have $u_{m-1}^+ = u_{m-1}^-$ (cf. (4.2)) and, thus,

$$(u_{m-1}^+,\varphi_{m-1}^+) = (u_{m-1}^-,\varphi_{m-1}^+). \tag{4.12}$$

The substitution of (4.12) into (4.11) and the integration by parts (in the reverse manner) yield

$$\begin{aligned} \int_{I_m} (u',\varphi)\,\mathrm{d}t &= -\int_{I_m} (u,\varphi')\,\mathrm{d}t + (u_m^-,\varphi_m^-) - (u_{m-1}^-,\varphi_{m-1}^+) \\ &= \int_{I_m} (u',\varphi)\,\mathrm{d}t + (u_{m-1}^+,\varphi_{m-1}^+) - (u_{m-1}^-,\varphi_{m-1}^+) \\ &= \int_{I_m} (u',\varphi)\,\mathrm{d}t + \big(\{u\}_{m-1},\varphi_{m-1}^+\big)\,. \end{aligned} \tag{4.13}$$

Remark 4.1. *Identity* (4.13) *makes sense also for a function u, which is piecewise polynomial with respect to t on I_m, $m = 1,\dots,r$. Then the equality* (4.12) *can be interpreted in such a way that the value of the function u at t_{m-1} from the right (on the new time interval) is approximated by the $L^2(\Omega)$-projection of the value of u at t_{m-1} from the left (on the previous time interval). Therefore, we can speak about the "upwinding" with respect to time – compare with the "space upwinding" in* (2.16).

The discretization of the diffusion term and the right-hand side in (4.10) is the same as in Chapter 1. Hence, in virtue of

(1.41) – (1.42) and (1.50) – (1.53), we define the diffusion, penalty and right-hand side forms as

$$a_{h,m}(w,\varphi) = \sum_{K\in\mathcal{T}_{h,m}} \int_K \nabla w\cdot\nabla\varphi\,\mathrm{d}x - \sum_{\Gamma\in\mathcal{F}_{h,m}^I}\int_\Gamma (\langle\nabla w\rangle\cdot\boldsymbol{n}[\varphi] + \Theta\langle\nabla\varphi\rangle\cdot\boldsymbol{n}\,[w])\,\mathrm{d}S$$
$$- \sum_{\Gamma\in\mathcal{F}_{h,m}^D}\int_\Gamma (\nabla w\cdot\boldsymbol{n}\,\varphi + \Theta\,\nabla\varphi\cdot\boldsymbol{n}\,w)\,\mathrm{d}S, \tag{4.14}$$

$$J_{h,m}^\sigma(w,\varphi) = \sum_{\Gamma\in\mathcal{F}_{h,m}^I}\frac{C_W}{h_\Gamma}\int_\Gamma [w]\,[\varphi]\,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_{h,m}^D}\frac{C_W}{h_\Gamma}\int_\Gamma w\,\varphi\,\mathrm{d}S, \tag{4.15}$$

$$A_{h,m}(w,\varphi) = \varepsilon a_{h,m}(w,\varphi) + \varepsilon J_{h,m}^\sigma(w,\varphi), \tag{4.16}$$

$$\ell_{h,m}(\varphi) = \int_\Omega g\,\varphi\,\mathrm{d}x + \int_{\partial\Omega_N} g_N\varphi\,\mathrm{d}S \tag{4.17}$$
$$- \varepsilon\Theta\sum_{\Gamma\in\mathcal{F}_{h,m}^D}\int_\Gamma \nabla\varphi\cdot\boldsymbol{n}u_D\,\mathrm{d}S + \varepsilon\sum_{\Gamma\in\mathcal{F}_{h,m}^D}\frac{C_W}{h_\Gamma}\int_\Gamma u_D\,\varphi\,\mathrm{d}S,$$

where $C_W > 0$ is a suitable constant and h_Γ characterizes the face Γ (cf. Lemma 1.5). Moreover, in (4.14) and (4.17), we take $\Theta = -1$, $\Theta = 0$ and $\Theta = 1$ and obtain the nonsymmetric (NIPG), incomplete (IIPG) and symmetric (SIPG) variants of the approximation of the diffusion terms, respectively. Obviously, forms (4.14) – (4.17) make sense for $v, w, \varphi \in H^2(\Omega, \mathcal{T}_{h,m})$.

In virtue of (4.10), (4.13) and (4.14) – (4.17), the exact regular solution u satisfies the identity

$$\int_{I_m}((u',\varphi) + A_{h,m}(u,\varphi))\,\mathrm{d}t + \big(\{u\}_{m-1},\varphi_{m-1}^+\big) = \int_{I_m}\ell_{h,m}(\varphi)\,\mathrm{d}t \quad \forall\,\varphi\in S_{h,\tau}^{p,q},$$
$$\text{with } u(0-) = u^0. \tag{4.18}$$

Based on (4.18), we introduce the approximate solution.

Definition 4.2. *We say that a function U is a ST-DG approximate solution of problem* (4.1)*, if $U \in S_{h,\tau}^{p,q}$ and*

$$\int_{I_m}((U',\varphi) + A_{h,m}(U,\varphi))\,\mathrm{d}t + \big(\{U\}_{m-1},\varphi_{m-1}^+\big) = \int_{I_m}\ell_{h,m}(\varphi)\,\mathrm{d}t \tag{4.19}$$
$$\forall\,\varphi\in S_{h,\tau}^{p,q},\ m = 1,\dots,r,\quad \textit{with}\ \ U_0^- := \Pi_{h,0}u^0,$$

where $U' = \partial U/\partial t$. We call (4.19) *the space-time discontinuous Galerkin discrete problem.*

Remark 4.3. *The expression $\big(\{U\}_{m-1},\varphi_{m-1}^+\big)$ in* (4.19) *patches together the approximate solution on neighbouring intervals I_{m-1} and I_m. At time $t = t_0 = 0$ we have $\{U\}_0 = U_0^+ - \Pi_{h,m}u^0$. It is also possible to consider $q = 0$. In this case, scheme* (4.19) *represents a variant of the backward Euler method analyzed in Section* **??***. Therefore, we shall assume that $q \geq 1$.*

Remark 4.4. *With respect to the notation in previous chapters, we should denote the approximate solution by $u_{h\tau}$, which would express that the approximate solution depends of the space and time discretization parameters h and τ. However, for the sake of simplicity we use the symbol U.*

Theorem 4.5. *There exists a unique approximate solution of* (4.19)*.*

Proof. Let $m \in \{1,\dots,r\}$ be fixed and let U_{m-1}^- be given either by the initial condition or from the previous interval I_{m-1}. Identity (4.19) can be written in the form

$$\mathcal{R}(U,\varphi) = \int_{I_m}\ell_{h,m}(\varphi)\,\mathrm{d}t + \big(U_{m-1}^-,\varphi_{m-1}^+\big),\quad \varphi\in S_{h,\tau,m}^{p,q}, \tag{4.20}$$

where

$$\mathcal{R}(U,\varphi) := \int_{I_m}((U',\varphi) + A_{h,m}(U,\varphi))\,\mathrm{d}t + \big(U_{m-1}^+,\varphi_{m-1}^+\big) \tag{4.21}$$

and

$$S_{h,\tau,m}^{p,q} := \Big\{\varphi\in L^2(\Omega\times I_m);\ \varphi(x,t) = \sum_{i=0}^q t^i\,\varphi_i(x)\ \text{with}\ \varphi_{m,i}\in S_{h,m}^p,\ i = 0,\dots,q\Big\}. \tag{4.22}$$

Obviously, the form $\mathcal{R}$ is a bilinear form on the finite dimension space $S^{p,q}_{h,\tau,m}$ and the right-hand side of (4.20) is a linear functional depending on $\varphi \in S^{p,q}_{h,\tau,m}$. Then, in virtue of Corollary 0.7, it is sufficient to prove the coercivity of the form $\mathcal{R}$ on $S^{p,q}_{h,\tau,m}$ with respect to a suitable norm. Hence, using (2.85), the coercivity of $A_{h,m}$ following from (1.140) and integration over I_m, we obtain

$$\begin{aligned}\mathcal{R}(\varphi,\varphi) &= \int_{I_m} \left((\varphi',\varphi) + A_{h,m}(\varphi,\varphi)\right)\,\mathrm{d}t + \left(\varphi^+_{m-1},\varphi^+_{m-1}\right) \\ &= \int_{I_m} \left(\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\varphi\|^2_{L^2(\Omega)} + A_{h,m}(\varphi,\varphi)\right)\,\mathrm{d}t + \left\|\varphi^+_{m-1}\right\|^2_{L^2(\Omega)} \\ &= \frac{1}{2}\left(\left\|\varphi^-_m\right\|^2_{L^2(\Omega)} - \left\|\varphi^+_{m-1}\right\|^2_{L^2(\Omega)}\right) + \int_{I_m} A_{h,m}(\varphi,\varphi)\,\mathrm{d}t + \left\|\varphi^+_{m-1}\right\|^2_{L^2(\Omega)} \\ &\geq \frac{1}{2}\left(\left\|\varphi^-_m\right\|^2_{L^2(\Omega)} + \left\|\varphi^+_{m-1}\right\|^2_{L^2(\Omega)}\right) + \varepsilon C_C \int_{I_m} |||\varphi|||^2\,\mathrm{d}t =: \|\varphi\|^2_\star.\end{aligned} \tag{4.23}$$

It is possible to show that $\|\cdot\|_\star$ is a norm on the space $S^{p,q}_{h,\tau,m}$ and, thus, the form $\mathcal{R}$ is coercive. Then Corollary 0.7 implies the existence and uniqueness of the approximate solution. □ □

Exercise 4.6. *Show that $\|\cdot\|_\star$ defined in* (4.23) *is a norm on the space $S^{p,q}_{h,\tau,m}$.*

Our main goal will be the investigation of qualitative properties of the ST-DG scheme (4.19). In particular, we shall be concerned with the analysis of error estimates.

Theorem 4.7. *Let u be the exact solution of problem* (4.1) *satisfying the regularity condition* (**??**) *and let $U \in S^{p,q}_{h,\tau}$ be its approximation given by* (4.19). *Let inequality*

$$\tau_m \geq C_S h_m^2 \tag{4.24}$$

hold for all $m = 1,\dots,r$ and let the shape regularity assumption (**??**) *and the equivalence condition* (**??**) *be satisfied. Then there exists a constant $C_{17} > 0$ independent of h, τ and u such that*

$$\begin{aligned}&\|e^-_m\|^2_{L^2(\Omega)} + \frac{\varepsilon}{2}\sum_{j=1}^m \int_{I_j} |||e|||^2_j\,\mathrm{d}t \\ &\leq C_{17}\varepsilon\left(h^{2(\mu-1)}|u|^2_{C([0,T];H^\mu(\Omega))} + \tau^{2(q+\gamma)}|u|^2_{H^{q+1}(0,T;H^1(\Omega))}\right), \\ &\qquad\qquad h \in (0,\bar{h}), \quad m = 1,\dots,r.\end{aligned} \tag{4.25}$$

*Here $\gamma = 0$, if (**??**) holds and the function u_D from the boundary condition* (4.1b) *has a general behaviour. If u_D is defined by (**??**), then $\gamma = 1$ and condition* (**??**) *is not required. The symbol $|\cdot\|$ is defined by* (**??**).

Theorem 4.8. *Let u be the exact solution of problem* (4.1) *satisfying the regularity condition*

$$u \in W^{q+1,\infty}(0,T;L^2(\Omega)) \cap C([0,T];H^s(\Omega)), \tag{4.26}$$

where $s \geq 2$ is an integer and $\mu = \min(p+1,s)$. Let $U \in S^{p,q}_{h,\tau}$ be its approximation given by (4.19). *Let* (4.24) *hold for all $m = 1,\dots,r$ and let the shape regularity assumption* (**??**) *and the equivalence condition* (**??**) *be satisfied. Then there exists a constant $C_{18} > 0$ independent of h, τ and u such that*

$$\begin{aligned}&\sup_{t\in I_m} \|u(t) - U(t)\|^2_{L^2(\Omega)} \\ &\leq C_{18}\left(h^{2(\mu-1)}|u|^2_{C([0,T];H^\mu(\Omega))} + \tau_m^{2(q+1)}|u|^2_{W^{q+1,\infty}(0,T;L^2(\Omega))}\right), \\ &\qquad\qquad h \in (0,\bar{h}), \quad m = 1,\dots,r.\end{aligned} \tag{4.27}$$

4.2 Space-time DGM for nonlinear convection-diffusion problems

In this section we shall extend the space-time discontinuous Galerkin method (ST-DGM), explained in the previous section on a simple initial-boundary value problem for the heat equation, to the solution of a more general problem for a convection-diffusion equation with *nonlinear convection* and *nonlinear diffusion*. We shall derive the error estimates in the $L^2(0,T;L^2(\Omega))$-norm and the DG-norm formed by the $L^2(0,T;H^1(\Omega))$-norm and penalty terms.

Let $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) be a bounded polygonal or polyhedral domain with Lipschitz boundary and $T > 0$. We consider the following initial-boundary value problem: Find $u : Q_T = \Omega \times (0,T) \to \mathbb{R}$ such that

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} - \nabla \cdot (\beta(u)\nabla u) = g \quad \text{in } Q_T, \tag{4.28a}$$

$$u\big|_{\partial\Omega\times(0,T)} = u_D, \tag{4.28b}$$

$$u(x,0) = u^0(x), \quad x \in \Omega. \tag{4.28c}$$

We assume that g, u_D, u^0, f_s are given functions and $f_s \in C^1(\mathbb{R}), |f_s'| \le C, \ s = 1,\dots,d$. Moreover, let

$$\beta : \mathbb{R} \to [\beta_0, \beta_1], \quad 0 < \beta_0 < \beta_1 < \infty, \tag{4.29a}$$

$$|\beta(u_1) - \beta(u_2)| \le L_\beta |u_1 - u_2| \quad \forall u_1, u_2 \in \mathbb{R}. \tag{4.29b}$$

Remark 4.9. *In this section we consider problem* (4.28) *with a Dirichlet boundary condition only. This means that* $\partial\Omega_D = \partial\Omega$, $\partial\Omega_N = \emptyset$, $\mathcal{F}_h^D = \mathcal{F}_h^B$ *and* $\mathcal{F}_h^N = \emptyset$. *The analysis of the problem with mixed Dirichlet-Neumann boundary conditions is more complicated due to the properties of the convection form* b_h *derived in Section 2.3.2 and represents an open challenging subject.*

In the derivation and analysis of the discrete problem we assume that the exact solution is regular in the following sense:

$$u \in L^2(0,T;H^2(\Omega)), \quad \frac{\partial u}{\partial t} \in L^2(0,T;H^1(\Omega)), \tag{4.30}$$

$$\|\nabla u(t)\|_{L^\infty(\Omega)} \le C_B \ \text{ for } \ t \in (0,T). \tag{4.31}$$

4.2.1 Discretization of the problem

We employ the same notation as in Section 4.1. Hence, we consider a partition $0 = t_0 < t_1 < \cdots < t_r = T$ of the time interval $[0,T]$, time subintervals $I_m = (t_{m-1}, t_m), \ m = 1,\dots,r$, and triangulations $\mathcal{T}_{h,m}, \ m = 0,\dots,r$, of the domain Ω associated with the time instants $t_m, \ m = 0,\dots,r$, and intervals $I_m, \ m = 1,\dots,r$. Further, we consider function spaces $S_{h,m}^p$ defined by (4.5) and $S_{h,\tau}^{p,q}$ defined by (4.8) and the projections Π_m and π - see (4.7) and (**??**), respectively.

For the derivation of the space-time discontinuous Galerkin discretization we assume that $u \in C^1((0,T);H^2(\Omega))$ is an exact solution of problem (4.28). We multiply (4.28a) by $\varphi \in S_{h,\tau}^{p,q}$, integrate over $K \times I_m$, sum over all $K \in \mathcal{T}_{h,m}$ and perform some manipulation. The time derivative term is discretized in the same manner as in (4.11) – (4.13). The discretization of the convection term and the source term (4.10) is the same as in Chapter 2.

The discretization of the diffusion term is a little more complicated due to the nonlinearity of the function β. Using the technique from Section 1.4, the application of Green's theorem to the diffusion term gives

$$\begin{aligned} &-\sum_{K\in\mathcal{T}_{h,m}} \int_K \nabla\cdot(\beta(u)\nabla u)\varphi \,\mathrm{d}x \\ &= \sum_{K\in\mathcal{T}_{h,m}} \int_K \beta(u)\nabla u\cdot\nabla\varphi \,\mathrm{d}x - \sum_{\Gamma\in\mathcal{F}_{h,m}^{IB}} \int_\Gamma \langle\beta(u)\nabla u\rangle\cdot\boldsymbol{n}[\varphi]\,\mathrm{d}S. \end{aligned} \tag{4.32}$$

In Section 1.4, we add to the right-hand side of (4.32) face integral terms, where the roles of the exact solution u and the test function φ are mutually exchanged. However, in contrast to the case of a linear diffusion (see, e.g., (4.14)), to the right-hand side we cannot add the expression

$$\Theta \sum_{\Gamma\in\mathcal{F}_{h,m}^I} \int_\Gamma \langle\beta(\varphi)\nabla\varphi\rangle\cdot\boldsymbol{n}[u]\,\mathrm{d}S + \Theta \sum_{\Gamma\in\mathcal{F}_{h,m}^B} \int_\Gamma \beta(\varphi)\nabla\varphi\cdot\boldsymbol{n}(u-u_D)\,\mathrm{d}S,$$

obtained by the mutual exchange of u and φ, because it is not linear with respect to the test function φ. Therefore, in the argument of β we keep the exact solution u, i.e., we use the expression

$$\Theta \sum_{\Gamma\in\mathcal{F}_{h,m}^I} \int_\Gamma \langle\beta(u)\nabla\varphi\rangle\cdot\boldsymbol{n}[u]\,\mathrm{d}S + \Theta \sum_{\Gamma\in\mathcal{F}_{h,m}^B} \int_\Gamma \beta(u)\nabla\varphi\cdot\boldsymbol{n}(u-u_D)\,\mathrm{d}S, \tag{4.33}$$

which vanishes for a regular function u satisfying the Dirichlet condition (4.28b).

Finally, we arrive at the definition of the following forms. If $v, w, \varphi \in H^2(\Omega, \mathcal{T}_{h,m})$ and $C_W > 0$ is a fixed constant, we define the diffusion, penalty, convection and righ-hand side forms

$$\begin{aligned} a_{h,m}(v,w,\varphi) = &\sum_{K\in\mathcal{T}_{h,m}} \int_K \beta(v)\nabla w\cdot\nabla\varphi\,\mathrm{d}x \\ &- \sum_{\Gamma\in\mathcal{F}^I_{h,m}} \int_\Gamma \left(\langle\beta(v)\nabla w\rangle\cdot\boldsymbol{n}[\varphi] + \Theta\langle\beta(v)\nabla\varphi\rangle\cdot\boldsymbol{n}\,[w]\right)\mathrm{d}S \\ &- \sum_{\Gamma\in\mathcal{F}^B_{h,m}} \int_\Gamma \left(\beta(v)\nabla w\cdot\boldsymbol{n}\,\varphi + \Theta\,\beta(v)\nabla\varphi\cdot\boldsymbol{n}\,(w-u_D)\right)\mathrm{d}S, \end{aligned} \tag{4.34}$$

$$J^\sigma_{h,m}(w,\varphi) = \sum_{\Gamma\in\mathcal{F}^I_{h,m}} \frac{C_W}{h_\Gamma}\int_\Gamma [w]\,[\varphi]\,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}^B_{h,m}} \frac{C_W}{h_\Gamma}\int_\Gamma w\,\varphi\,\mathrm{d}S, \tag{4.35}$$

$$A_{h,m}(w,v,\varphi) = a_{h,m}(w,v,\varphi) + \beta_0 J^\sigma_{h,m}(v,\varphi), \tag{4.36}$$

$$\begin{aligned} b_{h,m}(w,\varphi) = &-\sum_{K\in\mathcal{T}_{h,m}} \int_K \sum_{s=1}^d f_s(w)\frac{\partial\varphi}{\partial x_s}\,\mathrm{d}x + \sum_{\Gamma\in\mathcal{F}^I_{h,m}}\int_\Gamma H\left(w^{(L)}_\Gamma, w^{(R)}_\Gamma, \boldsymbol{n}\right)[\varphi]\,\mathrm{d}S \\ &+ \sum_{\Gamma\in\mathcal{F}^B_{h,m}}\int_\Gamma H\left(w^{(L)}_\Gamma, w^{(L)}_\Gamma, \boldsymbol{n}\right)\varphi\,\mathrm{d}S. \end{aligned} \tag{4.37}$$

$$\ell_{h,m}(\varphi) = (g,\varphi) + \beta_0 \sum_{\Gamma\in\mathcal{F}^B_{h,m}} \frac{C_W}{h_\Gamma}\int_\Gamma u_D\,\varphi\,\mathrm{d}S. \tag{4.38}$$

In (4.34), we take $\Theta = -1$, $\Theta = 0$ and $\Theta = 1$ and obtain the nonsymmetric (NIPG), incomplete (IIPG) and symmetric (SIPG) variants of the approximation of the diffusion terms, respectively. In (4.37), H is a numerical flux with the properties (2.18) – (2.20) introduced in Section 2.2.

Similarly as in Section 4.1, the exact regular solution u of (4.28) satisfies the identity

$$\begin{aligned} &\int_{I_m} \left((u',\varphi) + A_{h,m}(u,u,\varphi) + b_{h,m}(u,\varphi)\right)\mathrm{d}t + \left(\{u\}_{m-1}, \varphi^+_{m-1}\right) \\ &= \int_{I_m} \ell_{h,m}(\varphi)\,\mathrm{d}t \quad \forall\,\varphi\in S^{p,q}_{h,\tau}, \quad \text{with } u(0-) = u(0) = u^0. \end{aligned} \tag{4.39}$$

Here $u' := \partial u/\partial t$ and $(\cdot,\cdot)$ denotes the $L^2(\Omega)$-scalar product.

Based on (4.39), we proceed to the definition of the approximate solution.

Definition 4.10. *We say that a function U is an* ST-DG approximate solution *of problem* (4.28)*, if $U \in S^{p,q}_{h,\tau}$ and*

$$\begin{aligned} &\int_{I_m} \left((U',\varphi) + A_{h,m}(U,U,\varphi) + b_{h,m}(U,\varphi)\right)\mathrm{d}t + \left(\{U\}_{m-1}, \varphi^+_{m-1}\right) \\ &= \int_{I_m} \ell_{h,m}(\varphi)\,\mathrm{d}t \quad \forall\,\varphi\in S^{p,q}_{h,\tau}, \quad m = 1,\dots,r,\ U^-_0 := \Pi_{h,0}u^0. \end{aligned} \tag{4.40}$$

where $U' = \partial U/\partial t$. We call (4.40) the space-time discontinuous Galerkin discrete problem.

Exercise 4.11. *Formulate the ST-DG discrete problem in the case, when mixed Dirichlet-Neumann boundary conditions are used.*

In the sequel, we shall analyze the ST-DGM, namely we derive an estimate of the error $e = U - u$, where u is the exact solution of (4.28) and U is the approximate solution given by (4.40). We assume that the approximate solution U exists and is unique.

4.2.2 Auxiliary results

In the analysis of the ST-DGM for the nonlinear problem we proceed in a similar way as in Section 4.1 for the heat equation. We consider a system (**??**) of triangulations $\mathcal{T}_{h,m}$, satisfying the conditions of the *shape regularity* (**??**) and of the *equivalence* (**??**). Let $\pi : C([0,T];L^2(\Omega)) \to S^{p,q}_{h,\tau}$ be the projection operator given by (**??**). The error of the method is expressed again in the form

$$e = U - u = \xi + \eta, \tag{4.41}$$

where

$$\xi = U - \pi u \in S^{p,q}_{h,\tau}, \quad \eta \;=\; \pi u - u. \tag{4.42}$$

Then, subtracting (4.39) from (4.40), and using (4.41), for each $\varphi \in S^{p,q}_{h,\tau}$ we find that

$$\begin{aligned} &\int_{I_m} ((\xi',\varphi) + A_{h,m}(U,U,\varphi) - A_{h,m}(u,u,\varphi))\,\mathrm{d}t + \left(\{\xi\}_{m-1}, \varphi^+_{m-1}\right) \\ &= \int_{I_m} (b_{h,m}(u,\varphi) - b_{h,m}(U,\varphi))\,\mathrm{d}t - \int_{I_m} (\eta',\varphi)\,\mathrm{d}t - \left(\{\eta\}_{m-1}, \varphi^+_{m-1}\right). \end{aligned} \tag{4.43}$$

Hence, we need to estimate individual terms appearing in (4.43).

The convection form $b_{h,m}$ has the following property.

Lemma 4.12. *For each $k_b > 0$ there exists a constant $C_b > 0$ independent of U, u, h, τ, r and m such that*

$$\begin{aligned} &|b_{h,m}(U,\varphi) - b_{h,m}(u,\varphi)| \\ &\leq \frac{\beta_0}{k_b} |||\varphi|||^2_m + C_b \left(\|\xi\|^2_{L^2(\Omega)} + \|\eta\|^2_{L^2(\Omega)} + \sum_{K\in\mathcal{T}_{h,m}} h_K^2 |\eta|^2_{H^1(K)} \right). \end{aligned} \tag{4.44}$$

Let us note that in the following considerations in some places the simplified form of Young's inequality $ab \leq \frac{1}{\delta}a^2 + \delta b^2$ is used.

As for the coercivity of the forms $A_{h,m}$, we can prove the following result.

Lemma 4.13. *Let*

$$C_W > 0, \quad \text{for } \Theta = -1 \ (NIPG), \tag{4.45}$$

$$C_W \geq \left(\frac{4\beta_1}{\beta_0}\right)^2 C_{MI} \quad \text{for } \Theta = 1 \ (SIPG), \tag{4.46}$$

$$C_W \geq 2\left(\frac{2\beta_1}{\beta_0}\right)^2 C_{MI} \quad \text{for } \Theta = 0 \ (IIPG), \tag{4.47}$$

where $C_{MI} = C_M(C_I+1)C_G$. Then, for $m = 1, \ldots, r$,

$$a_{h,m}(U,U,\xi) - a_{h,m}(U,\pi u,\xi) + \beta_0 J^\sigma_{h,m}(\xi,\xi) \geq \frac{\beta_0}{2} |||\xi|||^2_m. \tag{4.48}$$

Exercise 4.14. *Prove that inequality (4.48) holds in the case $\Theta = 0$ under condition (4.47).*

Lemma 4.15. *There exists a constant $C > 0$ independent of U, ξ, φ, h such that*

$$a_{h,m}(U,U,\varphi) - a_{h,m}(U,\pi u,\varphi) + \beta_0 J^\sigma_{h,m}(\xi,\varphi) \leq C(|||\xi|||^2_m + |||\varphi|||^2_m) \tag{4.49}$$

for any $\varphi \in S^p_{h,m}$ and $m = 1, \ldots, r$.

Lemma 4.16. *For arbitrary $k_a, k_c > 0$ there exist constants $C_a = C_a(k_a),\ C_c = C_c(k_c) > 0$ independent of U, ξ, φ and h such that for each $\varphi \in S^p_{h,m}$ the following estimates hold:*

$$|a_{h,m}(U,\pi u,\varphi) - a_{h,m}(u,\pi u,\varphi)| \leq \frac{\beta_0}{k_a} |||\varphi|||^2_m + C_a(\|\xi\|^2_{L^2(\Omega)} + R_m(\eta)), \tag{4.50}$$

$$|a_{h,m}(u,\pi u,\varphi) - a_{h,m}(u,u,\varphi)| \leq \frac{\beta_0}{k_c} |||\varphi|||^2_m + C_c \tilde{R}_m(\eta), \tag{4.51}$$

where

$$R_m(\eta) = |||\eta|||^2_m + \|\eta\|^2_{L^2(\Omega)} + \sum_{K\in\mathcal{T}_{h,m}} \left(|\eta|^2_{H^1(K)} + h_K^2 |\eta|^2_{H^2(K)} \right), \tag{4.52}$$

$$\tilde{R}_m(\eta) = |||\eta|||^2_m + \sum_{K\in\mathcal{T}_{h,m}} \left(h_K^2 |\eta|^2_{H^2(K)} \right). \tag{4.53}$$

Remark 4.17. *In view of (4.52), estimate (4.44) can be written as*

$$|b_{h,m}(U,\varphi) - b_{h,m}(u,\varphi)| \leq \frac{\beta_0}{k_b} |||\varphi|||^2_m + C_b \left(\|\xi\|^2_{L^2(\Omega)} + R_m(\eta) \right). \tag{4.54}$$

4.2.3 Abstract error estimate

Estimate of ξ

In what follows, we shall use the conditions (**??**) of the *shape regularity*, (**??**) of the *equivalence* and assumptions from Lemma 4.13.

Let us substitute $\varphi := \xi$ in (4.43). From the definition (4.36) of the form $A_{h,m}$ it follows that

$$\begin{aligned}
&\int_{I_m} \left((\xi',\xi) + a_{h,m}(U,U,\xi) - a_{h,m}(U,\pi u,\xi) + \beta_0 J^\sigma_{h,m}(\xi,\xi)\right) \mathrm{d}t \\
&\quad + \left(\{\xi\}_{m-1}, \xi^+_{m-1}\right) \\
=&\int_{I_m} \left(-a_{h,m}(U,\pi u,\xi) + a_{h,m}(u,\pi u,\xi) - a_{h,m}(u,\pi u,\xi) + a_{h,m}(u,u,\xi)\right) \mathrm{d}t \\
&+ \int_{I_m} \left(b_{h,m}(u,\xi) - b_{h,m}(U,\xi) - \beta_0 J^\sigma_{h,m}(\eta,\xi) - (\eta',\xi)\right) \mathrm{d}t - \left(\{\eta\}_{m-1}, \xi^+_{m-1}\right).
\end{aligned} \tag{4.55}$$

By (**??**), we have

$$\begin{aligned}
&\int_{I_m} (\xi',\xi)\mathrm{d}t + \left(\{\xi\}_{m-1}, \xi^+_{m-1}\right) \\
=&\frac{1}{2}\left(\left\|\xi^-_m\right\|^2_{L^2(\Omega)} - \left\|\xi^-_{m-1}\right\|^2_{L^2(\Omega)} + \left\|\{\xi\}_{m-1}\right\|^2_{L^2(\Omega)}\right).
\end{aligned} \tag{4.56}$$

Moreover, (**??**) with $\delta := 1$ gives

$$\int_{I_m} (\eta',\varphi)\mathrm{d}t + \left(\{\eta\}_{m-1}, \varphi^+_{m-1}\right) \le \|\eta^-_{m-1}\|^2_{L^2(\Omega)} + \frac{1}{4}\|\{\varphi\}_{m-1}\|^2_{L^2(\Omega)}, \quad \varphi \in S^{p,q}_{h,m}. \tag{4.57}$$

The use of (**??**), (4.55), (4.56), (4.57), (4.54), Young's inequality and Lemmas 4.13 and 4.16 imply that for arbitrary $\delta, k_a, k_b, k_c > 0$ we have

$$\begin{aligned}
&\left\|\xi^-_m\right\|^2_{L^2(\Omega)} - \left\|\xi^-_{m-1}\right\|^2_{L^2(\Omega)} + \frac{1}{2}\|\{\xi\}_{m-1}\|^2_{L^2(\Omega)} \\
&+ \beta_0\left(1 - \frac{2}{k_a} - \frac{2}{k_b} - \frac{2}{k_c} - 2\delta\right)\int_{I_m} |||\xi|||^2_m \mathrm{d}t \\
\le& C\left(\int_{I_m} \|\xi\|^2_{L^2(\Omega)}\mathrm{d}t + \left\|\eta^-_{m-1}\right\|^2_{L^2(\Omega)} + \int_{I_m} R_m(\eta)\mathrm{d}t\right).
\end{aligned}$$

This and the choice $k_a = k_b = k_c = 16$ and $\delta = \frac{1}{16}$ imply that

$$\begin{aligned}
&\left\|\xi^-_m\right\|^2_{L^2(\Omega)} - \left\|\xi^-_{m-1}\right\|^2_{L^2(\Omega)} + \frac{1}{2}\|\{\xi\}_{m-1}\|^2_{L^2(\Omega)} + \frac{\beta_0}{2}\int_{I_m} |||\xi|||^2_m \,\mathrm{d}t \\
\le& C\left(\int_{I_m} \|\xi\|^2_{L^2(\Omega)}\,\mathrm{d}t + \left\|\eta^-_{m-1}\right\|^2_{L^2(\Omega)} + \int_{I_m} R_m(\eta)\,\mathrm{d}t\right), \quad m = 1,\dots,r.
\end{aligned} \tag{4.58}$$

Estimate of $\int_{I_m} \|\xi\|^2_{L^2(\Omega)}\,\mathrm{d}t$

An important task is the estimation of the term $\int_{I_m} \|\xi\|^2_{L^2(\Omega)}\,\mathrm{d}t$. The case, when $\beta(u) = \text{ const } > 0$, was analyzed in [**?**] using the approach from [**?**] based on the application of the so-called Gauss-Radau quadrature and interpolation. However, in the case of nonlinear diffusion, this technique is not applicable. It appears suitable to apply here the approach from [**?**] based on the concept of discrete characteristic functions constructed to ξ in Section **??**.

We shall proceed in several steps. Let us set

$$t_{m-1+l/q} = t_{m-1} + \frac{l}{q}(t_m - t_{m-1}) \quad \text{for} \quad l = 0, ..., q,$$

and use the notation $\xi_{m-1+l/q} = \xi(t_{m-1+l/q})$, $\xi_{m-1} = \xi^+_{m-1}$, $\xi_m = \xi^-_m$.

Lemma 4.18. *There exist constants $L_q, M_q > 0$ dependent on q only such that*

$$\sum_{l=0}^{q} \left\|\xi_{m-1+l/q}\right\|^2_{L^2(\Omega)} \ge \frac{L_q}{\tau_m}\int_{I_m} \|\xi\|^2_{L^2(\Omega)}\mathrm{d}t, \tag{4.59}$$

$$\left\|\xi^+_{m-1}\right\|^2_{L^2(\Omega)} \le \frac{M_q}{\tau_m}\int_{I_m} \|\xi\|^2_{L^2(\Omega)}\mathrm{d}t. \tag{4.60}$$

Further, we shall return to identity (4.43), where we set $\varphi := \xi$. It can be written in the form

$$\begin{aligned}
&\int_{I_m} \big((\xi',\xi) + a_{h,m}(U,U,\xi) - a_{h,m}(U,\pi u,\xi) + \beta_0 J^{\sigma}_{h,m}(\xi,\xi)\big)\,\mathrm{d}t + \ \big(\xi^+_{m-1},\xi^+_{m-1}\big)\\
=&\int_{I_m} \left(-a_{h,m}(U,\pi u,\xi) + a_{h,m}(u,\pi u,\xi) - a_{h,m}(u,\pi u,\xi) + a_{h,m}(u,u,\xi)\right)\mathrm{d}t\\
&+\int_{I_m} \big(-\beta_0 J^{\sigma}_{h,m}(\eta,\xi) + b_{h,m}(u,\xi) - b_{h,m}(U,\xi) - (\eta',\xi)\big)\,\mathrm{d}t\\
&-\big(\{\eta\}_{m-1},\xi^+_{m-1}\big) + \big(\xi^-_{m-1},\xi^+_{m-1}\big) \quad \forall\varphi\in S^{p,q}_{h,\tau}.
\end{aligned}$$

Using the relations (**??**) with $\varphi := \xi$ and

$$\int_{I_m} (\xi,\xi')\mathrm{d}t + \big(\xi^+_{m-1},\xi^+_{m-1}\big) = \frac{1}{2}\left(\big\|\xi^-_m\big\|^2_{L^2(\Omega)} + \big\|\xi^+_{m-1}\big\|^2_{L^2(\Omega)}\right), \tag{4.61}$$

we get

$$\begin{aligned}
&\frac{1}{2}\left(\big\|\xi^-_m\big\|^2_{L^2(\Omega)} + \big\|\xi^+_{m-1}\big\|^2_{L^2(\Omega)}\right)\\
&+\int_{I_m} \big(a_{h,m}(U,U,\xi) - a_{h,m}(U,\pi u,\xi) + \beta_0 J^{\sigma}_{h,m}(\xi,\xi)\big)\,\mathrm{d}t\\
\le&\int_{I_m} \big(|a_{h,m}(U,\pi u,\xi) - a_{h,m}(u,\pi u,\xi)| + |a_{h,m}(u,\pi u,\xi) - a_{h,m}(u,u,\xi)|\big)\,\mathrm{d}t\\
&+\int_{I_m} \Big(\beta_0\big|J^{\sigma}_{h,m}(\eta,\xi)\big| + |b_{h,m}(U,\xi) - b_{h,m}(u,\xi)|\Big)\,\mathrm{d}t\\
&+\big|\big(\eta^-_{m-1},\xi^+_{m-1}\big)\big| + \big|\big(\xi^-_{m-1},\xi^+_{m-1}\big)\big|.
\end{aligned}$$

Now, Lemmas 4.13, 4.16, inequalities (**??**), (4.54) and Young's inequality imply that

$$\begin{aligned}
&\frac{1}{2}\left(\big\|\xi^-_m\big\|^2_{L^2(\Omega)} + \big\|\xi^+_{m-1}\big\|^2_{L^2(\Omega)}\right) + \frac{\beta_0}{2}\int_{I_m} |||\xi|||^2_m\mathrm{d}t\\
\le&\int_{I_m}\left(\frac{\beta_0}{k_a}|||\xi|||^2_m + C_a\|\xi\|^2_{L^2(\Omega)} + C_a R_m(\eta) + \frac{\beta_0}{k_c}|||\xi|||^2_m + C_c\tilde{R}_m(\eta)\right)\mathrm{d}t\\
&+\int_{I_m}\left(\frac{\beta_0}{\delta}J^{\sigma}_{h,m}(\eta,\eta) + \delta\beta_0 J^{\sigma}_{h,m}(\xi,\xi) + \frac{\beta_0}{k_b}|||\xi|||^2_m + C_b\|\xi\|^2_{L^2(\Omega)} + C_b R_m(\eta)\right)\mathrm{d}t\\
&+\frac{\big\|\eta^-_{m-1}\big\|^2_{L^2(\Omega)}}{\delta_1} + \delta_1\big\|\xi^+_{m-1}\big\|^2_{L^2(\Omega)} + \frac{\big\|\xi^-_{m-1}\big\|^2_{L^2(\Omega)}}{\delta_1} + \delta_1\big\|\xi^+_{m-1}\big\|^2_{L^2(\Omega)}.
\end{aligned}$$

After some manipulation, taking into account that $\tilde{R}_m(\eta) \le R_m(\eta)$, we get

$$\begin{aligned}
&\big\|\xi^-_m\big\|^2_{L^2(\Omega)} + \big\|\xi^+_{m-1}\big\|^2_{L^2(\Omega)} + \beta_0\left(1 - \frac{2}{k_a} - \frac{2}{k_b} - \frac{2}{k_c} - 2\delta\right)\int_{I_m} |||\xi|||^2_m\mathrm{d}t\\
\le&2(C_a + C_b)\int_{I_m}\|\xi\|^2_{L^2(\Omega)}\mathrm{d}t + \left(2(C_a + C_b + C_c) + \frac{\beta_0}{\delta}\right)\int_{I_m} R_m(\eta)\ dt\\
&+2\frac{\big\|\eta^-_{m-1}\big\|^2_{L^2(\Omega)}}{\delta_1} + 2\frac{\big\|\xi^-_{m-1}\big\|^2_{L^2(\Omega)}}{\delta_1} + 4\delta_1\big\|\xi^+_{m-1}\big\|^2_{L^2(\Omega)}.
\end{aligned}$$

Finally, the choice $k_a = k_b = k_c = 16$ and $\delta = 1/16$ yields

$$\begin{aligned}
&\big\|\xi^-_m\big\|^2_{L^2(\Omega)} + \big\|\xi^+_{m-1}\big\|^2_{L^2(\Omega)} + \frac{\beta_0}{2}\int_{I_m} |||\xi|||^2_m\mathrm{d}t\\
\le&C_1\int_{I_m}\|\xi\|^2_{L^2(\Omega)}\mathrm{d}t + C_2\int_{I_m} R_m(\eta)\mathrm{d}t\\
&+2\frac{\big\|\eta^-_{m-1}\big\|^2_{L^2(\Omega)}}{\delta_1} + 2\frac{\big\|\xi^-_{m-1}\big\|^2_{L^2(\Omega)}}{\delta_1} + 4\delta_1\big\|\xi^+_{m-1}\big\|^2_{L^2(\Omega)},
\end{aligned} \tag{4.62}$$

with constants $C_1 = 2(C_a + C_b), C_2 = 2(C_a + C_b + C_c) + 16\beta_0$.

Now we prove the following important result.

Lemma 4.19. *There exist constants C, $C^* > 0$ such that*

$$\int_{I_m} \|\xi\|^2_{L^2(\Omega)}\,\mathrm{d}t \le C\,\tau_m \left(\left\|\xi^-_{m-1}\right\|^2_{L^2(\Omega)} + \left\|\eta^-_{m-1}\right\|^2_{L^2(\Omega)} + \int_{I_m} R_m(\eta)\,\mathrm{d}t \right), \tag{4.63}$$
$$m = 1, \dots, r,$$

where $R_m(\eta)$ is defined in (4.52)*, provided*

$$0 < \tau_m \le C^*. \tag{4.64}$$

Now we finish the derivation of the *abstract error estimate* of the ST-DGM.

Theorem 4.20. *Let* (4.30)*,* (4.31) *and* (4.64) *hold. Then there exists a constant $C_{AE} > 0$ such that the error $e = U - u$ satisfies the following estimates:*

$$\|e^-_m\|^2_{L^2(\Omega)} + \frac{\beta_0}{2} \sum_{j=1}^{m} \int_{I_j} |||e|||^2_j\,\mathrm{d}t \tag{4.65}$$
$$\le C_{AE} \Big(\sum_{j=1}^{m} \|\eta^-_{j-1}\|^2_{L^2(\Omega)} + \sum_{j=1}^{m} \int_{I_j} R_j(\eta)\,\mathrm{d}t \Big) + 2\|\eta^-_m\|^2_{L^2(\Omega)} + \beta_0 \sum_{j=1}^{m} \int_{I_j} |||\eta|||^2_j\,\mathrm{d}t,$$
$$m = 1, \dots, r, \quad h \in (0, \bar{h}),$$

and

$$\|e\|^2_{L^2(Q_T)} \le C_{AE} \sum_{m=1}^{r} \tau_m \Big(\|\eta^-_{m-1}\|^2_{L^2(\Omega)} + \int_{I_m} R_m(\eta)\,\mathrm{d}t \tag{4.66}$$
$$+ \sum_{j=1}^{r} \|\eta^-_{j-1}\|^2_{L^2(\Omega)} + \sum_{j=1}^{r} \int_{I_j} R_j(\eta)\,\mathrm{d}t \Big) + 2\|\eta\|^2_{L^2(Q_T)}, \quad h \in (0, \bar{h}),$$

where $R_m(\eta)$ is defined by (4.52).

Remark 4.21. *A detailed analysis shows that the constant C_{AE} from the abstract error estimate (4.65) behaves in dependence on β_0 as $\exp(C/\beta_0)$, which means that this constant blows up for $\beta_0 \to 0+$ and the obtained error estimates cannot be applied to the case of nonlinear singularly perturbed convection-diffusion problems with degenerated diffusion. Uniform error estimates with respect to the diffusion tending to zero were obtained, e.g. in [**?**] for the space-time DG approximations of linear convection-diffusion-reaction problems. This will be treated in Section* **??**.

4.2.4 Main result

Here we present the final error estimate of the ST-DGM applied to the nonlinear convection-diffusion equation. We assume that the exact solution satisfies the regularity conditions (4.31) and

$$u \in H^{q+1}\big(0, T; H^1(\Omega)\big) \cap C([0, T]; H^s(\Omega)) \tag{4.67}$$

with integers $s \ge 2$ and $q \ge 1$. We set $\mu = \min(p + 1, s)$. Obviously, $C([0, T]; H^s(\Omega)) \subset L^2(0, T; H^s(\Omega))$ and condition (4.30) is also satisfied.

Moreover, we assume that

$$\tau_m \ge C_S h^2_m, \quad m = 1, \dots, r. \tag{4.68}$$

Let us note that it will be shown in Remark 4.24 that this assumption is not necessary, if the meshes are not time-dependent, i.e., if all meshes $\mathcal{T}_{h,m},\ m = 1, \dots, r$, are identical.

We remind that the meshes are assumed to satisfy the *shape regularity assumption* (**??**) and the *equivalence condition* (**??**).

Now we can formulate the main results of the analysis of the error estimates for the ST-DGM.

Theorem 4.22. *Let u be the exact solution of problem* (4.28) *satisfying the regularity conditions* (4.31) *and* (4.67)*. Let the system of triangulation satisfy the shape regularity assumption* (**??**) *and the equivalence condition* (**??**) *and the time steps $\tau_m,\ m = 1, \dots, r$ satisfy the conditions* (4.64) *and* (4.68)*. Let U be the approximate solution to problem* (4.28) *obtained by scheme* (4.40)*. Then there exists a constant $C > 0$ independent of h, τ, m, r, u, U such that*

$$\|e^-_m\|^2_{L^2(\Omega)} + \frac{\beta_0}{2} \sum_{j=1}^{m} \int_{I_j} |||e|||^2_j\,\mathrm{d}t \tag{4.69}$$
$$\le C \left(h^{2(\mu-1)} |u|^2_{C([0,T];H^\mu(\Omega))} + \tau^{2(q+\gamma)} |u|_{H^{q+1}(0,T;H^1(\Omega))} \right),\ m = 1, \dots, r,\ h \in (0, \bar{h}),$$

and

$$\|e\|^2_{L^2(Q_T)} \le C\left(h^{2(\mu-1)}|u|^2_{L^2(0,T;H^\mu(\Omega))} + \tau^{2(q+\gamma)}|u|^2_{H^{q+1}(0,T;H^1(\Omega))}\right),\ h\in(0,\bar h). \tag{4.70}$$

Here $\gamma = 0$, if (??) holds and the function u_D from the boundary condition (4.28b) *has a general behaviour with respect to t. If u_D is defined by (??), then $\gamma = 1$ and condition (??) is not required. (The symbol $|\cdot\|$ is defined by (??).)*

Exercise 4.23. *Prove estimate* (4.70) *in detail.*

Remark 4.24. The case of identical meshes on all time levels. *Similarly as in Section ??, assumption* (4.24) *can be avoided, if all meshes $\mathcal{T}_{h,m}$, $m = 1\dots,r$, are identical. Then relations (??) and (??) are valid and it is possible to show that the expression $\sum_{j=1}^m \|\eta^-_{j-1}\|^2_{L^2(\Omega)}$ does not appear in estimate (??). We find that instead of (4.65) we get the abstract error estimate in the form*

$$\begin{aligned}&\|e^-_m\|^2_{L^2(\Omega)} + \frac{\beta_0}{2}\sum_{j=1}^m\int_{I_j}|||e|||^2_j\,\mathrm{d}t \\ \le& C\sum_{j=1}^m\int_{I_j}R_j(\eta)\,\mathrm{d}t + 2\|\eta^-_m\|^2_{L^2(\Omega)} + \beta_0\sum_{j=1}^m\int_{I_j}|||\eta|||^2_j\,\mathrm{d}t,\\ &m=1,\dots,r,\quad h\in(0,\bar h).\end{aligned} \tag{4.71}$$

Then Theorem 4.20 holds without assumption (4.68).

Remark 4.25. *The error estimate (4.70) in the L^2-norm is of order $O(h^{\mu-1})$ with respect to h, which is suboptimal in comparison to the interpolation error estimate (??) and one would expect the error estimate in the L^2-norm of order $O(h^\mu)$. This is a well-known phenomenon in the finite element method as well as in the DGM. In several discontinuous Galerkin techniques, similarly as in conforming finite elements (cf. [Cia79]), it is possible to prove the optimal error estimate in the L^2-norm in the case of the SIPG version with the aid of the Nitsche method, as for example in [Arn82], [?], [?] and [?]. See also Sections 1.7.2 and 2.5. The case, when the space-time DGM is applied to the nonlinear convection-diffusion problem, remains to be solved.*

Remark 4.26. *Similarly, as in Remark 4.21, it is possible to show that in the above error estimates, the constants C depend on β_0 as $\exp(c/\beta_0)$, which means that these constants blow up for $\beta_0 \to 0+$. Error estimates uniform with respect to the diffusion coefficient will be proven in Section ?? in the case of a linear convection-diffusion problem. The case with a nonlinear convection and linear diffusion was analyzed recently in [?] in the case, when backward Euler time discretization was used.*

Chapter 5

Generalization of the DGM

The aim of this chapter is to present some advanced aspects and special techniques of the discontinuous Galerkin method. First, we present the *hp*-discontinuous Galerkin method. Then the DGM over nonstandard nonsimplicial meshes will be treated. Finally, the effect of numerical integration in the DGM will be analyzed in the case of a nonstationary convection-diffusion problem with nonlinear convection.

5.1 *hp*-discontinuous Galerkin method

Since the DGM is based on discontinuous piecewise polynomial approximations, it is possible to use different polynomial degrees on different elements in a simple way. Then we speak of the *hp*-discontinuous Galerkin method (*hp*-DGM). A suitable adaptive mesh refinement combined with the choice of the polynomial approximation degrees, representing the *hp*-adaptation, can significantly increase the *efficiency* of the computational process. It allows us to achieve the given error tolerance with the aid of the low number of degrees of freedom. The origins of *hp* finite element methods date back to the pioneering work of Ivo Babuška et al., see the survey paper [BS94a]. Based on several theoretical works as, e.g., monographs [Sch98, Šol04] or papers [BS94a, DRD02, ŠD04], it is possible to expect that the error decreases to zero at an exponential rate with respect to the number of degrees of freedom.

We present here the analysis of error estimates for the *hp*-DGM in the case of a model of the Poisson boundary value problem. We underline the similarity and differences with analysis of the *h*-version of the DGM presented in Chapter 1. Mostly the same notation is used for several constants appearing also in Chapter 1, but some constants may have slightly different meaning. However, we suppose that there is no danger of misunderstanding. On the contrary, it helps us to adapt the techniques from Chapter 1 to this section.

The analysis of the *hp*-DGM can be directly extended to nonstationary convection-diffusion equations from Chapters 2 and **??**. See, e.g., [Dol08a, Hoz09, Hol10].

5.1.1 Formulation of a model problem

Similarly, as in Section 1.1, let Ω be a bounded polygonal or polyhedral domain in $\mathbb{R}^d$, $d = 2, 3$, with Lipschitz boundary $\partial\Omega$. We denote by $\partial\Omega_D$ and $\partial\Omega_N$ parts of the boundary $\partial\Omega$ such that $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, $\partial\Omega_D \cap \partial\Omega_N = \emptyset$ and $\mathrm{meas}_{d-1}(\partial\Omega_D) > 0$.

We consider the Poisson problem (1.1) to find a function $u : \Omega \to \mathbb{R}$ such that

$$-\Delta u = f \quad \text{in } \Omega, \tag{5.1a}$$

$$u = u_D \quad \text{on } \partial\Omega_D, \tag{5.1b}$$

$$\boldsymbol{n} \cdot \nabla u = g_N \quad \text{on } \partial\Omega_N, \tag{5.1c}$$

where f, u_D and g_N are given functions. The weak solution of problem (5.1) is given by Definition 1.1.

5.1.2 Discretization

In this section we introduce the *hp*-DGM numerical solution of problem (5.1). We start from the generalization of the function spaces defined in Chapter 1.

Function spaces

Let $\mathcal{T}_h$ ($h > 0$) be a triangulation of Ω. In the same way as in Chapter 1, by the symbols $\mathcal{F}_h, \mathcal{F}_h^I, \mathcal{F}_h^B, \mathcal{F}_h^D$ and $\mathcal{F}_h^{ID}$ we denote sets of faces of elements $K \in \mathcal{T}_h$. To each $K \in \mathcal{T}_h$, we assign a positive integer s_K – *local Sobolev index* and a positive integer

p_K – *local polynomial degree*. Then we define the sets

$$\boldsymbol{s} = \{s_K, K \in \mathcal{T}_h\}, \qquad \boldsymbol{p} = \{p_K, K \in \mathcal{T}_h\}. \tag{5.2}$$

Over the triangulation $\mathcal{T}_h$, we define (instead of (1.29)) the *broken Sobolev space* corresponding to the vector $\boldsymbol{s}$

$$H^{\boldsymbol{s}}(\Omega, \mathcal{T}_h) = \{v; v|_K \in H^{s_K}(K) \ \forall\, K \in \mathcal{T}_h\} \tag{5.3}$$

with the norm

$$\|v\|_{H^{\boldsymbol{s}}(\Omega,\mathcal{T}_h)} = \left(\sum_{K\in\mathcal{T}_h} \|v\|^2_{H^{s_K}(K)} \right)^{1/2} \tag{5.4}$$

and the seminorm

$$|v|_{H^{\boldsymbol{s}}(\Omega,\mathcal{T}_h)} = \left(\sum_{K\in\mathcal{T}_h} |v|^2_{H^{s_K}(K)} \right)^{1/2}, \tag{5.5}$$

where $\|\cdot\|_{H^{s_K}(K)}$ and $|\cdot|_{H^{s_K}(K)}$ denotes the norm and seminorm in the Sobolev space $H^{s_K}(K) = W^{s_K,2}(K)$, respectively. If $s_K = q \geq 1$ for all $K \in \mathcal{T}_h$, then we use the notation $H^q(\Omega, \mathcal{T}_h) = H^{\boldsymbol{s}}(\Omega, \mathcal{T}_h)$. Obviously,

$$H^{\bar{s}}(\Omega, \mathcal{T}_h) \subset H^{\boldsymbol{s}}(\Omega, \mathcal{T}_h) \subset H^{\underline{s}}(\Omega, \mathcal{T}_h), \tag{5.6}$$

where $\bar{s} = \max\{s_K,\ s_K \in \boldsymbol{s}\}$ and $\underline{s} = \min\{s_K,\ s_K \in \boldsymbol{s}\}$.

Furthermore, we define (instead of (1.34)) the space of discontinuous piecewise polynomial functions associated with the vector $\boldsymbol{p}$ by

$$S_{h\boldsymbol{p}} = \{v \in L^2(\Omega);\ v|_K \in P_{p_K}(K) \ \forall\, K \in \mathcal{T}_h\}, \tag{5.7}$$

where $P_{p_K}(K)$ denotes the space of all polynomials on K of degree $\leq p_K$. In the *hp*-error analysis we shall assume that there exists a constant $C_P \geq 1$ such that

$$\frac{p_K}{p_{K'}} \leq C_P \quad \forall\, K,\ K' \in \mathcal{T}_h \text{ such that } K \text{ and } K' \text{ are neighbours.} \tag{5.8}$$

Assumption (5.8) may seem rather restrictive. However, it appears that the application of the *hp*-methods to practical problems is efficient and accurate, if the polynomial degrees of approximation on neighbouring elements do not differ too much.

hp-variant of the penalty parameter

In Section 1.6.1 we introduced the penalty parameter $\sigma : \cup_{\Gamma\in\mathcal{F}_h^{ID}} \to \mathbb{R}$, which was proportional to $\mathrm{diam}(\Gamma)^{-1} \sim h_K^{-1}$ where $\Gamma \subset \partial K$, $\Gamma \in \mathcal{F}_h^{ID}$. However, the following numerical analysis shows that for the *hp*-DGM, the penalty parameter σ has to depend also on the degree of the polynomial approximation (see also [HRS05]). To this end, for each $K \in \mathcal{T}_h$ we define the parameter

$$d(K) = \frac{h_K}{p_K^2}, \quad K \in \mathcal{T}_h. \tag{5.9}$$

Now for each $\Gamma \in \mathcal{F}_h^{ID}$ we introduce the *hp*-analogue to the quantity h_Γ from Section 1.6.1, which is now denoted by $d(\Gamma)$. In the theoretical analysis, we require that the quantity $d(\Gamma),\ \Gamma \in \mathcal{F}_h,\ h \in (0, \bar{h})$, satisfies the *equivalence condition* with $d(K)$, i.e., there exist constants $C_T, C_G > 0$ independent of h, K and Γ such that

$$C_T\, d(K) \leq d(\Gamma) \leq C_G\, d(K), \quad K \in \mathcal{T}_h,\ \Gamma \in \mathcal{F}_h,\ \Gamma \subset \partial K. \tag{5.10}$$

Let $K_\Gamma^{(L)}$ and $K_\Gamma^{(R)}$ be the neighbouring elements sharing the face $\Gamma \in \mathcal{F}_h^I$. There are several possibilities how to define the parameter $d(\Gamma)$ for all interior faces $\Gamma \in \mathcal{F}_h^I$:

(i)

$$d(\Gamma) = \frac{2\,\mathrm{diam}(\Gamma)}{(p_{K_\Gamma^{(L)}})^2 + (p_{K_\Gamma^{(R)}})^2}, \quad \Gamma \in \mathcal{F}_h^I, \tag{5.11}$$

(ii)

$$d(\Gamma) = \max(d(K_\Gamma^{(L)}), d(K_\Gamma^{(R)})), \quad \Gamma \in \mathcal{F}_h^I, \tag{5.12}$$

(iii)

$$d(\Gamma) = \min(d(K_\Gamma^{(L)}), d(K_\Gamma^{(R)})), \quad \Gamma \in \mathcal{F}_h^I. \tag{5.13}$$

Moreover, for the boundary faces $\Gamma \in \mathcal{F}_h^D$, we put

$$d(\Gamma) = d(K_\Gamma^{(L)}), \tag{5.14}$$

where $K_\Gamma^{(L)}$ is the element adjacent to Γ.

In the sequel we consider a system $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ of triangulations of the domain Ω satisfying the shape-regularity assumption (1.19), i.e.,

$$\frac{h_K}{\rho_K} \le C_R, \quad K \in \mathcal{T}_h, \quad h \in (0,\bar{h}). \tag{5.15}$$

The following lemma characterizes the mesh assumptions and the choices of $d(\Gamma)$, which guarantees the equivalence condition (5.10).

Lemma 5.1. *Let $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ be a system of triangulations of the domain Ω satisfying assumption (5.15). Moreover, let $\boldsymbol{p}$ be the polynomial degree vector given by (5.2), satisfying assumption (5.8). Then condition (5.10) is satisfied in the following cases:*

(a) The triangulations $\mathcal{T}_h$, $h \in (0,\bar{h})$, are conforming (i.e., assumption (MA4) from Section 1.3.2 is satisfied) and $d(\Gamma)$ is defined by (5.11) or (5.12) or (5.13).

(b) The triangulations $\mathcal{T}_h$, $h \in (0,\bar{h})$, are, in general, nonconforming, assumption (A2) (i.e., (1.22) is satisfied and $d(\Gamma)$ is defined by (5.11).

(c) The triangulations $\mathcal{T}_h$, $h \in (0,\bar{h})$, are, in general, nonconforming, assumption (A1) is satisfied (i.e., the system $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ is locally quasi-uniform) and $d(\Gamma)$ is defined by (5.12) or (5.13).

Exercise 5.2. *Prove the above lemma and determine the constants C_T and C_G.*

Remark 5.3. *If $p_K = p \in \mathbb{N}$ for all $K \in \mathcal{T}_h$, then the constants C_T and C_G from (5.10) are identical with the constants from (1.20).*

Approximate solution

Now we are ready to introduce the *hp*-DGM approximate solution. Using the same process as in Chapter 1, we arrive at the definition of the following forms. For $u, v \in H^{\boldsymbol{s}}(\Omega,\mathcal{T}_h)$, where $s_K \ge 2$ for all $K \in \mathcal{T}_h$, we put

$$a_h(u,v) = \sum_{K\in\mathcal{T}_h} \int_K \nabla u \cdot \nabla v \,\mathrm{d}x - \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \left(\langle\nabla u\rangle \cdot \boldsymbol{n}[v] + \Theta\langle\nabla v\rangle \cdot \boldsymbol{n}[u]\right) \mathrm{d}S, \tag{5.16}$$

$$J_h^\sigma(u,v) = \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \sigma[u]\,[v]\,\mathrm{d}S, \tag{5.17}$$

$$\ell_h(v) = \int_\Omega g v \,\mathrm{d}x - \Theta \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma u_D \left(\nabla v \cdot \boldsymbol{n}\right) \mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \sigma u_D\, v \,\mathrm{d}S + \int_{\partial\Omega_N} g_N v \,\mathrm{d}S, \tag{5.18}$$

where the *penalty parameter* σ is given by

$$\sigma|_\Gamma = \sigma_\Gamma = \frac{C_W}{d(\Gamma)}, \quad \Gamma \in \mathcal{F}_h^{ID}, \tag{5.19}$$

with $d(\Gamma)$ introduced in (5.11) – (5.14), and a suitable constant $C_W > 0$. In contrast to the penalty parameter σ defined in Section 1.6.1, we have $\sigma|_\Gamma \sim p^2 h^{-1}$, where h and p correspond to the diameter of Γ and the degree of the polynomial approximation, respectively, in the vicinity of Γ.

Similarly as in Section 1.4, for $\Theta = -1$, $\Theta = 0$ and $\Theta = 1$ the form a_h (together with the form J_h^σ) represents the nonsymmetric variant (NIPG), incomplete variant (IIPG) and symmetric variant (SIPG), respectively, of the approximation of the diffusion term. Moreover, we put

$$A_h(u,v) = a_h(u,v) + J_h^\sigma(u,v), \quad u, v \in H^{\boldsymbol{s}}(\Omega,\mathcal{T}_h). \tag{5.20}$$

Now we shall define an approximate solution of problem (5.1).

Definition 5.4. *A function $u_h \in S_{h\boldsymbol{p}}$ is called an hp-*DG approximate solution *of problem (5.1), if it satisfies the identity*

$$A_h(u_h, v_h) = \ell_h(v_h) \quad \forall\, v_h \in S_{h\boldsymbol{p}}. \tag{5.21}$$

From the construction of the forms A_h and ℓ_h one can see that the exact solution $u \in H^2(\Omega)$ of problem (5.1) satisfies the identity

$$A_h(u, v) = \ell_h(v) \quad \forall\, v \in H^2(\Omega, \mathcal{T}_h), \tag{5.22}$$

which represents the *consistency* of the method. Identities (5.21) and (5.22) imply the *Galerkin orthogonality* of the error $e_h = u_h - u$ of the method:

$$A_h(e_h, v_h) = 0 \quad \forall\, v_h \in S_{h\boldsymbol{p}}, \tag{5.23}$$

which will be used in the analysis of error estimates. (Compare with (1.57).)

5.1.3 Theoretical analysis

This section is devoted to the error analysis of the *hp*-DGM introduced above. Namely, an error estimate in the analogue to the DG-norm introduced by (1.103) will be derived. We follow the analysis of the abstract method from Section 1.2 and present several "*hp*-variants" of results from Chapter 1. We use the same notation for constants, although they attain different values in Chapter 1 and Section 5.1.3.

Auxiliary results

Similarly as in Section 1.5, the numerical analysis is based on three fundamental results: the multiplicative trace inequality, the inverse inequality and the approximation properties.

The *multiplicative trace inequality* presented in Lemma 1.19 remains the same. This means that under the shape-regularity assumption (5.15), there exists a constant $C_M > 0$ independent of v, h and K such that

$$\|v\|^2_{L^2(\partial K)} \le C_M \left(\|v\|_{L^2(K)} \, |v|_{H^1(K)} + h_K^{-1} \|v\|^2_{L^2(K)} \right), \tag{5.24}$$
$$K \in \mathcal{T}_h,\ v \in H^1(K),\ h \in (0, \bar{h}).$$

The proof of Lemma 1.21 gives us the *hp*-version of the *inverse inequality*: Let the shape-regularity assumption (5.15) be satisfied. Then there exists a constant $C_I > 0$ independent of v, h, p_K, and K such that

$$|v|_{H^1(K)} \le C_I p_K^2 h_K^{-1} \|v\|_{L^2(K)}, \quad v \in P_{p_K}(K),\ K \in \mathcal{T}_h,\ h \in (0, \bar{h}). \tag{5.25}$$

Finally, we introduce the *hp*-version of approximation properties of spaces $S_{h\boldsymbol{p}}$. We present the results from [BS87]. Since the proof is very technical, we skip it and refer to the original work.

Lemma 5.5 (Approximation properties)**.** *There exists a constant $C_A > 0$ independent of v, h, K and p_K and a mapping $\pi^K_{p_K} : H^{s_K}(K) \to P_{p_K}(K)$, $s_K \ge 1$, such that the inequality*

$$\|\pi^K_{p_K} v - v\|_{H^q(K)} \le C_A \frac{h_K^{\mu_K - q}}{p_K^{s_K - q}} \|v\|_{H^{s_K}(K)} \tag{5.26}$$

holds for all $v \in H^{s_K}(K)$, $K \in \mathcal{T}_h$ and $h \in (0, \bar{h})$ with $\mu_K = \min(p_K + 1, s_K)$, $0 \le q \le s_K$,

Proof. See Lemma 4.5 in [BS87] for the case $d = 2$. If $d = 3$, the arguments are analogous. □ □

Definition 5.6. *Let $\boldsymbol{s}$ and $\boldsymbol{p}$ be the vectors introduced in* (5.2)*. We define the mapping $\Pi_{h\boldsymbol{p}} : H^{\boldsymbol{s}}(\Omega, \mathcal{T}_h) \to S_{h\boldsymbol{p}}$ by*

$$(\Pi_{h\boldsymbol{p}} u)\,|_K = \pi^K_{p_K}\,(u|_K) \quad \forall\, K \in \mathcal{T}_h, \tag{5.27}$$

where $\pi^K_{p_K} : H^{s_K}(K) \to P_{p_K}(K)$ is the mapping introduced in Lemma 5.5.

Lemma 5.7. *Let $\boldsymbol{s}$ and $\boldsymbol{p}$ be the vectors introduced in* (5.2) *and $\Pi_{h\boldsymbol{p}} : H^{\boldsymbol{s}}(\Omega, \mathcal{T}_h) \to S_{h\boldsymbol{p}}$ the corresponding mapping defined by* (5.27)*. If $v \in H^{\boldsymbol{s}}(\Omega, \mathcal{T}_h)$, then*

$$\|\Pi_{h\boldsymbol{p}} v - v\|^2_{H^q(\Omega, \mathcal{T}_h)} \le C_A^2 \sum_{K \in \mathcal{T}_h} \frac{h_K^{2(\mu_K - q)}}{p_K^{2s_K - 2q}} \|v\|^2_{H^{s_K}(K)}, \tag{5.28}$$

where $\mu_K = \min(p_K + 1, s_K)$, $K \in \mathcal{T}_h$ and $0 \le q \le \min_{s_K \in \boldsymbol{s}} s_K$ and C_A is the constant from Lemma 5.5.

Proof. Using definition (5.27) and the approximation properties (5.26), we obtain (5.28). □ □

Moreover, using the previous results, we prove some technical inequalities analogous to Lemma 1.27.

Lemma 5.8. *Let* (5.10) *be valid and let* σ *be defined by* (5.19). *Then for each* $v \in H^1(\Omega, \mathcal{T}_h)$ *we have*

$$\sum_{\Gamma \in \mathcal{F}_h^{ID}} d(\Gamma)^{-1} \int_\Gamma [v]^2 \,\mathrm{d}S \le \frac{2}{C_T} \sum_{K \in \mathcal{T}_h} d(K)^{-1} \int_{\partial K} |v|^2 \,\mathrm{d}S, \tag{5.29}$$

$$\sum_{\Gamma \in \mathcal{F}_h^{ID}} d(\Gamma) \int_\Gamma \langle v \rangle^2 \,\mathrm{d}S \le C_G \sum_{K \in \mathcal{T}_h} d(K) \int_{\partial K} |v|^2 \,\mathrm{d}S. \tag{5.30}$$

Hence,

$$\sum_{\Gamma \in \mathcal{F}_h^{ID}} \sigma_\Gamma \|[v]\|^2_{L^2(\Gamma)} \le \frac{2C_W}{C_T} \sum_{K \in \mathcal{T}_h} d(K)^{-1} \|v\|^2_{L^2(\partial K)}, \tag{5.31}$$

$$\sum_{\Gamma \in \mathcal{F}_h^{ID}} \frac{1}{\sigma_\Gamma} \|\langle v \rangle\|^2_{L^2(\Gamma)} \le \frac{C_G}{C_W} \sum_{K \in \mathcal{T}_h} d(K) \|v\|^2_{L^2(\partial K)}, \tag{5.32}$$

where the penalty parameter σ *is given by* (5.19).

Proof. (a) By definition (1.32), (1.33), inequality (1.110) and assumption (5.10), we have

$$\begin{aligned}
&\sum_{\Gamma \in \mathcal{F}_h^{ID}} d(\Gamma)^{-1} \int_\Gamma [v]^2 \,\mathrm{d}S \\
&= \sum_{\Gamma \in \mathcal{F}_h^{I}} d(\Gamma)^{-1} \int_\Gamma \left| v_\Gamma^{(L)} - v_\Gamma^{(R)} \right|^2 \mathrm{d}S + \sum_{\Gamma \in \mathcal{F}_h^{D}} d(\Gamma)^{-1} \int_\Gamma \left| v_\Gamma^{(L)} \right|^2 \mathrm{d}S \\
&\le 2 \sum_{\Gamma \in \mathcal{F}_h^{I}} d(\Gamma)^{-1} \int_\Gamma \left(\left| v_\Gamma^{(L)} \right|^2 + \left| v_\Gamma^{(R)} \right|^2 \right) \mathrm{d}S + \sum_{\Gamma \in \mathcal{F}_h^{D}} d(\Gamma)^{-1} \int_\Gamma \left| v_\Gamma^{(L)} \right|^2 \mathrm{d}S \\
&\le 2C_T^{-1} \sum_{\Gamma \in \mathcal{F}_h^{ID}} d(K_\Gamma^{(L)})^{-1} \int_\Gamma \left| v_\Gamma^{(L)} \right|^2 \mathrm{d}S + 2C_T^{-1} \sum_{\Gamma \in \mathcal{F}_h^{I}} d(K_\Gamma^{(R)})^{-1} \int_\Gamma \left| v_\Gamma^{(R)} \right|^2 \mathrm{d}S \\
&\le 2C_T^{-1} \sum_{K \in \mathcal{T}_h} d(K)^{-1} \int_{\partial K} |v|^2 \,\mathrm{d}S,
\end{aligned}$$

which proves (5.29). Moreover, using (5.19) we immediately obtain (5.31).

(b) In the proof of (5.30), we proceed in a similar way, using (1.32), (5.10) and (1.110). Inequality (5.32) is a direct consequence of (5.30) and (5.19). □ □

Analogously to Lemma 1.32, we present its *hp*-variant.

Lemma 5.9. *Let* $v \in H^1(\Omega, \mathcal{T}_h)$. *Then*

$$\begin{aligned}
J_h^\sigma(v,v) &\le \frac{2C_W C_M}{C_T} \sum_{K \in \mathcal{T}_h} \left(\frac{p_K^2}{h_K^2} \|v\|^2_{L^2(K)} + \frac{p_K^2}{h_K} \|v\|_{L^2(K)} |v|_{H^1(K)} \right) \\
&\le \frac{C_W C_M}{C_T} \sum_{K \in \mathcal{T}_h} \left(\frac{2p_K^2}{h_K^2} \|v\|^2_{L^2(K)} + \frac{p_K^3}{h_K^2} \|v\|^2_{L^2(K)} + p_K |v|^2_{H^1(K)} \right).
\end{aligned} \tag{5.33}$$

Proof. If $v \in H^1(\Omega, \mathcal{T}_h)$, then the definition (5.17) of the form J_h^σ, (5.31) and (5.9) imply that

$$\begin{aligned}
J_h^\sigma(v,v) &= \sum_{\Gamma \in \mathcal{F}_h^{ID}} \int_\Gamma \sigma [v]^2 \,\mathrm{d}S = \sum_{\Gamma \in \mathcal{F}_h^{ID}} \sigma_\Gamma \|[v]^2\|_{L^2(\Gamma)} \\
&\le \frac{2C_W}{C_T} \sum_{K \in \mathcal{T}_h} d(K)^{-1} \|v\|^2_{L^2(\partial K)} = \frac{2C_W}{C_T} \sum_{K \in \mathcal{T}_h} \frac{p_K^2}{h_K} \|v\|^2_{L^2(\partial K)}.
\end{aligned}$$

Now, using the multiplicative trace inequality (5.24), we get

$$J_h^\sigma(v,v) \le \frac{2C_W C_M}{C_T} \sum_{K\in\mathcal{T}_h} \left(\frac{p_K^2}{h_K^2}\|v\|_{L^2(K)}^2 + \frac{p_K^2}{h_K}\|v\|_{L^2(K)}|v|_{H^1(K)} \right),$$

which gives the first inequality in (5.33). Moreover, the application of Young's inequality yields the second one. □ □

Finally, we introduce the *hp*-variant of Lemma 1.34.

Lemma 5.10. *Under assumptions* (5.15) *and* (5.10)*, for any* $v \in H^2(\Omega,\mathcal{T}_h)$ *the following estimate holds:*

$$\begin{aligned}\sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v\rangle)^2\,\mathrm{d}S &\le \frac{C_G C_M}{C_W} \sum_{K\in\mathcal{T}_h} \frac{h_K}{p_K^2}\left(|v|_{H^1(K)}\,|v|_{H^2(K)} + h_K^{-1}|v|_{H^1(K)}^2 \right)\\ &\le \frac{C_G C_M}{2C_W} \sum_{K\in\mathcal{T}_h} \left(\frac{3}{p_K^2}|v|_{H^1(K)}^2 + \frac{h_K^2}{p_K^2}|v|_{H^2(K)}^2 \right). \end{aligned} \tag{5.34}$$

Moreover, for $v_h \in S_{h\boldsymbol{p}}$ *we have*

$$\sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v_h\rangle)^2\,\mathrm{d}S \le \frac{C_G C_M}{C_W}(C_I+1)|v_h|_{H^1(\Omega,\mathcal{T}_h)}^2. \tag{5.35}$$

Proof. Using (5.32), the multiplicative trace inequality (5.24) and notation (5.9), we find that

$$\begin{aligned}\sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v\rangle)^2\,\mathrm{d}S &\le \frac{C_G}{C_W}\sum_{K\in\mathcal{T}_h} d(K)\|\nabla v\|_{L^2(\partial K)}^2\\ &\le \frac{C_G C_M}{C_W}\sum_{K\in\mathcal{T}_h}\frac{h_K}{p_K^2}\left(\|\nabla v\|_{L^2(K)}\,|\nabla v|_{H^1(K)} + h_K^{-1}\|\nabla v\|_{L^2(K)}^2\right),\\ &= \frac{C_G C_M}{C_W}\sum_{K\in\mathcal{T}_h}\frac{h_K}{p_K^2}\left(|v|_{H^1(K)}\,|v|_{H^2(K)} + h_K^{-1}|v|_{H^1(K)}^2\right),\end{aligned}$$

which is the first inequality in (5.34). The second one is obtained by the application of Young's inequality.

Further, for $v_h \in S_{h\boldsymbol{p}}$, estimate (5.34), the inverse inequality (5.25) and the inequality $1/p_K^2 \le 1$ give

$$\begin{aligned}&\sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v_h\rangle)^2\,\mathrm{d}S\\ &\quad\le \frac{C_G C_M}{C_W}\sum_{K\in\mathcal{T}_h}\frac{h_K}{p_K^2}\left(\|\nabla v_h\|_{L^2(K)}\,|\nabla v_h|_{H^1(K)} + h_K^{-1}\|\nabla v_h\|_{L^2(K)}^2\right),\\ &\quad\le \frac{C_G C_M}{C_W}\sum_{K\in\mathcal{T}_h}\frac{h_K}{p_K^2}\left(C_I p_K^2 h_K^{-1}\|\nabla v_h\|_{L^2(K)}\,\|\nabla v_h\|_{L^2(K)} + h_K^{-1}\|\nabla v_h\|_{L^2(K)}^2\right),\\ &\quad\le \frac{C_G C_M}{C_W}(C_I+1)\sum_{K\in\mathcal{T}_h}\|\nabla v_h\|_{L^2(K)}^2 = \frac{C_G C_M}{C_W}(C_I+1)|v_h|_{H^1(\Omega,\mathcal{T}_h)}^2,\end{aligned}$$

which implies (5.35). □ □

Continuity of the bilinear forms

Now, we prove the continuity of the bilinear form A_h defined by (5.20). In the space S_{hp} we again employ the DG-norm

$$|||u||| = \left(|u|_{H^1(\Omega,\mathcal{T}_h)}^2 + J_h^\sigma(u,u)\right)^{1/2}. \tag{5.36}$$

Comparing (5.36) with (1.103), both relations are formally identical. However, the norm in (5.36) is p-dependent, because σ depends on the polynomial degrees $p_K,\ K\in\mathcal{T}_h$.

Exercise 5.11. *Prove that* $|||\cdot|||$ *is a norm in the spaces* $H^{\boldsymbol{s}}(\Omega,\mathcal{T}_h)$ *and* $S_{h\boldsymbol{p}}$.

Furthermore, due to (1.122), we have

$$|A_h(u,v)| \le 2\|u\|_{1,\sigma}\|v\|_{1,\sigma} \quad \forall\, u,v \in H^2(\Omega,\mathcal{T}_h), \tag{5.37}$$

where

$$\begin{aligned}\|v\|_{1,\sigma}^2 =& |||v|||^2 + \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v\rangle)^2\,\mathrm{d}S \\ =& |v|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(v,v) + \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla v\rangle)^2\mathrm{d}S.\end{aligned} \tag{5.38}$$

Now, we derive the *hp*-estimate of the $\|\cdot\|_{1,\sigma}$-norm, compare with Lemma 1.35.

Lemma 5.12. *Let* (5.10) *be valid and let* σ *be defined by* (5.19)*. Then, there exist constants* C_σ, $\tilde{C}_\sigma > 0$ *such that*

$$J_h^\sigma(u,u)^{1/2} \le |||u||| \le \|u\|_{1,\sigma} \le C_\sigma\, R_a(u) \quad \forall\, u \in H^2(\Omega,\mathcal{T}_h),\ h\in(0,\bar{h}), \tag{5.39}$$

$$J_h^\sigma(v_h,v_h)^{1/2} \le |||v_h||| \le \|v_h\|_{1,\sigma} \le \tilde{C}_\sigma |||v_h||| \quad \forall\, v_h \in S_{h\boldsymbol{p}},\ h\in(0,\bar{h}), \tag{5.40}$$

where

$$R_a(u) = \left(\sum_{K\in\mathcal{T}_h}\left(\frac{p_K^3}{h_K^2}\|u\|^2_{L^2(K)} + p_K|u|^2_{H^1(K)} + \frac{h_K^2}{p_K^2}|u|^2_{H^2(K)}\right)\right)^{1/2},\ u\in H^2(\Omega,\mathcal{T}_h). \tag{5.41}$$

Proof. The first two inequalities in (5.39) as well as in (5.40) follow immediately from the definition of the DG-norm (5.36) and $\|\cdot\|_{1,\sigma}$-norm (5.38). Moreover, in view of (5.38), (5.4), (5.33) and (5.34), for $u\in H^2(\Omega,\mathcal{T}_h)$, we have

$$\begin{aligned}\|u\|^2_{1,\sigma} &= |u|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(u,u) + \sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma \sigma^{-1}(\boldsymbol{n}\cdot\langle\nabla u\rangle)^2\mathrm{d}S \\ &\le \sum_{K\in\mathcal{T}_h}|u|^2_{H^1(K)} + \frac{C_W C_M}{C_T}\sum_{K\in\mathcal{T}_h}\left(\frac{2p_K^2}{h_K^2}\|u\|^2_{L^2(K)} + \frac{p_K^3}{h_K^2}\|u\|^2_{L^2(K)} + p_K|u|^2_{H^1(K)}\right) \\ &\quad + \frac{C_G C_M}{2C_W}\sum_{K\in\mathcal{T}_h}\left(\frac{3}{p_K^2}|u|^2_{H^1(K)} + \frac{h_K^2}{p_K^2}|u|^2_{H^2(K)}\right).\end{aligned}$$

Now, using the inequalities $p_k \ge 1$ and $1/p_K \le 1$, we get

$$\begin{aligned}\|u\|^2_{1,\sigma} \le \sum_{K\in\mathcal{T}_h}\Bigg(&\Big(1 + \frac{3C_G\,C_M}{2C_W} + \frac{C_W\,C_M}{C_T}\Big)p_K|u|^2_{H^1(K)} \\ &+ \frac{C_G\,C_M}{2C_W}\,\frac{h_K^2}{p_K^2}\,|u|^2_{H^2(K)} + \frac{3C_W\,C_M}{C_T}\frac{p_K^3}{h_k^2}\,\|u\|^2_{L^2(K)}\Bigg).\end{aligned}$$

Hence, (5.39) holds with

$$C_\sigma = \left(\max\left(1 + \frac{3C_G\,C_M}{2C_W} + \frac{C_W\,C_M}{C_T},\ \frac{C_G\,C_M}{2C_W},\ \frac{3C_W\,C_M}{C_T}\right)\right)^{1/2}.$$

Further, if $v_h \in S_{h\boldsymbol{p}}$, then (5.38) and (5.35) immediately imply (5.40) with $\tilde{C}_\sigma = (1 + C_G\,C_M(C_I+1)/C_W)^{1/2}$. □ □

Lemma 5.12 directly implies the continuity of the form A_h:

Corollary 5.13. *Let* (5.10) *be valid and let* σ *be defined by* (5.19)*. Then there exist constants* $C_B > 0$ *and* $\tilde{C}_B > 0$ *such that the forms* A_h *defined by* (5.20) *satisfies the estimates*

$$|A_h(u_h,v_h)| \le C_B|||u_h|||\,|||v_h||| \quad \forall\, u_h, v_h \in S_{hp}, \tag{5.42}$$

$$|A_h(u,v_h)| \le \tilde{C}_B\, R_a(u)\,|||v_h||| \quad \forall\, u\in H^2(\Omega,\mathcal{T}_h)\ \forall\, v_h\in S_{hp}\ \forall\, h(0,\bar{h}), \tag{5.43}$$

where R_a *is defined by* (5.41)*.*

Proof. Estimates (5.37), (5.39) and (5.40) give (5.42) with $C_B = 2\tilde{C}_\sigma^2$. Moreover, by (5.37) and (5.39),

$$|A_h(u,v_h)| \le 2\|u\|_{1,\sigma}\|v_h\|_{1,\sigma} \le 2C_\sigma\tilde{C}_\sigma R_a(u)|||v_h|||,$$

which is (5.43) with $\tilde{C}_B = 2C_\sigma\tilde{C}_\sigma$. □ □

Coercivity of the bilinear forms

In order to derive error estimates of the approximate solution (5.21), we need the coercivity of the form A_h. To this end, we shall present here the generalization of the results from Section 1.6.3.

Lemma 5.14 (NIPG coercivity)**.** *For any $C_W > 0$ the bilinear form A_h defined by* (5.20) *with $\Theta = -1$ in* (5.16) *satisfies the coercivity condition*

$$A_h(v,v) \geq |||v|||^2 \qquad \forall\, v \in H^2(\Omega, \mathcal{T}_h). \tag{5.44}$$

Proof. If $\Theta = -1$, then from (5.16) and (5.20) it immediately follows that

$$A_h(v,v) = a_h(v,v) + J_h^\sigma(v,v) = |v|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(v,v) = |||v|||^2, \tag{5.45}$$

which we wanted to prove. □ □

The proof of coercivity of the symmetric bilinear form A_h defined by (5.16) with $\Theta = 1$ is more complicated.

Lemma 5.15 (SIPG coercivity)**.** *Let assumptions* (5.15) *and* (5.10) *be satisfied, let*

$$C_W \geq 4C_G C_M(1 + C_I), \tag{5.46}$$

where C_M, C_I and C_G are the constants from (5.24), (5.25) *and* (5.10)*, respectively, and let the penalty parameter σ be given by* (5.19) *for all $\Gamma \in \mathcal{F}_h^{ID}$. Then the bilinear form A_h defined by* (5.20) *and* (5.16) *with $\Theta = 1$ satisfies the coercivity condition*

$$A_h(v_h, v_h) \geq \frac{1}{2}|||v_h|||^2 \quad \forall\, v_h \in S_{h\boldsymbol{p}},\ \forall\, h \in (0, \bar{h}).$$

Proof. Let $\delta > 0$. Then (5.17), (5.19), (5.16) with $\Theta = 1$ and the Cauchy and Young's inequalities imply that

$$\begin{aligned}
&a_h(v_h, v_h) \qquad\qquad (5.47)\\
&= |v_h|^2_{H^1(\Omega,\mathcal{T}_h)} - 2\sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \boldsymbol{n}\cdot\langle\nabla v_h\rangle[v_h]\mathrm{d}S\\
&\geq |v_h|^2_{H^1(\Omega,\mathcal{T}_h)} - 2\left\{\frac{1}{\delta}\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma d(\Gamma)(\boldsymbol{n}\cdot\langle\nabla v_h\rangle)^2\mathrm{d}S\right\}^{\frac{1}{2}}\left\{\delta\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma\frac{[v_h]^2}{d(\Gamma)}\mathrm{d}S\right\}^{\frac{1}{2}}\\
&\geq |v_h|^2_{H^1(\Omega,\mathcal{T}_h)} - \omega - \frac{\delta}{C_W}J_h^\sigma(v_h,v_h),
\end{aligned}$$

where

$$\omega = \frac{1}{\delta}\sum_{\Gamma\in\mathcal{F}_h^{ID}}\int_\Gamma d(\Gamma)|\langle\nabla v_h\rangle|^2\,\mathrm{d}S. \tag{5.48}$$

Further, from (5.9), assumption (5.10), inequality (5.30), the multiplicative trace inequality (5.24), the inverse inequality (5.25) and the inequality $p_K^{-2} \leq 1$, we get

$$\begin{aligned}
\omega &\leq \frac{C_G}{\delta}\sum_{K\in\mathcal{T}_h}\frac{h_K}{p_K^2}\|\nabla v_h\|^2_{L^2(\partial K)} \qquad (5.49)\\
&\leq \frac{C_G C_M}{\delta}\sum_{K\in\mathcal{T}_h}\frac{h_K}{p_K^2}\left(|v_h|_{H^1(K)}|\nabla v_h|_{H^1(K)} + h_K^{-1}|v_h|^2_{H^1(K)}\right)\\
&\leq \frac{C_G C_M}{\delta}\sum_{K\in\mathcal{T}_h}\frac{h_K}{p_K^2}\left(C_I p_K^2 h_K^{-1}|v_h|^2_{H^1(K)} + h_K^{-1}|v_h|^2_{H^1(K)}\right)\\
&\leq \frac{C_G C_M(1+C_I)}{\delta}|v_h|^2_{H^1(\Omega,\mathcal{T}_h)}.
\end{aligned}$$

Now let us choose

$$\delta = 2C_G C_M(1 + C_I). \tag{5.50}$$

Then it follows from (5.46) and (5.47) – (5.50) that

$$\begin{aligned} a_h(v_h, v_h) \geq & \frac{1}{2}\left(|v_h|^2_{H^1(\Omega,\mathcal{T}_h)} - \frac{4C_G C_M(1+C_I)}{C_W} J_h^\sigma(v_h, v_h)\right) \\ \geq & \frac{1}{2}\left(|v_h|^2_{H^1(\Omega,\mathcal{T}_h)} - J_h^\sigma(v_h, v_h)\right). \end{aligned} \tag{5.51}$$

Finally, from the definition (5.20) of the form A_h and from (5.51) we have

$$\begin{aligned} A_h(v_h, v_h) = & a_h(v_h, v_h) + J_h^\sigma(v_h, v_h) \\ \geq & \frac{1}{2}\left(|v_h|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(v_h, v_h)\right) = \frac{1}{2}|||v_h|||^2, \end{aligned} \tag{5.52}$$

which we wanted to prove. □ □

Lemma 5.16 (IIPG coercivity)**.** *Let assumptions* (5.15) *and* (5.10) *be satisfied, let*

$$C_W \geq C_G C_M(1+C_I), \tag{5.53}$$

where C_M, C_I and C_G are constants from (5.24), (5.25) *and* (5.10)*, respectively, and let the penalty parameter σ be given by* (5.19) *for all $\Gamma \in \mathcal{F}_h^{ID}$. Then the bilinear form A_h defined by* (5.20) *and* (5.16) *with $\Theta = 0$ satisfies the coercivity condition*

$$A_h(v_h, v_h) \geq \frac{1}{2}|||v_h|||^2 \quad \forall\, v_h \in S_{h\boldsymbol{p}}.$$

Proof. The proof is almost identical with the proof of the previous lemma. □ □

Corollary 5.17. *We can summarize the above results in the following way. We have*

$$A_h(v_h, v_h) \geq C_C |||v_h|||^2 \quad \forall\, v_h \in S_{h\boldsymbol{p}}, \tag{5.54}$$

with

$$\begin{array}{llll} C_C & = 1 & \text{for } \Theta = -1, & \text{if } C_W > 0, \\ C_C & = 1/2 & \text{for } \Theta = 1, & \text{if } C_W \geq 4C_G C_M(1+C_I), \\ C_C & = 1/2 & \text{for } \Theta = 0, & \text{if } C_W \geq C_G C_M(1+C_I). \end{array}$$

Corollary 5.18. *By virtue of Corollary 0.7, the coercivity of the form A_h implies the existence and uniqueness of the solution of the discrete problem.*

Error estimates in the DG-norm

In this section we will be concerned with the derivation of the error estimates of the *hp*-discontinuous Galerkin method (5.21). Let u and u_h denote the exact solution of problem (5.1) and the approximate solution obtained by method (5.21), respectively. The error $e_h = u_h - u$ can be written in the form

$$e_h = \xi + \eta, \quad \text{with } \xi = u_h - \Pi_{h\boldsymbol{p}} u \in S_{h\boldsymbol{p}},\ \eta = \Pi_{h\boldsymbol{p}} u - u, \tag{5.55}$$

where $\Pi_{h\boldsymbol{p}}$ is the $S_{h\boldsymbol{p}}$-interpolation defined by (5.27). The estimation of the error e_h will be carried out in several steps.

We suppose that the system of triangulations $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ satisfies the shape-regularity assumption (5.15) and that the relations (5.10) between $d(\Gamma)$ and $d(K)$ are valid.

First, we prove the *abstract error estimate*, representing a bound of the error in terms of the S_{hp}-interpolation error η, cf. Theorem 1.43.

Theorem 5.19. *Let* (5.10) *be valid, let σ be defined by* (5.19) *and let the exact solution of problem* (5.1) *satisfy the condition $u \in H^2(\Omega)$. Then there exists a constant $C_{\mathrm{AE}} > 0$ such that*

$$|||e_h||| \leq C_{\mathrm{AE}}\, R_a(\eta) = C_{\mathrm{AE}}\, R_a(\Pi_{hp} u - u) \quad \forall\, h \in (0, \bar{h}), \tag{5.56}$$

where $R_a(u)$ is given by (5.41).

Proof. The proof is completely identical with the proof of Theorem 1.43. We obtain again $C_{\mathrm{AE}} = C_\sigma + \tilde{C}_B/C_C$, where C_σ and $\tilde{C}_B$ and C_C are constants from (5.39) and (5.43) and (5.54). □ □

The abstract error estimate is the basis for the estimation of the error e_h in terms of the mesh size h.

Theorem 5.20 (DG-norm error estimate)**.** *Let $\{\mathcal{T}_h\}_{h\in(0,\bar{h})}$ be a system of triangulations of the domain Ω satisfying the shape-regularity assumption* (5.15)*. Let $\boldsymbol{s}$ and $\boldsymbol{p}$ be the vectors* (5.2) *such that $s_K \geq 2$, $p_K \geq 1$ and $\mu_K = \min(p_K + 1, s_K)$ for each $K \in \mathcal{T}_h$. Let the condition of equivalence* (5.10) *between $d(\Gamma)$ and $d(K)$ be valid (cf. Lemma 5.1). Let u be the solution of problem* (5.1) *such that $u \in H^2(\Omega) \cap H^{\boldsymbol{s}}(\Omega, \mathcal{T}_h)$ for any $h \in (0, \bar{h})$. Moreover, let the penalty constant C_W satisfy the conditions from Corollary 5.17. Let $u_h \in S_{h\boldsymbol{p}}$ be the approximate solution obtained by means of method* (5.21)*. Then the error $e_h = u_h - u$ satisfies the estimate*

$$|||e_h||| \leq \tilde{C} \left(\sum_{K\in\mathcal{T}_h} \frac{h_K^{2(\mu_K-1)}}{p_K^{2s_K-3}} \|u\|^2_{H^{s_K}(K)} \right)^{\frac{1}{2}}, \quad h \in (0, \bar{h}), \tag{5.57}$$

where $\tilde{C}$ is a constant independent of h and $\boldsymbol{p}$.

Proof. It is enough to use the abstract error estimate (5.56), where the expressions $|\eta|_{H^1(K)}$, $|\eta|_{H^2(K)}$ and $\|\eta\|_{L^2(K)}$, $K \in \mathcal{T}_h$, are estimated on the basis of the approximation properties (5.26), rewritten for $\eta|_K = (\Pi_{hp}u - u)|_K = \pi_{K,p}(u|_K) - u|_K$ and $K \in \mathcal{T}_h$:

$$\begin{aligned} \|\eta\|_{L^2(K)} \leq & C_A \frac{h_K^{\mu_K}}{p_K^{s_K}} \|u\|_{H^\mu(K)}, \\ |\eta|_{H^1(K)} \leq & C_A \frac{h_K^{\mu_K-1}}{p_K^{s_K-1}} \|u\|_{H^\mu(K)}, \\ |\eta|_{H^2(K)} \leq & C_A \frac{h_K^{\mu_K-2}}{p_K^{s_K-2}} \|u\|_{H^{\mu_K}(K)}. \end{aligned} \tag{5.58}$$

The above, the definition (5.41) of the expression R_a and the inequalities $1/p_K^{2s-2} \leq 1/p_K^{2s-3}$, $p_K \geq 1$ imply

$$\begin{aligned} R_a(\eta)^2 = & \sum_{K\in\mathcal{T}_h} \left(\frac{p_K^3}{h_K^2} \|\eta\|^2_{L^2(K)} + p_K |\eta|^2_{H^1(K)} + \frac{h_K^2}{p_K^2} |\eta|^2_{H^2(K)} \right) \\ \leq & C_A^2 \sum_{K\in\mathcal{T}_h} \left(\frac{p_K^3}{h_K^2} \frac{h_K^{2\mu_K}}{p_K^{2s_K}} + p_K \frac{h_K^{2(\mu_K-1)}}{p_K^{2s_K-2}} + \frac{h_K^2}{p_K^2} \frac{h_K^{2\mu_K-4}}{p_K^{2s_K-4}} \right) \|u\|^2_{H^{\mu_K}(K)} \\ \leq & C_A^2 \sum_{K\in\mathcal{T}_h} \left(\frac{h_K^{2(\mu_K-1)}}{p_K^{2s_K-3}} + \frac{h_K^{2(\mu_K-1)}}{p_K^{2s_K-3}} + \frac{h_K^{2(\mu_K-1)}}{p_K^{2s_K-3}} \right) \|u\|^2_{H^{\mu_K}(K)} \\ = & 3C_A^2 \sum_{K\in\mathcal{T}_h} \frac{h_K^{2(\mu_K-1)}}{p_K^{2s_K-3}} \|u\|^2_{H^{\mu_K}(K)}. \end{aligned}$$

Together with (5.56) this gives (5.57) with the constant $\tilde{C} = \sqrt{3} C_{\mathrm{AE}}\, C_A$. □ □

Comparing error estimate (5.57) with the approximation property (5.28) with $q = 1$, we see that (5.57) is suboptimal with respect to the polynomial degrees p_K, $K \in \mathcal{T}_h$. This is caused by the presence of the interior penalty form J_h^σ , see the last two terms in the second inequality in (5.33), namely the terms

$$\frac{p_K^3}{h_K^2} \|v\|^2_{L^2(K)} + p_K |v|^2_{H^1(K)} = p_K \left(\frac{p_K^2}{h_K^2} \|v\|^2_{L^2(K)} + |v|^2_{H^1(K)} \right), \quad K \in \mathcal{T}_h.$$

The error estimates optimal with respect to p were derived in [GS05] using an augmented Sobolev space.

As for the analysis of further subjects concerned with the *hp*-DGM, we refer to several works, namely [HSW08], [HSW07] dealing with the *hp*-DGM for quasilinear elliptic problems, [Geo06], [GHH07] dealing with the *hp*-DGM on anisotropic meshes, [WFS03] proving the exponential rate of the convergence of the *hp*-DGM, [HSS02], [CCSS02] dealing with the *hp*-DGM for convection-diffusion problems and [Tos02], [SW03] analyzing the *hp*-DGM for the Stokes problem.

5.1.4 Computational performance of the *hp*-DGM

In the previous sections we analyzed the *hp*-DGM, where the mesh $\mathcal{T}_h$ and the approximation polynomial degrees p_K, $K \in \mathcal{T}_h$, were given in advance. In practice, the *hp*-DGM can be applied in the combination with an adaptive algorithm, where the size h_K of the elements $K \in \mathcal{T}_h$ as well as the polynomial degrees p_K on elements $K \in \mathcal{T}_h$ are adaptively determined. The aim of this section is to demonstrate the ability of the *hp*-DGM to deal with refined grids and with different polynomial degrees on different $K \in \mathcal{T}_h$. We present one numerical example showing the efficiency and a possible potential of the *hp*-DGM.

Mesh adaptation — an overview

Numerical examples presented in Section 1.8.2 show that if the exact solution of the given problem is not sufficiently regular, then the experimental order of convergence of the DGM is low for any polynomial approximation degree. Therefore, a high number of *degrees of freedom* (DOF) (=$\dim S_{hp}$) has to be used in order to achieve a given accuracy. A significant reduction of the number of DOF can be achieved by a local mesh refinement of the given grid $\mathcal{T}_h$, in which we look for elements $K \in \mathcal{T}_h$, for which the computational error is too large. Then these marked elements are refined. In practice, for each element $K \in \mathcal{T}_h$ we define an *error estimator* η_K such that

$$\|u - u_h\|_K \approx \eta_K, \tag{5.59}$$

where $\|\cdot\|_K$ denotes a suitable norm of functions defined on $K \in \mathcal{T}_h$. The elements, where η_K is larger than a prescribed tolerance, are split into several daughter elements. E.g., for $d = 2$, by connecting the mid points of edges of the triangle marked for refinement, new four daughter triangles arise in place of the original one. This refinement strategy leads to hanging nodes, see Section 1.3.1. Figure 5.2 shows a sequence of adaptively refined triangular grids.

There exist a number of works dealing with strategies for the error estimation and the corresponding mesh adaptive techniques. Since a posteriori error analysis and mesh adaption are out of the scope of this book, we refer only to [EEHJ95], where an introduction to adaptive methods for partial differential equations can be found. Moreover, an overview of standard approaches was presented in [Ver96], [Ver13] and [Voh10].

Here we use the *residual error estimator* $\eta_K,\ K \in \mathcal{T}_h$, developed in [Dol13b], which is based on the approximation of the computational error measured in the dual norm. We suppose that similar results can be obtained by any other reasonable error estimator. However, a single error estimator η_K cannot simultaneously decide whether it is better to accomplish h or p refinement. Several strategies for making this decision have been proposed. See, e.g., [HS05] or [EM07] for a survey.

In the following numerical examples, we employ the approach from [Dol13b], where the regularity indicator is based on measuring the interelement jumps of the DG solution.

Numerical example

We illustrate the efficiency of the *hp*-discontinuous Galerkin method by the following example. Let $\Omega = (0,1)\times(0,1)$, $\partial\Omega_D := \partial\Omega$. We consider the Poisson problem (5.1), where the right-hand side f and the Dirichlet boundary condition u_D are chosen so that the exact solution has the form

$$u(x_1, x_2) = 2(x_1^2 + x_2^2)^{-3/4}\, x_1 x_2 (1 - x_1)(1 - x_2), \tag{5.60}$$

cf. Section 1.8.2. The function u has a singularity at the origin and, hence, $u \in H^1(\Omega)$ but $u \notin H^2(\Omega)$. Numerical examples presented in Section 1.8.2 showed that the experimental order of convergence of DGM in the $H^1(\Omega, \mathcal{T}_h)$-seminorm is approximately $O(h^{1/2})$ for any tested polynomial approximation degree.

In order to study the computational properties of the *hp*-DGM, we carried out three types of calculations:

- fix-DGM: $P_p,\ p = 1, 3, 5$, *approximations on uniformly refined grids*, i.e., the computation with fixed polynomial approximation degree ($p_K = p$ for all $K \in \mathcal{T}_h$) on uniform triangular grids with $h_\ell = 1/2^{2+\ell},\ \ell = 0, 1, \dots$. Figure 5.1 shows the uniformly refined grids for $\ell = 0, 2, 4$.
- *h*-DGM: *h-adaptive DGM for* $P_p,\ p = 1, 3, 5$, *polynomial approximations*, i.e., the computation with fixed polynomial approximation degree ($p_K = p$ for all $K \in \mathcal{T}_h$) on adaptively (locally) refined grids. Figure 5.2 shows the example of the sequence of meshes generated by the *h*-refinement algorithm for $p = 3$ together with details at the singularity corner.
- *hp*-DGM: *hp-adaptive DGM*, i.e., the computation with adaptively chosen polynomial approximation degree $p_K,\ K \in \mathcal{T}_h$, on adaptively (locally) refined grids using the algorithm from [Dol13b]. Figure 5.3 shows the *hp*-grids generated by this algorithm for selected levels of adaptation. Each $K \in \mathcal{T}_h$ is marked by the colour corresponding to the used polynomial approximation degree.

Our aim is to identify the *experimental order of convergence* (EOC), similarly as in Section 1.8. Since we employ locally adaptive grids and possible different polynomial approximation degrees on $K \in \mathcal{T}_h$, it does not make sense to use formula (1.176) and to define the EOC by (1.177). Therefore, we expect that the computational error $e_h = u_h - u$ behaves according to the formula

$$\|e_h\| \approx C N_h^{-\frac{\mathrm{EOC}}{d}}, \tag{5.61}$$

where $\|e_h\|$ is the computational error in the (semi-)norm of interest, $d = 2$ is the space dimension, $C > 0$ is a constant, $\mathrm{EOC} \in \mathbb{R}$ is the experimental order of convergence and N_h is the number of degrees of freedom given by (cf., e.g., [BS94b, Chapter 3] or [Cia79])

$$N_h = \dim S_{h\boldsymbol{p}} = \sum_{K \in \mathcal{T}_h} \frac{1}{d!} \prod_{j=1}^{d} (p_K + j). \tag{5.62}$$

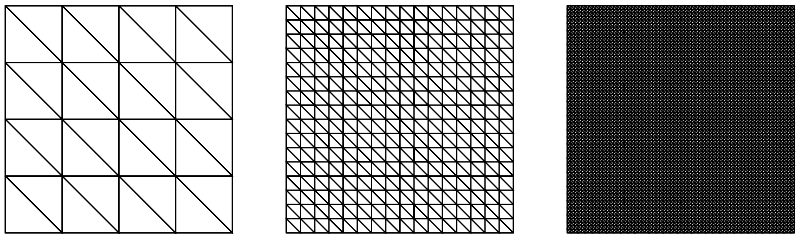

Figure 5.1: Computation fix-DGM: the uniformly refined computational grids for $\ell = 0, 2, 4$.

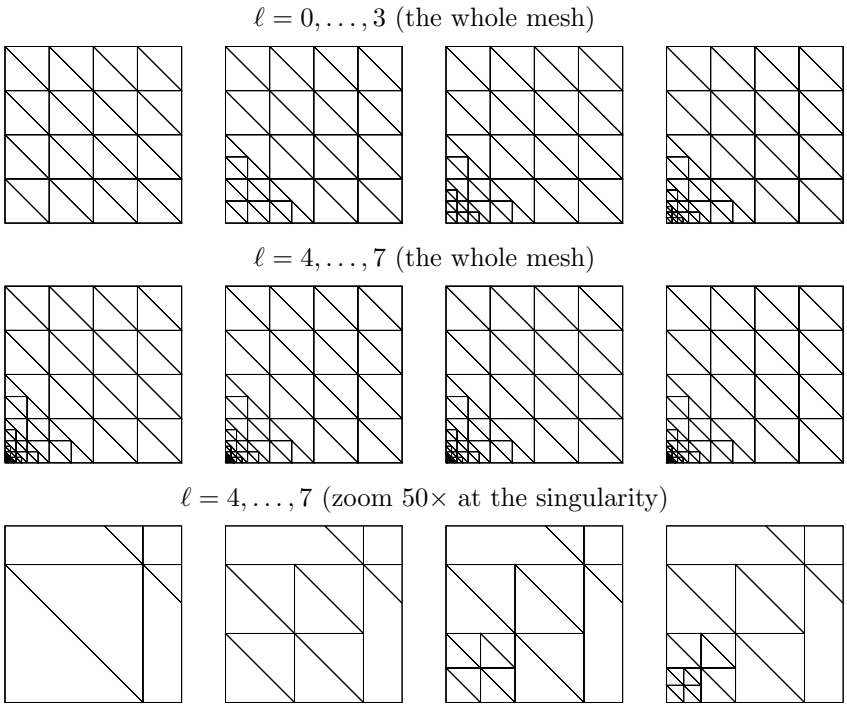

Figure 5.2: Computation h-DGM: example of the sequence of the meshes $\ell = 0, \ldots, 7$, generated by the h-refinement algorithm for $p = 3$; the last row shows the details at the singularity corner.

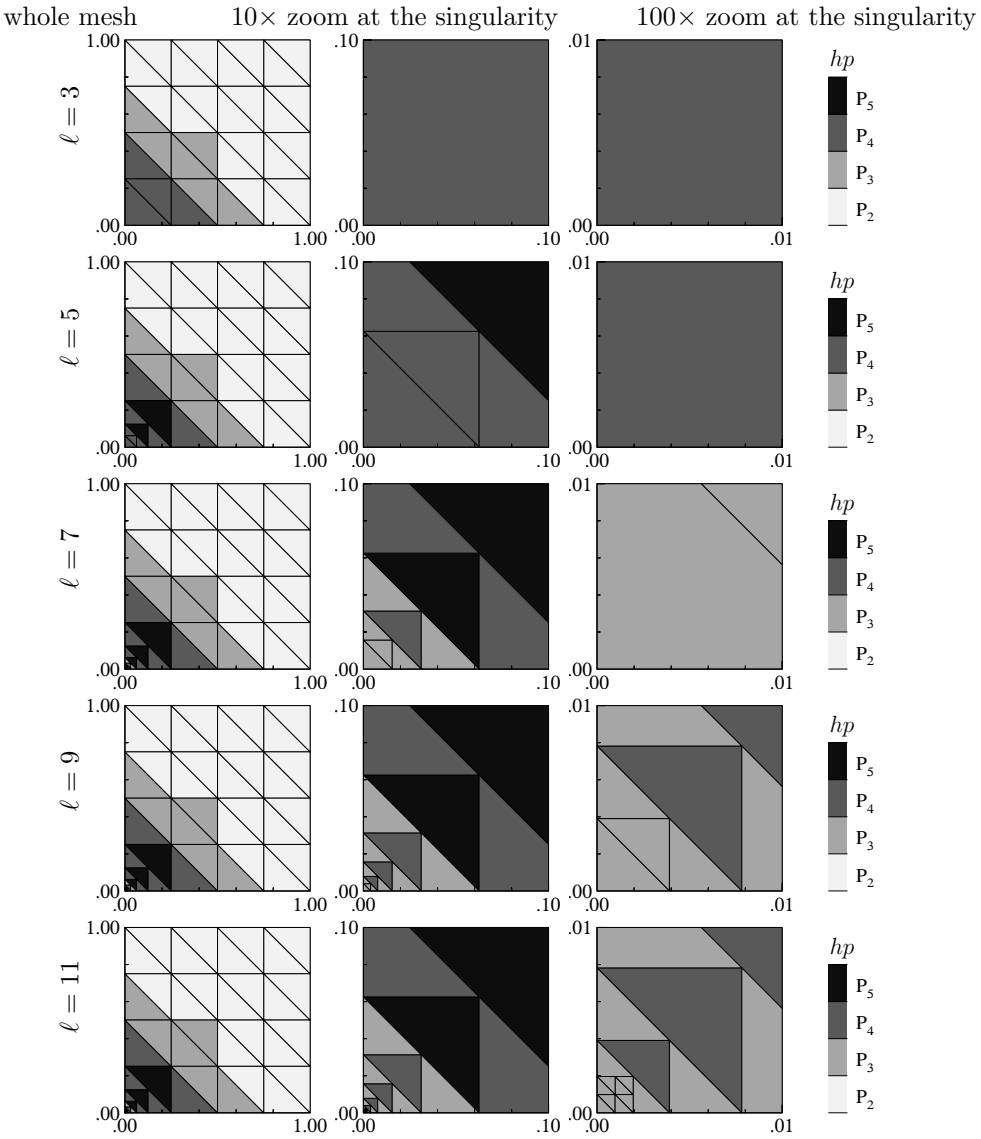

Figure 5.3: Computation hp-DGM: the hp-meshes for the levels of adaptation $\ell = 3, 5, 7, 9, 11$; each $K \in \mathcal{T}_h$ is marked by the colour corresponding to the used polynomial approximation degree; the whole domain (left), zooms 10× and 100× at the singularity corner (center and right), respectively.

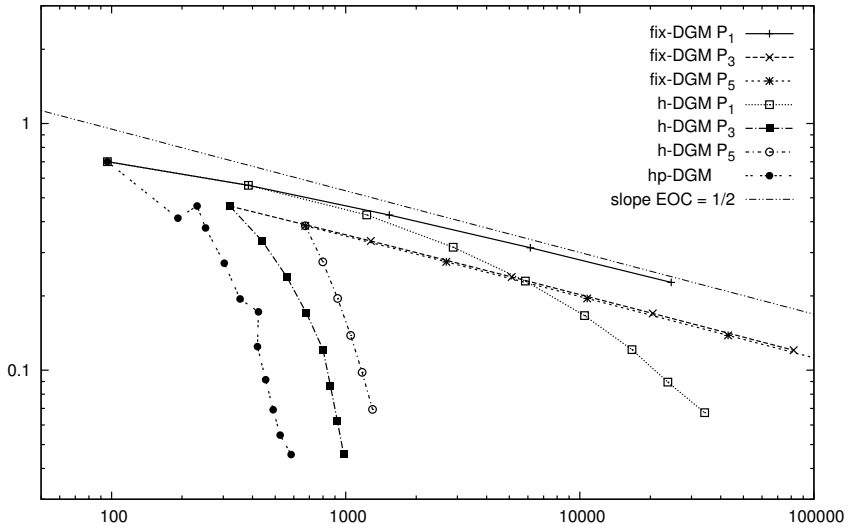

Figure 5.4: Convergence of errors in the $H^1(\Omega, \mathcal{T}_h)$-seminorm with respect to the number of DOF for fix-DGM, h-DGM, hp-DGM computations. Moreover the slope corresponding to EOC= 1/2 is plotted.

Obviously, if the mesh $\mathcal{T}_h$ is quasi-uniform (cf. Remark 1.3) and $p_K = p$ for all $K \in \mathcal{T}_h$, then the experimental orders of convergence defined by (5.61) and by (1.176) are identical.

Since the exact solution is known and, therefore, $\|e_h\|$ can be exactly evaluated, it is possible to determine the EOC in the following way. Let $\|e_{h_1}\|$ and $\|e_{h_2}\|$ be the computational errors of numerical solutions obtained on two different meshes $\mathcal{T}_{h_1}$ and $\mathcal{T}_{h_2}$ having the numbers of degrees of freedom N_{h_1} and N_{h_2}, respectively. Then eliminating the constant C from (5.61), we come to the definition of the EOC in the form

$$\text{EOC} = -\frac{\log(\|e_{h_1}\|/\|e_{h_2}\|)}{\log((N_{h_1}/N_{h_2})^{1/d})}. \tag{5.63}$$

Table 5.1 shows the results of all types of computations (fix-DGM, h-DGM, hp-DGM), namely, the computational errors in the $L^\infty(\Omega)$-norm, the $L^2(\Omega)$-norm and the $H^1(\Omega, \mathcal{T}_h)$-seminorm and the corresponding EOC together with the computational time in seconds. The results with the error in the $H^1(\Omega, \mathcal{T}_h)$-seminorm are visualized in Figure 5.4. We observe that the fix-DGM computations give a low experimental order of convergence in agreement with results in Tables 1.5 and 1.6. Moreover, the h-mesh refinements h-DGM achieve the same error level with smaller number of DOF. Namely, for P_3 and P_5 approximation the decrease of the number of DOF is essential. Finally, the hp-adaptive strategy hp-DGM leads to the lower number of DOF (and a shorter computational time) in comparison to h-DGM.

We observe that in some cases EOC is negative for the hp-DGM. The relation (5.63) gives EOC < 0 in two situations:

- The adaptive algorithm increases the number of degrees of freedom N_h but the computational error e_h increases too. This is the usual property of hp-adaptive methods, when at the beginning of the adaptation algorithm we use high polynomial degrees on coarse grids. The polynomial approximation oscillates and thus e_h is large.
- The adaptive algorithm reduces the number of degrees of freedom N_h together with a decrease of the computational error e_h (see level 7 of hp-DGM in Table 5.1). This is in fact a positive property of the used algorithm.

Furthermore, from Table 5.1, we find out that for the hp-DGM computations, the error in the $L^2(\Omega)$-norm is almost constant for the levels $\ell = 8$, 9, 10 and 11, whereas the errors in the $L^\infty(\Omega)$-norm and in the $H^1(\Omega, \mathcal{T}_h)$-seminorm are decreasing. This is caused be the fact that the piecewise constant function $F^0 : \Omega \to \mathbb{R}$ given by

$$F^0|_K = \|u - u_h\|_{L^2(K)}, \quad K \in \mathcal{T}_h$$

attains the maximal values for K far from the singularity (if the mesh is already sufficiently refined), whereas the piecewise constant function $F^1 : \Omega \to \mathbb{R}$ given by

$$F^1|_K = |u - u_h|_{H^1(K)}, \quad K \in \mathcal{T}_h$$

attains the maximal values for K near the singularity even for sufficiently refined grids. Figure 5.3 shows that for $\ell \geq 5$ only elements near the singularity are adapted, and hence the error in the $L^2(\Omega)$-norm cannot be further decreased.

The presented numerical experiments show that the hp-DGM can treat locally refined grids with hanging nodes and different approximation polynomial degrees generated by an hp-adaptive technique. This approach allows us to achieve the given error tolerance with the aid of a low number of DOF.

fix-DGM

level	p	$\#\mathcal{T}_h$	DOF	$\|e_h\|_{L^\infty(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{H^1(\Omega,\mathcal{T}_h)}$	EOC	CPU(s)
0	1	32	96	2.47E-01	–	4.22E-02	–	7.01E-01	–	0.3
1	1	128	384	1.99E-01	0.3	1.83E-02	1.2	5.61E-01	0.3	0.5
2	1	512	1536	1.50E-01	0.4	7.28E-03	1.3	4.26E-01	0.4	1.4
3	1	2048	6144	1.09E-01	0.5	2.77E-03	1.4	3.14E-01	0.4	6.3
4	1	8192	24576	7.84E-02	0.5	1.02E-03	1.4	2.27E-01	0.5	38.9
0	3	32	320	1.51E-01	–	5.79E-03	–	4.63E-01	–	0.4
1	3	128	1280	1.07E-01	0.5	2.13E-03	1.4	3.34E-01	0.5	1.0
2	3	512	5120	7.55E-02	0.5	7.71E-04	1.5	2.39E-01	0.5	3.9
3	3	2048	20480	5.34E-02	0.5	2.76E-04	1.5	1.70E-01	0.5	16.8
4	3	8192	81920	3.78E-02	0.5	9.83E-05	1.5	1.20E-01	0.5	82.2
0	5	32	672	2.29E-01	–	5.09E-03	–	3.85E-01	–	0.6
1	5	128	2688	1.62E-01	0.5	1.81E-03	1.5	2.75E-01	0.5	2.2
2	5	512	10752	1.15E-01	0.5	6.42E-04	1.5	1.95E-01	0.5	9.2
3	5	2048	43008	8.12E-02	0.5	2.28E-04	1.5	1.38E-01	0.5	41.2
4	5	8192	172032	5.74E-02	0.5	8.05E-05	1.5	9.80E-02	0.5	235.3

h-DGM

level	p	$\#\mathcal{T}_h$	DOF	$\|e_h\|_{L^\infty(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{H^1(\Omega,\mathcal{T}_h)}$	EOC	CPU(s)
0	1	32	96	2.47E-01	–	4.22E-02	–	7.01E-01	–	0.3
1	1	128	384	1.99E-01	0.3	1.83E-02	1.2	5.61E-01	0.3	0.5
2	1	410	1230	1.50E-01	0.5	7.34E-03	1.6	4.26E-01	0.5	1.3
3	1	959	2877	1.09E-01	0.7	2.89E-03	2.2	3.15E-01	0.7	3.2
4	1	1952	5856	7.84E-02	0.9	1.15E-03	2.6	2.30E-01	0.9	8.2
5	1	3491	10473	5.59E-02	1.2	5.28E-04	2.7	1.67E-01	1.1	21.1
6	1	5567	16701	3.96E-02	1.5	3.11E-04	2.3	1.21E-01	1.4	47.8
7	1	7922	23766	2.81E-02	2.0	2.40E-04	1.5	8.95E-02	1.7	86.0
8	1	11387	34161	1.99E-02	1.9	1.77E-04	1.7	6.73E-02	1.6	168.6
0	3	32	320	1.51E-01	–	5.79E-03	–	4.63E-01	–	0.4
1	3	44	440	1.07E-01	2.2	2.14E-03	6.3	3.34E-01	2.0	0.7
2	3	56	560	7.55E-02	2.9	7.99E-04	8.2	2.39E-01	2.8	1.0
3	3	68	680	5.34E-02	3.6	3.42E-04	8.7	1.70E-01	3.5	1.3
4	3	80	800	3.78E-02	4.3	2.23E-04	5.3	1.21E-01	4.2	1.8
5	3	86	860	2.67E-02	9.5	2.03E-04	2.6	8.67E-02	9.3	2.2
6	3	92	920	1.89E-02	10.3	2.00E-04	0.4	6.25E-02	9.7	2.7
7	3	98	980	1.34E-02	11.0	2.00E-04	0.1	4.57E-02	9.9	3.1
0	5	32	672	2.29E-01	–	5.09E-03	–	3.85E-01	–	0.6
1	5	38	798	1.62E-01	4.0	1.81E-03	12.0	2.75E-01	3.9	1.0
2	5	44	924	1.15E-01	4.7	6.43E-04	14.1	1.95E-01	4.7	1.5
3	5	50	1050	8.12E-02	5.4	2.29E-04	16.1	1.38E-01	5.4	2.1
4	5	56	1176	5.74E-02	6.1	8.53E-05	17.5	9.80E-02	6.1	2.8
5	5	62	1302	4.06E-02	6.8	3.99E-05	15.0	6.94E-02	6.8	3.6

hp-DGM

level	p	$\#\mathcal{T}_h$	DOF	$\|e_h\|_{L^\infty(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{H^1(\Omega,\mathcal{T}_h)}$	EOC	CPU(s)
0	–	32	96	2.47E-01	–	4.22E-02	–	7.01E-01	–	0.3
1	–	32	192	1.14E-01	2.2	8.68E-03	4.6	4.14E-01	1.5	0.4
2	–	32	232	1.51E-01	-3.0	5.86E-03	4.1	4.63E-01	-1.2	0.5
3	–	32	252	2.01E-01	-7.0	5.98E-03	-0.5	3.77E-01	5.0	0.6
4	–	35	303	1.43E-01	3.7	2.25E-03	10.6	2.71E-01	3.6	0.8
5	–	38	354	1.01E-01	4.4	1.03E-03	10.0	1.95E-01	4.3	1.0
6	–	44	424	5.34E-02	7.1	7.40E-04	3.7	1.72E-01	1.3	1.2
7	–	44	420	3.78E-02	-81.5	6.93E-04	-15.4	1.25E-01	-76.1	1.4
8	–	47	455	2.67E-02	8.7	6.86E-04	0.2	9.15E-02	7.7	1.7
9	–	50	490	1.89E-02	9.4	6.86E-04	0.0	6.91E-02	7.6	1.9
10	–	53	525	1.34E-02	10.1	6.85E-04	0.0	5.46E-02	6.9	2.2
11	–	59	585	9.45E-03	6.4	6.85E-04	0.0	4.55E-02	3.3	2.4

Table 5.1: Computational errors in the $L^\infty(\Omega)$-norm, the $L^2(\Omega)$-norm and the $H^1(\Omega,\mathcal{T}_h)$-seminorm, the corresponding EOC and the CPU time for all types of computations.

Chapter 6

Inviscid compressible flow

In previous chapters we introduced and analyzed the *discontinuous Galerkin method* (DGM) for the numerical solution of several scalar equations. However, many practical problems are described by systems of partial differential equations. In the second part of this book, we present the application of the DGM to solving compressible flow problems. The numerical schemes, analyzed for a scalar equation, are extended to a system of equations and numerically verified. We also deal with an efficient solution of resulting systems of algebraic equations.

One of the models used for the numerical simulation of a compressible (i.e., gas) flow is based on the assumption that the flow is inviscid and adiabatic. This means that in gas we neglect the internal friction and heat transfer. Inviscid adiabatic flow is described by the *continuity equation*, the *Euler equations* of motion and the *energy equation*, to which we add closing thermodynamical relations. See, for example, [FFS03, Section 1.2]. This complete system is usually called the Euler equations.

The Euler equations, similarly as other nonlinear hyperbolic systems of conservation laws, may have discontinuous solutions. This is one of the reasons that the finite volume method (FVM) using piecewise constant approximations became very popular for the numerical solution of compressible flow. For a detailed treatment of finite volume techniques, we can refer to [EGH00] and [Krö97]. See also [Fei93] and [FFS03]. Moreover, the FVM is applicable on general polygonal meshes and its algorithmization is relatively easy. Therefore, many fluid dynamics codes and program packages are based on the FVM. However, the standard FVM is only of the first order, which is not sufficient in a number of applications. The increase of accuracy in finite volume schemes applied on unstructured and/or anisotropic meshes seems to be problematic and is not theoretically sufficiently justified.

As for the finite element method (FEM), the standard conforming finite element techniques were considered to be suitable for the numerical solution of elliptic and parabolic problems, linear elasticity and incompressible viscous flow, when the exact solution is sufficiently regular. Of course, there are also conforming finite element techniques applied to the solution of compressible flow, but the treatment of discontinuous solutions is rather complicated. For a survey, see [FFS03, Section 4.3].

A combination of ideas and techniques of the FV and FE methods yields the discontinuous Galerkin method using advantages of both approaches and allowing to obtain schemes with a higher-order accuracy in a natural way. In this chapter we present the application of the DGM to the Euler equations. We describe the discretization, a special attention is paid to the choice of boundary conditions and we also discuss an efficient solution of the resulting discrete problem.

6.1 Formulation of the inviscid flow problem

6.1.1 Governing equations

We shall consider the unsteady compressible inviscid adiabatic flow in a domain $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) and time interval $(0, T)$, $0 < T < \infty$. In what follows, we present only the governing equations, their derivation can be found, e.g., in [FFS03, Section 3.1].

We use the standard notation: ρ - density, p - pressure (symbol p denotes the degree of polynomial approximation), E - total energy, v_s, $s = 1, \dots, d$ - components of the velocity vector $\boldsymbol{v} = (v_1, \dots, v_d)^{\mathrm{T}}$ in the directions x_s, θ - absolute temperature, $c_v > 0$ - specific heat at constant volume, $c_{\mathrm{p}} > 0$ - specific heat at constant pressure, $\gamma = c_{\mathrm{p}}/c_v > 1$ - Poisson adiabatic constant, $R = c_{\mathrm{p}} - c_v > 0$ - gas constant. We shall be concerned with the flow of a perfect gas, for which the equation of state has the form

$$\mathrm{p} = R\rho\theta, \tag{6.1}$$

and assume that c_{p}, c_v are constants. Since the gas is light, we neglect the outer volume force.

The system of governing equations formed by the continuity equation, the Euler equations of motion and the energy equation

(see [FFS03, Section 3.1]) considered in the space-time cylinder $Q_T = \Omega \times (0,T)$ can be written in the form

$$\frac{\partial \rho}{\partial t} + \sum_{s=1}^{d} \frac{\partial(\rho v_s)}{\partial x_s} = 0, \tag{6.2}$$

$$\frac{\partial(\rho v_i)}{\partial t} + \sum_{s=1}^{d} \frac{\partial(\rho v_i v_s + \delta_{is}\mathrm{p})}{\partial x_s} = 0, \quad i = 1,\dots,d, \tag{6.3}$$

$$\frac{\partial E}{\partial t} + \sum_{s=1}^{d} \frac{\partial((E+\mathrm{p})v_s)}{\partial x_s} = 0. \tag{6.4}$$

To the above system, we add the thermodynamical relations defining the pressure

$$\mathrm{p} = (\gamma - 1)(E - \rho|\boldsymbol{v}|^2/2), \tag{6.5}$$

and the total energy

$$E = \rho(c_v\theta + |\boldsymbol{v}|^2/2), \tag{6.6}$$

in terms of other quantities.

We define the *speed of sound* a and the *Mach number* M by

$$a = \sqrt{\gamma \mathrm{p}/\rho}, \qquad M = \frac{|\boldsymbol{v}|}{a}. \tag{6.7}$$

The flow is called *subsonic* and *supersonic* in a region ω, if $M < 1$ and $M > 1$, respectively, in ω. If $M \gg 1$, we speak about *hypersonic flow*. If there are two subregions ω_1 and ω_2 in the flow field such that $M < 1$ in ω_1 and $M > 1$ in ω_2, the flow is called *transonic*.

Exercise 6.1. *Derive (6.5) from* (6.1) *and (6.6).*

System (6.2)–(6.4) has $m = d + 2$ equations and it can be written in the form

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{s=1}^{d} \frac{\partial \boldsymbol{f}_s(\boldsymbol{w})}{\partial x_s} = 0, \tag{6.8}$$

where

$$\boldsymbol{w} = (w_1,\dots,w_m)^{\mathrm{T}} = (\rho, \rho v_1, \dots, \rho v_d, E)^{\mathrm{T}} \in \mathbb{R}^m, \tag{6.9}$$

is the so-called *state vector*, and

$$\begin{aligned} \boldsymbol{f}_s(\boldsymbol{w}) &= \begin{pmatrix} f_{s,1}(\boldsymbol{w}) \\ f_{s,2}(\boldsymbol{w}) \\ \vdots \\ f_{s,m-1}(\boldsymbol{w}) \\ f_{s,m}(\boldsymbol{w}) \end{pmatrix} = \begin{pmatrix} \rho v_s \\ \rho v_1 v_s + \delta_{1s}\mathrm{p} \\ \vdots \\ \rho v_d v_s + \delta_{ds}\mathrm{p} \\ (E+\mathrm{p})v_s \end{pmatrix} \\ &= \begin{pmatrix} w_{s+1} \\ \frac{w_2 w_{s+1}}{w_1} + \delta_{1s}(\gamma-1)\left(w_m - \frac{1}{2w_1}\sum_{i=2}^{m-1} w_i^2\right) \\ \vdots \\ \frac{w_{m-1} w_{s+1}}{w_1} + \delta_{ds}(\gamma-1)\left(w_m - \frac{1}{2w_1}\sum_{i=2}^{m-1} w_i^2\right) \\ \frac{w_{s+1}}{w_1}\left(\gamma w_m - \frac{\gamma-1}{2w_1}\sum_{i=2}^{m-1} w_i^2\right) \end{pmatrix}, \end{aligned} \tag{6.10}$$

is the *flux* of the quantity $\boldsymbol{w}$ in the direction x_s, $s = 1,\dots,d$. By δ_{ij} we denote the Kronecker symbol. Often, $\boldsymbol{f}_s$, $s = 1,\dots,d$, are called *inviscid Euler fluxes*.

Usually, system (6.2)–(6.4), i.e., (6.8), is called the system of the Euler equations, or simply the *Euler equations*. The functions $\rho, v_1, \dots, v_d, \mathrm{p}$ are called *primitive* (or physical) *variables*, whereas $w_1 = \rho$, $w_2 = \rho v_1, \dots, w_{m-1} = \rho v_d$, $w_m = E$ are

conservative variables. It is easy to show that

$$\begin{aligned} v_i =& w_{i+1}/w_1,\ i=1,\dots,d, \\ \mathrm{p} =& (\gamma-1)\left(w_m - \sum_{i=2}^{m-1} w_i^2/(2w_1)\right), \\ \theta =& \left(w_m/w_1 - \frac{1}{2}\sum_{i=2}^{m-1}(w_i/w_1)^2\right)/c_v. \end{aligned} \tag{6.11}$$

The domain of definition of the vector-valued functions $\boldsymbol{f}_s,\ s=1,\dots,d$, is the open set $\mathcal{D}\subset\mathbb{R}^m$ of vectors $\boldsymbol{w}=(w_1,\dots,w_m)^{\mathrm{T}}$ such that the corresponding density and pressure are positive:

$$\mathcal{D}=\Big\{\boldsymbol{w}\in\mathbb{R}^m;\ w_1=\rho>0,\ w_m-\sum_{i=2}^{m-1} w_i^2/(2w_1)=\mathrm{p}/(\gamma-1)>0\Big\}. \tag{6.12}$$

Obviously, $\boldsymbol{f}_s\in(C^1(\mathcal{D}))^m$.

Differentiation in (6.8) and the use of the chain rule lead to a *first-order quasilinear system* of partial differential equations

$$\frac{\partial \boldsymbol{w}}{\partial t}+\sum_{s=1}^{d}\mathbb{A}_s(\boldsymbol{w})\frac{\partial \boldsymbol{w}}{\partial x_s}=0, \tag{6.13}$$

where $\mathbb{A}_s(\boldsymbol{w})$ is the $m\times m$ Jacobi matrix of the mapping $\boldsymbol{f}_s$ defined for $\boldsymbol{w}\in\mathcal{D}$:

$$\mathbb{A}_s(\boldsymbol{w})=\quad \frac{\mathrm{D}\boldsymbol{f}_s(\boldsymbol{w})}{\mathrm{D}\boldsymbol{w}}=\left(\frac{\partial f_{s,i}(\boldsymbol{w})}{\partial w_j}\right)_{i,j=1}^{m},\ s=1,\dots,d. \tag{6.14}$$

Let

$$\mathrm{B}_1=\{\boldsymbol{n}\in\mathbb{R}^d;\ |\boldsymbol{n}|=1\} \tag{6.15}$$

denote the unit sphere in $\mathbb{R}^d$. For $\boldsymbol{w}\in\mathcal{D}$ and $\boldsymbol{n}=(n_1,\dots,n_d)^{\mathrm{T}}\in\mathrm{B}_1$ we denote

$$\boldsymbol{P}(\boldsymbol{w},\boldsymbol{n})=\sum_{s=1}^{d}\boldsymbol{f}_s(\boldsymbol{w})n_s, \tag{6.16}$$

which is the *physical flux* of the quantity $\boldsymbol{w}$ in the direction $\boldsymbol{n}$. Obviously, the Jacobi matrix $\mathrm{D}\boldsymbol{P}(\boldsymbol{w},\boldsymbol{n})/\mathrm{D}\boldsymbol{w}$ can be expressed in the form

$$\frac{\mathrm{D}\boldsymbol{P}(\boldsymbol{w},\boldsymbol{n})}{\mathrm{D}\boldsymbol{w}}=\mathbb{P}(\boldsymbol{w},\boldsymbol{n}):=\sum_{s=1}^{d}\mathbb{A}_s(\boldsymbol{w})n_s. \tag{6.17}$$

Exercise 6.2. *Let $d=2$. Prove that the Jacobi matrices $\mathbb{A}_s,\ s=1,2$, have the form*

$$\mathbb{A}_1(\boldsymbol{w})=\begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2-v_1^2 & (3-\gamma)v_1 & -\gamma_1 v_2 & \gamma_1 \\ -v_1v_2 & v_2 & v_1 & 0 \\ v_1\left(\gamma_1|\boldsymbol{v}|^2-\gamma\frac{E}{\rho}\right) & \gamma\frac{E}{\rho}-\gamma_1 v_1^2-\frac{\gamma_1}{2}|\boldsymbol{v}|^2 & -\gamma_1 v_1v_2 & \gamma v_1 \end{pmatrix}, \tag{6.18}$$

$$\mathbb{A}_2(\boldsymbol{w})=\begin{pmatrix} 0 & 0 & 1 & 0 \\ -v_1v_2 & v_2 & v_1 & 0 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2-v_2^2 & -\gamma_1 v_1 & (3-\gamma)v_2 & \gamma_1 \\ v_2\left(\gamma_1|\boldsymbol{v}|^2-\gamma\frac{E}{\rho}\right) & -\gamma_1 v_1v_2 & \gamma\frac{E}{\rho}-\gamma_1 v_2^2-\frac{\gamma_1}{2}|\boldsymbol{v}|^2 & \gamma v_2 \end{pmatrix}, \tag{6.19}$$

where $\gamma_1=\gamma-1$.

Exercise 6.3. *Let $d=2$. With the aid of* (6.18) – (6.19) *show that the matrix $\mathbb{P}(\boldsymbol{w},\boldsymbol{n})$ has the form*

$$\mathbb{P}(\boldsymbol{w},\boldsymbol{n})=\left(\begin{array}{c|c|c|c} 0 & n_1 & n_2 & 0 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2n_1-v_1\boldsymbol{v}\cdot\boldsymbol{n} & -\gamma_2 v_1n_1+\boldsymbol{v}\cdot\boldsymbol{n} & v_1n_2-\gamma_1v_2n_1 & \gamma_1n_1 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2n_2-v_2\boldsymbol{v}\cdot\boldsymbol{n} & v_2n_1-\gamma_1v_1n_2 & -\gamma_2v_2n_2+\boldsymbol{v}\cdot\boldsymbol{n} & \gamma_1n_2 \\ \left(\gamma_1|\boldsymbol{v}|^2-\frac{\gamma E}{\rho}\right)\boldsymbol{v}\cdot\boldsymbol{n} & Gn_1-\gamma_1v_1\boldsymbol{v}\cdot\boldsymbol{n} & Gn_2-\gamma_1v_2\boldsymbol{v}\cdot\boldsymbol{n} & \gamma\boldsymbol{v}\cdot\boldsymbol{n} \end{array}\right), \tag{6.20}$$

where $\boldsymbol{n}=(n_1,n_2)$, $\gamma_1=\gamma-1$, $\gamma_2=\gamma-2$ and $G=\gamma\frac{E}{\rho}-\frac{\gamma_1}{2}|\boldsymbol{v}|^2$.

Exercise 6.4. *Let $d = 3$. Prove that the Jacobi matrices $\mathbb{A}_s$, $s = 1, 2, 3$, have the form*

$$\mathbb{A}_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2 - v_1^2 & (3-\gamma)v_1 & -\gamma_1 v_2 & -\gamma_1 v_3 & \gamma_1 \\ -v_1 v_2 & v_2 & v_1 & 0 & 0 \\ -v_1 v_3 & v_3 & 0 & v_1 & 0 \\ v_1\left(\gamma_1|\boldsymbol{v}|^2 - \gamma\frac{E}{\rho}\right) & \gamma\frac{E}{\rho} - \gamma_1 v_1^2 - \frac{\gamma_1}{2}|\boldsymbol{v}|^2 & -\gamma_1 v_1 v_2 & -\gamma_1 v_1 v_3 & \gamma v_1 \end{pmatrix}, \tag{6.21}$$

$$\mathbb{A}_2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ -v_1 v_2 & v_2 & v_1 & 0 & 0 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2 - v_2^2 & -\gamma_1 v_1 & (3-\gamma)v_2 & -\gamma_1 v_3 & \gamma_1 \\ -v_2 v_3 & 0 & v_3 & v_2 & 0 \\ v_2\left(\gamma_1|\boldsymbol{v}|^2 - \gamma\frac{E}{\rho}\right) & -\gamma_1 v_1 v_2 & \gamma\frac{E}{\rho} - \gamma_1 v_2^2 - \frac{\gamma_1}{2}|\boldsymbol{v}|^2 & -\gamma_1 v_2 v_3 & \gamma v_2 \end{pmatrix}, \tag{6.22}$$

$$\mathbb{A}_3 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ -v_1 v_3 & v_3 & 0 & v_1 & 0 \\ -v_2 v_3 & 0 & v_3 & v_2 & 0 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2 - v_3^2 & -\gamma_1 v_1 & -\gamma_1 v_2 & (3-\gamma)v_3 & \gamma_1 \\ v_3\left(\gamma_1|\boldsymbol{v}|^2 - \gamma\frac{E}{\rho}\right) & -\gamma_1 v_1 v_3 & -\gamma_1 v_2 v_3 & \gamma\frac{E}{\rho} - \gamma_1 v_3^2 - \frac{\gamma_1}{2}|\boldsymbol{v}|^2 & \gamma v_3 \end{pmatrix}, \tag{6.23}$$

where $\gamma_1 = \gamma - 1$.

Exercise 6.5. *Let $d = 3$. With the aid of (6.21) – (6.23), show that the matrix $\mathbb{P}(\boldsymbol{w}, \boldsymbol{n})$ has the form*

$$\mathbb{P}(\boldsymbol{w}, \boldsymbol{n}) = \tag{6.24}$$
$$\left(\begin{array}{c|c|c|c|c} 0 & n_1 & n_2 & n_3 & 0 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2 n_1 - v_1 \boldsymbol{v}\cdot\boldsymbol{n} & -\gamma_2 v_1 n_1 + \boldsymbol{v}\cdot\boldsymbol{n} & v_1 n_2 - \gamma_1 v_2 n_1 & v_1 n_3 - \gamma_1 v_3 n_1 & \gamma_1 n_1 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2 n_2 - v_2 \boldsymbol{v}\cdot\boldsymbol{n} & v_2 n_1 - \gamma_1 v_1 n_2 & -\gamma_2 v_2 n_2 + \boldsymbol{v}\cdot\boldsymbol{n} & v_2 n_3 - \gamma_1 v_3 n_2 & \gamma_1 n_2 \\ \frac{\gamma_1}{2}|\boldsymbol{v}|^2 n_3 - v_3 \boldsymbol{v}\cdot\boldsymbol{n} & v_3 n_1 - \gamma_1 v_1 n_3 & v_3 n_2 - \gamma_1 v_2 n_3 & -\gamma_2 v_3 n_3 + \boldsymbol{v}\cdot\boldsymbol{n} & \gamma_1 n_3 \\ \left(\gamma_1|\boldsymbol{v}|^2 - \frac{\gamma E}{\rho}\right)\boldsymbol{v}\cdot\boldsymbol{n} & G n_1 - \gamma_1 v_1 \boldsymbol{v}\cdot\boldsymbol{n} & G n_2 - \gamma_1 v_2 \boldsymbol{v}\cdot\boldsymbol{n} & G n_3 - \gamma_1 v_3 \boldsymbol{v}\cdot\boldsymbol{n} & \gamma \boldsymbol{v}\cdot\boldsymbol{n} \end{array}\right),$$

where $\boldsymbol{n} = (n_1, n_2, n_3)$, $\gamma_1 = \gamma - 1$, $\gamma_2 = \gamma - 2$ and $G = \gamma\frac{E}{\rho} - \frac{\gamma_1}{2}|\boldsymbol{v}|^2$.

Let us summarize some important properties of the system of the Euler equations (6.8).

Lemma 6.6. *(a) The vector-valued functions $\boldsymbol{f}_s$ defined by (6.10) are homogeneous mappings of order 1:*

$$\boldsymbol{f}_s(\alpha \boldsymbol{w}) = \alpha \boldsymbol{f}_s(\boldsymbol{w}), \quad \alpha > 0. \tag{6.25}$$

Moreover, we have

$$\boldsymbol{f}_s(\boldsymbol{w}) = \mathbb{A}_s(\boldsymbol{w})\boldsymbol{w}. \tag{6.26}$$

(b) Similarly,

$$\boldsymbol{P}(\alpha \boldsymbol{w}, \boldsymbol{n}) = \alpha \boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}), \quad \alpha > 0, \tag{6.27}$$
$$\boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}) = \mathbb{P}(\boldsymbol{w}, \boldsymbol{n})\boldsymbol{w}. \tag{6.28}$$

(c) The system of the Euler equations is diagonally hyperbolic. This means that the matrix $\mathbb{P} = \sum_{j=1}^{d} \mathbb{A}_j(\boldsymbol{w}) n_j$ has only real eigenvalues $\lambda_i = \lambda_i(\boldsymbol{w}, \boldsymbol{n})$, $i = 1, \ldots, m$, and is diagonalizable: there exists a nonsingular matrix $\mathbb{T} = \mathbb{T}(\boldsymbol{w}, \boldsymbol{n})$ such that

$$\mathbb{T}^{-1}\mathbb{P}\,\mathbb{T} = \boldsymbol{\Lambda} = \boldsymbol{\Lambda}(\boldsymbol{w}, \boldsymbol{n}) = \operatorname{diag}(\lambda_1, \ldots, \lambda_m) = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \ldots & 0 & \lambda_{m-1} & 0 \\ 0 & 0 & \ldots & 0 & \lambda_m \end{pmatrix}. \tag{6.29}$$

The columns of the matrix $\mathbb{T}$ are the eigenvectors of the matrix $\mathbb{P}$.

(d) The eigenvalues of the matrix $\mathbb{P}(\boldsymbol{w},\boldsymbol{n})$, $\boldsymbol{w}\in\mathcal{D}$, $\boldsymbol{n}\in \mathrm{B}_1$ *have the form*

$$\begin{aligned}\lambda_1(\boldsymbol{w},\boldsymbol{n}) =& \boldsymbol{v}\cdot\boldsymbol{n}-a,\\ \lambda_2(\boldsymbol{w},\boldsymbol{n}) =& \cdots=\lambda_{d+1}(\boldsymbol{w},\boldsymbol{n}) \;=\; \boldsymbol{v}\cdot\boldsymbol{n},\\ \lambda_m(\boldsymbol{w},\boldsymbol{n}) =& \boldsymbol{v}\cdot\boldsymbol{n}+a,\end{aligned} \tag{6.30}$$

where $a=\sqrt{\gamma \mathrm{p}/\rho}$ *is the speed of sound and* $\boldsymbol{v}$ *is the velocity vector given by* $\boldsymbol{v}=(w_2/w_1,\, w_3/w_1,\ldots,w_{d+1}/w_1)^{\mathrm{T}}$.

(e) The system of the Euler equations is rotationally invariant. Namely, for $\boldsymbol{n}=(n_1,\ldots,n_d)\in \mathrm{B}_1, \boldsymbol{w}\in\mathcal{D}$ *it holds*

$$\boldsymbol{P}(\boldsymbol{w},\boldsymbol{n})=\sum_{s=1}^{d}\boldsymbol{f}_s(\boldsymbol{w})n_s=\mathbb{Q}^{-1}(\boldsymbol{n})\boldsymbol{f}_1(\mathbb{Q}(\boldsymbol{n})\boldsymbol{w}), \tag{6.31}$$

$$\mathbb{P}(\boldsymbol{w},\boldsymbol{n})=\sum_{s=1}^{d}\mathbb{A}_s(\boldsymbol{w})n_s=\mathbb{Q}^{-1}(\boldsymbol{n})\mathbb{A}_1(\mathbb{Q}(\boldsymbol{n})\boldsymbol{w})\mathbb{Q}(\boldsymbol{n}), \tag{6.32}$$

where $\mathbb{Q}(\boldsymbol{n})$ *is the* $m\times m$ *matrix corresponding to* $\boldsymbol{n}\in \mathrm{B}_1$ *given by*

$$\mathbb{Q}(\boldsymbol{n})=\begin{pmatrix}1 & \mathbf{0} & 0\\ \mathbf{0}^{\mathrm{T}} & \mathbb{Q}_0(\boldsymbol{n}) & \mathbf{0}^{\mathrm{T}}\\ 0 & \mathbf{0} & 1\end{pmatrix}, \tag{6.33}$$

where the $d\times d$ *rotation matrix* $\mathbb{Q}_0(\boldsymbol{n})$ *is defined for* $d=2$ *by*

$$\mathbb{Q}_0(\boldsymbol{n})=\begin{pmatrix}n_1 & n_2\\ -n_2 & n_1\end{pmatrix},\qquad \boldsymbol{n}=(n_1,n_2), \tag{6.34}$$

and for $d=3$ *by*

$$\mathbb{Q}_0(\boldsymbol{n})=\begin{pmatrix}\cos\alpha\cos\beta & \sin\alpha\cos\beta & \sin\beta\\ -\sin\alpha & \cos\alpha & 0\\ -\cos\alpha\sin\beta & -\sin\alpha\sin\beta & \cos\beta\end{pmatrix}, \tag{6.35}$$

$$\boldsymbol{n}=(\cos\alpha\cos\beta,\sin\alpha\cos\beta,\sin\beta),\ \alpha\in[0,2\pi),\ \beta\in[-\pi/2,\pi/2].$$

By $\mathbf{0}$ *we denote the vector* $(0,0)$*, if* $d=2$*, and* $(0,0,0)$*, if* $d=3$.

Proof. See [FFS03, Lemma 3.1, Lemma 3.3, Theorem 3.4]. □ □

6.1.2 Initial and boundary conditions

In order to formulate the problem of inviscid compressible flow, the system of the Euler equations (6.8) has to be equipped with initial and boundary conditions. Let $\Omega\subset\mathbb{R}^d$, $d=2,3$, be a bounded computational domain with a piecewise smooth Lipschitz boundary $\partial\Omega$. We prescribe the *initial condition*

$$\boldsymbol{w}(x,0)=\boldsymbol{w}^0(x),\quad x\in\Omega, \tag{6.36}$$

where $\boldsymbol{w}^0:\Omega\to\mathcal{D}$ is a given vector-valued function. Moreover, the *boundary conditions* are given formally by

$$\mathscr{B}(\boldsymbol{w})=0\quad\text{on } \partial\Omega\times(0,T), \tag{6.37}$$

where $\mathscr{B}$ is a boundary operator.

The choice of appropriate boundary conditions is a very important and delicate question in the numerical simulation of fluid flow. Determining of boundary conditions is, basically, a physical problem, but it must correspond to the mathematical character of the solved equations. Great care is required in their numerical implementation. Usually two types of boundaries are considered: *reflective* and *transparent* or *transmissive*. The reflective boundaries usually consist of fixed walls. Transmissive or transparent boundaries arise from the need to replace unbounded or rather large physical domains by bounded or sufficiently small computational domains. The corresponding boundary conditions are devised so that they allow the passage of waves without any effect on them. For 1D problems the objective is reasonably well attained. For multidimensional problems this is a substantial area of current research, usually referred to *open-end* boundary conditions, *transparent* boundary conditions, *far-field* boundary conditions, *radiation* boundary conditions or *non-reflecting* boundary conditions. Useful publications dealing with boundary conditions are [BT80], [Hed79], [Roe89], [Gil90], [GF87], [GF88], [GK79], [HH88], [Krö91], [GR96, Chapter V]. A rigorous mathematical theory of boundary conditions to conservation laws was developed only for a scalar equation in [BLN79].

The choice of well-posed boundary conditions for the Euler equations (or, in general, of conservation laws) is a delicate question, not completely satisfactorily solved (see, e.g., the paper [BLN79] dealing with the boundary conditions for a scalar equation). We discuss the choice of the boundary conditions in Section 6.3 in relation to the definition of the numerical solution of (6.8).

Let us only mention that we distinguish several disjoint parts of the boundary $\partial\Omega$, namely *inlet* $\partial\Omega_i$, *outlet* $\partial\Omega_o$ and *impermeable walls* $\partial\Omega_W$, i.e., $\partial\Omega = \partial\Omega_i \cup \partial\Omega_o \cup \partial\Omega_W$. In some situations the inlet and outlet parts are considered together. Therefore, we speak about the inlet/outlet part of the boundary. On $\partial\Omega_W$ we prescribe the impermeability condition

$$\boldsymbol{v} \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega_W, \tag{6.38}$$

where $\boldsymbol{n}$ denotes the outer unit normal to $\partial\Omega_W$ and $\boldsymbol{v}$ is the velocity vector.

Concerning the inlet/outlet part of the boundary $\partial\Omega_i \cup \partial\Omega_o$, the boundary conditions are usually chosen in such a way that problem (6.8) is linearly well–posed. (See, e.g., [FFS03, Section 3.3.6].) Practically it means that the number of prescribed boundary conditions is equal to the number of negative eigenvalues of the matrix $\mathbb{P}(\boldsymbol{w}, \boldsymbol{n})$ defined by (6.31). See Section 6.3.

6.2 DG space semidiscretization

In the following, we shall deal with the discretization of the Euler equations (6.8) by the DGM. We recall some notation introduced in Chapters 1 and 2. Similarly as in Chapter 2, we shall derive the DG space semidiscretization leading to a system of ordinary differential equations. Moreover, we develop a (semi-)implicit time discretization technique which is based on a formal linearization of nonlinear terms. We shall also pay attention to some further aspects of the DG discretization of the Euler equations, namely the choice of boundary conditions, the approximation of nonpolygonal boundary and the shock capturing.

6.2.1 Notation

We shall recall and extend notation introduced in Chapters 1 and 2. In the finite element method, the computational domain Ω is usually approximated by a polygonal (if $d = 2$) or polyhedral (if $d = 3$) domain Ω_h, which is the domain of definition of the approximate solution. For the sake of simplicity, we shall assume that the domain Ω is polygonal, and thus $\Omega_h = \Omega$. By $\mathcal{T}_h$ we denote a partition of Ω consisting of closed d-dimensional simplexes with mutually disjoint interiors. We call $\mathcal{T}_h$ the triangulation of Ω.

By $\mathcal{F}_h$ we denote the set of all open $(d-1)$-dimensional faces (open edges when $d = 2$ or open faces when $d = 3$) of all elements $K \in \mathcal{T}_h$. Further, the symbol $\mathcal{F}_h^I$ stands for the set of all $\Gamma \in \mathcal{F}_h$ that are contained in Ω (inner faces). Moreover, we define $\mathcal{F}_h^W$, $\mathcal{F}_h^i$ and $\mathcal{F}_h^o$ as the sets of all $\Gamma \in \mathcal{F}_h$ such that $\Gamma \subset \partial\Omega_W$, $\Gamma \subset \partial\Omega_i$ and $\Gamma \subset \partial\Omega_o$, respectively. In order to simplify the notation, we put $\mathcal{F}_h^{io} = \mathcal{F}_h^i \cup \mathcal{F}_h^o$ and $\mathcal{F}_h^B = \mathcal{F}_h^W \cup \mathcal{F}_h^i \cup \mathcal{F}_h^o$. Finally, for each $\Gamma \in \mathcal{F}_h$ we define a unit normal vector $\boldsymbol{n}_\Gamma = (n_{\Gamma,1}, \dots, n_{\Gamma,d})$. We assume that for $\Gamma \in \mathcal{F}_h^B$ the vector $\boldsymbol{n}_\Gamma$ has the same orientation as the outer normal of $\partial\Omega$. For each $\Gamma \in \mathcal{F}_h^I$, the orientation of $\boldsymbol{n}_\Gamma$ is arbitrary but fixed.

Over the triangulation $\mathcal{T}_h$ we define the *broken Sobolev space* of vector-valued functions (cf. (1))

$$\boldsymbol{H}^1(\Omega, \mathcal{T}_h) = (H^1(\Omega, \mathcal{T}_h))^m, \tag{6.39}$$

where

$$H^1(\Omega, \mathcal{T}_h) = \{v;\ v : \Omega \to \mathbb{R},\ v|_K \in H^1(K)\ \forall\, K \in \mathcal{T}_h\} \tag{6.40}$$

is the broken Sobolev space of scalar functions introduced by (1.29).

For each $\Gamma \in \mathcal{F}_h^I$ there exist two elements $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathcal{T}_h$ such that $\Gamma \subset K_\Gamma^{(L)} \cap K_\Gamma^{(R)}$. We use again the convention that $K_\Gamma^{(R)}$ lies in the direction of $\boldsymbol{n}_\Gamma$ and $K_\Gamma^{(L)}$ in the opposite direction of $\boldsymbol{n}_\Gamma$, see Figure 1.2.

In agreement with Section 1.3.3, for $\boldsymbol{u} \in \boldsymbol{H}^1(\Omega, \mathcal{T}_h)$ and $\Gamma \in \mathcal{F}_h^I$, we introduce the notation:

$$\boldsymbol{u}_\Gamma^{(L)} \text{ is the trace of } \boldsymbol{u}|_{K_\Gamma^{(L)}} \text{ on } \Gamma, \qquad \boldsymbol{u}_\Gamma^{(R)} \text{ is the trace of } \boldsymbol{u}|_{K_\Gamma^{(R)}} \text{ on } \Gamma \tag{6.41}$$

and

$$\langle \boldsymbol{u} \rangle_\Gamma = \frac{1}{2}\left(\boldsymbol{u}_\Gamma^{(L)} + \boldsymbol{u}_\Gamma^{(R)}\right), \tag{6.42}$$

$$[\boldsymbol{u}]_\Gamma = \boldsymbol{u}_\Gamma^{(L)} - \boldsymbol{u}_\Gamma^{(R)}. \tag{6.43}$$

In case that $[\cdot]_\Gamma$, $\langle\cdot\rangle_\Gamma$ and $\boldsymbol{n}_\Gamma$ are arguments of $\int_\Gamma \dots \mathrm{d}S$, $\Gamma \in \mathcal{F}_h$, we usually omit the subscript Γ and write simply $[\cdot]$, $\langle\cdot\rangle$ and $\boldsymbol{n}$, respectively. The value $[\boldsymbol{u}]_\Gamma$ depends on the orientation of $\boldsymbol{n}_\Gamma$, but the value $[\boldsymbol{u}]_\Gamma \cdot \boldsymbol{n}_\Gamma$ is independent of this orientation.

Finally, for $\boldsymbol{u} \in \boldsymbol{H}^1(\Omega, \mathcal{T}_h)$ and $\Gamma \in \mathcal{F}_h^B$, we denote by $\boldsymbol{u}_\Gamma^{(L)}$ the trace of $\boldsymbol{u}|_{K^{(L)}}$ on Γ, where $K^{(L)} \in \mathcal{T}_h$ such that $\Gamma \subset K^{(L)} \cap \partial\Omega$.

The discontinuous Galerkin (DG) approximate solution of (6.8) is sought in a finite-dimensional subspace of $\boldsymbol{H}^1(\Omega, \mathcal{T}_h)$ which consists of piecewise polynomial functions. Hence, over the triangulation $\mathcal{T}_h$ we define the space of vector-valued discontinuous piecewise polynomial functions

$$\boldsymbol{S}_{hp} = (S_{hp})^m, \tag{6.44}$$

where

$$S_{hp} = \{v \in L^2(\Omega); v|_K \in P_p(K) \ \forall\, K \in \mathcal{T}_h\} \tag{6.45}$$

is the space of scalar functions defined by (1.34). Here $P_p(K)$ denotes the space of all polynomials on K of degree $\le p,\ K \in \mathcal{T}_h$. Obviously, $\boldsymbol{S}_{hp} \subset \boldsymbol{H}^1(\Omega, \mathcal{T}_h)$.

6.2.2 Discontinuous Galerkin space semidiscretization

In order to derive the discrete problem, we assume that there exists an exact solution $\boldsymbol{w} \in C^1([0,T]; \boldsymbol{H}^1(\Omega, \mathcal{T}_h))$ of the Euler equations (6.8). Then we multiply (6.8) by a test function $\boldsymbol{\varphi} \in \boldsymbol{H}^1(\Omega, \mathcal{T}_h)$, integrate over any element $K \in \mathcal{T}_h$, apply Green's theorem and sum over all $K \in \mathcal{T}_h$. Then we get

$$\sum_{K\in\mathcal{T}_h} \int_K \frac{\partial \boldsymbol{w}}{\partial t} \cdot \boldsymbol{\varphi}\, \mathrm{d}x - \sum_{K\in\mathcal{T}_h} \int_K \sum_{s=1}^d \boldsymbol{f}_s(\boldsymbol{w}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_s}\, \mathrm{d}x + \sum_{K\in\mathcal{T}_h} \int_{\partial K} \sum_{s=1}^d \boldsymbol{f}_s(\boldsymbol{w}) n_s \cdot \boldsymbol{\varphi}\, \mathrm{d}S = 0, \tag{6.46}$$

where $\boldsymbol{n} = (n_1, \dots, n_d)$ denotes the outer unit normal to the boundary of $K \in \mathcal{T}_h$. Similarly as in Section 1.4, we rewrite (6.46) in the form

$$\begin{aligned} &\sum_{K\in\mathcal{T}_h} \int_K \frac{\partial \boldsymbol{w}}{\partial t} \cdot \boldsymbol{\varphi}\, \mathrm{d}x - \sum_{K\in\mathcal{T}_h} \int_K \sum_{s=1}^d \boldsymbol{f}_s(\boldsymbol{w}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_s}\, \mathrm{d}x \\ &+ \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \sum_{s=1}^d \boldsymbol{f}_s(\boldsymbol{w}) n_{\Gamma,s} \cdot [\boldsymbol{\varphi}]\, \mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^B} \int_\Gamma \sum_{s=1}^d \boldsymbol{f}_s(\boldsymbol{w}) n_{\Gamma,s} \cdot \boldsymbol{\varphi}\, \mathrm{d}S = 0. \end{aligned} \tag{6.47}$$

The crucial point of the DG approximation of conservation laws is the evaluation of the integrals over $\Gamma \in \mathcal{F}_h$ in (6.47). These integrals are approximated with the aid of the *numerical flux* $\mathbf{H} : \mathcal{D} \times \mathcal{D} \times \mathrm{B}_1 \to \mathbb{R}^m$ by

$$\int_\Gamma \sum_{s=1}^d \boldsymbol{f}_s(\boldsymbol{w})\, n_{\Gamma,s} \cdot \boldsymbol{\varphi}\, \mathrm{d}S \approx \int_\Gamma \mathbf{H}(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{w}_\Gamma^{(R)}, \boldsymbol{n}_\Gamma) \cdot \boldsymbol{\varphi}\, \mathrm{d}S, \tag{6.48}$$

where the functions $\boldsymbol{w}_\Gamma^{(L)}$ and $\boldsymbol{w}_\Gamma^{(R)}$ are defined by (6.41) and B_1 by (6.15). The meaning of $\boldsymbol{w}_\Gamma^{(R)}$ for $\Gamma \in \mathcal{F}_h^B$ will be specified later in the treatment of boundary conditions in Section 6.3. The numerical flux is an important concept in the finite volume method (see, e.g., [FFS03, Section 3.2] or [Wes01]). It has to satisfy some basic conditions:

- *continuity*: $\mathbf{H}(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{n})$ is locally Lipschitz-continuous with respect to the variables $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$,
- *consistency*:

$$\mathbf{H}(\boldsymbol{w}, \boldsymbol{w}, \boldsymbol{n}) = \sum_{s=1}^d \boldsymbol{f}_s(\boldsymbol{w})\, n_s, \quad \boldsymbol{w} \in \mathcal{D},\ \boldsymbol{n} = (n_1, \dots, n_d) \in \mathrm{B}_1, \tag{6.49}$$

- *conservativity*:

$$\mathbf{H}(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{n}) = -\mathbf{H}(\boldsymbol{w}_2, \boldsymbol{w}_1, -\boldsymbol{n}), \quad \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{D},\ \boldsymbol{n} \in \mathrm{B}_1. \tag{6.50}$$

Examples of numerical fluxes can be found, e.g., in [Fei93], [FFS03], [Krö97], [Tor97].

Now, we complete the DG space semidiscretization of (6.8). Approximating the face integrals in (6.47) by (6.48) and interchanging the derivative and integral in the first term, we obtain the identity

$$\frac{\mathrm{d}}{\mathrm{d}t}\, (\boldsymbol{w}(t), \boldsymbol{\varphi}) + \boldsymbol{b}_h(\boldsymbol{w}(t), \boldsymbol{\varphi}) = 0 \quad \forall\, \boldsymbol{\varphi} \in \boldsymbol{H}^1(\Omega, \mathcal{T}_h)\ \forall\, t \in (0,T), \tag{6.51}$$

where

$$(\boldsymbol{w},\boldsymbol{\varphi}) = \int_\Omega \boldsymbol{w}\cdot\boldsymbol{\varphi}\,\mathrm{d}x, \tag{6.52}$$

$$\boldsymbol{b}_h(\boldsymbol{w},\boldsymbol{\varphi}) = \sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma \mathbf{H}(\boldsymbol{w}_\Gamma^{(L)},\boldsymbol{w}_\Gamma^{(R)},\boldsymbol{n}_\Gamma)\cdot[\boldsymbol{\varphi}]\,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^B}\int_\Gamma \mathbf{H}(\boldsymbol{w}_\Gamma^{(L)},\boldsymbol{w}_\Gamma^{(R)},\boldsymbol{n}_\Gamma)\cdot\boldsymbol{\varphi}\,\mathrm{d}S$$
$$-\sum_{K\in\mathcal{T}_h}\int_K\sum_{s=1}^d \boldsymbol{f}_s(\boldsymbol{w})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,\mathrm{d}x. \tag{6.53}$$

The meaning of $\boldsymbol{w}_\Gamma^{(R)}$ for $\Gamma\in\mathcal{F}_h^B$ will be specified in Section 6.3. We call $\boldsymbol{b}_h$ the *convection* (or inviscid) *form*. The expressions in (6.51) – (6.53) make sense for $\boldsymbol{w},\boldsymbol{\varphi}\in\boldsymbol{H}^1(\Omega,\mathcal{T}_h)$. The approximation of the exact solution $\boldsymbol{w}(t)$ of (6.8) will be sought in the finite-dimensional space $\boldsymbol{S}_{hp}\subset\boldsymbol{H}^1(\Omega,\mathcal{T}_h)$ for each $t\in(0,T)$. Therefore, using (6.51), we immediately arrive at the definition of an approximate solution.

Definition 6.7. *We say that a function $\boldsymbol{w}_h:\Omega\times(0,T)\to\mathbb{R}^m$ is the* space semidiscrete solution *of the Euler equations (6.8), if the following conditions are satisfied:*

$$\boldsymbol{w}_h\in C^1([0,T];\boldsymbol{S}_{hp}), \tag{6.54a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h)+\boldsymbol{b}_h(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h)=0\quad\forall\boldsymbol{\varphi}_h\in\boldsymbol{S}_{hp}\ \forall t\in(0,T), \tag{6.54b}$$

$$\boldsymbol{w}_h(0)=\Pi_h\boldsymbol{w}^0, \tag{6.54c}$$

where $\Pi_h\boldsymbol{w}^0$ is the $\boldsymbol{S}_{hp}$-approximation of the function $\boldsymbol{w}^0$ from the initial condition (6.36). Usually it is defined as the L^2-projection of $\boldsymbol{w}^0$ on the space $\boldsymbol{S}_{hp}$.

Problem (6.54) represents a system of N_{hp} ordinary differential equations (ODEs), where N_{hp} is equal to the dimension of the space $\boldsymbol{S}_{hp}$. Its solution will be discussed in Section 6.4.

Remark 6.8. *If we consider the case $p=0$ (i.e., the approximate solution is piecewise constant on $\mathcal{T}_h$), then the numerical scheme (6.54) represents the standard* finite volume method. *See, e.g., [FFS03], [Wes01], [Krö97]. Actually, for $p=0$ we choose the basis functions of $\boldsymbol{S}_{h0}$ as characteristic functions χ_K of $K\in\mathcal{T}_h$. Let us recall that $\chi_K=1$ on K and $\chi_K=0$ elsewhere. Therefore, putting $\boldsymbol{\varphi}_h=\chi_K$, $K\in\mathcal{T}_h$, in* (6.54b)*, we obtain*

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{w}_K(t)+\sum_{K'\in\mathcal{N}(K)}|\Gamma_{K,K'}|\,\mathbf{H}(\boldsymbol{w}_K(t),\boldsymbol{w}_{K'}(t),\boldsymbol{n}_{K,K'})=0, \tag{6.55}$$

where

$$\boldsymbol{w}_K=\frac{1}{|K|}\int_K\boldsymbol{w}_h\,\mathrm{d}x,\quad K\in\mathcal{T}_h, \tag{6.56}$$

and $\mathcal{N}(K)=\{K',\ \partial K\cap\partial K'\in\mathcal{F}_h\}$ is the set of all elements K' having a common face $\Gamma_{K,K'}$ with K. The set $\mathcal{N}(K)$ contains also fictitious elements outside of Ω having a common face $\partial K\cap\Omega$ with $K\in\mathcal{T}_h$. In this case, the value $\boldsymbol{w}_{K'}$ in the numerical flux $\mathbf{H}$ is determined from boundary conditions. By $|\Gamma_{K,K'}|$ and $|K|$ we denote the $(d-1)$-Lebesgue measure of the common face $\Gamma_{K,K'}$ between K and K' and the d-dimensional measure of the element K, respectively. The symbol $\boldsymbol{n}_{K,K'}$ denotes the outer unit normal to ∂K on $\Gamma_{K,K'}$.

6.3 Numerical treatment of boundary conditions

If $\Gamma\in\mathcal{F}_h^B$, then it is necessary to specify the boundary state $\boldsymbol{w}_\Gamma^{(R)}$ appearing in the numerical flux $\mathbf{H}$ in the definition (6.53) of the convection form $\boldsymbol{b}_h$. In what follows, we shall describe the treatment of the boundary conditions for impermeable walls and the inlet/outlet part of the boundary. The boundary conditions should be theoretically determined at all boundary points. In practical computations, when the integrals are evaluated with the aid of quadrature formulae, it is enough to consider the boundary conditions at only integration boundary points. Therefore, for the sake of simplicity, the symbol $\boldsymbol{w}_\Gamma^{(R)}$ will mean the value of this function at a boundary point in consideration.

6.3.1 Boundary conditions on impermeable walls

For $\Gamma\in\mathcal{F}_h^W$ we should interpret in a suitable way the impermeability condition (6.38), i.e., $\boldsymbol{v}\cdot\boldsymbol{n}=0$, where $\boldsymbol{v}$ is the velocity vector and $\boldsymbol{n}$ the outer unit normal to $\partial\Omega_W$. This condition has to be incorporated in some sense into the expression $\mathbf{H}(\boldsymbol{w}_\Gamma^{(L)},\boldsymbol{w}_\Gamma^{(R)},\boldsymbol{n}_\Gamma)$ appearing in the definition (6.53) of the form $\boldsymbol{b}_h$.

We shall describe two possibilities. The first one is based on the direct use of the impermeability condition in the physical flux $\boldsymbol{P}(\boldsymbol{w},\boldsymbol{n})$. The second one applies the *mirror operator* to the state $\boldsymbol{w}$.

Direct use of the impermeability condition

Let $\boldsymbol{n} = (n_1, \dots, n_d) \in \mathrm{B}_1$. Then from (6.16) and (6.10) we have

$$\boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}) = \sum_{s=1}^{d} \boldsymbol{f}_s(\boldsymbol{w}) n_s = \sum_{s=1}^{d} \begin{pmatrix} f_{s,1}(\boldsymbol{w}) \\ f_{s,2}(\boldsymbol{w}) \\ \vdots \\ f_{s,m-1}(\boldsymbol{w}) \\ f_{s,m}(\boldsymbol{w}) \end{pmatrix} n_s = \begin{pmatrix} \rho \boldsymbol{v} \cdot \boldsymbol{n} \\ \rho v_1 \boldsymbol{v} \cdot \boldsymbol{n} + \mathrm{p} n_1 \\ \vdots \\ \rho v_d \boldsymbol{v} \cdot \boldsymbol{n} + \mathrm{p} n_d \\ (E + \mathrm{p}) \boldsymbol{v} \cdot \boldsymbol{n} \end{pmatrix}. \tag{6.57}$$

Using the condition $\boldsymbol{v} \cdot \boldsymbol{n} = 0$ in (6.57), we obtain

$$\boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}) = \sum_{s=1}^{d} \boldsymbol{f}_s(\boldsymbol{w}) n_s = (0, \mathrm{p}\, n_1, \dots, \mathrm{p}\, n_d, 0)^{\mathrm{T}} =: \boldsymbol{f}_{\mathrm{W}}^1(\boldsymbol{w}, \boldsymbol{n}), \tag{6.58}$$

where the pressure satisfies the relation $\mathrm{p} = (\gamma - 1)(w_m - (w_2^2 + \cdots + w_{m-1}^2)/(2w_1))$. Then, taking into account (6.48) and (6.58), for $\Gamma \in \mathcal{F}_h^W$ we can put

$$\int_\Gamma \mathbf{H}(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{w}_\Gamma^{(R)}, \boldsymbol{n}_\Gamma) \cdot \boldsymbol{\varphi}_h \, \mathrm{d}S = \int_\Gamma \boldsymbol{f}_{\mathrm{W}}^1(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{n}_\Gamma) \cdot \boldsymbol{\varphi}_h \, \mathrm{d}S, \quad \Gamma \in \mathcal{F}_h^W. \tag{6.59}$$

For the purpose of the solution strategy developed in Section 6.4, we introduce a linearization of $\boldsymbol{f}_{\mathrm{W}}^1$. By virtue of (6.28), we have

$$\sum_{s=1}^{d} \boldsymbol{f}_s(\boldsymbol{w})\, n_s = \boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}) = \mathbb{P}(\boldsymbol{w}, \boldsymbol{n}) \boldsymbol{w} \quad \forall\, \boldsymbol{w} \in \mathcal{D}\ \forall\, \boldsymbol{n} = (n_1, \dots, n_d) \in \mathrm{B}_1. \tag{6.60}$$

Our aim is to introduce a matrix (denoted by $\mathbb{P}_W$ hereafter), which is the simplest possible and such that

$$\mathbb{P}(\boldsymbol{w}, \boldsymbol{n}) \boldsymbol{w} = \mathbb{P}_W(\boldsymbol{w}, \boldsymbol{n}) \boldsymbol{w} \tag{6.61}$$

provided that $\boldsymbol{w} \in \mathcal{D}$ and $\boldsymbol{n} \in \mathrm{B}_1$ satisfy the impermeability condition $\boldsymbol{v} \cdot \boldsymbol{n} = 0$, where $\boldsymbol{v}$ is the velocity vector corresponding to $\boldsymbol{w}$. Taking into account the explicit expression (6.24) for $\mathbb{P}$, we remove some of its entries and define the matrix

$$\mathbb{P}_W(\boldsymbol{w}, \boldsymbol{n}) = (\gamma - 1) \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ |\boldsymbol{v}|^2\, n_1/2 & -v_1 n_1 & \dots & -v_d n_1 & n_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ |\boldsymbol{v}|^2\, n_d/2 & -v_1 n_d & \dots & -v_d n_d & n_d \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}, \tag{6.62}$$

where $\boldsymbol{w} \in \mathcal{D}$, $\boldsymbol{n} = (n_1, \dots, n_d) \in \mathrm{B}_1$, $v_j = w_{j+1}/w_1$, $j = 1, \dots, d$, are the components of the velocity vector and $|\boldsymbol{v}|^2 = v_1^2 + \cdots + v_d^2$. We can verify by a simple calculation that (6.61) is valid.

Moreover, we define the linearized form of $\boldsymbol{f}_{\mathrm{W}}^1$ by

$$\boldsymbol{f}_{\mathrm{W}}^{1,\mathrm{L}}(\bar{\boldsymbol{w}}, \boldsymbol{w}, \boldsymbol{n}) = \mathbb{P}_W(\bar{\boldsymbol{w}}, \boldsymbol{n}) \boldsymbol{w}, \quad \bar{\boldsymbol{w}}, \boldsymbol{w} \in \mathcal{D},\ \boldsymbol{n} \in \mathrm{B}_1, \tag{6.63}$$

which is linear with respect to the argument $\boldsymbol{w}$. Obviously, due to (6.58), (6.61) and (6.63), we find that under the condition $\boldsymbol{v} \cdot \boldsymbol{n} = 0$, the linearized form $\boldsymbol{f}_{\mathrm{W}}^{1,\mathrm{L}}$ is consistent with $\boldsymbol{f}_{\mathrm{W}}^1$, i.e.,

$$\boldsymbol{f}_{\mathrm{W}}^{1,\mathrm{L}}(\boldsymbol{w}, \boldsymbol{w}, \boldsymbol{n}) = \boldsymbol{f}_{\mathrm{W}}^1(\boldsymbol{w}, \boldsymbol{n}) \quad \forall\, \boldsymbol{w} \in \mathcal{D}\ \forall\, \boldsymbol{n} \in \mathrm{B}_1 \text{ such that } \boldsymbol{v} \cdot \boldsymbol{n} = 0. \tag{6.64}$$

Exercise 6.9. *Verify relation* (6.61) *for* $\mathbb{P}_W$ *given by* (6.62)*, provided* $\boldsymbol{v} \cdot \boldsymbol{n} = 0$.

Inviscid mirror boundary conditions

This approach is based on the definition of the state vector $\boldsymbol{w}_\Gamma^{(R)}$, $\Gamma \in \mathcal{F}_h^W$ in the form

$$\boldsymbol{w}_\Gamma^{(R)} = \mathscr{M}(\boldsymbol{w}_\Gamma^{(L)}), \tag{6.65}$$

where the boundary operator $\mathscr{M}$, called the *inviscid mirror operator*, is defined in the following way. If $\boldsymbol{w} \in \mathcal{D}$, $\boldsymbol{w} = (\rho, \rho \boldsymbol{v}, E)^{\mathrm{T}}$ and $\boldsymbol{n} \in \mathrm{B}_1$ is the outer unit normal to $\partial\Omega$ at a point in consideration lying on $\partial\Omega_W$, then we set

$$\boldsymbol{v}^\perp = \boldsymbol{v} - 2(\boldsymbol{v} \cdot \boldsymbol{n}) \boldsymbol{n}, \tag{6.66}$$

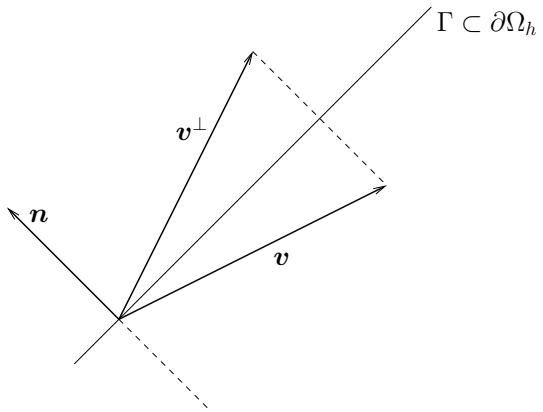

Figure 6.1: Impermeability conditions defined by the mirror operator, vectors of velocity of $\boldsymbol{v}$ and $\boldsymbol{v}^\perp = \boldsymbol{v} - 2(\boldsymbol{v}\cdot\boldsymbol{n})\boldsymbol{n}$.

and

$$\mathscr{M}(\boldsymbol{w}) = (\rho, \rho\boldsymbol{v}^\perp, E)^{\mathrm{T}}. \tag{6.67}$$

The vectors $\boldsymbol{v}$ and $\boldsymbol{v}^\perp$ have the same tangential component but opposite normal components, see Figure 6.1. Obviously, the operator $\mathscr{M}$ is linear.

Now we define the mapping $\boldsymbol{f}^2_{\mathrm{W}} : \mathcal{D} \times \mathrm{B}_1 \to \mathbb{R}^m$ by

$$\boldsymbol{f}^2_{\mathrm{W}}(\boldsymbol{w}, \boldsymbol{n}) = \mathbf{H}(\boldsymbol{w}, \mathscr{M}(\boldsymbol{w}), \boldsymbol{n}) \tag{6.68}$$

and, if $\Gamma \in \mathcal{F}_h^W$, then in (6.53) we have

$$\mathbf{H}(\boldsymbol{w}^{(L)}_\Gamma, \boldsymbol{w}^{(R)}_\Gamma, \boldsymbol{n}_\Gamma) = \boldsymbol{f}^2_{\mathrm{W}}(\boldsymbol{w}^{(L)}_\Gamma, \boldsymbol{n}_\Gamma). \tag{6.69}$$

6.3.2 Boundary conditions on the inlet and outlet

The definition of the boundary state $\boldsymbol{w}^{(R)}_\Gamma$ in (6.53) for $\Gamma \in \mathcal{F}_h^{io} \subset \partial\Omega_i \cup \partial\Omega_o$ (i.e., $\Gamma \subset \partial\Omega_i \cup \partial\Omega_o$) is more delicate. The determination of the inlet/outlet boundary conditions is usually based on a given state vector function $\boldsymbol{w}_{\mathrm{BC}}$ prescribed on $(\partial\Omega_i \cup \partial\Omega_o) \times (0, T)$. For example, when we solve flow around an isolated profile, the state vector $\boldsymbol{w}_{\mathrm{BC}}$ corresponds to the unperturbed far-field flow (flow at infinity). For flow in a channel, the state vector $\boldsymbol{w}_{\mathrm{BC}}$ may correspond to a flow at the inlet and outlet of the channel.

However, since system (6.8) is hyperbolic, we cannot simply put $\boldsymbol{w}^{(R)}_\Gamma = \boldsymbol{w}_{\mathrm{BC}}$. As we shall show later (see also [FFS03]), for a *linear hyperbolic* system with one space variable

$$\frac{\partial \boldsymbol{q}}{\partial t} + \bar{\mathbb{A}}\frac{\partial \boldsymbol{q}}{\partial x} = 0, \quad (x,t) \in (-\infty, 0) \times (0, \infty), \tag{6.70}$$

where $\boldsymbol{q} : (-\infty, 0) \times [0, \infty) \to \mathbb{R}^m$ and $\bar{\mathbb{A}}$ is a constant $m \times m$ matrix, only some quantities defining $\boldsymbol{q}$ at $x = 0$ can be prescribed, whereas other quantities have to be extrapolated from the interior of the computational domain. We shall see that the number of prescribed components of $\boldsymbol{q}$ is equal to the number of negative eigenvalues of $\bar{\mathbb{A}}$.

However, for *nonlinear hyperbolic* systems the theory is missing. Therefore, a usual approach is to choose the boundary conditions in such a way that a linearized initial-boundary value problem is well-posed, i.e., it has a unique solution. We describe this method in the following part of this section.

Approach based on the solution of the linearized Riemann problem

Let $\Gamma \in \mathcal{F}_h^{io}$ and let $x_\Gamma \in \Gamma$ be a point in consideration, at which we want to determine boundary conditions. We introduce a new coordinate system $(\tilde{x}_1, \ldots, \tilde{x}_d)$ such that the coordinate origin lies at the point x_Γ, the axis $\tilde{x}_1$ is parallel to the normal direction $\boldsymbol{n}$ to the boundary, and the coordinate axes $\tilde{x}_2, \ldots, \tilde{x}_d$ are tangential to the boundary, see Figure 6.2. This transformation of the space coordinates is carried out by the mapping $\tilde{x} = \mathbb{Q}_0(\boldsymbol{n})(x - x_\Gamma)$, where $\mathbb{Q}_0(\boldsymbol{n})$ is the rotation matrix defined by (6.34) for $d = 2$ and (6.35) for $d = 3$.

Let $\boldsymbol{w}^{(L)}_\Gamma$ be the value of the trace of the state vector $\boldsymbol{w}$ on Γ from the interior of Ω at the point x_Γ and let

$$\boldsymbol{q}^{(L)}_\Gamma = \mathbb{Q}(\boldsymbol{n}_\Gamma)\boldsymbol{w}^{(L)}_\Gamma, \tag{6.71}$$

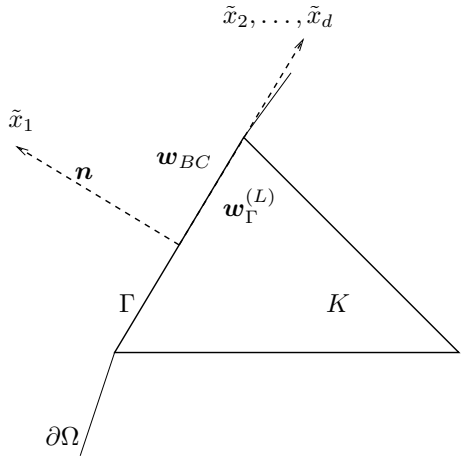

Figure 6.2: The new coordinate system $(\tilde{x}_1, \ldots, \tilde{x}_d)$.

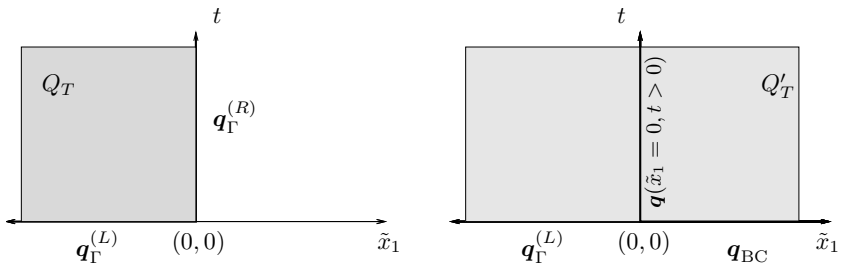

Figure 6.3: Initial-boundary value problem (6.72)–(6.73) (left) and the Riemann problem (6.74)–(6.75) (right), the computational domains $(-\infty, 0) \times (0, \infty)$ and $(-\infty, \infty) \times (0, \infty)$ are grey.

where $\mathbb{Q}(\boldsymbol{n}_\Gamma)$ is given by (6.33).

Using rotational invariance of the Euler equations introduced in Lemma 6.6, e), we transform them to the coordinates $\tilde{x}_1, \ldots, \tilde{x}_d$, neglect the derivative with respect to $\tilde{x}_j,\ j = 2, \ldots, d$, and linearize the resulting system around the state $\boldsymbol{q}_\Gamma^{(L)}$. Then we obtain the linear system

$$\frac{\partial \boldsymbol{q}}{\partial t} + \mathbb{A}_1(\boldsymbol{q}_\Gamma^{(L)}) \frac{\partial \boldsymbol{q}}{\partial \tilde{x}_1} = 0, \quad (\tilde{x}_1, t) \in (-\infty, 0) \times [0, \infty) \tag{6.72}$$

for the transformed vector-valued function $\boldsymbol{q} = \mathbb{Q}(\boldsymbol{n}_\Gamma)\boldsymbol{w}$, see Figure 6.3, left. To this system we add the initial and boundary conditions

$$\begin{aligned} \boldsymbol{q}(\tilde{x}_1, 0) &= \boldsymbol{q}_\Gamma^{(L)}, \quad \tilde{x}_1 < 0, \\ \boldsymbol{q}(0, t) &= \boldsymbol{q}_\Gamma^{(R)}, \quad t > 0, \end{aligned} \tag{6.73}$$

where $\boldsymbol{q}_\Gamma^{(L)}$ is given by (6.71) and the unknown state vector $\boldsymbol{q}_\Gamma^{(R)}$ should be determined in such a way that it reflects the state vector $\boldsymbol{q}_{\rm BC} = \mathbb{Q}(\boldsymbol{n}_\Gamma)\boldsymbol{w}_{\rm BC}$ with a prescribed state $\boldsymbol{w}_{\rm BC}$, and the initial-boundary value problem (6.72) – (6.73) is well–posed, i.e., has a unique solution.

In order to find the vector $\boldsymbol{q}_\Gamma^{(R)}$, we consider the *linearized Riemann problem*

$$\frac{\partial \boldsymbol{q}}{\partial t} + \mathbb{A}_1(\boldsymbol{q}_\Gamma^{(L)}) \frac{\partial \boldsymbol{q}}{\partial \tilde{x}_1} = 0, \quad (\tilde{x}_1, t) \in (-\infty, \infty) \times [0, \infty) \tag{6.74}$$

with the initial condition

$$\boldsymbol{q}(\tilde{x}_1, 0) = \begin{cases} \boldsymbol{q}_\Gamma^{(L)}, & \text{if } \tilde{x}_1 < 0, \\ \boldsymbol{q}_{\rm BC}, & \text{if } \tilde{x}_1 > 0, \end{cases} \tag{6.75}$$

see Figure 6.3 (right).

The exact solution of problem (6.74)–(6.75) can be found by the method of characteristics in the following way: Let $\boldsymbol{g}_s,\ s=1,\dots,m$, be the eigenvectors corresponding to the eigenvalues $\tilde{\lambda}_s,\ s=1,\dots,m$, of the matrix $\mathbb{A}_1=\mathbb{A}_1(\boldsymbol{q}_\Gamma^{(L)})$. Hence, $\mathbb{A}_1\boldsymbol{g}_s=\tilde{\lambda}_s\boldsymbol{g}_s,\ s=1,\dots,m$.

Taking into account (6.32), we see that the eigenvalues of the matrices $\mathbb{A}_1(\boldsymbol{q}_\Gamma^{(L)})$ and $\mathbb{P}(\boldsymbol{w}_\Gamma^{(L)},\boldsymbol{n}_\Gamma)$ attain the same values, i.e.,

$$\tilde{\lambda}_s=\lambda_s,\quad s=1,\dots,m, \tag{6.76}$$

where λ_s are the eigenvalues of the matrix $\mathbb{P}(\boldsymbol{w}_\Gamma^{(L)},\boldsymbol{n}_\Gamma)$.

The explicit formulae for the eigenvectors $\boldsymbol{g}_s,\ s=1,\dots,m$, can be found in [FFS03], Section 3.1. These eigenvectors form a basis of $\mathbb{R}^m$, and thus the exact solution of (6.74) can be written in the form

$$\boldsymbol{q}(\tilde{x}_1,t)=\sum_{s=1}^{m}\mu_s(\tilde{x}_1,t)\boldsymbol{g}_s,\quad \tilde{x}_1\in\mathbb{R},\ t>0, \tag{6.77}$$

where $\mu_s,\ s=1,\dots,m$, are unknown functions defined in $(-\infty,\infty)\times[0,\infty)$. Similarly, the initial states from (6.75) can be expressed as

$$\boldsymbol{q}_\Gamma^{(L)}=\sum_{s=1}^{m}\alpha_s\boldsymbol{g}_s,\quad \boldsymbol{q}_{\mathrm{BC}}=\sum_{s=1}^{m}\beta_s\boldsymbol{g}_s. \tag{6.78}$$

The vectors $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_m)$ and $\boldsymbol{\beta}=(\beta_1,\dots,\beta_m)$ are given by the relations

$$\boldsymbol{\alpha}=\mathbb{T}^{-1}\boldsymbol{q}_\Gamma^{(L)},\quad \boldsymbol{\beta}=\mathbb{T}^{-1}\boldsymbol{q}_{\mathrm{BC}}, \tag{6.79}$$

where $\mathbb{T}$ is the $m\times m$-matrix whose columns are the eigenvectors $\boldsymbol{g}_s,\ s=1,\dots,m$. The functions $\mu_s,\ s=1,\dots,m$, are called the *characteristic variables*.

Substituting (6.77) into (6.74), we get

$$0=\sum_{s=1}^{m}\left(\frac{\partial\mu_s}{\partial t}+\tilde{\lambda}_s\frac{\partial\mu_s}{\partial\tilde{x}_1}\right)\boldsymbol{g}_s,\quad s=1,\dots,m, \tag{6.80}$$

which holds if and only if

$$\frac{\partial\mu_s}{\partial t}+\tilde{\lambda}_s\frac{\partial\mu_s}{\partial\tilde{x}_1}=0,\quad \tilde{x}_1\in\mathbb{R},\ t>0,\ s=1,\dots,m. \tag{6.81}$$

These equations are equipped with initial conditions following from (6.75) and (6.78)

$$\mu_s(\tilde{x}_1,0)=\bar{\mu}_s(\tilde{x}_1):=\begin{cases}\alpha_s, & \tilde{x}_1<0,\\ \beta_s, & \tilde{x}_1>0,\end{cases}\quad s=1,\dots,m. \tag{6.82}$$

We can simply verify that the exact solution of (6.81) – (6.82) reads

$$\mu_s(\tilde{x}_1,t)=\bar{\mu}_s(\tilde{x}_1-\tilde{\lambda}_s t),\quad \tilde{x}_1\in\mathbb{R},\ t\geq 0,$$

where $\bar{\mu}_s$ is given by (6.82). This together with (6.82) gives

$$\mu_s(\tilde{x}_1,t)=\begin{cases}\alpha_s, & \text{if } \tilde{x}_1-\tilde{\lambda}_s t<0,\\ \beta_s, & \text{if } \tilde{x}_1-\tilde{\lambda}_s t>0,\end{cases}\quad s=1,\dots,m. \tag{6.83}$$

We define the sought state $\boldsymbol{q}_\Gamma^{(R)}$ as the solution of problem (6.74)–(6.75) at $\tilde{x}_1=0$. Hence, we put $\boldsymbol{q}_\Gamma^{(R)}=\boldsymbol{q}(0,t)$, and by (6.77) and (6.83), we get

$$\boldsymbol{q}_\Gamma^{(R)}=\sum_{s=1}^{m}\eta_s\boldsymbol{g}_s,\quad \eta_s=\begin{cases}\alpha_s, & \tilde{\lambda}_s\geq 0,\\ \beta_s, & \tilde{\lambda}_s<0.\end{cases} \tag{6.84}$$

Finally, we introduce the inlet/outlet *boundary operator* based on the solution of the linearized Riemann problem

$$\mathscr{B}^{\mathrm{LRP}}(\boldsymbol{w}_\Gamma^{(L)},\boldsymbol{w}_{\mathrm{BC}}):=\mathbb{Q}^{-1}(\boldsymbol{n}_\Gamma)\boldsymbol{q}_\Gamma^{(R)}. \tag{6.85}$$

Then we define the sought boundary state

$$\boldsymbol{w}_\Gamma^{(R)}:=\mathscr{B}^{\mathrm{LRP}}(\boldsymbol{w}_\Gamma^{(L)},\boldsymbol{w}_{\mathrm{BC}}). \tag{6.86}$$

flow regime	$m_{\rm pr}$	$m_{\rm ex}$
supersonic inlet	m	0
subsonic inlet	$m-1$	1
subsonic outlet	1	$m-1$
supersonic outlet	0	m

Table 6.1: Boundary conditions based on the well-posedness of the linearized problem: number of prescribed $m_{\rm pr}$ and extrapolated $m_{\rm ex}$ components of $\boldsymbol{w}$ for subsonic/supersonic inlet/outlet.

Remark 6.10. *From the above process (taking into account (6.76)) we can conclude that the sought boundary state $\boldsymbol{w}_\Gamma^{(R)}$ is determined using $m_{\rm pr}$ quantities characterizing the prescribed boundary state $\boldsymbol{w}_{\rm BC}$, where $m_{\rm pr}$ is the number of negative eigenvalues of the matrix $\mathbb{P}(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{n}_\Gamma)$, whereas we extrapolate $m_{\rm ex}$ quantities defining the state $\boldsymbol{w}_\Gamma^{(L)}$, where $m_{\rm ex} = m - m_{\rm pr}$ is the number of nonnegative eigenvalues of the matrix $\mathbb{P}(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{n}_\Gamma)$.*

This observation is in agreement with the definitions of boundary conditions on impermeable walls. Taking into account that by (6.30) the eigenvalues of the matrix $\mathbb{P}(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{n}_\Gamma)$ read

$$\lambda_1 = \boldsymbol{v}\cdot\boldsymbol{n} - a, \quad \lambda_2 = \ldots = \lambda_{d+1} = \boldsymbol{v}\cdot\boldsymbol{n}, \quad \lambda_{d+2} = \boldsymbol{v}\cdot\boldsymbol{n} + a, \tag{6.87}$$

where $\boldsymbol{v}$ and a represent the velocity vector and the speed of sound, respectively, corresponding to the state $\boldsymbol{w}_\Gamma^{(L)}$, and $\boldsymbol{n} = \boldsymbol{n}_\Gamma$. Then the impermeability condition $\boldsymbol{v}\cdot\boldsymbol{n} = 0$ implies that $\lambda_1 < 0$, $\lambda_2 = \ldots = \lambda_{d+1} = 0$, $\lambda_{d+2} > 0$. Hence, in this case we prescribe only one quantity, namely $\boldsymbol{v}\cdot\boldsymbol{n} = 0$ or the opposite normal component $-\boldsymbol{v}\cdot\boldsymbol{n}$ of the velocity vector and the remaining quantities defining the state $\boldsymbol{w}_\Gamma^{(R)}$ are obtained by extrapolation.

Approach based on physical properties of the flow

It follows from the above considerations and the form (6.87) of eigenvalues λ_s, $s = 1, \ldots, m = d+2$, that in the case of the inlet or outlet, on which $\boldsymbol{v}\cdot\boldsymbol{n} < 0$ or $\boldsymbol{v}\cdot\boldsymbol{n} > 0$, respectively, it is necessary to distinguish between the subsonic or supersonic regime, when $|\boldsymbol{v}\cdot\boldsymbol{n}| < a$ or $|\boldsymbol{v}\cdot\boldsymbol{n}| > a$, respectively. The number of prescribed and extrapolated boundary conditions for the mentioned possibilities is shown in Table 6.1.

On the basis of these results, it is possible to introduce a widely used method for determining the inlet/outlet boundary conditions based on the use of physical variables. In this approach we extrapolate or prescribe directly some physical variables. Particularly, we distinguish the following cases:

- *supersonic inlet*, $m_{\rm pr} = m$, we prescribe all components of the boundary state $\boldsymbol{w}_\Gamma^{(R)}$. Hence, we set $\boldsymbol{w}_\Gamma^{(R)} = \boldsymbol{w}_{\rm BC}$,
- *subsonic inlet*, $m_{\rm pr} = m-1$, we extrapolate the pressure from the interior of the domain, and prescribe the density and the components of the velocity on the boundary,
- *subsonic outlet*, $m_{\rm pr} = 1$, we prescribe the pressure and extrapolate the density and the components of the velocity on the boundary,
- *supersonic outlet*, $m_{\rm pr} = 0$, we extrapolate all components of $\boldsymbol{w}$ from the interior of Ω on the boundary. This means that we set $\boldsymbol{w}_\Gamma^{(R)} = \boldsymbol{w}_\Gamma^{(L)}$.

Hence, we define the inlet/outlet boundary operator based on physical variables:

$$\mathscr{B}^{\rm phys}(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{w}_{\rm BC}) = \begin{cases} \boldsymbol{w}_{\rm BC} & \text{if } \boldsymbol{v}\cdot\boldsymbol{n} < -a & \text{supersonic inlet} \\ \mathsf{Phys}(\rho_{\rm BC},\, \boldsymbol{v}_{\rm BC},\, \mathrm{p}_\Gamma^{(L)}) & \text{if } -a \le \boldsymbol{v}\cdot\boldsymbol{n} < 0 & \text{subsonic inlet} \\ \mathsf{Phys}(\rho_\Gamma^{(L)},\, \boldsymbol{v}_\Gamma^{(L)},\, \mathrm{p}_{\rm BC}) & \text{if } 0 < \boldsymbol{v}\cdot\boldsymbol{n} \le a & \text{subsonic outlet} \\ \boldsymbol{w}_\Gamma^{(L)} & \text{if } \boldsymbol{v}\cdot\boldsymbol{n} > a & \text{supersonic outlet} \end{cases} \tag{6.88}$$

where $\rho_{\rm BC}$, $\boldsymbol{v}_{\rm BC}$, $\mathrm{p}_{\rm BC}$ are the density, the velocity vector and the pressure, respectively, corresponding to the prescribed state $\boldsymbol{w}_{\rm BC}$ and $\rho_\Gamma^{(L)}$, $\boldsymbol{v}_\Gamma^{(L)}$, $\mathrm{p}_\Gamma^{(L)}$ denote the density, the velocity vector and the pressure corresponding to $\boldsymbol{w}_\Gamma^{(L)}$. The symbol Phys denotes the transformation from the physical variables to the conservative ones, namely, for $\rho > 0$, $\mathrm{p} > 0$ and $\boldsymbol{v} \in \mathbb{R}^d$ we set

$$\mathsf{Phys}(\rho, \boldsymbol{v}, \mathrm{p}) = \big(\rho,\ \rho\boldsymbol{v},\ \mathrm{p}/(\gamma-1) + \rho|\boldsymbol{v}|^2/2\big)^{\mathrm{T}} \in \mathbb{R}^m. \tag{6.89}$$

This approach is usually used with success for the transonic flow. However, its application to low Mach number flows does not give reasonable results, because these boundary conditions are not transparent for acoustic waves coming from inside of

the computational domain Ω. In numerical simulations, we observe some reflection from the inlet/outlet parts of the boundary. Therefore, in a low Mach number flow, it is suitable to apply the method based on the solution of a linearized Riemann problem. This means that the boundary state $\boldsymbol{w}_\Gamma^{(R)}$ is defined by (6.86). Another more sophisticated method will be treated in the following section.

Boundary conditions based on the exact solution of the nonlinear Riemann problem

The generalization of the method based on the solution of the linearized Riemann problem uses the exact solution of the nonlinear Riemann problem. The only difference is that we do not linearize the system of the Euler equations around the state $\boldsymbol{w}_\Gamma^{(L)}$, but instead of (6.72) we consider the nonlinear system

$$\frac{\partial \boldsymbol{q}}{\partial t} + \mathbb{A}_1(\boldsymbol{q})\frac{\partial \boldsymbol{q}}{\partial \tilde{x}_1} = 0, \quad (\tilde{x}_1, t) \in (-\infty, 0) \times [0, \infty) \tag{6.90}$$

with the initial and boundary conditions (6.73). This means that instead of (6.74), we consider the *Riemann problem*

$$\frac{\partial \boldsymbol{q}}{\partial t} + \mathbb{A}_1(\boldsymbol{q})\frac{\partial \boldsymbol{q}}{\partial \tilde{x}_1} = 0, \quad (\tilde{x}_1, t) \in (-\infty, \infty) \times [0, \infty) \tag{6.91}$$

equipped with the initial condition (6.75). The solution of problem (6.91), (6.75) is much more complicated than the solution of the linearized problem (6.74) – (6.75) but for the Euler equations it can be constructed analytically, see e.g., [FFS03, Section 3.1.6] or [Wes01, Section 10.2]. This analytical solution contains an implicit formula for the pressure p, which has to be obtained iteratively.

When the solution $\boldsymbol{q}$ of the Riemann problem (6.91), (6.75) is obtained, then we define the inlet/outlet boundary operator based on the solution of the nonlinear Riemann problem as

$$\mathscr{B}^{\mathrm{RP}}(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{w}_{\mathrm{BC}}) := \mathbb{Q}^{-1}(\boldsymbol{n}_\Gamma)\boldsymbol{q}(0, t) \tag{6.92}$$

and set $\boldsymbol{w}_\Gamma^{(R)} := \mathscr{B}^{\mathrm{RP}}(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{w}_{\mathrm{BC}})$.

Finally, based on the presented approaches to the choice of boundary conditions we specify the definition (6.53) of the form $\boldsymbol{b}_h$ by

$$\begin{aligned}
\boldsymbol{b}_h(\boldsymbol{w}, \boldsymbol{\varphi}) = &- \sum_{K \in \mathcal{T}_h} \int_K \sum_{s=1}^{d} \boldsymbol{f}_s(\boldsymbol{w}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_s} \,\mathrm{d}x \\
&+ \sum_{\Gamma \in \mathcal{F}_h^I} \int_\Gamma \mathbf{H}(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{w}_\Gamma^{(R)}, \boldsymbol{n}_\Gamma) \cdot [\boldsymbol{\varphi}] \,\mathrm{d}S \\
&+ \sum_{\Gamma \in \mathcal{F}_h^W} \int_\Gamma \boldsymbol{f}_{\mathrm{W}}^i(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{n}_\Gamma) \cdot \boldsymbol{\varphi} \,\mathrm{d}S \\
&+ \sum_{\Gamma \in \mathcal{F}_h^{io}} \int_\Gamma \mathbf{H}\left(\boldsymbol{w}_\Gamma^{(L)}, \mathscr{B}(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{w}_{\mathrm{BC}}), \boldsymbol{n}_\Gamma\right) \cdot \boldsymbol{\varphi} \,\mathrm{d}S,
\end{aligned} \tag{6.93}$$

where $i = 1$ or $i = 2$, if we use the impermeability boundary condition (6.58) or (6.68), respectively. Moreover, the inlet/outlet boundary operator $\mathscr{B}$ represents $\mathscr{B}^{\mathrm{phys}}$, $\mathscr{B}^{\mathrm{LRP}}$ and $\mathscr{B}^{\mathrm{RP}}$ given by (6.88), (6.85) and (6.92), respectively.

Remark 6.11. *The definitions of the boundary operators $\mathscr{B}^{\mathrm{phys}}$, $\mathscr{B}^{\mathrm{LRP}}$ and $\mathscr{B}^{\mathrm{RP}}$ and of the form $\boldsymbol{b}_h$ and their evaluations may seem to be rather complicated and CPU time demanding. However, it is necessary to take into account that the integrals appearing in the definition of the form $\boldsymbol{b}_h$ are computed with the aid of numerical integration and the boundary conditions have to be determined only at integration points.*

6.4 Time discretization

The space semidiscrete problem (6.54) represents a system of ordinary differential equations (ODEs), which has to be solved with the aid of suitable numerical schemes. In the framework of the finite difference and finite volume methods, the explicit Euler or Runge–Kutta time discretization is very popular for solving the Euler equations. In early works on the DGM for the Euler equations ([CS89], [BR97b], [BO99]), explicit time discretization was also used. Their advantage is a simple algorithmization, but on the other hand, the size of the time step τ is strongly restricted by the *Courant–Friedrichs–Lewy (CFL) stability condition* written, for example, in the form

$$\tau \le \mathrm{CFL} \min_{\substack{K \in \mathcal{T}_h \\ \Gamma \subset \partial K}} \frac{|K|}{\varrho(\mathbb{P}(\boldsymbol{w}_h, \boldsymbol{n})|_\Gamma)|\Gamma|}, \tag{6.94}$$

where $\varrho(\mathbb{P}(\boldsymbol{w}_h,\boldsymbol{n})|_\Gamma)$ denotes the spectral radius of the matrix $\mathbb{P}(\boldsymbol{w}_h,\boldsymbol{n})|_\Gamma$ given by (6.17) and evaluated at the points of $\Gamma\in\mathcal{F}_h$, $|K|$ is the d-dimensional measure of $K\in\mathcal{T}_h$ and $|\Gamma|$ denotes the $(d-1)$-dimensional measure of $\Gamma\in\mathcal{F}_h$. Moreover, $0<\mathrm{CFL}\le 1$ is the Courant–Friedrichs–Lewy (CFL) number. Our numerical experiments indicate that whereas the value $\mathrm{CFL}=0.85$ was sufficient for almost all flow regimes in finite volume computations, the P_1 discontinuous Galerkin approximation requires the value $\mathrm{CFL}\approx 0.15$ in order to guarantee stability. Moreover, the stability condition (6.94) becomes more and more restrictive with increasing polynomial approximation degree p.

Therefore, it is suitable to consider implicit methods for numerically solving compressible flow problems, see, e.g., [BR00], [BBHN09], [HH06a], [HH06b]. It is well known that the use of implicit methods contributes to improving the efficiency of numerical schemes for solving the Euler equations in some cases, because implicit methods allow using longer time steps. In the framework of the finite volume methods, implicit schemes were used, for example in [Sto85], [FS89] and [Mei98]. The drawback of the implicit schemes is having to solve a large nonlinear algebraic system on each time level. To this end, the Newton method is often applied leading to a sequence of linear discrete problems. One variant of this approach is a well-known Δ-scheme by Beam and Warming [BW76], [BW78], see also [Hir88]. This approach is often combined with multigrid techniques, see e.g., [HS86], [KH91], [Dic91].

The application of the Newton schemes requires, of course, the differentiability of the numerical flux and the computation of its partial derivatives, which is usually rather complicated. This is the reason that some authors use artificial pseudo-time-integration, as was applied together with multigrid in [vdVvdV02a] and [vdVvdV02b] for the DG discrete problem. Multigrid techniques usually require using structured meshes and, in the case of the mesh refinement, a sequence of nested meshes. This is not the case when the anisotropic mesh adaptation (AMA) method is used. Then the algebraic multigrid would have to be applied, but its efficiency is not high. Therefore, one often uses the Krylov subspace methods for solving linear systems in linearized schemes for the Euler equations (cf., e.g., [Mei98]).

In the following we will be concerned with developing several numerical schemes for the full space-time discretization of the Euler equations. The presented techniques were developed on the basis of results from [DF03], [DF04a], [DFS03], [FDK06], [FDK07], [FK07].

6.4.1 Backward Euler method

The implicit *backward Euler* time discretization of (6.54) is the simplest implicit method for numerically solving ODEs. It can be formally considered either as the first-order implicit Runge–Kutta method or as the first-order backward difference formula (BDF), or as the first-order time discontinuous Galerkin method, see [HNW00], [Tho06]. The higher-order time discretizations will be discussed in Section 6.4.5.

In what follows we consider a partition $0=t_0<t_1<t_2\cdots<t_r=T$ of the time interval $[0,T]$ and set $\tau_k=t_k-t_{k-1}$ for $k=1,\dots,r$. We use the symbol $\boldsymbol{w}_h^k$ for the approximation of $\boldsymbol{w}_h(t_k)$, $k=1,\dots,r$.

Using the backward Euler scheme for the time discretization of (6.54), we can define the following method for the numerical solution of problem (6.8).

Definition 6.12. *We say that the finite sequence of functions $\boldsymbol{w}_h^k,\ k=0,\dots,r$, is an* approximate solution *of problem (6.8) obtained by the* backward Euler – discontinuous Galerkin method *(BE-DGM) if the following conditions are satisfied:*

$$\boldsymbol{w}_h^k\in\boldsymbol{S}_{hp},\ k=0,1,\dots,r, \tag{6.95a}$$

$$\frac{1}{\tau_k}\left(\boldsymbol{w}_h^k-\boldsymbol{w}_h^{k-1},\boldsymbol{\varphi}_h\right)+\boldsymbol{b}_h(\boldsymbol{w}_h^k,\boldsymbol{\varphi}_h)=0\quad\forall\,\boldsymbol{\varphi}_h\in\boldsymbol{S}_{hp},\ k=1,\dots,r, \tag{6.95b}$$

$$\boldsymbol{w}_h^0=\Pi_h\boldsymbol{w}^0, \tag{6.95c}$$

where $\Pi_h\boldsymbol{w}^0$ is the $\boldsymbol{S}_{hp}$-approximation (usually $L^2(\Omega)$-projection on the space $\boldsymbol{S}_{hp}$) of the function $\boldsymbol{w}^0$ from the initial condition (6.36).

Remark 6.13. *The BE-DGM has formally the order of convergence $O(h^{p+1}+\tau)$ in the $L^\infty(0,T;(L^2(\Omega))^m)$-norm and the order of convergence $O(h^p+\tau)$ in the $L^2(0,T;(H^1(\Omega))^m)$-seminorm, provided that the exact solution is sufficiently regular. These results were verified numerically in [DF04a] and [Dol13a].*

Problem (6.95) represents a nonlinear algebraic system for each $k=1,\dots,r$. Its solution will be discussed in the following sections. First, we shall present its solution with the aid of the standard Newton method [Deu04]. Then we shall develop a Newton-like method based on the approximation of the Jacobi matrix by the flux matrix.

6.4.2 Newton method based on the Jacobi matrix

In order to develop the solution strategy for the nonlinear systems (6.95b), we introduce their algebraic representation. Let N_{hp} denote the dimension of the space $\boldsymbol{S}_{hp}$ and let $\mathrm{B}_{hp}=\{\boldsymbol{\varphi}_i(x),\ i=1,\dots,N_{hp}\}$ denote a set of linearly independent functions forming a basis of $\boldsymbol{S}_{hp}$. It is possible to construct a basis B_{hp} as a composition of local bases constructed separately for each $K\in\mathcal{T}_h$. See Section 6.4.8, where one possibility is described in detail.

Any function $\boldsymbol{w}_h^k \in \boldsymbol{S}_{hp}$ can be expressed in the form

$$\boldsymbol{w}_h^k(x) = \sum_{j=1}^{N_{hp}} \xi^{k,j} \boldsymbol{\varphi}_j(x) \in \boldsymbol{S}_{hp} \iff \boldsymbol{\xi}_k = (\xi^{k,j})_{j=1}^{N_{hp}} \in \mathbb{R}^{N_{hp}},\ k = 1, \dots, r, \tag{6.96}$$

where $\xi^{k,j} \in \mathbb{R},\ j = 1, \dots, N_{hp},\ k = 1, \dots, r$, are its basis coefficients. Obviously, (6.96) defines an isomorphism between $\boldsymbol{w}_h^k \in \boldsymbol{S}_{hp}$ and $\boldsymbol{\xi}_k \in \mathbb{R}^{N_{hp}}$. We call $\boldsymbol{\xi}_k$ the *algebraic representation* of $\boldsymbol{w}_h^k$.

In order to rewrite the nonlinear algebraic systems (6.95b), we define the vector-valued function $\boldsymbol{F}_h : \mathbb{R}^{N_{hp}} \times \mathbb{R}^{N_{hp}} \to \mathbb{R}^{N_{hp}}$ by

$$\boldsymbol{F}_h\left(\boldsymbol{\xi}_{k-1}; \boldsymbol{\xi}_k\right) = \left(\frac{1}{\tau_k} \left(\boldsymbol{w}_h^k - \boldsymbol{w}_h^{k-1}, \boldsymbol{\varphi}_i \right) + \boldsymbol{b}_h(\boldsymbol{w}_h^k, \boldsymbol{\varphi}_i) \right)_{i=1}^{N_{hp}},\ k = 1, \dots, r, \tag{6.97}$$

where $\boldsymbol{\xi}_{k-l} \in \mathbb{R}^{N_{hp}}$ is the algebraic representation of $\boldsymbol{w}_h^{k-l} \in \boldsymbol{S}_{hp}$ for $l = 0, 1$. We do not emphasize that $\boldsymbol{F}_h$ depends explicitly on τ_k. Therefore, the algebraic representation of the systems (6.95b) reads: For a given vector $\boldsymbol{\xi}_{k-1} \in \mathbb{R}^{N_{hp}}$ find $\boldsymbol{\xi}_k \in \mathbb{R}^{N_{hp}}$ such that

$$\boldsymbol{F}_h(\boldsymbol{\xi}_{k-1}; \boldsymbol{\xi}_k) = \boldsymbol{0}, \quad k = 1, \dots, r. \tag{6.98}$$

Here $\boldsymbol{0}$ denotes a generic zero vector (i.e., all entries of $\boldsymbol{0}$ are equal to zero) and $\boldsymbol{\xi}_0$ is given by the initial condition (6.95c) and the isomorphism (6.96). Systems (6.98) are strongly nonlinear and their efficient and accurate solution is demanding.

A natural strategy is to apply the (damped) *Newton method* ([Deu04]) which generates a sequence of approximations $\boldsymbol{\xi}_k^l,\ l = 0, 1, \dots$, to the actual numerical solution $\boldsymbol{\xi}_k$ using the following algorithm. Given an iterate $\boldsymbol{\xi}_k^l \in \mathbb{R}^{N_{hp}}$, the update of $\boldsymbol{\xi}_k^l$ reads

$$\boldsymbol{\xi}_k^{l+1} = \boldsymbol{\xi}_k^l + \lambda^l \boldsymbol{\delta}^l, \tag{6.99}$$

where $\boldsymbol{\delta}^l \in \mathbb{R}^{N_{hp}}$ is defined as the solution of the system

$$\mathbb{D}_h(\boldsymbol{\xi}_k^l)\boldsymbol{\delta}^l = -\boldsymbol{F}_h(\boldsymbol{\xi}_{k-1}; \boldsymbol{\xi}_k^l). \tag{6.100}$$

Here $\lambda^l \in (0, 1]$ is the damping parameter (for its choice see Section 6.4.4) and $\mathbb{D}_h$ is the *Jacobi matrix* of the vector-valued function $\boldsymbol{F}_h$ given by (6.97), i.e.,

$$\mathbb{D}_h(\boldsymbol{\xi}_k^l) = \frac{\mathrm{D}\boldsymbol{F}_h(\boldsymbol{\xi}_{k-1}; \boldsymbol{\xi}_k^l)}{\mathrm{D}\boldsymbol{\xi}_k^l}. \tag{6.101}$$

From (6.96), (6.97) and (6.101) we obtain

$$\begin{aligned} \mathbb{D}_h(\boldsymbol{\xi}_k) =& (d_{ij}(\boldsymbol{\xi}_k))_{i,j=1}^{N_{hp}}, \\ d_{ij}(\boldsymbol{\xi}_k) =& \frac{1}{\tau_k}(\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i) + \frac{\partial \boldsymbol{b}_h\left(\sum_{l=1}^{N_{hp}} \xi^{k,l} \boldsymbol{\varphi}_l, \boldsymbol{\varphi}_i\right)}{\partial \xi^{k,j}}, \quad i, j = 1, \dots, N_{hp}. \end{aligned} \tag{6.102}$$

For $\lambda^l = 1$ we get the standard Newton method. This technique was also successfully applied in [HH06a], [BR00] for computing viscous flow.

Evaluating of the Jacobi matrix $\mathbb{D}_h$ is not quite easy, since the form $\boldsymbol{b}_h$ depends nonlinearly on its first argument. Moreover, there are difficulties with the differentiability of the mapping $\boldsymbol{F}_h$, because the numerical flux $\mathbf{H}$ is sometimes only Lipschitz-continuous, but not differentiable.

In the following section we present an alternative approach inspired by the semi-implicit technique from [DF04a], [FK07] and based on the so-called flux matrix.

6.4.3 Newton-like method based on the flux matrix

Evaluating of the Jacobi matrix $\mathbb{D}_h$ in (6.100) can be avoided with the aid of a formal linearization of the convection form $\boldsymbol{b}_h$. The aim is to define the form $\boldsymbol{b}_h^L : \boldsymbol{S}_{hp} \times \boldsymbol{S}_{hp} \times \boldsymbol{S}_{hp} \to \mathbb{R}$ such that it is linear with respect to its second and third arguments and is consistent with $\boldsymbol{b}_h$, i.e.,

$$\boldsymbol{b}_h(\boldsymbol{w}_h, \boldsymbol{\varphi}_h) = \boldsymbol{b}_h^L(\boldsymbol{w}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) - \tilde{\boldsymbol{b}}_h(\boldsymbol{w}_h, \boldsymbol{\varphi}_h) \qquad \forall\, \boldsymbol{w}_h, \boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp}, \tag{6.103}$$

where $\tilde{\boldsymbol{b}}_h : \boldsymbol{S}_{hp} \times \boldsymbol{S}_{hp} \to \mathbb{R}$ is some "residual" form, vanishing for the majority of functions $\boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp}$, see (6.121).

By (6.93), we defined the form

$$\begin{aligned}
\boldsymbol{b}_h(\boldsymbol{w}_h, \boldsymbol{\varphi}_h) = &- \sum_{K\in\mathcal{T}_h} \int_K \sum_{s=1}^{d} \boldsymbol{f}_s(\boldsymbol{w}_h) \cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s} \,\mathrm{d}x && (=: \eta_1) \\
&+ \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \mathbf{H}(\boldsymbol{w}_{h\Gamma}^{(L)}, \boldsymbol{w}_{h\Gamma}^{(R)}, \boldsymbol{n}_\Gamma) \cdot [\boldsymbol{\varphi}_h] \,\mathrm{d}S && (=: \eta_2) \\
&+ \sum_{\Gamma\in\mathcal{F}_h^W} \int_\Gamma \boldsymbol{f}_{\mathrm{W}}^i(\boldsymbol{w}_{h\Gamma}^{(L)}, \boldsymbol{n}) \cdot \boldsymbol{\varphi}_h \,\mathrm{d}S && (=: \eta_3) \\
&+ \sum_{\Gamma\in\mathcal{F}_h^{io}} \int_\Gamma \mathbf{H}\left(\boldsymbol{w}_{h\Gamma}^{(L)}, \mathscr{B}(\boldsymbol{w}_{h\Gamma}^{(L)}, \boldsymbol{w}_{\mathrm{BC}}), \boldsymbol{n}_\Gamma\right) \cdot \boldsymbol{\varphi}_h \mathrm{d}S && (=: \eta_4),
\end{aligned} \tag{6.104}$$

where $\boldsymbol{w}_{h\Gamma}^{(L)}$ and $\boldsymbol{w}_{h\Gamma}^{(R)}$ denote the traces of $\boldsymbol{w}_h$ on $\Gamma \in \mathcal{F}_h$, cf. (6.41). The individual terms $\eta_1, \dots, \eta_4$ will be partially linearized.

For η_1 we use the property (6.26) of the Euler fluxes and define the form $\eta_1^{\mathrm{L}} : \boldsymbol{S}_{hp} \times \boldsymbol{S}_{hp} \times \boldsymbol{S}_{hp} \to \mathbb{R}$ by

$$\eta_1^{\mathrm{L}}(\bar{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) = - \sum_{K\in\mathcal{T}_h} \int_K \sum_{s=1}^{d} \mathbb{A}_s(\bar{\boldsymbol{w}}_h)\boldsymbol{w}_h \cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s} \,\mathrm{d}x. \tag{6.105}$$

Obviously, $\eta_1^{\mathrm{L}}(\boldsymbol{w}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) = \eta_1$ and η_1^{L} is linear with respect to its second and third arguments.

Linearizing of the term η_2 can be carried out on the basis of a suitable choice of the numerical flux $\mathbf{H}$. For example, let us use in (6.104) the Vijayasundaram numerical flux, see [Vij86], [Fei93, Section 7.3] or [FFS03, Section 3.3.4]. It is defined in the following way. By (6.29), the matrix $\mathbb{P} = \mathbb{P}(\boldsymbol{w}, \boldsymbol{n})$ defined in (6.17) is diagonalizable: there exists a nonsingular matrix $\mathbb{T} = \mathbb{T}(\boldsymbol{w}, \boldsymbol{n})$ such that

$$\mathbb{P} = \mathbb{T}\boldsymbol{\Lambda}\mathbb{T}^{-1}, \tag{6.106}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}\,(\lambda_1, \dots, \lambda_m)$ and $\lambda_1, \dots, \lambda_m$ are the eigenvalues of $\mathbb{P}$. The columns of the matrix $\mathbb{T}$ are the eigenvectors of the matrix $\mathbb{P}$. We define the "positive" and "negative" part of $\mathbb{P}$ by

$$\mathbb{P}^{\pm} = \mathbb{T}\boldsymbol{\Lambda}^{\pm}\mathbb{T}^{-1},\ \boldsymbol{\Lambda}^{\pm} = \mathrm{diag}\,(\lambda_1^{\pm}, \dots, \lambda_m^{\pm}), \tag{6.107}$$

where $a^+ = \max(a, 0)$ and $a^- = \min(a, 0)$ for $a \in \mathbb{R}$. Then the *Vijayasundaram numerical flux* reads as

$$\mathbf{H}_{VS}(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{n}) = \mathbb{P}^+\left(\frac{\boldsymbol{w}_1 + \boldsymbol{w}_2}{2}, \boldsymbol{n}\right)\boldsymbol{w}_1 + \mathbb{P}^-\left(\frac{\boldsymbol{w}_1 + \boldsymbol{w}_2}{2}, \boldsymbol{n}\right)\boldsymbol{w}_2. \tag{6.108}$$

We can characterize the properties of the Vijayasundaram numerical flux.

Lemma 6.14. *The Vijayasundaram numerical flux $\mathbf{H}_{VS} = \mathbf{H}(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{n})$ is Lipschitz-continuous with respect to $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{D}$ and satisfies conditions* (6.49) *and* (6.50)*, i.e., it is consistent and conservative.*

Proof. (a) From (6.10) it follows that the entries of the matrix $\mathbb{P}$ are continuously differentiable. This fact, the definition of the matrices $\mathbb{P}^{\pm}$, definition (6.108) and the Lipschitz-continuity of the functions $\lambda \in \mathbb{R} \to \lambda^+$ and $\lambda \in \mathbb{R} \to \lambda^-$ imply that the Vijayasundaram numerical flux is locally Lipschitz-continuous.

(b) The consistency of $\mathbf{H}_{VS}$ is a consequence of the relations (6.16), (6.28) and $\mathbb{P}(\boldsymbol{w}, \boldsymbol{n}) = \mathbb{P}^+(\boldsymbol{w}, \boldsymbol{n}) + \mathbb{P}^-(\boldsymbol{w}, \boldsymbol{n})$.

(c) The proof of the consistency of $\mathbf{H}_{VS}$ is more complicated. First, we show that

$$\mathbb{P}^{\pm}(\boldsymbol{w}, -\boldsymbol{n}) = -\mathbb{P}^{\mp}(\boldsymbol{w}, \boldsymbol{n}) \tag{6.109}$$

for $\boldsymbol{w} \in \mathcal{D}$ and $\boldsymbol{n} = (n_1, \dots, n_d) \in B_1$. It follows from (6.16) that

$$\boldsymbol{P}(\boldsymbol{w}, -\boldsymbol{n}) = -\boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}).$$

By differentiation,

$$\mathbb{P}(\boldsymbol{w}, -\boldsymbol{n}) = -\mathbb{P}(\boldsymbol{w}, \boldsymbol{n}),$$

and thus

$$\mathbb{P}^{\pm}(\boldsymbol{w}, -\boldsymbol{n}) = (-\mathbb{P}(\boldsymbol{w}, \boldsymbol{n}))^{\pm}. \tag{6.110}$$

Further, by (6.106),

$$\pm\mathbb{P}(\boldsymbol{w},\,\boldsymbol{n}) = \mathbb{T}(\boldsymbol{w},\boldsymbol{n})\,(\pm\boldsymbol{\Lambda}(\boldsymbol{w},\,\boldsymbol{n}))\,\mathbb{T}^{-1}(\boldsymbol{w},\,\boldsymbol{n}),$$

where

$$\boldsymbol{\Lambda}(\boldsymbol{w},\,\boldsymbol{n}) = \mathrm{diag}\,(\lambda_1(\boldsymbol{w},\,\boldsymbol{n}),\ldots,\lambda_m(\boldsymbol{w},\,\boldsymbol{n}))\,.$$

Thus,

$$\mathbb{P}^{\pm}(\boldsymbol{w},\,\boldsymbol{n}) = \mathbb{T}(\boldsymbol{w},\,\boldsymbol{n})\,\boldsymbol{\Lambda}^{\pm}(\boldsymbol{w},\,\boldsymbol{n})\,\mathbb{T}^{-1}(\boldsymbol{w},\,\boldsymbol{n}) \tag{6.111}$$

and

$$(-\mathbb{P}(\boldsymbol{w},\,\boldsymbol{n}))^{\pm} = \mathbb{T}(,\boldsymbol{w},\,\boldsymbol{n})\,(-\boldsymbol{\Lambda}(\boldsymbol{w},\,\boldsymbol{n}))^{\pm}\,\mathbb{T}^{-1}(\boldsymbol{w},\,\boldsymbol{n}). \tag{6.112}$$

Here

$$\begin{aligned}\boldsymbol{\Lambda}^{\pm}(\boldsymbol{w},\,\boldsymbol{n}) &= \mathrm{diag}\left(\lambda_1^{\pm}(\boldsymbol{w},\,\boldsymbol{n}),\ldots,\lambda_m^{\pm}(\boldsymbol{w},\,\boldsymbol{n})\right),\\ (-\boldsymbol{\Lambda}(\boldsymbol{w},\,\boldsymbol{n}))^{\pm} &= \mathrm{diag}\left((-\lambda_1)^{\pm}(\boldsymbol{w},\,\boldsymbol{n}),\ldots,(-\lambda_m)^{\pm}(\boldsymbol{w},\,\boldsymbol{n})\right),\end{aligned}$$

It is easy to find that $(-\lambda)^{\pm} = -\lambda^{\mp}$, which implies that

$$(-\boldsymbol{\Lambda}(\boldsymbol{w},\,n))^{\pm} = -\boldsymbol{\Lambda}^{\mp}(\boldsymbol{w},\,n).$$

The above, (6.111) and (6.112) yield

$$\begin{aligned}(-\mathbb{P}(\boldsymbol{w},\,\boldsymbol{n}))^{\pm} &= -\mathbb{T}(\boldsymbol{w},\,\boldsymbol{n})\,\boldsymbol{\Lambda}^{\mp}(\boldsymbol{w},\,n)\,\mathbb{T}(\boldsymbol{w},\,\boldsymbol{n}) \\ &= -\mathbb{P}^{\mp}(\boldsymbol{w},\,\boldsymbol{n}).\end{aligned} \tag{6.113}$$

Now, by (6.110) and (6.113) we get (6.109).

Finally, by virtue of (6.109), for $\boldsymbol{w}_1,\boldsymbol{w}_2 \in \mathcal{D}$ and $\boldsymbol{n} \in B_1$,

$$\begin{aligned}&\mathbf{H}_{VS}(\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{n}) = \mathbb{P}^{+}\left(\frac{w_1+\boldsymbol{w}_2}{2},\,\boldsymbol{n}\right)\boldsymbol{w}_1 + \mathbb{P}^{-}\left(\frac{w_1+\boldsymbol{w}_2}{2},\,\boldsymbol{n}\right)\boldsymbol{w}_2\\ &= -\mathbb{P}^{-}\left(\frac{w_1+\boldsymbol{w}_2}{2},\,-\boldsymbol{n}\right)\boldsymbol{w}_1 - \mathbb{P}^{+}\left(\frac{w_1+\boldsymbol{w}_2}{2},\,-\boldsymbol{n}\right)\boldsymbol{w}_2 = -\mathbf{H}_{VS}(\boldsymbol{w}_2,\boldsymbol{w}_1,-\boldsymbol{n}),\end{aligned}$$

which is what we wanted to prove. □ □

The form of $\mathbf{H}_{VS}$ is a way of defining the form $\eta_2^{\mathrm{L}} : \boldsymbol{S}_{hp}\times\boldsymbol{S}_{hp}\times\boldsymbol{S}_{hp}\to\mathbb{R}$ by

$$\eta_2^{\mathrm{L}}(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \sum_{\Gamma\in\mathcal{F}_h^I}\int_{\Gamma}\left[\mathbb{P}^{+}\left(\langle\bar{\boldsymbol{w}}_h\rangle_{\Gamma},\boldsymbol{n}_{\Gamma}\right)\boldsymbol{w}_{h\Gamma}^{(L)} + \mathbb{P}^{-}\left(\langle\bar{\boldsymbol{w}}_h\rangle_{\Gamma},\boldsymbol{n}_{\Gamma}\right)\boldsymbol{w}_{h\Gamma}^{(R)}\right]\cdot\boldsymbol{\varphi}_h\mathrm{d}S, \tag{6.114}$$

where $\langle\bar{\boldsymbol{w}}_h\rangle_{\Gamma}$ denotes the mean value of $\bar{\boldsymbol{w}}_h$ on $\Gamma\in\mathcal{F}_h$ defined by (6.42). Obviously, $\eta_2^{\mathrm{L}}(\boldsymbol{w}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \eta_2$ and η_2^{L} is linear with respect to its second and third arguments.

Concerning the term η_3 in (6.104), we distinguish between the direct use of the impermeability condition and the inviscid mirror boundary condition presented in Section 6.3.1. For the former case we define the form

$$\eta_3^{\mathrm{L}}(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \sum_{\Gamma\in\mathcal{F}_h^W}\int_{\Gamma}\boldsymbol{f}_{\mathrm{W}}^{1,\mathrm{L}}(\bar{\boldsymbol{w}}_{h\Gamma}^{(L)},\boldsymbol{w}_{h\Gamma}^{(L)},\boldsymbol{n})\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S, \tag{6.115}$$

where $\boldsymbol{f}_{\mathrm{W}}^{1,\mathrm{L}}$ is defined by (6.63), i.e.,

$$\boldsymbol{f}_{\mathrm{W}}^{1,\mathrm{L}}(\bar{\boldsymbol{w}},\boldsymbol{w},\boldsymbol{n}) = \mathbb{P}_W(\bar{\boldsymbol{w}},\boldsymbol{n})\boldsymbol{w},\quad \bar{\boldsymbol{w}},\boldsymbol{w}\in\mathcal{D},\ \boldsymbol{n}\in\mathrm{B}_1, \tag{6.116}$$

with $\mathbb{P}_W$ given in (6.62).

In the case of inviscid mirror boundary conditions we use relations (6.68) and (6.108) and put

$$\boldsymbol{f}_{\mathrm{W}}^{2,\mathrm{L}}(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{n}) = \mathbb{P}^{+}\left(\bar{\boldsymbol{w}}_h,\boldsymbol{n}\right)\boldsymbol{w}_h + \mathbb{P}^{-}\left(\bar{\boldsymbol{w}}_h,\boldsymbol{n}\right)\mathscr{M}(\boldsymbol{w}_h), \tag{6.117}$$

where $\mathbb{P}^{\pm}$ are defined by (6.107). Now, on the basis of (6.68), (6.69) and (6.117), we put

$$\eta_3^{\mathrm{L}}(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \sum_{\Gamma\in\mathcal{F}_h^W}\int_{\Gamma}\boldsymbol{f}_{\mathrm{W}}^{2,\mathrm{L}}(\bar{\boldsymbol{w}}_{h\Gamma}^{(L)},\boldsymbol{w}_{h\Gamma}^{(L)},\boldsymbol{n})\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S. \tag{6.118}$$

Therefore, (6.115) and (6.118) can be written as

$$\eta_3^{\mathrm{L}}(\bar{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) = \sum_{\Gamma \in \mathcal{F}_h^W} \int_\Gamma \boldsymbol{f}_{\mathrm{W}}^{\alpha,\mathrm{L}}(\bar{\boldsymbol{w}}_{h\Gamma}^{(L)}, \boldsymbol{w}_{h\Gamma}^{(L)}, \boldsymbol{n}) \cdot \boldsymbol{\varphi}_h \,\mathrm{d}S, \tag{6.119}$$

where $\alpha = 1$ for directly using the impermeability condition and $\alpha = 2$ for the inviscid mirror boundary condition. It follows from (6.116) – (6.119) and the linearity of the operator $\mathscr{M}$ that η_3^{L} is linear with respect to its second and third arguments. Moreover, $\eta_3^{\mathrm{L}}(\boldsymbol{w}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) = \eta_3$.

Finally, η_4 is approximated with the aid of the forms

$$\eta_4^{\mathrm{L}}(\bar{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) = \sum_{\Gamma \in \mathcal{F}_h^{io}} \int_\Gamma \left(\mathbb{P}^+(\bar{\boldsymbol{w}}_{h\Gamma}^{(L)}, \boldsymbol{n}_\Gamma) \boldsymbol{w}_{h\Gamma}^{(L)} \right) \cdot \boldsymbol{\varphi}_h \mathrm{d}S, \tag{6.120}$$

and

$$\tilde{\boldsymbol{b}}_h(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_h) = -\sum_{\Gamma \in \mathcal{F}_h^{io}} \int_\Gamma \left(\mathbb{P}^-(\bar{\boldsymbol{w}}_{h\Gamma}^{(L)}, \boldsymbol{n}_\Gamma) \mathscr{B}(\bar{\boldsymbol{w}}_\Gamma^{(L)}, \boldsymbol{w}_{\mathrm{BC}}) \right) \cdot \boldsymbol{\varphi}_h \mathrm{d}S, \tag{6.121}$$

where $\mathscr{B}$ represents the boundary operators $\mathscr{B}^{\mathrm{phys}}$, $\mathscr{B}^{\mathrm{LRP}}$ and $\mathscr{B}^{\mathrm{RP}}$ given by (6.88), (6.85) and (6.92), respectively. Let us underline that in the arguments of $\mathbb{P}^\pm$ we do not use the mean value of the state vectors from the left- and right-hand side of Γ as in (6.108). Moreover, if $\operatorname{supp} \boldsymbol{\varphi}_h \cap (\partial\Omega_i \cup \partial\Omega_o) = \emptyset$, then $\tilde{\boldsymbol{b}}_h(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_h) = 0$.

Obviously, due to (6.93) and (6.120), we have

$$\eta_4^{\mathrm{L}}(\boldsymbol{w}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) - \tilde{\boldsymbol{b}}_h(\boldsymbol{w}_h, \boldsymbol{\varphi}_h) = \eta_4. \tag{6.122}$$

Taking into account (6.93), (6.105), (6.114), (6.119) and (6.120), we introduce the form

$$\begin{aligned} \boldsymbol{b}_h^L(\bar{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) = &-\sum_{K \in \mathcal{T}_h} \int_K \sum_{s=1}^d \mathbb{A}_s(\bar{\boldsymbol{w}}_h) \boldsymbol{w}_h \cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s} \,\mathrm{d}x \\ &+ \sum_{\Gamma \in \mathcal{F}_h^I} \int_\Gamma \left[\mathbb{P}^+ \left(\langle \bar{\boldsymbol{w}}_h \rangle_\Gamma, \boldsymbol{n}_\Gamma \right) \boldsymbol{w}_{h\Gamma}^{(L)} + \mathbb{P}^- \left(\langle \bar{\boldsymbol{w}}_h \rangle_\Gamma, \boldsymbol{n}_\Gamma \right) \boldsymbol{w}_{h\Gamma}^{(R)} \right] \cdot \boldsymbol{\varphi}_h \mathrm{d}S \\ &+ \sum_{\Gamma \in \mathcal{F}_h^W} \int_\Gamma \boldsymbol{f}_{\mathrm{W}}^{\alpha,\mathrm{L}}(\bar{\boldsymbol{w}}_{h\Gamma}^{(L)}, \boldsymbol{w}_{h\Gamma}^{(L)}, \boldsymbol{n}) \cdot \boldsymbol{\varphi}_h \,\mathrm{d}S \\ &+ \sum_{\Gamma \in \mathcal{F}_h^{io}} \int_\Gamma \left(\mathbb{P}^+ \left(\bar{\boldsymbol{w}}_{h\Gamma}^{(L)}, \boldsymbol{n}_\Gamma \right) \boldsymbol{w}_{h\Gamma}^{(L)} + \mathbb{P}^-(\bar{\boldsymbol{w}}_{h\Gamma}^{(L)}, \boldsymbol{n}_\Gamma) \mathscr{B}(\bar{\boldsymbol{w}}_\Gamma^{(L)}, \boldsymbol{w}_{\mathrm{BC}}) \right) \cdot \boldsymbol{\varphi}_h \mathrm{d}S. \end{aligned} \tag{6.123}$$

From the definitions (6.93) of $\boldsymbol{b}_h$, (6.123) of $\boldsymbol{b}_h^L$ and (6.121) of $\tilde{\boldsymbol{b}}_h$ we can see that relation (6.103) is valid. Moreover, the form $\boldsymbol{b}_h^L$ is linear with respect to the arguments $\boldsymbol{w}_h$ and $\boldsymbol{\varphi}_h$.

Now we introduce the Newton-like method for solving systems (6.98) based on the flux matrix. We again return to the algebraic representation of the method. Using notation from Section 6.4.2, we define the $N_{hp} \times N_{hp}$ *flux matrix*

$$\mathbb{C}_h\left(\bar{\boldsymbol{\xi}}\right) = \left(\frac{1}{\tau_k} \left(\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i \right) + \boldsymbol{b}_h^L(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i) \right)_{i,j=1}^{N_{hp}} \tag{6.124}$$

and the vector

$$\boldsymbol{d}_h\left(\boldsymbol{\xi}_{k-1}, \bar{\boldsymbol{\xi}}\right) = \left(\frac{1}{\tau_k} \left(\boldsymbol{w}_h^{k-1}, \boldsymbol{\varphi}_i \right) + \tilde{\boldsymbol{b}}_h(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_i) \right)_{i=1}^{N_{hp}}, \tag{6.125}$$

where $\boldsymbol{\varphi}_i \in \mathrm{B}_{hp},\ i = 1, \ldots, N_{hp}$, are the basis functions in the space $\boldsymbol{S}_{hp}$, $\bar{\boldsymbol{\xi}} \in \mathbb{R}^{N_{hp}}$ and $\boldsymbol{\xi}_{k-l} \in \mathbb{R}^{N_{hp}},\ l = 0, 1$, are the algebraic representations of $\bar{\boldsymbol{w}}_h \in \boldsymbol{S}_{hp}$ and $\boldsymbol{w}_h^{k-l} \in \boldsymbol{S}_{hp},\ l = 0, 1$, respectively. (We do not emphasize that $\mathbb{C}_h$ and $\boldsymbol{d}_h$ depend explicitly on τ_k.) Finally, using (6.97), (6.103) and (6.124) – (6.125), we have

$$\boldsymbol{F}_h(\boldsymbol{\xi}_{k-1}; \boldsymbol{\xi}_k) = \mathbb{C}_h(\boldsymbol{\xi}_k)\boldsymbol{\xi}_k - \boldsymbol{d}_h(\boldsymbol{\xi}_{k-1}, \boldsymbol{\xi}_k), \quad k = 1, \ldots, r. \tag{6.126}$$

Obviously, the sparsity of $\mathbb{C}_h$ is identical with the sparsity of the Jacobi matrix $\mathbb{D}_h$ introduced in (6.101). Therefore, in the following Newton-like method for solving systems (6.98), we use $\mathbb{C}_h$ as the approximation of $\mathbb{D}_h$ in the definition of our iterative Newton-like method, which is represented as the following algorithm.

If the approximate solution $\boldsymbol{w}_h^{k-1} \in \boldsymbol{S}_{hp}$, represented by $\boldsymbol{\xi}_{k-1}$, was already computed, then we set $\boldsymbol{\xi}_k^0 = \boldsymbol{\xi}_{k-1}$ and apply the iterative process

$$\boldsymbol{\xi}_k^{l+1} = \boldsymbol{\xi}_k^l + \lambda^l \boldsymbol{\delta}^l, \quad l = 0, 1, \ldots, \tag{6.127}$$

with $\boldsymbol{\delta}^l$ defined by

$$\mathbb{C}_h(\boldsymbol{\xi}_k^l)\boldsymbol{\delta}^l = -\boldsymbol{F}_h(\boldsymbol{\xi}_{k-1}; \boldsymbol{\xi}_k^l). \tag{6.128}$$

The term $\lambda^l \in (0, 1]$ is a damping parameter which should ensure convergence of (6.127) – (6.128) in case when the initial guess $\boldsymbol{\xi}_k^0$ is far from the solution of (6.98). The initial guess $\boldsymbol{\xi}_k^0$ can be defined as

$$\boldsymbol{\xi}_k^0 = \boldsymbol{\xi}_{k-1}, \quad k = 1, \ldots, r, \tag{6.129}$$

where $\boldsymbol{\xi}_{k-1}$ corresponds to the approximate solution $\boldsymbol{w}_h^{k-1}$.

In the following section we discuss several aspects of the iterative method (6.127) – (6.128).

Remark 6.15. *Let us note that if we carry out only one Newton-like iteration at each time level, put $\lambda^0 = 1$, and the matrix $\mathbb{C}_h$ is updated at each time step, then the implicit method* (6.95) *reduces to the* semi-implicit time discretization *approach presented in [DF04a] and [FK07]. It can be formulated in the following way: We seek the finite sequence of functions $\boldsymbol{w}_h^k$, $k = 0, 1, \ldots, r$, such that*

$$\boldsymbol{w}_h^k \in \boldsymbol{S}_{hp}, \ k = 0, 1, \ldots, r, \tag{6.130a}$$

$$\frac{1}{\tau_k}\left(\boldsymbol{w}_h^k - \boldsymbol{w}_h^{k-1}, \boldsymbol{\varphi}_h\right) + \hat{\boldsymbol{b}}_h(\boldsymbol{w}_h^{k-1}, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h) = 0 \quad \forall \boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp}, \ k = 1, \ldots, r, \tag{6.130b}$$

$$\boldsymbol{w}_h^0 = \Pi_h \boldsymbol{w}^0, \tag{6.130c}$$

where $\Pi_h \boldsymbol{w}^0$ is the $\boldsymbol{S}_{hp}$-approximation of $\boldsymbol{w}^0$ from the initial condition (6.36) and

$$\hat{\boldsymbol{b}}_h(\bar{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) = \boldsymbol{b}_h^L(\bar{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) - \tilde{\boldsymbol{b}}_h(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_h) \tag{6.131}$$

with $\boldsymbol{b}_h^L$ and $\tilde{\boldsymbol{b}}_h$ given by (6.123) *and* (6.121)*, respectively.*

6.4.4 Realization of the iterative algorithm

In this section we mention some aspects of the Newton-like iterative process (6.127) – (6.128).

Choice of damping parameters

The damping parameters λ^l, $l = 0, 1, \ldots$, should guarantee convergence of the iterative process (6.127) – (6.128). Following the analysis presented in [Deu04], we start from the value $\lambda^l = 1$ and evaluate a monitoring function

$$\kappa^l = \frac{\left\| \boldsymbol{F}_h(\boldsymbol{\xi}_{k-1}; \boldsymbol{\xi}_k^{l+1}) \right\|}{\left\| \boldsymbol{F}_h(\boldsymbol{\xi}_{k-1}; \boldsymbol{\xi}_k^{l}) \right\|}, \tag{6.132}$$

where $\|\cdot\|$ is a norm in the space $\mathbb{R}^{N_{hp}}$. If $\kappa^l < 1$, we proceed to the next Newton-like iteration. Otherwise, we put $\lambda^l = \lambda^l/2$ and repeat the actual l^{th} Newton-like iteration.

Update of the flux matrix

As numerical experiments show in the iterative process it is not necessary to update the flux matrix $\mathbb{C}_h(\boldsymbol{\xi}_k^l)$ at each Newton-like iteration $l = 1, 2, \ldots$ and each time level $k = 1, \ldots, r$. Computational costs of the evaluation of $\boldsymbol{F}_h$ are much smaller than the evaluation of $\mathbb{C}_h$. For simplicity, let us consider the case $d = 2$ and assume that $\mathcal{T}_h$ is a conforming triangulation. By $\#\mathcal{T}_h$ we denote the number of elements of $\mathcal{T}_h$. Then $\boldsymbol{F}_h$ has $N_{hp} = \#\mathcal{T}_h(p+1)(p+1)/2$ components and $\mathbb{C}_h$ has approximately $4\#\mathcal{T}_h((p+1)(p+1)/2)^2$ non-vanishing components. Hence, the evaluation of $\boldsymbol{F}_h$ is approximately $2(p+1)(p+2)$-times cheaper than the evaluation of $\mathbb{C}_h$.

Therefore, it is more efficient to perform more Newton-like iterations than to update the matrix $\mathbb{C}_h$. In practice, we update $\mathbb{C}_h$, when either the damping parameter λ achieves a minimal prescribed value (using the algorithm described in Section 6.4.4) or the prescribed maximal number of Newton-like iterations is achieved.

Termination of the iterative process

The iterative process (6.127) – (6.128) is terminated if a suitable *algebraic stopping criterion* is achieved. The standard approach is based on the condition

$$\left\| \boldsymbol{F}_h(\boldsymbol{\xi}_{k-1}; \boldsymbol{\xi}_k^l) \right\| \leq \mathrm{TOL}, \tag{6.133}$$

where $\|\cdot\|$ is a norm in $\mathbb{R}^{N_{hp}}$ and TOL is a given tolerance. However, it is difficult to choose TOL in order to guarantee the accuracy of the solution and to avoid a too long iterative process. The optimal stopping criterion, which balances the accuracy and efficiency, should be derived from a posteriori estimates taking into account algebraic errors. This is out of the scope of this monograph and we refer, for example, to [CS07] and [AEV11], dealing with this subject. In [Dol13a] a heuristic stopping criterion solving this problem was proposed.

Solution of the linear algebraic systems (6.128)

The linear algebraic systems (6.128) can be solved by a direct solver (e.g., UMFPACK [DD99]) in case that the number of unknowns is not high (the limit value is usually 10^5). In general, iterative solvers are more efficient, because a good initial approximation is obtained from the previous Newton-like iteration or the previous time level. Usually it is necessary to compute only a few iterations. Among the iterative solvers, very efficient are the Krylov subspace methods, see [LS13].

It is possible to apply, e.g., the GMRES method ([SS86]) with block diagonal or block ILU(0) preconditioning ([DHH11]). Usually, the GMRES iterative process is stopped, when the preconditioned residuum is two times smaller than the initial one. This criterion may seem to be too weak, but numerical experiments show that it is sufficient in a number of applications.

6.4.5 Higher-order time discretization

In Section 6.4.1, we have introduced the space-time discretization of the Euler equations (6.8) with the aid of the backward Euler – discontinuous Galerkin method (BE-DGM). However, by virtue of Remark 6.13, this method is only of the first order in time. In solving nonstationary flows, it is necessary to apply schemes that are sufficiently accurate in space as well as in time. There are several possibilities (see, e.g., [HNW00], [Tho06]) how to obtain a higher-order time discretizations.

We shall mention three techniques having the order n with respect to the time discretization, i.e., the error is of order $O(\tau^n)$:

- *backward difference formula* (BDF) method, which is a multistep method using computed approximate solutions from n previous time levels. On each time level, it is necessary to solve one nonlinear algebraic system with N_{hp} equations, where N_{hp} is the dimension of the space $\boldsymbol{S}_{hp}$. Hence, the BDF method has (approximately) the same computational costs as the backward Euler method.

- *implicit Runge–Kutta* (IRK) method, which is a one-step method and it evaluates several (at least n) stages within one time step. This means that we solve (at least) n-nonlinear algebraic systems with N_{hp} equations at each time level. Hence, the IRK method has approximately n-times higher computational cost than the backward Euler method.

- *time discontinuous Galerkin* (TDG) method, which is based on a polynomial approximation of degree $n-1$ with respect to time. The TDG method was introduced in Section 4.2 for a scalar equation. We solve one nonlinear algebraic system with $n\,N_{hp}$ equations at each time level. As we see, the TDG method has approximately n^2-times higher computational cost than the backward Euler method or the BDF method.

The BDF, IRK and TDG time discretizations reduce to backward Euler method for the limit case $n = 1$. An overview of theoretical aspects of the higher-order time discretization in combination with the DG space discretization can be found in [Vla10].

It follows from the above discussion that the cheapest approach is the BDF technique, which will be described in this section. Again let $0 = t_0 < t_1 < t_2 < \ldots t_r = T$ be a partition of the time interval $[0, T]$, $\tau_k = t_k - t_{k-1},\ k = 1, \ldots r$, and let $\boldsymbol{w}_h^k \in \boldsymbol{S}_{hp}$ denote a piecewise polynomial approximation of $\boldsymbol{w}_h(t_k),\ k = 0, 1, \ldots, r$. We define the following scheme.

Definition 6.16. *We say that the finite sequence of functions $\boldsymbol{w}_h^k,\ k = 0, \ldots, r$, is the* approximate solution *of (6.8) computed by the n-*step backward difference formula – discontinuous Galerkin method *(BDF-DGM) if the following conditions are satisfied:*

	constant time step			variable time step		
	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$
$\alpha_{n,0}$	1	$\frac{3}{2}$	$\frac{11}{6}$	1	$\frac{2\,\theta_k+1}{\theta_k+1}$	$\frac{\theta_k\theta_{k-1}}{\theta_k\theta_{k-1}+\theta_{k-1}+1}+\frac{2\theta_k+1}{\theta_k+1}$
$\alpha_{n,1}$	-1	-2	-3	-1	$-(\theta_k+1)$	$-\frac{(\theta_k+1)(\theta_k\theta_{k-1}+\theta_{k-1}+1)}{\theta_{k-1}+1}$
$\alpha_{n,2}$		$\frac{1}{2}$	$\frac{3}{2}$		$\frac{\theta_k^2}{\theta_k+1}$	$\frac{\theta_k^2(\theta_k\theta_{k-1}+\theta_{k-1}+1)}{\theta_k+1}$
$\alpha_{n,3}$			$-\frac{1}{3}$			$-\frac{(\theta_k+1)\theta_k^2\theta_{k-1}^3}{(\theta_{k-1}+1)(\theta_k\theta_{k-1}+\theta_{k-1}+1)}$

Table 6.2: Values of $\alpha_{n,l},\ l=0,\dots,n$, for $n=1,2,3$ for constant and variable time steps, $\theta_k=\tau_k/\tau_{k-1},\ k=1,2,\dots,r$.

	$n=1$	$n=2$	$n=3$
$\alpha_{n,0}$	1	$\frac{2\tau_k+\tau_{k-1}}{\tau_k+\tau_{k-1}}$	$\frac{(2\tau_k+\tau_{k-1})(2\tau_k+\tau_{k-1}+\tau_{k-2})-\tau_k^2}{(\tau_k+\tau_{k-1})(\tau_k+\tau_{k-1}+\tau_{k-2})}$
$\alpha_{n,1}$	-1	$-\frac{\tau_k+\tau_{k-1}}{\tau_{k-1}}$	$-\frac{(\tau_k+\tau_{k-1})(\tau_k+\tau_{k-1}+\tau_{k-2})}{\tau_{k-1}(\tau_{k-1}+\tau_{k-2})}$
$\alpha_{n,2}$		$\frac{\tau_k^2}{\tau_{k-1}(\tau_k+\tau_{k-1})}$	$\frac{\tau_k^2(\tau_k+\tau_{k-1}+\tau_{k-2})}{\tau_{k-1}\tau_{k-2}(\tau_k+\tau_{k-1})}$
$\alpha_{n,3}$			$-\frac{\tau_k^2(\tau_k+\tau_{k-1})}{\tau_{k-2}(\tau_k+\tau_{k-1}+\tau_{k-2})(\tau_{k-1}+\tau_{k-2})}$

Table 6.3: Values of the coefficients $\alpha_{n,l}$ expressed in terms of the time steps.

$$\boldsymbol{w}_h^k \in \boldsymbol{S}_{hp},\ k=0,1,\dots,r, \tag{6.134a}$$

$$\frac{1}{\tau_k}\left(\sum_{l=0}^{n}\alpha_{n,l}\boldsymbol{w}_h^{k-l},\boldsymbol{\varphi}_h\right)+\boldsymbol{b}_h\left(\boldsymbol{w}_h^k,\boldsymbol{\varphi}_h\right)=0\quad\forall\boldsymbol{\varphi}_h\in\boldsymbol{S}_{hp},\ k=n,\dots,r, \tag{6.134b}$$

$$\boldsymbol{w}_h^0\ \textit{is the } \boldsymbol{S}_{hp}\textit{-approximation (usually } L^2(\Omega)\textit{-projection on } \boldsymbol{S}_{hp})\ \textit{of the initial condition } \boldsymbol{w}^0, \tag{6.134c}$$

$$\boldsymbol{w}_h^l\in\boldsymbol{S}_{hp},\ l=1,\dots,n-1,\ \textit{are determined by a suitable } q\textit{-step method with } q\le l\ \textit{ or by an explicit Runge–Kutta method.} \tag{6.134d}$$

Some Runge–Kutta schemes can be found in Section **??**. Their application to a system of partial differential equations can be written in the same form.

The BDF coefficients $\alpha_{n,l},\ l=0,\dots,n$, depend on time steps $\tau_{k-l},\ l=0,\dots,n$. They can be derived from the Lagrange interpolation of pairs $(t_{k-l},\boldsymbol{w}_h^{k-l}),\ l=0,\dots,n$, see, e.g. [HNW00]. Table 6.2 shows their values in the case of constant and variable time steps for $n=1,2,3$. Obviously, the one-step BDF-DGM is identical with the BE-DGM defined by (6.95). In Table 6.3 these coefficients are expressed directly in terms of the time steps τ_j.

Remark 6.17 (Stability of the BDF-DGM)**.** *The n-step BDF method is unconditionally stable for $n=1$ and $n=2$, and for increasing n the region of stability decreasing. For $n>7$ this method is unconditionally unstable, see [HNW00, Section III.5]. In practice, the n-BDF-DGM with $n=1,2,3$ is usually used.*

Remark 6.18 (Accuracy of the BDF-DGM)**.** *The n-step BDF-DGM has formally the order of convergence $O(h^{p+1}+\tau^n)$ in the $L^\infty(0,T;(L^2(\Omega))^m)$-norm and $O(h^p+\tau^n)$ in the $L^2(0,T;(H^1(\Omega))^m)$-seminorm, provided that the exact solution is sufficiently regular. These orders of convergence were numerically verified for a scalar equation.*

Problem (6.134) represents a nonlinear algebraic system for each $k=1,\dots,r$, which can be solved with the strategy presented in Section 6.4.3.

Again, let N_{hp} denote the dimension of the space $\boldsymbol{S}_{hp}$ of the piecewise polynomial functions and let $\mathrm{B}_{hp}=\{\boldsymbol{\varphi}_i(x),\ i=1,\dots,N_{hp}\}$ be a basis of $\boldsymbol{S}_{hp}$. Using the isomorphism (6.96) between $\boldsymbol{w}_h^k\in\boldsymbol{S}_{hp}$ and $\boldsymbol{\xi}_k\in\mathbb{R}^{N_{hp}}$, we define the vector-valued

function $\boldsymbol{F}_h : (\mathbb{R}^{N_{hp}})^n \times \mathbb{R}^{N_{hp}} \to \mathbb{R}^{N_{hp}}$ by

$$\boldsymbol{F}_h\left(\{\boldsymbol{\xi}_{k-l}\}_{l=1}^n;\boldsymbol{\xi}_k\right) = \left(\frac{1}{\tau_k}\left(\sum_{l=0}^{n}\alpha_{n,l}\boldsymbol{w}_h^{k-l},\boldsymbol{\varphi}_i\right) + \boldsymbol{b}_h(\boldsymbol{w}_h^k,\boldsymbol{\varphi}_i)\right)_{i=1}^{N_{hp}}, \quad k=1,\ldots,r, \tag{6.135}$$

where $\boldsymbol{\xi}_{k-l} \in \mathbb{R}^{N_{hp}}$ is the algebraic representation of $\boldsymbol{w}_h^{k-l} \in \boldsymbol{S}_{hp}$ for $l = 1,\ldots,n$. Then scheme (6.134) has the following algebraic representation. If $\boldsymbol{\xi}_{k-l}$, $l = 1,\ldots,n$, $(k = 1,\ldots,r)$ are given vectors, then we want to find $\boldsymbol{\xi}_k \in \mathbb{R}^{N_{hp}}$ such that

$$\boldsymbol{F}_h(\{\boldsymbol{\xi}_{k-l}\}_{l=1}^n;\boldsymbol{\xi}_k) = \boldsymbol{0}. \tag{6.136}$$

System (6.136) is strongly nonlinear. It can be solved with the aid of the Newton-like method based on the flux matrix, presented in Section 6.4.3. Let $\boldsymbol{b}_h^L$ and $\tilde{\boldsymbol{b}}_h$ be the forms defined by (6.123) and (6.121), respectively. Then (6.103) implies the consistency

$$\boldsymbol{b}_h(\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \boldsymbol{b}_h^L(\boldsymbol{w}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) - \tilde{\boldsymbol{b}}_h(\boldsymbol{w}_h,\boldsymbol{\varphi}_h) \qquad \forall\, \boldsymbol{w}_h,\boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp}, \tag{6.137}$$

where the form $\boldsymbol{b}_h^L$ is defined in (6.123).

We see that instead of (6.124) and (6.125), we define the *flux matrix* $\mathbb{C}_h$ and the vector $\boldsymbol{d}_h$ by

$$\mathbb{C}_h\left(\bar{\boldsymbol{\xi}}\right) = \left(\frac{\alpha_{n,0}}{\tau_k}\left(\boldsymbol{\varphi}_j,\boldsymbol{\varphi}_i\right) + \boldsymbol{b}_h^L(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_j,\boldsymbol{\varphi}_i)\right)_{i,j=1}^{N_{hp}} \tag{6.138}$$

and

$$\boldsymbol{d}_h\left(\{\boldsymbol{\xi}_{k-l}\}_{l=1}^n,\bar{\boldsymbol{\xi}}\right) = \left(\frac{1}{\tau_k}\left(\sum_{i=1}^{n}\alpha_{n,i}\boldsymbol{w}_h^{k-l},\boldsymbol{\varphi}_i\right) + \tilde{\boldsymbol{b}}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_i)\right)_{i=1}^{N_{hp}}, \tag{6.139}$$

respectively. Here $\boldsymbol{\varphi}_i \in \mathrm{B}_{hp}$, $i = 1,\ldots,N_{hp}$, are the basis functions, $\bar{\boldsymbol{\xi}} \in \mathbb{R}^{N_{hp}}$ and $\boldsymbol{\xi}_{k-l} \in \mathbb{R}^{N_{hp}}$, $l = 1,\ldots,n$, are the algebraic representations of $\bar{\boldsymbol{w}}_h \in \boldsymbol{S}_{hp}$ and $\boldsymbol{w}_h^{k-l} \in \boldsymbol{S}_{hp}$, $l = 1,\ldots,n$, respectively. Finally, using (6.135) and (6.137) – (6.139), we have

$$\boldsymbol{F}_h(\{\boldsymbol{\xi}_{k-l}\}_{l=1}^n;\boldsymbol{\xi}_k) = \mathbb{C}_h(\boldsymbol{\xi}_k)\boldsymbol{\xi}_k - \boldsymbol{d}_h(\{\boldsymbol{\xi}_{k-l}\}_{l=1}^n,\boldsymbol{\xi}_k), \quad k=1,\ldots,r. \tag{6.140}$$

Let us note that the flux matrix $\mathbb{C}_h$ given by (6.138) has the same block structure as the matrix $\mathbb{C}_h$ defined by (6.124). The sequence of nonlinear algebraic systems can be solved by the damped Newton-like iterative process (6.127) – (6.128) treated in Section 6.4.4.

Concerning the initial guess $\boldsymbol{\xi}_k^0$ for the iterative process (6.127) – (6.128), we use either the value known from the previous time level given by (6.129), i.e, $\boldsymbol{\xi}_k^0 = \boldsymbol{\xi}_{k-1}$, $k = 1,\ldots,r$, or it is possible to apply a higher-order extrapolation from previous time levels similarly as in the high-order semi-implicit time discretization from [Dol08b]. Hence, we put

$$\boldsymbol{\xi}_k^0 = \sum_{l=1}^{n}\beta_{n,l}\boldsymbol{\xi}_{k-l}, \quad k=1,\ldots,r, \tag{6.141}$$

where $\boldsymbol{\xi}_{k-l}$, $l = 1,\ldots,n$, correspond to the solution $\boldsymbol{w}_h^{k-l}$ at the time level t_{k-l} and $\beta_{n,l}$, $l = 1,\ldots,n$, are coefficients depending on time steps τ_{k-l}, $l = 0,\ldots,n$. Table 6.4 shows the values of $\beta_{n,l}$, $l = 1,\ldots,n$, for $n = 1,2,3$. In Table 6.5, these coefficients are expressed in terms of the time steps.

Remark 6.19. *Similarly as in Remark 6.15, if we carry out only one Newton-like iteration at each time level, put* $\lambda^0 = 1$, *the matrix* $\mathbb{C}$ *is updated at each time step and use the extrapolation* (6.141)*; then the implicit method* (6.134) *reduces to the* high-order semi-implicit time discretization *approach presented in [DF04a] and [FK07], which can be formulated in the following way: We seek the finite sequence of functions* $\{\boldsymbol{w}_h^k\}_{k=0}^r$ *such that*

$$\boldsymbol{w}_h^k \in \boldsymbol{S}_{hp},\ k=0,1,\ldots,r, \tag{6.142a}$$

$$\frac{1}{\tau_k}\left(\sum_{l=0}^{n}\alpha_{n,l}\boldsymbol{w}_h^{k-l},\boldsymbol{\varphi}_h\right) + \hat{\boldsymbol{b}}_h\left(\sum_{l=1}^{n}\beta_{n,l}\boldsymbol{w}_h^{k-l},\boldsymbol{w}_h^k,\boldsymbol{\varphi}_h\right) = 0 \tag{6.142b}$$

$$\forall\,\boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp},\ k=1,\ldots,r.$$

Similarly as in (6.134), $\boldsymbol{w}_h^0,\ldots,\boldsymbol{w}_h^{n-1}$ *are defined by* (6.134c) *and* (6.134d). *Here,* $\beta_{n,l}$, $l = 1,\ldots,n$, *are coefficients introduced above and* $\hat{\boldsymbol{b}}_h$ *is the form given by* (6.131)*, i.e.,*

$$\hat{\boldsymbol{b}}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \boldsymbol{b}_h^L(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) - \tilde{\boldsymbol{b}}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_h), \quad \boldsymbol{w}_h,\boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp}.$$

Obviously, $\hat{\boldsymbol{b}}_h$ *is consistent with* $\boldsymbol{b}_h$ *because* $\boldsymbol{b}_h(\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \hat{\boldsymbol{b}}_h(\boldsymbol{w}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h)$ *for all* $\boldsymbol{w}_h,\boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp}$. *Problem* (6.142) *represents a sequence of systems of linear algebraic equations.*

	constant time step			variable time step		
	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$
$\beta_{n,1}$	1	2	3	1	$1+\theta_k$	$(1+\theta_k)\frac{\theta_k\theta_{k-1}+\theta_{k-1}+1}{\theta_{k-1}+1}$
$\beta_{n,2}$		-1	-3		$-\theta_k$	$-\theta_k(\theta_k\theta_{k-1}+\theta_{k-1}+1)$
$\beta_{n,3}$			1			$\theta_k\theta_{k-1}\frac{\theta_k\theta_{k-1}+\theta_{k-1}}{\theta_{k-1}+1}$

Table 6.4: Values of $\beta_{n,l},\ l=0,\dots,n$, for $n=1,2,3$ for constant and variable time steps, $\theta_k=\tau_k/\tau_{k-1},\ k=1,2,\dots,r$.

	$n=1$	$n=2$	$n=3$
$\beta_{n,1}$	1	$\frac{\tau_k+\tau_{k-1}}{\tau_{k-1}}$	$\frac{(\tau_k+\tau_{k-1}+\tau_{k-2})(\tau_k+\tau_{k-1})}{\tau_{k-1}(\tau_{k-1}+\tau_{k-2})}$
$\beta_{n,2}$		$-\frac{\tau_k}{\tau_{k-1}}$	$-\frac{\tau_k(\tau_k+\tau_{k-1}+\tau_{k-2})}{\tau_{k-1}\tau_{k-2}}$
$\beta_{n,3}$			$\frac{\tau_k(\tau_k+\tau_{k-1})}{\tau_{k-2}(\tau_{k-1}+\tau_{k-2})}$

Table 6.5: Values of $\beta_{n,l}$ expressed in terms of time steps.

6.4.6 Choice of the time step

The choice of the time step has a great influence on the efficiency of the BDF-DGM. We already mentioned that the implicit time discretization allows us to choose the time step many times larger than an explicit scheme. Too large time step causes the loss of accuracy and too small time step reduces the efficiency of the computation.

On the other hand, in the beginning of the computation, we usually start from a nonphysical initial condition and a large time step may cause failure of the computational process. Therefore, the aim is to develop a sufficiently robust algorithm which automatically increases the time step from small values in the beginning of the computation to larger values, but which also ensures accuracy with respect to time.

The standard ODE strategy chooses the size of the time step so that the corresponding *local discretization error* is below a given tolerance, see, e.g., [HNW00]. Very often, the local discretization error is estimated by a difference of two numerical solutions obtained by two time integration methods. However, we have to solve two nonlinear algebraic systems at each time level which leads to higher computational costs, see [DK08].

In this section we present a strategy, which is based on a very low cost estimation of the local discretization error. For simplicity, we deal only with the first-order method, but these considerations can be simply extended to higher-order schemes. Let us consider the ordinary differential equation

$$y' := \frac{\mathrm{d}y}{\mathrm{d}t} = f(y), \quad y(0) = y_0, \tag{6.143}$$

where $y:[0,T]\to\mathbb{R}$, $f:\mathbb{R}\to\mathbb{R}$ and $y_0\in\mathbb{R}$. We assume that problem (6.143) has a unique solution $y\in C^2([0,T])$. Moreover, let $0=t_0<t_1<t_2<\cdots<t_r=T$ be a partition of $[0,T]$. We denote by $y_k\approx y(t_k)$ an approximation of the solution y at $t_k,\ k=1,\dots,r$. The *backward Euler method* reads as

$$y_k = y_{k-1} + \tau_k f(y_k), \qquad k=1,2,\dots,r, \tag{6.144}$$

where $\tau_k = t_k - t_{k-1}$. By the Taylor theorem, there exists $\theta_k\in[t_{k-1},t_k]$ such that the corresponding local discretization error L_k has the form

$$L_k = \frac{1}{2}\tau_k^2 y''(\theta_k), \qquad \theta_k\in(t_{k-1},t_k), \tag{6.145}$$

where y'' denotes the second-order derivative of y.

Our idea is the following. We define the quadratic function $\tilde{y}_k:[t_{k-2},t_k]\to\mathbb{R}$ such that $\tilde{y}_k(t_{k-l})=y_{k-l},\ l=0,1,2$. The second-order derivative of $\tilde{y}_k$ is constant on (t_{k-2},t_k). We use the approximation

$$|L_k| \approx L_k^{\mathrm{app}} = \frac{1}{2}\tau_k^2|\tilde{y}_k''|. \tag{6.146}$$

Let $\omega>0$ be a given tolerance for the local discretization error. Our aim is to choose the time step as large as possible but guaranteeing the condition $L_k^{\mathrm{app}}\le\omega,\ k=1,\dots,r$. On the basis of (6.146), we shall assume that

$$\omega \approx \frac{1}{2}(\tau_k^{\mathrm{opt}})^2|\tilde{y}_k''|, \tag{6.147}$$

where $\tau_k^{\rm opt}$ denotes the optimal size of τ_k. We express $|\tilde{y}_k''|$ from (6.146), insert it in (6.147) and express $\tau_k^{\rm opt}$ as

$$\tau_k^{\rm opt} := \tau_k \left(\frac{\omega}{L_k^{\rm app}} \right)^{1/2}. \tag{6.148}$$

On the basis of the above considerations, we define the following
Adaptive time step algorithm

(1) let $\omega > 0$, $k > 1$, $y_{k-1}, y_{k-2} \in \mathbb{R}$ and $\tau_k > 0$ be given,

(2) compute y_k by (6.144),

(3) from $[t_{k-l}, y_{k-l}]$, $l = 0, 1, 2$, construct $\tilde{y}_k$,

(4) compute $\tau_k^{\rm opt}$ by (6.146) and (6.148),

(5) *if* $\tau_k^{\rm opt} \geq \tau_k$
then

(i) put $\tau_{k+1} = \min(\tau_k^{\rm opt},\ c_1\tau_k,\ \tau^{\max})$,
(ii) put $k = k + 1$
(iii) go to step 2)

else

(i) put $\tau_k = \tau_k^{\rm opt}$,
(ii) go to step 2).

The constant $c_1 > 1$ restricts the maximal ratio of two successive time steps. It is possible to use the value $c_1 = 2.5$. The value $\tau^{\max}$ restricts the maximal size of the time step for practical reasons. For example, $\tau^{\max} = 2\tau_0 10^{12}$, but any sufficiently large value yields similar results. If the *else* branch in step (5) of the algorithm is reached, then on each time level we solve more than one algebraic problem, which is expensive. However, this branch is reached very rarely in practice. It may occur only if the initial time step τ_0 or the constant c_1 are chosen too large.

This approach is extended to a system of ODEs in the following way. Let $\boldsymbol{y}_k \in \mathbb{R}^N$ be an approximation of the solution of the system of ODEs at t_k, $k = 0, 1, \ldots$. For each time level t_k, we define a vector-valued quadratic function $\tilde{\boldsymbol{y}}_k(t) : [t_{k-2}, t_k] \to \mathbb{R}^N$ such that $\tilde{\boldsymbol{y}}_k(t_{k-l}) = \boldsymbol{y}_{k-l}$, $l = 0, 1, 2$. Then the optimal time step is given by (6.148) with the approximation of the local discretization error

$$L_k^{\rm app} = \frac{1}{2}\tau_k^2 |\tilde{\boldsymbol{y}}_k''|, \tag{6.149}$$

where $\tilde{\boldsymbol{y}}_k'' \in \mathbb{R}^N$ denotes the second-order derivative of $\tilde{\boldsymbol{y}}_k(t)$ with respect to t. The adaptive time stepping algorithm remains the same, $\tilde{y}_k$ is replaced by $\tilde{\boldsymbol{y}}_k$ and (6.146) is replaced by (6.149).

Concerning the choice of the first two time steps in the case of the solution of the Euler equations, we use the relation (6.94), namely

$$\tau_k = {\rm CFL} \min_{K \in \mathcal{T}_h} \frac{|K|}{\max_{\Gamma \subset \partial K} \varrho(\mathbb{P}(\boldsymbol{w}_h^k|_\Gamma))|\Gamma|}, \quad k = 0, 1, \tag{6.150}$$

where $\varrho(\mathbb{P}(\boldsymbol{w}_h^k|_\Gamma))$ is the spectral radius of the matrix $\mathbb{P}(\boldsymbol{w}_h^k|_\Gamma, \boldsymbol{n}_\Gamma)$ given by (6.17) on $\Gamma \in \mathcal{F}_h$ and the value CFL is the initial Courant–Friedrichs–Lewy number. In order to avoid drawback resulting from a nonphysical initial condition (which is the usual case), we put CFL = 0.5. Thus τ_0 and τ_1 correspond to the time steps used for the explicit time discretization with this CFL value. This choice may be underestimated in some cases, but based on our numerical experiments, it is robust with respect to the flow regime.

Remark 6.20. *The presented technique can be simply extended to n-step BDF-DGM. For $n \geq 1$ we derive (instead of* (6.146)*) the relation $L_k^{\rm app} = \gamma_n \tau_k^{n+1} |\tilde{y}_k^{(n+1)}|$, where $\gamma_n > 0$. Then relations* (6.147) *and* (6.148) *have to be modified.*

Remark 6.21. *In order to accelerate the convergence to the steady state solutions, it is possible to apply local time stepping. However, our aim is to develop a scheme which can also be applied to nonstationary problems. Therefore, we consider only global time stepping.*

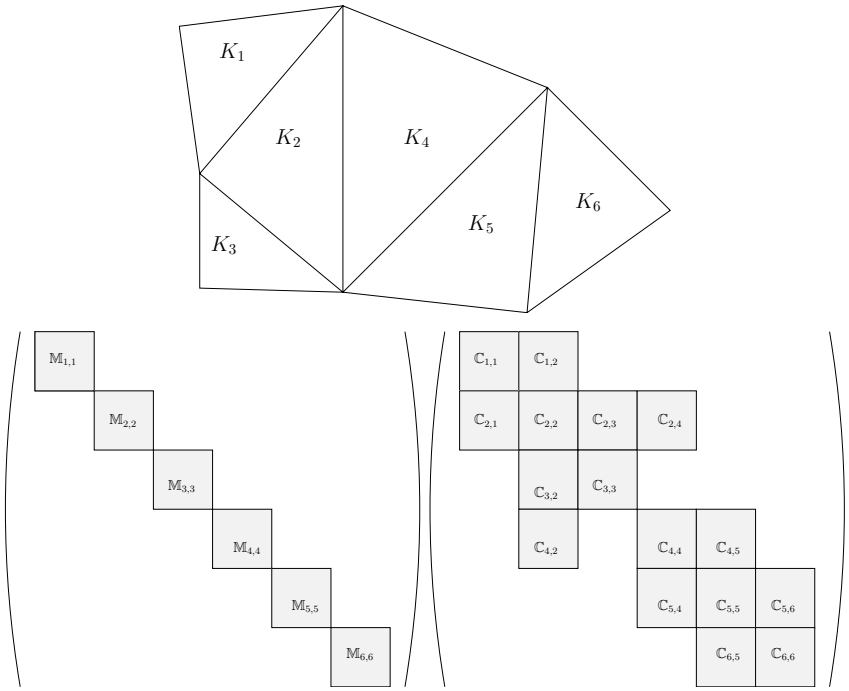

Figure 6.4: Example of a triangular mesh with elements $K_\mu,\ \mu = 1,\dots,6$ (top) and the corresponding block structure of the matrices $\mathbb{M}_h$ (left bottom) and $\mathbb{C}_h$ (right bottom).

6.4.7 Structure of the flux matrix

The flux matrix $\mathbb{C}_h$ given by (6.124) can be written in the form

$$\mathbb{C}_h(\bar{\boldsymbol{\xi}}) = \frac{1}{\tau_k}\mathbb{M}_h + \mathbb{B}_h(\bar{\boldsymbol{\xi}}), \tag{6.151}$$

where

$$\mathbb{M}_h = \big((\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i)\big)_{i,j=1}^{N_{hp}}, \qquad \mathbb{B}_h(\bar{\boldsymbol{\xi}}) = \big(\boldsymbol{b}_h^L(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i)\big)_{i,j=1}^{N_{hp}}. \tag{6.152}$$

The matrix $\mathbb{M}_h$ is called the *mass matrix*. If the basis in $\boldsymbol{S}_{hp}$ is constructed elementwise (i.e., the support of each basis function is just one simplex from $\mathcal{T}_h$), then $\mathbb{M}_h$ is block diagonal. Similarly, the matrices $\mathbb{B}_h$ and therefore $\mathbb{C}_h$ have a block structure. By virtue of (6.123), we easily find that each block-row of $\mathbb{B}_h$ corresponds to one element $K \in \mathcal{T}_h$ and contains a diagonal block and several off-diagonal blocks. Each off-diagonal block corresponds to one face $\Gamma \in \mathcal{F}_h$. See Figure 6.4, where an illustrative mesh and the corresponding block structures of matrices $\mathbb{M}_h$ and $\mathbb{C}_h$ are shown.

Similarly, the vector $\boldsymbol{d}_h$ from (6.125) can be written as

$$\boldsymbol{d}_h(\boldsymbol{\xi}_{k-1}, \bar{\boldsymbol{\xi}}) = \frac{1}{\tau_k}\boldsymbol{m}_h(\boldsymbol{\xi}_{k-1}) + \boldsymbol{u}_h(\bar{\boldsymbol{\xi}}), \tag{6.153}$$

where

$$\boldsymbol{m}_h(\boldsymbol{\xi}_{k-1}) = \big((\boldsymbol{w}_h^{k-l}, \boldsymbol{\varphi}_i)\big)_{i=1}^{N_{hp}}, \quad \boldsymbol{u}_h(\bar{\boldsymbol{\xi}}) = \Big(\tilde{\boldsymbol{b}}_h(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_i)\Big)_{i=1}^{N_{hp}}. \tag{6.154}$$

If the time step τ_k in (6.151) is small enough, then the matrix $\mathbb{M}_h/\tau_k$ dominates over $\mathbb{B}_h$. Hence, if we construct a basis of $\boldsymbol{S}_{hp}$ which is orthonormal with respect to the L^2-scalar product, then $\mathbb{M}_h$ is the identity matrix and the linear algebraic problems (6.128) is solved easily for small τ_k.

Remark 6.22. *On the other hand, there exists a limit value $\tau^\infty \gg 1$, such that for any $\tau_k \geq \tau^\infty$ we have*

$$\mathbb{C}_h(\bar{\boldsymbol{\xi}}) \doteq \mathbb{B}_h(\bar{\boldsymbol{\xi}}), \quad \bar{\boldsymbol{\xi}} \in \mathbb{R}^{N_{hp}}, \tag{6.155}$$

where the symbol $\doteq$ denotes the equality in the finite precision arithmetic. Similarly, for any $\tau_k \geq \tau^\infty$ from (6.153) – (6.154) *we obtain the relation*

$$\boldsymbol{d}_h(\boldsymbol{\xi}_{k-1}, \bar{\boldsymbol{\xi}}) \doteq \boldsymbol{u}_h(\bar{\boldsymbol{\xi}}). \tag{6.156}$$

This means that $\mathbb{C}_h$ *as well as* $\boldsymbol{d}_h$ *are independent of the size of* τ_k*. Moreover, by virtue of* (6.126)*, problem* (6.98)

$$\begin{aligned}\mathbf{0} = \boldsymbol{F}_h(\boldsymbol{\xi}_{k-1};\boldsymbol{\xi}_k) =& \mathbb{C}_h(\boldsymbol{\xi}_k)\boldsymbol{\xi}_k - \boldsymbol{d}_h(\boldsymbol{\xi}_{k-1},\boldsymbol{\xi}_k) \\ \doteq& \mathbb{B}_h(\boldsymbol{\xi}_k)\boldsymbol{\xi}_k - \boldsymbol{u}_h(\boldsymbol{\xi}_k), \quad k = 1,\dots,r,\end{aligned} \tag{6.157}$$

is independent (in the finite precision arithmetic) on the size of τ_k *provided that* $\tau_k \geq \tau^\infty$*. Our numerical experiments indicated that limit value* $\tau_\infty \approx 10^{12}$ *in the double precision arithmetic.*

6.4.8 Construction of the basis in the space $\boldsymbol{S}_{hp}$

In this section we present one possibility, how to construct a basis $\mathrm{B}_{hp} = \{\boldsymbol{\varphi}_i(x),\ i = 1,\dots,N_{hp}\}$ in the space $\boldsymbol{S}_{hp}$, in order to solve efficiently the Euler equations with the aid of the DGM. Obviously, it is advantageous to use functions from B_{hp} with small supports. Since $\boldsymbol{S}_{hp}$ consists of discontinuous functions, for each element $K \in \mathcal{T}_h$ it is possible to define a local basis

$$\mathrm{B}_K = \left\{\boldsymbol{\psi}_{K,i} \in \boldsymbol{S}_{hp};\ \mathrm{supp}(\boldsymbol{\psi}_{K,i}) \subset K,\ i = 1,\dots,\hat{N}\right\}, \tag{6.158}$$

with $\boldsymbol{\psi}_{K,i} \in (P_p(K))^m$ (= the space of vector-valued polynomials of degree $\leq p$ on $K \in \mathcal{T}_h$), where $\hat{N} = \frac{d+2}{d!}\Pi_{j=1}^d(p+j)$ is its dimension. Then the basis B_{hp} will be a composition of the local bases $\mathrm{B}_K,\ K \in \mathcal{T}_h$.

Let

$$\hat{K} = \{(\hat{x}_1,\dots,\hat{x}_d);\ \hat{x}_i \geq 0,\ i = 1,\dots,d,\ \sum_{i=1}^d \hat{x}_i \leq 1\} \tag{6.159}$$

be the *reference simplex*. We consider affine mappings

$$F_K : \hat{K} \to \mathbb{R}^d, \qquad F_K(\hat{K}) = K, \quad K \in \mathcal{F}_h. \tag{6.160}$$

(In Section 6.6 we deal with curved elements. In this case F_K is a polynomial mapping of degree > 1.)

On the reference element $\hat{K}$ we define a basis in the space of vector-valued polynomials of degree $\leq p$ by

$$\begin{aligned}&\hat{\boldsymbol{S}}_p = (\hat{S}_p)^m, \\ &\hat{S}_p = \left\{\phi_{n_1,\dots,n_d}(\hat{x}_1,\dots,\hat{x}_d) = \Pi_{i=1}^d(\hat{x}_i - \hat{x}_i^c)^{n_i};\quad n_1,\dots,n_d \geq 0,\ \sum_{j=1}^d n_j \leq p\right\},\end{aligned} \tag{6.161}$$

where $(\hat{x}_1^c,\dots,\hat{x}_d^c)$ is the barycenter of $\hat{K}$. The dimension of the space spanned over the set $\hat{\boldsymbol{S}}_p$ is $\hat{N} = \frac{d+2}{d!}\Pi_{j=1}^d(p+j)$. By the Gram–Schmidt $L^2(\hat{K})$-orthonormalization process applied to $\hat{\boldsymbol{S}}_p$ we obtain the orthonormal system $\{\hat{\phi}_j,\ j = 1,\dots,\hat{N}\}$. The Gram–Schmidt orthonormalization on the reference element can be easily computed, because $\hat{N}$ is small (moreover, the orthonormalization can be done for each component of $\boldsymbol{S}_{hp}$ independently). Hence, this orthonormalization does not cause any essential loss of accuracy.

Furthermore, let $F_K,\ K \in \mathcal{T}_h$, be the mapping introduced in (6.160). We put

$$\mathrm{B}_K = \{\boldsymbol{\psi}_{K,j};\ \boldsymbol{\psi}_{K,j}(x) = \hat{\phi}_j(F_K^{-1}(x)),\ x \in K,\ j = 1,\dots,\hat{N}\}, \tag{6.162}$$

which defines a local basis B_K for each element $K \in \mathcal{T}_h$ separately. For an affine mapping F_K the basis B_K is $L^2(K)$-orthogonal with respect to the L^2-scalar product and the blocks $\mathbb{M}_{K,K}$ of the mass matrix $\mathbb{M}$ given by (6.152) are diagonal. If F_K is not afine, then the orthogonality of B_K is violated. However, in practical applications, the curved face $K_K \cap \partial\Omega$ is close to a straight (polygonal) one (see Section 6.6), and thus the matrix block $\mathbb{M}_{K,K}$ is strongly diagonally dominant.

Finally, a composition of the local bases $\mathrm{B}_K,\ K \in \mathcal{T}_h$, defines a basis of $\boldsymbol{S}_{hp}$, i.e.,

$$\mathrm{B}_{hp} = \{\boldsymbol{\psi}_{K,j};\ \boldsymbol{\psi}_{K,j} \in \mathrm{B}_K,\ j = 1,\dots,\hat{N},\ K \in \mathcal{T}_h\}, \tag{6.163}$$

which is, for affine mappings $F_K,\ K \in \mathcal{T}_h$, the L^2-orthogonal basis of $\boldsymbol{S}_{hp}$. In case that F_K is not an affine mapping for some $K \in \mathcal{T}_h$, the L^2-orthogonality is violated, i.e., $\left(\boldsymbol{\psi}_{K,i},\boldsymbol{\psi}_{K,j}\right) \neq 0$ for $i,j = 1,\dots,\hat{N}$, $i \neq j$. However, since F_K is usually close to an affine mapping, we have $\left|\left(\boldsymbol{\psi}_{K,i},\boldsymbol{\psi}_{K,j}\right)\right| \ll \left|\left(\boldsymbol{\psi}_{K,i},\boldsymbol{\psi}_{K,i}\right)\right|$ for $i,j = 1,\dots,\hat{N}$, $i \neq j$.

Remark 6.23. *It is possible to find that every entry of* $\boldsymbol{F}_h$ *and/or* $\mathbb{C}_h$ *depends on* $\boldsymbol{w}_h$ *on at most two neighbouring elements. This is a favourable property which simplifies the parallelization of the algorithm.*

6.4.9 Steady-state solution

Very often, we are interested in the solution of the *stationary Euler equations*, i.e., we seek $\boldsymbol{w} : \Omega \to \mathcal{D}$ ($\mathcal{D}$ is given by (6.12)) such that

$$\sum_{s=1}^{d} \frac{\partial \boldsymbol{f}_s(\boldsymbol{w})}{\partial x_s} = 0, \tag{6.164}$$

where $\boldsymbol{w}$ is the steady-state vector and $\boldsymbol{f}_s,\ s = 1, \dots, d$, are the Euler fluxes defined in (6.9) and (6.10), respectively. This system is equipped with boundary conditions (6.37), discussed in detail in Section 6.3.

The stationary Euler equations can be discretized in the same way as the non-stationary ones, omitting only the approximation of the time derivative.

Definition 6.24. *We say that $\boldsymbol{w}_h \in \boldsymbol{S}_{hp}$ is a DG approximate solution of (6.164) if*

$$\boldsymbol{b}_h(\boldsymbol{w}_h, \boldsymbol{\varphi}_h) = 0 \quad \forall\, \boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp}, \tag{6.165}$$

where $\boldsymbol{b}_h$ is given by (6.93)*. We call $\boldsymbol{w}_h$ the* steady-state solution *of the Euler equations.*

With the aid of the notation introduced in Section 6.4.2, we can formulate (6.165) as the algebraic problem to find $\boldsymbol{\xi} \in \mathbb{R}^{N_{hp}}$ such that

$$\boldsymbol{F}_h^{\mathrm{SS}}(\boldsymbol{\xi}) = \boldsymbol{0}, \tag{6.166}$$

where $\boldsymbol{\xi}$ is the algebraic representation of $\boldsymbol{w}_h$ by the isomorphism (6.96) and

$$\boldsymbol{F}_h^{\mathrm{SS}}(\boldsymbol{\xi}) = \left(\boldsymbol{b}_h\Big(\sum_{j=1}^{N_{hp}} \xi_j \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i \Big) \right)_{i=1}^{N_{hp}} \in \mathbb{R}^{N_{hp}}. \tag{6.167}$$

By virtue of (6.137), (6.152) and (6.154), we have

$$\boldsymbol{F}_h^{\mathrm{SS}}(\boldsymbol{\xi}) = \mathbb{B}(\boldsymbol{\xi})\boldsymbol{\xi} - \boldsymbol{u}_h(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \in \mathbb{R}^{N_{hp}}. \tag{6.168}$$

Problem (6.166) represents a system of nonlinear algebraic equations. It can be solved directly by the (damped) Newton method, see [HH02]. Another very often used possibility is to apply the *time-marching* (or *time stabilization*) method based on the solution of the nonstationary Euler equations (6.8) and to seek the *steady-state* solution as a limit of the nonstationary solution for $t \to \infty$. This means that the methods for solving unsteady flow are applied as iterative processes, assuming that $\boldsymbol{w}_h = \lim_{k\to\infty} \boldsymbol{w}_h^k$. The nonstationary computational process is stopped, when a suitable *steady-state criterion* is achieved.

The usual steady-state criterion often used for explicit time discretization reads (for an othonormal basis) as

$$\left\| \frac{\partial \boldsymbol{w}_h}{\partial t} \right\|_{L^2(\Omega)} \approx \eta_k = \frac{1}{\tau_k} \|\boldsymbol{w}_h^k - \boldsymbol{w}_h^{k-1}\|_{L^2(\Omega)} = \frac{1}{\tau_k} |\boldsymbol{\xi}_k - \boldsymbol{\xi}_{k-1}| \le \mathrm{TOL}, \tag{6.169}$$

where $\boldsymbol{w}_h^{k-l},\ l = 0, 1$, denote the values of the approximate solution at time levels $t_{k-l},\ l = 0, 1$, $\boldsymbol{\xi}_{k-l},\ l = 0, 1$, are their algebraic representations given by the isomorphism (6.96) and TOL is a given tolerance.

Criterion (6.169) is not suitable for the implicit time discretization, when very large time steps are used, see [DHH11, Section 4.3.1.]. Then it is suitable to use the *steady-state residual criterion*

$$|\boldsymbol{F}_h^{\mathrm{SS}}(\boldsymbol{\xi}_k)| = |\mathbb{B}(\boldsymbol{\xi}_k)\boldsymbol{\xi}_k - \boldsymbol{u}_h(\boldsymbol{\xi}_k)| \le \mathrm{TOL}, \tag{6.170}$$

which is independent of τ_k and measures the residuum of the nonlinear algebraic system (6.167).

However, it is an open question as to how to choose the tolerance TOL in (6.170), since the residuum depends on the size of the computational domain Ω, on the magnitude of components of $\boldsymbol{w}_h^k$, etc. Therefore, from the practical reasons, we use the *relative residuum steady-state criterion*

$$\mathrm{SSres}(k) := \frac{|\boldsymbol{F}_h^{\mathrm{SS}}(\boldsymbol{\xi}_k)|}{|\boldsymbol{F}_h^{\mathrm{SS}}(\boldsymbol{\xi}_0)|} \le \mathrm{TOL}, \tag{6.171}$$

which already does not suffer from the mentioned drawbacks. Here $\boldsymbol{\xi}_0$ is the algebraic representation of the initial state $\boldsymbol{w}_h^0$.

Another possibility are the stopping criteria which follow from the physical nature of the considered problem. E.g., in aerodynamics, when we solve flow around a 2D profile, we are often interested in the *aerodynamic coefficients* of the considered

flow, namely coefficients of *drag* (c_D), *lift* (c_L) and *momentum* (c_M). In the 2D case, the coefficients c_D and c_L are defined as the first and second components of the vector

$$\frac{1}{\frac{1}{2}\rho_\infty|\boldsymbol{v}_\infty|^2 L_{\mathrm{ref}}}\int_{\Gamma_{\mathrm{prof}}} \mathrm{p}\boldsymbol{n}\,\mathrm{d}S, \tag{6.172}$$

where ρ_∞ and $\boldsymbol{v}_\infty$ are the far-field density and velocity, respectively, L_{ref} is the reference length, Γ_{prof} is the profile, $\boldsymbol{n}$ is outer unit normal to the profile pointing into the profile and p is the pressure. Moreover, c_M is given by

$$\frac{1}{\frac{1}{2}\rho_\infty|\boldsymbol{v}_\infty|^2 L_{\mathrm{ref}}^2}\int_{\Gamma_{\mathrm{prof}}} (x - x_{\mathrm{ref}}) \times \mathrm{p}\boldsymbol{n}\,\mathrm{d}S, \tag{6.173}$$

where x_{ref} is the moment reference point. We adopt the notation $x \times y = x_1y_2 - x_2y_1$ for $x = (x_1, x_2), y = (y_1, y_2) \in \mathbb{R}^2$.

Then it is natural to stop the computation when these coefficients achieve a given tolerance tol, e.g.,

$$\Delta c_\alpha(k) \le \mathrm{tol}, \qquad \Delta c_\alpha(k) = \max_{l=\overline{k},\dots,k} c_\alpha(l) - \min_{l=\overline{k},\dots,k} c_\alpha(l), \tag{6.174}$$

where $\alpha = D$, L and M (for the drag, lift and momentum), $c_\alpha(k)$ is the value of the corresponding aerodynamic coefficient at the k^{th}-time level and $\overline{k}$ is the entire part of the number $0.9k$. This means that the minimum and maximum in (6.174) are taken over the last 10% of the number of time levels.

In contrast to the tolerance TOL in (6.171), which has to be chosen empirically, the tolerance tol in (6.174) can be chosen only on the basis of our accuracy requirements (without any previous numerical experiments). Since the absolute values of aerodynamic coefficient are (usually) less than one, the stopping criterion (6.174) with tolerance, e.g., $\mathrm{tol} = 10^{-4}$, gives accuracy of the aerodynamic coefficients for 3 decimal digits.

Finally, let us note that since we seek only the steady-state solution, we do not need to take care of an accurate approximation of the evolution process. Therefore, we can choose the time step τ_k relatively large. Hence, the tolerance ω appearing in (6.148) can also be large.

6.5 Shock capturing

In higher-order numerical methods applied to the solution of high speed flows with shock waves and contact discontinuities the Gibbs phenomenon appears manifested by spurious (nonphysical) oscillations in computed quantities propagating from discontinuities. In the standard Galerkin finite element methods, these oscillations propagate far into the computational domain. However, in DG numerical solutions the Gibbs phenomenon is manifested only by spurious overshoots and undershoots appearing in the vicinity of discontinuities. These phenomena do not occur in low Mach number regimes, when the exact solution is regular, but in the high-speed flow they cause instabilities in the numerical solution and collapse of the computational process.

In order to cure this undesirable feature, in the framework of higher-order finite volume methods one uses suitable limiting procedures. They should preserve the higher-order accuracy of the method in regions where the solution is regular, and decrease the order to 1 in a neighbourhood of discontinuities or steep gradients. These methods are based on the use of the flux limiter. See e.g., [FFS03] and citations therein. In [CS89] and [CHS90], the finite volume limiting procedures were generalized also to DGM.

Here we present another technique, based on the concept of artificial viscosity applied locally on the basis of a suitable *jump (discontinuity) indicator*.

6.5.1 Jump indicators

Approximate solutions obtained by the DGM are, in general, discontinuous on interfaces between neighbouring elements. If the exact solution is sufficiently regular, then the jumps in the approximate solution are small and, as follows from the theory as well as numerical experiments, tend to zero if $h \to 0$.

The DG solution of inviscid flow can contain large inter-element jumps in subdomains, where the solution is not sufficiently smooth, i.e., in areas with discontinuities (shock waves or contact discontinuities). Numerical experiments show that the inter-element jumps in the approximate solution are $[\boldsymbol{w}_h]_\Gamma = O(1)$ on discontinuities, but $[\boldsymbol{w}_h]_\Gamma = O(h^{p+1})$ in the areas where the solution is regular. This inspires us to define a *jump indicator* , which evaluates the inter-element jumps of the approximate solution. On general unstructured grids, it appears to be suitable to measure the magnitude of inter-element jumps in the integral form by

$$\int_{\partial K\cap\Omega} [w_{h,1}]^2\,\mathrm{d}S, \quad K \in \mathcal{T}_h \tag{6.175}$$

on interior faces $\Gamma \in \mathcal{F}_h^I$, where $w_{h,1}$ denotes the first component, i.e., the density ρ_h corresponding to the state $\boldsymbol{w}_h$. (Here we take into account that the density is discontinuous both on shock waves and contact discontinuities.)

This leads us to the definition of the *jump indicator* in the form

$$g_K(\boldsymbol{w}_h) = \frac{\int_{\partial K\cap\Omega}[w_{h,1}]^2\,\mathrm{d}S}{|K|\,\sum_{\Gamma\subset\partial K\cap\Omega}\mathrm{diam}(\Gamma)}, \quad K\in\mathcal{T}_h, \tag{6.176}$$

where $|K|$ denotes the d-dimensional measure of K and $\mathrm{diam}(\Gamma)$ is the diameter of Γ. We see that we have

$$g_K(\boldsymbol{w}_h) = \begin{cases} O(h^{2p}) & \text{for } K\in\mathcal{T}_h, \text{ where the solution is smooth,} \\ O(h^{-2}) & \text{for } K\in\mathcal{T}_h \text{ near discontinuities.} \end{cases} \tag{6.177}$$

Thus, $g_K\to 0$ for $h\to 0$ in the case when $K\in\mathcal{T}_h$ is in a subdomain where the solution is regular, and $g_K\to\infty$ for $h\to 0$ in the case when $K\in\mathcal{T}_h$ is in the vicinity of a discontinuity.

There are various modifications of this indicator, as for example,

$$g_K(\boldsymbol{w}_h) = \int_{\partial K}[w^k_{h,1}]^2\,\mathrm{d}S\big/(h_K|K|^{3/4}), \quad K\in\mathcal{T}_h, \tag{6.178}$$

in the 2D case, proposed in [DFS03] and applied in [FK07]. The indicator g_K was constructed in such a way that it takes an anisotropy of the computational mesh into account. It was shown in [DFS03] that the indicator $g_K(\boldsymbol{w}_h)$ identifies discontinuities safely on unstructured and anisotropic meshes.

Now we introduce the *discrete jump (discontinuity) indicator*

$$G_K(\boldsymbol{w}_h) = 0, \quad \text{if } g_K(\boldsymbol{w}_h)<1, \quad G_K(\boldsymbol{w}_h)=1, \quad \text{if } g_K(\boldsymbol{w}_h)\ge 1, \quad K\in\mathcal{T}_h. \tag{6.179}$$

Numerical experiments show that under the assumption that the mesh space size $h<1$, it is possible to indicate the areas without discontinuities checking the condition $G_K(\boldsymbol{w}_h)<1$. On the other hand, if $G_K(\boldsymbol{w}_h)>1$, the element K is lying in a neighbourhood of a discontinuity.

However, it appears that the above discrete discontinuity indicators and the artificial viscosity forms (6.181) and (6.182) introduced in the following section are too strict. Particularly, it may happen in some situations that the value of g_K in (6.176) is close to 1 and then during the computational process the value G_K from (6.179) oscillates between 1 and 0. This can be disabled to achieve a steady-state solution. Therefore, it is suitable to introduce some "smoothing" of the discrete indicator (6.179). Namely we set

$$G_K(\boldsymbol{w}_h) = \begin{cases} 0, & \text{if } g_K(\boldsymbol{w}_h)<\xi_{\min}, \\ \frac{1}{2}\sin\left(\pi\frac{g_K(\boldsymbol{w}_h)-(\xi_{\max}-\xi_{\min})}{2(\xi_{\max}-\xi_{\min})}\right)+\frac{1}{2}, & \text{if } g_K(\boldsymbol{w}_h)\in[\xi_{\min};\xi_{\max}), \\ 1, & \text{if } g_K(\boldsymbol{w}_h)\ge\xi_{\max}, \end{cases} \tag{6.180}$$

where $0\le\xi_{\min}<\xi_{\max}$. In practical applications, it is suitable to set $\xi_{\min}=0.5$ and $\xi_{\max}=1.5$.

6.5.2 Artificial viscosity shock capturing

On the basis of the discrete discontinuity indicator we introduce local artificial viscosity forms, which are included in the numerical schemes for solving inviscid compressible flow. For example, we define the artificial viscosity form $\boldsymbol{\beta}_h : \boldsymbol{S}_{hp}\times\boldsymbol{S}_{hp}\times\boldsymbol{S}_{hp}\to\mathbb{R}$ by

$$\boldsymbol{\beta}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \nu_1\sum_{K\in\mathcal{T}_h} h_K G_K(\bar{\boldsymbol{w}}_h)\int_K \nabla\boldsymbol{w}_h\cdot\nabla\boldsymbol{\varphi}_h\,\mathrm{d}\boldsymbol{x} \tag{6.181}$$

with $\nu_1=O(1)$. Since this artificial viscosity form is rather local, we propose to augment it by the form $\boldsymbol{\gamma}_h : \boldsymbol{S}_{hp}\times\boldsymbol{S}_{hp}\times\boldsymbol{S}_{hp}\to\mathbb{R}$ defined as

$$\boldsymbol{\gamma}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \nu_2\sum_{\Gamma\in\mathcal{F}_h^I}\frac{1}{2}\big(G_{K_\Gamma^{(L)}}(\bar{\boldsymbol{w}}_h)+G_{K_\Gamma^{(R)}}(\bar{\boldsymbol{w}}_h)\big)\int_\Gamma[\boldsymbol{w}_h]\cdot[\boldsymbol{\varphi}_h]\,\mathrm{d}S, \tag{6.182}$$

where $\nu_2=O(1)$ and $K_\Gamma^{(L)}, K_\Gamma^{(R)}\in\mathcal{T}_h$ are the elements sharing the inner face $\Gamma\in\mathcal{F}_h^I$. This form allows strengthening the influence of neighbouring elements and improves the behaviour of the method in the case, when strongly unstructured and/or anisotropic meshes are used. These artificial viscosity forms were introduced in [FK07], where the indicator (6.179) was used.

Because of the reasons mentioned already above, using the discontinuity indicator (6.180), we also introduce more sophisticated artificial viscosity forms $\boldsymbol{\beta}_h,\boldsymbol{\gamma}_h : \boldsymbol{S}_{hp}\times\boldsymbol{S}_{hp}\times\boldsymbol{S}_{hp}\to\mathbb{R}$, defined as

$$\boldsymbol{\beta}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \nu_1\sum_{K\in\mathcal{T}_h} G_K(\bar{\boldsymbol{w}}_h)\,h_K^{\alpha_1}\int_K \nabla\boldsymbol{w}_h\cdot\nabla\boldsymbol{\varphi}_h\,\mathrm{d}x, \tag{6.183}$$

and

$$\boldsymbol{\gamma}_h(\bar{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) = \nu_2 \sum_{\Gamma \in \mathcal{F}_h^I} \frac{1}{2}\big(G_{K_\Gamma^{(L)}}(\bar{\boldsymbol{w}}_h) + G_{K_\Gamma^{(R)}}(\bar{\boldsymbol{w}}_h)\big)\, h_\Gamma^{\alpha_2} \int_\Gamma [\boldsymbol{w}_h] \cdot [\boldsymbol{\varphi}_h]\, \mathrm{d}S, \tag{6.184}$$

with the parameters α_1, α_2, ν_1, $\nu_2 = O(1)$.

The described approach was partly motivated by the theoretical paper [JJS95]. However, the artificial viscosity was applied there in the whole domain, which can lead to a nonphysical entropy production. In our case, it is important that the discrete indicators G_K vanish in regions where the solution is regular and the artificial viscosity acts only locally in the vicinity of discontinuities. Therefore, the scheme does not produce any nonphysical entropy in regions where the exact solution is regular.

The artificial viscosity forms $\boldsymbol{\beta}_h$ and $\boldsymbol{\gamma}_h$ are added to the left-hand side of the numerical schemes presented in previous sections. For example, the backward Euler - discontinuous Galerkin method with shock capturing now reads as

$$\begin{aligned}\frac{1}{\tau_k}\left(\boldsymbol{w}_h^k - \boldsymbol{w}_h^{k-1}, \boldsymbol{\varphi}_h\right) + \boldsymbol{b}_h(\boldsymbol{w}_h^k, \boldsymbol{\varphi}_h) + \boldsymbol{\beta}_h(\boldsymbol{w}_h^k, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h) + \boldsymbol{\gamma}_h(\boldsymbol{w}_h^k, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h) = 0 \\ \forall\, \boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp},\ k = 1, \dots, r.\end{aligned} \tag{6.185}$$

Equalities (6.185) represent a system on nonlinear algebraic equations. In the case when the artificial viscosity forms $\boldsymbol{\beta}_h$ and $\boldsymbol{\gamma}_h$ are defined with the aid of the jump indicator (6.180), the discrete problem can be solved by the Newton-like method, presented in Section 6.4.3. Namely, in (6.97), we replace $\boldsymbol{b}_h(\boldsymbol{w}_h^k, \boldsymbol{\varphi}_i)$ by

$$\boldsymbol{b}_h(\boldsymbol{w}_h^k, \boldsymbol{\varphi}_i) + \boldsymbol{\beta}_h(\boldsymbol{w}_h^k, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_i) + \boldsymbol{\gamma}_h(\boldsymbol{w}_h^k, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_i),$$

and, in (6.124), we replace $\boldsymbol{b}_h^L(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i)$ by

$$\boldsymbol{b}_h^L(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i) + \boldsymbol{\beta}_h(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i) + \boldsymbol{\gamma}_h(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i).$$

Also in other schemes we proceed in a similar way. The discrete problem with higher-order time discretization and shock capturing reads as

$$\boldsymbol{w}_h^k \in \boldsymbol{S}_{hp},\ k = 0, 1, \dots, r, \tag{6.186a}$$

$$\begin{aligned}\frac{1}{\tau_k}\left(\sum_{l=0}^{n} \alpha_{n,l} \boldsymbol{w}_h^{k-l}, \boldsymbol{\varphi}_h\right) + \boldsymbol{b}_h\left(\boldsymbol{w}_h^k, \boldsymbol{\varphi}_h\right) + \boldsymbol{\beta}_h(\boldsymbol{w}_h^k, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h) + \boldsymbol{\gamma}_h(\boldsymbol{w}_h^k, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h) = 0 \\ \forall\, \boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp},\ k = n, \dots, r,\end{aligned} \tag{6.186b}$$

where $\boldsymbol{w}_h^0, \dots, \boldsymbol{w}_h^{n-1}$ are defined by (6.134c) and (6.134d).

Similarly we formulate the higher-order semi-implicit scheme with shock capturing:

$$\boldsymbol{w}_h^k \in \boldsymbol{S}_{hp},\ k = 0, 1, \dots, r, \tag{6.187a}$$

$$\begin{aligned}&\frac{1}{\tau_k}\left(\sum_{l=0}^{n} \alpha_{n,l} \boldsymbol{w}_h^{k-l}, \boldsymbol{\varphi}_h\right) + \hat{\boldsymbol{b}}_h\left(\sum_{l=1}^{n} \beta_{n,l} \boldsymbol{w}_h^{k-l}, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h\right) + \boldsymbol{\beta}_h\Big(\sum_{l=1}^{n} \beta_{n,l} \boldsymbol{w}_h^{k-l}, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h\Big) \\ &\quad + \boldsymbol{\gamma}_h\Big(\sum_{l=1}^{n} \beta_{n,l} \boldsymbol{w}_h^{k-l}, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h\Big) = 0 \qquad \forall\, \boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp},\ k = n, \dots, r,\end{aligned} \tag{6.187b}$$

where $\boldsymbol{w}_h^0, \dots, \boldsymbol{w}_h^{n-1}$ are defined by (6.134c) and (6.134d). Problem (6.187) represents again a sequence of systems of linear algebraic equations. In this case the artificial viscosity can be defined by any jump indicator introduced in Section 6.5.1.

6.5.3 Numerical examples

In this section we present the solution of some test problems showing the performance of the shock capturing technique introduced above.

We consider transonic inviscid flow past the profile NACA 0012 given by the parametrization

$$\left[x, \pm\frac{0.12}{0.6}(0.2969\sqrt{x} - 0.126x - 0.3516x^2 + 0.2843x^3 - 0.1015x^4)\right],\ x \in [0, 1],$$

see Figure 6.5. We consider the far-field Mach number $M_\infty = 0.8$ (see (6.7)) and the angle of attack $\alpha = 1.25°$. (Let us note that $\tan\alpha = v_2/v_1$, where (v_1, v_2) is the far-field velocity vector.) This flow regime leads to two shock waves (discontinuities in the solution). The shock wave on the upper side of the profile is stronger than the shock wave on the lower side.

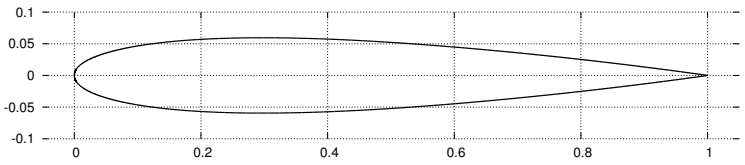

Figure 6.5: Geometry of the NACA 0012 profile.

We seek the steady-state solution of the Euler equations (6.8) with the aid of the time stabilization technique described in Section 6.4.9, using the backward Euler – discontinuous Galerkin method (BE-DGM) (6.95). The nonlinear algebraic systems are solved by the Newton-like iterative process (6.127) – (6.128).

We employ two unstructured triangular grids with piecewise polynomial approximation of the boundary described in Section 6.6. The first grid is formed by 2120 triangles and is not adapted. The second one with 2420 elements was adaptively refined along the shock waves by ANGENER code [Dol00] developed in papers [Dol98], [DF04b] and [Dol01]. See Figure 6.6. The problem was solved by the DGM using the P_p polynomial approximations with $p = 1, 2, 3$.

Figure 6.7 shows the Mach number isolines and the distribution of the Mach number along the profile in dependence on the horizontal component obtained with the aid of the P_1 and P_2 approximation on the non-adapted mesh without the shock capturing technique. We observe overshoots and undershoots in the approximate solution near the shock waves. Let us note that the P_3 computation failed.

Figure 6.8 shows the results obtained with the aid of the P_1, P_2 and P_3 approximations on the non-adapted mesh with the shock capturing technique. We can see that the nonphysical overshoots and undershoots are mostly suppressed. Finally, Figure 6.9 shows the results for P_1, P_2 and P_3 approximations on the adapted mesh with the shock capturing technique. We see that a very good resolution of the shock waves was obtained.

Further numerical experiment can be found in Section 6.7.4, where an example of the supersonic flow past the NACA 0012 profile is presented.

6.6 Approximation of a nonpolygonal boundary

In practical applications, the computational domain Ω is usually nonpolygonal, and thus its boundary has to be approximated in some way. In [BR00], Bassi and Rebay showed that a piecewise linear approximation of $\partial\Omega$ can lead to a nonphysical production of entropy and expansion waves at boundary corner points, leading to incorrect numerical solutions. In order to obtain an accurate and physically admissible solution, it is necessary to use a higher-order approximation of the boundary. We proceed in such a way that a reference triangle is transformed by a polynomial mapping onto the approximation of a curved triangle adjacent to the boundary $\partial\Omega$.

6.6.1 Curved elements

Here we describe only the two dimensional ($d = 2$) situation, the case $d = 3$ has to be generalized in a suitable way. Let K be a triangle with vertices $P_K^l,\ l = 1, 2, 3$, numbered in a such way that P_K^1 and P_K^2 lie on a curved part of $\partial\Omega$ and P_K^3 lies in the interior of Ω. By Γ we denote the edge $P_K^1 P_K^2$. Moreover, we assume that P_K^1 and P_K^2 are oriented in such a way that Ω is on the left-hand side of the oriented edge from P_K^1 to P_K^2, see Figure 6.10. We consider elements having at most one curved edge. The generalization to the case with elements having more curved edges is straightforward.

Let $q \geq 2$ be an integer denoting the *polynomial degree of the boundary approximation*. We define $q - 1$ nodes $P_K^{C,j},\ j = 1, \ldots, q-1$, lying on $\partial\Omega$ between P_K^1 and P_K^2 in such a way that nodes $P_K^{C,j},\ j = 1, \ldots, q-1$, divide the curved segment of $\partial\Omega$ between P_K^1 and P_K^2 into q parts having (approximately) the same length. We assume that $P_K^{C,j},\ j = 1, \ldots, q-1$, are ordered with an increasing index on the path along $\partial\Omega$ from P_K^1 to P_K^2. See Figure 6.10 showing a possible situation for $q = 2$ and $q = 3$.

Let

$$\hat{K} = \{(\hat{x}_1, \hat{x}_2);\ \hat{x}_i \geq 0,\ i = 1, 2,\ \hat{x}_1 + \hat{x}_2 \leq 1\} \tag{6.188}$$

be the *reference triangle*. In $\hat{K}$, we define the Lagrangian nodes of degree q by

$$\hat{P}^{\frac{i}{q};\frac{j}{q}} = [i/q; j/q], \quad 0 \leq i \leq q,\ 0 \leq j \leq q,\ 0 \leq i + j \leq q, \tag{6.189}$$

i.e., the vertices of $\hat{K}$ are the points $\hat{P}^{0;0}$, $\hat{P}^{0;1}$ and $\hat{P}^{1;0}$.

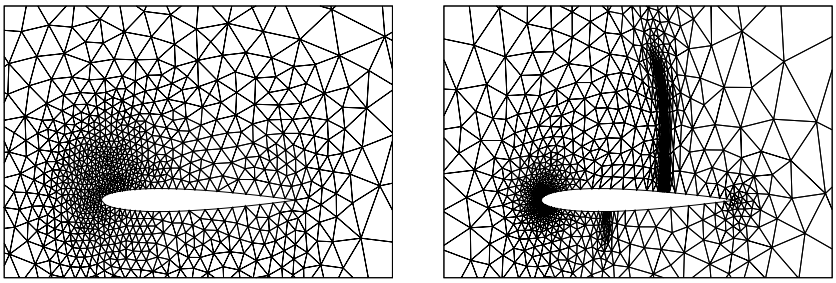

Figure 6.6: Transonic inviscid flow around the NACA 0012 profile ($M_\infty = 0.8$, $\alpha = 1.25°$): the non-adapted (left) and the adapted (right) computational meshes.

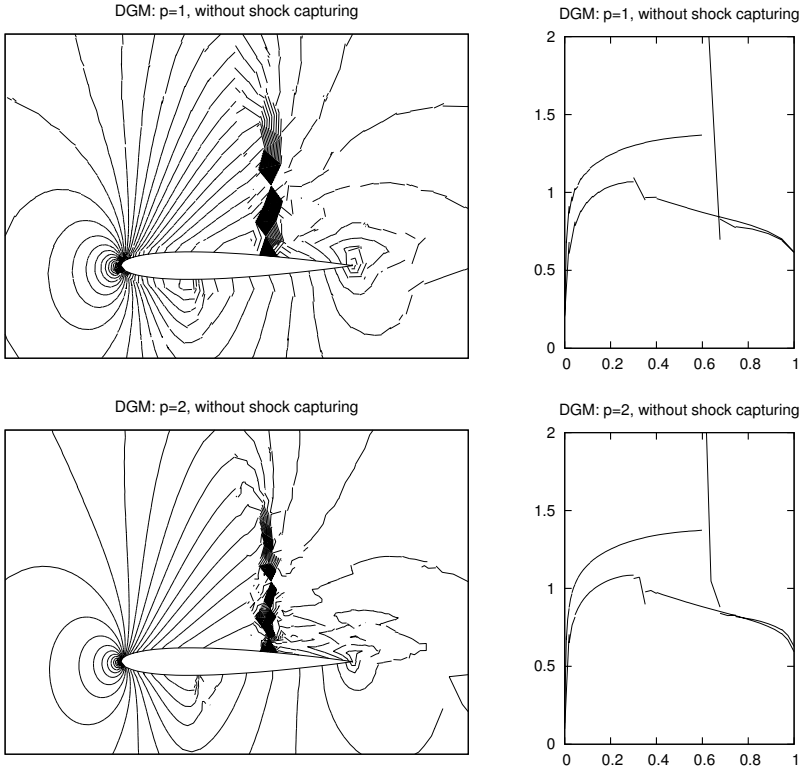

Figure 6.7: Transonic inviscid flow around the NACA 0012 profile ($M_\infty = 0.8$, $\alpha = 1.25°$): DGM with P_1 approximation (top) and P_2 approximation (bottom), Mach number isolines (left) and the distribution of the Mach number along the profile (right) on a non-adapted mesh without the shock capturing technique.

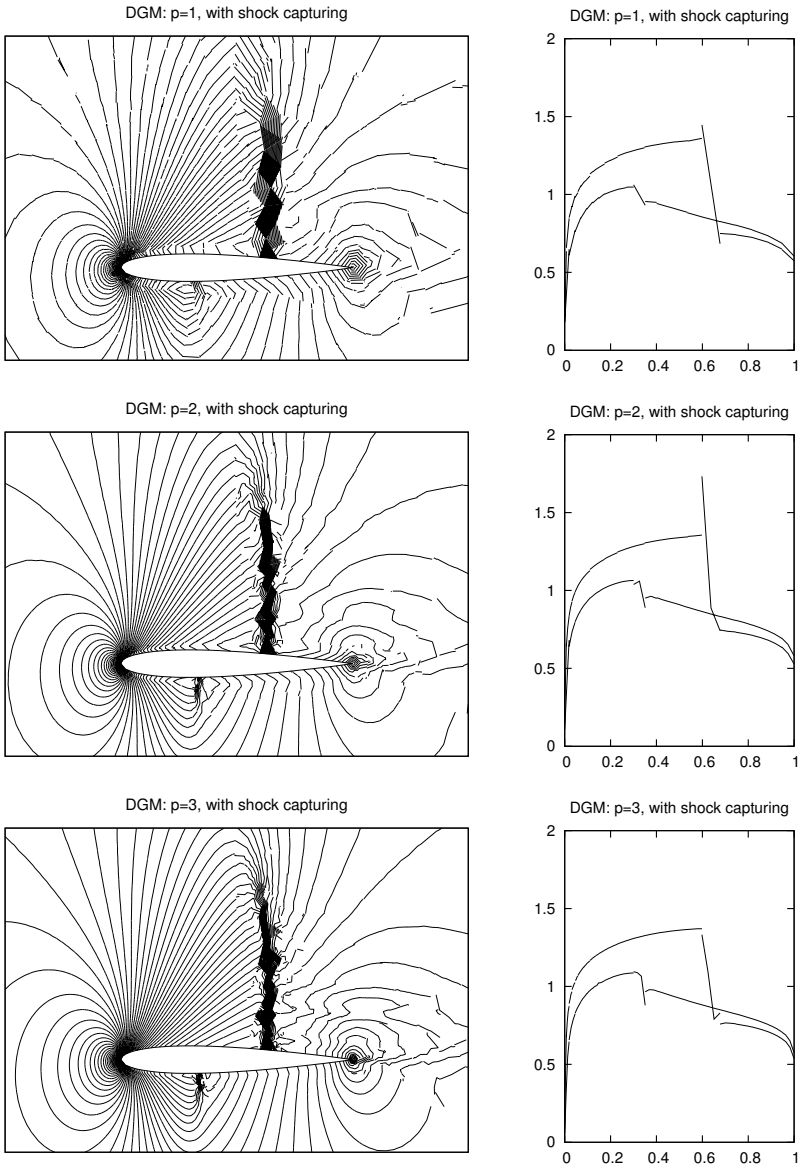

Figure 6.8: Transonic inviscid flow around the NACA 0012 profile ($M_\infty = 0.8$, $\alpha = 1.25°$): DGM with P_1 approximation (top), P_2 approximation (center) and P_3 approximation (bottom), Mach number isolines (left) and the distribution of the Mach number along the profile (right) on a non-adapted mesh with the shock capturing technique.

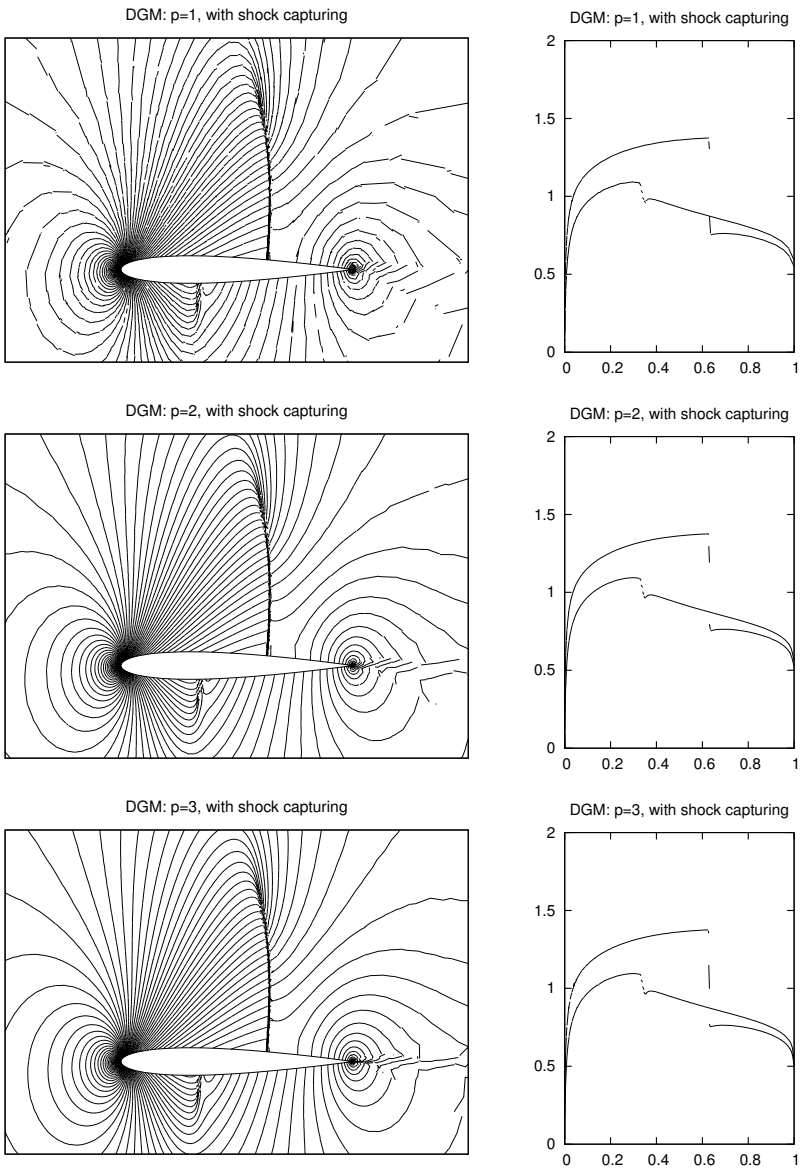

Figure 6.9: Transonic inviscid flow around the NACA 0012 profile ($M_\infty = 0.8$, $\alpha = 1.25°$): DGM with P_1 approximation (top), P_2 approximation (center) and P_3 approximation (bottom) and with boundary approximation, Mach number isolines (left) and the distribution of the Mach number along the profile (right) on an adapted mesh with the shock capturing technique.

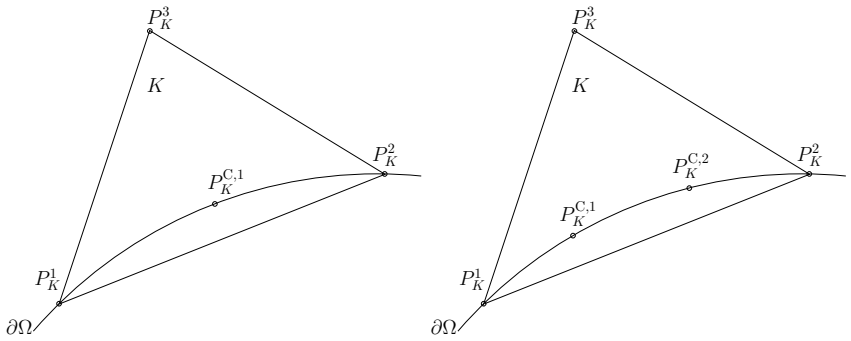

Figure 6.10: Triangle K with vertices P_k^1 and P_K^2 lying on a nonpolygonal part of $\partial\Omega$; adding one (left) and two (right) nodes on $\partial\Omega$.

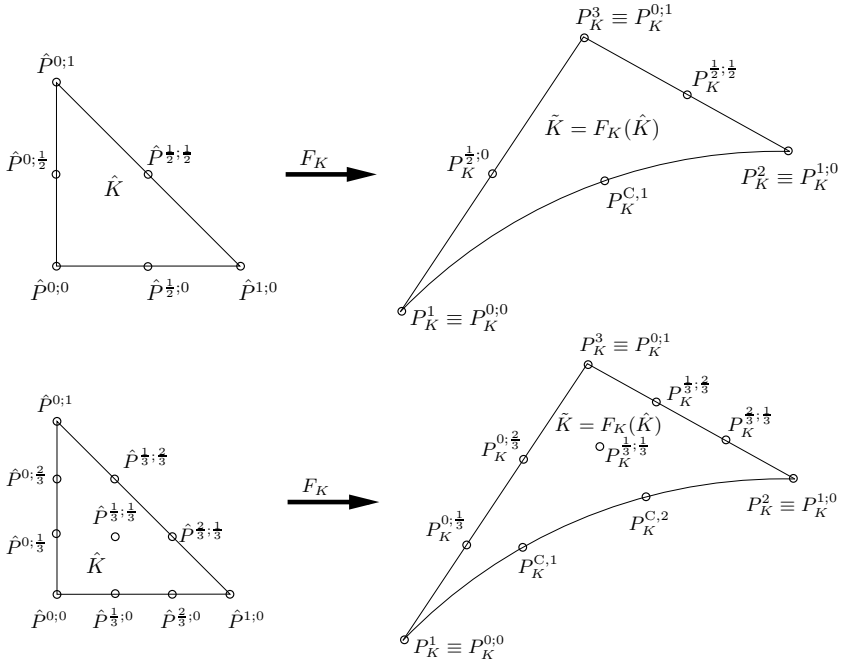

Figure 6.11: Mapping $F_K : \hat{K} \to \tilde{K}$: quadratic (top) and cubic (bottom).

Let K be the triangle with vertices $P_K^l,\ l = 1, 2, 3$, and let $P_K^{C,j} \in \partial\Omega,\ j = 1, \ldots, q-1$, be the points lying on $\partial\Omega$ between P_K^1 and P_K^2 as described above. We define the Lagrangian nodes of degree q of K by

$$P_K{}^{\frac{i}{q};\frac{j}{q}} = \frac{i}{q}P_K^1 + \frac{j}{q}P_K^2 + \frac{1-i-j}{q}P_K^3, \quad 0 \le i \le q,\ 0 \le j \le q,\ 0 \le i+j \le q. \tag{6.190}$$

Obviously, $P_K^{0;0} = P_K^1$, $P_K^{1;0} = P_K^2$ and $P_K^{0;1} = P_K^3$.

Then, there exists a unique polynomial mapping $F_K : \hat{K} \to \mathbb{R}^2$ of degree $\le q$ such that

$$\begin{aligned}
&F_K(\hat{P}^{0;0}) = P_K^1, \quad F_K(\hat{P}^{1;0}) = P_K^2, \quad F_K(\hat{P}^{0;1}) = P_K^3 \quad \text{are vertices,}\\
&F_K(\hat{P}^{\frac{i}{q};0}) = P_K^{C,i}, \quad i = 1, \ldots, q-1, \quad \text{are nodes on the curved edge,}\\
&F_K(\hat{P}^{\frac{i}{q};\frac{j}{q}}) = P_K^{\frac{i}{q};\frac{j}{q}}, \quad 0 \le i \le q,\ 1 \le j \le q-1,\ 0 \le i+j \le q, \text{ are other nodes.}
\end{aligned} \tag{6.191}$$

The existence and uniqueness of the mapping F_K follows from the fact that a polynomial mapping of degree q from $\mathbb{R}^2$ to $\mathbb{R}^2$ has $(q+1)(q+2)$ degrees of freedom equal to the number of conditions in (6.191). Then we obtain a linear algebraic system, which is regular, since the Lagrangian nodes on $\hat{K}$ are mutually different and at most q nodes belong to any straight line.

Then the triangle K will be replaced by the *curved triangle*

$$\tilde{K} = F_K(\hat{K}). \tag{6.192}$$

The set $\tilde{K}$ is a plane figure having two straight sides and one curved side $\tilde{\Gamma}$, which is an image of the *reference edge* $\hat{P}^{0;0}\,\hat{P}^{1;0}$, see Figure 6.11.

Using the described procedure, we get a partition $\tilde{\mathcal{T}}_h$ associated with the triangulation $\mathcal{T}_h$. The partition $\tilde{\mathcal{T}}_h$, called the *curved triangulation*, consists of triangles $K \in \mathcal{T}_h$ and curved elements $\tilde{K}$, associated with triangles $K \in \mathcal{T}_h$ with one edge approximating a curved part of $\partial\Omega$.

Remark 6.25. *Let us note that the considerations presented in this section make sense also for $q = 1$. In this case, any node $P_K^{C,i}$ that is not inserted on $\partial\Omega$, mapping F_K given by* (6.191) *is linear and $\tilde{K} = F_K(\hat{K}) = K$ is the triangle with straight edges.*

Remark 6.26. *The concept of the curved element can be extended also to 3D by defining a polynomial mapping F_K from a reference tetrahedron $\hat{K}_{3\mathrm{D}}$ into $\mathbb{R}^3$ for each tetrahedron K with one face approximating a curved part of $\partial\Omega$. Then K is replaced by $F_K(\hat{K}_{3\mathrm{D}})$.*

6.6.2 DGM over curved elements

Let $\tilde{\mathcal{T}}_h$ be a curved triangulation consisting of (non-curved) simplexes K as well as possible curved elements $\tilde{K}$. By virtue of Remark 6.25, a non-curved element can be considered as a special curved simplex obtained by a linear ($q = 1$) mapping F_K.

Therefore, we shall not distinguish between curved and non-curved elements in the following and we shall use the symbol K also for curved elements. Moreover, instead of $\tilde{\mathcal{T}}_h$, we shall write $\mathcal{T}_h$.

Since $\mathcal{T}_h$ may contain curved elements, we have to modify the definition (6.44) of the space $\boldsymbol{S}_{hp}$. For an integer $p \geq 0$, over the triangulation $\mathcal{T}_h$ we define the finite-dimensional function space

$$\boldsymbol{S}_{hp} = (S_{hp})^m, \quad S_{hp} = \{v;\ v \in L^2(\Omega),\ v|_K \circ F_K \in P_p(\hat{K})\ \forall\, K \in \mathcal{T}_h\}, \tag{6.193}$$

where $P_p(\hat{K})$ denotes the space of all polynomials of degree $\leq p$ on the reference element $\hat{K}$ and the symbol $\circ$ denotes the composition of mappings. Hence, instead of (6.44) and (6.45), we employ definition (6.193).

Remark 6.27. *The definition* (6.193) *of the space* $\boldsymbol{S}_{hp}$ *implies that for a curved element* K*, the function* $\boldsymbol{w}_h|_K$ *is not a polynomial of degree* $\leq p$*. Moreover, if all* $K \in \mathcal{T}_h$ *are non-curved (i.e.,* F_K *are linear for all* $K \in \mathcal{T}_h$*), then the spaces defined by* (6.193) *are identical with the spaces defined by* (6.44) *and* (6.45)*.*

Now let us describe how to evaluate the volume and boundary integrals over elements K and their sides Γ. We denote by

$$J_{F_K}(\hat{x}) = \frac{D\,F_K}{D\,\hat{x}}(\hat{x}), \quad \hat{x} \in \hat{K}, \tag{6.194}$$

the Jacobian matrix of the mapping F_K. Since F_K is a polynomial mapping of degree q, J_{F_K} is a polynomial mapping of degree $q-1$ in the variable $\hat{x} = (\hat{x}_1, \hat{x}_2)$. The components of the vector-valued test functions $\boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp}$ from (6.193) are defined on the curved elements K (adjacent to the boundary $\partial\Omega$) with the aid of the mapping F_K. Hence, for each $\boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp}$ and each $K \in \mathcal{T}_h$ there exists a function $\hat{\boldsymbol{\varphi}}_K \in (P_p(\hat{K}))^m$ such that

$$\hat{\boldsymbol{\varphi}}_K(\hat{x}) = \boldsymbol{\varphi}_h(F_K(\hat{x})), \quad \hat{x} \in \hat{K}. \tag{6.195}$$

In the following, we shall describe how to evaluate the volume and face integrals appearing in the definition of the forms $\boldsymbol{b}_h$ and $\boldsymbol{b}_h^L$ given by (6.93) and (6.123), respectively. Evaluating the integrals is based on the transformation to the reference element (or reference edge) with the aid of the substitution theorem.

Volume integrals

The volume integral of a product of two (or more) functions is simply expressed as

$$\int_K \boldsymbol{w}_h(x,t) \cdot \boldsymbol{\varphi}_h(x)\,\mathrm{d}x = \int_{\hat{K}} \hat{\boldsymbol{w}}_K(\hat{x},t) \cdot \hat{\boldsymbol{\varphi}}_K(\hat{x})|\det J_{F_K}(\hat{x})|\,\mathrm{d}\hat{x}, \quad K \in \mathcal{T}_h,\ t \in (0,T), \tag{6.196}$$

where $\hat{\boldsymbol{w}}_K(\hat{x},t) = \boldsymbol{w}_h|_K(F_K(\hat{x},t))$ and $\hat{\boldsymbol{\varphi}}_K$ is given by (6.195).

Moreover, the evaluation of the volume integral of a product of a function and the gradient of a function requires a transformation of the gradient with respect to the variable x to the gradient with respect to $\hat{x}$. Hence, we obtain

$$\begin{aligned}
&\int_K \sum_{s=1}^{d} \boldsymbol{f}_s(\boldsymbol{w}_h(x,t)) \cdot \frac{\partial \boldsymbol{\varphi}_h(x)}{\partial x_s}\,\mathrm{d}x \qquad (6.197)\\
&= \int_{\hat{K}} \sum_{s=1}^{d} \boldsymbol{f}_s(\hat{\boldsymbol{w}}_K(\hat{x},t)) \cdot \sum_{j=1}^{d} \frac{\partial \hat{\boldsymbol{\varphi}}_K(\hat{x})}{\partial \hat{x}_j} \frac{\partial F_{K,j}^{-1}(F_K(\hat{x}))}{\partial x_s} |\det J_{F_K}(\hat{x})|\,\mathrm{d}\hat{x},\ K \in \mathcal{T}_h,\ t \in (0,T),
\end{aligned}$$

where $F_{K,j}^{-1}$ denotes the j-th component of the inverse mapping F_K^{-1}. In order to compute the inverse mapping F_K^{-1}, we use the following relation written in the matrix form:

$$\frac{D\,F_K^{-1}}{Dx}(F_K(\hat{x})) = \left(\frac{D\,F_K}{D\hat{x}}(\hat{x})\right)^{-1} \tag{6.198}$$

following from the identity $x = F_K(F_K^{-1}(x))$. The computation of the inverse matrix in (6.198) is simpler than the evaluation of F_K^{-1}.

Face integrals

Finally, we describe the evaluation of face integrals along a curved edge in $\mathbb{R}^2$. The three-dimensional case can be generalized in a natural way. Let $\Gamma \in \mathcal{F}_h$ be a (possibly curved) edge of $K \in \mathcal{T}_h$. Our aim is to evaluate the integrals

$$\int_\Gamma f(x)\,\mathrm{d}S, \qquad \int_\Gamma \boldsymbol{f}(x) \cdot \boldsymbol{n}(x)\ \boldsymbol{\varphi}(x)\,\mathrm{d}S, \tag{6.199}$$

where $\boldsymbol{n}$ is the normal vector to Γ and $f : \Gamma \to \mathbb{R}$, $\boldsymbol{f} : \Gamma \to \mathbb{R}^2$ are given functions. Such type of integral appears in (6.93) in terms containing the numerical flux. Let us recall the definition of the face integral. If $\psi = (\psi_1, \psi_2) : [0,1] \to \Gamma$ is a parameterization of the edge Γ, then

$$\int_\Gamma f(x)\,\mathrm{d}S = \int_0^1 f(\psi(\xi))\sqrt{(\psi_1'(\xi))^2 + (\psi_2'(\xi))^2}\mathrm{d}\xi, \tag{6.200}$$

where $\psi_i'(\xi),\ i = 1, 2$, denotes the derivative of $\psi_i(\xi)$ with respect to ξ.

Integrals (6.199) are evaluated with the aid of a transformation to the reference element. Let $\hat{\Gamma}$ be an edge of the reference element $\hat{K}$ such that $K = F_K(\hat{K})$ and $\Gamma = F_K(\hat{\Gamma})$. We call $\hat{\Gamma}$ the *reference edge*. Let

$$x_{\hat{\Gamma}}(\xi) = (x_{\hat{\Gamma},1}(\xi),\, x_{\hat{\Gamma},2}(\xi)) : [0,1] \to \hat{\Gamma} \tag{6.201}$$

be a parametrization of the reference edge $\hat{\Gamma}$ preserving the counterclockwise orientation of the element boundary. Namely, the reference triangle given by (6.159) (with $d = 2$) has three reference edges parametrized by

$$\begin{aligned} x_{\hat{\Gamma}_1}(\xi) &= (\xi, 0), \quad \xi \in [0,1], \\ x_{\hat{\Gamma}_2}(\xi) &= (1-\xi, \xi), \quad \xi \in [0,1], \\ x_{\hat{\Gamma}_3}(\xi) &= (0, 1-\xi), \quad \xi \in [0,1]. \end{aligned} \tag{6.202}$$

Moreover, we use the notation $\dot{x}_{\hat{\Gamma}}(\xi) = \frac{\mathrm{d}}{\mathrm{d}\xi} x_{\hat{\Gamma}}(\xi) \in \mathbb{R}^2$ and have

$$\begin{aligned} \dot{x}_{\hat{\Gamma}_1} &= (1, 0), \quad \xi \in [0,1], \\ \dot{x}_{\hat{\Gamma}_2} &= (-1, 1), \quad \xi \in [0,1], \\ \dot{x}_{\hat{\Gamma}_3} &= (0, -1), \quad \xi \in [0,1]. \end{aligned} \tag{6.203}$$

Therefore, the edge Γ is parameterized by

$$\begin{aligned} x = F_K(x_{\hat{\Gamma}}(\xi)) &= \big(F_{K,1}(x_{\hat{\Gamma}}(\xi)), F_{K,2}(x_{\hat{\Gamma}}(\xi))\big) \\ &= \Big(F_{K,1}(\hat{x}_{\hat{\Gamma},1}(\xi), \hat{x}_{\hat{\Gamma},2}(\xi)), F_{K,2}(\hat{x}_{\hat{\Gamma},1}(\xi), \hat{x}_{\hat{\Gamma},2}(\xi))\Big), \quad \xi \in [0,1]. \end{aligned} \tag{6.204}$$

The first integral in (6.199) is transformed by

$$\begin{aligned} \int_\Gamma f(x)\,\mathrm{d}S &= \int_0^1 f(F_K(x_{\hat{\Gamma}}(\xi))) \left(\sum_{i=1}^2 \left(\frac{\mathrm{d}}{\mathrm{d}\xi} F_{K,i}(x_{\hat{\Gamma}}(\xi))\right)^2\right)^{1/2} \mathrm{d}\xi \\ &= \int_0^1 f(F_K(x_{\hat{\Gamma}}(\xi))) \left(\sum_{i,j=1}^2 \left(\frac{\partial F_{K,i}(x_{\hat{\Gamma}}(\xi))}{\partial \hat{x}_j} \dot{\hat{x}}_{\hat{\Gamma},j}(\xi)\right)^2\right)^{1/2} \mathrm{d}\xi \\ &= \int_0^1 f(F_K(x_{\hat{\Gamma}}(\xi))) \left|J_{F_K}(x_{\hat{\Gamma}}(\xi))\dot{x}_{\hat{\Gamma}}\right| \mathrm{d}\xi, \end{aligned} \tag{6.205}$$

where J_{F_K} is the Jacobian matrix of the mapping F_K multiplied by the vector $\dot{x}_{\hat{\Gamma}}$ given by (6.203) and $|\cdot|$ is the Euclidean norm of the vector. Let us note that if F_K is a linear mapping, then e is a straight edge and $\left|J_{F_K}(x_{\hat{\Gamma}}(\xi))\dot{x}_{\hat{\Gamma}}(\xi)\right|$ is equal to its length.

Now, we focus on the second integral from (6.199). Let $\boldsymbol{t}_\Gamma$ be the tangential vector to Γ defined by

$$\begin{aligned} \boldsymbol{t}_\Gamma(x(\xi)) =& (t_{\Gamma,1}(x(\xi)), t_{\Gamma,2}(x(\xi))) \\ =& \frac{\mathrm{d}}{\mathrm{d}\xi} F_K(x_{\hat{\Gamma}}(\xi)) = \left(J_{F_{K,1}}(x_{\hat{\Gamma}}(\xi))\dot{x}_{\hat{\Gamma}}(\xi), J_{F_{K,2}}(x_{\hat{\Gamma}}(\xi))\dot{x}_{\hat{\Gamma}}(\xi)\right). \end{aligned} \tag{6.206}$$

(If Γ is a straight line, then $\boldsymbol{t}_\Gamma$ is constant on Γ, it has the orientation of Γ and $|\boldsymbol{t}_\Gamma| = |\Gamma|$.) Now, by the rotation we obtain the normal vector $\boldsymbol{n}_\Gamma$ pointing outside of K, namely

$$\begin{aligned} &\boldsymbol{n}_\Gamma(x(\xi)) = (n_{\Gamma,1}(x(\xi)), n_{\Gamma,2}(x(\xi))), \\ &n_{\Gamma,1}(x(\xi)) = t_{\Gamma,2}(x(\xi)),\ n_{\Gamma,2}(x(\xi)) = -t_{\Gamma,1}(x(\xi)). \end{aligned} \tag{6.207}$$

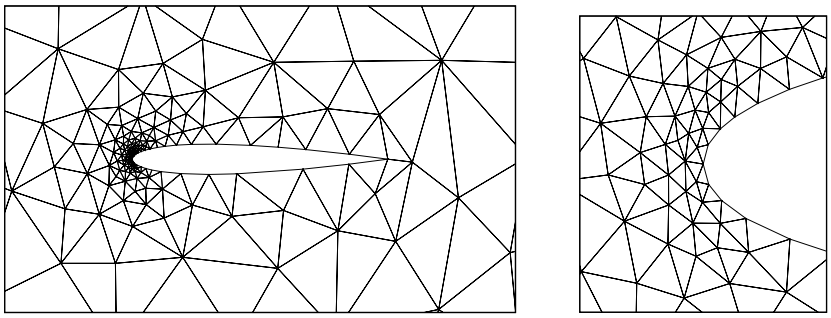

Figure 6.12: Subsonic inviscid flow around the NACA 0012 profile ($M_\infty = 0.5$, $\alpha = 2^\circ$): computational mesh, detail around the whole profile (left) and around the leading edge (right).

Here it is important that the counter-clockwise orientation of the elements is considered. Therefore, from (6.206) and (6.207), we have

$$\boldsymbol{n}_\Gamma(x(\xi)) = \left(J_{F_{K,2}}(x_{\hat\Gamma}(\xi))\dot{x}_{\hat\Gamma}(\xi), -J_{F_{K,1}}(x_{\hat\Gamma}(\xi))\dot{x}_{\hat\Gamma}(\xi)\right). \tag{6.208}$$

Let us note that because $\boldsymbol{n}_\Gamma(x(\xi))$ is not normalized, it is necessary to divide it by $|\boldsymbol{n}_\Gamma(x(\xi))| = |J_{F_K}(x_{\hat\Gamma}(\xi))\dot{x}_{\hat\Gamma}(\xi)|$. Finally, similarly as in (6.205), we obtain

$$\begin{aligned}
&\int_\Gamma \boldsymbol{f}(x)\cdot\boldsymbol{n}(x)\,\boldsymbol{\varphi}(x)\,\mathrm{d}S \qquad (6.209)\\
&= \int_0^1 \boldsymbol{f}(F_K(x_{\hat\Gamma}(\xi)))\cdot\frac{\boldsymbol{n}_\Gamma(x(\xi))}{|\boldsymbol{n}_\Gamma(x(\xi))|}\left|J_{F_K}(x_{\hat\Gamma}(\xi))\dot{x}_{\hat\Gamma}(\xi)\right|\,\boldsymbol{\varphi}(F_K(x_{\hat\Gamma}(\xi)))\,\mathrm{d}t\xi\\
&= \int_0^1 \boldsymbol{f}(F_K(x_{\hat\Gamma}(\xi)))\cdot\boldsymbol{n}_\Gamma(x(\xi))\,\hat{\boldsymbol{\varphi}}(x_{\hat\Gamma}(\xi))\,\mathrm{d}\xi,
\end{aligned}$$

where $\boldsymbol{n}_\Gamma(x(\xi))$ is given by (6.208) and $\hat{\boldsymbol{\varphi}}$ was obtained by transformation of the function $\boldsymbol{\varphi}$: $\hat{\boldsymbol{\varphi}}(\hat{x}) = \boldsymbol{\varphi}(F_K(\hat{x}))$. Let us note that if F_K is a linear mapping, then Γ is a straight edge and $|\boldsymbol{n}_\Gamma(x(\xi))|$ is equal to its length.

Implementation aspects of curved elements

The integrals over the reference triangle $\hat{K}$ and over the reference edge $\hat\Gamma$ in (6.196), (6.197), (6.205) and (6.209) are evaluated with the aid of suitable numerical quadratures. For the volume integrals we can employ the *Dunavant quadrature rules* [Dun85], which give the optimal order of accuracy of the numerical integration. For face integrals the well-known *Gauss quadrature rules*, having the maximal degree of approximation for the given number of integration nodes, can be used. For other possibilities, we refer to [ŠSD03].

Finally, let us mention the data structure in the implementation. Let $\hat{p}$ be an integer denoting the maximal implemented degree of the polynomial approximation in the DGM. We put $\hat{N} = (\hat{p}+1)(\hat{p}+2)/2$ denoting the corresponding maximal number of degrees of freedom for one element and one component of $\boldsymbol{w}$ for $d = 2$. Hence, in order to evaluate integrals appearing in (6.93) and (6.123) with the aid of the techniques presented above and with the aid of numerical quadratures, it is enough to evaluate (and store) the following quantities:

- for each $K \in \mathcal{T}_h$, the determinant $\det J_{F_K}$ of the Jacobi matrix and the transposed matrix to the inversion of the Jacobi matrix J_{F_K} evaluated at the used *edge and volume quadrature nodes*,
- the reference basis functions $\hat{\boldsymbol{\varphi}}_i(\hat{x}),\ i = 1,\dots,\hat{N}$, with their partial derivatives $\partial\hat{\boldsymbol{\varphi}}_i(\hat{x})/\partial\hat{x}_j,\ j = 1,2,\ i = 1,\dots\hat{N}$, on $\hat{K}$ evaluated at the used *edge and volume quadrature nodes*.

6.6.3 Numerical examples

In this section we present the results of numerical experiments demonstrating the influence of higher-order approximation of the nonpolygonal boundary. We consider an inviscid flow around the NACA 0012 profile with the far-field Mach number $M_\infty = 0.5$ (see (6.7)) and the angle of attack $\alpha = 2^\circ$. We seek the steady-state solution of the Euler equations (6.8) with the aid of the time stabilization described in Section 6.4.9, using the BE-DGM (6.95) combined with the Newton-like iterations (6.127) - (6.128).

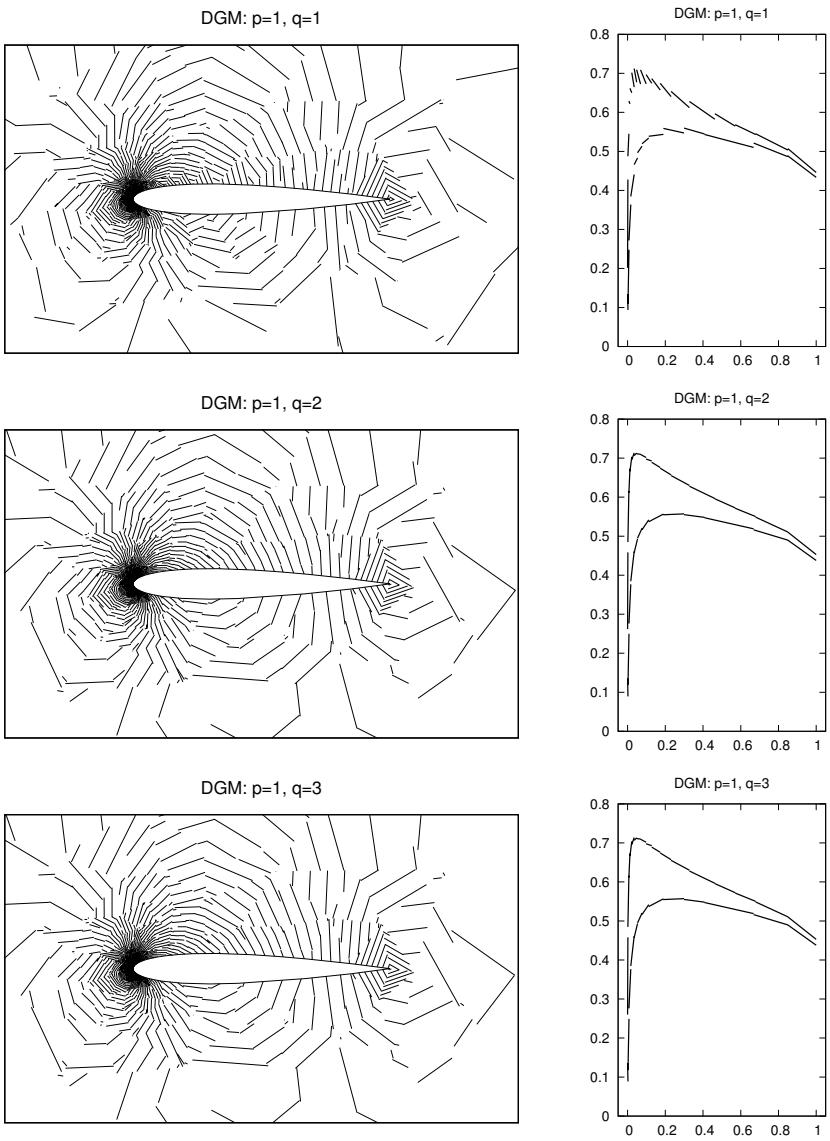

Figure 6.13: Subsonic inviscid flow around the NACA 0012 profile ($M_\infty = 0.5$, $\alpha = 2°$): DGM with polynomial approximation with $p = 1$, boundary approximation with $q = 1$ (top), $q = 2$ (center) and $q = 3$ (bottom), Mach number isolines (left) and the Mach number distribution around the profile (right).

The computation was performed on a coarse unstructured triangular grid having 507 elements, refined around the leading edge of the profile by the ANGENER code [Dol00] (see Figure 6.12). The polynomial approximations P_p, $p = 1, 3, 5$, in the DGM and the polynomial approximations P_q, $q = 1, 2, 3$, of the boundary described in Section 6.6 were used. Figures 6.13 – 6.15 show results of these computations, namely Mach number isolines and the Mach number distribution along the profile.

We observe that the P_1 approximation of the boundary produces nonphysical oscillations in the solution. This unpleasant behaviour disappears for P_2 or P_3 approximation of the boundary. There is almost no difference between P_2 and P_3. Finally, it is possible to see that the high-order DG approximation (P_5) gives very smooth isolines even on a coarse grid.

6.7 Numerical verification of the BDF-DGM

In this section we shall present computational results demonstrating the robustness and accuracy of the BDF-DGM for solving the Euler equations.

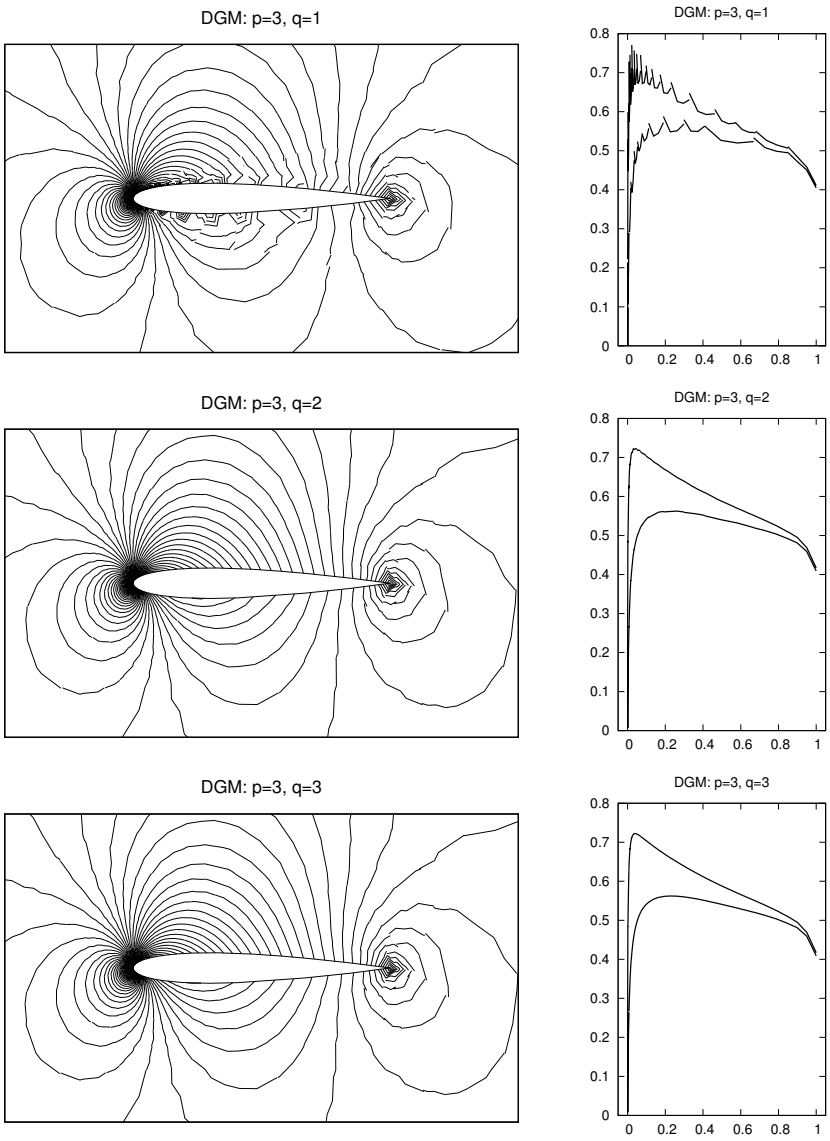

Figure 6.14: Subsonic inviscid flow around the NACA 0012 profile ($M_\infty = 0.5$, $\alpha = 2°$): DGM with polynomial approximation with $p = 3$, boundary approximation with $q = 1$ (top), $q = 2$ (center) and $q = 3$ (bottom), Mach number isolines (left) and the Mach number distribution around the profile (right).

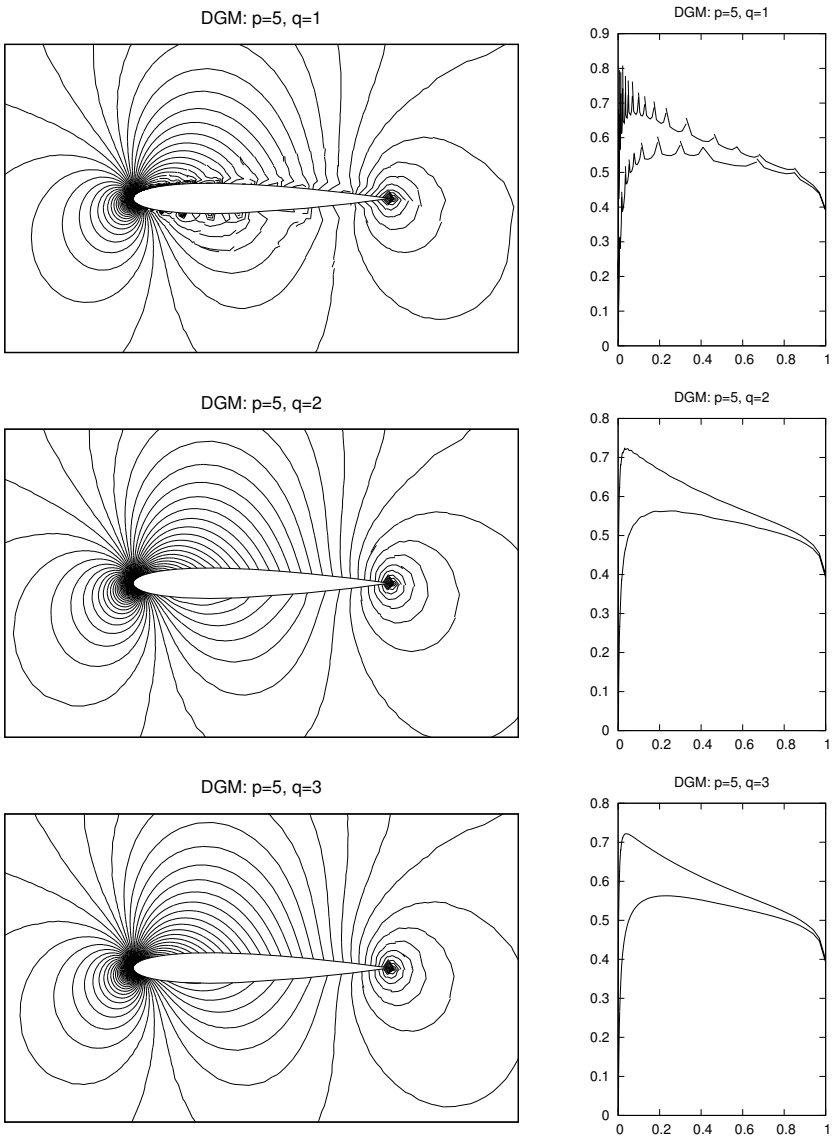

Figure 6.15: Subsonic inviscid flow around the NACA 0012 profile ($M_\infty = 0.5$, $\alpha = 2°$): DGM with polynomial approximation with $p = 5$, boundary approximation with $q = 1$ (top), $q = 2$ (center) and $q = 3$ (bottom), Mach number isolines (left) and the Mach number distribution around the profile (right).

6.7.1 Inviscid low Mach number flow

It is well-known that the numerical solution of low Mach number compressible flow is rather difficult. This is caused by the stiff behaviour of numerical schemes and acoustic phenomena appearing in low Mach number flows at incompressible limit. In this case, standard finite volume and finite element methods fail. This led to the development of special finite volume techniques allowing for the simulation of compressible flow at incompressible limit, which are based on modifications of the Euler or Navier–Stokes equations. We can mention works by Klein, Munz, Meister, Wesseling and their collaborators (see e.g. [Kle95], [RMGK97], [MS02, Chapter 5], or [Wes01, Chapter 14]). However, these techniques could not be applied to the solution of high speed flow. Therefore, further attempts were concentrated on extending these methods to solving flows at all speeds. A success in this direction was achieved by several authors. Let us mention, for example, the works by Wesseling et al. (e.g., [vdHVW03]), Parker and Munz ([PM05]), Meister ([Mei03]) and Darwish et al. ([DMS03]). The main ingredients of these techniques are finite volume schemes applied on staggered grids, combined with multigrid, the use of the pressure-correction, multiple pressure variables and flux preconditioning.

In 2007, in paper [FK07], it was discovered that the DG method described above allows the solution of compressible flow with practically all Mach numbers, without any modification of the governing equations, written in the conservative form with conservative variables. The robustness with respect to the magnitude of the Mach number of this method is based on the following ingredients:

- the application of the discontinuous Galerkin method for space discretization,
- special treatment of boundary conditions,
- (semi-)implicit time discretization,
- limiting of the order of accuracy in the vicinity of discontinuities based on the locally applied artificial viscosity,
- the use of curved elements near curved parts of the boundary.

In this section we present results of numerical examples showing that the described DG method allows for the low Mach number flow, nearly at incompressible limit. First, we solve stationary inviscid low Mach number flow around the NACA 0012 profile similarly as in [BBHN09]. The angle of attack is equal to zero and the far-field Mach number M_∞ is equal to 10^{-1}, 10^{-2}, 10^{-3} and 10^{-4}. The computation was carried out on a grid having 3587 elements (see Figure 6.16, bottom) with the aid of the 3-steps BDF-DGM with P_p, $p = 1, 2, 3, 4$, polynomial approximation in space. The computations are stop when the relative residuum steady-state criterion (6.171) is achieved for $\mathrm{TOL} = 10^{-5}$.

Table 6.6 shows the relative maximum pressure and density variations $(\mathrm{p}_{\max} - \mathrm{p}_{\min})/\mathrm{p}_{\max}$ and $(\rho_{\max} - \rho_{\min})/\rho_{\max}$, respectively, the *drag coefficient* c_D and the *lift coefficient* c_L, see (6.172). Let us note that

$$\mathrm{p}_{\max} = \max_{x\in\Omega} \mathrm{p}_h(x), \qquad \mathrm{p}_{\min} = \min_{x\in\Omega} \mathrm{p}_h(x), \qquad \rho_{\max} = \max_{x\in\Omega} \rho_h(x), \qquad \rho_{\min} = \min_{x\in\Omega} \rho_h(x),$$

where $\mathrm{p}_h(x)$ and $\rho_h(x)$ are the numerical approximations of the pressure and the density, respectively, evaluated from $\boldsymbol{w}_h$.

Both the pressure and density maximum variations are of order M_∞^2, which is in agreement with theoretical results in the analysis of compressible flow at incompressible limit. One can also see that the drag and lift coefficients attain small values, which correspond to the fact that in inviscid flow around a symmetric airfoil with zero angle of attack these quantities vanish. Figure 6.16 shows the pressure isolines obtained with the aid of P_1 and P_4 approximations.

6.7.2 Low Mach number flow at incompressible limit

It is well-known that compressible flow with a very low Mach number is very close to incompressible flow. This fact allows us to test the quality of numerical schemes for solving compressible low Mach number flow using a comparison with exact solutions of the corresponding incompressible flow, which are available in some cases. Here we present two examples of stationary compressible flow compared with incompressible flow. The steady-state solution was obtained with the aid of the time stabilization using the backward Euler linearized semi-implicit scheme (6.130). The computational grids were constructed with the aid of the anisotropic mesh adaptation technique by the ANGENER code [Dol00]. In both examples quadratic elements ($p = 2$) were applied.

Irrotational flow around a Joukowski profile

We consider flow around a negatively oriented Joukowski profile given by parameters $\Delta = 0.07, a = 0.5, h = 0.05$ (under the notation from [Fei93], Section 2.2.68) with zero angle of attack. The far-field quantities are constant, which implies that the flow is irrotational and homoentropic. Using the complex function method from [Fei93], we can obtain the exact solution of incompressible inviscid irrotational flow satisfying the Kutta–Joukowski trailing condition, provided the velocity circulation around the profile, related to the magnitude of the far-field velocity, $\gamma_{\mathrm{ref}} = 0.7158$. We assume that the far-field Mach number of

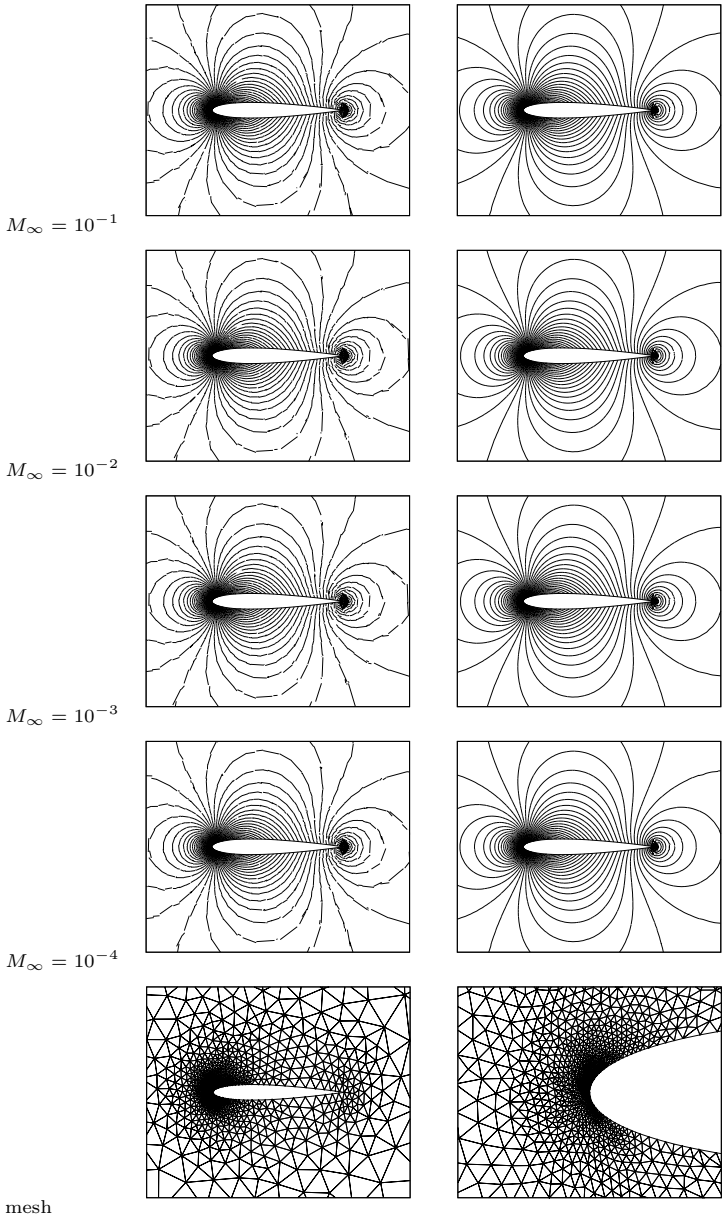

Figure 6.16: Low Mach number flow around the NACA 0012 profile for far-field Mach number $M_\infty = 10^{-1}$, 10^{-2}, 10^{-3} and 10^{-4}, with the aid of P_1 (left) and P_4 (right) polynomial approximation: pressure isolines and the used mesh with its detail (bottom).

M_∞	p	$\frac{p_{\max}-p_{\min}}{p_{\max}}$	$\frac{\rho_{\max}-\rho_{\min}}{\rho_{\max}}$	c_D	c_L
10^{-1}	1	9.89E-03	7.08E-03	2.57E-04	1.46E-03
10^{-1}	2	9.87E-03	7.09E-03	6.63E-05	1.20E-03
10^{-1}	3	9.87E-03	7.06E-03	4.26E-05	7.97E-04
10^{-1}	4	9.87E-03	7.06E-03	1.90E-05	6.83E-04
10^{-2}	1	9.92E-05	7.10E-05	3.80E-04	1.80E-03
10^{-2}	2	9.91E-05	7.11E-05	9.63E-05	1.22E-03
10^{-2}	3	9.90E-05	7.65E-05	4.68E-05	1.11E-03
10^{-2}	4	9.91E-05	7.13E-05	-5.73E-05	3.01E-04
10^{-3}	1	9.92E-07	7.11E-07	3.95E-04	1.57E-03
10^{-3}	2	9.93E-07	7.56E-07	3.74E-05	4.75E-04
10^{-3}	3	9.90E-07	7.08E-07	5.70E-05	8.96E-04
10^{-3}	4	9.90E-07	7.08E-07	3.69E-05	6.64E-04
10^{-4}	1	9.88E-09	4.84E-08	-1.69E-05	5.42E-04
10^{-4}	2	9.91E-09	8.29E-08	1.17E-04	1.10E-03
10^{-4}	3	9.90E-09	2.51E-08	-9.56E-06	5.02E-04
10^{-4}	4	9.93E-09	3.32E-08	-2.80E-04	3.17E-04

Table 6.6: Low Mach number flow around the NACA 0012 profile for far-field Mach number $M_\infty = 10^{-1}$, 10^{-2}, 10^{-3} and 10^{-4}, with the aid of P_p, $p = 1, \dots, 4$, polynomial approximation: ratios $(\mathrm{p}_{\max} - \mathrm{p}_{\min})/\mathrm{p}_{\max}$, $(\rho_{\max} - \rho_{\min})/\rho_{\max}$, drag coefficient c_D and lift coefficient c_L.

compressible flow $M_\infty = 10^{-4}$. The computational domain is of the form of a square with side of the length equal to 10 chords of the profile from which the profile is removed. The mesh (in the whole computational domain) was formed by 5418 triangular elements and refined towards the profile. Figure 6.17 (top) shows a detail near the profile of the velocity isolines for the exact solution of incompressible flow and for the approximate solution of compressible flow. Further, in Figure 6.17 (bottom), the distribution of the velocity related to the far-field velocity and the pressure coefficient distribution around the profile is plotted in the direction from the leading edge to the trailing edge. The pressure coefficient was defined as $10^7 \cdot (\mathrm{p} - \mathrm{p}_\infty)$, where p_∞ denotes the far-field pressure.

The maximum density variation is $1.04 \cdot 10^{-8}$. The computed velocity circulation related to the magnitude of the far-field velocity is $\gamma_{\mathrm{refcomp}} = 0.7205$, which gives the relative error 0.66% with respect to the theoretical value γ_{ref} obtained for incompressible flow.

In order to establish the quality of the computed pressure of the low Mach compressible flow in a quantitative way, we introduce the function

$$B = \frac{\mathrm{p}}{\rho} + \frac{1}{2}|\boldsymbol{v}|^2, \tag{6.210}$$

which is constant for incompressible, inviscid, irrotational flow, as follows from the Bernoulli equation. In the considered compressible case, the relative variation of the function B, i.e., $(B_{\max} - B_{\min})/B_{\max} = 3.84 \cdot 10^{-6}$, where $B_{\max} = \max_{x\in\Omega} B(x)$ and $B_{\min} = \min_{x\in\Omega} B(x)$. This means that the Bernoulli equation is satisfied with a small error in the case of the compressible low Mach number flow computed by the developed method.

Rotational flow past a circular half-cylinder

In the second example we present the comparison of the exact solution of incompressible inviscid rotational flow past a circular half-cylinder, with center at the origin and diameter equal to one, and with an approximate solution of compressible flow. The far-field Mach number is 10^{-4} and the far-field velocity has the components $v_1 = x_2, v_2 = 0$. The analytical exact solution was obtained in [Fra61]. This flow is interesting for its corner vortices. The computational domain was chosen in the form of a rectangle with length 10 and width 5, from which the half-cylinder was cut off. The mesh was formed by 3541 elements. We present here computational results in the vicinity of the half-cylinder. Figure 6.18 shows streamlines of incompressible and compressible flow. Figure 6.18 (bottom) shows the velocity distribution along the half-cylinder in dependence on the variable $\vartheta - \pi/2$, where $\vartheta \in [0, \pi]$ is the angle from cylindrical coordinates. The maximum density variation is $3.44 \cdot 10^{-9}$.

Accuracy of the method

An interesting question is the order of accuracy of the semi-implicit DG method. We tested numerically the accuracy of the piecewise quadratic DG approximations of the stationary inviscid flow past a circular cylinder with the far-field velocity parallel to the axis x_1 and the Mach number $M_\infty = 10^{-4}$. The problem was solved in a computational domain in the form of a square with sides of length equal to 20 diameters of the cylinder. Table 6.7 presents the behaviour of the error in the magnitude

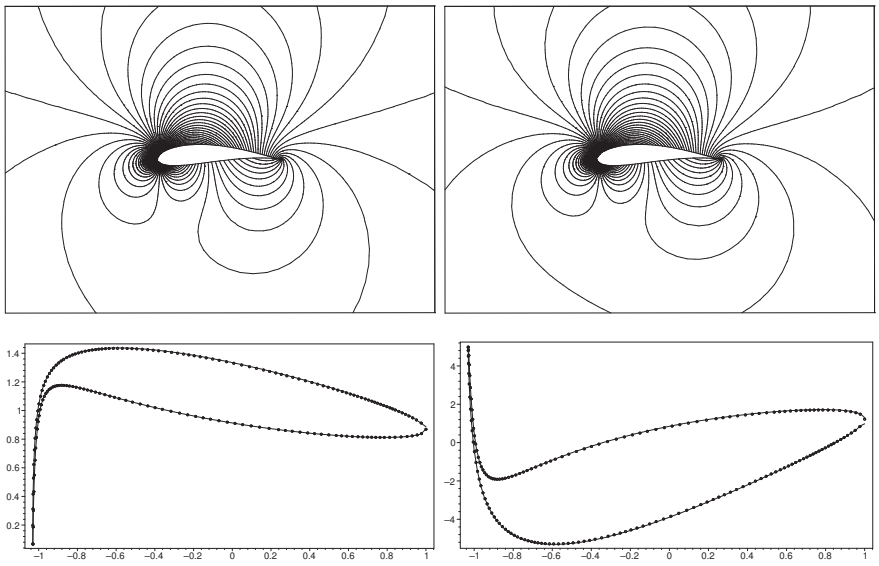

Figure 6.17: Flow around a Joukowski airfoil, velocity isolines for the exact solution of incompressible flow (top left) and approximate solution of compressible low Mach number flow (top right), velocity (left bottom) and pressure coefficient (right bottom) distribution along the profile: exact solution of incompressible flow (dots) and the approximate solution of compressible flow (full line).

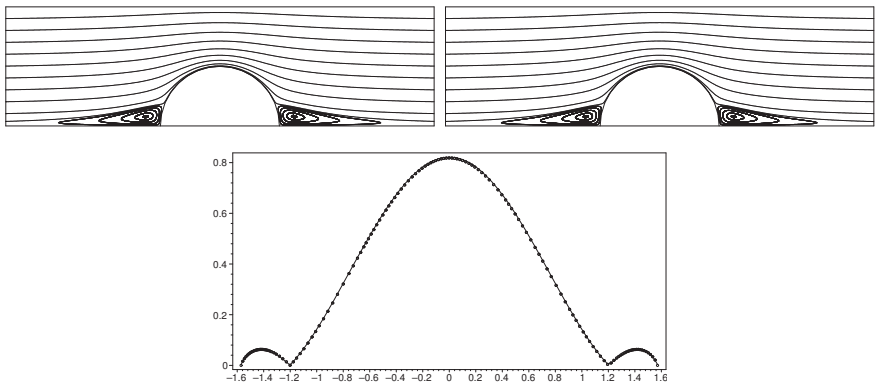

Figure 6.18: Flow past a half-cylinder, streamlines of rotational incompressible (top left) and compressible (top right) flows and the velocity distribution (bottom) on the half-cylinder incompressible flow (dots) and compressible flow (full line).

$\#\mathcal{T}_h$	$\|error\|_{L^\infty(\Omega)}$	EOC
1251	5.05E-01	–
1941	4.23E-01	0.406
5031	2.77E-02	2.86
8719	6.68E-03	2.59

Table 6.7: Error in the $L^\infty(\Omega)$-norm and corresponding experimental order of convergence for approximating incompressible flow by low Mach number compressible flow with respect to $h \to 0$.

of the velocity related to the far-field velocity and experimental order of convergence (EOC) for approximating of the exact incompressible solution by compressible low Mach number flow on successively refined meshes measured in the $L^\infty(\Omega)$-norm.

We see that the experimental order of convergence is close to 2.5, which is comparable to theoretical error estimate (in the $L^\infty(0,T;L^2(\Omega))$-norm) obtained in Section 2.6.

6.7.3 Isentropic vortex propagation

We consider the propagation of an isentropic vortex in a compressible inviscid flow, analyzed numerically in [Shu98]. This example is suitable for demonstrating the order of accuracy of the BDF-DGM, because the regular exact solution is known, and thus we can simply evaluate the computational error.

The computational domain is taken as $[0,10]\times[0,10]$ and extended periodically in both directions. The mean flow is $\bar\rho = 1$, $\bar{\boldsymbol{v}} = (1,1)$ (diagonal flow) and $\bar{\mathrm{p}} = 1$. To this mean flow we add an *isentropic vortex*, i.e., perturbation in $\boldsymbol{v}$ and the temperature $\theta = \mathrm{p}/\rho$, but no perturbation in the entropy $\eta = \mathrm{p}/\rho^\gamma$:

$$\delta\boldsymbol{v} = \frac{\varepsilon}{2\pi}\exp[(1-r^2)/2](-\bar x_2, \bar x_1), \quad \delta\theta = -\frac{(\gamma-1)\varepsilon^2}{8\gamma\pi^2}\exp[1-r^2], \quad \delta\eta = 0, \tag{6.211}$$

where $(-\bar x_2, \bar x_1) = (x_1 - 5, x_2 - 5)$, $r^2 = x_1^2 + x_2^2$, and the vortex strength $\varepsilon = 5$. The perturbations $\delta\rho$ and $\delta\mathrm{p}$ are obtained from the above relations according to

$$\bar\eta = \bar{\mathrm{p}}/\bar\rho^\gamma, \qquad \bar\theta = \bar{\mathrm{p}}/\bar\rho,$$
$$\delta\rho = \left(\frac{\bar\theta + \delta\theta}{\bar\eta}\right)^{1/(\gamma-1)} - \bar\rho, \qquad \delta\mathrm{p} = (\bar\rho + \delta\rho)(\bar\theta + \delta\theta) - \bar{\mathrm{p}}.$$

It is possible to see that the exact solution of the Euler equations with the initial conditions

$$\rho(x,0) = \bar\rho + \delta\rho, \quad \boldsymbol{v}(x,0) = \bar{\boldsymbol{v}} + \delta\boldsymbol{v}, \quad \mathrm{p}(x,0) = \bar{\mathrm{p}} + \delta\mathrm{p}, \tag{6.212}$$

and periodic boundary conditions is just the passive convection of the vortex with the mean velocity. Therefore, we are able to evaluate the computational error $\|\boldsymbol{w} - \boldsymbol{w}_{h\tau}\|$ over the space-time domain $Q_T := \Omega\times(0,T)$, where $\boldsymbol{w}$ is the exact solution and $\boldsymbol{w}_{h\tau}$ is the approximate solution obtained by the time interpolation of the approximate solution computed by the n-step BDF-DGM with the discretization parameters h and τ. This means that the function $\boldsymbol{w}_{h\tau}$ is defined by

$$\begin{aligned} &\boldsymbol{w}_{h\tau}(x,t_k) = \boldsymbol{w}_h^k(x),\ x\in\Omega,\ k = 0,\dots,r, \\ &\boldsymbol{w}_{h\tau}(x,t)|_{\Omega\times I_k} = \mathscr{L}^n(\boldsymbol{w}_h^{k+1}, \boldsymbol{w}_h^k, \dots, \boldsymbol{w}_k^{k-n+1})|_{\Omega\times I_k}, \end{aligned} \tag{6.213}$$

where $I_k = (t_{k-1}, t_k)$ and $\mathscr{L}^n$ is the Lagrange interpolation of degree n in the space $\mathbb{R}\times\boldsymbol{S}_{hp}$ constructed over the pairs

$$(t_{k-n+1}, \boldsymbol{w}_h^{k-n+1}),\ (t_{k-n+2}, \boldsymbol{w}_h^{k-n+2}),\ \dots,\ (t_k, \boldsymbol{w}_h^k),\ (t_{k+1}, \boldsymbol{w}_h^{k+1}).$$

In our computations we evaluate the following errors:

- $\|e_h(T)\|_{(L^2(\Omega))^m}$ – error over Ω at the final time T,
- $|e_h(T)|_{(H^1(\Omega))^m}$ – error over Ω at the final time T,
- $\|e_{h\tau}\|_{(L^2(Q_T))^m}$ – error over the space-time cylinder $\Omega\times(0,T)$,
- $\|e_{h\tau}\|_{(L^2(0,T;\,H^1(\Omega)))^m}$ – error over the space-time cylinder $\Omega\times(0,T)$.

h	τ	$k=n$	$\|e_h(T)\|_{L^2(\Omega)}$	$\|e_h(T)\|_{H^1(\Omega)}$	$\|e_{h\tau}\|_{L^2(Q_T)}$	$\|e_{h\tau}\|_{L^2(0,T;H^1(\Omega))}$
5.87E-01	1.00E-02	1	8.54E-01	1.69E+00	1.71E+00	4.01E+00
2.84E-01	5.00E-03	1	3.30E-01	7.56E-01	6.27E-01	1.81E+00
		EOC	(1.31)	(1.11)	(1.38)	(1.09)
1.41E-01	2.50E-03	1	1.50E-01	3.51E-01	2.82E-01	8.66E-01
		EOC	(1.13)	(1.10)	(1.15)	(1.06)
5.87E-01	1.00E-02	2	3.93E-02	2.40E-01	9.64E-02	7.10E-01
2.84E-01	5.00E-03	2	3.84E-03	5.05E-02	1.02E-02	1.61E-01
		EOC	(3.20)	(2.14)	(3.09)	(2.04)
1.41E-01	2.50E-03	2	6.69E-04	1.26E-02	1.55E-03	3.96E-02
		EOC	(2.51)	(1.99)	(2.70)	(2.01)
5.87E-01	1.00E-02	3	3.97E-03	3.75E-02	1.19E-02	1.30E-01
2.84E-01	5.00E-03	3	4.89E-04	5.04E-03	1.47E-03	1.56E-02
		EOC	(2.88)	(2.76)	(2.88)	(2.91)
1.41E-01	2.50E-03	3	1.14E-04	7.38E-04	3.45E-04	2.87E-03
		EOC	(2.09)	(2.76)	(2.08)	(2.43)

Table 6.8: Isentropic vortex propagation: computational errors and the corresponding EOC.

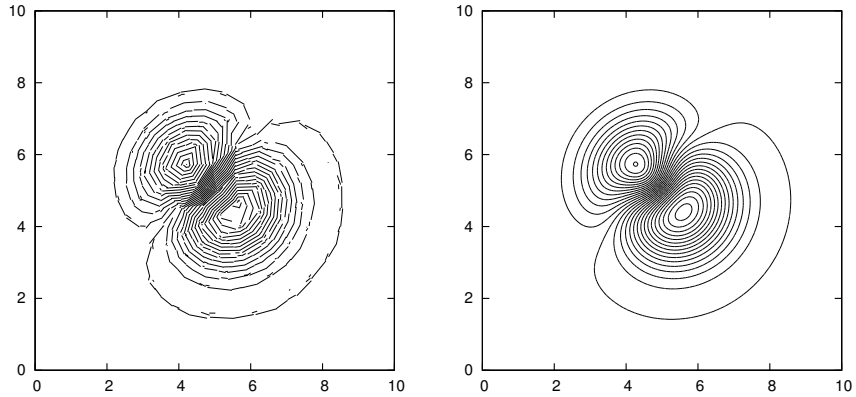

Figure 6.19: Isentropic vortex propagation: the isolines of the Mach number computed with the aid of P_1 approximation on the coarsest mesh (left) and P_3 approximation of the finest one (right).

We perform the computation on unstructured quasi-uniform triangular grids having 580, 2484 and 10008 elements, which corresponds to the average element size $h = 0.587$, $h = 0.284$ and $h = 0.141$, respectively. For each grid, we employ the k-step BDF-DGM with P_k polynomial approximation, $k = 1, 2, 3$. We use a fixed time step $\tau = 0.01$ on the coarsest mesh, $\tau = 0.005$ on the middle one and $\tau = 0.0025$ on the finest one. It means that the ratio h/τ is almost fixed for all computations. The final time was set $T = 10$.

Table 6.8 shows the computational errors in the norms mentioned above for each case and also the corresponding *experimental orders of convergence* (EOC). We observe that EOC measured in the H^1-seminorm is roughly $O(h^k)$ for $k = 1, 2, 3$, cf. Remarks 6.13 and 6.18. On the other hand, EOC measured in the L^2-norms are higher for $k = 2$ than for $k = 3$. However, the size of the error is smaller for $k = 3$ than for $k = 2$.

Moreover, Figure 6.19 shows the isolines of the Mach number for P_1 polynomial approximation on the coarsest mesh and for P_3 polynomial approximation on the finest mesh.

6.7.4 Supersonic flow

In order to demonstrate the applicability of the described DG schemes to the solution of supersonic flow with high Mach numbers, we present an inviscid supersonic flow around the NACA 0012 profile with the far-field Mach number $M_\infty = 2$ and the angle of attack $\alpha = 2°$. This flow produces a strong oblique shock wave in front the leading edge of the profile. The computation was performed on the anisotropically refined grid by the ANGENER code [Dol00] shown in Figure 6.20. We observe a strong refinement along shock waves. Some elements in front of the oblique shock wave are very obtuse, however the DGM was able to overcome this annoyance. Figure 6.21 shows the Mach number obtained with the aid of the P_3 approximation. Due to the applied shock capturing technique presented in Section 6.5 (with the same setting of all parameters α_1, α_2, ν_1 and ν_2), a good resolution of the shock waves is obtained.

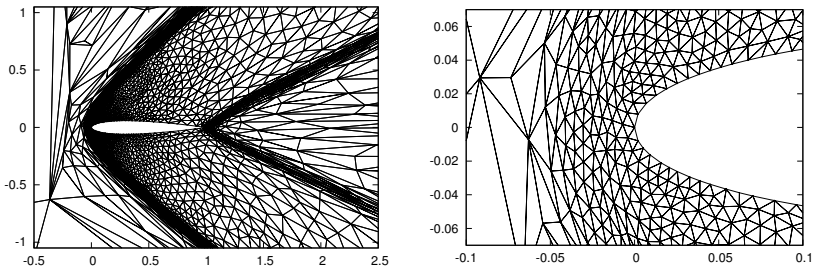

Figure 6.20: Supersonic flow around the NACA 0012 profile ($M_\infty = 2$, $\alpha = 2°$): the grid used, details around the profile (left) and the leading edge (right).

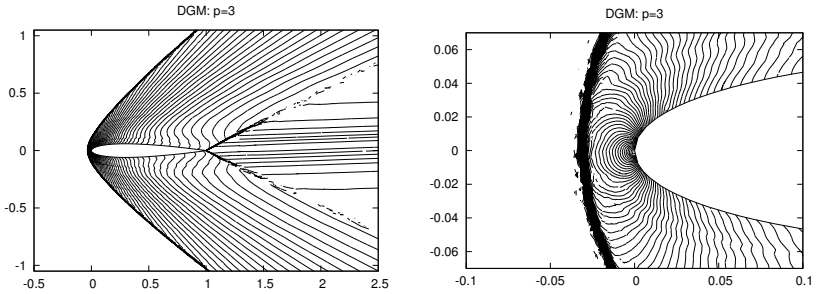

Figure 6.21: Supersonic flow around the NACA 0012 profile ($M_\infty = 2$, $\alpha = 2°$): Mach number isolines, around of the profile (left) and at the leading edge (right).

Chapter 7

Viscous compressible flow

This chapter is devoted to the numerical simulation of viscous compressible flow. The methods treated here represent the generalization of techniques for solving inviscid flow problems contained in Chapter 6. Viscous compressible flow is described by the continuity equation, the Navier–Stokes equations of motion and the energy equation, to which we add closing thermodynamical relations.

In the following, we introduce the DG space semidiscretization of the compressible Navier–Stokes equations with the aid of the interior penalty Galerkin (IPG) techniques. Since the convective terms were treated in detail in Chapter 6, we focus on discretization of viscous diffusion terms. We extend heuristically the approach developed in Chapter 1. Semidiscretization leads to a system of ordinary differential equations (ODEs), which is solved by the approach presented in Chapter 6 for the Euler equations. We demonstrate the accuracy, robustness and efficiency of the DG method in the solution of several flow problems.

7.1 Formulation of the viscous compressible flow problem

7.1.1 Governing equations

We shall consider unsteady compressible viscous flow in a domain $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) and time interval $(0, T)$ ($0 < T < \infty$). In what follows, we present the governing equations. Their derivation can be found, e.g., in [FFS03, Section 1.2].

We use the standard notation: ρ - density, p - pressure (symbol p denotes the degree of polynomial approximation), E - total energy, v_s - components of the velocity vector $\boldsymbol{v} = (v_1, \dots, v_d)^T$ in the directions x_s, $s = 1, \dots, d$, θ - absolute temperature, $c_v > 0$ - specific heat at constant volume, $c_{\mathrm{p}} > 0$ - specific heat at constant pressure, $\gamma = c_{\mathrm{p}}/c_v > 1$ - Poisson adiabatic constant, $R = c_{\mathrm{p}} - c_v > 0$ - gas constant, τ_{ij}^{V}, $i, j = 1, \dots, d$ - components of the viscous part of the stress tensor, $\boldsymbol{q} = (q_1, \dots, q_d)$ - heat flux. We will be concerned with the flow of a perfect gas, for which the equation of state (6.1) reads as

$$\mathrm{p} = R\rho\theta, \tag{7.1}$$

and assume that c_{p}, c_v are constants. Since the gas is light, we neglect the outer volume force and heat sources.

The system of governing equations formed by the continuity equation, the Navier–Stokes equations of motion and the energy equation (see [FFS03, Section 3.1]) considered in the space-time cylinder $Q_T = \Omega \times (0, T)$ can be written in the form

$$\frac{\partial \rho}{\partial t} + \sum_{s=1}^{d} \frac{\partial(\rho v_s)}{\partial x_s} = 0, \tag{7.2}$$

$$\frac{\partial(\rho v_i)}{\partial t} + \sum_{s=1}^{d} \frac{\partial(\rho v_i v_s + \delta_{is}\mathrm{p})}{\partial x_s} = \sum_{s=1}^{d} \frac{\partial \tau_{is}^{\mathrm{V}}}{\partial x_s}, \quad i = 1, \dots, d, \tag{7.3}$$

$$\frac{\partial E}{\partial t} + \sum_{s=1}^{d} \frac{\partial((E + \mathrm{p})v_s)}{\partial x_s} = \sum_{s,j=1}^{d} \frac{\partial(\tau_{sj}^{\mathrm{V}} v_j)}{\partial x_s} - \sum_{s=1}^{d} \frac{\partial q_s}{\partial x_s}, \tag{7.4}$$

$$\mathrm{p} = (\gamma - 1)(E - \rho|\boldsymbol{v}|^2/2). \tag{7.5}$$

As we see, system (7.2)–(7.4) consists of $m = d + 2$ partial differential equations. This whole system is usually simply called compressible Navier–Stokes equations. The total energy is defined by the relation

$$E = \rho(c_v\theta + |\boldsymbol{v}|^2/2). \tag{7.6}$$

The heat flux $\boldsymbol{q} = (q_1, \dots, q_d)$ satisfies the *Fourier law*

$$\boldsymbol{q} = -k\nabla\theta, \tag{7.7}$$

where $k > 0$ is the *heat conductivity* assumed here to be constant. This relation allows us to express the absolute temperature θ in terms of the quantities E, ρ and $|\boldsymbol{v}|^2$. Furthermore, we consider the Newtonian type of fluid, i.e., the viscous part of the stress tensor has the form

$$\tau^{\mathrm{V}}_{sk} = \mu \left(\frac{\partial v_s}{\partial x_k} + \frac{\partial v_k}{\partial x_s} \right) + \lambda \nabla \cdot \boldsymbol{v}\, \delta_{sk}, \quad s,k = 1, \ldots, d, \tag{7.8}$$

where δ_{sk} is the Kronecker symbol and $\mu > 0$ and λ are the viscosity coefficients. We assume that $\lambda = -\frac{2}{3}\mu$. It is valid, for example, for a monoatomic gas, but very often it is also used for more complicated gases.

Moreover, we recall the definition of the *speed of sound* a and the *Mach number* M by

$$a = \sqrt{\gamma \mathrm{p}/\rho}, \qquad M = |\boldsymbol{v}|/a. \tag{7.9}$$

It appears suitable to write and solve numerically the Navier–Stokes equations describing viscous compressible flow in a *dimensionless form*. We introduce the following positive *reference* (scalar) *quantities*: a reference length L^*, a reference velocity U^*, a reference density ρ^*. All other reference quantities can be derived from these basic ones: we choose L^*/U^* for t, $\rho^* U^{*2}$ for both p and E, U^{*3}/L^* for heat sources q, U^{*2}/c_v for θ. Then we can define the dimensionless quantities denoted here by primes:

$$\begin{aligned} &x_i' = x_i/L^*, \quad v_i' = v_i/U^*, \quad \boldsymbol{v}' = \boldsymbol{v}/U^*, \quad \rho' = \rho/\rho^*, \\ &\mathrm{p}' = \mathrm{p}/(\rho^* U^{*2}), \quad E' = E/(\rho^* U^{*2}), \quad \theta' = \frac{c_v \theta}{U^{*2}}, \quad t' = t\, U^*/L^*. \end{aligned} \tag{7.10}$$

Moreover, we introduce the *Reynolds number* Re and the *Prandtl number* Pr defined as

$$\mathrm{Re} = \rho^* U^* L^*/\mu, \quad \mathrm{Pr} = c_{\mathrm{p}}\, \mu/k. \tag{7.11}$$

In the sequel we denote the dimensionless quantities by the same symbols as the original dimensional quantities. This means that $\boldsymbol{v}$ will denote the dimensionless velocity, p will denote the dimensionless pressure, etc. Then system (7.2)–(7.4) can be written in the dimensionless form (cf. [FFS03])

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{s=1}^{d} \frac{\partial \boldsymbol{f}_s(\boldsymbol{w})}{\partial x_s} = \sum_{s=1}^{d} \frac{\partial \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w})}{\partial x_s} \quad \text{in } Q_T, \tag{7.12}$$

where

$$\boldsymbol{w} = (w_1, \ldots, w_{d+2})^{\mathrm{T}} = (\rho,\, \rho v_1, \ldots, \rho v_d,\, E)^{\mathrm{T}} \tag{7.13}$$

is the *state vector*,

$$\boldsymbol{f}_s(\boldsymbol{w}) = \begin{pmatrix} f_{s,1}(\boldsymbol{w}) \\ f_{s,2}(\boldsymbol{w}) \\ \vdots \\ f_{s,m-1}(\boldsymbol{w}) \\ f_{s,m}(\boldsymbol{w}) \end{pmatrix} = \begin{pmatrix} \rho v_s \\ \rho v_1 v_s + \delta_{1,s}\mathrm{p} \\ \vdots \\ \rho v_d v_s + \delta_{d,s}\mathrm{p} \\ (E + \mathrm{p}) v_s \end{pmatrix}, \ s = 1, \ldots, d, \tag{7.14}$$

are the *inviscid (Euler) fluxes* introduced already in (6.10). The expressions

$$\boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w}) = \begin{pmatrix} R_{s,1}(\boldsymbol{w}, \nabla \boldsymbol{w}) \\ R_{s,2}(\boldsymbol{w}, \nabla \boldsymbol{w}) \\ \vdots \\ R_{s,m-1}(\boldsymbol{w}, \nabla \boldsymbol{w}) \\ R_{s,m}(\boldsymbol{w}, \nabla \boldsymbol{w}) \end{pmatrix} = \begin{pmatrix} 0 \\ \tau^{\mathrm{V}}_{s1} \\ \vdots \\ \tau^{\mathrm{V}}_{sd} \\ \sum_{k=1}^{d} \tau^{\mathrm{V}}_{sk} v_k + \frac{\gamma}{\mathrm{Re\,Pr}} \frac{\partial \theta}{\partial x_s} \end{pmatrix}, \quad s = 1, \ldots, d, \tag{7.15}$$

represent the *viscous* and *heat conduction* terms, and

$$\tau^{\mathrm{V}}_{sk} = \frac{1}{\mathrm{Re}} \left(\frac{\partial v_s}{\partial x_k} + \frac{\partial v_k}{\partial x_s} - \frac{2}{3} \nabla \cdot \boldsymbol{v}\, \delta_{sk} \right), \quad s,k = 1, \ldots, d, \tag{7.16}$$

are the dimensionless components of the viscous part of the stress tensor. The dimensionless pressure and temperature are defined by

$$\mathrm{p} = (\gamma - 1)(E - \rho |\boldsymbol{v}|^2/2), \quad \theta = E/\rho - |\boldsymbol{v}|^2/2. \tag{7.17}$$

Of course, the set Q_T is obtained by the transformation of the original space-time cylinder using the relations for t' and x_i'.

The domain of definition of the vector-valued functions $\boldsymbol{f}_s$ and $\boldsymbol{R}_s$, $s = 1, \dots, d$, is the open set $\mathcal{D} \subset \mathbb{R}^m$ of vectors $\boldsymbol{w} = (w_1, \dots, w_m)^T$ such that the corresponding density and pressure are positive:

$$\mathcal{D} = \Big\{ \boldsymbol{w} \in \mathbb{R}^m; w_1 = \rho > 0,\ w_m - \sum_{i=2}^{m-1} w_i^2/(2w_1) = \mathrm{p}/(\gamma - 1) > 0 \Big\}. \tag{7.18}$$

Obviously, $\boldsymbol{f}_s, \boldsymbol{R}_s \in (C^1(\mathcal{D}))^m,\ s = 1, \dots, d$.

Similarly as in (6.13) – (6.17), the differentiation of the second term on the left-hand side of (7.12) and using the chain rule give

$$\sum_{s=1}^{d} \frac{\partial \boldsymbol{f}_s(\boldsymbol{w})}{\partial x_s} = \sum_{s=1}^{d} \mathbb{A}_s(\boldsymbol{w}) \frac{\partial \boldsymbol{w}}{\partial x_s}, \tag{7.19}$$

where $\mathbb{A}_s(\boldsymbol{w})$ is the $m \times m$ Jacobi matrix of the mapping $\boldsymbol{f}_s$ defined for $\boldsymbol{w} \in \mathcal{D}$:

$$\mathbb{A}_s(\boldsymbol{w}) = \quad \frac{\mathrm{D}\boldsymbol{f}_s(\boldsymbol{w})}{\mathrm{D}\boldsymbol{w}} = \left(\frac{\partial f_{s,i}(\boldsymbol{w})}{\partial w_j} \right)_{i,j=1}^{m},\ s = 1, \dots, d. \tag{7.20}$$

Moreover, let

$$\mathrm{B}_1 = \{ \boldsymbol{n} \in \mathbb{R}^d;\ |\boldsymbol{n}| = 1 \} \tag{7.21}$$

denote the unit sphere in $\mathbb{R}^d$. Then, for $\boldsymbol{w} \in \mathcal{D}$ and $\boldsymbol{n} = (n_1, \dots, n_d)^T \in \mathrm{B}_1$ we denote

$$\boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n}) = \sum_{s=1}^{d} \boldsymbol{f}_s(\boldsymbol{w}) n_s, \tag{7.22}$$

which is the *physical flux* of the quantity $\boldsymbol{w}$ in the direction $\boldsymbol{n}$. Obviously, the Jacobi matrix $\mathrm{D}\boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n})/\mathrm{D}\boldsymbol{w}$ can be expressed in the form

$$\frac{\mathrm{D}\boldsymbol{P}(\boldsymbol{w}, \boldsymbol{n})}{\mathrm{D}\boldsymbol{w}} = \mathbb{P}(\boldsymbol{w}, \boldsymbol{n}) = \sum_{s=1}^{d} \mathbb{A}_s(\boldsymbol{w}) n_s. \tag{7.23}$$

The explicit form of the matrices $\mathbb{A}_s,\ s = 1, \dots, d$, and $\mathbb{P}$ is given in Exercises 6.2 – 6.5.

Furthermore, the viscous terms $\boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w})$ can be expressed in the form

$$\boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w}) = \sum_{k=1}^{d} \mathbb{K}_{s,k}(\boldsymbol{w}) \frac{\partial \boldsymbol{w}}{\partial x_k}, \quad s = 1, \dots, d, \tag{7.24}$$

where $\mathbb{K}_{s,k}(\cdot)$ are $m \times m$ matrices ($m = d + 2$) dependent on $\boldsymbol{w}$. These matrices $\mathbb{K}_{s,k} := (K_{s,k}^{(\alpha,\beta)})_{\alpha,\beta=1}^{d+2},\ s, k = 1, \dots, d$, have for $d = 3$ the following form:

$$\mathbb{K}_{1,1}(\boldsymbol{w}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ -\frac{4}{3}\frac{w_2}{\mathrm{Re}\, w_1^2} & \frac{4}{3}\frac{1}{\mathrm{Re}\, w_1} & 0 & 0 & 0 \\ -\frac{w_3}{\mathrm{Re}\, w_1^2} & 0 & \frac{1}{\mathrm{Re}\, w_1} & 0 & 0 \\ -\frac{w_4}{\mathrm{Re}\, w_1^2} & 0 & 0 & \frac{1}{\mathrm{Re}\, w_1} & 0 \\ K_{1,1}^{(5,1)} & \frac{1}{\mathrm{Re}}(\frac{4}{3} - \frac{\gamma}{\mathrm{Pr}})\frac{w_2}{w_1^2} & \frac{1}{\mathrm{Re}}(1 - \frac{\gamma}{\mathrm{Pr}})\frac{w_3}{w_1^2} & \frac{1}{\mathrm{Re}}(1 - \frac{\gamma}{\mathrm{Pr}})\frac{w_4}{w_1^2} & \frac{\gamma}{\mathrm{Re}\,\mathrm{Pr}}\frac{1}{w_1} \end{pmatrix}, \tag{7.25}$$

with $K_{1,1}^{(5,1)} = -\frac{1}{\mathrm{Re}} \left(\frac{4}{3} w_2^2 + w_3^2 + w_4^2 \right) / w_1^3 + \frac{\gamma}{\mathrm{Re}\,\mathrm{Pr}} \left(-w_5/w_1^2 + (w_2^2 + w_3^2 + w_4^2)/w_1^3 \right)$,

$$\mathbb{K}_{2,2}(\boldsymbol{w}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ -\frac{w_2}{\mathrm{Re}\, w_1^2} & \frac{1}{\mathrm{Re}\, w_1} & 0 & 0 & 0 \\ -\frac{4}{3}\frac{w_3}{\mathrm{Re}\, w_1^2} & 0 & \frac{4}{3}\frac{1}{\mathrm{Re}\, w_1} & 0 & 0 \\ -\frac{w_4}{\mathrm{Re}\, w_1^2} & 0 & 0 & \frac{1}{\mathrm{Re}\, w_1} & 0 \\ K_{2,2}^{(5,1)} & \frac{1}{\mathrm{Re}}(1 - \frac{\gamma}{\mathrm{Pr}})\frac{w_2}{w_1^2} & \frac{1}{\mathrm{Re}}(\frac{4}{3} - \frac{\gamma}{\mathrm{Pr}})\frac{w_3}{w_1^2} & \frac{1}{\mathrm{Re}}(1 - \frac{\gamma}{\mathrm{Pr}})\frac{w_4}{w_1^2} & \frac{\gamma}{\mathrm{Re}\,\mathrm{Pr}}\frac{1}{w_1} \end{pmatrix}, \tag{7.26}$$

with $K_{2,2}^{(5,1)} = -\frac{1}{\mathrm{Re}}\left(w_2^2 + \frac{4}{3}w_3^2 + w_4^2\right)/w_1^3 + \frac{\gamma}{\mathrm{Re}\,\mathrm{Pr}}\left(-w_5/w_1^2 + (w_2^2+w_3^2+w_4^2)/w_1^3\right)$,

$$\mathbb{K}_{3,3}(\boldsymbol{w}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ -\frac{w_2}{\mathrm{Re}\,w_1^2} & \frac{1}{\mathrm{Re}\,w_1} & 0 & 0 & 0 \\ -\frac{w_3}{\mathrm{Re}\,w_1^2} & 0 & \frac{1}{\mathrm{Re}\,w_1} & 0 & 0 \\ -\frac{4}{3}\frac{w_4}{\mathrm{Re}\,w_1^2} & 0 & 0 & \frac{4}{3}\frac{1}{\mathrm{Re}\,w_1} & 0 \\ K_{3,3}^{(5,1)} & \frac{1}{\mathrm{Re}}(1-\frac{\gamma}{\mathrm{Pr}})\frac{w_2}{w_1^2} & \frac{1}{\mathrm{Re}}(1-\frac{\gamma}{\mathrm{Pr}})\frac{w_3}{w_1^2} & \frac{1}{\mathrm{Re}}(\frac{4}{3}-\frac{\gamma}{\mathrm{Pr}})\frac{w_4}{w_1^2} & \frac{\gamma}{\mathrm{Re}\,\mathrm{Pr}}\frac{1}{w_1} \end{pmatrix}, \tag{7.27}$$

with $K_{3,3}^{(5,1)} = -\frac{1}{\mathrm{Re}}\left(w_2^2 + w_3^2 + \frac{4}{3}w_4^2\right)/w_1^3 + \frac{\gamma}{\mathrm{Re}\,\mathrm{Pr}}\left(-w_5/w_1^2 + (w_2^2+w_3^2+w_4^2)/w_1^3\right)$,

$$\mathbb{K}_{1,2}(\boldsymbol{w}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \frac{2}{3}\frac{w_3}{\mathrm{Re}\,w_1^2} & 0 & -\frac{2}{3}\frac{1}{\mathrm{Re}\,w_1} & 0 & 0 \\ -\frac{w_2}{\mathrm{Re}\,w_1^2} & \frac{1}{\mathrm{Re}\,w_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{3}\frac{w_2\,w_3}{\mathrm{Re}\,w_1^3} & \frac{w_3}{\mathrm{Re}\,w_1^2} & -\frac{2}{3}\frac{w_2}{\mathrm{Re}\,w_1^2} & 0 & 0 \end{pmatrix}, \tag{7.28}$$

$$\mathbb{K}_{1,3}(\boldsymbol{w}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \frac{2}{3}\frac{w_4}{\mathrm{Re}\,w_1^2} & 0 & 0 & -\frac{2}{3}\frac{1}{\mathrm{Re}\,w_1} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -\frac{w_2}{\mathrm{Re}\,w_1^2} & \frac{1}{\mathrm{Re}\,w_1} & 0 & 0 & 0 \\ -\frac{1}{3}\frac{w_2\,w_4}{\mathrm{Re}\,w_1^3} & \frac{w_4}{\mathrm{Re}\,w_1^2} & 0 & -\frac{2}{3}\frac{w_2}{\mathrm{Re}\,w_1^2} & 0 \end{pmatrix}, \tag{7.29}$$

$$\mathbb{K}_{2,1}(\boldsymbol{w}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ -\frac{w_3}{\mathrm{Re}\,w_1^2} & 0 & \frac{1}{\mathrm{Re}\,w_1} & 0 & 0 \\ \frac{2}{3}\frac{w_2}{\mathrm{Re}\,w_1^2} & -\frac{2}{3}\frac{1}{\mathrm{Re}\,w_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{3}\frac{w_2\,w_3}{\mathrm{Re}\,w_1^3} & -\frac{2}{3}\frac{w_3}{\mathrm{Re}\,w_1^2} & \frac{w_2}{\mathrm{Re}\,w_1^2} & 0 & 0 \end{pmatrix}, \tag{7.30}$$

$$\mathbb{K}_{2,3}(\boldsymbol{w}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \frac{2}{3}\frac{w_4}{\mathrm{Re}\,w_1^2} & 0 & 0 & -\frac{2}{3}\frac{1}{\mathrm{Re}\,w_1} & 0 \\ -\frac{w_3}{\mathrm{Re}\,w_1^2} & 0 & \frac{1}{\mathrm{Re}\,w_1} & 0 & 0 \\ -\frac{1}{3}\frac{w_3\,w_4}{\mathrm{Re}\,w_1^3} & 0 & \frac{w_4}{\mathrm{Re}\,w_1^2} & -\frac{2}{3}\frac{w_2}{\mathrm{Re}\,w_1^2} & 0 \end{pmatrix}, \tag{7.31}$$

$$\mathbb{K}_{3,1}(\boldsymbol{w}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ -\frac{w_4}{\mathrm{Re}\,w_1^2} & 0 & 0 & \frac{1}{\mathrm{Re}\,w_1} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \frac{2}{3}\frac{w_2}{\mathrm{Re}\,w_1^2} & -\frac{2}{3}\frac{1}{\mathrm{Re}\,w_1} & 0 & 0 & 0 \\ -\frac{1}{3}\frac{w_2\,w_4}{\mathrm{Re}\,w_1^3} & -\frac{2}{3}\frac{w_4}{\mathrm{Re}\,w_1^2} & 0 & \frac{w_2}{\mathrm{Re}\,w_1^2} & 0 \end{pmatrix}, \tag{7.32}$$

$$\mathbb{K}_{3,2}(\boldsymbol{w}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -\frac{w_4}{\mathrm{Re}\,w_1^2} & 0 & 0 & \frac{1}{\mathrm{Re}\,w_1} & 0 \\ \frac{2}{3}\frac{w_3}{\mathrm{Re}\,w_1^2} & 0 & -\frac{2}{3}\frac{1}{\mathrm{Re}\,w_1} & 0 & 0 \\ -\frac{1}{3}\frac{w_3\,w_4}{\mathrm{Re}\,w_1^3} & 0 & -\frac{2}{3}\frac{w_4}{\mathrm{Re}\,w_1^2} & \frac{w_3}{\mathrm{Re}\,w_1^2} & 0 \end{pmatrix}. \tag{7.33}$$

Exercise 7.1. *Verify the form of* $\mathbb{K}_{s,k},\ s,k=1,2,3$, *given by* (7.25)–(7.33).

Exercise 7.2. *Derive the form of* $\mathbb{K}_{s,k},\ s,k=1,2$, *for* $d=2$.

7.1.2 Initial and boundary conditions

In order to formulate the problem of viscous compressible flow, the system of the Navier–Stokes equations (7.12) has to be equipped with initial and boundary conditions. Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded computational domain with a piecewise smooth boundary $\partial\Omega$. We prescribe the *initial condition*

$$\boldsymbol{w}(x, 0) = \boldsymbol{w}^0(x), \quad x \in \Omega, \tag{7.34}$$

where $\boldsymbol{w}^0 : \Omega \to \mathcal{D}$ is a given vector-valued function.

Concerning the *boundary conditions*, we distinguish (as in Section 6) the following disjoint parts of the boundary $\partial\Omega$: *inlet* $\partial\Omega_i$, *outlet* $\partial\Omega_o$ and *impermeable walls* $\partial\Omega_W$, i.e., $\partial\Omega = \partial\Omega_i \cup \partial\Omega_o \cup \partial\Omega_W$. We prescribe the following boundary conditions on individual parts of the boundary:

$$\rho = \rho_D, \quad \boldsymbol{v} = \boldsymbol{v}_D, \ \sum_{k=1}^{d} \left(\sum_{l=1}^{d} \tau_{lk}^{\mathrm{V}} n_l \right) v_k + \frac{\gamma}{\mathrm{Re}\,\mathrm{Pr}} \frac{\partial \theta}{\partial \boldsymbol{n}} = 0 \quad \text{on } \partial\Omega_i, \tag{7.35}$$

$$\sum_{k=1}^{d} \tau_{sk}^{\mathrm{V}} n_k = 0, \ s = 1, \ldots, d, \quad \frac{\partial \theta}{\partial \boldsymbol{n}} = 0 \quad \text{on } \partial\Omega_o, \tag{7.36}$$

$$\boldsymbol{v} = 0, \ \frac{\partial \theta}{\partial \boldsymbol{n}} = 0 \quad \text{on } \partial\Omega_W, \tag{7.37}$$

where ρ_D and $\boldsymbol{v}_D$ are given functions and $\boldsymbol{n} = (n_1, \ldots, n_d)$ is the outer unit normal to $\partial\Omega$. Another possibility is to replace the *adiabatic boundary condition* (7.37) by

$$\boldsymbol{v} = 0, \quad \theta = \theta_D \quad \text{on } \partial\Omega_W, \tag{7.38}$$

with a given function θ_D defined on $\partial\Omega_W$. Moreover, in the sequel we shall also apply boundary conditions in the discretization of the convective terms, similarly as in Section 6.3.

Finally, we introduce two relations, which we employ in the DG discretization. If $\boldsymbol{w}$ is the state vector satisfying the outlet boundary condition (7.36), then, using (7.15) and (7.24), on $\partial\Omega_o$ we have

$$\sum_{s=1}^{d} \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w})\, n_s \Big|_{\partial\Omega_o} = \begin{pmatrix} 0 \\ \sum_{s=1}^{d} \tau_{s1}^{\mathrm{V}} n_s \\ \vdots \\ \sum_{s=1}^{d} \tau_{sd}^{\mathrm{V}} n_s \\ \sum_{k,s=1}^{d} \tau_{sk}^{\mathrm{V}} n_k v_s + \frac{\gamma}{\mathrm{Re}\,\mathrm{Pr}} \sum_{s=1}^{d} \frac{\partial \theta}{\partial x_s} n_s \end{pmatrix} = 0. \tag{7.39}$$

Therefore, condition (7.36) represents the so-called "do-nothing" boundary condition.

Moreover, if $\boldsymbol{w}$ is the state vector satisfying the no-slip wall boundary condition (7.37), then using (7.15) we have

$$\sum_{s,k=1}^{d} \mathbb{K}_{s,k}(\boldsymbol{w}) \frac{\partial \boldsymbol{w}}{\partial x_k} n_s \Big|_{\partial\Omega_W} = \begin{pmatrix} 0 \\ \sum_{s=1}^{d} \tau_{1s}^{\mathrm{V}} n_s \\ \vdots \\ \sum_{s=1}^{d} \tau_{ds}^{\mathrm{V}} n_s \\ 0 \end{pmatrix} =: \sum_{s,k=1}^{d} \mathbb{K}_{s,k}^{W}(\boldsymbol{w}) \frac{\partial \boldsymbol{w}}{\partial x_k} n_s \Big|_{\partial\Omega_W}, \tag{7.40}$$

where τ_{ks}^{V} are the components of the viscous part of the stress tensor and $\mathbb{K}_{s,k}^{W}$, $s, k = 1, \ldots, d$, are the matrices that have the last row equal to zero and the other rows are identical with the rows of $\mathbb{K}_{s,k}$, $s, k = 1, \ldots, d$, i.e.,

$$\mathbb{K}_{s,k}^{W} = (\mathbb{K}_{s,k}^{W,(i,j)})_{i,j=1}^{m}, \quad \text{where} \tag{7.41}$$

$$\mathbb{K}_{s,k}^{W,(i,j)} = \begin{cases} \mathbb{K}_{s,k}^{(i,j)} & \text{for } i = 1, \ldots, m-1, \ j = 1, \ldots, m, \\ 0 & \text{for } i = m, \ j = 1, \ldots, m, \end{cases} \quad s, k = 1, \ldots, d,$$

where $\mathbb{K}_{s,k}$ are given by (7.24).

7.2 DG space semidiscretization

In the following, we describe the discretization of the Navier–Stokes equations (7.12) by the DGM. Similarly as in Chapter 6, we derive the DG space semidiscretization leading to a system of ordinary differential equations.

7.2.1 Notation

We use the same notation as in Section 6.2.1. It means that we assume that the domain Ω is polygonal (if $d = 2$) or polyhedral (if $d = 3$), $\mathcal{T}_h$ is a triangulation of Ω and $\mathcal{F}_h$ denotes the set of all faces of elements from $\mathcal{T}_h$. Further, $\mathcal{F}_h^I$, $\mathcal{F}_h^i$, $\mathcal{F}_h^o$ and $\mathcal{F}_h^W$ denote the set of all interior, inlet, outlet and wall faces, respectively. Moreover, we put $\mathcal{F}_h^B = \mathcal{F}_h^W \cup \mathcal{F}_h^i \cup \mathcal{F}_h^o$. Each face $\Gamma \in \mathcal{F}_h$ is associated with a unit normal $\boldsymbol{n}_\Gamma$, which is the outer unit normal to $\partial\Omega$ on $\Gamma \in {\mathcal{F}_h}^B$.

Further, over $\mathcal{T}_h$ we define the *broken Sobolev space* of vector-valued functions

$$\boldsymbol{H}^2(\Omega, \mathcal{T}_h) = (H^2(\Omega, \mathcal{T}_h))^m, \tag{7.42}$$

where

$$H^2(\Omega, \mathcal{T}_h) = \{v : \Omega \to \mathbb{R};\ v|_K \in H^2(K)\ \forall K \in \mathcal{T}_h\} \tag{7.43}$$

is the broken Sobolev space of scalar functions introduced by (1.29) (cf. (6.39)–(6.40)). The symbols $[\boldsymbol{u}]_\Gamma$ and $\langle \boldsymbol{u} \rangle_\Gamma$ denote the jump and the mean value of $\boldsymbol{u} \in \boldsymbol{H}^2(\Omega, \mathcal{T}_h)$ on $\Gamma \in \mathcal{F}_h^I$ and $[\boldsymbol{u}]_\Gamma = \langle \boldsymbol{u} \rangle_\Gamma = \boldsymbol{u}|_\Gamma$ for $\Gamma \in \mathcal{F}_h^B$. The approximate solution is sought in the *space of piecewise polynomial functions*

$$\boldsymbol{S}_{hp} = (S_{hp})^m, \tag{7.44}$$

where

$$S_{hp} = \left\{v \in L^2(\Omega); v|_K \in P_p(K)\ \forall K \in \mathcal{T}_h\right\}. \tag{7.45}$$

Finally, let us note that the inviscid Euler fluxes $\boldsymbol{f}_s,\ s = 1, \dots, d$, are discretized (including the boundary conditions) with the same approach as presented in Section 6.2.2. Therefore, we will pay attention here mainly to the discretization of the viscous terms.

7.2.2 DG space semidiscretization of viscous terms

In order to derive the discrete problem, we assume that there exists an exact solution $\boldsymbol{w} \in C^1([0, T]; \boldsymbol{H}^2(\Omega, \mathcal{T}_h))$ of the Navier–Stokes equations (7.12). We multiply (7.12) by a test function $\boldsymbol{\varphi} \in \boldsymbol{H}^2(\Omega, \mathcal{T}_h)$, integrate over an element $K \in \mathcal{T}_h$, apply Green's theorem and sum over all $K \in \mathcal{T}_h$. Then we can formally write

$$\sum_{K \in \mathcal{T}_h} \int_K \frac{\partial \boldsymbol{w}}{\partial t} \cdot \boldsymbol{\varphi}\, \mathrm{d}x + \mathrm{Inv} + \mathrm{Vis} = 0, \tag{7.46}$$

where

$$\mathrm{Inv} = \sum_{K \in \mathcal{T}_h} \int_{\partial K} \sum_{s=1}^d \boldsymbol{f}_s(\boldsymbol{w}) n_s \cdot \boldsymbol{\varphi}\, \mathrm{d}S - \sum_{K \in \mathcal{T}_h} \int_K \sum_{s=1}^d \boldsymbol{f}_s(\boldsymbol{w}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_s}\, \mathrm{d}x \tag{7.47}$$

$$\mathrm{Vis} = -\sum_{K \in \mathcal{T}_h} \int_{\partial K} \sum_{s=1}^d \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w}) n_s \cdot \boldsymbol{\varphi}\, \mathrm{d}S + \sum_{K \in \mathcal{T}_h} \int_K \sum_{s=1}^d \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_s}\, \mathrm{d}x \tag{7.48}$$

represent the inviscid and viscous terms and $(n_1, \dots, n_d)$ is the outer unit normal to ∂K.

The inviscid terms Inv are discretized by the technique presented in Chapter 6, namely, by (6.53). Hence,

$$\mathrm{Inv} \approx \boldsymbol{b}_h(\boldsymbol{w}, \boldsymbol{\varphi}), \tag{7.49}$$

where $\boldsymbol{b}_h$ is the *convection form*, given by (6.93). Let us mention that now the inviscid mirror boundary condition (6.68) is replaced by the *viscous mirror boundary condition* with the *viscous mirror operator*

$$\mathscr{M}(\boldsymbol{w}) = (\rho, -\rho \boldsymbol{v}, E)^{\mathrm{T}}, \tag{7.50}$$

replacing (6.67).

Here, we focus on the discretization of the viscous terms Vis. Similarly as in (1.36), we rearrange the first term in (7.48) according to the type of faces Γ, i.e.,

$$\sum_{K \in \mathcal{T}_h} \int_{\partial K} \sum_{s=1}^d \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w}) n_s \cdot \boldsymbol{\varphi}\, \mathrm{d}S \tag{7.51}$$

$$= \sum_{\Gamma \in \mathcal{F}_h^I} \int_\Gamma \sum_{s=1}^d \langle \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w}) \rangle\ n_s \cdot [\boldsymbol{\varphi}]\, \mathrm{d}S + \sum_{\Gamma \in \mathcal{F}_h^B} \int_\Gamma \sum_{s=1}^d \boldsymbol{R}_s(\boldsymbol{w}, \nabla \boldsymbol{w})\, n_s \cdot \boldsymbol{\varphi}\, \mathrm{d}S.$$

Let us deal with treating of the boundary conditions on the outlet, where only the "Neumann" boundary conditions are prescribed. With the aid of (7.39), we immediately get the relation

$$\sum_{\Gamma\in\mathcal{F}_h^o}\int_\Gamma\sum_{s=1}^d \boldsymbol{R}_s(\boldsymbol{w},\nabla\boldsymbol{w})\,n_s\cdot\boldsymbol{\varphi}\,\mathrm{d}S=0. \tag{7.52}$$

Concerning the boundary conditions on the inlet and fixed walls, the situation is more complicated, because both the Dirichlet and Neumann boundary conditions are prescribed there. However, using (7.48), (7.51), (7.52) and (7.24), we obtain

$$\begin{aligned}\mathrm{Vis}=&\sum_{K\in\mathcal{T}_h}\int_K\sum_{s=1}^d \boldsymbol{R}_s(\boldsymbol{w},\nabla\boldsymbol{w})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,\mathrm{d}x\\ &-\sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma\sum_{s=1}^d\left\langle\sum_{k=1}^d\mathbb{K}_{s,k}(\boldsymbol{w})\frac{\partial\boldsymbol{w}}{\partial x_k}\right\rangle n_s\cdot[\boldsymbol{\varphi}]\,\mathrm{d}S\\ &-\sum_{\Gamma\in\mathcal{F}_h^i}\int_\Gamma\sum_{s=1}^d\sum_{k=1}^d\mathbb{K}_{s,k}(\boldsymbol{w})\frac{\partial\boldsymbol{w}}{\partial x_k}\,n_s\cdot\boldsymbol{\varphi}\,\mathrm{d}S\\ &-\sum_{\Gamma\in\mathcal{F}_h^W}\int_\Gamma\sum_{s=1}^d\sum_{k=1}^d\mathbb{K}_{s,k}(\boldsymbol{w})\frac{\partial\boldsymbol{w}}{\partial x_k}\,n_s\cdot\boldsymbol{\varphi}\,\mathrm{d}S.\end{aligned} \tag{7.53}$$

In the last term of (7.53), we shall use relation (7.40) following from the wall boundary condition (7.37). Hence, we obtain

$$\begin{aligned}\mathrm{Vis}=&\sum_{K\in\mathcal{T}_h}\int_K\sum_{s=1}^d \boldsymbol{R}_s(\boldsymbol{w},\nabla\boldsymbol{w})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,\mathrm{d}x\\ &-\sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma\sum_{s=1}^d\left\langle\sum_{k=1}^d\mathbb{K}_{s,k}(\boldsymbol{w})\frac{\partial\boldsymbol{w}}{\partial x_k}\right\rangle n_s\cdot[\boldsymbol{\varphi}]\,\mathrm{d}S\\ &-\sum_{\Gamma\in\mathcal{F}_h^i}\int_\Gamma\sum_{s=1}^d\sum_{k=1}^d\mathbb{K}_{s,k}(\boldsymbol{w})\frac{\partial\boldsymbol{w}}{\partial x_k}\,n_s\cdot\boldsymbol{\varphi}\,\mathrm{d}S\\ &-\sum_{\Gamma\in\mathcal{F}_h^W}\int_\Gamma\sum_{s=1}^d\sum_{k=1}^d\mathbb{K}^W_{s,k}(\boldsymbol{w})\frac{\partial\boldsymbol{w}}{\partial x_k}\,n_s\cdot\boldsymbol{\varphi}\,\mathrm{d}S.\end{aligned} \tag{7.54}$$

Similarly as in Section 1.4, relation (1.44), we have to add to the relation (7.54) a *stabilization term*, which vanishes for a smooth solution satisfying the Dirichlet boundary conditions. Analogous to scalar problems, by the formal exchange of arguments $\boldsymbol{w}$ and $\boldsymbol{\varphi}$ in the second term of (7.54), for the interior faces we obtain the expression

$$-\Theta\sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma\sum_{s=1}^d\left\langle\sum_{k=1}^d\mathbb{K}_{s,k}(\boldsymbol{w})\frac{\partial\boldsymbol{\varphi}}{\partial x_k}\right\rangle n_s\cdot[\boldsymbol{w}]\,\mathrm{d}S \tag{7.55}$$

with $\Theta=-1$ or 1 depending on the type of stabilization, i.e., NIPG or SIPG variants. If we do not consider this stabilization, i.e., if $\Theta=0$, we get the simple IIPG variant. However, numerical experiments indicate that this choice of stabilization is not suitable. It is caused by the fact that for $\boldsymbol{\varphi}=(\varphi_1,0,\dots,0)^{\mathrm{T}},\ \varphi_1\in H^2(\Omega,\mathcal{T}_h),\ \varphi_1\neq\mathrm{const}$, we obtain a nonzero term (7.55), whereas all terms in (7.54) are equal to zero, because the first rows of $\boldsymbol{R}_s,\ \mathbb{K}_{s,k}, s,k=1,\dots,d$, vanish, see (7.15) and (7.25)–(7.33). This means that we would get nonzero additional terms on the right-hand side of the continuity equation, which is zero in the continuous problem. Therefore, in [BO99], [HH06a], [HH06b], the stabilization term

$$-\Theta\sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma\sum_{s=1}^d\left\langle\sum_{k=1}^d\mathbb{K}^{\mathrm{T}}_{s,k}(\boldsymbol{w})\frac{\partial\boldsymbol{\varphi}}{\partial x_k}\right\rangle n_s[\boldsymbol{w}]\,\mathrm{d}S \tag{7.56}$$

was proposed. This avoids the drawback mentioned above. Here, $\mathbb{K}^{\mathrm{T}}_{s,k}$ denotes the matrix transposed to $\mathbb{K}_{s,k},\ s,k=1,\dots,d$. Obviously, expression (7.56) vanishes for $\boldsymbol{w}(t)\in(H^2(\Omega))^m,\ t\in(0,T)$.

Moreover, similarly as in Section 1.4, we consider an extra stabilization term for the boundary faces, where at least one Dirichlet boundary condition is prescribed. Particularly, for the inlet part of the boundary, we add

$$-\Theta \sum_{\Gamma\in\mathcal{F}_h^i} \int_\Gamma \sum_{s,k=1}^{d} \mathbb{K}_{s,k}^{\mathrm{T}}(\boldsymbol{w}) \frac{\partial \boldsymbol{\varphi}}{\partial x_k} n_s(\boldsymbol{w}-\boldsymbol{w}_B)\,\mathrm{d}S, \tag{7.57}$$

where $\boldsymbol{w}_B$ is a *boundary state*. It is defined on the basis of the prescribed density ρ and the velocity $\boldsymbol{v}$ in condition (7.35) and the extrapolation of the absolute temperature. This yields the boundary state

$$\boldsymbol{w}_B|_\Gamma := (\rho_D, \rho_D v_{D,1}, \dots, \rho_D v_{D,d}, \rho_D \theta_\Gamma^{(L)} + \frac{1}{2}\rho_D |\boldsymbol{v}_D|^2)^{\mathrm{T}}, \quad \Gamma \in \mathcal{F}_h^i, \tag{7.58}$$

where $\theta_\Gamma^{(L)}$ is the trace of the temperature on $\Gamma \in \mathcal{F}_h^i$ from the interior of Ω, and ρ_D and $\boldsymbol{v}_D = (v_{D,1}, \dots, v_{D,d})$ are the prescribed density and velocity from (7.35), respectively.

In the case of the flow past an airfoil, when usually the far-field state vector $\boldsymbol{w}_{\mathrm{BC}}$ is prescribed, it is possible to define $\boldsymbol{w}_B$ to agree with the inviscid boundary conditions introduced in Section 6.3.2. In this case, we put

$$\boldsymbol{w}_B|_\Gamma := \mathscr{B}(\boldsymbol{w}_\Gamma^{(L)}, \boldsymbol{w}_{\mathrm{BC}}), \quad \Gamma \in \mathcal{F}_h^i, \tag{7.59}$$

where the inlet/outlet boundary operator $\mathscr{B}$ represents $\mathscr{B}^{\mathrm{phys}}$, $\mathscr{B}^{\mathrm{LRP}}$ and $\mathscr{B}^{\mathrm{RP}}$ given by (6.88), (6.85) and (6.92), respectively, and $\boldsymbol{w}_\Gamma^{(L)}$ is the trace of the state vector on $\Gamma \in \mathcal{F}_h^i$ from the interior of Ω.

The last term in (7.54) is stabilized by the expression

$$-\Theta \sum_{\Gamma\in\mathcal{F}_h^W} \int_\Gamma \sum_{s,k=1}^{d} (\mathbb{K}_{s,k}^{W}(\boldsymbol{w}))^{\mathrm{T}} \frac{\partial \boldsymbol{\varphi}}{\partial x_k} n_s(\boldsymbol{w}-\boldsymbol{w}_B)\,\mathrm{d}S, \tag{7.60}$$

where $(\mathbb{K}_{s,k}^{W}(\boldsymbol{w}))^{\mathrm{T}}$ is the transposed matrix to $\mathbb{K}_{s,k}^{W}(\boldsymbol{w})$, $s,k = 1,\dots,m$, and $\boldsymbol{w}_B$ is the prescribed boundary state vector. In the case of the adiabatic boundary condition (7.37), we define the *boundary state* as

$$\boldsymbol{w}_B|_\Gamma := (\rho_\Gamma^{(L)}, 0, \dots, 0, \rho_\Gamma^{(L)}\theta_\Gamma^{(L)})^{\mathrm{T}}, \quad \Gamma \in \mathcal{F}_h^W, \tag{7.61}$$

where $\rho_\Gamma^{(L)}$ and $\theta_\Gamma^{(L)}$ are the traces of the density and temperature on $\Gamma \in \mathcal{F}_h^W$ from the interior of Ω, respectively. In the case of the boundary condition (7.38), we put

$$\boldsymbol{w}_B|_\Gamma := (\rho_\Gamma^{(L)}, 0, \dots, 0, \rho_\Gamma^{(L)}\theta_D)^{\mathrm{T}}, \quad \Gamma \in \mathcal{F}_h^W, \tag{7.62}$$

where $\rho_\Gamma^{(L)}$ is the trace of the density on $\Gamma \in \mathcal{F}_h^W$ and θ_D is the prescribed temperature on the solid wall $\partial\Omega^W$.

As we see, the boundary state $\boldsymbol{w}_B$ depends partly on the unknown solution $\boldsymbol{w}$ and partly on the prescribed Dirichlet boundary conditions. Hence, we can write

$$\boldsymbol{w}_B = BC(\boldsymbol{w}, \boldsymbol{u}_D), \tag{7.63}$$

where $\boldsymbol{u}_D$ represents the Dirichlet boundary data and BC represents the definitions of boundary states (7.58), (7.59), (7.61) and (7.62).

Analogous to the DG discretization of the model problem in Section 1.4, for $\boldsymbol{w}$, $\boldsymbol{\varphi} \in \boldsymbol{H}^2(\Omega, \mathcal{T}_h)$ we define the *viscous form*

$$\begin{aligned}
\boldsymbol{a}_h(\boldsymbol{w}_h, \boldsymbol{\varphi}_h) = & \sum_{K\in\mathcal{T}_h} \int_K \sum_{s,k=1}^{d} \left(\mathbb{K}_{s,k}(\boldsymbol{w}_h) \frac{\partial \boldsymbol{w}_h}{\partial x_k} \right) \cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s} \,\mathrm{d}x \\
& - \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \sum_{s=1}^{d} \left\langle \sum_{k=1}^{d} \mathbb{K}_{s,k}(\boldsymbol{w}_h) \frac{\partial \boldsymbol{w}_h}{\partial x_k} \right\rangle n_s \cdot [\boldsymbol{\varphi}_h] \,\mathrm{d}S \\
& - \sum_{\Gamma\in\mathcal{F}_h^i} \int_\Gamma \sum_{s,k=1}^{d} \mathbb{K}_{s,k}(\boldsymbol{w}_h) \frac{\partial \boldsymbol{w}_h}{\partial x_k} \, n_s \cdot \boldsymbol{\varphi}_h \,\mathrm{d}S \\
& - \sum_{\Gamma\in\mathcal{F}_h^W} \int_\Gamma \sum_{s,k=1}^{d} \mathbb{K}^W_{s,k}(\boldsymbol{w}_h) \frac{\partial \boldsymbol{w}_h}{\partial x_k} \, n_s \cdot \boldsymbol{\varphi}_h \,\mathrm{d}S \\
& - \Theta \left(\sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \sum_{s,k=1}^{d} \left\langle \mathbb{K}^{\mathrm{T}}_{s,k}(\boldsymbol{w}_h) \frac{\partial \boldsymbol{\varphi}_h}{\partial x_k} \right\rangle n_s \cdot [\boldsymbol{w}_h] \,\mathrm{d}S \right. \\
& \quad + \sum_{\Gamma\in\mathcal{F}_h^i} \int_\Gamma \sum_{s,k=1}^{d} \mathbb{K}^{\mathrm{T}}_{s,k}(\boldsymbol{w}_h) \frac{\partial \boldsymbol{\varphi}_h}{\partial x_k} n_s \cdot (\boldsymbol{w}_h - \boldsymbol{w}_B) \,\mathrm{d}S \\
& \quad \left. + \sum_{\Gamma\in\mathcal{F}_h^W} \int_\Gamma \sum_{s,k=1}^{d} \left(\mathbb{K}^W_{s,k}(\boldsymbol{w}_h) \right)^{\mathrm{T}} \frac{\partial \boldsymbol{\varphi}_h}{\partial x_k} n_s \cdot (\boldsymbol{w}_h - \boldsymbol{w}_B) \,\mathrm{d}S \right).
\end{aligned} \tag{7.64}$$

We consider $\Theta = -1, 0, 1$ and get the NIPG, IIPG and SIPG variant of the viscous form, respectively.

Similarly as in Section 1.4, relations (1.41) – (1.42), in the scheme we include *interior and boundary penalty* terms, vanishing for the smooth solution satisfying the boundary conditions. Here we define the form

$$\begin{aligned}
\boldsymbol{J}_h^\sigma(\boldsymbol{w}_h, \boldsymbol{\varphi}_h) := & \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \sigma [\boldsymbol{w}_h] \cdot [\boldsymbol{\varphi}_h] \,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^i} \int_\Gamma \sigma (\boldsymbol{w}_h - \boldsymbol{w}_B) \cdot \boldsymbol{\varphi}_h \,\mathrm{d}S \\
& + \sum_{\Gamma\in\mathcal{F}_h^W} \int_\Gamma \sigma (\boldsymbol{w}_h - \boldsymbol{w}_B) \cdot \mathscr{V}(\boldsymbol{\varphi}_h) \,\mathrm{d}S,
\end{aligned} \tag{7.65}$$

where, in view of (7.63), $\boldsymbol{w}_B = BC(\boldsymbol{w}_h, \boldsymbol{u}_h)$ is the boundary state vector (given either by (7.58) or (7.59) for $\Gamma \in \mathcal{F}_h^i$ and either by (7.61) or (7.62) for $\Gamma \in \mathcal{F}_h^W$). The operator $\mathscr{V} : \mathbb{R}^{d+2} \to \mathbb{R}^{d+2}$ is defined as

$$\mathscr{V}(\boldsymbol{\varphi}) := (0, \varphi_2, \ldots, \varphi_{d+1}, 0)^{\mathrm{T}} \quad \text{for } \boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_{d+1}, \varphi_{d+2})^{\mathrm{T}}. \tag{7.66}$$

The role of $\mathscr{V}$ is to penalize only the components of $\boldsymbol{w}$, for which the Dirichlet boundary conditions are prescribed on fixed walls. Let us mention that we penalize all components of $\boldsymbol{w}$ on the inlet. It would also be possible to define a similar operator $\mathscr{V}$ for $\Gamma \in \mathcal{F}_h^i$. However, numerical experiments show that it is not necessary.

The penalty weight σ is chosen as

$$\sigma|_\Gamma = \frac{C_W}{\mathrm{diam}(\Gamma)\,\mathrm{Re}}, \quad \Gamma \in \mathcal{F}_h, \tag{7.67}$$

where Re is the Reynolds number of the flow, and $C_W > 0$ is a suitable constant which guarantees the stability of the method. Its choice depends on the variant of the DG method used (NIPG, IIPG or SIPG), see Section 7.4.1, where the choice of C_W is investigated with the aid of numerical experiments. The expression $\mathrm{diam}(\Gamma)$ can be replaced by the value h_Γ defined in Section 1.6. (Another possibility was used in [HH06a].)

We conclude that if $\boldsymbol{w}$ is a sufficiently regular exact solution of (7.12) satisfying the boundary conditions (7.35) - (7.37), then the viscous expression Vis from (7.48) can be rewritten in the form

$$\mathrm{Vis} = \boldsymbol{a}_h(\boldsymbol{w}, \boldsymbol{\varphi}) + \boldsymbol{J}_h^\sigma(\boldsymbol{w}, \boldsymbol{\varphi}) \quad \forall\, \boldsymbol{\varphi} \in \boldsymbol{H}^2(\Omega, \mathcal{T}_h). \tag{7.68}$$

7.2.3 Semidiscrete problem

Now, we complete the DG space semidiscretization of (7.12). By $(\cdot, \cdot)$ we denote the scalar product in the space $(L^2(\Omega))^{d+2}$:

$$(\boldsymbol{w}, \boldsymbol{\varphi}) = \int_\Omega \boldsymbol{w} \cdot \boldsymbol{\varphi} \,\mathrm{d}x, \quad \boldsymbol{w},\ \boldsymbol{\varphi} \in (L^2(\Omega))^{d+2}. \tag{7.69}$$

From (7.46), where we interchange the time derivative and integral in the first term, (7.47) and (7.68) we obtain the identity

$$\frac{\mathrm{d}}{\mathrm{d}t}\,(\boldsymbol{w}(t),\boldsymbol{\varphi}) + \boldsymbol{b}_h(\boldsymbol{w}(t),\boldsymbol{\varphi}) + \boldsymbol{a}_h(\boldsymbol{w}(t),\boldsymbol{\varphi}) + \boldsymbol{J}_h^\sigma(\boldsymbol{w}(t),\boldsymbol{\varphi}) = 0 \tag{7.70}$$
$$\forall\,\boldsymbol{\varphi}\in\boldsymbol{H}^2(\Omega,\mathcal{T}_h)\ \forall\,t\in(0,T),$$

In the discrete problem, because of the solution of high-speed flow containing discontinuities (shock waves and contact discontinuities, slightly smeared by the viscosity and heat conduction), we shall also consider the artificial viscosity forms $\boldsymbol{\beta}_h$ and $\boldsymbol{\gamma}_h$ introduced in (6.183) and (6.184), respectively. Therefore, we set

$$\begin{aligned}\boldsymbol{c}_h(\boldsymbol{w},\boldsymbol{\varphi}) =& \boldsymbol{b}_h(\boldsymbol{w},\boldsymbol{\varphi}) + \boldsymbol{a}_h(\boldsymbol{w},\boldsymbol{\varphi}) + \boldsymbol{J}_h^\sigma(\boldsymbol{w},\boldsymbol{\varphi}) \\ &+ \boldsymbol{\beta}_h(\boldsymbol{w},\boldsymbol{w},\boldsymbol{\varphi}) + \boldsymbol{\gamma}_h(\boldsymbol{w},\boldsymbol{w},\boldsymbol{\varphi}),\quad \boldsymbol{w},\boldsymbol{\varphi}\in\boldsymbol{H}^2(\Omega,\mathcal{T}_h),\end{aligned} \tag{7.71}$$

with the forms $\boldsymbol{b}_h$, $\boldsymbol{a}_h$, $\boldsymbol{J}_h^\sigma$, $\boldsymbol{\beta}_h$ and $\boldsymbol{\gamma}_h$ defined by (6.93), (7.64), (7.65), (6.183) and (6.184), respectively. The expressions in (7.70) – (7.71) make sense for $\boldsymbol{w},\boldsymbol{\varphi}\in\boldsymbol{H}^2(\Omega,\mathcal{T}_h)$. For each $t\in[0,T]$ the approximation of $\boldsymbol{w}(t)$ will be sought in the finite-dimensional space $\boldsymbol{S}_{hp}\subset\boldsymbol{H}^2(\Omega,\mathcal{T}_h)$ defined by (7.44) – (7.45). Using (7.70), we immediately arrive at the definition of an approximate solution.

Definition 7.3. *We say that a function $\boldsymbol{w}_h$ is the* space semidiscrete solution *of the compressible Navier–Stokes equations* (7.12)*, if the following conditions are satisfied:*

$$\boldsymbol{w}_h\in C^1([0,T];\boldsymbol{S}_{hp}), \tag{7.72a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\,(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h) + \boldsymbol{c}_h(\boldsymbol{w}_h(t),\boldsymbol{\varphi}_h) = 0\quad \forall\,\boldsymbol{\varphi}_h\in\boldsymbol{S}_{hp}\ \forall\,t\in(0,T), \tag{7.72b}$$

$$\boldsymbol{w}_h(0) = \Pi_h\boldsymbol{w}^0, \tag{7.72c}$$

where $\Pi_h\boldsymbol{w}^0$ is an $\boldsymbol{S}_{hp}$-approximation of $\boldsymbol{w}^0$ from the initial condition (7.34)*. Usually it is defined as the $L^2(\Omega)$-projection on the space $\boldsymbol{S}_{hp}$.*

7.3 Time discretization

The space semidiscrete problem (7.72) represents a system of N_{hp} ordinary differential equations (ODEs), where N_{hp} is equal to the dimension of the space $\boldsymbol{S}_{hp}$. This system has to be solved with the aid of a suitable numerical scheme. Often the Runge–Kutta methods are used. (See e.g., Section **??**.) However, they are conditionally stable and the CFL stability condition represents a strong restriction of the time step. This is the reason that we will be concerned with using implicit or semi-implicit time discretizations. We follow the approach developed in Section 6.4.1 and introduce the backward Euler and the BDF discretization of the ODE system (7.72). Then we develop the solution strategy of the corresponding nonlinear algebraic systems with the aid of the Newton-like method based on the flux matrix. In Chapter **??**, the full space-time discontinuous Galerkin method will be described and applied to the solution of flows in time-dependent domains and fluid-structure interaction problems.

7.3.1 Time discretization schemes

In what follows, we consider a partition $0=t_0<t_1<t_2\ldots<t_r=T$ of the time interval $[0,T]$ and set $\tau_k=t_k-t_{k-1},\ k=1,\ldots,r$. We use the symbol $\boldsymbol{w}_h^k$ for the approximation of $\boldsymbol{w}_h(t_k)$, $k=1,\ldots,r$.

Similarly as in Definitions 6.12 and 6.16, we define the following methods for the time discretization of (7.72).

Definition 7.4. *We say that the finite sequence of functions $\boldsymbol{w}_h^k,\ k=0,\ldots,r$, is an* approximate solution *of problem* (7.12) *obtained by the* backward Euler – discontinuous Galerkin *method (BE-DGM), if the following conditions are satisfied:*

$$\boldsymbol{w}_h^k\in\boldsymbol{S}_{hp},\ k=0,1,\ldots,r, \tag{7.73a}$$

$$\frac{1}{\tau_k}\left(\boldsymbol{w}_h^k-\boldsymbol{w}_h^{k-1},\boldsymbol{\varphi}_h\right) + \boldsymbol{c}_h(\boldsymbol{w}_h^k,\boldsymbol{\varphi}_h) = 0\quad \forall\,\boldsymbol{\varphi}_h\in\boldsymbol{S}_{hp},\ k=1,\ldots,r, \tag{7.73b}$$

$$\boldsymbol{w}_h^0 = \Pi_h\boldsymbol{w}^0, \tag{7.73c}$$

where $\Pi_h\boldsymbol{w}^0$ is the $\boldsymbol{S}_{hp}$-approximation of $\boldsymbol{w}^0$.

Definition 7.5. *We say that the finite sequence of functions $\boldsymbol{w}_h^k,\ k=0,\ldots,r$, is the* approximate solution *of* (7.12) *computed by the n-*step backward difference formula – discontinuous Galerkin method *(BDF-DGM) if the following conditions are satisfied:*

$$\boldsymbol{w}_h^k \in \boldsymbol{S}_{hp},\ k=0,1,\dots,r, \tag{7.74a}$$

$$\frac{1}{\tau_k}\left(\sum_{l=0}^{n}\alpha_{n,l}\boldsymbol{w}_h^{k-l},\boldsymbol{\varphi}_h\right)+\boldsymbol{c}_h\left(\boldsymbol{w}_h^k,\boldsymbol{\varphi}_h\right)=0\quad\forall\boldsymbol{\varphi}_h\in\boldsymbol{S}_{hp},\ k=n,\dots,r, \tag{7.74b}$$

$$\boldsymbol{w}_h^0=\Pi_h\boldsymbol{w}^0, \tag{7.74c}$$

$$\boldsymbol{w}_h^l\in\boldsymbol{S}_{hp},\ l=1,\dots,n-1,\ \textit{are determined by a suitable q-step method with } q<n \textit{ or by an explicit Runge–Kutta method – cf. Section ??.} \tag{7.74d}$$

The BDF coefficients $\alpha_{n,l},\ l=0,\dots,n$, depend on time steps $\tau_{k-l},\ l=0,\dots,n$. They can be derived from the Lagrange interpolation of pairs $[t_{k-l},\boldsymbol{w}_h^{k-l}],\ l=0,\dots,n$, see e.g. [HNW00]. Tables 6.2 and 6.3 show their values in the case of constant and variable time steps for $n=1,2,3$. One-step BDF-DGM is identical with BE-DGM defined by (7.73).

Remark 7.6. *By virtue of Remark 6.13 and Chapters 1–??, we expect that the BE-DGM has formally the order of accuracy $O(h^p+\tau)$ in the $L^\infty(0,T;L^2(\Omega))$-norm as well as in the $L^2(0,T;H^1(\Omega))$-seminorm, provided that the exact solution is sufficiently regular. Concerning the stability of the BDF-DGM, we refer to Remark 6.17.*

Schemes (7.73) and (7.74) represent nonlinear algebraic systems for each time level $t_k,\ k=1,\dots,r$, which should be solved by a suitable technique. It will be discussed in the following sections.

7.3.2 Solution strategy

Since the backward Euler method (7.73) is a special case of the BDF discretization (7.74), we deal here only with the latter case. The nonlinear algebraic system arising from (7.74) for each $k=n,\dots,r$ will be solved by the Newton-like method based on the approximation of the Jacobi matrix by the flux matrix, which was developed in Sections 6.4.3–6.4.5.

Again, let N_{hp} denote the dimension of the piecewise polynomial space $\boldsymbol{S}_{hp}$ and let $\mathrm{B}_{hp}=\{\boldsymbol{\varphi}_i(x),\ i=1,\dots,N_{hp}\}$ be a basis of $\boldsymbol{S}_{hp}$, see Section 6.4.8. Using the isomorphism (6.96) between $\boldsymbol{w}_h^k\in\boldsymbol{S}_{hp}$ and $\boldsymbol{\xi}_k\in\mathbb{R}^{N_{hp}}$, we define the vector-valued function $\boldsymbol{F}_h:(\mathbb{R}^{N_{hp}})^n\times\mathbb{R}^{N_{hp}}\to\mathbb{R}^{N_{hp}}$ by

$$\boldsymbol{F}_h\left(\left\{\boldsymbol{\xi}_{k-l}\right\}_{l=1}^n;\boldsymbol{\xi}_k\right)=\left(\frac{1}{\tau_k}\left(\sum_{l=0}^{n}\alpha_{n,l}\boldsymbol{w}_h^{k-l},\boldsymbol{\varphi}_i\right)+\boldsymbol{c}_h(\boldsymbol{w}_h^k,\boldsymbol{\varphi}_i)\right)_{i=1}^{N_{hp}},\quad k=n,\dots,r, \tag{7.75}$$

where $\boldsymbol{\xi}_{k-l}\in\mathbb{R}^{N_{hp}}$ is the algebraic representation of $\boldsymbol{w}_h^{k-l}\in\boldsymbol{S}_{hp}$ for $l=1,\dots,n$. We do not emphasize that $\boldsymbol{F}_h$ depends explicitly on τ_k. Then scheme (7.74) has the following algebraic representation. If $\boldsymbol{\xi}_{k-l},\ l=1,\dots,n,\ (k=1,\dots,r)$ are given vectors, then we want to find $\boldsymbol{\xi}_k\in\mathbb{R}^{N_{hp}}$ such that

$$\boldsymbol{F}_h(\left\{\boldsymbol{\xi}_{k-l}\right\}_{l=1}^n;\boldsymbol{\xi}_k)=\boldsymbol{0}. \tag{7.76}$$

System (7.76) is strongly nonlinear. In order to solve (7.76) with the aid of the Newton-like method based on the flux matrix, presented in Section 6.4.3, we have to linearize the form $\boldsymbol{c}_h$ similarly as the form $\boldsymbol{b}_h$ was linearized in (6.137).

To this end, on the basis of (7.64) we introduce the forms

$$\begin{aligned}\boldsymbol{a}_h^L(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h)=&\sum_{K\in\mathcal{T}_h}\int_K\sum_{s,k=1}^{d}\left(\mathbb{K}_{s,k}(\bar{\boldsymbol{w}}_h)\frac{\partial\boldsymbol{w}_h}{\partial x_k}\right)\cdot\frac{\partial\boldsymbol{\varphi}_h}{\partial x_s}\,\mathrm{d}x\\&-\sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma\sum_{s=1}^{d}\left\langle\sum_{k=1}^{d}\mathbb{K}_{s,k}(\bar{\boldsymbol{w}}_h)\frac{\partial\boldsymbol{w}_h}{\partial x_k}\right\rangle n_s\cdot[\boldsymbol{\varphi}_h]\,\mathrm{d}S\\&-\sum_{\Gamma\in\mathcal{F}_h^i}\int_\Gamma\sum_{s,k=1}^{d}\mathbb{K}_{s,k}(\bar{\boldsymbol{w}}_h)\frac{\partial\boldsymbol{w}_h}{\partial x_k}\,n_s\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S\\&-\sum_{\Gamma\in\mathcal{F}_h^W}\int_\Gamma\sum_{s,k=1}^{d}\mathbb{K}_{s,k}^W(\bar{\boldsymbol{w}}_h)\frac{\partial\boldsymbol{w}_h}{\partial x_k}\,n_s\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S\\&-\Theta\left(\sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma\sum_{s,k=1}^{d}\left\langle\mathbb{K}_{s,k}^{\mathrm{T}}(\bar{\boldsymbol{w}}_h)\frac{\partial\boldsymbol{\varphi}_h}{\partial x_k}\right\rangle n_s\cdot[\boldsymbol{w}_h]\,\mathrm{d}S\right.\end{aligned} \tag{7.77}$$

$$
\begin{aligned}
&+ \sum_{\Gamma\in\mathcal{F}_h^i}\int_\Gamma \sum_{s,k=1}^{d} \mathbb{K}_{s,k}^{\mathrm{T}}(\bar{\boldsymbol{w}}_h)\frac{\partial\boldsymbol{\varphi}_h}{\partial x_k}n_s\cdot\boldsymbol{w}_h\,\mathrm{d}S\\
&+ \sum_{\Gamma\in\mathcal{F}_h^W}\int_\Gamma \sum_{s,k=1}^{d} \left(\mathbb{K}_{s,k}^{W}(\bar{\boldsymbol{w}}_h)\right)^{\mathrm{T}}\frac{\partial\boldsymbol{\varphi}_h}{\partial x_k}n_s\cdot\boldsymbol{w}_h\,\mathrm{d}S\Bigg),\\
\tilde{\boldsymbol{a}}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_h) = &-\Theta\Bigg(\sum_{\Gamma\in\mathcal{F}_h^i}\int_\Gamma \sum_{s,k=1}^{d} \mathbb{K}_{s,k}^{\mathrm{T}}(\bar{\boldsymbol{w}}_h)\frac{\partial\boldsymbol{\varphi}_h}{\partial x_k}n_s\cdot\bar{\boldsymbol{w}}_B\,\mathrm{d}S\\
&+ \sum_{\Gamma\in\mathcal{F}_h^W}\int_\Gamma \sum_{s,k=1}^{d} \left(\mathbb{K}_{s,k}^{W}(\bar{\boldsymbol{w}}_h)\right)^{\mathrm{T}}\frac{\partial\boldsymbol{\varphi}_h}{\partial x_k}n_s\cdot\bar{\boldsymbol{w}}_B\,\mathrm{d}S\Bigg),
\end{aligned}
\tag{7.78}
$$

where $\bar{\boldsymbol{w}}_B = BC(\bar{\boldsymbol{w}}_h,\boldsymbol{u}_D)$ is the boundary state vector given either by (7.58) or (7.59) for $\Gamma\in\mathcal{F}_h^i$ and either by (7.61) or (7.62) for $\Gamma\in\mathcal{F}_h^W$. The above forms are consistent with the form $\boldsymbol{a}_h$:

$$
\boldsymbol{a}_h(\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \boldsymbol{a}_h^L(\boldsymbol{w}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) - \tilde{\boldsymbol{a}}_h(\boldsymbol{w}_h,\boldsymbol{\varphi}_h)\qquad \forall\,\boldsymbol{w}_h,\boldsymbol{\varphi}_h\in\boldsymbol{S}_{hp}. \tag{7.79}
$$

The form $\boldsymbol{a}_h^L$ is linear with respect to the second and third variables.

Furthermore, because of the penalty form $\boldsymbol{J}_h^\sigma$ given by (7.65), we introduce the forms

$$
\begin{aligned}
\boldsymbol{J}_h^{\sigma,L}(\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = &\sum_{\Gamma\in\mathcal{F}_h^I}\int_\Gamma \sigma[\boldsymbol{w}_h]\cdot[\boldsymbol{\varphi}_h]\,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^i}\int_\Gamma \sigma\boldsymbol{w}_h\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S\\
&+ \sum_{\Gamma\in\mathcal{F}_h^W}\int_\Gamma \sigma\boldsymbol{w}_h\cdot\mathscr{V}(\boldsymbol{\varphi}_h)\,\mathrm{d}S,
\end{aligned}
\tag{7.80}
$$

$$
\tilde{\boldsymbol{J}}_h^{\sigma}(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_h) = \sum_{\Gamma\in\mathcal{F}_h^i}\int_\Gamma \sigma\bar{\boldsymbol{w}}_B\cdot\boldsymbol{\varphi}_h\,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^W}\int_\Gamma \sigma\bar{\boldsymbol{w}}_B\cdot\mathscr{V}(\boldsymbol{\varphi}_h)\,\mathrm{d}S, \tag{7.81}
$$

where $\bar{\boldsymbol{w}}_B = BC(\bar{\boldsymbol{w}}_h,\boldsymbol{u}_D)$ is the boundary state vector corresponding to the function $\bar{\boldsymbol{w}}_h$. Obviously,

$$
\boldsymbol{J}_h^{\sigma}(\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \boldsymbol{J}_h^{\sigma,L}(\boldsymbol{w}_h,\boldsymbol{\varphi}_h) - \tilde{\boldsymbol{J}}_h^{\sigma}(\boldsymbol{w}_h,\boldsymbol{\varphi}_h)\qquad \forall\,\boldsymbol{w}_h,\boldsymbol{\varphi}_h\in\boldsymbol{S}_{hp}. \tag{7.82}
$$

Finally, let $\boldsymbol{b}_h$, $\boldsymbol{b}_h^L$ and $\tilde{\boldsymbol{b}}_h$ be the forms defined by (6.93), (6.123) and (6.121), respectively. By virtue of (7.71), we define the forms

$$
\begin{aligned}
\boldsymbol{c}_h^L(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) =&\boldsymbol{b}_h^L(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}) + \boldsymbol{a}_h^L(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) + \boldsymbol{J}_h^{\sigma,L}(\boldsymbol{w}_h,\boldsymbol{\varphi}_h)\\
&+\boldsymbol{\beta}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) + \boldsymbol{\gamma}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h),\quad \bar{\boldsymbol{w}}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h\in\boldsymbol{S}_{hp},\\
\tilde{\boldsymbol{c}}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_h) =&\tilde{\boldsymbol{b}}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_h) + \tilde{\boldsymbol{a}}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_h) + \tilde{\boldsymbol{J}}_h^{\sigma}(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_h),\quad \bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_h\in\boldsymbol{S}_{hp},
\end{aligned}
\tag{7.83}
$$

which together with (6.137), (7.79) and (7.82) imply consistency:

$$
\boldsymbol{c}_h(\boldsymbol{w}_h,\boldsymbol{\varphi}_h) = \boldsymbol{c}_h^L(\boldsymbol{w}_h,\boldsymbol{w}_h,\boldsymbol{\varphi}_h) - \tilde{\boldsymbol{c}}_h(\boldsymbol{w}_h,\boldsymbol{\varphi}_h),\quad \boldsymbol{w}_h,\boldsymbol{\varphi}_h\in\boldsymbol{S}_{hp}. \tag{7.84}
$$

Following directly the approach from Section 6.4.5, we transform problem (7.88b) into a system of algebraic equations. Instead of (6.138) and (6.139), for $k=n,\ldots,r$ we define the *flux matrix* $\mathbb{C}_h$ and the vector $\boldsymbol{d}_h$ by

$$
\mathbb{C}_h\left(\bar{\boldsymbol{\xi}}\right) = \left(\frac{\alpha_{n,0}}{\tau_k}\left(\boldsymbol{\varphi}_j,\boldsymbol{\varphi}_i\right) + \boldsymbol{c}_h^L(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_j,\boldsymbol{\varphi}_i)\right)_{i,j=1}^{N_{hp}} \tag{7.85}
$$

and

$$
\boldsymbol{d}_h\left(\left\{\boldsymbol{\xi}_{k-l}\right\}_{l=1}^n,\bar{\boldsymbol{\xi}}\right) = \left(\frac{1}{\tau_k}\left(\sum_{i=1}^{n}\alpha_{n,i}\boldsymbol{w}_h^{k-l},\boldsymbol{\varphi}_i\right) + \tilde{\boldsymbol{c}}_h(\bar{\boldsymbol{w}}_h,\boldsymbol{\varphi}_i)\right)_{i=1}^{N_{hp}}, \tag{7.86}
$$

respectively. Here $\boldsymbol{\varphi}_i\in \mathrm{B}_{hp},\ i=1,\ldots,N_{hp}$, are the basis functions in the space $\boldsymbol{S}_{hp}$, $\bar{\boldsymbol{\xi}}\in\mathbb{R}^{N_{hp}}$ and $\boldsymbol{\xi}_{k-l}\in\mathbb{R}^{N_{hp}},\ l=1,\ldots,n$, are the algebraic representations of $\bar{\boldsymbol{w}}_h\in\boldsymbol{S}_{hp}$ and $\boldsymbol{w}_h^{k-l}\in\boldsymbol{S}_{hp},\ l=1,\ldots,n$, respectively. Then problem (7.74) is equivalent to the nonlinear systems (compare with (6.126))

$$
\boldsymbol{F}_h(\left\{\boldsymbol{\xi}_{k-l}\right\}_{l=1}^n;\boldsymbol{\xi}_k) = \mathbb{C}_h(\boldsymbol{\xi}_k)\boldsymbol{\xi}_k - \boldsymbol{d}_h(\left\{\boldsymbol{\xi}_{k-l}\right\}_{l=1}^n,\boldsymbol{\xi}_k) = 0,\quad k=n,\ldots,r. \tag{7.87}
$$

Let us note that the flux matrix $\mathbb{C}_h$ given by (7.85) has the same block structure as the matrix $\mathbb{C}_h$ given by (6.124). The sequence of nonlinear algebraic systems (7.87) can be solved by the damped Newton-like iterative process (6.127) – (6.128) treated in Section 6.4.4.

Concerning the initial guess $\boldsymbol{\xi}_k^0$ for the iterative process (6.127) – (6.128), we use either the value known from the previous time level given by (6.129), i.e, $\boldsymbol{\xi}_k^0 = \boldsymbol{\xi}_{k-1},\ k = 1, \dots, r$, or it is possible to apply a higher-order extrapolation from previous time levels given by (6.141).

Remark 7.7. *Similarly as in Remarks 6.15 and 6.19, if we carry out only one Newton iteration ($l = 0$) at each time level, put $\lambda^0 = 1$ and use the extrapolation* (6.141)*, then the implicit method* (7.74) *reduces to the* BDF-DG higher-order semi-implicit method *of the viscous compressible flow including the shock capturing, which can be formulated in the following way: We seek the finite sequence of functions $\{\boldsymbol{w}_h^k\}_{k=0}^r$ such that*

$$\boldsymbol{w}_h^k \in \boldsymbol{S}_{hp},\ k = 0, 1, \dots, r, \tag{7.88a}$$

$$\frac{1}{\tau_k}\left(\sum_{l=0}^{n} \alpha_{n,l}\boldsymbol{w}_h^{k-l}, \boldsymbol{\varphi}_h\right) + \boldsymbol{c}_h^L\left(\sum_{l=1}^{n} \beta_{n,l}\boldsymbol{w}_h^{k-l}, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h\right) = \tilde{\boldsymbol{c}}_h\left(\sum_{l=1}^{n} \beta_{n,l}\boldsymbol{w}_h^{k-l}, \boldsymbol{\varphi}_h\right)$$
$$\forall\, \boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp},\ k = n, \dots, r, \tag{7.88b}$$

$$\boldsymbol{w}_h^0 = \Pi_h \boldsymbol{w}^0, \tag{7.88c}$$

$$\boldsymbol{w}_h^l \in \boldsymbol{S}_{hp},\ l = 1, \dots, n-1,\ \textit{are determined by a suitable } q\textit{-step method}$$
$$\textit{with } q < n \ \textit{ or by an explicit Runge–Kutta method – cf. Section } \textbf{??}. \tag{7.88d}$$

Here $\Pi_h \boldsymbol{w}^0$ is the $\boldsymbol{S}_{hp}$-approximation of $\boldsymbol{w}^0$, $\alpha_{n,l},\ l = 0, \dots, n$, are the BDF coefficients and $\beta_{n,l},\ l = 0, \dots, n$, are the coefficients of the extrapolation (6.141)*. (See Tables 6.2, 6.3 and 6.4, 6.5, for $n = 1, 2, 3$.)*

Setting

$$\bar{\boldsymbol{w}}_h^k = \sum_{l=1}^{n} \beta_{n,l}\boldsymbol{w}_h^{k-l}, \quad \bar{\boldsymbol{\xi}}_k = \sum_{l=1}^{n} \beta_{n,l}\boldsymbol{\xi}_{k-l}, \tag{7.89}$$

problem (7.88) *is equivalent to the linear algebraic systems*

$$\boldsymbol{F}_h(\{\boldsymbol{\xi}_{k-l}\}_{l=1}^n; \boldsymbol{\xi}_k) = \mathbb{C}_h(\bar{\boldsymbol{\xi}}_k)\boldsymbol{\xi}_k - \boldsymbol{d}_h(\{\boldsymbol{\xi}_{k-l}\}_{l=1}^n, \boldsymbol{\xi}_k) = 0, \quad k = n, \dots, r. \tag{7.90}$$

Finally, because of our considerations in Chapter **??**, we introduce the notation

$$\hat{\boldsymbol{a}}_h(\bar{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) = \boldsymbol{a}_h^L(\bar{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) - \tilde{\boldsymbol{a}}_h(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_h), \tag{7.91}$$

$$\hat{\boldsymbol{J}}_h^\sigma(\bar{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) = \boldsymbol{J}_h^{\sigma,L}(\bar{\boldsymbol{w}}_h, \boldsymbol{w}_h, \boldsymbol{\varphi}_h) - \tilde{\boldsymbol{J}}_h^\sigma(\bar{\boldsymbol{w}}_h, \boldsymbol{\varphi}_h), \tag{7.92}$$

for the viscous and penalty forms. Then (7.88b), can be replaced by the identity

$$\frac{1}{\tau_k}\left(\sum_{l=0}^{n} \alpha_{n,l}\boldsymbol{w}_h^{k-l}, \boldsymbol{\varphi}_h\right) + \hat{\boldsymbol{b}}_h(\bar{\boldsymbol{w}}_h^k, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h) + \hat{\boldsymbol{a}}_h(\bar{\boldsymbol{w}}_h^k, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h) + \hat{\boldsymbol{J}}_h^\sigma(\bar{\boldsymbol{w}}_h^k, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h) \tag{7.93}$$
$$+ \boldsymbol{\beta}_h(\bar{\boldsymbol{w}}_h^k, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h) + \boldsymbol{\gamma}_h(\bar{\boldsymbol{w}}_h^k, \boldsymbol{w}_h^k, \boldsymbol{\varphi}_h) = 0,\ \forall\, \boldsymbol{\varphi}_h \in \boldsymbol{S}_{hp},\ k = n, \dots, r,$$

where $\hat{\boldsymbol{b}}_h$ is given by (6.131) and $\bar{\boldsymbol{w}}_h^k$ is defined in (7.89).

7.4 Numerical examples

This section is devoted to applications of the presented BDF-DG schemes to the numerical solution of several test problems for the compressible Navier–Stokes equations. First, we consider a low Mach number flow past an adiabatic flat plate, where the analytical solution of incompressible flow is known. This example shows that the developed method is sufficiently accurate and stable even for compressible flow at an incompressible limit. Further, we present several flow regimes around the NACA 0012 profile, demonstrate the high accuracy of the DG discretization and mention some possible problems in the simulation of unsteady flows with the aid of implicit time discretization. Finally, we present a simulation of the viscous shock-vortex interaction by high-order methods. For the steady-state problems, the backward Euler method is used for the time discretization.

7.4.1 Blasius problem

The so-called Blasius problem represents the well-known test case, when a low-speed laminar flow along an adiabatic flat plate is considered. In this case the exact analytical solution is known for incompressible flow, see [Bla08]. Since the flow speed is

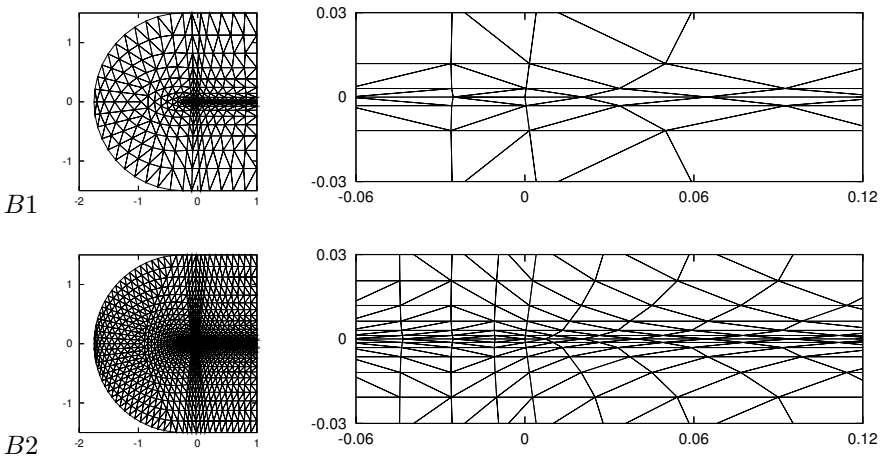

Figure 7.1: Blasius problem: computational grids – $B1$ with 662 elements (top) and $B2$ with 2648 elements (bottom), the whole computational domain (left) and their details around the leading edge (right).

low, similarly as in Section 6.7.2, we compare the compressible numerical solution with the exact solution of the corresponding incompressible flow.

We consider the laminar flow past the adiabatic flat plate $\{(x_1, x_2);\ 0 \le x_1 \le 1,\ x_2 = 0\}$ characterized by the freestream Mach number $M = 0.1$ and the Reynolds number $\mathrm{Re} = 10^4$. The computational domain is shown in Figure 7.1, where two used triangular grids are plotted together with their details around the leading edge. We prescribe the adiabatic boundary conditions (7.37) on the flat plate, the outflow boundary conditions (7.36) at $\{(x_1, x_2);\ x_1 = 1,\ -1.5 \le x_2 \le 1.5\}$ and the inflow boundary conditions (7.35) on the rest of the boundary.

We seek the steady-state solution by the time stabilization approach, in which the computational process is carried out for "$t \to \infty$". As a stopping criterion we use condition (6.171) (adapted to the viscous flow problem) with $\mathrm{TOL} = 10^{-6}$.

In the following, we investigate two items:

- the *stability of the method*, namely the influence of the value of the constant C_W in (7.67) on the convergence of the numerical scheme to the stationary solution,
- the *accuracy of the method*, namely the comparison of the numerical solutions with the exact solution of the incompressible flow.

Exercise 7.8. *Modify the stop criterion* (6.171) *to the viscous flow problem.*

Stability of the method

We compare the NIPG, IIPG, SIPG variants of the DGM using piecewise linear, quadratic and cubic space approximations. Our aim is to find a suitable value of the constant C_W in (7.67), which ensures the stability of the method and the convergence to the steady-state solution. First, we carried out computations for the values $C_W = 1,\ 5,\ 25,\ 125,\ 625,\ 3\,125$ and consequently, several additional values of C_W were chosen in order to find the limit value of C_W. These results obtained on the grid $B1$ are shown in Table 7.1, where an indication of the convergence of the appropriate variant of the DGM with a given value C_W is marked, namely,

- "convergence" (C): the stopping condition (6.171) was achieved after less than 200 time steps,
- "quasiconvergence" (qC): the stopping condition (6.171) was achieved after more than 200 time steps,
- "no-convergence" (NC): the stopping condition (6.171) was not achieved after 500 time steps.

The "quasiconvergence" in fact means that the appropriate value C_W is just under the limit value ensuring the convergence to the steady-state solution.

From Table 7.1 we can find that

- NIPG variant converges for any $C_W \ge 1$ independently of the degree of polynomial approximation,
- IIPG variant requires higher values of C_W for P_2 and P_3 approximations, namely $C_W = 5$ and $C_W = 10$ are sufficient, respectively. On the other hand, P_1 approximation converges for any $C_W \ge 1$.

	NIPG			IIPG			SIPG		
C_W	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3
1	C	C	C	C	NC	NC	NC	NC	NC
5	C	C	C	C	C	NC	NC	NC	NC
10	-	-	-	-	C	C	-	-	-
25	C	C	C	C	C	C	NC	NC	NC
100	-	-	-	-	-	-	NC	-	-
125	C	C	C	C	C	C	C	NC	NC
150	-	-	-	-	-	-	C	-	-
250	-	-	-	-	-	-	-	NC	-
300	-	-	-	-	-	-	-	qC	-
400	-	-	-	-	-	-	-	C	NC
500	-	-	-	-	-	-	-	C	NC
625	C	C	C	C	C	C	C	C	qC
1 000	-	-	-	-	-	-	-	-	C
3 125	C	C	C	C	C	C	C	C	C

Table 7.1: Blasius problem: the convergence (C), non-convergence (NC) or quasiconvergence (qC) of the NIPG, IIPG and SIPG variants of the DGM for P_1, P_2 and P_3 approximations for different values of C_W (symbol "-" means that the corresponding case was not tested).

- SIPG variant requires significantly higher values of C_W. We observe that $C_W \geq 125$ for P_1, $C_W \geq 400$ for P_1 and $C_W \geq 1\,000$ for P_3. This is in a good agreement with theoretical results from [HRS05] carried out for a scalar quasilinear elliptic problem, where the dependence $C_W = cp^2$ with a constant $c > 0$ is derived (p denotes the degree of the polynomial approximation).

Figure 7.2 shows the convergence history to the steady-state solution (i.e., the dependence of the steady-state residuum defined as in (6.170) on the number of time steps) for some interesting cases from Table 7.1.

Accuracy of the method

In order to analyze the accuracy of the method at incompressible limit, we compare the numerical solution of the Blasius problem for viscous compressible flow with its incompressible analytical solution. To this end, we introduce the dimensionless velocities in the streamwise direction and in the direction orthogonal to the stream by

$$v_1^\star := \frac{v_1(\eta)}{|\boldsymbol{v}_\infty|} \qquad \text{and} \qquad v_2^\star := \sqrt{\mathrm{Re}_x}\,\frac{v_2(\eta)}{|\boldsymbol{v}_\infty|}, \tag{7.94}$$

respectively, where

$$\eta := \sqrt{\mathrm{Re}_x}\,\frac{x_2}{x_1}, \qquad \mathrm{Re}_x := |\boldsymbol{v}_\infty|\,\mathrm{Re}\;\; x_1, \tag{7.95}$$

Re is the Reynolds number and $\boldsymbol{v}_\infty$ is the freestream velocity.

Figures 7.3 – 7.6 show the velocity profiles $v_1^\star$ and $v_2^\star$ obtained by P_1, P_2 and P_3 approximations on the meshes $B1$ and $B2$ at $x_1 = 0.1$, $x_1 = 0.3$ and $x_1 = 0.5$ in comparison with the exact solution. We present here results obtained by the NIPG method with $C_W = 25$. (The difference between the results obtained by the NIPG, SIPG and IIPG variants are negligible.) We observe a very accurate capturing of the $v_1^\star$-profile and a reasonable capturing of the $v_2^\star$-profile. An increase of accuracy for an increasing degree of approximation and a decreasing mesh size is evident.

Moreover, Figure 7.7 shows the comparison of the skin friction coefficient c_f computed by P_1, P_2 and P_3 approximations on the meshes $B1$ and $B2$ with the exact solution given by the Blasius formula. The *skin friction coefficient* is defined by

$$c_f = \frac{2\boldsymbol{t}\cdot(T^{\mathrm{V}}\boldsymbol{n})}{\rho_\infty|\boldsymbol{v}_\infty|^2 L_{\mathrm{ref}}}, \tag{7.96}$$

where ρ_∞ and $\boldsymbol{v}_\infty$ are the freestream density and velocity, respectively, L_{ref} is the reference length, $\boldsymbol{n}$ and $\boldsymbol{t}$ are the unit normal and tangential vectors to the wall and $T^{\mathrm{V}} = (\tau_{ij}^{\mathrm{V}})_{i,j=1}^2$ is the viscous part of the stress tensor. (The components τ_{ij}^{V} are defined in (7.8)).

We observe good agreement with the Blasius solution. The P_2 and P_3 approximations give the same value of c_f at the first element on the flat plate. Similar results were obtained in [BR97a, Fig. 2], where the improvement of the quality of

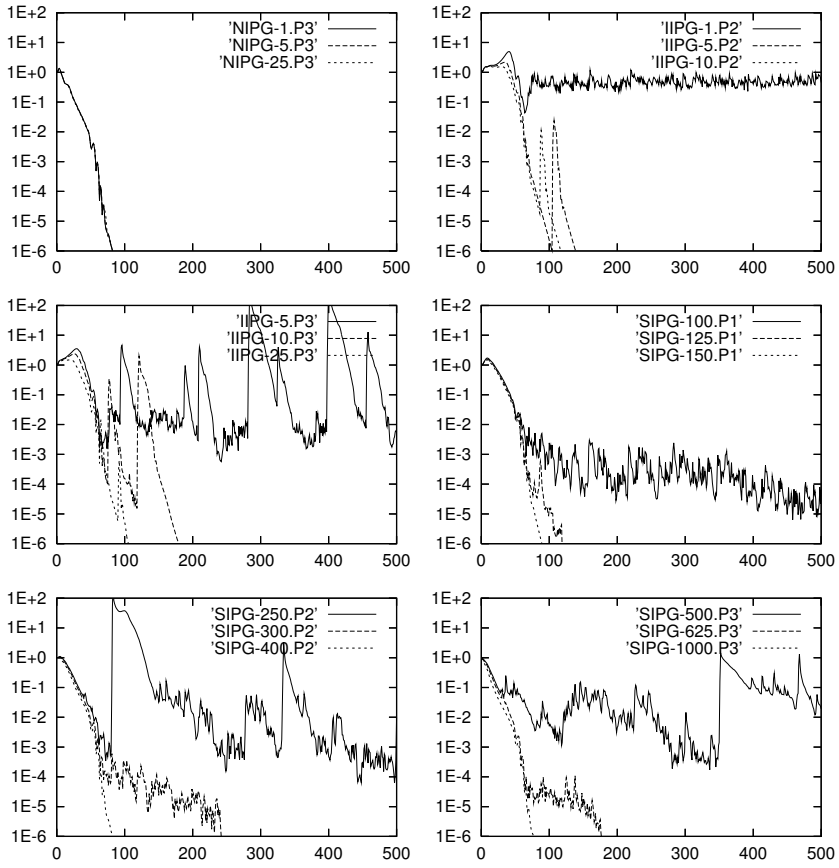

Figure 7.2: Blasius problem: the convergence of the steady-state residuum (6.170) in the logarithmic scale on the number of time steps for some computations from Table 7.1, (e.g., 'NIPG-625.P3' means the NIPG variant of the DGM with $C_W = 625$ and P_3 approximation).

the approximate solution on the first cell of the flat plate obtained by increasing the polynomial degree $p = 1, 2, 3$ is almost negligible. It is caused by the singularity in the solution at the leading edge of the flat plate at the point $(x_1, x_2) = (0, 0)$, which causes the decrease of the local order of accuracy of the DG method. This phenomenon was numerically verified also for a scalar nonlinear equation in Chapter 1.

7.4.2 Stationary flow around the NACA 0012 profile

We consider laminar steady-state viscous subsonic flow around the NACA 0012 profile for three different flow regimes characterized by the far-field Mach number M_∞, angle of attack α and the Reynolds number Re:

$$\begin{array}{lllll} \text{(C1)} & M_\infty = 0.50, & \alpha = 2^\circ, & \text{Re} = 500, \\ \text{(C2)} & M_\infty = 0.50, & \alpha = 2^\circ, & \text{Re} = 2\,000, \\ \text{(C3)} & M_\infty = 0.85, & \alpha = 2^\circ, & \text{Re} = 2\,000. \end{array}$$

We carried out computations on four triangular grids $N1 - N4$. Figure 7.8 shows these grids around the NACA 0012 profile and their zooms around the trailing and leading edges.

We evaluate the *aerodynamic coefficients drag* (c_D), *lift* (c_L) and *moment* (c_M). The coefficients c_D and c_L are defined as the first and the second components of the vector

$$\frac{1}{\frac{1}{2}\rho_\infty |\boldsymbol{v}_\infty|^2 L_{\text{ref}}} \int_{\Gamma_{\text{prof}}} (\text{p}\mathbb{I} - T^{\text{V}})\boldsymbol{n}\, \text{d}S, \tag{7.97}$$

where ρ_∞ and $\boldsymbol{v}_\infty$ are the far-field density and velocity, respectively, L_{ref} is the reference length, Γ_{prof} is the profile, p is the pressure, $\mathbb{I}$ is the identity matrix and T^{V} is the viscous part of the stress tensor given by (7.8). Moreover, c_M is given by

$$\frac{1}{\frac{1}{2}\rho_\infty |\boldsymbol{v}_\infty|^2 L_{\text{ref}}^2} \int_{\Gamma_{\text{prof}}} (x - x_{\text{ref}}) \times \big((\text{p}\mathbb{I} - T^{\text{V}})\boldsymbol{n}\big)\, \text{d}S, \tag{7.98}$$

where $x_{\text{ref}} = (\frac{1}{4}, 0)$ is the moment reference point. We use the notation $x \times y = x_1 y_2 - x_2 y_1$ for $x = (x_1, x_2), y = (y_1, y_2) \in \mathbb{R}^2$.

For each flow regime C1, C2 and C3, we carried out computations with polynomial approximation P_p, $p = 1, 3, 5$, on grids $N1 - N4$. We apply the stopping criterion (6.174) with tolerance $\text{tol} = 10^{-4}$.

Tables 7.2 and 7.4 show the values of the corresponding drag, lift and moment coefficients for each computation. These tables show also the number N_h of elements of each mesh and corresponding number of degrees of freedom N_{hp}. We observe that the high degree polynomial approximation gives a sufficiently accurate solution even on coarse grids. On the other hand, P_1 polynomial approximation is not sufficiently accurate even for the finest mesh.

Further, Figures 7.9 – 7.14 show Mach number isolines and the distribution of the skin friction coefficient (7.96) obtained for each flow regime on the meshes $N1$ and $N4$.

The presented numerical results of examples C1, C2 and C3 show that the high-order DG method is suitable for the numerical solution of the compressible viscous flow. With the aid of the P_5 polynomial approximation we obtain the aerodynamic coefficients with sufficient accuracy even on the coarsest grid.

Finally, we demonstrate the stability of the time discretization schemes with respect to the size of the time steps. According to (6.150), we define the value

$$\text{CFL}_k = \frac{\tau_k}{\min_{K \in \mathcal{T}_h} \left(|K|^{-1} \max_{\Gamma \in \partial K} \varrho(\mathbb{P}(\boldsymbol{w}_h^k|_\Gamma))|\Gamma| \right)}, \quad k = 0, 1, \ldots, r, \tag{7.99}$$

which measures how many times the time step is larger in comparison to the time step for an explicit time discretization. Here $\varrho(\mathbb{P}(\boldsymbol{w}_h^k|_\Gamma))$ denotes the spectral radius of the matrix $\mathbb{P}(\boldsymbol{w}_h^k|_\Gamma)$ defined by (7.23). Figure 7.15 shows the dependence of CFL_k on the parameter k for the flow regime C1, C2 and C3 using P_1 polynomial approximation on grid $N4$. We observe that very large values CFL_k are attained, and hence the BDF-DGFE method is practically unconditionally stable.

7.4.3 Unsteady flow

We consider a transonic flow around the NACA 0012 profile with the far-field Mach number $M_\infty = 0.85$, angle of attack $\alpha = 0^\circ$ and the Reynolds number $\text{Re} = 10\,000$. In this case the flow is unsteady with a periodic propagation of vortices behind the profile, see [Mit98].

In the numerical simulation of nonstationary processes, it is necessary to use a sufficiently small time step in order to guarantee accuracy with respect to time. In our computations the time step was chosen adaptively with the aid of the adaptive algorithm presented in Section 6.4.6 with the tolerance $\omega = 10^{-2}$ in (6.148).

We applied the 3-step BDF-DGM with the P_2 polynomial approximation on the mesh from Figure 7.16. The computation was carried out for the dimensionless time $t \in (0, 90)$. Figure 7.17 shows the dependence of the lift, drag and moment coefficients

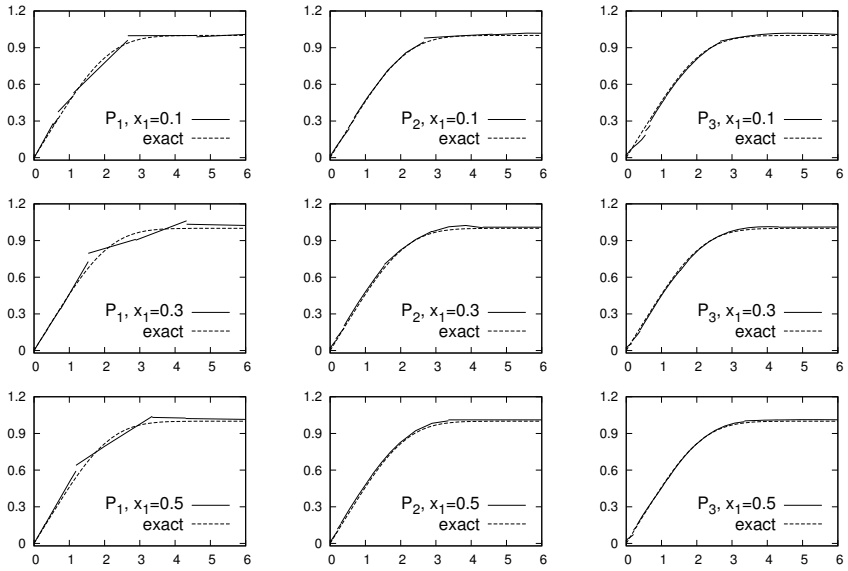

Figure 7.3: Blasius problem: mesh $B1$, velocity profiles $v_1^\star = v_1^\star(\eta)$ for P_1, P_2 and P_3 approximations at $x_1 = 0.1$, $x_1 = 0.3$ and $x_1 = 0.5$ in comparison with the exact solution.

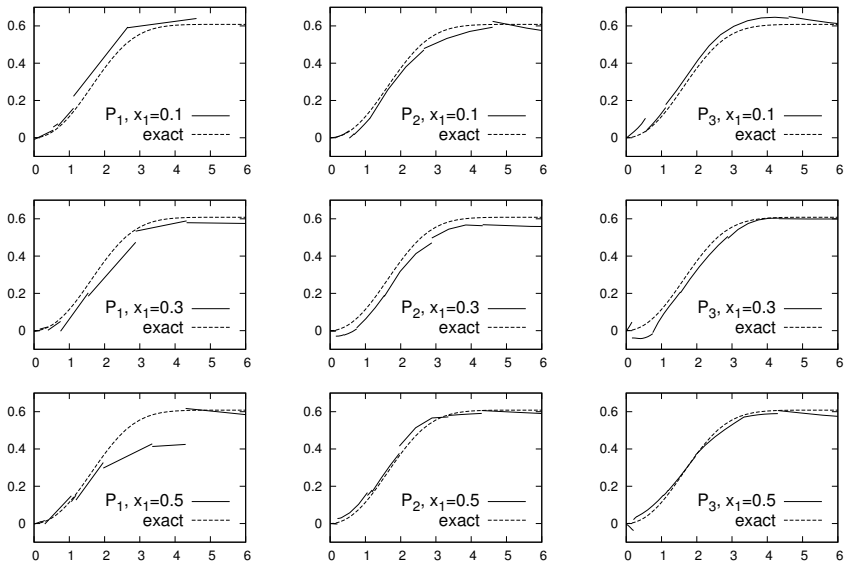

Figure 7.4: Blasius problem: mesh $B1$, velocity profiles $v_2^\star = v_2^\star(\eta)$ for P_1, P_2 and P_3 approximations at $x_1 = 0.1$, $x_1 = 0.3$ and $x_1 = 0.5$ in comparison with the exact solution.

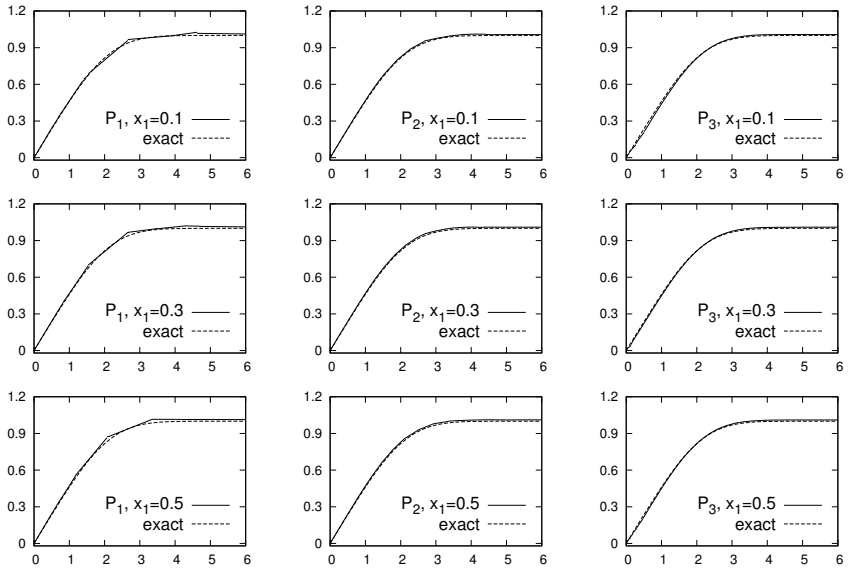

Figure 7.5: Blasius problem: mesh $B2$, velocity profiles $v_1^\star = v_1^\star(\eta)$ for P_1, P_2 and P_3 approximations at $x_1 = 0.1$, $x_1 = 0.3$ and $x_1 = 0.5$ in comparison with the exact solution.

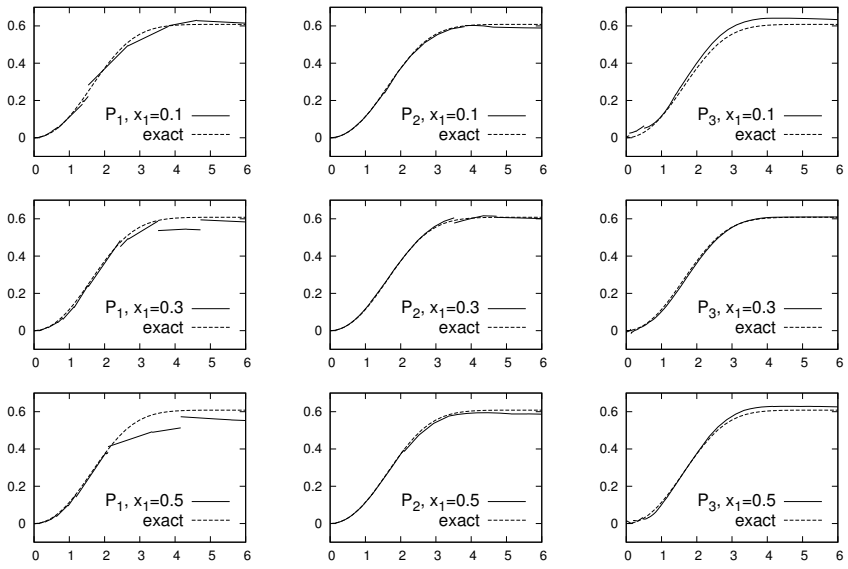

Figure 7.6: Blasius problem: mesh $B2$, velocity profiles $v_2^\star = v_2^\star(\eta)$ for P_1, P_2 and P_3 approximations at $x_1 = 0.1$, $x_1 = 0.3$ and $x_1 = 0.5$ in comparison with the exact solution.

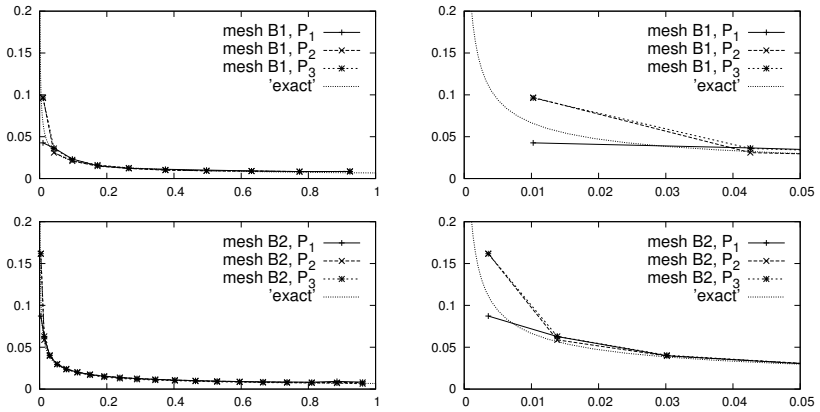

Figure 7.7: Blasius problem: skin friction coefficient computed on the meshes $B1$ (top) and $B2$ (bottom) by P_1, P_2 and P_3 approximation in comparison with the Blasius formula (exact), distributions along the whole plate (left), their details around $x_1 = 0$ (right).

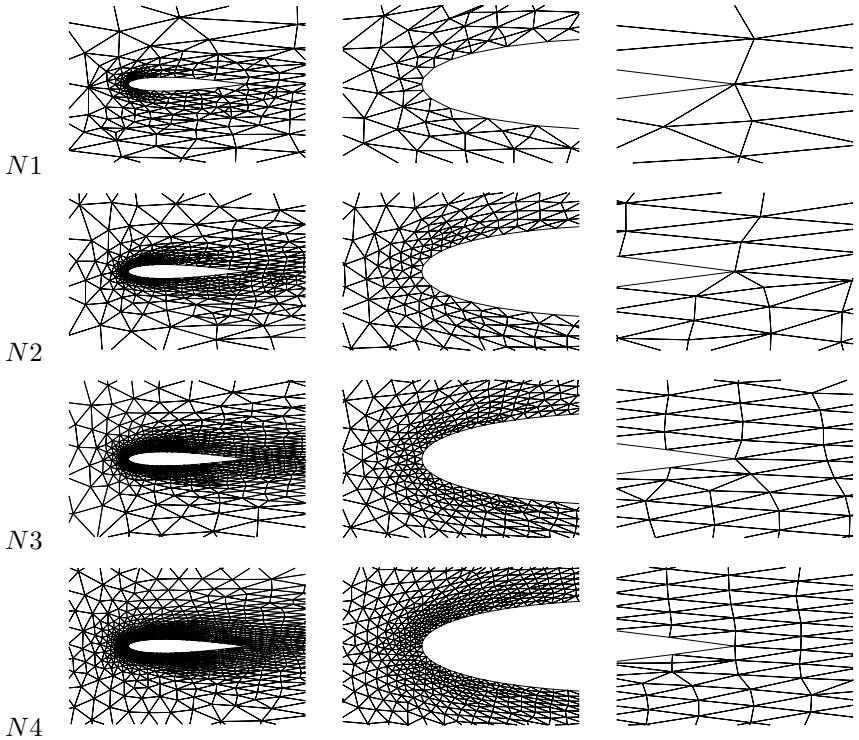

Figure 7.8: Computational grids $N1 - N4$ around the NACA 0012 profile (left) with details around the leading (middle) and trailing edges (right) used for steady-state examples.

p	N_h	N_{hp}	c_D	c_L	c_M
1	782	9384	1.7416E-01	1.0260E-01	-3.3278E-03
1	1442	17304	1.7632E-01	1.1225E-01	-2.8440E-03
1	2350	28200	1.7767E-01	1.1291E-01	-2.8089E-03
1	3681	44172	1.7775E-01	1.1338E-01	-2.8734E-03
3	782	31280	1.8086E-01	1.1283E-01	-3.1439E-03
3	1442	57680	1.8093E-01	1.1284E-01	-3.1186E-03
3	2350	94000	1.8080E-01	1.1322E-01	-3.0036E-03
3	3681	147240	1.8085E-01	1.1302E-01	-3.0590E-03
5	782	65688	1.8077E-01	1.1269E-01	-3.1054E-03
5	1442	121128	1.8085E-01	1.1299E-01	-3.0896E-03
5	2350	197400	1.8087E-01	1.1310E-01	-3.0601E-03
5	3681	309204	1.8088E-01	1.1304E-01	-3.0719E-03

Table 7.2: NACA 0012 ($M_\infty = 0.5$, $\alpha = 0°$, Re $= 500$): the values of the drag, lift and moment coefficient obtained by the BDF-DGM for P_p, $p = 1, 3, 5$, polynomial approximations on grids $N1 - N4$.

p	N_h	N_{hp}	c_D	c_L	c_M
1	782	9384	8.5405E-02	9.0263E-02	-6.7673E-03
1	1442	17304	8.5231E-02	8.2415E-02	-9.7498E-03
1	2350	28200	8.6387E-02	8.0999E-02	-1.0283E-02
1	3681	44172	8.6219E-02	8.2633E-02	-1.0149E-02
3	782	31280	8.7319E-02	8.5077E-02	-1.0116E-02
3	1442	57680	8.8193E-02	8.4048E-02	-1.0124E-02
3	2350	94000	8.8148E-02	8.4091E-02	-1.0079E-02
3	3681	147240	8.8264E-02	8.4082E-02	-1.0094E-02
5	782	65688	8.8124E-02	8.4008E-02	-1.0048E-02
5	1442	121128	8.8281E-02	8.4201E-02	-1.0091E-02
5	2350	197400	8.8283E-02	8.4290E-02	-1.0075E-02
5	3681	309204	8.8284E-02	8.4317E-02	-1.0068E-02

Table 7.3: NACA 0012 ($M_\infty = 0.5$, $\alpha = 0°$, Re $= 2\,000$): the values of the drag, lift and moment coefficient obtained by the BDF-DGM for P_p, $p = 1, 3, 5$, polynomial approximations on grids $N1 - N4$.

p	N_h	N_{hp}	c_D	c_L	c_M
1	782	9384	1.1610E-01	4.4091E-02	-1.4702E-02
1	1442	17304	1.1444E-01	3.8107E-02	-1.5934E-02
1	2350	28200	1.1605E-01	3.4837E-02	-1.6923E-02
1	3681	44172	1.1566E-01	3.3338E-02	-1.7027E-02
3	782	31280	1.1809E-01	3.1726E-02	-1.7463E-02
3	1442	57680	1.1892E-01	3.1212E-02	-1.7163E-02
3	2350	94000	1.1887E-01	3.0834E-02	-1.7164E-02
3	3681	147240	1.1898E-01	3.0918E-02	-1.7142E-02
5	782	65688	1.1885E-01	3.1034E-02	-1.7048E-02
5	1442	121128	1.1899E-01	3.1056E-02	-1.7128E-02
5	2350	197400	1.1899E-01	3.0971E-02	-1.7154E-02
5	3681	309204	1.1899E-01	3.0981E-02	-1.7148E-02

Table 7.4: NACA 0012 ($M_\infty = 0.85$, $\alpha = 0°$, Re $= 2\,000$): the values of the drag, lift and moment coefficient obtained by the BDF-DGM for P_p, $p = 1, 3, 5$, polynomial approximations on grids $N1 - N4$.

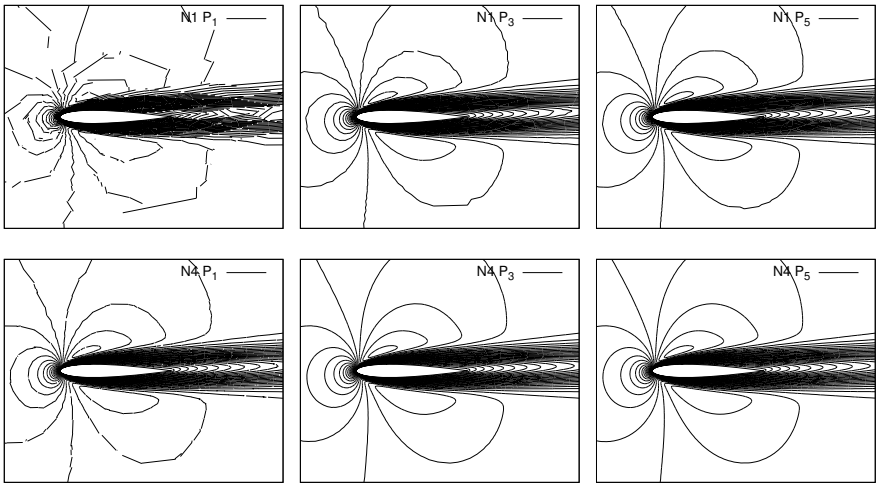

Figure 7.9: NACA 0012 ($M_\infty = 0.5$, $\alpha = 2°$, Re $= 500$): Mach number isolines for P_1, P_3 and P_5 polynomial approximations on grids $N1$ and $N4$.

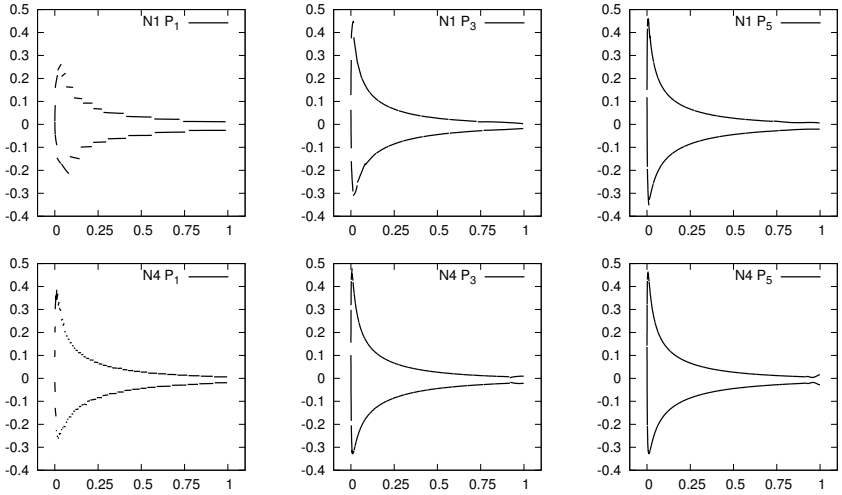

Figure 7.10: NACA 0012 ($M_\infty = 0.5$, $\alpha = 2°$, Re $= 500$): distribution of the skin friction coefficient for P_1, P_3 and P_5 polynomial approximations on grids $N1$ and $N4$.

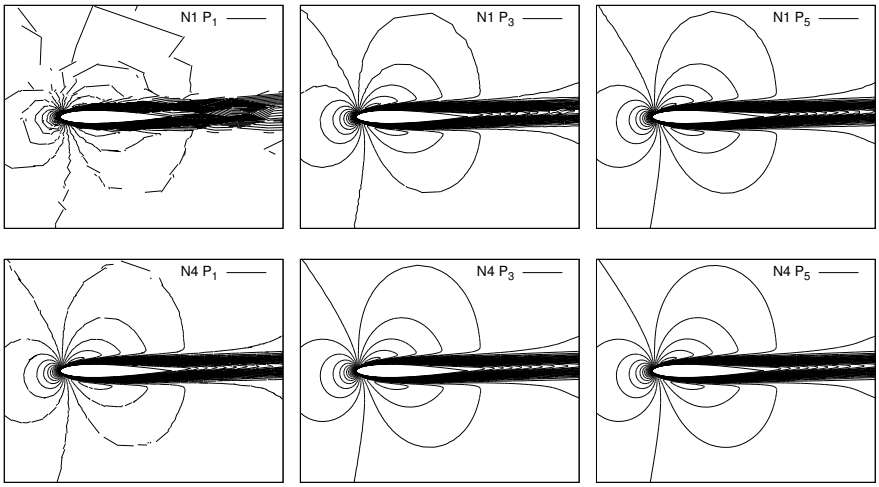

Figure 7.11: NACA 0012 ($M_\infty = 0.5$, $\alpha = 2°$, Re $= 2000$): Mach number isolines for P_1, P_3 and P_5 polynomial approximations on grids $N1$ and $N4$.

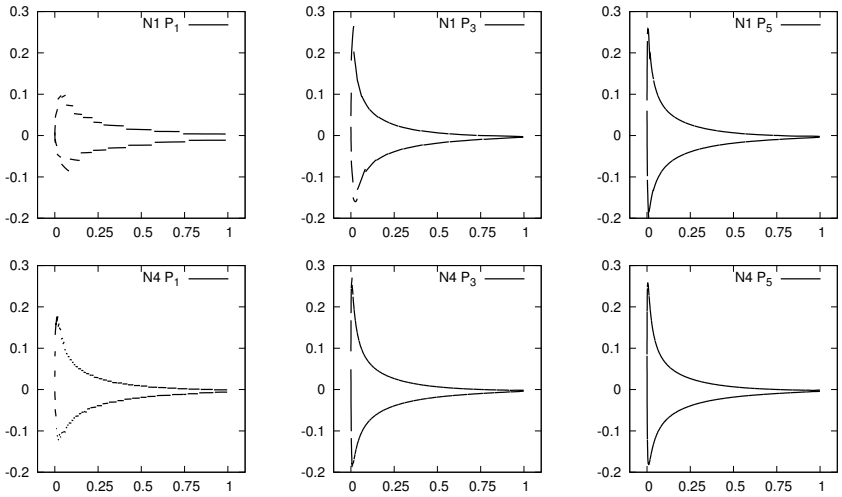

Figure 7.12: NACA 0012 ($M_\infty = 0.5$, $\alpha = 2°$, Re $= 2000$): distribution of the skin friction coefficient for P_1, P_3 and P_5 polynomial approximations on grids $N1$ and $N4$.

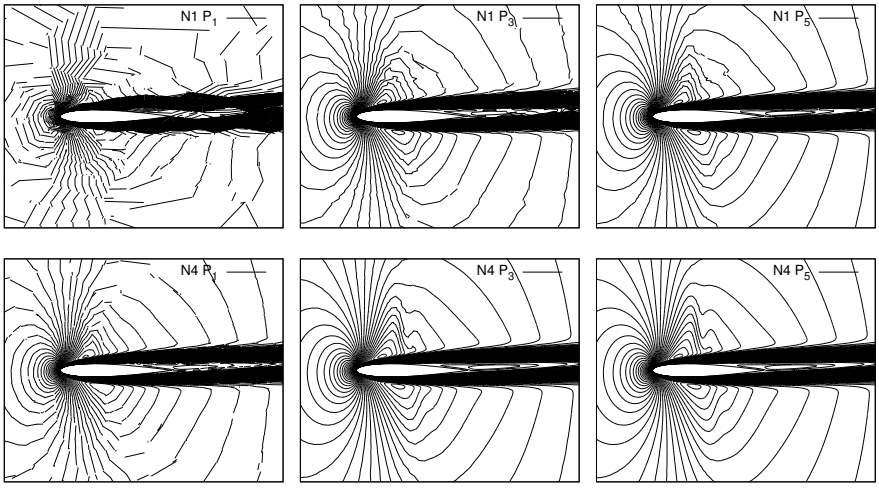

Figure 7.13: NACA 0012 ($M_\infty = 0.85$, $\alpha = 2°$, Re $= 2000$): Mach number isolines for P_1, P_3 and P_5 polynomial approximations on grids $N1$ and $N4$.

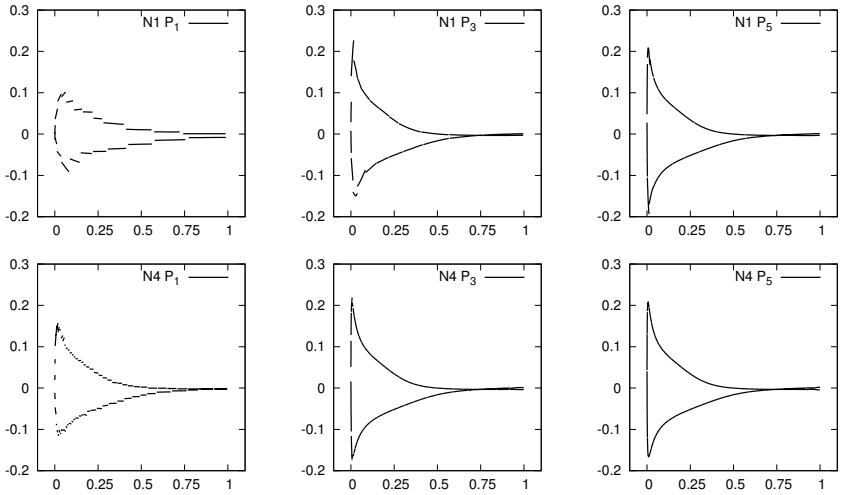

Figure 7.14: NACA 0012 ($M_\infty = 0.85$, $\alpha = 2°$, Re $= 2000$): distribution of the skin friction coefficient for P_1, P_3 and P_5 polynomial approximations on grids $N1$ and $N4$.

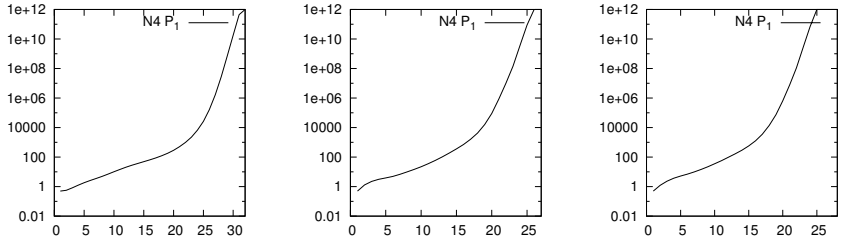

Figure 7.15: The dependence of the value CFL_k on the parameter k for the flow regimes C1 (left), C2 (center) and C3 (right).

on time $t \in (80, 90)$. We observe periodic oscillations of c_L and c_M with period $\Delta t \approx 0.7$. Figure 7.18 shows the Mach number isolines at time instants $t_i = 89.3 + i\Delta t/7,\ i = 1, 2, \dots 7$, demonstrating the periodic propagation of vortices behind the profile. These results are in a good agreement with results from [Dol08b] and [Mit98].

This example demonstrates that the presented BDF-DGM is able to resolve steady as well as unsteady flow without any modification of the scheme. It is very important in the case, when it is not a priori known, whether the considered flow is steady or unsteady.

7.4.4 Steady vs. unsteady flow

The numerical examples presented in the previous sections lead us to the conclusion that the presented BDF-DGM is robust with respect to the magnitude of the Mach number and is practically unconditionally stable. This means that large time steps can be used, cf. Figure 7.15. However, there is a danger that the use of too long time steps can lead to qualitatively different results.

As an example we consider a laminar viscous subsonic flow around the NACA 0012 profile with the far-field Mach number $M_\infty = 0.5$, angle of attack $\alpha = 2^\circ$ and the Reynolds number $\mathrm{Re} = 5\,000$. This flow is close to a limit between the steady and unsteady flow regimes. In [Dol08b] and [DHH11], we presented steady-state solutions for this flow regime computed using several degrees of polynomial approximation and several grids.

Here we present computations carried out by the 3-step BDF-DGM with P_3 and P_4 polynomial approximation, applied on an unstructured mesh shown in Figure 7.19. The time steps were chosen adaptively with the aid of the adaptive algorithm presented in Section 6.4.6 with two different tolerances $\omega = 1$ and $\omega = 10^{-4}$ in (6.148). This means that in the former case we do not take care of the accuracy with respect to time. In the latter case, the problem was solved with a high accuracy with respect to time. Of course, the computation needs much longer CPU time.

Figure 7.20 shows the convergence of the steady-state residuum (cf. the criterion (6.171) adapted to the viscous flow problem) and the corresponding value CFL_k (cf. (7.99)) for both settings $\omega = 1$ and $\omega = 10^{-4}$.

It can be seen that for $\omega = 1$ a steady-state solution is obtained. On the other hand, for $\omega = 10^{-4}$ the resolution in time is much more accurate and an unsteady solution is obtained. Moreover, Figure 7.21 shows the dependence of the lift coefficient c_L on the dimensionless time for P_3 and P_4 polynomial approximations with $\omega = 10^{-4}$ in (6.148). The constant value c_L-'steady' was obtained with the same method but with $\omega = 1$. Finally, Figure 7.22 shows Mach number isolines for P_3 and P_4 polynomial approximations and for $\omega = 1$ and $\omega = 10^{-4}$.

These experiments indicate that an insufficiently accurate resolution with respect to time can lead to different flow regimes (steady vs. unsteady). These results are in agreement with [KBD+10], where this example was solved by several research groups. They achieved mostly the steady state solution using steady-state solvers or implicit time discretizations with large time steps. Only a sufficiently accurate (explicit) time discretization (carried out at the University of Stuttgart) gave the unsteady flow regime, see [KBD+10, Chapter 5].

7.4.5 Viscous shock-vortex interaction

This example represents a challenging unsteady viscous flow simulation. Similarly as in [DT04], [Für01] and [TGS00], we consider the viscous interaction of a plane weak shock wave with a single isentropic vortex. During the interaction, acoustic waves are produced, and we investigate the ability of the numerical scheme to capture these waves. The computational domain is $\Omega = (0, 2) \times (0, 2)$ with the periodic extension in the x_2-direction. A stationary plane shock wave is located at $x_1 = 1$. The prescribed pressure jump through the shock is $\mathrm{p}_R - \mathrm{p}_L = 0.4$, where p_L and p_R are the pressure values from the left and right of the shock wave, respectively, corresponding to the inlet (left) Mach number $M_L = 1.1588$. The reference density and velocity are those of the free uniform flow at infinity. In particular, we define the initial density, x_1-component of velocity and pressure by

$$\rho_L = 1,\ u_L = M_L \gamma^{1/2},\ \mathrm{p}_L = 1, \quad \rho_R = \rho_L K_1,\ u_R = u_L K_1^{-1},\ \mathrm{p}_R = \mathrm{p}_1 K_2, \tag{7.100}$$

where

$$K_1 = \frac{\gamma+1}{2} \frac{M_L^2}{1 + \frac{\gamma-1}{2} M_L^2}, \quad K_2 = \frac{2}{\gamma+1} \left(\gamma M_L^2 - \frac{\gamma-1}{2} \right). \tag{7.101}$$

Here, the subscripts $_L$ and $_R$ denote the quantities at $x < 1$ and $x > 1$, respectively, $\gamma = 1.4$ is the Poisson constant. The Reynolds number is 2000. An isolated isentropic vortex centered at $(0.5, 1)$ is added to the basic flow. The angular velocity in the vortex is given by

$$v_\theta = c_1 r \exp(-c_2 r^2), \quad c_1 = u_c / r_c, \quad c_2 = r_c^{-2}/2, \quad r = ((x_1 - 0.5)^2 - (x_2 - 1)^2)^{1/2}, \tag{7.102}$$

where we set $r_c = 0.075$ and $u_c = 0.5$. Computations are stopped at the dimensionless time $T = 0.7$.

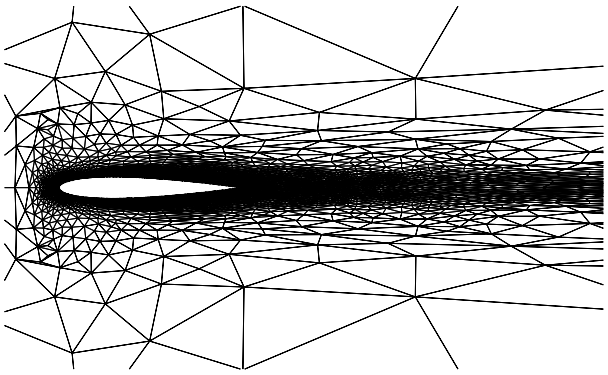

Figure 7.16: NACA 0012 , $M_\infty = 0.85$, $\alpha = 0°$ and $\mathrm{Re} = 10\,000$: triangular grid.

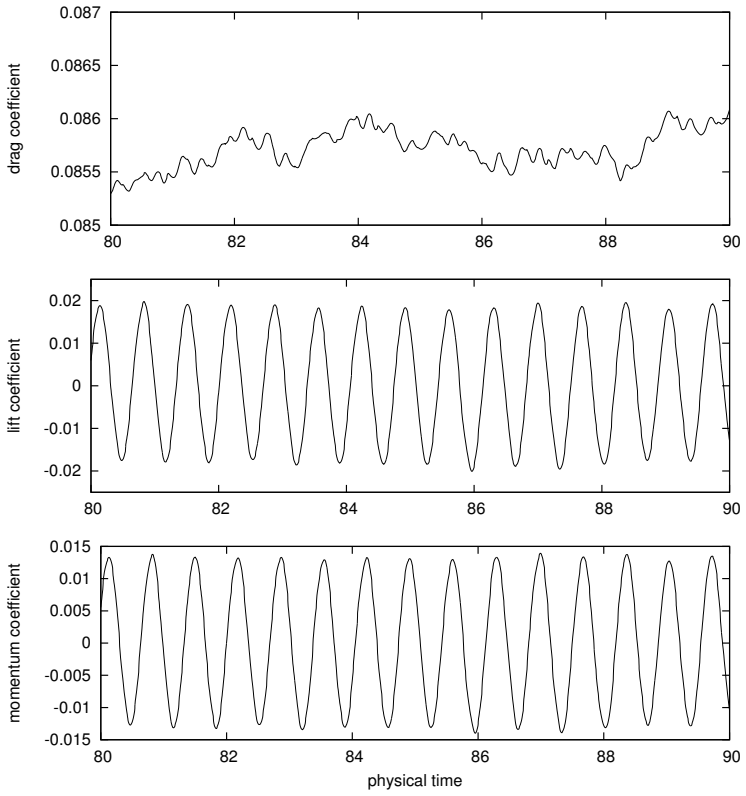

Figure 7.17: NACA 0012 , $M_\infty = 0.85$, $\alpha = 0°$ and $\mathrm{Re} = 10\,000$: dependence of the drag coefficient c_D, lift coefficient c_L and moment coefficient c_M on the dimensionless time $t \in (80, 90)$.

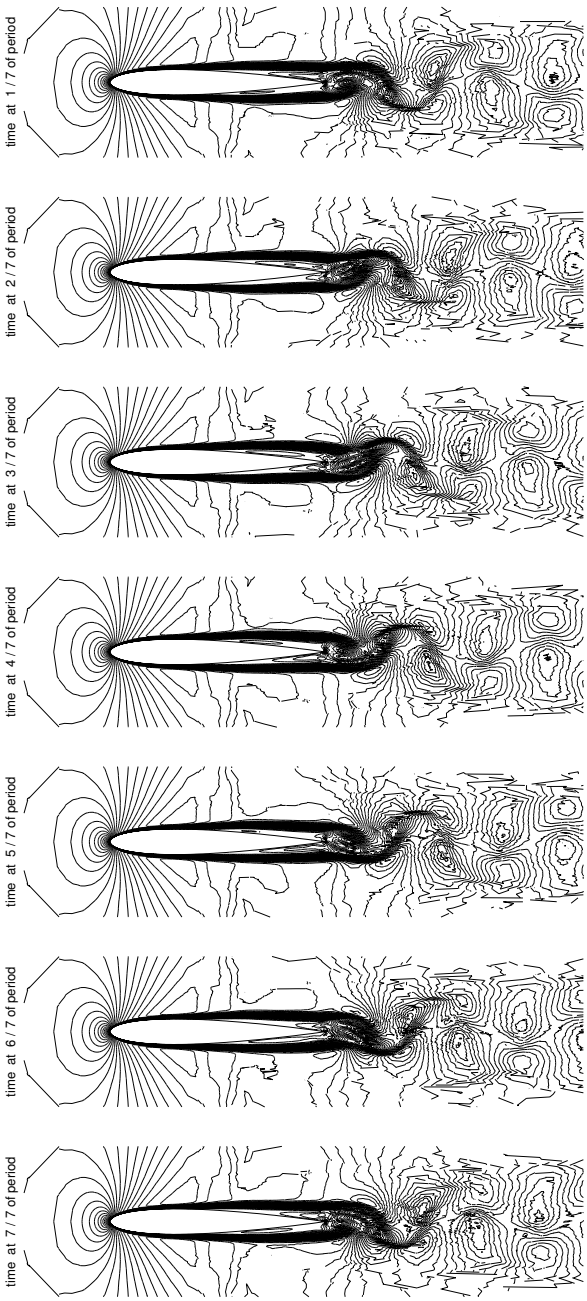

Figure 7.18: NACA 0012, $M_\infty = 0.85$, $\alpha = 0°$ and Re = 10 000: Mach number isolines at the time instants $t_i = 89.3 + i\Delta t/7$, $i = 1, \ldots, 7$, in one period.

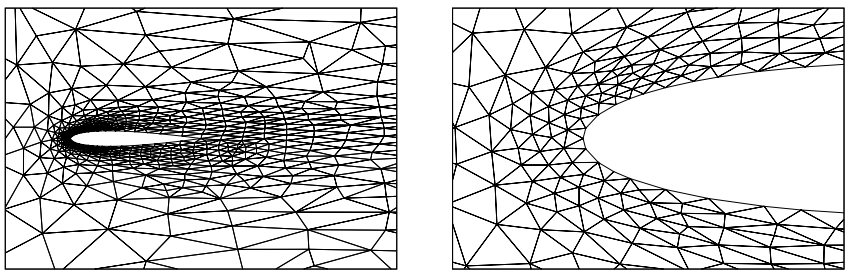

Figure 7.19: NACA 0012, $M_\infty = 0.5$, $\alpha = 0°$and Re = 5 000: computational grid, around the profile (left) and a detail at the leading edge (right).

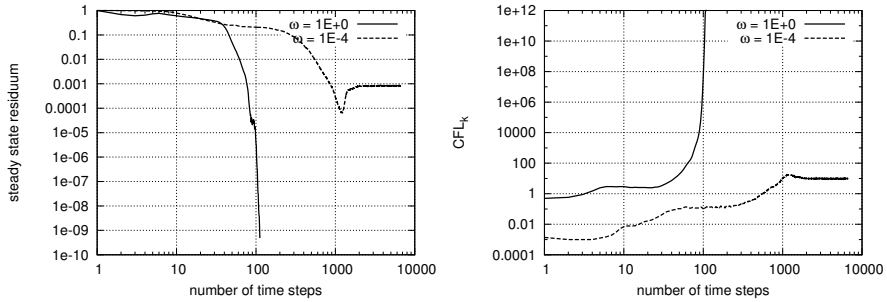

Figure 7.20: NACA 0012, $M_\infty = 0.5$, $\alpha = 0°$and $\mathrm{Re} = 5\,000$, P_4 approximation, $\omega = 1$ and $\omega = 10^{-4}$: steady-state residuum (left) and the value CFL_k (right) with respect to the number of time steps.

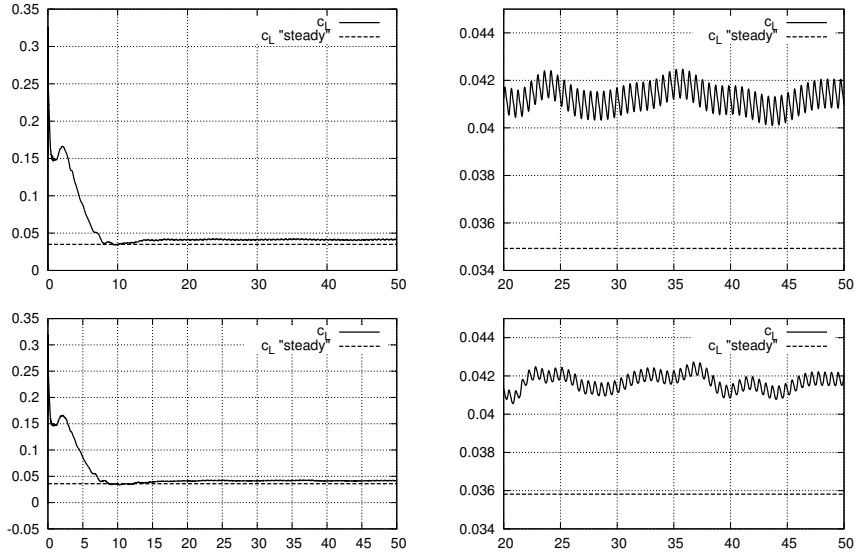

Figure 7.21: NACA 0012, $M_\infty = 0.5$, $\alpha = 0°$and $\mathrm{Re} = 5\,000$: P_3 (top) and P_4 (bottom) approximation, time evolution of the lift coefficient c_L with respect to the physical time for the setting $\omega = 10^{-4}$ (left) and its detail (right), the value c_L "steady" was obtain with $\omega = 1$.

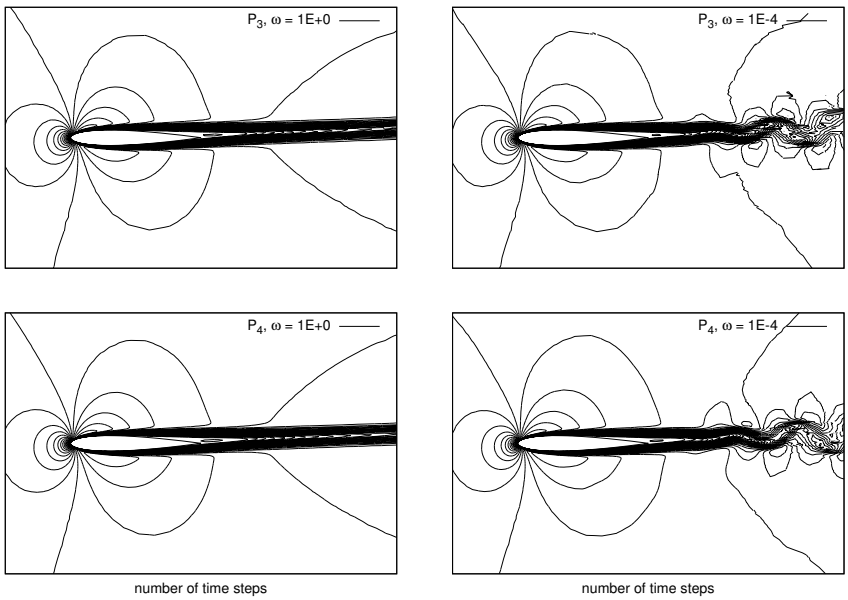

Figure 7.22: NACA 0012, $M_\infty = 0.5$, $\alpha = 0°$and $\mathrm{Re} = 5\,000$ for P_3 and P_4 polynomial approximations and for $\omega = 1$ and $\omega = 10^{-4}$: Mach number isolines.

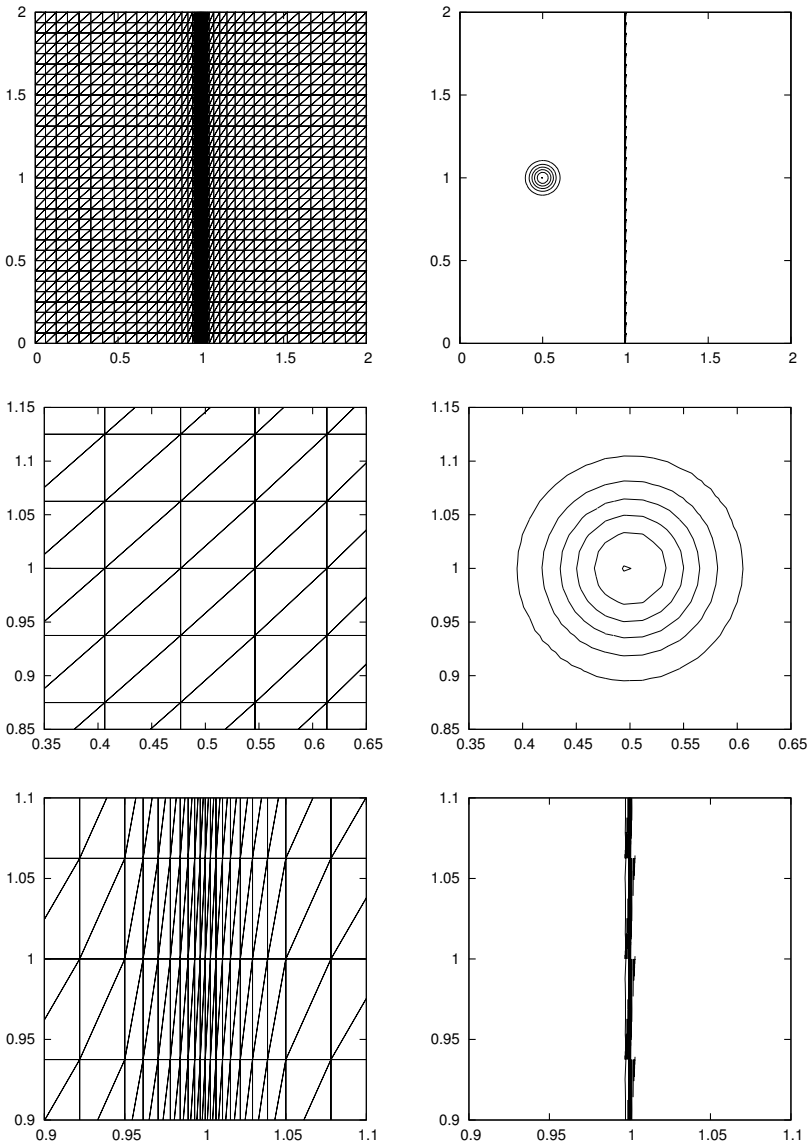

Figure 7.23: Viscous shock-vortex interaction: the used grid (left) and pressure isolines (right) at $t = 0$, the total view (top), its details near the vortex (center) and the shock wave (bottom).

We solved this problem with the aid of the 3-steps BDF-DGM (7.74) with P_4 polynomial approximation in space. The computational grid with 3 072 triangles was a priori refined in the vicinity of the stationary shock wave, see Figure 7.23. This figure shows also the initial setting of the shock wave and the isentropic vortex with their details.

Figures 7.24 and 7.25 show the results of the simulation of viscous shock-vortex interaction, namely, the isolines of the pressure and the pressure distribution along $x_2 = 1$ at several time instants. We observe a capturing of the shock-vortex interaction with the appearance of incident and reflected acoustic waves. These results are in agreement with results presented in [DT04], [Für01] and [TGS00]. Hence, we can conclude that the DGM is able to capture such complicated physical phenomena as shock-vortex interaction.

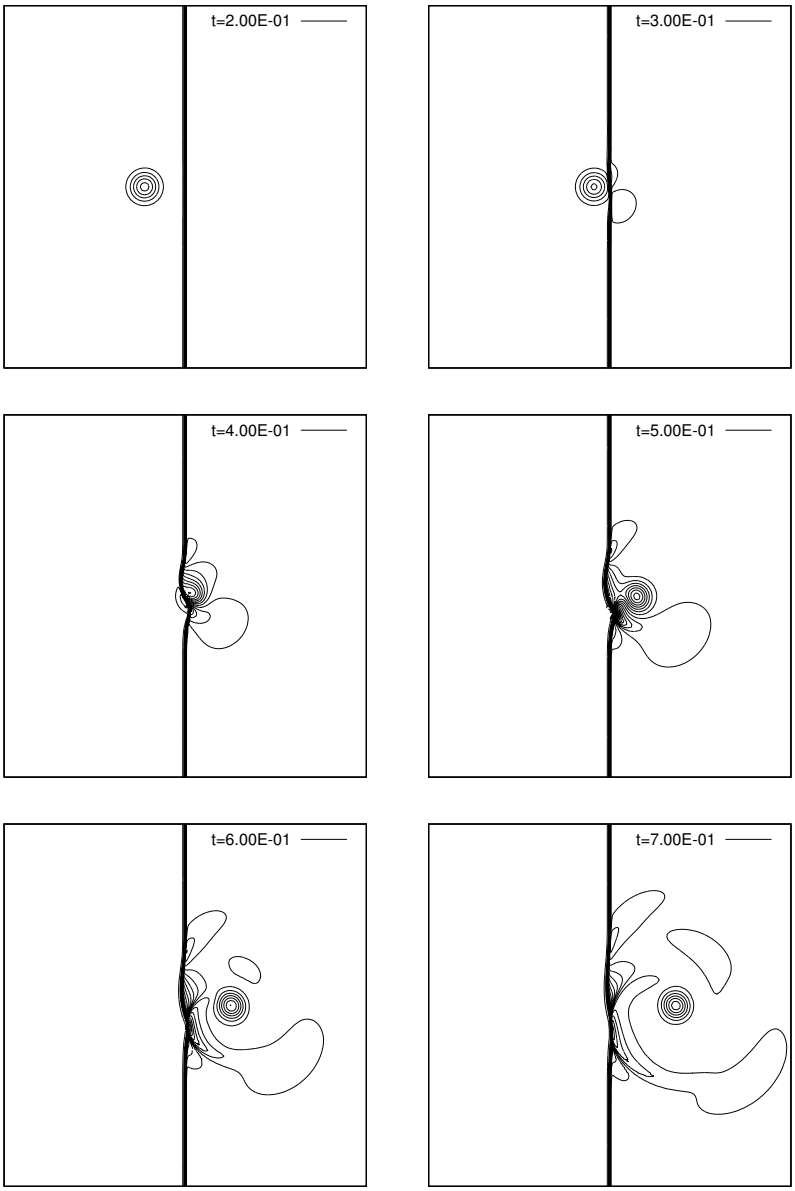

Figure 7.24: Viscous shock-vortex interaction: pressure isolines at $t = 0.2$, 0.3, 0.4, 0.5, 0.6 and 0.7.

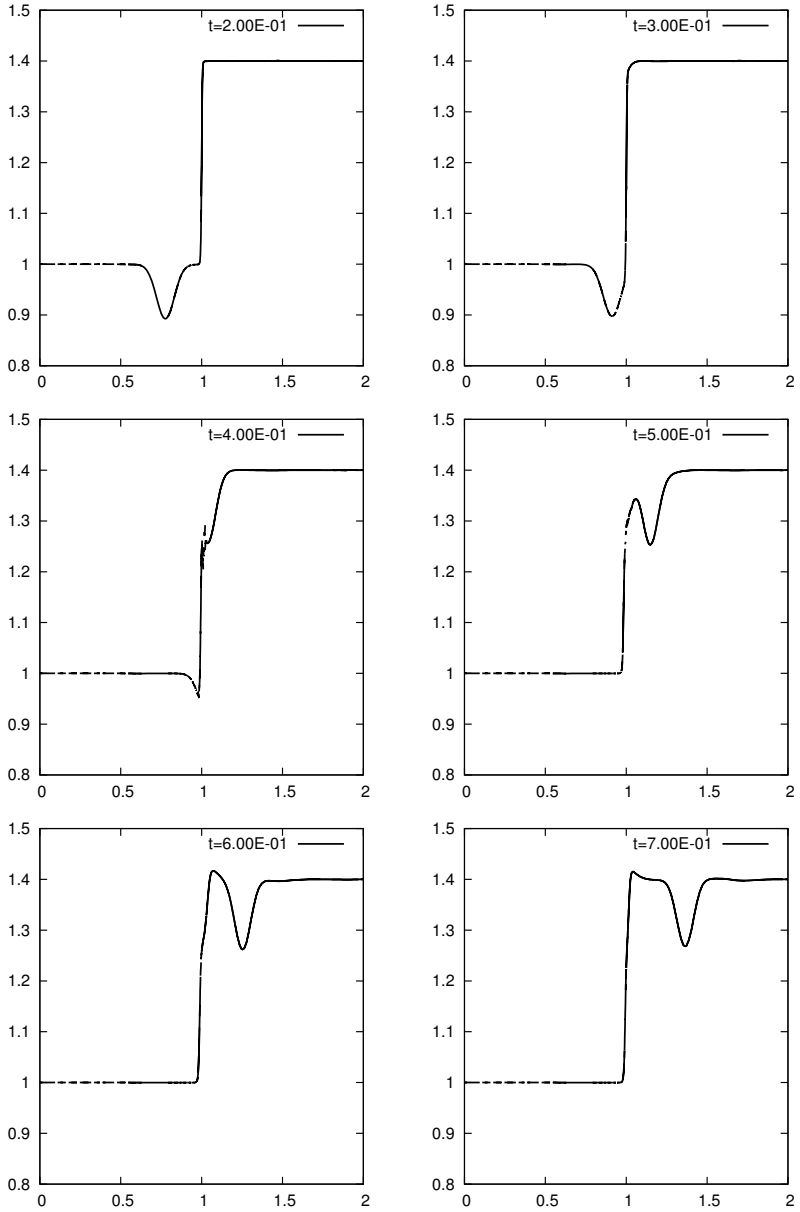

Figure 7.25: Viscous shock-vortex interaction: the pressure distribution along the line $x_2 = 1$ at $t = 0.2$, 0.3, 0.4, 0.5, 0.6 and 0.7.

Bibliography

[AEV11] L. El Alaoui, A. Ern, and M. Vohralík. Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems. *Comput. Methods Appl. Mech. Engrg*, 200:2782–2795, 2011.

[AF03] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*. Academic Press, 2003.

[Arn82] D. N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19(4):742–760, 1982.

[BBHN09] F. Bassi, C. De Bartolo, R. Hartmann, and A. Nigro. A discontinuous Galerkin method for inviscid low Mach number flows. *J. Comput. Phys.*, 228:3996–4011, 2009.

[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.

[Bla08] H. Blasius. Grenzschichten in flüssigkeiten mit kleiner reibung. *Z. Math. Phys.*, 56:1–37, 1908.

[BLN79] C. Bardos, A.-Y. Le Roux, and J.-C. Nedelec. First order quasilinear equations with boundary conditions. *Commun. Partial. Differ. Equations*, 4:1017–1034, 1979.

[BO99] C. E. Baumann and J. T. Oden. A discontinuous *hp* finite element method for the Euler and Navier-Stokes equations. *Int. J. Numer. Methods Fluids*, 31:79–95, 1999.

[BR97a] 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.

[BR97b] F. Bassi and S. Rebay. High-order accurate discontinuous finite element solution of the 2D Euler equations. *J. Comput. Phys.*, 138:251–285, 1997.

[BR00] F. Bassi and S. Rebay. A high order discontinuous Galerkin method for compressible turbulent flow. In B. Cockburn, G. E. Karniadakis, and C.-W. Shu, editors, *Discontinuous Galerkin Method: Theory, Computations and Applications*, Lecture Notes in Computational Science and Engineering 11, pages 113–123. Springer-Verlag, 2000.

[Bre03] S. C. Brenner. Poincare-Friedrichs inequalities for piecewise H-1 functions. *SIAM Journal on Numerical Analysis*, 41(1):306–324, 2003.

[BS87] I. Babuška and M. Suri. The *hp*-version of the finite element method with quasiuniform meshes. *M^2AN Math. Model. Numer. Anal.*, 21:199–238, 1987.

[BS90] I. Babuška and M. Suri. The *p*- and *hp*- versions of the finite element method. An overview. *Comput. Methods Appl. Mech. Eng.*, 80:5–26, 1990.

[BS94a] I. Babuška and M. Suri. The *p*- and *hp*-FEM a survey. *SIAM Review*, 36:578–632, 1994.

[BS94b] S. Brenner and R. L. Scott. *The Mathematical Theory of Finite Element Methods*. Spriger, New York, 1994.

[BS01] I. Babuška and T. Strouboulis. *The Finite Element Method and its Reliability*. Clarendon Press, Oxford, 2001.

[BT80] A. Bayliss and E. Turkel. Radiation boundary conditions for wave-like equations. *Commun. Pure Appl. Math.*, 33:708–725, 1980.

[BW76] R. M. Beam and R. F. Warming. An implicit finite-difference algorithm for hyperbolic systems in conservation-law form. *J. Comput. Phys*, 22,:87–110, 1976.

[BW78] R. M. Beam and R. F. Warming. An implicit factored scheme for the compressible Navier-Stokes equations. *AIAA J.*, 16:393–402, 1978.

[Cas02] Paul Castillo. Performance of discontinuous Galerkin methods for elliptic PDEs. *SIAM J. Sci. Comput.*, 24(2):524–547, February 2002.

[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.

[Che06] H. Chen. Superconvergence properties of discontinuous Galerkin methods for two-point boundary value problems. *Int. J. Numer. Anal. Model.*, 3(2):163–185, 2006.

[CHS90] B. Cockburn, S. Hou, and C. W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element for conservation laws IV: The multi-dimensional case. *Math. Comp.*, 54:545–581, 1990.

[Cia79] P. G. Ciarlet. *The Finite Elements Method for Elliptic Problems*. North-Holland, Amsterdam, New York, Oxford, 1979.

[CS89] B. Cockburn and C. W. Shu. TVB Runge–Kutta local projection discontinuous Galerkin finite element for conservation laws II: General framework. *Math. Comput.*, 52:411–435, 1989.

[CS07] A. Chaillou and M. Suri. A posteriori estimation of the linearization error for strongly monotone nonlinear operators. *J. Comput. Appl. Math.*, 205(1):72–87, 2007.

[DD99] T. A. Davis and I. S. Duff. A combined unifrontal/multifrontal method for unsymmetric sparse matrices. *ACM Transactions on Mathematical Software*, 25:1–19, 1999.

[dDBHM12] B. Ayuso de Dios, F. Brezzi, O. Havle, and L. D. Marini. L^2-estimates for the DG IIPG-0 scheme. *Numer. Meth. Partial Diff. Eqs*, 28(5):1440–1465, 2012.

[Deu04] P. Deuflhard. *Newton Methods for Nonlinear Problems*, volume 35 of *Springer Series in Computational Mathematics*. Springer, 2004.

[DEV13] V. Dolejší, A. Ern, and M. Vohralík. A framework for robust a posteriori error control in unsteady nonlinear advection-diffusion problems. *SIAM J. Numer. Anal.*, 51(2):773–793, 2013.

[DF03] V. Dolejší and M. Feistauer. On the discontinuous Galerkin method for the numerical solution of compressible high-speed flow. In F. Brezzi, A. Buffa, S. Corsaro, and A. Murli, editors, *Numerical Mathematics and Advanced Applications, ENUMATH 2001*, pages 65–84. Springer-Verlag, Italia, Milano, 2003.

[DF04a] V. Dolejší and M. Feistauer. Semi-implicit discontinuous Galerkin finite element method for the numerical solution of inviscid compressible flow. *J. Comput. Phys.*, 198(2):727–746, 2004.

[DF04b] V. Dolejší and J. Felcman. Anisotropic mesh adaptation and its application for scalar diffusion equations. *Numer. Methods Partial Differ. Equations*, 20:576–608, 2004.

[DFS03] V. Dolejší, M. Feistauer, and C. Schwab. On some aspects of the discontinuous Galerkin finite element method for conservation laws. *Math. Comput. Simul.*, 61:333–346, 2003.

[DH79] L.M. Delves and C.A. Hall. An implicit matching principle for global element calculations. *J. Inst. Math. Appl.*, 23:223–234, 1979.

[DH10] V. Dolejší and O. Havle. The L^2-optimality of the IIPG method for odd degrees of polynomial approximation in 1D. *J. Sci. Comput.*, 42(1):122–143, 2010.

[DHH11] V. Dolejší, M. Holík, and J. Hozman. Efficient solution strategy for the semi-implicit discontinuous Galerkin discretization of the Navier-Stokes equations. *J. Comput. Phys.*, 230:4176–4200, 2011.

[Dic91] E. Dick. Second-order formulation of a multigrid method for steady Euler equations through defect-correction. *J. Comput. Appl. Math. 35, No.1-3, 159-168 (1991*, 35(1-3):159–168, 1991.

[DK08] V. Dolejší and P. Kůs. Adaptive backward difference formula – discontinuous Galerkin finite element method for the solution of conservation laws. *Int. J. Numer. Methods Eng.*, 73(12):1739–1766, 2008.

[DMS03] M. Darwish, F. Moukalled, and B. Sekar. A robust multi-grid pressure-based algorithm for multi-fluid flow at all speeds. *Int. J. Numer. Methods Fluids*, 41:1221–1251, 2003.

[Dol98] V. Dolejší. Anisotropic mesh adaptation for finite volume and finite element methods on triangular meshes. *Comput. Vis. Sci.*, 1(3):165–178, 1998.

[Dol00] V. Dolejší. *ANGENER – software package*. Charles University Prague, Faculty of Mathematics and Physics, 2000. www.karlin.mff.cuni.cz/ dolejsi/angen.html.

[Dol01] V. Dolejší. Anisotropic mesh adaptation technique for viscous flow simulation. *East-West J. Numer. Math.*, 9(1):1–24, 2001.

[Dol08a] V. Dolejší. Analysis and application of IIPG method to quasilinear nonstationary convection-diffusion problems. *J. Comp. Appl. Math.*, 222:251–273, 2008.

[Dol08b] V. Dolejší. Semi-implicit interior penalty discontinuous Galerkin methods for viscous compressible flows. *Commun. Comput. Phys.*, 4(2):231–274, 2008.

[Dol13a] V. Dolejší. A design of residual error estimates for a high order BDF-DGFE method applied to compressible flows. *Int. J. Numer. Meth. Fluids*, 73(6):523–559, 2013.

[Dol13b] V. Dolejší. *hp*-DGFEM for nonlinear convection-diffusion problems. *Math. Comput. Simul.*, 87:87–118, 2013.

[DP80] L.M. Delves and C. Phillips. A fast implementation of the global element methods. *J. Inst. Math. Appl.*, 25:179–197, 1980.

[DRD02] L. Demkowicz, W. Rachowicz, and Ph. Devloo. A fully automatic *hp*-adaptivity. *J. Sci. Comput.*, 17(1-4):117–142, 2002.

[DSW04] C. N. Dawson, S. Sun, and M. F. Wheeler. Compatible algorithms for coupled flow and transport. *Comput. Meth. Appl. Mech. Engng.*, 193:2565–2580., 2004.

[DT04] V. Daru and C. Tenaud. High order one-step monotonicity-preserving schemes for unsteady compressible flow calculations. *J. Comput. Phys.*, 193(2):563–594, 2004.

[Dun85] D. A. Dunavant. High degree efficient symmetrical gaussian quadrature rules for the triangle. *Int. J. Numer. Methods Eng.*, 21:1129–1148, 1985.

[Edw65] R. E. Edwards. *Functional Analysis, Theory and Applications*. Holt, Rinehart and Winston, New York, 1965.

[EEHJ95] Kenneth Eriksson, Don Estep, Peter Hansbo, and Claes Johnson. Introduction to adaptive methods for differential equations. *Acta Numerica*, 4:105–158, 0 1995.

[EEHJ96] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. *Computational Differential Equations*. Cambridge University Press, Cambridge, 1996.

[EGH00] R. Eymard, T. Gallouët, and R. Herbin. *Solution of equations in R^n (Part 3). Techniques of scientific computing (Part 3).*, chapter Finite volume methods, pages 713–1020. Handbook of numerical analysis. Amsterdam: North-Holland/ Elsevier, 2000.

[EM07] T. Eibner and J. M. Melenk. An adaptive strategy for *hp*-FEM based on testing for analyticity. *Comput. Mech.*, 39(5):575–595, 2007.

[EV10] A. Ern and M. Vohralík. Aposteriori error estimation based on potential and flux reconstruction for the heat equation,. *SIAM J. Numer. Anal.*, 48:198–223, 2010.

[FDK06] M. Feistauer, V. Dolejší, and V. Kučera. On a semi-implicit discontinuous Galerkin FEM for the nonstationary compressible Euler equations. In F. Asukara, H. Aiso, S. Kawashima, A. Matsumura, S. Nishibata, and K. Nishihara, editors, *Hyperbolic Probelms: Theory Numerics and Applications*, pages 391–398, Tenth international conference in Osaka, September 2004,, 2006. Yokohama Publishers, Inc.

[FDK07] M. Feistauer, V. Dolejší, and V. Kučera. On the discontinuous Galerkin method for the simulation of compressible flow with wide range of Mach numbers. *Comput Visual Sci*, 10:17–27, 2007.

[Fei89] M. Feistauer. On the finite element approximation of functions with noninteger derivatives. *Numer. Funct. Anal. and Optimiz.*, 10:91–110, 1989.

[Fei93] M. Feistauer. *Mathematical Methods in Fluid Dynamics*. Longman Scientific & Technical, Harlow, 1993.

[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.

[FFS03] M. Feistauer, J. Felcman, and I. Straškraba. *Mathematical and Computational Methods for Compressible Flow*. Oxford University Press, Oxford, 2003.

[FHH+11] M. Fabian, P. Habala, P. Hájek, V. Montesinos, and V. Zizler. *Banach Space Theory*. Springer-Verlag, New York, 2011.

[FK07] M. Feistauer and V. Kučera. On a robust discontinuous Galerkin technique for the solution of compressible flow. *J. Comput. Phys.*, 224:208–221, 2007.

[Fra61] L. Frankel. On corner eddies in plane inviscid shear flow. *J. Fluid Mech.*, 11:400–406, 1961.

[FS89] L. Fezoui and B. Stoufflet. A class of implicit upwind schemes for Euler simulations with unstructured meshes. *J. Comput. Phys.*, 84(1):174–206, 1989.

[FŠ04] M. Feistauer and K. Švadlenka. Discontinuous Galerkin method of lines for solving nonstationary singularly perturbed linear problems. *J. Numer. Math.*, 12:97–118, 2004.

[FS12] Miloslav Feistauer and Anna-Margarete Sändig. Graded mesh refinement and error estimates of higher order for DGFE solutions of elliptic boundary value problems in polygons. *Numer. Methods Partial Differential Equations*, 28(4):1124–1151, 2012.

[Für01] J. Fürst. *Modélisation numérique d'écoulements transsoniques avec des schémas TVD et ENO*. PhD thesis, Universite Mediterranee, Marseille and Czech Technical University Prague, 2001.

[Geo06] E. H. Georgoulis. *hp*-version interior penalty discontinuous Galerkin finite element methods on anisotropic meshes. *Int. J. Numer. Anal. Model.*, 3(1):52–79, 2006.

[GF87] B. Gustafsson and L. Fern. Far fields boundary conditions for steady state solutions to hyperbolic systems. In *Nonlinear hyperbolic problems, Proceedings, St Etienne 1986*, number 1270 in Lecture Notes in Mathematics, pages 238–252. Springer, Berlin, 1987.

[GF88] B. Gustafsson and L. Fern. Far fields boundary conditions for time-dependent hyperbolic systems. *SIAM J. Sci. Statist. Comput.*, 9:812–848, 1988.

[GHH07] E. H. Georgoulis, E. Hall, and P. Houston. Discontinuous Galerkin methods on *hp*-anisotropic meshes I: A priori error analysis. *Int. J. Comput. Sci. Math*, 1(2-3):221–244, 2007.

[Gil90] M. B. Giles. Non-reflecting boundary conditions for Euler equation calculations. *AIAA J.*, 42:2050–2058, 1990.

[GK79] B. Gustafsson and H.-O. Kreiss. Boundary conditions for time-dependent problems with an artificial boundary. *J. Comput. Phys.*, 30:333–351, 1979.

[GR96] E. Godlewski and P. A. Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, volume 118 of *Applied Mathematical Sciences*. Springer, New York, 1996.

[GR09] J. Guzmán and B. Rivière. Sub-optimal convergence of non-symmetric discontinuous Galerkin method for odd polynomial approximations. *J. Sci. Comput.*, 40(1-3):273–280, 2009.

[Gri92] P. Grisvard. *Singularities in Boundary Value Problems*. pringer, Berlin, 1992.

[GS05] E. H. Georgoulis and E. Süli. Optimal error estimates for the *hp*-version interior penalty discontinuous Galerkin finite element method. *IMA J. Numer. Anal.*, 25(1):205–220, 2005.

[HD79] J.A. Hendry and L.M. Delves. The global element method applied to a harmonic mixed boundary value problem. *J. Comput. Phys.*, 33:33–44, 1979.

[HDP79] J.A. Hendry, L.M. Delves, and C. Phillips. Numerical experience with the global element method. In *Mathematics of finite elements and applications III*, pages 341–348. Academic Press, London-New York, 1979.

[Hed79] G. W. Hedstrom. Nonreflecting boundary conditions for nonlinear hyperbolic systems. *J. Comput. Phys.*, 30:222–237, 1979.

[HH88] T. Hagstrom and S. I. Hariharan. Accurate boundary conditions for exterior problems in gas dynamics. *Math. Comput.*, 51:581–597, 1988.

[HH02] R. Hartmann and P. Houston. Adaptive discontinuous Galerkin finite element methods for the compressible Euler equations. *J. Comput. Phys.*, 183(2):508–532, 2002.

[HH06a] R. Hartmann and P. Houston. Symmetric interior penalty DG methods for the compressible Navier-Stokes equations I: Method formulation. *Int. J. Numer. Anal. Model.*, 1:1–20, 2006.

[HH06b] R. Hartmann and P. Houston. Symmetric interior penalty DG methods for the compressible Navier-Stokes equations II: Goal-oriented a posteriori error estimation. *Int. J. Numer. Anal. Model.*, 3:141–162, 2006.

[Hir88] C. Hirch. *Numerical computation of internal and external flows. Volume 1: Fundamentals of numerical discretization.* Wiley Series in Numerical Methods in Engineering. Wiley-Interscience Publication. Chichester, 1988.

[HNW00] E. Hairer, S. P. Norsett, and G. Wanner. *Solving ordinary differential equations I, Nonstiff problems.* Number 8 in Springer Series in Computational Mathematics. Springer Verlag, 2000.

[Hol10] M. Holík. *Discontinuous Galerkin Method for Convection-Diffusion Problems.* PhD thesis, Charles University Prague, 2010.

[Hoz09] J. Hozman. *Discontinuous Galerkin Method for Convection-Diffusion Problems.* PhD thesis, Charles University Prague, Faculty of Mathematics and Physics, 2009.

[HRS05] P. Houston, J. Robson, and E. Süli. Discontinuous Galerkin finite element approximation of quasilinear elliptic boundary value problems I: The scalar case. *IMA J. Numer. Anal.*, 25:726–749, 2005.

[HS86] P. W. Hemker and S.P. Spekreijse. Multiple grid and osher's scheme for the efficient solution of the steady Euler equations. *Appl. Numer. Math.*, 2:475–493, 1986.

[HS05] P. Houston and E. Süli. A note on the design of *hp*–adaptive finite element methods for elliptic partial differential equations. *Comput. Methods Appl. Mech. Engrg.*, 194:229–243, 2005.

[HSS02] P. Houston, C. Schwab, and E. Süli. Discontinuous *hp*-finite element methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 39(6):2133–2163, 2002.

[HSW07] P. Houston, D. Schötzau, and T. P. Wihler. Energy norm a posteriori error estimation of *hp*-adaptive discontinuous Galerkin methods for elliptic problems. *Math. Models Methods Appl. Sci.*, 17(1):33–62, 2007.

[HSW08] P. Houston, E. Süli, and T. P. Wihler. A posteriori error analysis of *hp*-version discontinuous Galerkin finite element methods for second-order quasilinear elliptic problems. *IMA J. Numer. Anal.*, 28:245–273, 2008.

[JJS95] J. Jaffre, C. Johnson, and A. Szepessy. Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws. *Math. Models Methods Appl. Sci*, 5(3):367–286, 1995.

[JK07] V. John and P. Knobloch. On spurious oscillations at layer diminishing (SOLD) methods for convection–diffusion equations: Part I – A review. *Comput. Methods Appl. Mech Engrg.*, 196:2197–2215, 2007.

[JSV10] P. Jiránek, Z. Strakoš, and M. Vohralík. A posteriori error estimates including algebraic error and stopping criteria for iterative solvers. *SIAM J. Sci. Comput.*, 32(3):1567–1590, 2010.

[KBD+10] N. Kroll, H. Bieler, H. Deconinck, V. Couallier, H. van der Ven, and K. Sorensen, editors. *ADIGMA - A European Initiative on the Development of Adaptive Higher-Order Variational Methods for Aerospace Applications*, volume 113 of *Notes on Numerical Fluid Mechanics and Multidisciplinary Design.* Springer Verlag, 2010.

[KH91] B. Koren and P. W. Hemker. Damped, direction-dependent multigrid for hypersonic flow computations. *Appl. Numer. Math.*, 7(4):309–328, 1991.

[KJk77] A. Kufner, O. John, and S. Fučík. *Function Spaces.* Academia, Prague, 1977.

[Kle95] R. Klein. Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics 1: one-dimensional flow. *J. Comput. Phys.*, 121:213–237, 1995.

[Krö91] D. Kröner. Absorbing boundary conditions for the linearized Euler equations in 2-D. *Math. Comput.*, 57:153–167, 1991.

[Krö97] D. Kröner. *Numerical Schemes for Conservation Laws.* Wiley Teubner, Stuttgart, 1997.

[KS87] A. Kufner and A.-M. Sändig. *Some applications of weighted Sobolev spaces*. Teubner, Leipzig, 1987.

[Leo09] G. Leoni. *A First Course in Sobolev Spaces*. Graduate Studiesin Mathematics Vol. 105. AMS, Providence, 2009.

[Lio96] P. L. Lions. *Mathematical Topics in Fluid Mechanics*. Oxford Science Publications, 1996.

[LN04] M. G. Larson and A. J. Niklasson. Analysis of a family of discontinuous Galerkin methods for elliptic problems: the one dimensional case. *Numer. Math.*, 99(1):113–130, 2004.

[LS13] Jörg Liesen and Zdeněk Strakoš. *Krylov subspace methods (Principles and analysis)*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013.

[Mei98] A. Meister. Comparison of different Krylov subspace methods embedded in an implicit finite volume scheme for the computation of viscous and inviscid flow fields on unstructured grids. *J. Comput. Phys.*, 140:311–345, 1998.

[Mei03] A. Meister. Viscous flow fields at all speeds: Analysis and numerical simulation. *J. Appl. Math. Phys.*, 54:1010–1049, 2003.

[Mit98] S. Mittal. Finite element computation of unsteady viscous compressible flows. *Comput. Methods Appl. Mech. Eng.*, 157:151–175, 1998.

[MS02] A. Meister and J. Struckmeier. *Hyperbolic Partial Differential Equations, Theory, Numerics and Applications*. Vieweg, Brauschweig/Wiesbaden, 2002.

[Neč67] J. Nečas. *Les Méthodes Directes en Thèorie des Equations Elliptiques*. Academia, Prague, 1967.

[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.

[PM05] J.H. Park and C.-D. Munz. Multiple pressure variables methods for fluid flow at all Mach numbers. *Int. J. Numer. Methods Fluids*, 49:905–931, 2005.

[QV99] Alfio Quarteroni and Alberto Valli. *Domain decomposition methods for partial differential equations*. Numerical Mathematics and Scientific Computation. The Clarendon Press Oxford University Press, New York, 1999. Oxford Science Publications.

[Rek82] K. Rektorys. *The Method of Discretization in Time and Partial Differential Equations*. Reidel, Dodrecht, 1982.

[Riv08] B. Rivière. *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation*. Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 2008.

[RMGK97] S. Roller, C.-D. Munz, K.J. Geratz, and R. Klein. The multiple pressure variables method for weakly compressible fluids. *Z. Angew. Math. Mech.*, 77:481–484, 1997.

[Roe89] P. L. Roe. Remote boundary conditions for unsteady multidimensional aerodynamic computations. *Comput. Fluids*, 17:221–231, 1989.

[Rou05] T. Roubíček. *Nonlinear Partial Differential Equations with Applications*. Birkhäuser, Basel, Boston, Berlin, 2005.

[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.

[RST08] H.-G. Roos, M. Stynes, and L. Tobiska. *Robust Numerical Methods for Singularly Perturbed Differential Equations*. Springer Series in Computational Mathematics. Springer-Verlag, Berlin Heidelberg, 2008.

[Rud87] W. Rudin. *Real and complex analysis*. McGraw-Hill, New York, 3rd edition, 1987.

[RWG99] B. Rivière, M. F. Wheeler, and V. Girault. Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. I. *Comput. Geosci.*, 3(3-4):337–360, 1999.

[RWG01] B. Rivière, M. F. Wheeler, and V. Girault. A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems. *SIAM J. Numer. Anal.*, 39(3):902–931, 2001.

[Sch98] C. Schwab. *p- and hp-Finite Element Methods*. Clarendon Press, Oxford, 1998.

[Sch00] C. Schwab. Discontinuous Galerkin method. Technical report, ETH Zürich, 2000.

[ŠD04] P. Šolín and L. Demkowicz. Goal-oriented *hp*-adaptivity for elliptic problems. *Comput. Methods Appl. Mech. Engrg.*, 193:449–468, 2004.

[Shu98] C.W. Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. In A. Quarteroni et al, editor, *Advanced numerical approximation of nonlinear hyperbolic equations*, Lect. Notes Math. 1697, pages 325–432. Berlin: Springer, 1998.

[Sob11] V. Sobotíková. Error analysis of a DG method employing ideal elements applied to a nonlinear convection-diffusion problem. *J. Numer. Math.*, 19(2):137–163, 2011.

[Šol04] P. Šolín. *Partial differential equations and the finite element method.* Pure and Applied Mathematics. Wiley-Interscience, New York, 2004.

[SS86] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 7:856–869, 1986.

[ŠSD03] P. Šolín, K. Segeth, and I. Doležel. *Higher-Order Finite Element Methods.* Chapman & Hall/CRC Press, 2003.

[Sto85] B. Stoufflet. Implicit finite element methods for the Euler equations. In *Numerical methods for the Euler equations of fluid dynamics*, Proc. INRIA Workshop, Rocquencourt/France 1983, pages 409–434, 1985.

[Sun03] S. Sun. *Discontinuous Galerkin methods for reactive transport in porous media.* PhD thesis, The University of Texas, Austin, 2003.

[SW03] D. Schötzau and T. P. Wihler. Exponential convergence of mixed *hp*-DGFEM for Stokes flow in polygons. *Numer. Math.*, 96:339–361, 2003.

[SW05] S. Sun and M. F. Wheeler. Symmetric and nonsymmetric discontinuous Galerkin methods for reactive transport in porous media. *SIAM J. Numer. Anal.*, 43(1):195–219, 2005.

[TGS00] C. Tenaud, E. Garnier, and P. Sagaut. Evaluation of some high-order shock capturing schemes for direct numerical simulation of unsteady two-dimensional free flows. *Int. J. Numer. Meth. Fluids*, 126:202–228, 2000.

[Tho06] V. Thomée. *Galerkin finite element methods for parabolic problems. 2nd revised and expanded ed.* Berlin, Springer, 2006.

[Tor97] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics.* Springer-Verlag, Berlin, 1997.

[Tos02] A. Toselli. *hp* discontinuous Galerkin approximations for the Stokes problem. *Math. Models Methods Appl. Sci*, 12(11):1565–1597, 2002.

[vdHVW03] D.R. van der Heul, C. Vuik, and P. Wesseling. A conservative pressure-correction method for flow at all speeds. *Comput. Fluids*, 32:1113–1132, 2003.

[vdVvdV02a] J. J. W. van der Vegt and H. van der Ven. Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flows. I: General formulation. *J. Comput. Phys.*, 182(2):546–585, 2002.

[vdVvdV02b] H. van der Ven and J. J. W. van der Vegt. Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flows II. efficient flux quadrature. *Comput. Methods Appl. Mech. Engrg.*, 191:4747–4780, 2002.

[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.

[Ver13] R. Verfürth. *A Posteriori Error Estimation Techniques for Finite Element Methods.* Numerical Mathematics and Scientific Computation. Oxford University Press, 2013.

[Vij86] G. Vijayasundaram. Transonic flow simulation using upstream centered scheme of Godunov type in finite elements. *J. Comput. Phys.*, 63:416–433, 1986.

[Vla10] M. Vlasák. *Numerical solution of convection–diffusion problems by discontinuous Galerkin method.* PhD thesis, Charles University Prague, 2010.

[Voh10] M. Vohralík. *A posteriori error estimates, stopping criteria and inexpensive implementation.* Habilitation thesis, Université Pierre et Marie Curie – Paris 6, 2010.

[Wes01] P. Wesseling. *Principles of Computational Fluid Dynamics.* Springer, Berlin, 2001.

[WFS03] T. P. Wihler, O. Frauenfelder, and C. Schwab. Exponential convergence of the *hp*-DGFEM for diffusion problems. *Comput. Math. Appl.*, 46:183–205, 2003.

[Whe78] M.F. Wheeler. An elliptic collocation-finite element method with interior penalties. *SIAM J. Numer. Anal*, 15(4):152–161, 1978.

[Wih02] T. P. Wihler. *Discontinuous Galerkin FEM for Elliptic Problems in Polygonal Domains*. PhD thesis, ETH Zürich, 2002.

[Žen90] A . Ženíšek. *Nonlinear Elliptic and Evolution Problems and Their Finite Element Approximations*. Academic Press, London, 1990.