Charles University

Faculty of Mathematics and Physics

DOCTORAL THESIS

Oto Havle

Numerical analysis of partial differential equations with applications in mathematical modeling

Department of Numerical Mathematics

Supervisor: *Doc. RNDr. Jiří Felcman, CSc.*

Title: Numerical analysis of partial differential equations with applications in mathematical modeling
Author: Oto Havle
Department: Department of Numerical Mathematics
Supervisor: Doc. RNDr. Jiří Felcman, CSc.
Author's e-mail address: havle@karlin.mff.cuni.cz
Supervisor's e-mail address: felcman@karlin.mff.cuni.cz

Abstract: The thesis is concerned with selected tools useful in the numerical analysis of partial differential equations of elliptic and hyperbolic type, namely the broken Sobolev spaces, discontinuous Galerkin method and finite volume method. The properties of broken Sobolev spaces $W^{1,p}(\Omega, \mathcal{T}_h)$, which underlie the discontinuous Galerkin methods are studied, namely traces, imbeddings into $L^q(\Omega)$, and $BV(\Omega)$, imbeddings into certain Besov spaces and interpolation between the spaces $W^{1,p}(\Omega, \mathcal{T}_h)$ with varying p. Apriori error analysis of the Interior Penalty discontinuous Galerkin method is presented, with emphasis to the convergence in L^2-norm. Optimal order of convergence of the Incomplete Interior Penalty Galerkin method on one-dimensional non-uniform meshes is proved. Finally, a numerical flux of Vijayasundaram type for the shallow water equations is constructed. Shallow water equations represent a hyperbolic system of conservation laws with a source term. The numerical flux is employed in the finite volume method. The resulting numerical scheme preserves certain class of stationary solutions.

Keywords: discontinuous Galerkin method, finite volume method, broken Sobolev spaces, shallow water equations

Název práce: Numerická analýza parciálních diferenciálních rovnic
s aplikacemi v matematickém modelování
Autor: Oto Havle
Katedra: Katedra numerické matematiky
Školitel: Doc. RNDr. Jiří Felcman, CSc.
e-mail autora: havle@karlin.mff.cuni.cz
e-mail školitele: felcman@karlin.mff.cuni.cz

Abstrakt: Dizertace se zabývá vybranými nástroji pro numerickou analýzu eliptických a hyperbolických parciálních diferenciálních rovnic, zejména nespojitou Galerkinovou metodou, metodou konečných objemů a prostory po částech sobolevovských funkcí. Teorie prostorů $W^{1,p}(\Omega, \mathcal{T}_h)$ po částech sobolevovských funkcí obsahuje věty o stopách, vnoření do prostorů $L^q(\Omega)$, $BV(\Omega)$ a Běsovových prostorů a interpolace mezi prostory $W^{1,p}(\Omega, \mathcal{T}_h)$ s různým exponentem p. Dále práce obsahuje analýzu konvergence nespojité Galerkinovy metody s vnitřní penalizací, s důrazem na konvergenci v normě prostoru $L^2(\Omega)$. Je dokázán optimální řád konvergence metody s neúplnou penalizací na nerovnoměrných jednorozměných sítích. V poslední části práce je uvedena konstrukce numerického toku Vijayasundaramova typu pro tzv. rovnice mělké vody (shallow water equations), které jsou příkladem hyperbolického systému se zdrojovým členem. Výsledné numerické schema respektuje jistou množinu stacionárních stavů.

Klíčová slova: nespojitá Galerkinova metoda, metoda konečných objemů, prostory po částech sobolevovkých funkcí, rovnice mělké vody

Acknowledgments

I would like to thank Vít Dolejší for introducing me to the subject of discontinuous Galerkin methods, and my supervisor Jiří Felcman for guidance and numerous corrections.

Research presented in this thesis was partially supported by the project LC06052 of the Ministry of Education of the Czech Republic (Jindřich Nečas Center for Mathematical Modelling).

Contents

List of Figures

List of Tables

Introduction

The broken Sobolev spaces $W^{1,p}(\Omega, \mathcal{T}_h)$ are spaces of functions whose restrictions to elements K of a given mesh $\mathcal{T}_h$ belong to the corresponding classical Sobolev spaces $W^{1,p}(K)$. Broken Sobolev spaces are useful generalization of the spaces of piecewise polynomial functions, which are employed as trial and test spaces in the discontinuous Galerkin (DG) methods. Analysis of these spaces is unavoidable prerequisite when studying the DG methods. The properties of the norms and relations to the classical Sobolev spaces justify the choice of penalization in the Interior Penalty discontinuous Galerkin (IPG) method.

In the first chapter, we review the properties of broken Sobolev spaces and the corresponding mesh-dependent norms. We generalize the global multiplicative trace theorem, published for $W^{1,2}(\Omega, \mathcal{T}_h)$ in (Dolejší, Feistauer and Havle, 2009), to the spaces $W^{1,p}(\Omega, \mathcal{T}_h)$ with arbitrary $p \in [1, \infty]$ (Theorem 1.7). We also show that the spaces $W^{1,p}(\Omega, \mathcal{T}_h)$ form a scale of interpolation spaces, using the real K-method of interpolation (Theorem 1.10). Up to author's knowledge, interpolation between the spaces $W^{1,p}(\Omega, \mathcal{T}_h)$ was not yet studied in literature. As an example of application of the interpolation result, we prove imbeddings of broken Sobolev spaces in certain Besov spaces (Lemma 1.24).

In the second chapter, we analyze the Interior Penalty discontinuous Galerkin methods for a model elliptic problem. We review the results on convergence in the broken H^1-seminorm, but our main focus is the convergence in L^2-norm. It is well known that the non-symmetric variants, namely the so called Nonsymmetric Interior Penalty discontinuous Galerkin method (NIPG) and the Incomplete Interior Penalty discontinuous Galerkin method (IIPG) exhibit suboptimal order of convergence in L^2-norm. In general, the suboptimality is attributed to the lack of adjoint consistency of the method (see Arnold et al., 2002). However, the adjoint consistency is not necessary condition for optimality. Moreover, the optimality depends on the parity of the degree of piecewise-polynomial discontinuous trial and test functions employed in the IPG methods (see Rivière, 2008). The adjoint-consistency clearly cannot explain this phenomenon.

Theoretical results concerning L^2-convergence of NIPG and IIPG were limited only to one-dimensional case and special multidimensional cases (see Larson and Niklasson, 2004; Burman and Stamm, 2008; Wang et al., 2009). Even in the one-dimensional cases, the analysis was restricted to uniform meshes. We present full analysis of IIPG method on one-dimensional non-uniform meshes (Theorem 2.6), which was originally published by Dolejší and Havle (2010).

Although the DG techniques are applicable to wide range of PDEs (see Cockburn et al., 2000, and references given therein), the applications to hyperbolic systems of conservation laws are most prevalent. In fact, the traditional Finite Volume (FV) method is a special case of DG. When discretizing first-order hyperbolic equations with FVM or DG, we need a so called numerical flux, which is consistent with the given PDE system.

In the third chapter, we present the construction of numerical flux for the Shallow Water Equations (SWE), which was originally published in (Felcman and Havle, 2010). The construction is inspired by the well-known Vijayasundaram flux from the context of compressible Euler equations. We prove that the flux is consistent, conservative and continuous (Theorem 3.4). We show that FV method employing this numerical flux preserves certain class of stationary solutions (Theorem 3.7).

Let us finish the introduction with few remarks about organization of the thesis. In chapters 1 to 3, we formulate the original results in Lemmas and Theorems. The results adopted from other authors are marked as Properties. In most cases, the Properties are stated without proofs. In few cases, we include proofs of such Properties for clarity. The appendix contain numerical experiments which illustrate the theoretical findings.

Chapter 1

Broken Sobolev Spaces

The spaces of piecewise polynomial functions play a crucial role in the analysis of discontinuous Galerkin methods. It is useful to have results independent of the polynomial order. That's why we analyze first the more general functions, whose restrictions to elements K of a given mesh $\mathcal{T}_h$ belong to the Sobolev space $W^{1,p}(K)$. Spaces of such functions are the so called called *broken Sobolev spaces.* The broken Sobolev space $W^{1,p}(\Omega, \mathcal{T}_h)$ is defined by

$$W^{1,p}(\Omega, \mathcal{T}_h) = \left\{ u \in L^p(\Omega) : u\big|_K \in W^{1,p}(K), \ K \in \mathcal{T}_h \right\} \tag{1.1}$$

The weak solutions of elliptic and parabolic PDEs belong to Sobolev spaces $W^{1,p}(\Omega)$, with suitable $p \in (1, \infty)$. The broken Sobolev spaces replace the spaces $W^{1,p}(\Omega)$ for the purposes of discontinuous Galerkin methods. Naturally, the broken Sobolev space $W^{1,p}(\Omega, \mathcal{T}_h)$ should in a sense approximate the Sobolev space $W^{1,p}(\Omega)$, at least in the limit $h \to 0$. Later in this chapter, we will see that many properties of the space $W^{1,p}(\Omega)$ are shared by the space $W^{1,p}(\Omega, \mathcal{T}_h)$ and our requirement is thus satisfied.

We have to respect this requirement when defining the norm in the space $W^{1,p}(\Omega, \mathcal{T}_h)$. Let us first show some heuristic arguments, which motivate the choice of norms. For the sake of argument, suppose v is a piecewise constant function. Let $v_K = v\big|_K$ denote the value of the function on the element K and x_K denote a representative point in K (e.g. the center of gravity of the element K). Concerning the case $p = 1$, $W^{1,1}(\Omega, \mathcal{T}_h) \subset BV(\Omega)$. The definition of norm on $W^{1,1}(\Omega, \mathcal{T}_h)$ should respect this inclusion. One can show that

$$|v|_{BV(\Omega)} = \sum_{\Gamma \in \mathcal{F}_h^I} \|[v]\|_{L^1(\Gamma)}\,, \tag{1.2}$$

where the sum is taken over all interior faces Γ of the partition $\mathcal{T}_h$, and $[v] = v_K - v_L$ is the jump of the function v on the common face Γ shared by two neighboring elements K and L.

Concerning the case $p = \infty$, the space $W^{1,\infty}(\Omega)$ is equal to the space $Lip(\Omega)$ of Lipschitz-continuous functions on Ω, that is

$$|v(x) - v(y)| \le C \,\|v\|_{W^{1,\infty}(\Omega)} \,|x - y|, \qquad \text{for a. a. } x, y \in \Omega, \ v \in W^{1,\infty}(\Omega). \tag{1.3}$$

Obviously, $W^{1,\infty}(\Omega, \mathcal{T}_h) \not\subset Lip(\Omega)$, since the functions from $W^{1,\infty}(\Omega, \mathcal{T}_h)$ need not be continuous. The natural generalization of the Lipschitz condition is the inequality

$$|v_K - v_L| \leq M|x_K - x_L|, \qquad K, L \in \mathcal{T}_h, \tag{1.4}$$

where M is a constant. Under suitable assumptions on the mesh $\mathcal{T}_h$ (see section 1.1 below), it follows from (1.4) that

$$\left|\frac{[v]}{h_\Gamma}\right| \leq C\frac{|v_K - v_L|}{|x_K - x_L|} \leq CM, \tag{1.5}$$

for all elements K, L sharing a common face Γ, where $h_\Gamma = \mathrm{diam}\,(\Gamma)$ is the diameter of the face Γ.

In the discussion above, we considered only piecewise-constant function. When defining the norm for arbitrary functions $v \in W^{1,p}(\Omega, \mathcal{T}_h)$, we must take into account not only the inter-element jumps, but also the behavior in the interior of each element. The equality (1.2) and the inequality (1.5) motivate us to define a norm in the space $W^{1,p}(\Omega, \mathcal{T}_h)$ by following formulae

$$\|v\|_{W^{1,p}(\Omega,\mathcal{T}_h)} = \left(\|v\|^p_{L^p(\Omega)} + |v|^p_{W^{1,p}(\Omega,\mathcal{T}_h)}\right)^{1/p}, \tag{1.6}$$

$$|v|_{W^{1,p}(\Omega,\mathcal{T}_h)} = \left(\sum_{K\in\mathcal{T}_h}\int_K |v|_{W^{1,p}(K)} + \sum_{\Gamma\in\mathcal{F}^I_h} h_\Gamma \left\|\frac{[v]}{h_\Gamma}\right\|^p_{L^p(\Gamma)}\right)^{1/p}, \qquad p \in [1, \infty), \tag{1.7}$$

$$\|v\|_{W^{1,\infty}(\Omega,\mathcal{T}_h)} = \max\left(\|v\|_{L^\infty(\Omega)}, |v|_{W^{1,\infty}(\Omega,\mathcal{T}_h)}\right), \tag{1.8}$$

$$|v|_{W^{1,\infty}(\Omega,\mathcal{T}_h)} = \max\left(\max_{K\in\mathcal{T}_h} |v|_{W^{1,\infty}(K)}, \max_{\Gamma\in\mathcal{F}^I_h}\left\|\frac{[v]}{h_\Gamma}\right\|_{L^\infty(\Gamma)}\right) \qquad p = \infty. \tag{1.9}$$

In the rest of this chapter, we state and prove properties of the space $W^{1,p}(\Omega)$, equipped with the norm (1.6)-(1.9).

1.1 Basic assumptions and notation

We assume that $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, is a given bounded domain with Lipschitz continuous boundary. By $L^p(\Omega)$ and $W^{s,p}(\Omega)$, $p \in [1, \infty]$, $s \in \{1, 2, \dots\}$ we denote the Lebesgue and Sobolev spaces, respectively, equipped with standard norms and seminorms. We use the standard abbreviations $W^{0,p}(\Omega) = L^p(\Omega)$ and $H^s(\Omega) = W^{s,2}(\Omega)$. By $BV(\Omega)$ we denote the space of functions with bounded variation, equipped with the norm and the seminorm

$$\|u\|_{BV(\Omega)} = \|u\|_{L^1(\Omega)} + |u|_{BV(\Omega)}, \tag{1.10}$$

$$|u|_{BV(\Omega)} = \sup_{\substack{\boldsymbol{\varphi}\in\left(C^1_0(\Omega)\right)^d \\ \|\boldsymbol{\varphi}\|_{(L^\infty(\Omega))^d}\leq 1}} \int_\Omega u \,\mathrm{div}\,\boldsymbol{\varphi}\,\mathrm{d}x, \qquad u \in BV(\Omega), \tag{1.11}$$

where $C_0^\infty(\Omega)$ is the space of continuously differentiable functions with compact support in Ω. Properties of these function spaces can be found in (Adams and Fournier, 2003; Giusti, 1984).

In order to keep proofs straightforward, we consider only conforming simplicial partitions of the domain Ω. We do not consider hanging nodes, general polygonal, polyhedral, or curvilinear elements. We refer to (Buffa and Ortner, 2009; Dolejší et al., 2002; Brenner, 2003, and references given therein).

Let us recall basic properties of simplexes. Consider d points $a^1, a^2, \dots, a^{d+1} \in \mathbb{R}^d$ and the convex hull $K = \operatorname{conv}\left\{a^1, \dots, a^{d+1}\right\}$. If the d-dimensional Lebesgue measure $|K|$ is not zero, we say that K is a *simplex* in $\mathbb{R}^d$. Then, the points $a^1, \dots, a^{d+1}$ are called *vertices* of K. The sets

$$\Gamma_j = \operatorname{conv}\left\{a^1, \dots, a^{j-1}, a^{j+1}, \dots, a^d\right\}, \qquad j = 1, \dots, d,$$

are called *faces* of the simplex K. The simplex K is a closed set. If $d = 1$, K is a closed bounded interval. If $d = 2$, K is a triangle, and if $d = 3$, K is a tetrahedron. Consequently, the faces are points (real numbers) for $d = 1$, line segments for $d = 2$ and triangles for $d = 3$.

By definition, every $x \in K$ is a convex combination of the vertices,

$$x = \sum_{i=1}^{d+1} \lambda_i a^i. \tag{1.12}$$

Moreover, the coefficients λ_i are unique, and satisfy

$$0 \le \lambda_i \le 1, \qquad \sum_{i=1}^{d+1} \lambda_i = 1. \tag{1.13}$$

The coefficients λ_i are called *barycentric coordinates* of the point x.

Definition 1.1. Let $\mathcal{T}_h$ be a finite set of simplices in $\mathbb{R}^d$. Let $\mathcal{F}_h$ denote the set of all $(d-1)$-dimensional faces of all elements $K \in \mathcal{T}_h$. We say that $\mathcal{T}_h$ is a *conforming partition* of Ω if

(A) $\overset{\circ}{K} \cap \overset{\circ}{L} = \emptyset$ for all $K, L \in \mathcal{T}_h$, $K \neq L$.

(B) $\bigcup_{K \in \mathcal{T}_h} K = \overline{\Omega}$

(C) For each pair of elements $K, L \in \mathcal{T}_h$, either the intersection $K \cap L$ is a face $\Gamma \in \mathcal{F}_h$, or the $(d-1)$-dimensional measure of $K \cap L$ is zero.

(D) Let $\mathcal{F}_h^{\partial\Omega} = \{\Gamma \in \mathcal{F}_h : \Gamma \subset \partial\Omega\}$ and $\mathcal{F}_h^I = \mathcal{F}_h \setminus \mathcal{F}_h^{\partial\Omega}$. There exist mappings

$$K^{\mathcal{L}}_{(\cdot)} : \mathcal{F}_h \to \mathcal{T}_h, \qquad\qquad K^{\mathcal{R}}_{(\cdot)} : \mathcal{F}_h^I \to \mathcal{T}_h$$

such that

$$\begin{aligned} K_\Gamma^{\mathcal{L}} \neq K_\Gamma^{\mathcal{R}}, \quad K_\Gamma^{\mathcal{L}} \cap K_\Gamma^{\mathcal{R}} &= \Gamma, && \Gamma \in \mathcal{F}_h^I, \\ K_\Gamma^{\mathcal{L}} \cap \partial\Omega &= \Gamma, && \Gamma \in \mathcal{F}_h^{\partial\Omega}. \end{aligned}$$

(E) There exist mapping $\boldsymbol{n}_{(\cdot)} : \mathcal{F}_h \to \mathbb{R}^d$, such that $\boldsymbol{n}_\Gamma$ is a unit normal vector to the face Γ, which points ,,outwards" of the element $K^{\mathcal{L}}_\Gamma$, i.e.

$$x + t\boldsymbol{n}_\Gamma \notin K^{\mathcal{L}}_\Gamma, \qquad \text{for all } x \in \Gamma, t > 0.$$

The mappings $K^{\mathcal{L}}_{(\cdot)}$, $K^{\mathcal{R}}_{(\cdot)}$ and $\boldsymbol{n}_{(\cdot)}$ define a orientation of each face Γ. The condition **(B)** imply that Ω is a polygonal (or polyhedral for $d = 3$) domain and **(E)** imply that for $\Gamma \in \mathcal{F}^{\partial\Omega}_h$, $\boldsymbol{n}_\Gamma$ is equal to the unit outer normal to the boundary $\partial\Omega$.

For each element $K \in \mathcal{T}_h$, we set $h_K = \operatorname{diam}(K)$, and for each face $\Gamma \in \mathcal{F}_h$, we put $h_\Gamma = \operatorname{diam}(\Gamma)$. We define the global mesh size $h = \max_{K\in\mathcal{T}_h} h_K$. By ρ_K we denote the radius of the largest d-dimensional ball inscribed into K. We define

$$\mathcal{D}(K) = \{L \in \mathcal{T}_h : L \cap K \neq \emptyset\}. \tag{1.14}$$

Definition 1.2. Let $C_r > 0$. We say that $\mathcal{T}_h$ is a C_r-regular partition of Ω, if $\mathcal{T}_h$ is a conforming partition of Ω, and

$$\frac{h_K}{\rho_K} \le C_r, \qquad\qquad K \in \mathcal{T}_h, \tag{1.15}$$

$$\frac{\max\{h_L, h_R\}}{\min\{h_L, h_R\}} \le C_r, \qquad\qquad L = K^{\mathcal{L}}_\Gamma,\ R = K^{\mathcal{R}}_\Gamma,\ \Gamma \in \mathcal{F}^I_h. \tag{1.16}$$

In the following considerations, we implicitly work with a family of partitions $\{\mathcal{T}_h\}_{h\in(0,h_0)}$. For simplicity of the notation, we assume that the partitions are parametrized with the global mesh size $h \in (0, h_0)$, where $h_0 > 0$. We assume the all partitions $\mathcal{T}_h$ under consideration are C_r-regular, with fixed constant C_r. The assumptions of shape regularity (1.15) and local quasi-uniformity (1.16) are standard in theory of the finite element method, see (Brenner and Scott, 2002; Ciarlet, 1978). In the following, we will use the symbol C to denote a generic constant, which does not depend on h or $\mathcal{T}_h$, but can depend only on Ω and the regularity constant C_r. The symbol C might denote different constant on different places. For example, we have

$$h_\Gamma \le h_K \le Ch_\Gamma, \qquad K \in \mathcal{T}_h,\ \Gamma \in \mathcal{F}_h,\ \Gamma \subset \partial K, \tag{1.17}$$

and

$$h^{d-1}_\Gamma \le C|\Gamma|, \qquad h^d_K \le C|K| \qquad K \in \mathcal{T}_h, \Gamma \in \mathcal{F}_h, \tag{1.18}$$

$$h_L \le Ch_K, \qquad \operatorname{card}\mathcal{D}(K) \le C, \qquad K \in \mathcal{T}_h,\ L \in \mathcal{D}(K). \tag{1.19}$$

Let us now turn to the definition of mesh-dependent function spaces. By $P^k(K)$ we denote the space of d-variate polynomials of degree at most k restricted to the element $K \in \mathcal{T}_h$. We set

$$\mathcal{S}_{h,k} = \left\{v \in L^1(\Omega) : v\big|_K \in P^k(K)\right\}, \qquad k = 0, 1, \ldots \tag{1.20}$$

For $p \in [1, \infty]$ and $s = 0, 1, \ldots$, we define the Broken Sobolev Space

$$W^{s,p}(\Omega, \mathcal{T}_h) = \left\{ u \in L^p(\Omega) : u\big|_K \in W^{s,p}(K),\ K \in \mathcal{T}_h \right\}. \tag{1.21}$$

For $p = 2$, we use the abbreviation $H^s(\Omega, \mathcal{T}_h) = W^{s,2}(\Omega, \mathcal{T}_h)$. For $v \in W^{1,1}(\Omega, \mathcal{T}_h)$, we set

$$\begin{aligned} v\big|_\Gamma^{\mathcal{L}} &= \text{ the trace of } v\big|_{K_\Gamma^{\mathcal{L}}} \text{ on } \Gamma, && \Gamma \in \mathcal{F}_h, \\ v\big|_\Gamma^{\mathcal{R}} &= \text{ the trace of } v\big|_{K_\Gamma^{\mathcal{L}}} \text{ on } \Gamma, && \Gamma \in \mathcal{F}_h^I. \end{aligned}$$

We define the *mean value* and *jump* of v on a face $\Gamma \in \mathcal{F}_h^I$ by

$$\langle v \rangle_\Gamma = \frac{1}{2}\left(v\big|_\Gamma^{\mathcal{L}} + v\big|_\Gamma^{\mathcal{R}} \right), \qquad [v]_\Gamma = v\big|_\Gamma^{\mathcal{L}} - v\big|_\Gamma^{\mathcal{R}}, \tag{1.22}$$

and for $\Gamma \in \mathcal{F}_h^{\partial\Omega}$ by $\langle v \rangle_\Gamma = [v]_\Gamma = v\big|_\Gamma^{\mathcal{L}}$. By Sobolev trace theorem, $\langle v \rangle_\Gamma, [v]_\Gamma \in L^p(\Gamma)$ for all $v \in W^{1,p}(\Omega, \mathcal{T}_h)$ and $\Gamma \in \mathcal{F}_h$. We often omit the subscript Γ and abbreviate the notation as $\langle v \rangle$ or $[v]$. The norm $\|\cdot\|_{W^{1,p}(\Omega,\mathcal{T}_h)}$ on the space $W^{1,p}(\Omega, \mathcal{T}_h)$ is defined by (1.6)-(1.9).

Let us recall several properties of the spaces $P^k(K)$.

Property 1.3 (Inverse inequality)**.** *For all $k, s_1, s_2 \in \mathbb{N}_0$ and $p_1, p_2 \in [1, \infty]$, there exists $C > 0$ such that*

$$|v|_{W^{s_1,p_1}(K)} \leq C h_K^{\left(s_2 - \frac{d}{p_2}\right) - \left(s_1 - \frac{d}{p_1}\right)} |v|_{W^{s_2,p_2}(K)}, \qquad v \in P^k(K),\ K \in \mathcal{T}_h. \tag{1.23}$$

Proof. See (Brenner and Scott, 2002, Lemma 4.5.3). □

Property 1.4 (Approximation properties of $P^k(K)$)**.** *Let $k \in \mathbb{N}_0, p \in [1, \infty]$, $K \in \mathcal{T}_h$ and $v \in W^{k,p}(K)$. Then there exists $q \in P^k(K)$ such that*

$$|v - q|_{W^{j,p}(K)} \leq C h_K^{k-j} |v|_{W^{k,p}(K)}, \qquad j = 0, \ldots, k. \tag{1.24}$$

Proof. Follows from the Bramble-Hilbert lemma, see (Brenner and Scott, 2002, Lemma 4.3.8). □

1.2 The multiplicative trace inequality

Lemma 1.5. *Let $G \subset \mathbb{R}^d$ be a bounded domain with Lipschitz continuous boundary. Then there exists a vector-valued function $\boldsymbol{\varphi} \in W^{1,\infty}(G)^d$ such that*

$$\boldsymbol{\varphi} \cdot \boldsymbol{n} \geq 1 \qquad \textit{almost everywhere on } \partial G, \tag{1.25}$$

where $\boldsymbol{n}$ is the unit outer normal vector to the boundary ∂G.

Proof. Since G is bounded, it follows from the Lipschitz-continuity of ∂G that there exist a finite open cover $\{U_j\}_{j=1}^m$ of ∂G. Moreover, for each $j = 1, \dots, m$, there exist a Cartesian coordinate system $X_j \equiv (\xi_{j,1}, \dots, \xi_{j,d})$ and a Lipschitz-continuous function $f_j : \Delta_j \subset \mathbb{R}^{d-1} \to \mathbb{R}$, such that the set $\Omega \cap U_j$ is represented by the inequality

$$\xi_{j,d} < f_j(\xi_{j,1}, \dots, \xi_{j,d-1}), \qquad (\xi_{j,1}, \dots, \xi_{j,d-1}) \in \Delta_j. \tag{1.26}$$

The inequality (1.26) represents $\Omega \cap U_j$ locally in the vicinity of the boundary $\partial\Omega$, but the definition of Lipschitz-continuity of the boundary is more involved, we refer to (Adams and Fournier, 2003, section 4.9) for details. We denote the coordinates of a point $x \in \mathbb{R}^d$ with respect to the coordinate system X_j by $\{x\}_{X_j}$. Since f_j is Lipschitz-continuous, it is differentiable almost everywhere in Δ_j. The outer unit normal to the set ∂G exists almost everywhere (with respect to the $(d-1)$-dimensional measure) and

$$\{\boldsymbol{n}(x)\}_{X_j} = \frac{1}{\sqrt{1 + |\nabla f_j(\xi_{j,1}, \dots, \xi_{j,d-1})|^2}} \left(-\nabla f_j(\xi_{j,1}, \dots, \xi_{j,d-1}), 1\right), \tag{1.27}$$

where $(\xi_{j,1}, \dots, \xi_{j,d}) = \{x\}_{X_j}$, and $\xi_{j,d} = f_j(\xi_{j,1}, \dots, \xi_{j,d-1})$. Let L be the Lipschitz constant of the functions f_k, $k = 1, \dots, m$. Let $\boldsymbol{z}_j$ be the vector, whose coordinates are $\{\boldsymbol{z}_j\}_{X_j} = (0, \dots, 0, \sqrt{1 + L^2})$. Since the coordinate systems are Cartesian,

$$\boldsymbol{z}_j \cdot \boldsymbol{n}(x) = \{\boldsymbol{z}_j\}_{X_j} \cdot \{\boldsymbol{n}(x)\}_{X_j} = \frac{\sqrt{1 + L^2}}{\sqrt{1 + |\nabla f_j(\xi_{j,1}, \dots, \xi_{j,d-1})|^2}} \geq 1 \tag{1.28}$$

almost everywhere on $\partial G \cap U_j$. By theorem on partition of unity (see Adams and Fournier, 2003, Theorem 3.15), there exist functions $\psi_j \in C_0^\infty(U_j)$, $j = 1, \dots, m$, such that

$$0 \leq \psi_j \leq 1, \ j = 1, \dots, m, \qquad \sum_{j=1}^m \psi_j(x) = 1, \ x \in \partial G.$$

Now we define the function $\boldsymbol{\varphi}$ by

$$\boldsymbol{\varphi}(x) = \sum_{j=1}^m \psi_j(x) \boldsymbol{z}_j, \qquad x \in \mathbb{R}^d. \tag{1.29}$$

Obviously, $\boldsymbol{\varphi} \in C_0^\infty(\mathbb{R}^d)^d$. Moreover, using (1.28) and the property $\operatorname{supp} \psi_j \subset U_j$, we have

$$\psi_j(x) \boldsymbol{z}_j \cdot \boldsymbol{n}(x) \geq \psi_j(x), \qquad \text{for all } j = 1, \dots, m, \text{ and almost all } x \in \partial G,$$

and

$$\boldsymbol{\varphi}(x) \cdot \boldsymbol{n}(x) = \sum_{j=1}^m \psi_j(x) \boldsymbol{z}_j \cdot \boldsymbol{n}(x) \geq \sum_{j=1}^m \psi_j(x) = 1, \qquad \text{for a.a. } x \in \partial G. \qquad \square$$

We can obtain the auxiliary function φ explicitly for particular examples of the domain G.

Example. *Let G be interior of a simplex K in $\mathbb{R}^d$. Let $x_0 \in G$ be arbitrary. Then the function*

$$\boldsymbol{\varphi}(x) = \frac{x - x_0}{\operatorname{dist}(x_0, \partial G)} \tag{1.30}$$

satisfies (1.25).

Proof. Let $\Gamma_1, \dots, \Gamma_{d+1}$ denote the faces of the simplex K, and $\boldsymbol{n}_i$ be the unit outer normal vector on Γ_i, $i = 1, \dots, n$. The proof is based on the observation of (Dolejší et al., 2002), that if $x \in \Gamma_i$, then the scalar product $(x - x_0) \cdot \boldsymbol{n}_i$ is equal to the distance of x_0 to the hyperplane containing Γ_i. For each $x \in \Gamma_i$, we have

$$\boldsymbol{\varphi}(x) \cdot \boldsymbol{n} = \boldsymbol{\varphi}(x) \cdot \boldsymbol{n}_i = \frac{(x - x_0) \cdot \boldsymbol{n}_i}{\operatorname{dist}(x_0, \partial G)} = \frac{\operatorname{dist}(x_0, \Gamma_i)}{\operatorname{dist}(x_0, \partial G)} \geq 1. \qquad \square$$

In comparison to the construction in Lemma 1.5, the formula (1.30) is surprisingly simple. Moreover, similar formula is valid for more general domains. One can prove, that if $G \subset \mathbb{R}^d$ is bounded domain with Lipschitz continuous boundary and $x_0 \in \mathbb{R}^d$, $\rho_0 > 0$, then the following two assertions are equivalent

(i) The function $\boldsymbol{\varphi}(x) = \rho_0^{-1}(x - x_0)$ satisfies (1.25).

(ii) The domain G is star-shaped with respect to all points $y \in \mathcal{B}(x_0, \rho_0)$, where $\mathcal{B}(x_0, \rho_0)$ denotes the ball centered at x_0 with radius ρ_0, i.e. the line segment $\operatorname{conv}\{x, y\}$ lies in G for all $x \in G$ and $y \in \mathcal{B}(x_0, \rho_0)$.

In (Feng and Karakashian, 2001), the domains satisfying the assumption **(i)** are called *star-like*. We will not present full proof of the equivalence of **(i)** and **(ii)**. The main idea of the proof is as follows. Let $x \in \partial G$, let $\boldsymbol{n}$ be the outer unit normal at x and $y = x_0 + \rho_0 \boldsymbol{n} \in \mathcal{B}(x_0, \rho_0)$. If G is star-shaped with respect to y, then the line segment connecting x with y lies in G. Vaguely speaking, this means that the vector $x - y$ points outwards of the domain G, or $(x - y) \cdot \boldsymbol{n} \geq 0$. Then

$$(x - x_0) \cdot \boldsymbol{n} = (x - y) \cdot \boldsymbol{n} + \rho_0 \boldsymbol{n} \cdot \boldsymbol{n} \geq \rho_0.$$

Next, we use the auxiliary function (1.30) and Gauss theorem to prove the local multiplicative trace inequality. Our proof follow closely (Dolejší et al., 2002, Lemma 3.1), see also (Feng and Karakashian, 2001). One can prove the theorem also using a finite element scaling argument, see (Arnold, 1982).

Theorem 1.6 (Local multiplicative trace inequality)**.** *For each $p \in [1, \infty)$, there exist a constant C_M such that*

$$\|v\|^p_{L^p(\partial\Omega)} \leq C_M \left(|v|_{W^{1,p}(K)} \|v\|^{p-1}_{L^p(K)} + h_K^{-1} \|v\|^p_{L^p(K)} \right), \quad K \in \mathcal{T}_h, v \in W^{1,p}(K). \tag{1.31}$$

Proof. There exists $x_0 \in K$ such that

$$\operatorname{dist}(x_0, \partial K) = \rho_K. \tag{1.32}$$

Let $\boldsymbol{\varphi}$ be given by (1.30). Let $w \in W^{1,1}(K)$ be arbitrary nonnegative function. Then,

$$\begin{aligned}
\|w\|_{L^1(\partial K)} &\le \int_{\partial K} w\boldsymbol{\varphi}\cdot\boldsymbol{n}\,\mathrm{dS} && [\text{ by (1.25) }]\\
&= \int_K \operatorname{div}(w\boldsymbol{\varphi})\,\mathrm{d}x = \int_K (\boldsymbol{\varphi}\cdot\nabla w + w\operatorname{div}\boldsymbol{\varphi})\,\mathrm{d}x && [\text{ by Gauss theorem }]\\
&\le \|\boldsymbol{\varphi}\|_{L^\infty(K)}\,|w|_{W^{1,1}(K)} + |\boldsymbol{\varphi}|_{W^{1,\infty}(K)}\,\|w\|_{L^1(K)}\\
&\le \frac{h_K}{\rho_K}\,|w|_{W^{1,1}(K)} + \frac{d}{\rho_K}\,\|w\|_{L^1(K)} && [\text{ by (1.30) and (1.32) }]\\
&\le (1+d)C_r\left(|w|_{W^{1,1}(K)} + h_K^{-1}\,\|w\|_{L^1(K)}\right). && [\text{ by (1.15) }]
\end{aligned}$$

We conclude the proof by putting $w = |v|^p$ and using the inequality $|\nabla w| \le p|v|^{p-1}|\nabla v|$. □

Theorem 1.7 (Global multiplicative trace inequality)**.** *For each $p \in [1, \infty)$, there exists a constant C'_M such that*

$$\|v\|^p_{L^p(\partial\Omega)} \le C'_M\left(|v|_{W^{1,p}(\Omega,\mathcal{T}_h)}\left(\|v\|^p_{L^p(\Omega)} + \sum_{K\in\mathcal{T}_h} h_K\,\|v\|^p_{L^p(\partial K)}\right)^{1-1/p} + \|v\|^p_{L^p(\Omega)}\right) \tag{1.33}$$

holds for all $v \in W^{1,p}(\Omega, \mathcal{T}_h)$.

Proof. Let $\boldsymbol{\varphi}$ be a function satisfying (1.25). Let $w \in W^{1,1}(\Omega, \mathcal{T}_h)$ be arbitrary nonnegative function. Similarly to the proof of Theorem 1.6, we use Gauss theorem on each element to obtain the inequality

$$\int_{\partial K} w\boldsymbol{\varphi}\cdot\boldsymbol{n}\,\mathrm{dS} \le \|\boldsymbol{\varphi}\|_{W^{1,\infty}(K)}\left(|w|_{W^{1,1}(K)} + \|w\|_{L^1(K)}\right). \tag{1.34}$$

Summing (1.34) over all elements K, we get

$$\begin{aligned}
\|w\|_{L^1(\partial\Omega)} &\le \int_{\partial\Omega} w\boldsymbol{\varphi}\cdot\boldsymbol{n}\,\mathrm{dS} && [\text{ by (1.25) }]\\
&= \sum_{\Gamma\in\mathcal{F}_h^{\partial\Omega}} \int_\Gamma [w]\,\boldsymbol{\varphi}\cdot\boldsymbol{n}\,\mathrm{dS} && [\text{ by the definition of } [w] \text{ and } \boldsymbol{n}_\Gamma\ ]\\
&= \sum_{\Gamma\in\mathcal{F}_h} \int_\Gamma [w]\,\boldsymbol{\varphi}\cdot\boldsymbol{n}\,\mathrm{dS} - \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma [w]\,\boldsymbol{\varphi}\cdot\boldsymbol{n}\,\mathrm{dS} && \left[\text{ because } \mathcal{F}_h = \mathcal{F}_h^{\partial\Omega}\cup\mathcal{F}_h^I\ \right]\\
&= \sum_{K\in\mathcal{T}_h} \int_{\partial K} w\boldsymbol{\varphi}\cdot\boldsymbol{n}_{\partial K}\,\mathrm{dS} - \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma [w]\,\boldsymbol{\varphi}\cdot\boldsymbol{n}\,\mathrm{dS}
\end{aligned}$$

$$\leq \|\boldsymbol{\varphi}\|_{W^{1,\infty}(\Omega)} \|w\|_{W^{1,1}(\Omega,\mathcal{T}_h)} . \quad [\text{ by (1.25) and the definition of the norm (1.6) }]$$

We put $w = |v|^p$. It remains to prove that the norm $\|w\|_{W^{1,1}(\Omega,\mathcal{T}_h)}$ is bounded by the right hand side of (1.33). Again, we need the inequalities $|\nabla w| \leq p|v|^{p-1}|\nabla v|$ and

$$|[w]_\Gamma| = |v_L^p - v_R^p| = p \left| \int_{v_R}^{v_L} |s|^{p-1} \, \mathrm{d}s \right| \leq p \, |[v]_\Gamma| \max \left(|v_L|^{p-1}, |v_R|^{p-1} \right) . \tag{1.35}$$

We estimate jump term of the norm $\|w\|_{W^{1,1}(\Omega,\mathcal{T}_h)}$ by

$$\begin{aligned}\sum_{\Gamma\in\mathcal{F}_h} \|[w]\|_{L^1(\Gamma)} &\leq p \sum_{\Gamma\in\mathcal{F}_h} \|[v]\|_{L^p(\Gamma)} \max \left(\|v_L\|_{L^p(\Gamma)}^{p-1}, \|v_R\|_{L^p(\Gamma)}^{p-1} \right) \\ &\leq p \left(\sum_{\Gamma\in\mathcal{F}_h} h_\Gamma^{1-p} \|[v]\|_{L^p(\Gamma)}^p \right)^{1/p} \left(\sum_{\Gamma\in\mathcal{F}_h} h_\Gamma \max \left(\|v_L\|_{L^p(\Gamma)}^p, \|v_R\|_{L^p(\Gamma)}^p \right) \right)^{1-1/p} . \\ &\leq p \, |v|_{W^{1,p}(\Omega,\mathcal{T}_h)} \left(2 \sum_{K\in\mathcal{T}_h} h_K \|v_L\|_{L^p(\partial K)}^p \right)^{1-1/p} .\end{aligned}$$

and the remaining terms of the norm $\|w\|_{W^{1,1}(\Omega,\mathcal{T}_h)}$ by

$$\sum_{K\in\mathcal{T}_h} \|w\|_{W^{1,1}(K)} \leq \|v\|_{L^p(\Omega)}^p + p \sum_{K\in\mathcal{T}_h} \|w\|_{L^p(K)}^{p-1} |v|_{W^{1,p}(K)} .$$

□

The global multiplicative trace inequality (1.33) for $p = 2$ was used in (Dolejší et al., 2009). However, with the help of (1.31), we can easily prove a trace inequality in simpler form

$$\|v\|_{L^p(\partial\Omega)} \leq C \|v\|_{W^{1,p}(\Omega,\mathcal{T}_h)} , \qquad v \in W^{1,p}(\Omega, \mathcal{T}_h). \tag{1.36}$$

For $p = 1$, the result follows also from the imbedding $W^{1,1}(\Omega, \mathcal{T}_h) \subset BV(\Omega)$, see section 1.4. If $p > 1$, (1.36) is not optimal with respect to the function space on $\partial\Omega$. A sharper bound

$$\|v_h\|_{L^{p^\#}(\partial\Omega)} \leq C \|v_h\|_{W^{1,p}(\Omega,\mathcal{T}_h)} , \quad p^\# = \frac{p(d-1)}{d-p}, \; p \in (1, d),$$

was proved for in (Buffa and Ortner, 2009, Theorem 4.4) for piecewise polynomial functions $v_h \in \mathcal{S}_{h,k}$. The proof in (Buffa and Ortner, 2009) is based on a reconstruction operator $\mathcal{S}_{h,k} \to W^{1,p}(\Omega)$, and is different from the proof presented here.

1.3 An interpolation result

In this section, we will discuss the relationship between the broken Sobolev spaces $W^{1,q}(\Omega, \mathcal{T}_h)$ for different $q \in [1, \infty]$. We will show that, analogously to the case of classical Sobolev spaces, the spaces $W^{1,q}(\Omega, \mathcal{T}_h)$ form a scale of interpolation spaces.

Review of results on interpolation of Banach Spaces

Before discussing this topic further, let us first recall the notion of interpolation of Banach spaces, namely the so called real K-method of interpolation.

Definition 1.8. (following (Tartar, 2007, Definition 22.1)) Let X_0 and X_1 be two normed spaces, continuously imbedded into a topological vector space $\mathcal{X}$ so that

$$\begin{aligned} &X_0 \cap X_1 \text{ is equipped with the norm} \qquad &&\|x\|_{X_0\cap X_1} = \max\left(\|x\|_{X_0}, \|x\|_{X_1}\right), \\ &X_0 + X_1 \text{ is equipped with the norm} \qquad &&\|x\|_{X_0+X_1} = \inf_{x=x_0+x_1}\left(\|x\|_{X_0} + \|x\|_{X_1}\right). \end{aligned}$$

For $x \in X_0 + X_1$ and $t > 0$ one defines

$$K(t, x, X_0, X_1) = \inf_{x=x_0+x_1}\left(\|x_0\|_{X_0} + t\,\|x_1\|_{X_1}\right), \tag{1.37}$$

and for $0 < \theta < 1$ and $1 \le p \le \infty$ (or for $\theta = 0, 1$ and $p = \infty$), one writes

$$\begin{aligned} (X_0, X_1)_{\theta,p} &= \left\{x \in X_0 + X_1 : t^{-\theta}K(t, x, X_0, X_1) \in L^p(0,\infty;\, \mathrm{d}t/t)\right\}, \\ \text{with the norm } \|x\|_{(X_0,X_1)_{\theta,p}} &= \left\|t^{-\theta}K(t, x, X_0, X_1)\right\|_{L^p(0,\infty;\,\mathrm{d}t/t)}. \end{aligned} \tag{1.38}$$

The notation X_0+X_1 stands for the set of all vectors $x \in \mathcal{X}$, which can be decomposed into a sum $x_0 + x_1$, where $x_0 \in X_0$ and $x_1 \in X_1$. The infimum in (1.37) is taken over all such decompositions. The symbol $L^p(0,\infty;\, \mathrm{d}t/t)$ denotes the weighted Lebesgue space for p-integrable functions on the interval $(0,\infty)$, with the weight $t \mapsto 1/t$,

$$L^p(0,\infty;\, \mathrm{d}t/t) = \left\{ f \text{ measurable function } : \int\limits_0^\infty |f(t)|^p \frac{\mathrm{d}t}{t} < \infty \right\}.$$

The norm of (1.38) can be also written as

$$\|x\|_{(X_0,X_1)_{\theta,p}} = \begin{cases} \left(\displaystyle\int_0^\infty t^{-p\theta-1}K^p(t, x, X_0, X_1)\,\mathrm{d}t\right)^{1/p}, & 1 \le p < \infty, \\ \operatorname{ess\,sup}_{t\in(0,\infty)} t^{-\theta}K(t, x, X_0, X_1), & p = \infty. \end{cases}$$

The basic result is following interpolation property of linear operators (see (Tartar, 2007, Lemma 22.3)).

Property 1.9. *If $A : X_0 + X_1 \to Y_0 + Y_1$ is a linear operator and maps X_0 to Y_0 with*

$$\|Ax\|_{Y_0} \le M_0\,\|x\|_{X_0}, \qquad x \in X_0, \tag{1.39}$$

and maps X_1 to Y_1 with

$$\|Ax\|_{Y_1} \le M_1\,\|x\|_{X_1}, \qquad x \in X_1, \tag{1.40}$$

then A is linear continuous operator from $(X_0, X_1)_{\theta,p}$ into $(Y_0, Y_1)_{\theta,p}$ for all θ and p and

$$\|Ax\|_{(Y_0,Y_1)_{\theta,p}} \le M_0^{1-\theta}M_1^{\theta}\,\|x\|_{(X_0,X_1)_{\theta,p}}, \qquad x \in (X_0, X_1)_{\theta,p}. \tag{1.41}$$

For various choices of the spaces X_0 and X_1, the exact characterization of the interpolation space $(X_0, X_1)_{\theta,p}$ is known. For example, the interpolation space between $X_0 = L^1(\Omega)$ and $X_1 = L^\infty(\Omega)$ with parameters $p \in (1,\infty)$ and $\theta = 1 - \frac{1}{p}$ is the space $L^p(\Omega)$,

$$\big(L^1(\Omega), L^\infty(\Omega)\big)_{1-\frac{1}{p},p} = L^p(\Omega), \tag{1.42}$$

and the corresponding interpolation norm is equivalent to the norm $\|\cdot\|_{L^p(\Omega)}$. Moreover, the K-functional is

$$K(t, f; L^1(\Omega), L^\infty(\Omega)) = t f^{\star\star}(t), \qquad t > 0, \tag{1.43}$$

where

$$f^{\star\star}(t) = \frac{1}{t}\int_0^t f^\star(s)\,\mathrm{d}s, \tag{1.44}$$

$$f^\star(s) = \inf\left\{r \geq 0 : m(r,f) \leq t\right\}, \tag{1.45}$$

$$m(r,f) = \left|\{x \in \Omega : |f(x)| > r\}\right|. \tag{1.46}$$

The function $f^\star$ is the *non-increasing rearrangement* of f and $m(\cdot, f)$ is the *distribution function* of f. The function $f^\star$ is non-increasing on the interval $(0,\infty)$ and is *equi-measurable* to the function f, i.e.

$$\left|\{x \in \Omega : |f(x)| > r\}\right| = \left|\{s \in (0,\infty) : f^\star(s) > r\}\right|.$$

The function $f^{\star\star}$ is continuous and nonincreasing on $(0,\infty)$. Moreover,

$$\|f\|_{L^p(\Omega)} = \|f^\star\|_{L^p(0,\infty)} \leq \|f^{\star\star}\|_{L^p(0,\infty)} \leq \frac{p}{p-1}\|f\|_{L^p(\Omega)}, \qquad f \in L^p(\Omega). \tag{1.47}$$

See (Bergh and Löfström, 1976, sections 1.3 and 5.2) or (Adams and Fournier, 2003, Corollary 7.27) for proof and further properties.

A result similar to (1.42) was proved for Sobolev spaces in (DeVore and Scherer, 1979),

$$W^{k,p}(\Omega) = \big(W^{k,1}(\Omega), W^{k,\infty}(\Omega)\big)_{1-\frac{1}{p},p}, \qquad k = 1, 2, \ldots,\ p \in (1,\infty), \tag{1.48}$$

with the aid of a characterization of the K-functional:

$$C_1 t \sum_{|\alpha|\leq k} (D^\alpha u)^{\star\star}(t) \leq K(t, u, W^{k,1}(\Omega), W^{k,\infty}(\Omega)) \leq C_2 t \sum_{|\alpha|\leq k} (D^\alpha u)^{\star\star}(t), \tag{1.49}$$

for all $t > 0$, where $D^\alpha = \frac{\partial^{\alpha_1+\cdots+\alpha_d}}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}$ denotes the partial derivative of the order given by the multiindex $\alpha = (\alpha_1, \ldots, \alpha_d)$.

Proof of the interpolation theorem

We shall generalize (1.48) to the broken Sobolev spaces.

Theorem 1.10. *For each $p \in (1, \infty)$,*

$$W^{1,p}(\Omega, \mathcal{T}_h) = \big(W^{1,1}(\Omega, \mathcal{T}_h), W^{1,\infty}(\Omega, \mathcal{T}_h)\big)_{1-\frac{1}{p},p} \tag{1.50}$$

with equivalent norms,

$$C_1 \|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)} \leq \|u\|_{(W^{1,1}(\Omega,\mathcal{T}_h),W^{1,\infty}(\Omega,\mathcal{T}_h))_{1-\frac{1}{p},p}} \leq C_2 \|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)}, \quad u \in W^{1,p}(\Omega).$$

The constants C_1, C_2 do not depend on h.

The proof is given in a sequence of lemmas. The proof can be summarized in three steps:

- **Step 1.** For each $u \in W^{1,p}(\Omega, \mathcal{T}_h)$, we define an auxiliary function $g_u \in L^p(\Omega)$, which represents the magnitude of u, and the derivative of u, including both the elementwise part ∇u and the jump part $[u]$. We define an auxiliary norm $\|\cdot\|_{h,p}$, equivalent to the norm $\|\cdot\|_{W^{1,p}(\Omega,\mathcal{T}_h)}$, and show that

$$\|u\|_{h,p} = \|g_u\|_{L^p(\Omega)}.$$

  This step is covered by Lemmas 1.11 – 1.17. The definition of g_u is given in (1.85).

- **Step 2.** We estimate the K-functional from above using the nonincreasing rearrangement $g_u^\star$ of the auxiliary function g_u. To this end, we construct a suitable decomposition $u = v + w$ for each $u \in W^{1,p}(\Omega, \mathcal{T}_h)$, depending on the parameter $t > 0$. This part of the proof is presented in Lemmas 1.18 – 1.20.

- **Step 3.** We prove the corresponding estimate of the K-functional from below (Lemma 1.21). We establish an inequality analogous to (1.49).

In this section, we denote the nonincreasing rearrangement of arbitrary function $f \in L^1(\Omega)$ by $f^\star$. We also use the notation (1.44)-(1.46).

Lemma 1.11. ***(i)*** *For each $c_D > 0$, there exists $c'_D > 0$ such that for all $K, L \in \mathcal{T}_h$, $K \neq L$, and $x \in K$, the following implication holds:*

$$\operatorname{dist}(x, \partial K) \geq c_D h_K \implies \operatorname{dist}(x, L) \geq c'_D h_L. \tag{1.51}$$

(ii) *There exists a constant c such that*

$$\operatorname{dist}\left(K, \Omega \setminus \bigcup_{L \in \mathcal{D}(K)} L\right) \geq c h_K, \qquad K \in \mathcal{T}_h. \tag{1.52}$$

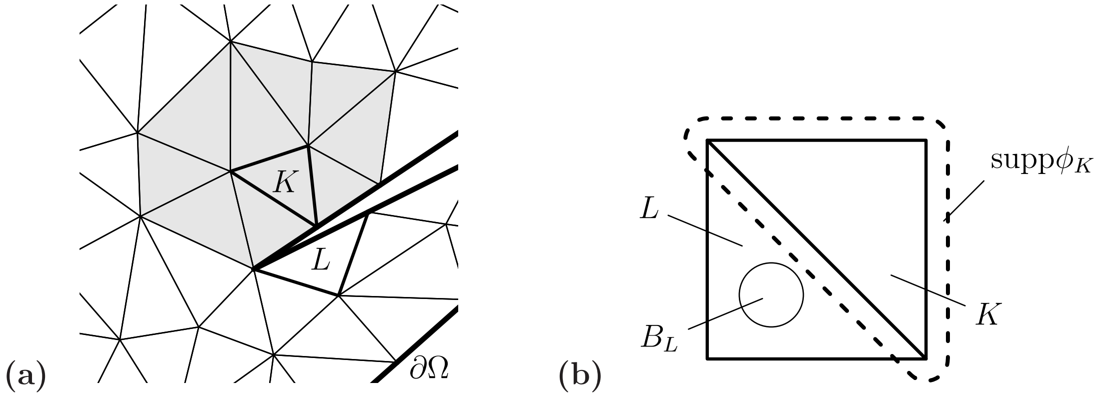

Figure 1.1: Illustration for proofs of Lemma 1.11 and Lemma 1.13. **(a)** The set $\mathcal{D}(K)$ and an element $L \notin \mathcal{D}(K)$. **(b)** Relation of $\operatorname{supp}\phi_K = \operatorname{supp}\psi_K$ with B_L for two elements $K \neq L$.

Proof. Let us prove **(i)**. Let $L, K \in \mathcal{T}_h$, $L \neq K$, be arbitrary elements. Note that in general, the elements L and K are not neighbors. Let $a_K^1, \dots, a_K^{d+1} \in \mathbb{R}^d$ be the vertices of the simplex K. Every $x \in K$ is a convex combination of the vertices,

$$x = \sum_{i=1}^{d+1} \lambda_i a_K^i,$$

where the coefficients λ_i are the *barycentric coordinates* of the point x, satisfying (1.13). First we prove following implication

$$x \in K,\ \operatorname{dist}(x, \partial K) \geq c_D h_K \implies \lambda_i \geq c_D,\ i = 1, \dots, d+1. \tag{1.53}$$

Fix $x \in K$, $\operatorname{dist}(x, \partial K) \geq c_D h_K$ and $i = 1, \dots, d+1$. Let $x' \in K$ be a point with barycentric coordinates

$$\lambda'_j = \begin{cases} 0, & j = i, \\ \frac{\lambda_j}{1-\lambda_i}, & j \neq i. \end{cases}$$

Then

$$x - x' = \sum_{j=1}^{d+1} (\lambda_j - \lambda'_j) a_K^j = \sum_{j=1}^{d+1} (\lambda_j - \lambda'_j)(a_K^j - a_K^i) = -\sum_{\substack{j=1 \\ j \neq i}}^{d+1} \frac{\lambda_i \lambda_j}{1 - \lambda_i} (a_K^j - a_K^i).$$

$$|x - x'| \leq h_K \sum_{\substack{j=1 \\ j \neq i}}^{d+1} \frac{\lambda_i \lambda_j}{1 - \lambda_i} = h_K \lambda_i.$$

Since $x' \in \partial K$, we have $c_D h_K \leq |x - x'| \leq h_K \lambda_i$, and $c_D \leq \lambda_i$. This proves (1.53)

Since $\mathcal{T}_h$ is conforming, there exists a piecewise-linear function $\chi_L \in \mathcal{S}_{h,1}$ such that $\chi_L(x) = 1$, for $x \in L$, and $\chi_L(x) = 0$, for $x \in \Omega \setminus \bigcup_{K' \in \mathcal{D}(L)} K'$. Obviously $\|\chi_L\|_{W^{1,\infty}(\Omega)} \leq C_1 h_L^{-1}$, where C_1 depends only on C_r. In virtue of the Lipschitz continuity of $\partial\Omega$,

$$|\chi_L(x) - \chi_L(y)| \leq C_2 \|\chi_K\|_{W^{1,\infty}(\Omega)} |x-y| \leq C_1 C_2 h_L^{-1} |x-y|, \qquad x, y \in \Omega.$$

where C_2 depends only on Ω. Now, let $x \in K$ be an arbitrary, and $\operatorname{dist}(x, \partial K) \geq c_D h_K$. Since $K \neq L$, $a_K^j \notin L$ for some index j. By the definition of χ_L and (1.53),

$$\chi_L(x) = \sum_{i=1}^{d+1} \lambda_i \chi_L(a_K^i) = \sum_{\substack{i=1 \\ i \neq j}}^{d+1} \lambda_i \chi_L(a_K^i) \leq \sum_{\substack{i=1 \\ i \neq j}}^{d+1} \lambda_i = 1 - \lambda_j \leq 1 - c_D.$$

Let $y \in L$ be arbitrary. Then $\chi_L(y) = 1$ and

$$|y - x| \geq C_1^{-1} C_2^{-1} h_L |\chi_L(y) - \chi_L(x)| = C_1^{-1} C_2^{-1} h_L (1 - \chi_L(x)) \geq C_1^{-1} C_2^{-1} c_D h_L.$$

Therefore, (1.51) holds with $c_D' = C_1^{-1} C_2^{-1} c_D$.

Let us prove **(ii)**. Let $K \in \mathcal{T}_h$ and $y \in \Omega \setminus \bigcup_{K' \in \mathcal{D}(K)} K'$ be arbitrary. There exists an element $L \in \mathcal{T}_h \setminus \mathcal{D}(K)$, such that $y \in L$ (see Fig. 1.1(a)). Let χ_L be defined as in the first part of the proof. Let $x \in K$ be arbitrary. Then

$$|x - y| \geq C_1^{-1} C_2^{-1} h_L |\chi_L(x) - \chi_L(y)| = C_1^{-1} C_2^{-1} h_L.$$

The assertion (1.52) holds with $c = C_1^{-1} C_2^{-1}$. □

Definition 1.12. Let $K \in \mathcal{T}_h$ and $x_K \in K$ be the center of the ball $\mathcal{B}(x_K, \rho_K) \subset K$ inscribed into K with maximal radius ρ_K. We set

$$B_K = \mathcal{B}(x_K, \rho_K/2). \tag{1.54}$$

Lemma 1.13. *For each $K \in \mathcal{T}_h$, there exists a Lipschitz continuous function ψ_K such that*

$$\psi_K(x) = 1, \qquad x \in B_K,\ K \in \mathcal{T}_h, \tag{1.55}$$

$$\overline{\Omega} \cap \operatorname{supp} \psi_K \subset \bigcup_{K' \in \mathcal{D}(K)} K', \qquad K \in \mathcal{T}_h, \tag{1.56}$$

$$|\nabla \psi_K(x)| \leq C h_K^{-1}, \qquad \text{for a. a. } x \in \Omega,\ K \in \mathcal{T}_h. \tag{1.57}$$

Moreover, the system $\{\psi_K\}_{K \in \mathcal{T}_h}$ is a partition of unity on Ω,

$$0 \leq \psi_K(x) \leq 1, \qquad \sum_{L \in \mathcal{T}_h} \psi_L(x) = 1, \qquad x \in \Omega,\ K \in \mathcal{T}_h. \tag{1.58}$$

Proof. Let $c_D = C_r^{-1}/2$ and let c'_D be given by Lemma 1.11. We set

$$\phi_K(x) = \operatorname{dist}\left(x, \left\{y \in \mathbb{R}^d : \operatorname{dist}(y, K) \geq c'_D h_K\right\}\right), \qquad x \in \mathbb{R}^d,\ K \in \mathcal{T}_h. \tag{1.59}$$

The function ϕ_K is Lipschitz continuous, with the Lipschitz constant equal to 1. Moreover, for each $K, L \in \mathcal{T}_h$, $L \neq K$ and $x \in B_L$, we have $\operatorname{dist}(x, \partial L) \geq \rho_L/2 \geq c_D h_L$, $\operatorname{dist}(x, K) \geq c'_D h_K$ by (1.51) and $\phi_K(x) = 0$ by (1.59), see Fig. 1.1(b). By (1.52), $\phi_K(x) = 0$ on the set $\Omega \setminus \bigcup_{K' \in \mathcal{D}(K)} K'$. On the other hand, for $x \in K$, we have $\phi_K(x) > 0$.

We set

$$\psi_K(x) = \frac{\phi_K(x)}{\sum_{L \in \mathcal{T}_h} \phi_L(x)}. \tag{1.60}$$

The denominator is positive on Ω. The properties (1.55), (1.56), (1.58) follow from the properties of ϕ_K stated above.

Let us prove (1.57). Let $x \in \mathbb{R}^d$ be arbitrary. If $\operatorname{dist}(x, K) > c'_D h_K$, then $\nabla \psi_K = 0$. Suppose $\operatorname{dist}(x, K) < c'_D h_K$. By (1.52), $x \in L$ for some $L \in \mathcal{D}(K)$. Then

$$\begin{aligned}
|\nabla \psi_K(x)| &\leq \frac{|\nabla \phi_K(x)|}{\sum_{R \in \mathcal{T}_h} \phi_R(x)} + \frac{\phi_K(x)}{\left(\sum_{R \in \mathcal{T}_h} \phi_R(x)\right)^2} \sum_{R \in \mathcal{T}_h} |\nabla \phi_R(x)| && [\text{ by differentiating (1.60) }] \\
&\leq 2 \frac{\sum_{R \in \mathcal{T}_h} |\nabla \phi_R(x)|}{\sum_{R \in \mathcal{T}_h} \phi_R(x)} \leq 2 \frac{\operatorname{card}\{R : \operatorname{supp} \phi_R \ni x\}}{\phi_L(x)} && [\text{ using } |\nabla \phi_R| \leq 1 \ ] \\
&\leq 2 \frac{\operatorname{card} \mathcal{D}(L)}{c'_D h_L} \leq C h_K^{-1}. && [\text{ using (1.19) }]
\end{aligned}$$

□

Definition 1.14. We define a projection operator $P_h : L^1(\Omega) \to \mathcal{S}_{h,0}$, a reconstruction operator $R_h : \mathcal{S}_{h,0} \to W^{1,\infty}(R^d)$ and a variation operator $G_h : \mathcal{S}_{h,0} \to \mathcal{S}_{h,0}$ by

$$(P_h u)\big|_K = \frac{1}{|B_K|} \int_{B_K} u(x)\, \mathrm{d}x, \qquad K \in \mathcal{T}_h,\ u \in L^1(\Omega), \tag{1.61}$$

$$(R_h u_h)(x) = \sum_{K \in \mathcal{T}_h} u_{h,K} \psi_K(x), \qquad x \in \Omega,\ u_h \in \mathcal{S}_{h,0}, \tag{1.62}$$

$$(G_h(u_h))\big|_K = h_K^{-1} \max_{L \in \mathcal{D}(K)} |u_{h,L} - u_{h,K}|, \qquad K \in \mathcal{T}_h,\ u_h \in \mathcal{S}_{h,0}. \tag{1.63}$$

where $u_{h,K} = u_h\big|_K$ denotes the constant value of $u_h \in \mathcal{S}_{h,0}$ on K, B_K is given by (1.54) and ψ_K is the function defined in Lemma 1.13. Moreover, for each $p \in [1, \infty]$ and $u \in W^{1,p}(\Omega, \mathcal{T}_h)$, we define a norm $\|\cdot\|_{h,p}$ by

$$\|u\|_{h,p}^p = \|P_h u\|_{L^p(\Omega)}^p + \|G_h(P_h u)\|_{L^p(\Omega)}^p + \sum_{K \in \mathcal{T}_h} \left(h_K^{-1} \|u - P_h u\|_{L^p(K)}^p + |u|_{W^{1,p}(K)}^p\right), \tag{1.64}$$

if $p < \infty$, and

$$\|u\|_{h,p} = \max\left\{\|P_h u\|_{L^\infty(\Omega)}, \|G_h(P_h u)\|_{L^\infty(\Omega)}, \max_{K \in \mathcal{T}_h} \max\left(h_K^{-1} \|u - P_h u\|_{L^\infty(K)}, |u|_{W^{1,\infty}(K)}\right)\right\} \tag{1.65}$$

if $p = \infty$.

Lemma 1.15. *For each $p \in [1, \infty]$, there exists a constant C such that*

$$\|P_h u\|_{L^p(\Omega)} \le C \|u\|_{L^p(\Omega)}\,, \qquad u \in L^p(\Omega), \tag{1.66}$$

$$\|G_h(P_h u)\|_{L^p(\Omega)} \le C |u|_{W^{1,p}(\Omega,\mathcal{T}_h)}\,, \qquad u \in W^{1,p}(\Omega,\mathcal{T}_h), \tag{1.67}$$

$$\|u - P_h u\|_{L^p(K)} \le C h_K |u|_{W^{1,p}(K)}\,, \qquad u \in W^{1,p}(\Omega,\mathcal{T}_h),\ K \in \mathcal{T}_h. \tag{1.68}$$

Proof. The inequality (1.66) follows from the definition (1.61) by Hölder inequality.

Let us prove (1.67). For each element $K \in \mathcal{T}_h$ and each face $\Gamma \subset \partial K$, we define a functional $F_{K,\Gamma} : W^{1,1}(K) \to \mathbb{R}$ by

$$F_{K,\Gamma}(v) = \frac{1}{|B_K|}\int_{B_K} v \,\mathrm{d}x - \frac{1}{|\Gamma|}\int_{\Gamma} v \,\mathrm{dS}, \qquad v \in W^{1,1}(K). \tag{1.69}$$

By Property 1.4, for each $w \in W^{1,r}(K)$ there exists $q_w \in P^0(K)$ such that

$$|w - q_w|_{W^{k,r}(K)} \le C_1 h_K^{1-k} |w|_{W^{k,r}(K)}\,, \qquad k = 0, 1,\ r \in [1, \infty], \tag{1.70}$$

where C_1 depends only on the dimension d and the mesh regularity constant C_r. We have the estimate

$$\begin{aligned}
|F_{K,\Gamma}(v)| &= |F_{K,\Gamma}(v - q_v)| && [\ F_{K,\Gamma}(q_v) = 0\ ]\\
&\le |B_K|^{-1} \|v - q_v\|_{L^1(K)} + |\Gamma|^{-1} \|v - q_v\|_{L^1(\Gamma)} && [\ \text{by (1.69)}\ ]\\
&\le (|B_K|^{-1} + C_M h_K^{-1}|\Gamma|^{-1}) \|v - q_v\|_{L^1(K)} + C_M|\Gamma|^{-1} |v - q_v|_{W^{1,1}(K)} && [\ \text{by (1.31)}\ ]\\
&\le C_2 h_\Gamma^{1-d} |v|_{W^{1,1}(K)}\,, && [\ \text{by (1.70)}\ ]
\end{aligned}$$

where C_2 depends on C_1, C_M and C_r. Let $u \in W^{1,p}(\Omega,\mathcal{T}_h)$ and $\Gamma \in \mathcal{F}_h^I$ be arbitrary. Then

$$\begin{aligned}
\left|(P_h u)\big|_{K_\Gamma^{\mathcal{L}}} - (P_h u)\big|_{K_\Gamma^{\mathcal{R}}}\right| &\le |F_{K_\Gamma^{\mathcal{L}},\Gamma}(u) - F_{K_\Gamma^{\mathcal{R}},\Gamma}(u)| + |\Gamma|^{-1} \|[u]\|_{L^1(\Gamma)}\\
&\le C_3 h_\Gamma^{1-d} \left(|u|_{W^{1,1}(K_\Gamma^{\mathcal{L}})} + |u|_{W^{1,1}(K_\Gamma^{\mathcal{R}})} + \|[u]\|_{L^1(\Gamma)}\right).
\end{aligned}$$

For all $K \in \mathcal{T}_h$, we get the estimate

$$G_h(P_h u)\big|_K \le C_4 h_K^{-d} \left(\sum_{L \in \mathcal{D}(K)} |u|_{W^{1,1}(L)} + \sum_{L \in \mathcal{D}(K)} \sum_{\Gamma \subset \partial L} \|[u]\|_{L^1(\Gamma)} \right).$$

Using Hölder inequality, we obtain

$$G_h(P_h u)\big|_K \le C_5 h_K^{-\frac{d}{p}} \left(\sum_{L \in \mathcal{D}(K)} |u|^p_{W^{1,p}(L)} + \sum_{L \in \mathcal{D}(K)} \sum_{\Gamma \subset \partial L} h_\Gamma \left\|h_\Gamma^{-1} [u]\right\|^p_{L^p(\Gamma)} \right)^{1/p}, \quad \text{for } p < \infty,$$

$$G_h(P_h u)\big|_K \le C_5 \left(\sum_{L \in \mathcal{D}(K)} |u|_{W^{1,\infty}(L)} + \sum_{L \in \mathcal{D}(K)} \sum_{\Gamma \subset \partial L} \left\|h_\Gamma^{-1} [u]\right\|_{L^\infty(\Gamma)} \right), \qquad \text{for } p = \infty.$$

Now (1.67) follows immediately.

The inequality (1.68) follows from (1.70) and the fact $P_h(q) = 0$ for all constant functions q. □

Lemma 1.16. *For each $p \in [1, \infty]$, the norms $\|\cdot\|_{W^{1,p}(\Omega,\mathcal{T}_h)}$ and $\|\cdot\|_{h,p}$ are equivalent.*

Proof. The inequality $\|u\|_{h,p} \le C \|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)}$ is immediate consequence of (1.66)-(1.67). The inequality $\|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)} \le C \|u\|_{h,p}$ follows from

$$\|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)} \le \|u - P_h u\|_{W^{1,p}(\Omega,\mathcal{T}_h)} + \|P_h u\|_{W^{1,p}(\Omega,\mathcal{T}_h)}$$

using multiplicative trace inequality (1.31) and (1.66)-(1.68). □

Lemma 1.17. *There exists a constant C such that*

$$\|R_h u_h\|_{L^\infty(K)} \le \max_{L\in\mathcal{D}(K)} \|u_h\|_{L^\infty(L)}\,, \qquad K \in \mathcal{T}_h, \tag{1.71}$$

$$\|R_h u_h\|_{W^{1,\infty}(K)} \le C \|G_h(u_h)\|_{L^\infty(K)}\,, \qquad K \in \mathcal{T}_h, \tag{1.72}$$

$$\|R_h u_h\|_{W^{1,p}(\Omega)} \le C \|u_h\|_{W^{1,p}(\Omega,\mathcal{T}_h)}\,, \qquad p \in [1,\infty], \tag{1.73}$$

for all $u_h \in \mathcal{S}_{h,0}$.

Proof. Bounds (1.71) and (1.72) follow from the definition (1.62) of R_h using the properties (1.55)-(1.58) of the functions ψ_K. The last inequality (1.73) follows from (1.71), (1.72) and Lemma 1.16. □

Lemma 1.18. *There exists a constant C such that*

$$\inf_{\substack{(v_h,w_h)\in\mathcal{S}_{h,0}\\ v_h+w_h=u_h}} \Big(\|v_h\|_{h,1} + t\,\|w_h\|_{h,\infty}\Big) \le C\,t\,g_h^{\star\star}(t), \qquad t \in (0,\infty),\ u_h \in \mathcal{S}_{h,0}, \tag{1.74}$$

where $g_h(x) = |u_h(x)| + G_h(u_h)(x)$, $x \in \Omega$.

Proof. Let $t_0, \varepsilon > 0$ be arbitrary. Put $u = R_h u_h$. From (1.49) there exist functions $v \in W^{1,1}(\Omega)$ and $w \in W^{1,\infty}(\Omega)$ such that

$$C_1 t_0 g^{\star\star}(t_0) \le \|v\|_{W^{1,1}(\Omega)} + t_0 \|w\|_{W^{1,\infty}(\Omega)} \le C_2 t_0 g^{\star\star}(t_0) + \varepsilon, \tag{1.75}$$

where $g(x) = |u(x)| + |\nabla u(x)|$. Let $v_h = P_h v$ and $w_h = P_h w$. Using (1.71), (1.72), we obtain the estimate

$$g(x) \le C g_h(x), \qquad x \in \Omega. \tag{1.76}$$

By (1.62) and (1.55),

$$v_h + w_h = P_h v + P_h w = P_h(v + w) = P_h u = P_h R_h u = u_h.$$

Moreover,

$$\|v_h\|_{h,1} + t_0 \|w_h\|_{h,\infty} = \|v_h\|_{L^1(\Omega)} + \|G_h(v_h)\|_{L^1(\Omega)}$$

$$
\begin{aligned}
&+ t_0 \max\left\{\|w_h\|_{L^\infty(\Omega)}, \|G_h(w_h)\|_{L^\infty(\Omega)}\right\} && [\text{ since } v_h, w_h \in \mathcal{S}_{h,0} \text{ }] \\
&\le C\left(\|v\|_{W^{1,1}(\Omega)} + t_0 \|w\|_{W^{1,\infty}(\Omega)}\right) && [\text{ by (1.66), (1.67) }] \\
&\le C\, t_0\, g^{\star\star}(t_0) + C\varepsilon && [\text{ by (1.75) }] \\
&\le C\, t_0\, g_h^{\star\star}(t_0) + C\varepsilon. && [\text{ by (1.76) }]
\end{aligned}
$$

Since t_0 and ε were arbitrary, the proof of (1.74) is thus finished. □

Lemma 1.19. *For each $K \in \mathcal{T}_h$, $u \in W^{1,1}(K)$, and $t > 0$ there exist $v \in W^{1,1}(K)$ and $w \in W^{1,\infty}(K)$ such that $u = v + w$, and*

$$
\|v\|_{L^1(K)} + h_K |v|_{W^{1,1}(K)} + t \max\left(\|w\|_{L^\infty(K)}, h_K |w|_{W^{1,\infty}(K)}\right) \le C t g^{\star\star}_{K,u}(t), \qquad t > 0, \tag{1.77}
$$

where $g_{K,u}(x) = |u(x)| + h_K |\nabla u(x)|$ and C does not depend on h, u, K.

Proof. Let $\hat{K}$ be a fixed simplex in $\mathbb{R}^d$. There exists an affine mapping $F_K : \mathbb{R}^d \to \mathbb{R}^d$ such that $F_K(\hat{K}) = K$. We put $\hat{u} = u \circ F_K^{-1}$. Let $t, \varepsilon > 0$ be arbitrary. Let

$$
\hat{t} = \frac{|\hat{K}|}{|K|} t.
$$

By (1.49), there exist $\hat{v} \in W^{1,1}(\hat{K})$ and $\hat{w} \in W^{1,\infty}(\hat{K})$ such that

$$
\|\hat{v}\|_{W^{1,1}(\hat{K})} + \hat{t}\, \|\hat{w}\|_{W^{1,\infty}(\hat{K})} \le C \hat{t} \hat{g}(\hat{t})^{\star\star} + \varepsilon, \qquad s > 0, \tag{1.78}
$$

where

$$
\hat{g}(\hat{x}) = |\hat{u}(\hat{x})| + |\nabla \hat{u}(\hat{x})|, \qquad \hat{x} \in \hat{K}. \tag{1.79}
$$

We put $v = \hat{v} \circ F_K$ and $w = \hat{w} \circ F_K$. Using the shape regularity (1.15), standard scaling argument gives us

$$
h_K^{-d} \|v\|_{L^1(K)} + h_K^{1-d} |v|_{W^{1,1}(K)} \le C \|\hat{v}\|_{W^{1,1}(\hat{K})}, \tag{1.80}
$$

$$
\max\left(\|w\|_{L^\infty(K)}, h_K |w|_{W^{1,\infty}(K)}\right) \le C \|\hat{w}\|_{W^{1,\infty}(\hat{K})}. \tag{1.81}
$$

Note that if $f \in L^1(K)$ and $\hat{f} = f \circ F_K^{-1}$, then the corresponding distribution and rearrangement function (recall the definitions (1.44)-(1.46)) satisfy

$$
m(r, f) = \frac{|K|}{|\hat{K}|} m(r, f), \;\; r \ge 0, \qquad \hat{f}^\star(\hat{t}) = f^\star(t), \qquad \hat{f}^{\star\star}(\hat{t}) = f^{\star\star}(t). \tag{1.82}
$$

In consequence

$$
\hat{g}^{\star\star}(t) \le C g^{\star\star}(t). \tag{1.83}
$$

The assertion (1.77) follows immediately from (1.78)-(1.81) and (1.83). □

Lemma 1.20. *For each $u \in W^{1,1}(\Omega, \mathcal{T}_h)$ and $t > 0$ there exist $v \in W^{1,1}(\Omega, \mathcal{T}_h)$ and $w \in W^{1,\infty}(\Omega, \mathcal{T}_h)$ such that $u = v + w$ and*

$$\|v\|_{h,1} + t\,\|w\|_{h,\infty} \leq C\,t\,g_u^{\star\star}(t), \tag{1.84}$$

where C does not depend on h, u, t and $g_u \in W^{1,1}(\Omega, \mathcal{T}_h)$ is defined by

$$g_u\big|_K(x) = |P_h u(x)| + G_h(P_h u)(x) + h_K^{-1}|u(x) - P_h u(x)| + |\nabla u(x)|. \tag{1.85}$$

Proof. Let $t_0 > 0$ and $u \in W^{1,1}(\Omega, \mathcal{T}_h)$ be arbitrary. We set $u^0 = P_h u$, $u^1 = u - u^0$,

$$\begin{aligned} g^0(x) &= |P_h u(x)| + G_h(P_h u)(x), && x \in \Omega, \\ g_K^1(x) &= h_K^{-1}|u(x) - P_h u(x)| + |\nabla u(x)|, && x \in K,\ K \in \mathcal{T}_h. \end{aligned}$$

Let $g^1 \in W^{1,1}(\Omega, \mathcal{T}_h)$ be such a function that $g^1\big|_K = g_K^1$ for all $K \in \mathcal{T}_h$. Let $u_K^1 = u^1\big|_K$.

For each element K, we define real number $t_K \geq 0$ and functions v_K^1, w_K^1 in the following way:

(a) If $\operatorname{ess\,sup}_{x\in K} g_K^1 \leq g^{1\star\star}(t_0)$, we set $t_K = 0$, $v_K^1 = 0$, $w_K^1 = u_K^1$.

(b) Otherwise, we can find $t_K > 0$ such that $g^{1\star\star}(t_0) = g_K^{1\star\star}(t_K)$. According to Lemma 1.19 there exists a decomposition $u_K^1 = \tilde{v}_K^1 + \tilde{w}_K^1$ such that

$$h_K^{-1}\left\|\tilde{v}_K^1\right\|_{L^1(K)} + \left|\tilde{v}_K^1\right|_{W^{1,1}(K)} + t_K \max\left(h_K^{-1}\left\|\tilde{w}_K^1\right\|_{L^\infty(K)}, \left|\tilde{w}_K^1\right|_{W^{1,\infty}(K)}\right) \leq C t_K g_K^{1\star\star}(t_K).$$

We set

$$v_K^1(x) = \tilde{v}_K^1(x) - \frac{1}{|B_K|}\int_{B_K} \tilde{v}_K^1(y)\,\mathrm{d}y, \tag{1.86}$$

$$w_K^1(x) = \tilde{w}_K^1(x) - \frac{1}{|B_K|}\int_{B_K} \tilde{w}_K^1(y)\,\mathrm{d}y, \qquad x \in K. \tag{1.87}$$

Since $P_h u^1 = 0$, we have

$$v_K^1(x) + w_K^1(x) = u_K^1(x) - \frac{1}{|B_K|}\int_{B_K} u_K^1(y)\,\mathrm{d}y = u_K^1(x).$$

Moreover, $\|v_K^1\|_{L^1(K)} \leq 2\,\|\tilde{v}_K^1\|_{L^1(K)}$, $\|w_K^1\|_{L^\infty(K)} \leq 2\,\|\tilde{w}_K^1\|_{L^\infty(K)}$, and

$$h_K^{-1}\left\|v_K^1\right\|_{L^1(K)} + \left|v_K^1\right|_{W^{1,1}(K)} + t_K \max\left(h_K^{-1}\left\|w_K^1\right\|_{L^\infty(K)}, \left|w_K^1\right|_{W^{1,\infty}(K)}\right) \leq C t_K g_K^{1\star\star}(t_K).$$

We define $v^1 \in W^{1,1}(\Omega, \mathcal{T}_h)$ and $w^1 \in W^{1,\infty}(\Omega, \mathcal{T}_h)$ such that $v^1\big|_K = v^1$, $w^1\big|_K = w^1$. Note $P_h v^1 = P_h w^1 = 0$.

We claim that

$$\sigma := \sum_{K\in\mathcal{T}_h} t_K \leq t_0. \tag{1.88}$$

The inequality (1.88) holds in the case $\sigma = 0$. Let us assume $\sigma > 0$. From the definition of t_K and $g^{1\star\star}$, we get

$$t_K g^{1\star\star}(t_0) = t_K g_K^{1\star\star}(t_K) = \int_0^{t_K} g_K^{1\star}(\tau)\,\mathrm{d}\tau. \tag{1.89}$$

By summing (1.89) over $K \in \mathcal{T}_h$, we get (using the fact that $g^{1\star}$ is nonincreasing rearrangement) the inequality

$$\sigma g^{1\star\star}(t_0) = \sum_{K\in\mathcal{T}_h} \int_0^{t_K} g_K^{1\star}(\tau)\,\mathrm{d}\tau \le \int_0^{\sigma} g^{1\star}(\tau)\,\mathrm{d}\tau = \sigma g^{1\star\star}(\sigma).$$

We have $g^{1\star\star}(t_0) \le g^{1\star\star}(\sigma)$. The function $g^{1\star\star}$ is nonincreasing, therefore $t_0 \ge \sigma$. The inequality (1.88) is proven.

We estimate

$$\begin{aligned}
\left\|v^1\right\|_{h,1} &= \sum_{K\in\mathcal{T}_h} \left(h_K^{-1}\left\|v_K^1\right\|_{L^1(K)} + \left|v_K^1\right|_{L^1(K)}\right) && \left[\text{ using } P_h v^1 = 0\ \right]\\
&\le C \sum_{K\in\mathcal{T}_h} t_K g_K^{1\star\star}(t_K) && \left[\text{ by the definition of } v_K^1\ \right]\\
&= C \sum_{K\in\mathcal{T}_h} t_K g^{1\star\star}(t_0) && \left[\text{ by the definition of } t_K\ \right]\\
&\le C t_0 g^{1\star\star}(t_0). && \left[\text{ by (1.88) }\right]
\end{aligned}$$

From the definition of w_K^1, we get

$$\left\|w^1\right\|_{h,\infty} = \max_{K\in\mathcal{T}_h} \max\left(h_K^{-1}\left\|w_K^1\right\|_{L^\infty(K)}, \left\|\nabla w_K^1\right\|_{L^\infty(K)}\right) \le g_K^{1\star\star}(t_0).$$

From Lemma 1.18, we get the decomposition $u^0 = v^0 + w^0$,

$$\left\|v^0\right\|_{h,1} + t_0\left\|w^0\right\|_{h,\infty} \le C\, t_0\, g^{0\star\star}(t_0).$$

We conclude the proof by setting $v = v^0 + v^1$, $w = w^0 + w^1$. □

Lemma 1.21. *There exists a constant C such that*

$$t\, g_u^{\star\star}(t) \le C \inf_{u=v+w}\left(\|v\|_{h,1} + t\,\|w\|_{h,\infty}\right), \tag{1.90}$$

for all $u \in W^{1,1}(\Omega,\mathcal{T}_h)$ and $t > 0$, where g_u is defined by (1.85)*. The infimum is taken over all decompositions $u = v + w$, with $v \in W^{1,1}(\Omega,\mathcal{T}_h)$ and $w \in W^{1,\infty}(\Omega,\mathcal{T}_h)$.*

Proof. Let $v \in W^{1,1}(\Omega,\mathcal{T}_h)$ and $w \in W^{1,\infty}(\Omega,\mathcal{T}_h)$ be arbitrary functions such that $u = v+w$ holds. Let g_v, g_w, respectively be defined by (1.85) with u replaced by v, w, respectively. Then

$$|g_u(x) - g_v(x)| \le C|g_w(x)|, \qquad x \in \Omega, \tag{1.91}$$

with a constant C independent of u, v, w, h, and

$$\begin{aligned} t\, g_u^{\star\star}(t) &= K\left(t, g_u, L^1(\Omega), L^\infty(\Omega)\right) && [\text{ by } (1.43)\ ] \\ &\le \|g_v\|_{L^1}(\Omega) + t\, \|g_u - g_v\|_{L^\infty(\Omega)} && [\text{ by } (1.37)\ ] \\ &\le \|g_v\|_{L^1}(\Omega) + Ct\, \|g_w\|_{L^\infty(\Omega)} && [\text{ by } (1.91)\ ] \\ &\le C\left(\|v\|_{h,1} + t\, \|w\|_{h,\infty}\right). && [\text{ by } (1.64),\ (1.65)\ ] \end{aligned}$$

The inequality (1.90) follows by taking infimum over all decompositions $u = v + w$. □

Lemma 1.21, together with Lemma 1.20, give a similar characterization of the K-functional for interpolation between $W^{1,1}(\Omega, \mathcal{T}_h)$ and $W^{1,\infty}(\Omega, \mathcal{T}_h)$, as in (1.49). The proof of Theorem 1.10 is finished by following chain of norm equivalences:

$$\begin{aligned} \|u\|_{(W^{1,1}(\Omega,\mathcal{T}_h), W^{1,p}(\Omega,\mathcal{T}_h))_{1-\frac{1}{p},p}} &\sim \|g_u^{\star\star}\|_{L^p(0,\infty)} && [\text{ by Lemmas 1.20, 1.21 }] \\ &\sim \|g_u\|_{L^p(\Omega)} && [\text{ by (1.47) }] \\ &\sim \|u\|_{h,p} && [\text{ by definition (1.85) of } g_u\ ] \\ &\sim \|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)} && [\text{ by Lemma 1.16 }]. \end{aligned}$$

1.4 Imbedding theorems

The definition of spaces $W^{1,p}(\Omega, \mathcal{T}_h)$ depend on the partition $\mathcal{T}_h$. If $\mathcal{T}_{h_1}$ and $\mathcal{T}_{h_2}$ are two different partitions of Ω, the spaces $W^{1,p}(\Omega, \mathcal{T}_{h_1})$ and $W^{1,p}(\Omega, \mathcal{T}_{h_2})$ are different as well. There exist examples of partitions such that $W^{1,p}(\Omega, \mathcal{T}_{h_1}) \cap W^{1,p}(\Omega, \mathcal{T}_{h_2}) = W^{1,p}(\Omega)$. However, the broken Sobolev spaces are imbedded into some mesh-independent function spaces, such as $BV(\Omega)$, the Lebesgue spaces $L^{p^\star}(\Omega)$ for suitable $p^\star \ge p$, and Besov spaces $B^{1/p;p,\infty}(\Omega)$. Besov spaces are defined by interpolation

$$B^{s;p,q}(\Omega) = \left(L^p(\Omega), W^{1,p}(\Omega)\right)_{s,q}, \qquad s \in (0,1),\ p, q \in [1, \infty], \tag{1.92}$$

see (Adams and Fournier, 2003, Section 7.32).

Property 1.22. *For each $p \in [1, \infty]$, $W^{1,p}(\Omega, \mathcal{T}_h) \subset BV(\Omega)$ and*

$$\|u\|_{BV(\Omega)} \le C\, \|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)}, \qquad u \in W^{1,p}(\Omega, \mathcal{T}_h). \tag{1.93}$$

Proof. The imbedding was proved in (Buffa and Ortner, 2009, Lemma 2), see also (Pietro and Ern, 2010, Lemma 6.2) and references therein. The proof reduces to the estimate of the BV-norm (1.10). For each $u \in W^{1,1}(\Omega)$ and $\boldsymbol{\varphi} \in [C_0^1(\Omega)]^d$,

$$\int_\Omega u \operatorname{div} \boldsymbol{\varphi} \,\mathrm{d}x = \sum_{\Gamma \in \mathcal{F}_h^I} \int_\Gamma [u]\, \boldsymbol{\varphi} \cdot \boldsymbol{n} \,\mathrm{d}x - \sum_{K \in \mathcal{T}_h} \int_K \boldsymbol{\varphi} \cdot \nabla u \,\mathrm{d}x \le \|\boldsymbol{\varphi}\|_{L^\infty(\Omega)} \, |u|_{W^{1,1}(\Omega,\mathcal{T}_h)}.$$

Taking supremum and substituting the result into (1.10), we get (1.93) for $p = 1$ with the constant $C = 1$. For $p > 1$, we use the inequality $\|u\|_{W^{1,1}(\Omega,\mathcal{T}_h)} \le C\, \|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)}$. □

Property 1.23 (Broken Sobolev imbedding)**.** *Let* $p \in [1, \infty]$ *and*

(i) $p^\star = dp/(d-p)$, *if* $p \in [1, d)$,

(ii) $p^\star \in (p, \infty)$ *be arbitrary, if* $p = d$,

(iii) $p^\star = \infty$, *if* $p > d$.

Then $W^{1,p}(\Omega, \mathcal{T}_h) \subset L^{p^\star}(\Omega)$, *and*

$$\|u\|_{L^{p^\star}(\Omega)} \leq C \|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)}, \qquad u \in W^{1,p}(\Omega, \mathcal{T}_h), \tag{1.94}$$

Proof. The choice of the exponent $p^\star$ ensures the imbedding $W^{1,p}(\Omega) \subset L^{p^\star}(\Omega)$ and also $W^{1,p}(K) \subset L^{p^\star}(K)$ for all $K \in \mathcal{T}_h$ (see (Adams and Fournier, 2003, Theorem 4.12)). Therefore, the set inclusion $W^{1,p}(\Omega, \mathcal{T}_h) \subset L^{p^\star}(\Omega)$ follows immediately. The bound (1.94) follows from the inequalities

$$\|u - P_h u\|_{L^{p^\star}(\Omega)} \leq C \|u - P_h u\|_{W^{1,p}(\Omega,\mathcal{T}_h)}, \tag{1.95}$$

$$\|P_h u\|_{L^{p^\star}(\Omega)} \leq C \|P_h u\|_{W^{1,p}(\Omega,\mathcal{T}_h)}, \tag{1.96}$$

$$\|P_h u\|_{W^{1,p}(\Omega)} \leq C \|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)}. \tag{1.97}$$

First, we prove (1.95). Recall, that by definition (1.61) of P_h,

$$\int_{B_K} (u - P_h u) \, \mathrm{d}x = 0.$$

A standard finite element scaling argument yields

$$\|u - P_h u\|_{L^{p^\star}(K)} \leq C h_K^{1-\frac{d}{p}+\frac{d}{p^\star}} |u - P_h u|_{W^{1,p}(K)} \leq C |u - P_h u|_{W^{1,p}(K)}, \qquad K \in \mathcal{T}_h.$$

If $p < \infty$ and $p^\star < \infty$, then

$$\begin{aligned}
\|u - P_h u\|_{L^{p^\star}(\Omega)}^{p^\star} &\leq C \sum_{K\in\mathcal{T}_h} |u - P_h u|_{W^{1,p}(K)}^{p^\star} \\
&\leq C \left(\max_{K\in\mathcal{T}_h} |u - P_h u|_{W^{1,p}(K)}^{p} \right)^{\frac{p^\star - p}{p}} \sum_{K\in\mathcal{T}_h} |u - P_h u|_{W^{1,p}(K)}^{p} \\
&\leq C \left(\sum_{K\in\mathcal{T}_h} |u - P_h u|_{W^{1,p}(K)}^{p} \right)^{\frac{p^\star}{p}} \\
&\leq C \|u - P_h u\|_{W^{1,p}(\Omega,\mathcal{T}_h)}^{\frac{p^\star}{p}}
\end{aligned}$$

The other cases $p = \infty$ or $p^\star = \infty$ are similar.

Now, let us prove (1.96). Since $P_h u = R_h P_h u$ on B_K,

$$\|P_h u\|_{L^{p^\star}(K)} = \left(\frac{|K|}{|B_K|}\right)^{1/p^\star} \|R_h P_h u\|_{L^{p^\star}(B_K)} \le C \|R_h P_h u\|_{L^{p^\star}(K)} .$$

By definition, $R_h P_h u \in W^{1,\infty}(\Omega)$. Using standard Sobolev imbedding, and (1.73), we obtain

$$\|R_h P_h u\|_{L^{p^\star}(\Omega)} \le C \|R_h P_h u\|_{W^{1,p}(\Omega)} \le C \|P_h u\|_{W^{1,p}(\Omega,\mathcal{T}_h)} .$$

The last inequality (1.97) follows from (1.66)-(1.68), similarly to the proof of Lemma 1.17. □

Property 1.23, for the case **(i)** and $u \in \mathcal{S}_{h,k}$, was proved in (Buffa and Ortner, 2009), using a different reconstruction operator. A more direct proof of Property 1.23 for the cases **(i), (ii)** can be found in (Pietro and Ern, 2010). We have included the case **(iii)** for completeness.

In (Brenner and Scott, 2002, Section 14.5), authors note that piecewise smooth functions lie in the space $(L^2(\Omega), W^{1,2}(\Omega))_{1/2,\infty}$. Concerning the broken Sobolev space, we formulate an analogous property as a continuous imbedding into the Besov space $B^{1/p;p,\infty}(\Omega)$.

Lemma 1.24. *Let $p \in (1,\infty)$. Then $W^{1,p}(\Omega,\mathcal{T}_h) \subset B^{1/p;p,\infty}(\Omega)$, and*

$$\|u\|_{B^{1/p;p,\infty}(\Omega)} \le C \|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)} , \qquad u \in W^{1,p}(\Omega,\mathcal{T}_h), \tag{1.98}$$

were C does not depend on h and u.

Proof. Since Ω has Lipschitz continuous boundary, the elements of the space $B^{s;p,q}(\Omega)$ are restrictions of functions of the space $B^{s;p,q}(\mathbb{R}^d)$, see (Adams and Fournier, 2003, Section 7.32). By (Adams and Fournier, 2003, Theorem 7.47), $v \in B^{s;p,\infty}(\mathbb{R}^d)$ if and only if

$$v \in L^p(\mathbb{R}^d), \qquad \operatorname{ess\,sup}_{z\in\mathbb{R}^d} |z|^{-s} \|\Delta_z v\|_{L^p(\mathbb{R}^d)} < \infty, \tag{1.99}$$

where Δ_z denote the finite difference operator

$$\Delta_z w(x) = w(x) - w(x-z), \qquad w \in L^1(\mathbb{R}^d),\ z \in \mathbb{R}^d.$$

For each $u \in L^1(\Omega)$, we define $E_0 u \in L^1(\mathbb{R}^d)$ as a zero extension of the function u. Obviously,

$$\|\Delta_z E_0 u\|_{L^\infty(\Omega)} \le 2 \|u\|_{W^{1,\infty}(\Omega,\mathcal{T}_h)} . \tag{1.100}$$

Using similar technique as in the proof of (1.93), we can prove

$$\|E_0 u\|_{BV(\mathbb{R}^d)} \le C \|u\|_{W^{1,1}(\Omega,\mathcal{T}_h)} . \tag{1.101}$$

By (Tartar, 2007, Lemma 37.4),

$$\|\Delta_z v\|_{L^1(\mathbb{R}^d)} \le |z|\ |v|_{BV(\mathbb{R}^d)} , \qquad v \in BV(\mathbb{R}^d). \tag{1.102}$$

Combining (1.100)-(1.102), the the interpolation Theorem 1.10 and the operator interpolation property (1.41) with $X_0 = W^{1,1}(\Omega, \mathcal{T}_h)$, $X_1 = W^{1,\infty}(\Omega, \mathcal{T}_h)$, $Y_0 = L^1(\Omega)$, $Y_1 = L^\infty(\Omega)$, $\theta = 1 - \frac{1}{p}$, we get

$$\|\Delta_z E_0 u\|_{L^p(\mathbb{R}^d)} \le C|z|^{1/p} \|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)} . \tag{1.103}$$

Therefore, $E_0 u \in B^{1/p;p,\infty}(\mathbb{R}^d)$ and $u \in B^{1/p;p,\infty}(\Omega)$. □

Recall that $\|\Delta_z v\|_{L^p(\mathbb{R}^d)} = \mathcal{O}(|z|)$ for $v \in W^{1,p}(\mathbb{R}^d)$, see (Evans, 1998, Section 5.8.2). The bound in (1.103) is only of order $\mathcal{O}\left(|z|^{1/p}\right)$. As the following lemma shows, it can be sharpened. However, the sharper bound is no longer h-independent.

Lemma 1.25. *Let $u \in W^{1,p}(\Omega, \mathcal{T}_h)$, $z \in \mathbb{R}^d$. Let Ω' be a subset of Ω such that*

$$\operatorname{dist}(\Omega', \partial\Omega) > |z|.$$

Then

$$\|u(\cdot) - u(\cdot - z)\|_{L^p(\Omega')} \le C|z|^{1/p} \left(|z| + h\right)^{1-1/p} \|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)} . \tag{1.104}$$

Proof. Let us consider the case $p = \infty$. Let $x \in \Omega'$ be arbitrary. Put $y = x - z$. There exist elements $K, L \in \mathcal{T}_h$ such that $x \in K$, $y \in L$. Let x_K denote the center of the ball B_K and y_L denote the center of the ball B_L. Then

$$\begin{aligned} |u(x) - u(x-z)| &\le |u(x) - P_h u(x)| + |u(y) - P_h u(y)| + |R_h P_h u(x_K) - R_h P_h u(y_L)| \\ &\le 2\|u - P_h u\|_{L^\infty(\Omega)} + C\|R_h P_h u\|_{W^{1,\infty}(\Omega)} |x_K - y_L| \\ &\le C(h + |x-y|) \|u\|_{W^{1,\infty}(\Omega,\mathcal{T}_h)} . \qquad [\text{ by (1.68), (1.67), (1.72) }] \end{aligned}$$

Taking essential supremum over $x \in \Omega$ gives us (1.104).

Now, let us consider the case $p = 1$. By (Eymard et al., 2000, Lemma 6.9, see also Tartar, 2007, Lemma 37.4),

$$\|u(\cdot) - u(\cdot - z)\|_{L^1(\Omega')} \le |z| \; |u|_{BV(\Omega)} , \qquad u \in BV(\Omega). \tag{1.105}$$

Using the imbedding (1.93), we get (1.104).

Finally, we prove (1.104) for $p \in (1, \infty)$. So far we have

$$\|\Delta_z u\|_{L^1(\Omega')} \le C|z| \|u\|_{W^{1,1}(\Omega,\mathcal{T}_h)} , \qquad \|\Delta_z u\|_{L^\infty(\Omega')} \le C(|z| + h) \|u\|_{W^{1,\infty}(\Omega,\mathcal{T}_h)} .$$

By the interpolation Theorem 1.10 and the operator interpolation property (1.41),

$$\|\Delta_z u\|_{L^p(\Omega')} \le C|z|^{1/p}(|z| + h)^{1-1/p} \|u\|_{W^{1,p}(\Omega,\mathcal{T}_h)} .$$

□

Remark. The bound (1.104) was proved for $p = 2$ and for piecewise constant functions in (Eymard et al., 2000, Lemma 3.3).

The inequality (1.104) shows that the imbedding $W^{1,p}(\Omega, \mathcal{T}_h) \subset L^p(\Omega)$ is compact in following sense

Property 1.26. *Let $C_r > 0$ and $p \in (1, \infty)$. Let $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ be a family of C_r-regular partitions of Ω. Let $\{u_h\}_{h\in(0,h_0)}$ be a family of functions $u_h \in W^{1,p}(\Omega, \mathcal{T}_h)$. Suppose there exists $M > 0$ such that*

$$\|u_h\|_{W^{1,p}(\Omega,\mathcal{T}_h)} \le M, \qquad h \in (0, h_0).$$

Then there exist $u \in W^{1,p}(\Omega)$ and a sequence $\{h_j\}_{j=1}^{\infty}$ such that $u_{h_j} \to u$ in $L^p(\Omega)$ and

$$\|u\|_{W^{1,p}(\Omega)} \le C_c M, \tag{1.106}$$

where C_c depends only on Ω, C_r and p.

Proof. Let us prove Property 1.26 using (1.104). Another proof can be found in (Buffa and Ortner, 2009; Pietro and Ern, 2010).

Let $\{h_i\}_{i=1}^{\infty}$ be an arbitrary sequence of real numbers satisfying $0 < h_i \le h_0$ and $h_i \to 0$. By (1.104), the set $\{u_h : h \in (0, h_0)\}$ is precompact in $L^p(\Omega)$, see (Adams and Fournier, 2003, Theorem 2.32). Considering the sequence $\{u_{h_i}\}_{i=1}^{\infty}$, there exists a L^p-convergent subsequence $\{u_{h_{i_j}}\}_{j=1}^{\infty}$. Let $u \in L^p(\Omega)$ be the limit function. By (1.104),

$$\|u(\cdot) - u(\cdot - z)\|_{L^p(\Omega')} \le CM \lim_{h\to 0} |z|^{1/p}(h + |z|)^{1-1/p} \le CM|z|,$$

for all subdomains $\Omega' \subset \Omega$, $\operatorname{dist}(\Omega', \partial\Omega) > |z|$. By (Evans, 1998, Section 5.8.2, Theorem 3), $u \in W^{1,p}(\Omega)$ and $\|u\|_{W^{1,p}(\Omega)} \le CM$. □

Chapter 2

Interior Penalty Discontinuous Galerkin Method

The *Discontinuous Galerkin* (DG) method is a versatile technique for numerical solution of partial differential equations. DG is characterized by piecewise polynomial, discontinuous approximate solution. From the historical perspective, DG is a generalization of the more traditional *finite element* (FE) and *finite volume* (FV) techniques. The DG combines the local conservation properties of FV and the high-order approximation properties of FE. Although the computational cost is often high, the flexibility with respect to local mesh refinement and domain decomposition make DG attractive in various applications. The DG method is applicable to hyperbolic systems of conservation laws and also partial differential equations of elliptic and parabolic type, (see Cockburn et al., 2000; Arnold et al., 2002, and references given therein).

There exist several variants of DG discretizations of linear elliptic problems (see Arnold et al., 2002). We focus on the approaches based on the primal formulation, namely SIPG (symmetric interior penalty Galerkin, see Arnold 1982), NIPG (nonsymmetric interior penalty Galerkin, see Rivière et al. 1999), and IIPG (incomplete interior penalty Galerkin, see Dawson et al. 2004) techniques. These *interior penalty* methods are characterized by the presence of interior and boundary penalties of order $O(h^{-1})$, where h is the mesh size.

As a model problem, we consider the Poisson equation with Dirichlet and Neumann boundary conditions. The classical formulation of our model problem reads: Find $u : \overline{\Omega} \to \mathbb{R}$ such that

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

$$u = u_D, \qquad \text{on } \partial\Omega_D, \tag{2.2}$$

$$\frac{\partial u}{\partial n} = g_N. \qquad \text{on } \partial\Omega_N, \tag{2.3}$$

where $\Omega \subset \mathbb{R}^d$, and the boundary $\partial\Omega$ is a disjoint union of the *Dirichlet* part $\partial\Omega_D$ and the *Neumann* part $\partial\Omega_N$. We assume $|\partial\Omega_D| \neq \emptyset$.

Moreover, we assume that a C_r-regular partition $\mathcal{T}_h$ of Ω is available (see Definition 1.2). We assume that the set of boundary faces $\mathcal{F}_h^{\partial\Omega}$ is equal to disjoint union of the *Dirichlet* *boundary faces* $\mathcal{F}_h^D$ and *Neumann boundary faces* $\mathcal{F}_h^N$,

$$\mathcal{F}_h^D = \left\{\Gamma \in \mathcal{F}_h^{\partial\Omega} : |\Gamma \cap \partial\Omega_D| \neq 0\right\}, \tag{2.4}$$

$$\mathcal{F}_h^N = \left\{\Gamma \in \mathcal{F}_h^{\partial\Omega} : |\Gamma \cap \partial\Omega_N| \neq 0\right\}. \tag{2.5}$$

The set of Dirichlet boundary faces $\mathcal{F}_h^D$ is not empty. We set $\mathcal{F}_h^{ID} = \mathcal{F}_h^I \cup \mathcal{F}_h^D$. The discrete problem reads: Find $u_h \in \mathcal{S}_{h,p}$ such that

$$B_h(u_h, v_h) = L_h(v_h), \quad v_h \in \mathcal{S}_{h,p}. \tag{2.6}$$

where

$$\begin{aligned} B_h(u_h, v_h) = &\sum_{K\in\mathcal{T}_h} \int_K \nabla u_h \cdot \nabla v_h \,\mathrm{d}x - \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \left(\langle \boldsymbol{n} \cdot \nabla u_h\rangle [v_h] + \theta \langle \boldsymbol{n} \cdot \nabla v_h\rangle [u_h]\right) \mathrm{dS} \\ &+ \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \frac{c_W}{h_\Gamma} [u_h] [v_h] \,\mathrm{dS}, \end{aligned} \tag{2.7}$$

$$\begin{aligned} L_h(v_h) = &\sum_{K\in\mathcal{T}_h} \int_K f v_h \,\mathrm{d}x + \sum_{\Gamma\in\mathcal{F}_h^N} \int_\Gamma g_N v_h \,\mathrm{dS} + \theta \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma (\boldsymbol{n} \cdot \nabla v_h) u_D \,\mathrm{dS} \\ &+ \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \frac{c_W}{h_\Gamma} u_D v_h \,\mathrm{dS}. \end{aligned} \tag{2.8}$$

In order to L_h be well-defined, we require $f \in L^2(\Omega)$, $u_D \in L^2(\Omega)$ and $g_N \in L^2(\Omega)$. Moreover, let us assume that (2.1)-(2.3) admits a strong solution $u \in H^2(\Omega)$. Then, the discrete problem (2.6) is consistent with (2.1)-(2.3),

$$B_h(u, v_h) = L_h(v_h), \qquad v_h \in \mathcal{S}_{h,p}. \tag{2.9}$$

We can derive (2.9) by multiplying (2.1) by arbitrary test function $v_h \in \mathcal{S}_{h,p}$, applying the Green theorem and using the fact $[u] = 0$. Now, the *Galerkin orthogonality* property

$$B_h(u_h - u, v_h) = 0, \qquad v_h \in \mathcal{S}_{h,p}. \tag{2.10}$$

follows easily. Until now, we did not impose any constraints on the parameters θ, c_W and h_Γ, $\Gamma \in \mathcal{F}_h^{ID}$.

Penalty parameters h_Γ. Let us first discuss the penalty parameters h_Γ. The term

$$J_h(u_h, v_h) = \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \frac{c_W}{h_\Gamma} [u_h] [v_h] \,\mathrm{dS} \tag{2.11}$$

in (2.7) penalizes both the inter-element jumps and the deviation of the discrete solution from the Dirichlet boundary condition. Let us consider the seminorm associated with the bilinear form $J_h(\cdot, \cdot)$,

$$|v_h|_{J_h}^2 = J_h(v_h, v_h) = c_W \sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma h_\Gamma^{-1} [v_h]^2 \,\mathrm{dS}, \qquad v_h \in \mathcal{S}_{h,p}.$$

In Chapter 1, we have seen such term in the norm of the broken Sobolev space $H^1(\Omega, \mathcal{T}_h) = W^{1,2}(\Omega, \mathcal{T}_h)$, with $h_\Gamma = \mathrm{diam}\,(\Gamma)$. However, other choices are possible, for example

$$h_\Gamma = \frac{h_L + h_R}{2}, \ \max\,(h_L, h_R)\,,\ |\Gamma|^{1/(d-1)}, \ \text{ etc. },$$

where $h_L = \mathrm{diam}\,\big(K_\Gamma^{\mathcal{L}}\big)$ and $h_R = \mathrm{diam}\,\big(K_\Gamma^{\mathcal{L}}\big)$. The resulting seminorms $|\cdot|_{J_h}$ are equivalent, since $\mathcal{T}_h$ is C_r-regular, see section 1.1. The choice of h_Γ plays a important role in the analysis of convergence L^2 norm (see section 2.2). In the following, we assume that there exists a constant C_P such that

$$C_P^{-1} h_K \leq h_\Gamma \leq C_P h_K, \tag{2.12}$$

whenever $\Gamma \in \mathcal{F}_h$ is a face of the element $K \in \mathcal{T}_h$. We will not discuss over-penalized variants, where $h_\Gamma = (\mathrm{diam}\,(\Gamma))^\alpha$, $\alpha > 1$ (see Rivière et al., 1999; Brenner et al., 2008).

Parameters θ and c_W. Note that the term

$$-\,\theta \sum_{\Gamma \in \mathcal{F}_h^{ID}} \int_\Gamma \langle \boldsymbol{n} \cdot \nabla v_h \rangle\, [u_h]\ \mathrm{dS} \tag{2.13}$$

in (2.7) is added artificially, in order to obtain special properties of the bilinear form $B_h(\cdot,\cdot)$. There are only three meaningful choices for parameter θ.

(i) $\theta = 1$: Symmetric Interior Penalty Galerkin Method (SIPG).
If $\theta = 1$, the bilinear form $B_h(\cdot,\cdot)$ is symmetric, i.e.

$$B_h(w, v) = B_h(v, w), \qquad w, v \in W^{2,2}(\Omega, \mathcal{T}_h). \tag{2.14}$$

Thus the symmetry of the Laplace operator Δ is preserved in the discretization. Let us note that when applied to more general PDEs, the SIPG discretization is *adjoint consistent* in the sense of (Arnold et al., 2002). Later, we will show that in order to get well-posed discrete problem (2.6), the penalty parameter c_W must be large enough.

(ii) $\theta = -1$: Nonsymmetric Interior Penalty Galerkin Method (NIPG).
If $\theta = -1$, the bilinear form $B_h(\cdot,\cdot)$ is positive definite for all choices of the penalty parameter $c_W > 0$. However, the theoretically attractive symmetry of $B_h(\cdot,\cdot)$ is lost.

(iii) $\theta = 0$: Incomplete Interior Penalty Galerkin Method (IIPG).
If $\theta = 0$, the bilinear form $B_h(\cdot,\cdot)$ does not contain the artificial term (2.13). The penalty parameter c_W must be large enough. The IIPG might be more suitable for some physical problems (see Dawson et al., 2004). Moreover, IIPG gives simpler discretization than SIPG or NIPG for nonlinear problems.

2.1 Apriori Error Analysis

The error estimates for interior penalty discontinuous Galerkin methods are now standard (Arnold et al., 2002). We present these results for completeness. We assume that u is the strong solution of (2.1)-(2.3) and $u \in H^s(\Omega)$, where $s \geq 2$. However, the following theory remains virtually unchanged, even if we use weaker assumption $s > 3/2$ (see Rivière, 2008).

In the analysis, we use the results on broken Sobolev spaces presented in Chapter 1. We use the shorter notation $H^s(\Omega, \mathcal{T}_h) = W^{s,2}(\Omega, \mathcal{T}_h)$, $s = 1, 2, \ldots$ and $\|\cdot\|_{H^1(\Omega,\mathcal{T}_h)} = \|\cdot\|_{W^{1,2}(\Omega,\mathcal{T}_h)}$. To avoid confusion, we emphasize that in this chapter, the symbol p denotes the degree of polynomial approximation, related to the discrete space $\mathcal{S}_{h,p}$. We assume $p \geq 1$.

Definition 2.1.

$$\|\!|v|\!\|^2 = \sum_{K\in\mathcal{T}_h} |v|^2_{H^1(K)} + \sum_{\Gamma\in\mathcal{F}_h^{ID}} \frac{c_W}{h_\Gamma} \|[v]\|^2_{L^2(\Gamma)}, \qquad v \in H^1(\Omega, \mathcal{T}_h), \tag{2.15}$$

$$\|\!|v|\!\|^2_\star = \|\!|v|\!\|^2 + \sum_{\Gamma\in\mathcal{F}_h^{ID}} \frac{h_\Gamma}{c_W} \|\langle \boldsymbol{n}\cdot\nabla v\rangle\|^2_{L^2(\Gamma)}, \qquad v \in H^2(\Omega, \mathcal{T}_h). \tag{2.16}$$

By the broken Poincare-Friedrichs inequality (Brenner, 2003)

$$\|v\|^2_{L^2(\Omega)} \leq C\|\!|v|\!\|^2, \qquad v \in H^1(\Omega, \mathcal{T}_h), \tag{2.17}$$

$\|\!|\cdot|\!\|$ is a norm on the space $H^2(\Omega, \mathcal{T}_h)$. The norm $\|\!|\cdot|\!\|$ is stronger than the norm $\|\cdot\|_{H^1(\Omega,\mathcal{T}_h)}$ introduced in Chapter 1. The space $H^1(\Omega, \mathcal{T}_h)$, equipped with the norm $\|\!|\cdot|\!\|$, is an analogue to the the space

$$H^1_0(\Omega) = \left\{ u \in H^1(\Omega) : u\big|_{\partial\Omega_D} = 0 \right\}.$$

Using the trace inequality (1.31) and the approximation properties (1.24), we obtain

$$\inf_{v_h\in\mathcal{S}_{h,q}} \|\!|v_h - v|\!\|_\star \leq Ch^q |v|_{H^{q+1}(\Omega,\mathcal{T}_h)}, \qquad v \in H^{q+1}(\Omega, \mathcal{T}_h),\ q = 1, 2, \ldots \tag{2.18}$$

First, we prove an auxiliary estimate:

Lemma 2.2. *There exists a constant $C_A > 0$ such that*

$$\sum_{\Gamma\in\mathcal{F}_h^{ID}} h_\Gamma \|\langle \boldsymbol{n}\cdot\nabla v_h\rangle\|^2_{L^2(\Gamma)} \leq C_A^2 \sum_{K\in\mathcal{T}_h} |v_h|^2_{H^1(K)}, \qquad v_h \in \mathcal{S}_{h,p}. \tag{2.19}$$

Proof. Let $v_h \in \mathcal{S}_{h,p}$ be arbitrary. Then

$$\sum_{\Gamma\in\mathcal{F}_h^{ID}} h_\Gamma \|\langle \boldsymbol{n}\cdot\nabla v_h\rangle\|^2_{L^2(\Gamma)}$$

$$
\begin{aligned}
&\leq \sum_{\Gamma\in\mathcal{F}_h^{ID}} h_\Gamma \left\|\nabla v_h\big|_\Gamma^{\mathcal{L}}\right\|_{L^2(\Gamma)^d}^2 + \sum_{\Gamma\in\mathcal{F}_h^{I}} h_\Gamma \left\|\nabla v_h\big|_\Gamma^{\mathcal{R}}\right\|_{L^2(\Gamma)^d}^2 \\
&\leq C \sum_{K\in\mathcal{T}_h} h_K \left\|\nabla v_h\right\|_{L^2(\partial K)^d}^2 && [\text{ by (2.12) }] \\
&\leq C \sum_{K\in\mathcal{T}_h} h_K \left(\left\|\nabla v_h\right\|_{L^2(K)^d} \left|\nabla v_h\right|_{H^1(K)^d} + h_K^{-1}\left\|\nabla v_h\right\|_{L^2(K)^d}^2\right) && [\text{ by trace ineq. (1.31) }] \\
&\leq C \sum_{K\in\mathcal{T}_h} \left(\left\|\nabla v_h\right\|_{L^2(K)^d}^2 + h_K^2\left|\nabla v_h\right|_{H^1(K)^d}^2\right) && [\text{ by Young inequality }] \\
&\leq C \sum_{K\in\mathcal{T}_h} \left\|\nabla v_h\right\|_{L^2(K)^d}^2 . && [\text{ by inverse ineq. (1.23) }]
\end{aligned}
$$

□

The inequality (2.19) implies following norm equivalence

$$
|||v_h|||^2 \leq |||v_h|||_\star^2 \leq (1 + c_W^{-1}C_A^2)|||v_h|||^2, \qquad v_h \in \mathcal{S}_{h,p}. \tag{2.20}
$$

Moreover, the inequality (2.19) is a key ingredient for the coercivity of the bilinear form $B_h(\cdot,\cdot)$ on the discrete space $\mathcal{S}_{h,p}$.

Property 2.3 (Boundedness and coercivity)**.** *There exists $c_{W,0} > 0$ and $C_c > 0$ such that for each $c_W > C_{W,0}$,*

$$
\begin{aligned}
B_h(w,v) &\leq C_c |||w|||_\star |||v|||_\star, && w,v \in H^2(\Omega,\mathcal{T}_h), && (2.21)\\
|||v_h|||^2 &\leq C_c B_h(v_h,v_h), && v_h \in \mathcal{S}_{h,p}. && (2.22)
\end{aligned}
$$

Proof. By repeated use of Cauchy-Schwarz inequality, we obtain

$$
|B_h(w,v)| \leq (1+|\theta|)|||w|||_\star |||v|||_\star .
$$

The inequality (2.21) holds with the constant $C = 2$, since $\theta \in \{-1,0,1\}$.

Now, let us prove (2.22). If $\theta = -1$, then $B_h(v_h,v_h) = |||v_h|||^2$. In this case, (2.22) holds with the constant $C = 1$. Let us now turn to the case $\theta \in \{0,1\}$. First, we estimate

$$
\begin{aligned}
&\sum_{\Gamma\in\mathcal{F}_h^{ID}} \int_\Gamma \Big(\langle \boldsymbol{n}\cdot\nabla v_h\rangle\,[v_h] + \theta\,\langle \boldsymbol{n}\cdot\nabla v_h\rangle\,[v_h]\Big)\,\mathrm{dS} \\
&\leq (1+\theta)\left(\sum_{\Gamma\in\mathcal{F}_h^{ID}} \frac{h_\Gamma}{c_W}\left\|\langle \boldsymbol{n}\cdot\nabla v_h\rangle\right\|_{L^2(\Gamma)}^2\right)^{1/2}\left(\sum_{\Gamma\in\mathcal{F}_h^{ID}} \frac{c_W}{h_\Gamma}\left\|[v]\right\|_{L^2(\Gamma)}^2\right)^{1/2} && [\text{ by Cauchy-Schwarz }] \\
&\leq \frac{(1+\theta)C_A}{\sqrt{c_W}}\left(\sum_{K\in\mathcal{T}_h} |v_h|_{W^{1,2}(K)}^2\right)^{1/2}\left(\sum_{\Gamma\in\mathcal{F}_h^{ID}} \frac{c_W}{h_\Gamma}\left\|[v]\right\|_{L^2(\Gamma)}^2\right)^{1/2} && [\text{ by (2.19) }]
\end{aligned}
$$

$$\leq \frac{(1+\theta)C_A}{2\sqrt{c_W}} \|v_h\|^2. \qquad \text{[by Young inequality]}$$

Then,

$$B_h(v_h, v_h) \geq \left(1 - \frac{(1+\theta)C_A}{2\sqrt{c_W}}\right) \|v_h\|^2.$$

The constant is positive, if $c_W > C_{W,0} := (1+\theta)^2 C_A^2/4$. □

Property 2.4 (broken H^1 error estimate)**.** *Let $u \in H^s(\Omega)$ with $s \geq 2$, $s \in \mathbb{N}$. Then there exist unique solution $u_h \in \mathcal{S}_{h,p}$ of the discrete problem* (2.6)*, and*

$$\|u_h - u\| \leq Ch^{\mu-1} \|u\|_{H^s(\Omega)}, \qquad \mu = \min\{p+1, s\}. \tag{2.23}$$

Proof. Since $\dim \mathcal{S}_{h,p} < \infty$, the discrete problem (2.6) is equivalent to a system of linear equations. The coercivity property (2.22) implies that the linear system is nonsingular, and thus uniquely solvable.

Let u_h be the solution of the discrete problem (2.6). Consider arbitrary decomposition of the error

$$e_h := u_h - u = \xi + \eta, \qquad \xi = u_h - v_h \in \mathcal{S}_{h,p},\ \eta = v_h - u \in H^2(\Omega, \mathcal{T}_h).$$

By triangle inequality

$$\|e_h\| \leq \|\xi\| + \|\eta\| \leq \|\xi\| + \|\eta\|_\star. \tag{2.24}$$

By coercivity (2.22), Galerkin orthogonality (2.10), boundedness (2.21), and norm equivalence (2.20) properties,

$$\frac{1}{C_c}\|\xi\|^2 \leq B_h(\xi, \xi) = -B_h(\xi, \eta) \leq C\|\xi\|_\star \|\eta\|_\star \leq C\|\xi\| \|\eta\|_\star. \tag{2.25}$$

Combining (2.24) and (2.25), and taking infimum over all approximations v_h of u, we get

$$\|e_h\| \leq C \inf_{v_h \in \mathcal{S}_{h,p}} \|v_h - u\|_\star. \tag{2.26}$$

Therefore, the discretization error, measured in the mesh-dependent broken H_0^1-seminorm, is bounded by approximation error in the auxiliary norm $\| \cdot \|_\star$. Using the approximation property (2.18), we obtain

$$\inf_{v_h \in \mathcal{S}_{h,p}} \|v_h - u\|_\star \leq \inf_{v_h \in \mathcal{S}_{h,\mu-1}} \|v_h - u\|_\star \leq Ch^{\mu-1} |u|_{H^\mu(\Omega,\mathcal{T}_h)} \leq Ch^{\mu-1} \|u\|_{H^s(\Omega)}. \tag{2.27}$$

□

If u is sufficiently regular ($u \in H^{p+1}(\Omega)$), then (2.23) gives us an error estimate of order $\mathcal{O}(h^p)$ in the broken H^1-norm $\| \cdot \|$. Moreover, the error estimate is optimal, because the best approximation of u in $\mathcal{S}_{h,p}$ is of the same order, i.e.

$$\inf_{v_h \in \mathcal{S}_{h,p}} \|u - v_h\| = \mathcal{O}(h^p).$$

By applying the broken Poincaré-Friedrichs inequality (2.17), we immediately get

$$\|u_h - u\|_{L^2(\Omega)} = \mathcal{O}\left(h^p\right). \tag{2.28}$$

However, (2.28) is not optimal, since

$$\inf_{v_h \in \mathcal{S}_{h,p}} \|v_h - u\|_{L^2(\Omega)} = \mathcal{O}\left(h^{p+1}\right).$$

In order to prove an optimal error estimate $\mathcal{O}\left(h^{p+1}\right)$ in the L^2-norm, the Aubin-Nitsche trick is usually employed. However, there the symmetry of the corresponding bilinear form is required which is satisfied only for the SIPG method.

Property 2.5 (L^2 error estimate)**.** *Let $\theta = 1$ and $u \in H^s(\Omega)$. Moreover, assume following regularity of the homogeneous continuous problem* (2.1)-(2.3)*: For each $r \in L^2(\Omega)$ there exists $\psi \in H^2(\Omega)$ such that*

$$-\Delta\psi = r \text{ in } \Omega, \quad \psi\big|_{\partial\Omega_D} = 0, \quad \frac{\partial\psi}{\partial n}\Big|_{\partial\Omega_N} = 0. \tag{2.29}$$

Then

$$\|u_h - u\|_{L^2(\Omega)} \le Ch^\mu \|u\|_{H^s(\Omega)}, \qquad \mu = \min\{p+1, s\}. \tag{2.30}$$

Proof. Let $r := e_h = u_h - u$ and let ψ be the solution of (2.29). Then

$$\begin{aligned}
\|e_h\|^2_{L^2(\Omega)} &= \int_\Omega r e_h = B_h(\psi, e_h) && [\text{ by consistency (2.9) }]\\
&= B_h(e_h, \psi) && [\text{ by symmetry (2.14) of } B_h(\cdot,\cdot)\text{ }]\\
&= B_h(e_h, \psi - \psi_h). && [\text{ by Galerkin orthogonality (2.10), } \psi_h \in \mathcal{S}_{h,1} \text{ arbitrary }]\\
&\le C|\!|\!|e_h|\!|\!| \inf_{\psi_h \in \mathcal{S}_{h,1}} |\!|\!|\psi - \psi_h|\!|\!|_\star && [\text{ by boundedness (2.21) of } B_h(\cdot,\cdot)\text{ }]\\
&\le Ch|\!|\!|e_h|\!|\!| \, \|\psi\|_{H^2(\Omega)}. && [\text{ by approximation property (2.18) }]\\
&\le Ch^\mu \|u\|_{H^s(\Omega)} \|\psi\|_{H^2(\Omega)}. && \left[\text{ by the broken } H^1\text{-estimate (2.23) }\right]
\end{aligned}$$

□

Numerical experiments carried out on uniform grids for NIPG and IIPG techniques (with sufficiently regular exact solution u) give the L^2-experimental orders of convergence (EOC) equal to $O(h^p)$ for even p but $O(h^{p+1})$ for odd p (see Babuška et al., 1999; Rivière, 2008, and references given therein). The optimal order of convergence for the odd degrees of approximation was theoretically justified in (Larson and Niklasson, 2004), where NIPG and IIPG methods were analyzed for uniform partitions of the one-dimensional domain. Similar result were presented in (Chen., 2006).

On the other hand, several examples of special non-uniform (but quasi-uniform) meshes were presented in (Guzmán and Rivière, 2009), where NIPG method gives EOC in the L^2-norm equal to $O(h^p)$ even for odd p. The sub-optimal EOC can be obtained also for

IIPG method using these meshes, see (Rivière, 2008, Section 1.5, Table 1.2). Optimal error estimates were shown for IIPG on arbitrary locally quasi-uniform meshes in 1D by (Dolejší and Havle, 2010). This result will be presented in the next section.

Theoretical results concerning NIPG and IIPG for 2D and 3D problems are very limited. In (Burman and Stamm, 2008), the optimal order of convergence in L^2-norm was proved for NIPG in 2D and 3D, with slightly modified penalization term. However, the proof is valid only for piecewise-linear approximation ($p = 1$) and simplicial meshes, which are *asymptotically uniform*, i.e.

$$\big| |K_\Gamma^{\mathcal{L}}| - |K_\Gamma^{\mathcal{R}}| \big| \leq C h_\Gamma^\zeta |K_\Gamma^{\mathcal{L}}|, \qquad \Gamma \in \mathcal{F}_h^I,$$

where ζ does not depend on h, and $\zeta \geq 2$. In (Wang et al., 2009), optimal estimates was proved for NIPG and IIPG on uniform rectangular meshes with piecewise bilinear approximation in 2D and piecewise trilinear approximation in 3D.

2.2 L^2-norm Error Estimate for IIPG Method in 1D

In this section, we show that if the penalty parameter h_Γ is specially chosen then IIPG method gives optimal error estimates in the L^2-norm for odd degrees of polynomial approximation for arbitrary locally quasi-uniform partitions. Moreover, we prove that any other choice of the penalty parameter of order $O(h)$ depending on the size of two neighboring elements does not lead to the optimal order of convergence in the L^2-norm. However, a choice of h_Γ for NIPG method, which guarantees optimal order of convergence in the L^2-norm for odd degrees of polynomial approximation, is still open. Numerical experiments which verify theoretical results can be found in appendix A.

Let $d = 1$ and $\Omega = (0, 1)$ be the one-dimensional computational domain. We consider the Poisson problem with mixed boundary conditions:

$$-u'' = f \quad \text{in } \Omega, \quad -u'(0) = g_N, \quad u(1) = u_D, \tag{2.31}$$

where $f : \Omega \to \mathbb{R}$, $u_D \in \mathbb{R}$ and $g_N \in \mathbb{R}$ are given. If $f \in H^s(\Omega)$, $s > 0$, then there exists $u \in H^{s+2}(\Omega)$ which is the unique strong solution of (2.31).

Remark. It is possible to consider the Dirichlet boundary conditions in both endpoints, i.e. $u(0) = u(1) = 0$. In this case Theorem 2.6 (the main result of this chapter) is valid. However, the proof of $(A) \implies (B)$ has to be slightly modified, the proof of $(B) \implies (A)$ rests the same.

Since Ω is one-dimensional, the elements of the partition $\mathcal{T}_h$ are intervals

$$\mathcal{T}_h = \{K_k : k = 0, \dots, N-1\}, \qquad \text{where } K_k = [x_k, x_{k+1}].$$

We assume that the nodes x_k are ordered by

$$0 = x_0 < x_1 < \cdots < x_{N-1} < x_N = 1.$$

We set $h_k = x_{k+1} - x_k$. Obviously, $h = \max_{k=0,\dots,N-1} h_k$ is the maximal element diameter. The partition $\mathcal{T}_h$ is C_r-regular in the sense of (1.15), (1.16), if and only if

$$h_k \le C_r h_{k+1}, \quad k = 0, \dots, N-2, \qquad h_k \le C_r h_{k-1}, \quad k = 1, \dots, N-1. \tag{2.32}$$

Let χ_k denote the characteristic function of element K_k, $k = 0, \dots, N-1$, i.e.

$$\chi_k(x) = \begin{cases} 1, & x \in K_k, \\ 0, & x \notin K_k. \end{cases}$$

The *jumps* and *mean values*, defined by (1.22), reduce to

$$[v]_k := [v]_{x_k} = \begin{cases} v(0^+), & k = 0, \\ v(x_k^-) - v(x_k^+), & k = 1, \dots, N-1, \\ v(1^-), & k = N \end{cases}$$
$$\langle v \rangle_k := \langle v \rangle_{x_k} = \begin{cases} v(0^+), & k = 0, \\ \frac{1}{2}\left(v(x_k^-) + v(x_k^+)\right), & k = 1, \dots, N-1, \\ v(1^-), & k = N, \end{cases}$$

where

$$v(x_k^-) = \lim_{\substack{x \to x_k \\ x < x_k}} v(x), \qquad k = 1, \dots, N,$$
$$v(x_k^+) = \lim_{\substack{x \to x_k \\ x > x_k}} v(x), \qquad k = 0, \dots, N-1.$$

We abbreviate the notation for the penalization parameters by $H_k := h_\Gamma$ where $\Gamma = x_k$. We consider the IIPG method ($\theta = 0$), thus the DG forms are

$$B_h(u_h, v_h) = \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} u_h' v_h' \,\mathrm{d}x - \sum_{k=1}^{N} \langle u_h' \rangle_k [v_h]_k + \sum_{k=1}^{N} \frac{c_W}{H_k} [u_h]_k [v_h]_k, \tag{2.33}$$
$$L_h(v_h) = \int_0^1 f v_h \,\mathrm{d}x + g_N v_h(0^+) + \frac{c_W}{H_N} u_D v_h(1^-), \qquad u_h, v_h \in \mathcal{S}_{h,p}, \tag{2.34}$$

and the discrete problem reads: Find $u_h \in \mathcal{S}_{h,p}$ such that

$$B_h(u_h, v_h) = L_h(v_h), \quad v_h \in \mathcal{S}_{h,p}. \tag{2.35}$$

All results of section 2.1, namely the broken H^1-error estimate (2.23), apply. In the following we deal with the optimality of the L^2-error estimate. Here, the choice of the parameters H_k, $k = 1, \dots, N$ is important. We assume that the parameters H_k are given by means of a given function $\mathcal{H} : (0, \infty) \times (0, \infty) \to \mathbb{R}$,

$$\begin{aligned} H_k &= \mathcal{H}(h_{k-1}, h_k), \ k = 1, \dots, N-1, \\ H_N &= \mathcal{H}(h_{N-1}, h_{N-1}). \end{aligned} \tag{2.36}$$

We assume that the function $\mathcal{H}(\cdot,\cdot)$ is continuous and satisfies

$$\begin{aligned} \mathcal{H}(a,b) &> 0, \\ \mathcal{H}(a,b) &= \mathcal{H}(b,a), \\ \mathcal{H}(\kappa a,\kappa b) &= \kappa\mathcal{H}(a,b), \qquad \kappa, a, b > 0. \end{aligned} \tag{2.37}$$

The assumptions (2.37) imply that inequalities (2.12) are satisfied with

$$C_P = \max_{\xi\in[C_r^{-1},C_r]} \max\left(\mathcal{H}(\xi,1), \frac{1}{\mathcal{H}(\xi,1)}\right).$$

The assumptions (2.36) – (2.37) are natural, e.g., in (Guzmán and Rivière, 2009) the values $H_k = (h_{k-1}+h_k)/2,\ k = 1,\dots,N-1$ and $H_N = h_{N-1}$ are used.

The main result of this chapter reads:

Theorem 2.6. *Let $p \in \mathbb{N}$ and a continuous function $H(\cdot,\cdot)$ satisfying (2.37) be given. Then two following assertions are equivalent.*

(A) *For each $C_r > 0$ there exists $C_{W,0} > 0$ such that for all $c_W > C_{W,0}$, $f \in H^p(0,1)$, $g_N, u_D \in \mathbb{R}$ there exists a constant $C_E > 0$ such that for any C_r-regular partition $\mathcal{T}_h$ the discrete problem (2.35) with the problem data f, g_N, u_D and the parameters c_W and $H_k,\ k = 1,\dots,N$ given by (2.36) has unique solution $u_h \in \mathcal{S}_{h,p}$ satisfying*

$$\|u_h - u\|_{L^2(0,1)} \le C_E h^{p+1}, \tag{2.38}$$

where u is the strong solution of (2.31).

(B) *The degree of approximation p is an odd number and the function $\mathcal{H}$ is a multiple of*

$$\mathcal{H}_p(a,b) = \begin{cases} \frac{a^{p+1}-b^{p+1}}{a^p-b^p}, & a \neq b, \\ \frac{p+1}{p}a, & a = b. \end{cases} \tag{2.39}$$

Theorem 2.6 implies that IIPG method gives optimal order of convergence for $p = 1$ if and only if $\mathcal{H}(a,b) = c(a+b),\ c > 0,\ c = \text{const}$. Hence for $p = 1$ the penalty parameters have to be chosen in the same was as, e.g, in (Guzmán and Rivière, 2009). However, for $p > 1$ the relation for "optimal" $\mathcal{H}$ is different.

Auxiliary results

Within this section we derive several auxiliary results which are the base of the proof of Theorem 2.6. In order to examine the penalization term, we construct representations of the *jump functionals*

$$\Phi_k(v_h) = \frac{c_W}{H_k}[v_h]_k, \qquad v_h \in \mathcal{S}_{h,p}, \qquad k = 1,\dots,N. \tag{2.40}$$

Since the bilinear form $B_h(\cdot,\cdot)$ is not symmetric, there are two natural choices of representation:

$$w_{h,p,k} \in \mathcal{S}_{h,p} \; : \quad B_h(w_{h,p,k}, v_h) = \Phi_k(v_h), \quad v_h \in \mathcal{S}_{h,p},\; k = 1,\dots,N, \tag{2.41}$$
$$w^\star_{h,p,k} \in \mathcal{S}_{h,p} : \quad B_h(v_h, w^\star_{h,p,k}) = \Phi_k(v_h), \quad v_h \in \mathcal{S}_{h,p},\; k = 1,\dots,N. \tag{2.42}$$

The existence and uniqueness of functions $w_{h,p,k}$ and $w^\star_{h,p,k}$ immediately follow from the coercivity (2.22), using the same reasoning as in the proof of Property 2.4.

The functions $w_{h,p,k}$ take a particularly simple form

$$w_{h,p,k} = \sum_{j=0}^{k-1} \chi_j. \tag{2.43}$$

The functions $w^\star_{h,p,k}$ can be expressed analytically for $p = 1$,

$$w^\star_{h,1,k}(x) = \frac{1}{2}\left(\frac{x - x_{k-1}}{h_{k-1}}\chi_{k-1}(x) + \frac{x - x_{k+1}}{h_{k+1}}\chi_k(x)\right), \qquad k = 1,\dots,N-1, \tag{2.44}$$
$$w^\star_{h,1,N}(x) = \frac{x - x_{k-1}}{h_{k-1}}\chi_{N-1}(x). \tag{2.45}$$

For general $p \geq 2$, the analytical expression of $w^\star_{h,p,k}$ is not easy to obtain, see also (Larson and Niklasson, 2004, paragraph 3.3), for special cases and different formulation. However, two following lemmas will be sufficient for our purposes.

Lemma 2.7. *For each $p = 1, 2, \dots$ there exists a polynomial $w^\star_p \in P^p(0,1)$ such that*

$$w^\star_{h,p,k}(x) = \frac{1}{2}\left(w^\star_p\left(\frac{x - x_{k-1}}{h_{k-1}}\right)\chi_{k-1}(x) - w^\star_p\left(\frac{x_{k+1} - x}{h_k}\right)\chi_k(x)\right), \tag{2.46}$$
$$k = 1,\dots,N-1,$$
$$w^\star_{h,p,N}(x) = w^\star_p\left(\frac{x - x_{N-1}}{h_{N-1}}\right)\chi_{N-1}(x), \tag{2.47}$$

where $w^\star_{h,p,k}$, $k = 1,\dots,N$ are defined by (2.42).

Proof. For $p = 1$, we use (2.44), (2.45). The polynomial $w^\star_1(t) = t$ satisfies (2.46), (2.47).

Let us now consider the case $p \geq 2$. Let $V_p = P^p(0,1) \cap H^1_0(0,1)$. Let $\tilde{w}_p \in V_p$ be the solution of the following symmetric, positive definite, finite-dimensional variational problem

$$\int_0^1 \tilde{w}'_p(t) v'_p(t)\,\mathrm{d}t = v'_p(1), \qquad v_p \in V_p.$$

We define

$$w^\star_p(t) = t + \tilde{w}_p(t). \tag{2.48}$$

First, we prove following auxiliary relation

$$\int_0^1 (w_p^\star)'(t)v'(t)\,\mathrm{d}t = v'(1), \qquad v \in P^p(0,1). \tag{2.49}$$

Let $v \in P^p(0,1)$ be arbitrary. There exist real numbers c_0, c_1 and a polynomial $v_p \in V_p$ such that $v(t) = c_0 + c_1 t + v_p(t)$. We have $v_p(0) = v_p(1) = \tilde{w}_p(0) = \tilde{w}_p(1) = 0$ and

$$\int_0^1 (w_p^\star)'(t)v'(t)\,\mathrm{d}t = \int_0^1 (1 + \tilde{w}_p'(t))(c_1 + v_p'(t))\,\mathrm{d}t = c_1 + v_p'(1) = v'(1)$$

Let us now prove (2.46). Let $v_h \in \mathcal{S}_{h,p}$ be arbitrary. Let $z_{h,k}$ denote the right-hand side of (2.46), where we set $w_p^\star$ as defined by (2.48). Using (2.49), we get

$$\begin{aligned}
\int_{x_{k-1}}^{x_k} v_h'(x)z_{h,k}'(x)\,\mathrm{d}x &= \frac{h_{k-1}}{2}\int_0^1 v_h'(x_{k-1} + th_{k-1})(w_p^\star(t))'\,\mathrm{d}t = \frac{v_h'(x_k^-)}{2},\\
\int_{x_k}^{x_{k+1}} v_h'(x)z_{h,k}'(x)\,\mathrm{d}x &= \frac{h_k}{2}\int_0^1 v_h'(x_{k+1} - th_k)(w_p^\star(t))'\,\mathrm{d}t = \frac{v_h'(x_k^+)}{2},\\
[z_{h,k}]_k &= 1,
\end{aligned}$$

and

$$B_h(v_h, z_{h,k}) = \frac{1}{2}\left(v_h'(x_k^-) + v_h'(x_k^+)\right) - \langle v_h'\rangle_k\, [z_{h,k}]_k + \frac{c_W}{H_k}[v_h]_k\,[z_{h,k}]_k = \frac{c_W}{H_k}[v_h]_k\,.$$

This proves the equation (2.46). We omit the proof of (2.47), since it is quite similar. □

Lemma 2.8. *The functions $w_{h,p,k}^\star$ defined by* (2.42) *have following properties*

$$\left[w_{h,p,k}^\star\right]_\ell = \delta_{k,\ell}, \qquad k, l = 1, \dots, N. \tag{2.50}$$

$$\int_{K_\ell} w_{h,p,k}^\star(x)g(x)\,\mathrm{d}x = 0, \qquad g \in P^{p-2}(K_\ell),\ p \geq 2, \tag{2.51}$$

$$\ell = 0, \dots, N-1,\ k = 1, \dots, N.$$

Proof. Putting $v_h := w_{h,p,k}^\star$ in (2.41), $v_h := w_{h,p,k}$ in (2.42) and using (2.40) and (2.43), we obtain

$$\frac{c_W}{H_\ell}\left[w_{h,p,k}^\star\right]_\ell = \Phi_\ell(w_{h,p,k}^\star) = B_h(w_{h,p,\ell}, w_{h,p,k}^\star) = \Phi_k(w_{h,p,\ell}) = \frac{c_W}{H_k}[w_{h,p,\ell}]_k = \frac{c_W}{H_k}\delta_{k,\ell},$$

which immediately gives (2.50).

Moreover, let $\ell = 0, \dots, N-1$ and let g be an arbitrary polynomial of degree less than or equal to $p-2$. Let $\psi \in H^2(\Omega)$ be the solution of the following boundary-value problem

$$\begin{aligned}
-\psi''(x) &= g(x)\chi_\ell(x), \qquad x \in (0,1),\\
\psi'(0) &= \psi(1) = 0.
\end{aligned}$$

Obviously, $\psi \in \mathcal{S}_{h,p}$ and $[\psi]_k = 0$. We have

$$\int_{K_\ell} g(x) w^\star_{h,p,k}(x)\,\mathrm{d}x = B_h(\psi, w^\star_{h,p,k}) = \Phi_k(\psi) = \frac{c_W}{H_k}\,[\psi]_k = 0.$$

□

The representation $w^\star_{h,p,k}$ allows us to express and estimate jumps of the approximate solution u_h. Substituting $w^\star_{h,p,k}$ as a test function to (2.35), we get

$$\frac{c_W}{H_k}\,[u_h]_k = B_h(u_h, w^\star_{h,p,k}) = L_h(w^\star_{h,p,k}),\ \ k = 1, \ldots, N. \tag{2.52}$$

In the following lemma, we identify the leading term on the the right-hand side $L_h(w^\star_{h,p,k})$.

Lemma 2.9. *Let* $w^\star_{h,p,k},\ k = 1, \ldots, N$ *be defined by* (2.42). *Then*

$$L_h(w^\star_{h,p,k}) = K_{0,p}\left((-1)^{p-1} h^p_{k-1} - h^p_k\right) f^{(p-1)}(x_k) + K_{2,p}\varepsilon_k, \qquad k = 1, \ldots, N-1, \tag{2.53}$$

$$L_h(w^\star_{h,p,N}) = \frac{c_W}{H_N} u_D + K_{1,p}(-1)^{p-1} h^p_{k-1} f^{(p-1)}(1) + K_{2,p}\varepsilon_N, \tag{2.54}$$

where

$$|\varepsilon_k| \le h^p \left\| f^{(p)} \right\|_{L^1(x_{k-1}, x_{k+1})}, \qquad k = 1, \ldots, N-1, \tag{2.55}$$

$$|\varepsilon_N| \le h^p \left\| f^{(p)} \right\|_{L^1(x_{N-1}, x_N)}, \tag{2.56}$$

$K_{0,p}, K_{1,p}, K_{2,p} \in \mathbb{R}$ *are constants and* $K_{0,p} \ne 0$.

Proof. Let $k = 1, \ldots, N-1$. Using the Taylor theorem, we have

$$f(x) = \sum_{j=0}^{p-1} \frac{f^{(j)}(x_k)}{j!}(x - x_k)^j + R(x), \qquad R(x) = \int_{x_k}^{x} f^{(p)}(\xi) \frac{(x-\xi)^p}{p!}\,\mathrm{d}\xi.$$

From (2.34), (2.46), (2.50) and (2.51) we get

$$\begin{aligned} L_h(w^\star_{h,p,k}) &= \int_{x_{k-1}}^{x_{k+1}} f(x) w^\star_{h,p,k}(x)\,\mathrm{d}x \\ &= \frac{f^{(p-1)}(x_k)}{(p-1)!} \int_{x_{k-1}}^{x_{k+1}} (x - x_k)^{p-1} w^\star_{h,p,k}(x)\,\mathrm{d}x \ + \int_{x_{k-1}}^{x_{k+1}} R(x) w^\star_{h,p,k}(x)\,\mathrm{d}x. \end{aligned}$$

Using (2.46), we obtain

$$\int_{x_{k-1}}^{x_{k+1}} (x - x_k)^{p-1} w^\star_{h,p,k}(x)\,\mathrm{d}x = \frac{(-1)^{p-1} h^p_{k-1} - h^p_k}{2} \int_0^1 (1-t)^{p-1} w^\star_p(t)\,\mathrm{d}t.$$

Finally, we estimate third term by

$$\left| \int\limits_{x_{k-1}}^{x_{k+1}} R(x) w^{\star}_{h,p,k}(x) \, \mathrm{d}x \right| \leq \frac{h^p}{p!} \left\| w^{\star}_p \right\|_{L^\infty(0,1)} \left\| f^{(p)} \right\|_{L^1(x_{k-1},x_{k+1})} .$$

Therefore, (2.53) holds with the constants

$$K_{0,p} = \frac{1}{2(p-1)!} \int\limits_0^1 (1-t)^{p-1} w^{\star}_p(t) \, \mathrm{d}t, \qquad K_{2,p} = \frac{1}{p!} \left\| w^{\star}_p \right\|_{L^\infty(0,1)} . \tag{2.57}$$

Now we prove by contradiction that $K_{0,p}$ is not zero. Let us assume that $K_{0,p} = 0$ then it follows from (2.57), (2.46) and (2.51) that $w^{\star}_p$ is orthogonal to $P^{p-1}(0,1)$. Thus $w^{\star}_p$ is a multiplicand of p^{th} element of the orthogonal basis of $P^p(0,1)$ and according, e.g., (Powell, 1981, Theorem 12.2) all roots of $w^{\star}_p$ are simple and lie in the interior of $(0,1)$. Then $w^{\star}_p(0) \neq 0$, but by (2.48) $w^{\star}_p(0) = 0$ which is in contradiction and thus $K_{0,p} \neq 0$.

The proof of (2.54) can be done by the similar technique. □

The representation $w_{h,p,k}$ is useful to quantify the influence of discretization parameters c_W and $\{H_k\}_{k=1}^N$ to the approximate solution u_h. Let us consider two sets of discretization parameters $(c_W, \{H_k\}_{k=1}^N)$ and $(\tilde{c}_W, \{\tilde{H}_k\}_{k=1}^N)$ and the corresponding approximate solutions u_h, $\tilde{u}_h$.

Lemma 2.10. *Let u_h and $\tilde{u}_h$ be the unique solutions of the discrete problem* (2.35) *with the discretization parameters $(c_W, \{H_k\}_{k=1}^N)$ and $(\tilde{c}_W, \{\tilde{H}_k\}_{k=1}^N)$, respectively. Then*

$$u_h - \tilde{u}_h = \sum_{k=1}^N \left(1 - \frac{c_W \tilde{H}_k}{\tilde{c}_W H_k} \right) [u_h]_k \, w_{h,p,k}. \tag{2.58}$$

Proof. By $\tilde{B}_h(\cdot,\cdot)$, $\tilde{L}_h(\cdot)$ we denote the bilinear and linear forms (2.33)-(2.34) corresponding to the discretization parameters $(\tilde{c}_W, \{\tilde{H}_k\}_{k=1}^N)$. Without any loss of generality, we assume $u_D = g_N = 0$. Then

$$B_h(u_h, v_h) = L_h(v_h) = \tilde{L}_h(v_h) = \tilde{B}_h(\tilde{u}_h, v_h), \qquad v_h \in \mathcal{S}_{h,p}. \tag{2.59}$$

By (2.41), the representation function $w_{h,p,k}$ does not depend on the choice of penalization parameters, and (2.41) holds with the form B_h replaced by $\tilde{B}_h$. Let r_h be the right-hand side of (2.58). Let $v_h \in \mathcal{S}_{h,p}$ be arbitrary. Then

$$\begin{aligned} \tilde{B}_h(r_h, v_h) &= \sum_{k=1}^N \left(1 - \frac{c_W \tilde{H}_k}{\tilde{c}_W H_k} \right) [u_h]_k \, \tilde{B}_h(w_{h,p,k}, v_h) \qquad \left[\text{ using bilinearity of } \tilde{B}_h \right] \\ &= \sum_{k=1}^N \left(1 - \frac{c_W \tilde{H}_k}{\tilde{c}_W H_k} \right) [u_h]_k \, \frac{\tilde{c}_W}{\tilde{H}_k} [v_h]_k \qquad [\text{ by } (2.41)] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{N} \frac{\tilde{c}_W}{\tilde{H}_k} [u_h]_k [v_h]_k - \sum_{k=1}^{N} \frac{c_W}{H_k} [u_h]_k [v_h]_k \\
&= \tilde{B}_h(u_h, v_h) - B_h(u_h, v_h) && [\text{ using the definition (2.33) of } B_h \ ] \\
&= \tilde{B}_h(u_h, v_h) - \tilde{B}_h(\tilde{u}_h, v_h). && [\text{ by (2.59) }]
\end{aligned}
$$

We have $\tilde{B}_h(r_h - (u_h - \tilde{u}_h), v_h) = 0$ for all $v_h \in \mathcal{S}_{h,p}$. Substituting $v_h = r_h - (u_h - \tilde{u}_h)$ and using the coercivity (2.22), we get $r_h - (u_h - \tilde{u}_h) = 0$. □

Proof of $(B) \implies (A)$

Assume **(B)**. Without any loss of generality, we assume $\mathcal{H}(a,b) = \mathcal{H}_p(a,b)$. We put $e_h = u_h - u \in L^2(\Omega)$. Let $\psi \in H^2(\Omega)$ be the weak solution of the boundary-value problem

$$
\begin{aligned}
-\psi'' &= e_h \quad \text{in } \Omega, \\
\psi'(0) &= 0, \quad \psi(1) = 0.
\end{aligned}
\tag{2.60}
$$

The function ψ is continuous and $[\psi]_k = 0$ for all $k = 1, \ldots, N$. A straightforward manipulation yields

$$
B_h(e_h, \psi) - \sum_{k=1}^{N} \langle \psi' \rangle_k [e_h]_k = B_h(\psi, e_h) = \int_\Omega e_h e_h \, \mathrm{d}x.
$$

The solution u of (2.31) is continuous as well. Therefore

$$
\|e_h\|^2_{L^2(\Omega)} = B_h(e_h, \psi) - \sum_{k=1}^{N} \langle \psi' \rangle_k [e_h]_k = B_h(e_h, \psi - \psi_h) - \sum_{k=1}^{N} \langle \psi' \rangle_k [e_h]_k \,, \tag{2.61}
$$

where ψ_h is a discontinuous piecewise linear approximation to ψ satisfying

$$
|\!|\!| \psi_h - \psi |\!|\!|_\star \le Ch \, |\psi|_{H^2(\Omega)} \,, \tag{2.62}
$$

see (2.18). So far, we followed the standard Nitsche trick. It is clear that we need to prove the inequality

$$
\left| \sum_{k=1}^{N} \psi'(x_k) \, [e_h]_k \right| \le Ch^{p+1} \, \|\psi\|_{H^2(\Omega)} \, \|f\|_{H^p(\Omega)} \,, \tag{2.63}
$$

since then we are able to estimate the right-hand side of (2.61) using (2.62), (2.21), (2.23) and (2.63) by

$$
\begin{aligned}
\|e_h\|^2_{L^2(\Omega)} &\le Ch |\!|\!| e_h |\!|\!| \, |\psi|_{H^2(\Omega)} + Ch^{p+1} \, \|\psi\|_{H^2(\Omega)} \, \|f\|_{H^p(\Omega)} \\
&\le C \left(\|u\|_{H^{p+1}(\Omega)} + \|f\|_{H^p(\Omega)} \right) h^{p+1} \, \|e_h\|_{L^2(\Omega)} \,.
\end{aligned}
$$

In order to prove (2.63), we use the functions $w^\star_{h,p,k},\ k=1,\dots,N$ given by (2.42). From (2.52) – (2.54), we get

$$\begin{aligned}
[e_h]_k &= [u_h]_k = \frac{H_k}{c_W} B_h(u_h, w^\star_{h,p,k}) = \frac{H_k}{c_W} L_h(w^\star_{h,p,k}) \\
&= \frac{K_{0,p}}{c_W} H_k \left((-1)^{p-1} h^p_{k-1} - h^p_k\right) f^{(p-1)}(x_k) + \frac{K_{2,p}}{c_W} H_k \varepsilon_k, \\
&\qquad\qquad\qquad\qquad\qquad\qquad k=1,\dots,N-1, \\
[e_h]_N &= [u_h]_N - u_D \\
&= \frac{K_{1,p}}{c_W} H_N (-1)^{p-1} h^p_{N-1} f^{(p-1)}(x_N) + \frac{K_{2,p}}{c_W} H_N \varepsilon_N.
\end{aligned}$$

By the assumption **(B)** of Theorem 2.6, p is odd and H_k are given by (2.36) and (2.39). Note that

$$H_k \left((-1)^{p-1} h^p_{k-1} - h^p_k\right) = h^{p+1}_{k-1} - h^{p+1}_k, \qquad k=1,\dots,N-1.$$

Therefore,

$$[e_h]_k = \frac{K_{0,p}}{c_W} \left(h^{p+1}_{k-1} - h^{p+1}_k\right) f^{(p-1)}(x_k) + \frac{K_{2,p}}{c_W} H_k \varepsilon_k, \quad k=1,\dots,N-1.$$

Let $\gamma(x) = \psi'(x) f^{(p-1)}(x)$. Using summation by parts

$$\begin{aligned}
\sum_{k=1}^{N-1} \left(h^{p+1}_{k-1} - h^{p+1}_k\right) \gamma(x_k) &= \sum_{k=1}^{N-2} h^{p+1}_k \left(\gamma(x_{k+1}) - \gamma(x_k)\right) + h^{p+1}_0 \gamma(x_1) - h^{p+1}_{N-1} \gamma(x_{N-1}) \\
&= \sum_{k=1}^{N-2} h^{p+1}_k \int\limits_{x_k}^{x_{k+1}} \gamma'(x) + h^{p+1}_0 \gamma(x_1) - h^{p+1}_{N-1} \gamma(x_{N-1}). \qquad (2.64)
\end{aligned}$$

Using the imbedding $W^{1,1}(0,1) \subset L^\infty(0,1)$, equivalence of $|\cdot|_{W^{1,1}(0,1)}$-seminorm with $\|\cdot\|_{W^{1,1}(0,1)}$-norm for $\gamma(0) = \psi'(0) f^{(p-1)}(0) = 0$ and the Cauchy inequality, we obtain

$$\|\gamma\|_{L^\infty(0,1)} \le C \|\gamma'\|_{L^1(0,1)} = C \left\|\psi'' f^{(p-1)} + \psi' f^{(p)}\right\|_{L^1(0,1)} \le C \|\psi\|_{H^2(0,1)} \|f\|_{H^p(0,1)},$$

which together with (2.64) yields

$$\left|\sum_{k=1}^{N-1} \left(h^{p+1}_{k-1} - h^{p+1}_k\right) \gamma(x_k)\right| \le C h^{p+1} \|\psi\|_{H^2(0,1)} \|f\|_{H^p(0,1)}$$

We complete the proof of (2.63) using (2.55), (2.56).

Proof of $(A) \implies (B)$

The assertion $A \implies B$ is proved in two steps:

- **Step 1.** by a contradiction we show that if **(A)** is valid then p has to be odd
- **Step 2.** assuming that **(A)** is valid and p is odd we show that $\mathcal{H}$ is a multiple of $\mathcal{H}_p$ given by (2.39).

Step 1. We prove by the contradiction, that the optimal L^2-estimate (2.38) does not hold for even p. Let us assume that **(A)** holds and p is even. Let $C_r = 1$, $f(x) = x^{p-1}$, $u_D = g_N = 0$. Let $C_{W,0}$ and C_E be the constants from **(A)**. Let $c_W = C_{W,0}$, $\tilde{c}_W = 2C_{W,0}$. Let $N \in \mathbb{N}$ be arbitrary and let $\mathcal{T}_h$ be an uniform partition with N elements, i.e. $h = 1/N$. Let u_h and $\tilde{u}_h$ respectively be the solution of (2.35) with discretization parameters $(c_W, \{H_k\}_{k=1}^N)$ and $(\tilde{c}_W, \{\tilde{H}_k\}_{k=1}^N)$, respectively, where $\tilde{H}_k = H_k = h\mathcal{H}(1,1)$, $k = 1, \ldots, N$. Let $r_h = u_h - \tilde{u}_h$. From (2.38), we have

$$\|r_h\|_{L^2(0,1)} \leq \|u_h - u\|_{L^2(0,1)} + \|u - \tilde{u}_h\|_{L^2(0,1)} \leq 2C_E h^{p+1}. \tag{2.65}$$

On the other hand, (2.58) and (2.43) implies, that the function r_h is constant on each element K_ℓ, $\ell = 1, \ldots, N$,

$$r_h\big|_{K_\ell}(x) = \frac{1}{2}\sum_{k=1}^{N} [u_h]_k\, w_{h,p,k}\big|_{K_\ell}(x) = \frac{1}{2}\sum_{k=1}^{N} [u_h]_k \sum_{j=0}^{k-1} \chi_j\big|_{K_\ell}(x) = \frac{1}{2}\sum_{k=\ell+1}^{N} [u_h]_k\,. \tag{2.66}$$

Using (2.52) – (2.56), the fact $f^{(p-1)}(x) = (p-1)!$ and $f^{(p)}(x) = 0$, we get

$$\begin{aligned} [u_h]_k &= M_1 h^{p+1}, \quad k = 1, \ldots, N-1, & M_1 &= \frac{2K_{0,p}(p-1)!H(1,1)}{c_W} \neq 0, \\ [u_h]_N &= M_2 h^{p+1}, & M_2 &= \frac{K_{1,p}(p-1)!H(1,1)}{c_W}. \end{aligned}$$

Therefore,

$$\begin{aligned} \|r_h\|_{L^2(0,1)} &\geq \|r_h\|_{L^1(0,1)} \geq \sum_{\ell=0}^{N-2} \|r_h\|_{L^1(x_\ell,x_{\ell+1})} = \sum_{\ell=0}^{N-2} h\left|M_1(N-\ell-1)h^{p+1} + M_2 h^{p+1}\right| \\ &\geq |M_1| h^{p+2} \sum_{\ell=0}^{N-2} (N-\ell-1) - |M_2| h^{p+2} \sum_{\ell=0}^{N-2} 1 \\ &\geq \left(|M_1|\frac{N(N-1)}{2} - |M_2|(N-1)\right) h^{p+2} \\ &\geq \frac{|M_1|}{2} h^p + \mathcal{O}\left(h^{p+1}\right), \end{aligned}$$

which is in contradiction with (2.65).

Step 2. Let us assume that **(A)** holds and p is odd. Let $C_r > 1$ is arbitrary, but fixed. Let $f(x) = x^{p-1}$, $u_D = g_N = 0$. Let $C_{W,0}$ and C_E be the constants from (A). Let $c_W = C_{W,0} + 1$, $\tilde{c}_W = 2C_{W,0}$. Let $\beta \in (C_r^{-1/2}, 1)$. Let $N \in \mathbb{N}$ be arbitrary multiple of three and let $\mathcal{T}_h$ be a partition with N elements, such that

$$h_{3j+\ell} = \frac{3\beta^\ell}{(1+\beta+\beta^2)N} \quad j = 0, \ldots, N/3-1,\ \ell = 0, 1, 2.$$

This partition satisfies (2.32) and

$$\begin{aligned} h &= \frac{M_0(\beta)}{N}, \qquad \text{where } M_0(\beta) = \frac{3}{1+\beta+\beta^2}, \\ h_{3j+\ell} &= h\beta^\ell, \quad j = 0,\dots,N/3-1, \ \ell = 0,1,2. \end{aligned} \tag{2.67}$$

Let u_h and $\tilde{u}_h$ respectively be the solution of (2.35) with discretization parameters $(c_W, \{H_k\}_{k=1}^N)$ and $(\tilde{c}_W, \{\tilde{H}_k\}_{k=1}^N)$, respectively, where $\tilde{H}_k = H_k = \mathcal{H}(h_{k-1}, h_k)$, $k = 1,\dots,N$. Let $r_h = u_h - \tilde{u}_h$.

Similarly as in Step 1, we have (2.65) and (2.66) and

$$\begin{aligned} [u_h]_k &= M_1 H_k (h_{k-1}^p - h_k^p), \quad k = 1,\dots,N-1, & M_1 &= \frac{K_{0,p}(p-1)!}{c_W}, \\ [u_h]_N &= M_2 H_N h_{N-1}^p, & M_2 &= \frac{K_p'(p-1)!}{c_W}. \end{aligned}$$

We estimate $\|r_h\|_{L^1(0,1)}$ from below by

$$\|r_h\|_{L^1(0,1)} \geq \sum_{\ell=1}^{N/3-1} h_{3\ell-1} \left| \sum_{k=3\ell}^{N-1} [u_h]_k + [u_h]_N \right| \geq M_3(h,\beta) - M_4(h,\beta), \tag{2.68}$$

where

$$\begin{aligned} M_3(h,\beta) &= \frac{|M_1|}{2} \sum_{\ell=1}^{N/3-1} h_{3\ell-1} \left| \sum_{k=3\ell}^{N-1} H_k (h_{k-1}^p - h_k^p) \right|, \\ M_4(h,\beta) &= \frac{|M_2|}{2} \sum_{\ell=1}^{N/3-1} h_\ell H_N h_{N-1}^p. \end{aligned}$$

The term $M_4(h,\beta)$ is $\mathcal{O}(h^{p+1})$ since

$$M_4(h,\beta) \leq \frac{|M_2|}{2} \mathcal{H}(1,1) h_{N-1}^{p+1} \sum_{\ell=1}^{N/3-1} h_\ell \leq \frac{|M_2|}{2} \mathcal{H}(1,1) h^{p+1}.$$

It remains to estimate term $M_3(h,\beta)$. Using $h_{3k-1} = h_{3k+2}$ and the homogeneity assumption (2.37), we get

$$\begin{aligned} \sum_{k=3\ell}^{N-1} H_k (h_{k-1}^p - h_k^p) &= \sum_{k=\ell}^{N/3-1} \sum_{j=0}^{2} H_{3k+j} (h_{3k+j-1}^p - h_{3k+j}^p) \\ &= \sum_{k=\ell}^{N/3-1} G(h_{3k}, h_{3k+1}, h_{3k+2}) = \frac{N-3\ell}{3} h^{p+1} G(1,\beta,\beta^2), \end{aligned}$$

where

$$G(\alpha_0,\alpha_1,\alpha_2)=\mathcal{H}(\alpha_0,\alpha_1)(\alpha_0^p-\alpha_1^p)+\mathcal{H}(\alpha_1,\alpha_2)(\alpha_1^p-\alpha_2^p)+\mathcal{H}(\alpha_2,\alpha_0)(\alpha_2^p-\alpha_0^p) \quad (2.69)$$

Therefore,

$$\begin{aligned}\sum_{\ell=1}^{N/3-1} h_{3\ell-1}\left|\sum_{k=3\ell}^{N-1} H_k(h_{k-1}^p-h_k^p)\right| &= \frac{h^{p+2}\beta^2}{3}|G(1,\beta,\beta^2)|\sum_{\ell=1}^{N/3-1}(N-3\ell)\\ &= \frac{(M_0^2(\beta)h^p-3M_0(\beta)h^{p+1})\beta^2}{18}|G(1,\beta,\beta^2)|,\end{aligned}$$

where $M_0(\beta)$ is given by (2.67). Then term $M_3(h,\beta)$ satisfies

$$M_3(h,\beta)=\frac{|M_1|}{2}\frac{(M_0^2(\beta)h^p-3M_0(\beta)h^{p+1})\beta^2}{18}|G(1,\beta,\beta^2)|.$$

However, estimates (2.65) and (2.68) implies that term $M_3(h,\beta)$ is $\mathcal{O}\left(h^{p+1}\right)$. Since $M_0(\beta)\neq 0$ and $M_1\neq 0$, it follows that

$$G(1,\beta,\beta^2)=0, \qquad \beta\in(C_r^{-1},1). \quad (2.70)$$

Finally, we have to prove that the property (2.70) implies that function $\mathcal{H}$ is a multiple of (2.39). Let us prove a technical lemma.

Lemma 2.11. *Let G be defined by* (2.69) *and there exists* $0<\varepsilon<1$ *such that*

$$G(1,\beta,\beta^2)=0, \qquad \beta\in(\varepsilon,1).$$

Then

$$\mathcal{H}(a,b)=\frac{\mathcal{H}(1,1)}{\mathcal{H}_p(1,1)}\mathcal{H}_p(a,b), \qquad a,b>0,\ \varepsilon^2<b/a<\varepsilon^{-2}, \quad (2.71)$$

where $\mathcal{H}_p$ is given by (2.39).

Proof. Let $F(\beta)=\mathcal{H}(1,\beta)$. A straightforward algebraic manipulation shows that

$$G(1,\beta,\beta^2)=(1-\beta^p)\left[(1+\beta^{p+1})F(\beta)-(1+\beta^p)F(\beta^2)\right].$$

Therefore, the function F satisfies the equation

$$(1+\beta^{p+1})F(\beta)=(1+\beta^p)F(\beta^2), \qquad \beta\in(\varepsilon,1). \quad (2.72)$$

Let $F_p(\beta)=\mathcal{H}_p(1,\beta)$. The function F_p satisfies the equation (2.72) as well,

$$(1+\beta^{p+1})F_p(\beta)=(1+\beta^p)F_p(\beta^2), \qquad \beta\in\mathbb{R}. \quad (2.73)$$

Dividing (2.72) by (2.73) we found that the ratio $Q(\beta)=F(\beta)/F_p(\beta)$ satisfies

$$Q(\beta)=Q(\beta^2), \qquad \beta\in(\varepsilon,1).$$

By continuity,

$$Q(\beta)=Q(\sqrt{\beta})=Q(\sqrt[4]{\beta})=\cdots=\lim_{r\to\infty}Q(\sqrt[2^r]{\beta})=Q(1), \qquad \beta\in(\varepsilon^2,1).$$

The equation (2.71) follows easily from the properties (2.37). □

Let us finish the proof of the main theorem. The equation (2.71) holds with $\varepsilon = C_r^{-1/2}$. The number $C_r > 1$ was arbitrary. Therefore $\mathcal{H}$ is indeed a multiple of $\mathcal{H}_p$,

$$\mathcal{H}(a,b) = \frac{\mathcal{H}(1,1)}{\mathcal{H}_p(1,1)}\mathcal{H}_p(a,b), \qquad a,b > 0.$$

Remark on the analysis of the NIPG method

There is a natural question if it is possible to use a similar technique for the determination of the penalty parameters which give optimal order of convergence for odd p also for the NIPG method on non-equidistant grids. Recall the one-dimensional form of the NIPG bilinear form (2.7) with $\theta = -1$

$$B_h^N(u_h, v_h) = \sum_{k=0}^{N-1} \int_{K_k} u_h' v_h' \,\mathrm{d}x - \sum_{k=1}^{N} \langle u_h' \rangle_k [v_h]_k + \sum_{k=1}^{N} \langle v_h' \rangle_k [u_h]_k + \sum_{k=1}^{N} \frac{c_W}{H_k} [u_h]_k [v_h]_k \quad (2.74)$$

and the corresponding NIPG linear form

$$L_h^N(v_h) = \int_0^1 f v_h \,\mathrm{d}x + g_N v_h(0) + u_D v_h'(1) + \frac{c_W}{H_N} u_D v_h(1), \quad (2.75)$$

where $c_W > 0$ and $H_k,\ k = 1,\dots,N$ are the given penalty parameters.

Replacing B_h by B_h^N in (2.41) – (2.42), we can define functions $w_{h,p,k}$ and $w_{h,p,k}^\star$ representing the jump functionals (2.40) for NIPG method. However, it is rather difficult to derive similar results as in Lemma 2.7 for NIPG method. Using Green's theorem and some technical manipulations, we obtain from (2.74) and (2.40) – (2.42) the identity

$$\begin{aligned}
\frac{c_W}{H_k} [v_h]_k =& \Phi_k(v_h) = B_h^N(v_h, w_{h,p,k}^\star) \\
=& -\sum_{j=0}^{N-1} \int_{K_j} v_h'' w_{h,p,k}^\star \,\mathrm{d}x - v_h'(0) w_{h,p,k}^\star(0) + \sum_{j=1}^{N-1} [v_h']_j \left\langle w_{h,p,k}^\star \right\rangle_j \\
&+ \sum_{j=1}^{N} \left\langle (w_{h,p,k}^\star)' \right\rangle_j [v_h]_j + \sum_{j=1}^{N} \frac{c_W}{H_j} [v_h]_j \left[w_{h,p,k}^\star\right]_j \quad \forall v_h \in \mathcal{S}_{h,p}.
\end{aligned}$$

By a contradiction it is possible to prove that function $w_{h,p,k}^\star$ has not support $[x_{k-1}, x_{k+1}]$ in contrast to the IIPG case (the proof is technical and it is based on suitable choices of $v_h \in \mathcal{S}_{h,p}$ which imply that $w_{h,p,k}^\star = 0$). This difference represents the main obstacle in the use of the jump functionals for the determination of the penalty parameters which give optimal order of convergence for the NIPG method.

Chapter 3

Finite Volume Methods for Shallow Water Equations

In this chapter, we present a finite volume discretization of the *Shallow Water Equations* (SWE) based on a Vijayasundaram numerical flux.

The SWE system (also called Saint-Venant equations) is an incompressible sub-model of the general governing equations for the dynamics of fluids. The system is derived from the incompressible Euler equations of fluid dynamics, neglecting the variations with respect to the vertical direction (see Toro, 1997, pg. 33). The equations represent the free-surface gravity flow in the three-dimensional channel with the bottom $x_3 = z(x_1, x_2)$ assumed fixed in time, and the free surface under gravity $x_3 = H(x_1, x_2, t)$ which depends on space and time (see Fig. 3.1). The flow is described by the height $h = h(x_1, x_2, t) = H(x_1, x_2, t) - z(x_1, x_2)$ and the two components of velocity $v_i = v_i(x_1, x_2, t)$, $i = 1, 2$. The SWE system

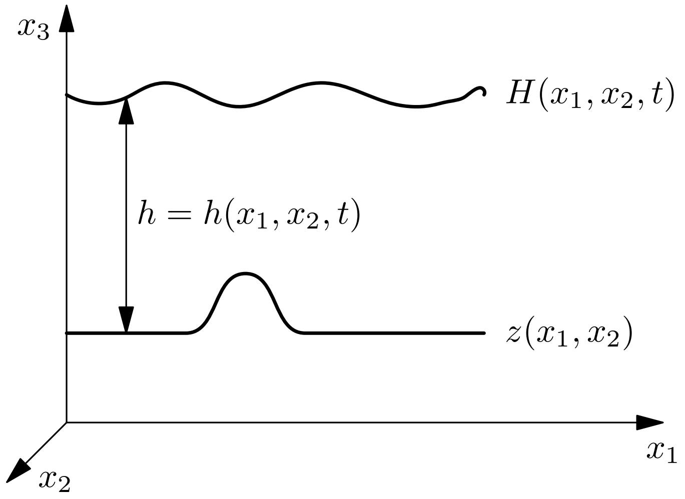

Figure 3.1: Notation for Shallow Water Equations.

reads

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x_1}(hv_1) + \frac{\partial}{\partial x_2}(hv_2) = 0, \tag{3.1}$$

$$\frac{\partial(hv_1)}{\partial t} + \frac{\partial}{\partial x_1}(hv_1^2 + \frac{1}{2}gh^2) + \frac{\partial}{\partial x_2}(hv_1v_2) = -gh\frac{\partial z}{\partial x_1}, \tag{3.2}$$

$$\frac{\partial(hv_2)}{\partial t} + \frac{\partial}{\partial x_1}(hv_1v_2) + \frac{\partial}{\partial x_2}(hv_2^2 + \frac{1}{2}gh^2) = -gh\frac{\partial z}{\partial x_2}. \tag{3.3}$$

The symbol g denotes a constant gravitational acceleration ($g \approx 9.8\text{ms}^{-2}$). The first equation (3.1) represents the conservation of mass, (3.2)-(3.3) represent the conservation of both horizontal components of the momentum.

We rewrite (3.1)-(3.3) as

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{s=1}^{2} \frac{\partial}{\partial x_s}\boldsymbol{f}_s(\boldsymbol{w}) = \boldsymbol{s}(x, \boldsymbol{w}), \tag{3.4}$$

where $\boldsymbol{w} = (h, hv_1, hv_2)^T$, and

$$\boldsymbol{f}_1(\boldsymbol{w}) = \begin{pmatrix} hv_1 \\ hv_1^2 + \frac{1}{2}gh^2 \\ hv_1v_2 \end{pmatrix}, \quad \boldsymbol{f}_2(\boldsymbol{w}) = \begin{pmatrix} hv_2 \\ hv_1v_2 \\ hv_2^2 + \frac{1}{2}gh^2 \end{pmatrix}, \quad \boldsymbol{s}(x, \boldsymbol{w}) = \begin{pmatrix} 0 \\ -gh\nabla z(x) \end{pmatrix}. \tag{3.5}$$

The fluxes $\boldsymbol{f}_s$ are defined on the domain $\mathcal{D} = \{(h, hv_1, hv_2) \in \mathbb{R}^3 : h > 0\}$.

If the channel bottom is flat ($z = \mathit{const.}$), then (3.4) becomes a system of conservation laws. Conservative methods, such as the Finite Volume or discontinuous Galerkin schemes, are natural choices for numerical solution. We present the construction of Finite Volume scheme for the flat bottom case in section 3.2, and for the general case in section 3.3.

It should be noted, that there is an important difficulty which we do not address in this work. During the evolution of the model, a state $h = 0$ may appear. The fluxes (3.5) are not well defined for $h = 0$. In this case, the numerical schemes discussed later may break down.

3.1 The continuous problem

Rotational Invariance. Let us consider an arbitrary orthogonal coordinate transformation

$$\tilde{x} = \mathbb{Q}_0 x + \tilde{x}_0$$

where $\mathbb{Q}_0 \in \mathbb{R}^{2\times 2}$ is orthogonal matrix, $\mathbb{Q}_0^T\mathbb{Q}_0 = \mathbb{Q}_0\mathbb{Q}_0^T = I$, and $\tilde{x}_0 \in \mathbb{R}^2$. The corresponding transformation of the vector $\boldsymbol{w}$ of conserved variables is

$$\tilde{\boldsymbol{w}} = \mathbb{Q}\boldsymbol{w}, \qquad \mathbb{Q} = \begin{pmatrix} 1 & 0 \\ 0 & \mathbb{Q}_0 \end{pmatrix} \in \mathbb{R}^{3\times 3}.$$

The SWE system (3.4) is *rotationally invariant*, i.e. a function $\boldsymbol{w} = \boldsymbol{w}(x,t)$ solves (3.4) if and only if the function

$$\tilde{\boldsymbol{w}}(\tilde{x},t) = \mathbb{Q}\boldsymbol{w}(\mathbb{Q}_0^{-1}(\tilde{x}-\tilde{x}_0),t)$$

solves the system (3.4), with the topography function z replaced by $\tilde{z}(\tilde{x}) = z(\mathbb{Q}_0^{-1}(\tilde{x}-\tilde{x}_0))$. The rotational invariance can be also stated as a property of the fluxes $\boldsymbol{f}_s$. The flux (of the quantity $\boldsymbol{w}$) in the direction $\boldsymbol{n} = (n_1,n_2)^T$ is (see Feistauer et al., 2003, section 3.1)

$$\mathcal{P}(\boldsymbol{w},\boldsymbol{n}) = \sum_{s=1}^{2} n_s \boldsymbol{f}_s(\boldsymbol{w}). \tag{3.6}$$

The flux is rotationally invariant, if

$$\mathcal{P}(\mathbb{Q}\boldsymbol{w},\mathbb{Q}_0\boldsymbol{n}) = \mathbb{Q}\mathcal{P}(\boldsymbol{w},\boldsymbol{n}) \tag{3.7}$$

for all $\boldsymbol{n} \in \mathbb{R}^2$ and all orthogonal matrices $\mathbb{Q}_0$. Let us prove (3.7) directly. Recall that $\boldsymbol{w} = (h, h\boldsymbol{v})^T = (h, hv_1, hv_2)^T$ and set $\boldsymbol{q} = h\boldsymbol{v}$.

$$\begin{aligned}
\mathcal{P}(\boldsymbol{w},\boldsymbol{n}) &= n_1 \begin{pmatrix} q_1 \\ h^{-1}q_1^2 + \frac{1}{2}gh^2 \\ h^{-1}q_1q_2 \end{pmatrix} + n_2 \begin{pmatrix} q_2 \\ h^{-1}q_1q_2 \\ h^{-1}q_2^2 + \frac{1}{2}gh^2 \end{pmatrix} = \frac{\boldsymbol{n}^T\boldsymbol{q}}{h}\begin{pmatrix} h \\ \boldsymbol{q} \end{pmatrix} + \frac{1}{2}gh^2\begin{pmatrix} 0 \\ \boldsymbol{n} \end{pmatrix}, \\
\mathcal{P}(\mathbb{Q}\boldsymbol{w},\mathbb{Q}_0\boldsymbol{n}) &= \frac{(\mathbb{Q}_0\boldsymbol{n})^T\mathbb{Q}_0\boldsymbol{q}}{h}\begin{pmatrix} h \\ \mathbb{Q}_0\boldsymbol{q} \end{pmatrix} + \frac{1}{2}gh^2\begin{pmatrix} 0 \\ \mathbb{Q}_0\boldsymbol{n} \end{pmatrix} \\
&= \frac{\boldsymbol{n}^T\boldsymbol{q}}{h}\mathbb{Q}\begin{pmatrix} h \\ \boldsymbol{q} \end{pmatrix} + \frac{1}{2}gh^2\mathbb{Q}\begin{pmatrix} 0 \\ \boldsymbol{n} \end{pmatrix} = \mathbb{Q}\mathcal{P}(\boldsymbol{w},\boldsymbol{n}).
\end{aligned}$$

This proves (3.7). The consequence of relation (3.7) is the following

Property 3.1. *Let $\boldsymbol{w} \in D$ and $\boldsymbol{n} \in \mathbb{R}^2$, $|\boldsymbol{n}| = 1$. Then*

$$\mathcal{P}(\boldsymbol{w},\boldsymbol{n}) = \mathbb{Q}^{-1}\boldsymbol{f}_1(\mathbb{Q}\boldsymbol{w}), \qquad \mathbb{Q}_0 = \mathbb{Q}_0(\boldsymbol{n}) = \begin{pmatrix} n_1 & n_2 \\ -n_2 & n_1 \end{pmatrix}. \tag{3.8}$$

Proof. The matrix $\mathbb{Q}_0$ is orthogonal and $\mathbb{Q}_0\boldsymbol{n} = (1,0)^T$ (matrix-vector product). It follows from (3.7)

$$\mathcal{P}(\boldsymbol{w},\boldsymbol{n}) = \mathbb{Q}^{-1}\mathcal{P}(\mathbb{Q}\boldsymbol{w},\mathbb{Q}_0\boldsymbol{n}) = \mathbb{Q}^{-1}\left[\underbrace{(\mathbb{Q}_0\boldsymbol{n})_1}_{=1}\boldsymbol{f}_1(\mathbb{Q}\boldsymbol{w}) + \underbrace{(\mathbb{Q}_0\boldsymbol{n})_2}_{=0}\boldsymbol{f}_2(\mathbb{Q}\boldsymbol{w})\right] = \mathbb{Q}^{-1}\boldsymbol{f}_1(\mathbb{Q}\boldsymbol{w}).$$

□

Hyperbolicity. Consider the case $z = const.$. Then (3.4) is a system of conservation laws. Hyperbolicity is an important property of system of conservative laws, which is crucial to the stability and well-posedness. By (Feistauer et al., 2003, Definition 2.3), the system is hyperbolic, if the Jacobi matrix (with respect to $\boldsymbol{w}$) of the directional flux (3.6) is diagonalizable and has real eigenvalues. By virtue of (3.8), it is sufficient to consider only the flux $\boldsymbol{f}_1$.

Property 3.2. *For all $\boldsymbol{w} = (h, hv_1, hv_2) \in \mathcal{D}$, the matrix*

$$\mathbb{A}_1(\boldsymbol{w}) = \begin{pmatrix} 0 & 1 & 0 \\ -v_1^2 + gh & 2v_1 & 0 \\ -v_1v_2 & v_2 & v_1 \end{pmatrix} \tag{3.9}$$

is the Jacobi matrix of $\boldsymbol{f}_1$. With the notation $c = \sqrt{gh}$, the eigenvalues and the corresponding eigenvalues of $\mathbb{A}_1(\boldsymbol{w})$ are

$$\lambda_1(\boldsymbol{w}) = v_1 - c, \qquad \lambda_2(\boldsymbol{w}) = v_1, \qquad \lambda_3(\boldsymbol{w}) = v_1 + c, \tag{3.10}$$

$$\boldsymbol{r}_1(\boldsymbol{w}) = \begin{pmatrix} 1 \\ v_1 - c \\ v_2 \end{pmatrix}, \qquad \boldsymbol{r}_2(\boldsymbol{w}) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \qquad \boldsymbol{r}_3(\boldsymbol{w}) = \begin{pmatrix} 1 \\ v_1 + c \\ v_2 \end{pmatrix}. \tag{3.11}$$

Moreover, the matrix $\mathbb{A}_1(\boldsymbol{w})$ is diagonalizable,

$$\mathbb{A}_1(\boldsymbol{w}) = \mathbb{T}\mathbb{\Lambda}\mathbb{T}^{-1}, \tag{3.12}$$

where

$$\mathbb{\Lambda} = \begin{pmatrix} v_1 - c & 0 & 0 \\ 0 & v_1 & 0 \\ 0 & 0 & v_1 + c \end{pmatrix}, \quad \mathbb{T} = \begin{pmatrix} 1 & 0 & 1 \\ v_1 - c & 0 & v_1 + c \\ v_2 & 1 & v_2 \end{pmatrix}, \quad \mathbb{T}^{-1} = \begin{pmatrix} \frac{v_1+c}{2c} & \frac{-1}{2c} & 0 \\ -v_2 & 0 & 1 \\ \frac{-v_1+c}{2c} & \frac{1}{2c} & 0 \end{pmatrix}. \tag{3.13}$$

Proof. By direct computation. □

The diagonal decomposition (3.12) shows that the system of conservation laws is hyperbolic. The explicit formulas for eigenvalues and eigenvectors (3.10)-(3.11) are important for the construction of numerical fluxes based on approximate Riemann solvers, such as the Vijayasundaram flux (see section 3.2).

Further properties of the continuous problem. We will make only few remarks. For example, one can show that the eigenvalues λ_1, λ_3 are *genuinely nonlinear*, and λ_2 is *linearly degenerate*. Also, theoretical results for the general case $z \neq$ const. can be found in literature, e.g. (Bernetti et al., 2008).

However, up to our knowledge, there are no results on global existence and uniqueness of the solution to (3.4). Therefore well-posed weak formulation of the PDE and the boundary conditions remains an open problem. Unfortunately, such questions are largely unsolved for systems of multidimensional nonlinear hyperbolic conservation laws, except certain special cases, 1D problems and Riemann problems (see Feistauer et al., 2003, section 2.3).

In order to have a starting point for the discretization, we need a formulation of the initial-boundary value problem (IBVP) on a bounded domain Ω and finite time interval $(0, T)$. Our treatment of the IBVP is rather formal. We will not discuss possible weak formulations, regularity assumption, and so forth. Without global well-posedness results, these discussions would be meaningless.

The initial-boundary value problem The shallow water equations represent a two-dimensional model. We consider the two-dimensional system (3.4) and the corresponding one-dimensional simplification (the so called split 2D version). Both cases are covered by the formulation

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{s=1}^{d} \frac{\partial}{\partial x_s} \boldsymbol{f}_s(\boldsymbol{w}) = \boldsymbol{s}(x, \boldsymbol{w}), \qquad x \in \Omega,\ t \in (0, T). \tag{3.14}$$

where $d \in \{1, 2\}$ and $\Omega \subset \mathbb{R}^d$. Note that in both cases, the vector of conserved variables has three components $\boldsymbol{w} = (h, hv_1, hv_2)$. We prescribe the initial condition

$$\boldsymbol{w}(x, 0) = \boldsymbol{w}^0(x), \qquad x \in \Omega, \tag{3.15}$$

where $\boldsymbol{w}^0$ is a given function.

The question of boundary conditions for nonlinear systems of conservation laws is delicate (see Dubois and Floch, 1988). Our notation closely matches the implementation on the numerical level. We prescribe boundary conditions in the form

$$\boldsymbol{w}(x, t) - \mathcal{B}\left(x, \boldsymbol{w}(x, t)\right) = 0, \tag{3.16}$$

where $\mathcal{B} : \partial\Omega \times \mathcal{D} \times \mathcal{D} \to \mathcal{D}$ is a given mapping. The motivation for this notation will be clear later in section 3.2, see (3.53). The mapping $\mathcal{B}$ represents the extrapolation procedure used in the finite volume method. In this way, several types of boundary conditions are possible, namely

(i) Prescribed water height

$$\mathcal{B}\left(x, (h, q_1, q_2)^T\right) = \mathcal{B}\left(x, (h^D(x), q_1, q_2)^T\right), \qquad x \in \partial\Omega,\ (h, q_1, q_2) \in \mathcal{D}. \tag{3.17}$$

where $h^D : \partial\Omega \to \mathbb{R}$ is a given function.

(ii) Prescribed discharge

$$\mathcal{B}\left(x, (h, q_1, q_2)^T\right) = \mathcal{B}\left(x, (h, q_1^D(x), q_2^D(x))^T\right), \qquad x \in \partial\Omega,\ (h, q_1, q_2) \in \mathcal{D}. \tag{3.18}$$

where $\boldsymbol{q}^D : \partial\Omega \to \mathbb{R}^2$ is a given vector-valued function.

(iii) Outflow boundary conditions

$$\mathcal{B}(x,\boldsymbol{w}) = \boldsymbol{w}, \qquad x \in \partial\Omega,\ \boldsymbol{w} \in \mathcal{D}. \tag{3.19}$$

Nothing is prescribed.

The zero-flux boundary conditions (e.g. impermeable wall) are not covered by (3.16). For simplicity, we do not consider such boundary conditions.

3.2 FV method for the case $z = const.$

We consider first the flat bottom case $z = const.$. A standard finite volume (FV) method (LeVeque, 1990; Eymard et al., 2000; Feistauer et al., 2003) can be applied. We retain the notation of Chapter 1. Let $\mathcal{T}_h$ be a partition of $\Omega \subset \mathbb{R}^d$. The elements of $\mathcal{T}_h$ are called *finite volumes* in this context. For the purposes of FV discretization, the requirements concerning $\mathcal{T}_h$ from section 1.1 can be relaxed. If $d = 2$, we assume that the finite volumes $K \in \mathcal{T}_h$ are closed polygons with mutually closed interiors. We introduce the notation

$$\mathcal{E}(L) = \{\Gamma \in \mathcal{F}_h : |\Gamma \cap L| \neq 0\}, \qquad L \in \mathcal{T}_h. \tag{3.20}$$

For each $\Gamma \in \mathcal{E}(L)$, we denote the outer unit normal to ∂L restricted to the face Γ by $\boldsymbol{n}_{L,\Gamma}$. We construct a partition $0 = t^0 < t^1 < \ldots < t^{N_T} = T$ of the time interval $[0,T]$ and denote by $\tau_j = t^{j+1} - t^j$ the time step between t^j and t^{j+1}.

Recall the derivation of the FV method. We integrate (3.14) over a finite volume $L \in \mathcal{T}_h$ and a time subinterval (t^j, t^{j+1}) and use Green's Theorem:

$$\int\limits_L w(x,t^{j+1})\,\mathrm{d}x - \int\limits_L w(x,t^j)\,\mathrm{d}x + \int\limits_{t^j}^{t^{j+1}} \int\limits_{\partial L} \mathcal{P}(\boldsymbol{w}(x,t),\boldsymbol{n})\,\mathrm{dS}\,\mathrm{d}t = 0,$$

$\mathcal{P}(\cdot,\cdot)$ is the directional flux (3.6). Now, we introduce the approximate solution

$$\boldsymbol{w}(x,t) \approx \boldsymbol{w}_K^j, \qquad x \in K,\ K \in \mathcal{T}_h,\ t \in [t^j, t^{j+1}),\ j = 0,1,\ldots,N_T - 1,$$

and replace the directional flux with the *numerical flux*

$$\mathcal{P}(\boldsymbol{w},\boldsymbol{n})\big|_\Gamma \approx \mathbf{H}\left(\boldsymbol{w}_L^j, \boldsymbol{w}_R^j, \boldsymbol{n}_{L,\Gamma}\right), \qquad \Gamma = L \cap R \in \mathcal{E}(L) \cap \mathcal{F}_h^I.$$

We discretize the boundary conditions (3.16) by

$$\mathcal{P}(\boldsymbol{w}(x,t),\boldsymbol{n})\big|_\Gamma \approx \mathbf{H}\left(\boldsymbol{w}_L^j, \mathcal{B}\left(x_\Gamma, \boldsymbol{w}_L^j\right), \boldsymbol{n}_{L,\Gamma}\right), \qquad \Gamma \in \mathcal{E}(L) \cap \mathcal{F}_h^{\partial\Omega}, \tag{3.21}$$

where x_Γ is the center of gravity of the face Γ. After simple algebraic manipulations, we obtain the explicit finite volume scheme

$$\boldsymbol{w}_L^{j+1} = \boldsymbol{w}_L^j - \tau^j \sum_{\substack{\Gamma \in \mathcal{E}(L) \cap \mathcal{F}_h^I \\ \Gamma = L \cap R, \\ R \in \mathcal{T}_h}} \frac{|\Gamma|}{|L|} \mathbf{H}\left(\boldsymbol{w}_L^j, \boldsymbol{w}_R^j, \boldsymbol{n}_{L,\Gamma}\right) \tag{3.22}$$

$$
\begin{aligned}
&- \tau^j \sum_{\Gamma\in\mathcal{E}(L)\cap\mathcal{F}_h^{\partial\Omega}} \frac{|\Gamma|}{|L|}\mathbf{H}\left(\boldsymbol{w}_L^j, \mathcal{B}\left(x_\Gamma, \boldsymbol{w}_L^j\right), \boldsymbol{n}_{L,\Gamma}\right), && L\in\mathcal{T}_h,\ j=0,1,\dots,N_T-1,\\
\boldsymbol{w}_L^0 &= \frac{1}{|L|}\int_L \boldsymbol{w}^0(x)\,\mathrm{d}x, && L\in\mathcal{T}_h. \qquad (3.23)
\end{aligned}
$$

It remains to specify the numerical flux $\mathbf{H}:\mathcal{D}\times\mathcal{D}\times\mathbb{R}^d\to\mathbb{R}^3$ and the time step τ^j. Let $\mathcal{S}_1=\left\{\boldsymbol{n}\in\mathbb{R}^d : |\boldsymbol{n}|=1\right\}$ be the unit sphere in $\mathbb{R}^d$. Following (Feistauer et al., 2003, section 3.3.3), we require that

(i) The numerical flux is defined and continuous on $\mathcal{D}\times\mathcal{D}\times\mathcal{S}_1$.

(ii) The numerical flux is *consistent* with the fluxes $\boldsymbol{f}_s$,

$$\mathbf{H}(\boldsymbol{w},\boldsymbol{w},\boldsymbol{n})=\mathcal{P}(\boldsymbol{w},\boldsymbol{n}), \qquad \boldsymbol{w}\in\mathcal{D},\ \boldsymbol{n}\in\mathcal{S}_1. \tag{3.24}$$

(iii) The numerical flux is *conservative*,

$$\mathbf{H}(\boldsymbol{w}_L,\boldsymbol{w}_R,\boldsymbol{n})=-\mathbf{H}(\boldsymbol{w}_R,\boldsymbol{w}_L,-\boldsymbol{n}), \qquad \boldsymbol{w}_L,\boldsymbol{w}_R\in\mathcal{D},\ \boldsymbol{n}\in\mathcal{S}_1. \tag{3.25}$$

Moreover, in analogy to the rotational invariance of the fluxes and (3.8), we assume that the numerical flux is given by means of a mapping $\mathbf{g}:\mathcal{D}\times\mathcal{D}\to\mathbb{R}^3$ such that

$$\mathbf{H}(\boldsymbol{w}_L,\boldsymbol{w}_R,\boldsymbol{n})=\mathbb{Q}^{-1}\mathbf{g}(\mathbb{Q}\boldsymbol{w}_L,\mathbb{Q}\boldsymbol{w}_R), \qquad \boldsymbol{w}_L,\boldsymbol{w}_R\in\mathcal{D},\ \boldsymbol{n}=(n_1,n_2)^T\in\mathbb{R}^2, \tag{3.26}$$
$$\text{where}\ \ \mathbb{Q}=\begin{pmatrix}1&0&0\\0&n_1&n_2\\0&-n_2&n_1\end{pmatrix}.$$

There are two fundamental approaches to construction of the numerical fluxes. The numerical flux can be derived from finite difference approximations to (3.14). This approach leads to central schemes, e.g. the Lax-Friedrichs flux, or the FORCE schemes (Toro et al., 2009; Canestrelli et al., 2009). The other approach is based on analysis of the solution of the Riemann problem. Examples are the Godunov numerical flux (exact Riemann solver) and various approximate Riemann solvers, e.g. the Roe-type fluxes (Gallouët et al., 2003), and the Osher-Solomon flux (Zhao et al., 1996).

We present a numerical flux, which is based on the well-known Vijayasundaram flux from the context of the compressible Euler equations (Vijayasundaram, 1982; Feistauer et al., 2003). The Vijayasundaram flux for the Euler equations reads

$$\mathbf{g}(\boldsymbol{w}_L,\boldsymbol{w}_R)=\mathbb{A}_1^+\left(\frac{\boldsymbol{w}_L+\boldsymbol{w}_R}{2}\right)\boldsymbol{w}_L+\mathbb{A}_1^-\left(\frac{\boldsymbol{w}_L+\boldsymbol{w}_R}{2}\right)\boldsymbol{w}_R, \tag{3.27}$$

where $\mathbb{A}_1^+$ (or $\mathbb{A}_1^-$) is the *positive part* (or the *negative part*, respectively) of the matrix $\mathbb{A}_1$. For scalar arguments, we set

$$a^+=\max(a,0), \qquad a^-=\min(a,0). \tag{3.28}$$

We generalize (3.28) for matrix arguments using the diagonal decomposition $\mathbb{A}_1 = \mathbb{T}\Lambda\!\!\setminus\mathbb{T}^{-1}$, $\Lambda\!\!\setminus = \text{diag}\left\{\lambda_1, \lambda_2, \lambda_3\right\}$, by

$$\mathbb{A}_1^{\pm} = \mathbb{T}\text{diag}\left\{\lambda_1^{\pm}, \lambda_2^{\pm}, \lambda_3^{\pm}\right\}\mathbb{T}^{-1}. \tag{3.29}$$

The use of (3.27) as a numerical flux is justified by the fact that (3.27) is an approximate Riemann solver. The formula (3.27) is equivalent to the exact Riemann solver for linear hyperbolic systems. We formulate this property for a scalar linear problem, but it is readily generalized to hyperbolic linear systems using the diagonal decomposition (see, e.g. LeVeque, 1990).

Property 3.3. *Let us consider the Riemann problem for a scalar linear equation*

$$\frac{\partial u}{\partial t} + \lambda\frac{\partial u}{\partial x} = 0, \qquad x \in \mathbb{R},\ t > 0, \tag{3.30}$$

$$u(x,0) = \begin{cases} u_L, & x < 0, \\ u_R, & x > 0. \end{cases} \tag{3.31}$$

Then, the value of the linear flux $f_\lambda(u) = \lambda u$ *on the line* $x = 0$ *is*

$$f_\lambda\left(u(0,t)\right) = \lambda^+ u_L + \lambda^- u_R. \tag{3.32}$$

Proof. The solution is $u(x,t) = u_0(x - \lambda t)$. If $\lambda > 0$, then $u(0,t) = u_L$. If $\lambda < u$, then $u(0,t) = u_R$. In both cases, (3.32) holds. □

The Vijayasundaram flux (3.27) is consistent in the sense of (3.24) if and only if $\boldsymbol{f}_1(\boldsymbol{w}) = \mathbb{A}_1(\boldsymbol{w})\boldsymbol{w}$ for all $\boldsymbol{w} \in \mathcal{D}$. This *homogeneity* property is valid for linear systems and for the Euler equations (Feistauer et al., 2003, Lemma 3.1), but it does not hold for SWE. Using (3.9), we can show

$$\mathbb{A}_1(\boldsymbol{w})\boldsymbol{w} = \boldsymbol{f}_1(\boldsymbol{w}) - \frac{1}{2}gh^2\boldsymbol{e}_2, \tag{3.33}$$

where $\boldsymbol{w} = (h, hv_1, hv_2)^T$ and $\boldsymbol{e}_2 = (0,1,0)^T$. We propose the new numerical flux of Vijayasundaram type, defined by

$$\mathbf{g}\left(\boldsymbol{w}_L, \boldsymbol{w}_R\right) = \mathbb{A}_1^+\left(\frac{\boldsymbol{w}_L + \boldsymbol{w}_R}{2}\right)\boldsymbol{w}_L + \mathbb{A}_1^-\left(\frac{\boldsymbol{w}_L + \boldsymbol{w}_R}{2}\right)\boldsymbol{w}_R - \frac{g}{2}\left(\frac{h_L + h_R}{2}\right)^2\boldsymbol{e}_2. \tag{3.34}$$

The following theorem shows that the numerical flux (3.27) is suitable for use in the FV scheme.

Theorem 3.4. *The numerical flux of Vijayasundaram type defined by* (3.34) *and* (3.26) *is continuous, consistent and conservative.*

Proof. The continuity of is obvious consequence of the continuous dependence of eigenvalues and eigenvectors on the entries of the matrix $\mathbb{A}_1$. The consistency (3.24) follows easily

$$\mathbf{H}\left(\boldsymbol{w}, \boldsymbol{w}, \boldsymbol{n}\right) = \mathbb{Q}^{-1}\mathbf{g}\left(\mathbb{Q}\boldsymbol{w}, \mathbb{Q}\boldsymbol{w}\right) \qquad \text{[by (3.26)]}$$

$$\begin{aligned}
&= \mathbb{Q}^{-1}\left(\mathbb{A}_1(\mathbb{Q}\boldsymbol{w})(\mathbb{Q}\boldsymbol{w}) - \frac{1}{2}gh^2\right) && [\text{ by (3.34), using } \mathbb{A}_1 = \mathbb{A}_1^+ + \mathbb{A}_1^- \ ]\\
&= \mathbb{Q}^{-1}\boldsymbol{f}_1(\mathbb{Q}\boldsymbol{w}) && [\text{ by (3.33) }]\\
&= \mathcal{P}(\boldsymbol{w},\boldsymbol{n}). && [\text{ by (3.8) }]
\end{aligned}$$

Before proving the conservation property (3.25), we need following lemma.

Lemma 3.5. *Let* $\mathbb{Z} = \operatorname{diag}\{1,-1,-1\}$. *Then it holds*

$$\mathbb{A}_1^+(\mathbb{Z}\boldsymbol{w}) = -\mathbb{Z}\mathbb{A}_1^-(\boldsymbol{w})\mathbb{Z}, \tag{3.35}$$

$$\mathbb{A}_1^-(\mathbb{Z}\boldsymbol{w}) = -\mathbb{Z}\mathbb{A}_1^+(\boldsymbol{w})\mathbb{Z}. \tag{3.36}$$

Proof of the lemma. Let $\mathbb{A}_1(\boldsymbol{w}) = \mathbb{T}\mathbb{\Lambda}\mathbb{T}$ be the diagonal decomposition of the matrix $\mathbb{A}_1$. First we transform the matrix $\mathbb{A}_1(\mathbb{Z}\boldsymbol{w})$ to the diagonal form. For $\boldsymbol{n} = (-1,0)^T$, the matrix $\mathbb{Q}$ from (3.8) satisfies $\mathbb{Q} = \mathbb{Q}^{-1} = \mathbb{Z}$ and the rotational invariance property gives

$$-\boldsymbol{f}_1(\boldsymbol{w}) = \mathcal{P}(\boldsymbol{w},\boldsymbol{n}) = \mathbb{Q}^{-1}\boldsymbol{f}_1(\mathbb{Q}\boldsymbol{w}) = \mathbb{Z}\boldsymbol{f}_1(\mathbb{Z}\boldsymbol{w}).$$

By differentiating this identity with respect to $\boldsymbol{w}$, we get $-\mathbb{A}_1(\boldsymbol{w}) = \mathbb{Z}\mathbb{A}_1(\mathbb{Z}\boldsymbol{w})\mathbb{Z}$, and

$$\mathbb{A}_1(\mathbb{Z}\boldsymbol{w}) = \mathbb{Z}\mathbb{T}(\boldsymbol{w})[-\mathbb{\Lambda}(\boldsymbol{w})][\mathbb{Z}\mathbb{T}(\boldsymbol{w})]^{-1}.$$

For all $a \in \mathbb{R}$,

$$(-a)^+ = -a^-, \qquad (-a)^- = -a^+,$$

and this property holds for diagonal matrices as well. So

$$\begin{aligned}
\mathbb{A}_1^+(\mathbb{Z}\boldsymbol{w}) = \mathbb{Z}\mathbb{T}(\boldsymbol{w})[-\mathbb{\Lambda}(\boldsymbol{w})]^+[\mathbb{Z}\mathbb{T}(\boldsymbol{w})]^{-1} = -\mathbb{Z}\mathbb{T}(\boldsymbol{w}))\mathbb{\Lambda}(\boldsymbol{w})^-[\mathbb{Z}\mathbb{T}(\boldsymbol{w})]^{-1} = -\mathbb{Z}\mathbb{A}_1^-(\boldsymbol{w})\mathbb{Z},\\
\mathbb{A}_1^-(\mathbb{Z}\boldsymbol{w}) = \mathbb{Z}\mathbb{T}(\boldsymbol{w})[-\mathbb{\Lambda}(\boldsymbol{w})]^-[\mathbb{Z}\mathbb{T}(\boldsymbol{w})]^{-1} = -\mathbb{Z}\mathbb{T}(\boldsymbol{w})\mathbb{\Lambda}(\boldsymbol{w})^+[\mathbb{Z}\mathbb{T}(\boldsymbol{w})]^{-1} = -\mathbb{Z}\mathbb{A}_1^+(\boldsymbol{w})\mathbb{Z}.
\end{aligned}$$

□

Now we can finish the proof of the theorem. We set $\boldsymbol{w}_\star = \frac{1}{2}(\boldsymbol{w}_L + \boldsymbol{w}_R)$ and $h_\star = \frac{1}{2}(h_L + h_R)$. We rewrite the numerical flux in the form $\mathbf{H} = \mathbf{H}_1 - \mathbf{H}_2$, where

$$\mathbf{H}_1(\boldsymbol{w}_L,\boldsymbol{w}_R,\boldsymbol{n}) = \mathbb{Q}^{-1}\left[\mathbb{A}_1^+(\mathbb{Q}\boldsymbol{w}_\star)\mathbb{Q}\boldsymbol{w}_L + \mathbb{A}_1^-(\mathbb{Q}\boldsymbol{w}_\star)\mathbb{Q}\boldsymbol{w}_R\right], \tag{3.37}$$

$$\mathbf{H}_2(\boldsymbol{w}_L,\boldsymbol{w}_R,\boldsymbol{n}) = \frac{g}{2}h_\star^2\mathbb{Q}^{-1}\boldsymbol{e}_2 \tag{3.38}$$

It is sufficient to prove the conservation property separately for $\mathbf{H}_1$ and $\mathbf{H}_2$.

For $\boldsymbol{H}_1$ we express $\mathbb{Q} = \mathbb{Q}_{\boldsymbol{n}}$ defined in (3.8) with the aid of $\mathbb{Q}_{-\boldsymbol{n}}$ and $\mathbb{Z}$ defined in Lemma 3.5 as

$$\mathbb{Q}_{\boldsymbol{n}} = \mathbb{Z}\mathbb{Q}_{-\boldsymbol{n}}, \qquad \mathbb{Q}_{\boldsymbol{n}}^{-1} = \mathbb{Q}_{-\boldsymbol{n}}^{-1}\mathbb{Z}.$$

By (3.35), (3.36)

$$\mathbf{H}_1(\boldsymbol{w}_L,\boldsymbol{w}_R,\boldsymbol{n}) = \quad \mathbb{Q}_{-\boldsymbol{n}}^{-1}\mathbb{Z}\left[\mathbb{A}_1^+(\mathbb{Z}\mathbb{Q}_{-\boldsymbol{n}}\boldsymbol{w}_\star)\mathbb{Z}\mathbb{Q}_{-\boldsymbol{n}}\boldsymbol{w}_L + \mathbb{A}_1^-(\mathbb{Z}\mathbb{Q}_{-\boldsymbol{n}}\boldsymbol{w}_\star)\mathbb{Z}\mathbb{Q}_{-\boldsymbol{n}}\boldsymbol{w}_R\right]$$

$$\begin{aligned}&= -\mathbb{Q}_{-\boldsymbol{n}}^{-1}\mathbb{Z}\left[\mathbb{A}_1^{-}\left(\mathbb{Z}\mathbb{Q}_{-\boldsymbol{n}}\boldsymbol{w}_{\star}\right)\mathbb{Z}\mathbb{Q}_{-\boldsymbol{n}}\boldsymbol{w}_L + \mathbb{A}_1^{+}\left(\mathbb{Z}\mathbb{Q}_{-\boldsymbol{n}}\boldsymbol{w}_{\star}\right)\mathbb{Z}\mathbb{Q}_{-\boldsymbol{n}}\boldsymbol{w}_R\right]\\ &= -\mathbf{H}_1\left(\boldsymbol{w}_R, \boldsymbol{w}_L, -\boldsymbol{n}\right).\end{aligned}$$

The conservation property of $\boldsymbol{H}_2$ follows immediately from the identity $\mathbb{Q}\boldsymbol{e}_2 = (0, n_1, n_2)^T$,

$$\mathbf{H}_2\left(\boldsymbol{w}_L, \boldsymbol{w}_R, \boldsymbol{n}\right) = \frac{1}{2}gh_{\star}^2\begin{pmatrix}0\\ \boldsymbol{n}\end{pmatrix} = -\frac{1}{2}gh_{\star}^2\begin{pmatrix}0\\ -\boldsymbol{n}\end{pmatrix} = \mathbf{H}_2\left(\boldsymbol{w}_R, \boldsymbol{w}_L, -\boldsymbol{n}\right).$$

□

Finally, we specify the time step τ^j in (3.22) using the Courant-Friedrichs-Levy (CFL) stability condition (see Feistauer et al., 2003, Section 3.3.11). We approximate the Lipschitz constant of the numerical flux $\mathbf{H}(\cdot,\cdot,\cdot)$ (with respect to the first two arguments) and the mapping $\mathbf{g}(\cdot,\cdot)$ by

$$\mathcal{L}_{\mathbf{H}}\left(\boldsymbol{w}_L, \boldsymbol{w}_R, \boldsymbol{n}\right) = \mathcal{L}_{\mathbf{g}}\left(\mathbb{Q}\boldsymbol{w}_L, \mathbb{Q}\boldsymbol{w}_R\right), \tag{3.39}$$

$$\mathcal{L}_{\mathbf{g}}\left(\boldsymbol{w}_L, \boldsymbol{w}_R\right) = \max\left|\lambda_i\left(\frac{\boldsymbol{w}_L + \boldsymbol{w}_R}{2}\right)\right|, \tag{3.40}$$

where the eigenvalues λ_i are given by (3.10). Then, we approximate the maximal speed of propagation in the finite volume L by

$$\mathcal{L}_L^j = \sum_{\substack{\Gamma\in\mathcal{E}(L)\cap\mathcal{F}_h^I\\ \Gamma=L\cap R,\\ R\in\mathcal{T}_h}} \frac{|\Gamma|}{|L|}\mathcal{L}_{\mathbf{H}}\left(\boldsymbol{w}_L^j, \boldsymbol{w}_R^j, \boldsymbol{n}_{L,\Gamma}\right) + \sum_{\Gamma\in\mathcal{E}(L)\cap\mathcal{F}_h^{\partial\Omega}} \frac{|\Gamma|}{|L|}\mathcal{L}_{\mathbf{H}}\left(\boldsymbol{w}_L^j, \mathcal{B}\left(x_\Gamma, \boldsymbol{w}_L^j\right), \boldsymbol{n}_{L,\Gamma}\right). \tag{3.41}$$

The CFL condition reads

$$\tau^j = \frac{\mathrm{CFL}}{\max_{L\in\mathcal{T}_h}\mathcal{L}_L^j}, \qquad j = 0, 1, \ldots, N_T - 1, \tag{3.42}$$

where CFL $\in (0,1)$.

3.3 FV method for the case $z \neq const.$

We now consider the general case $z \neq const.$ We seek a finite volume scheme, which preserves a class of stationary solutions:

$$h(x,t) = H_0 - z(x), \qquad \boldsymbol{v}(x,t) = 0. \tag{3.43}$$

The function $\boldsymbol{w} = (h, h\boldsymbol{v})^T$ with components given by (3.43) is a solution to (3.4) (the lake at rest solution). As in (Gallouët et al., 2003), we use piecewise-constant approximation of the topography function z,

$$z_K = \frac{1}{|K|}\int_K z(x)\,\mathrm{d}x, \qquad K \in \mathcal{T}_h. \tag{3.44}$$

The discrete version of (3.43) is

$$h_K^j = H_0 - z_K, \qquad \boldsymbol{v}_K^j = 0, \qquad K \in \mathcal{T}_h,\ j = 0, 1, \ldots, N_T - 1. \tag{3.45}$$

In the previous section, we have presented a discretization of the "convective" fluxes $\boldsymbol{f}_s(\cdot)$. It turns out that the numerical flux defined above must be extended in order to preserve the stationary solution (3.45). Then, we must approximate the source term $\boldsymbol{s}(\cdot)$ as well.

Assuming the piecewise-constant topography (3.44), the source term $-gh\nabla z$ is a Dirac distribution concentrated on the mesh faces. We start with the analysis of linear Riemann problem similar to (3.30), (3.31), with Dirac distribution δ (concentrated at the point $x = 0$) as the source term.

Lemma 3.6. *Let $A, B, u_L, u_R \in \mathbb{R}$. Consider the Riemann problem*

$$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} = B\delta, \qquad \textit{in } \mathbb{R} \times (0, \infty), \tag{3.46}$$

$$u(x, 0) = u_0(x) = \begin{cases} u_L, & x < 0, \\ u_R, & x > 0. \end{cases} \tag{3.47}$$

Then there exists u, which solves(3.46), (3.46) *in the following sense:*

(i) $u \in C([0, \infty), \mathfrak{D}')$*, where $\mathfrak{D}'$ is the space of distributions on $\mathbb{R}$,*

(ii) *u satisfies* (3.46) *in the sense of distributions,*

(iii) $u(0) = u_0$.

Moreover, the solution is unique. If $A \neq 0$, then

$$u(x, t) = \begin{cases} u_0(x - At) + B/|A|, & (x - At)x < 0, \\ u_0(x - At), & (x - At)x \geq 0. \end{cases} \tag{3.48}$$

Proof. If $A \neq 0$, then the method of characteristics gives (3.48). If $A = 0$, then the solution is $u = u_0 + Bt\delta$. By direct computation, one can show that **(i)** - **(iii)** indeed holds. The uniqueness can be shown also using the method of characteristics. □

In order to derive an approximate Riemann solver, we need the value of the flux Au on the line $x = 0$. We neglect the fact that, unlike the homogeneous case, the flux is not continuous on the line $x = 0$ and use mean value,

$$\frac{1}{2}\left(\lim_{x\to 0+} Au(x, t) + \lim_{x \to 0-} Au(x, t)\right) = A^+ u_L + A^- u_R + \frac{1}{2} B \operatorname{sgn} A.$$

Note that the result is equivalent to the Vijayasundaram flux (3.27) with additional correction term $\frac{1}{2}B \operatorname{sgn} A$. We generalize the correction term to the full system using the diagonal decomposition of $\mathbb{A}_1$. We put

$$B\delta = -gh(\boldsymbol{n} \cdot \nabla z) \approx -gh(z_R - z_L)\delta$$

and thus derive the derive the numerical flux for the convective part of the inhomogeneous shallow water equations.

$$\begin{aligned}\mathbf{g}_{conv}\left(\boldsymbol{w}_L,\boldsymbol{w}_R,z_L,z_R\right) = \mathbb{A}_1^+(\boldsymbol{w}_\star)\boldsymbol{w}_L + \mathbb{A}_1^-(\boldsymbol{w}_\star)\boldsymbol{w}_R \\ - \frac{1}{2}gh_\star(z_R - z_L)(\operatorname{sgn}\mathbb{A}_1(\boldsymbol{w}_\star))\boldsymbol{e}_2 - \frac{1}{2}gh_{\star\star}^2\boldsymbol{e}_2\end{aligned} \tag{3.49}$$

The matrix $\operatorname{sgn}\mathbb{A}_1$ is defined analogously as the matrices $\mathbb{A}_1^\pm$,

$$\operatorname{sgn}\mathbb{A}_1 = \mathbb{T}\ \mathrm{diag}\left\{\operatorname{sgn}\lambda_1,\operatorname{sgn}\lambda_2,\operatorname{sgn}\lambda_3\right\}\ \mathbb{T}^{-1}. \tag{3.50}$$

where $\mathbb{T}$ and $\Lambda = \mathrm{diag}\left\{\lambda_1,\lambda_2,\lambda_3\right\}$ are given by (3.13). As before, we put $\boldsymbol{w}_\star = \frac{1}{2}\left(\boldsymbol{w}_L + \boldsymbol{w}_R\right)$ and $h_\star = \frac{1}{2}\left(\boldsymbol{w}_L + \boldsymbol{w}_R\right)$. We will specify the value $h_{\star\star} = h_{\star\star}(\boldsymbol{w}_L,\boldsymbol{w}_R)$ later.

Now we are ready to derive finite volume discretization. We integrate (3.14) over the finite volume $L \in \mathcal{T}_h$ and time subinterval (t^j, t^{j+1}). On the left hand side, we follow the same path as in section 3.2,

$$\begin{aligned}&\int_{t^j}^{t^{j+1}}\int_L \left(\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{s=1}^{2}\frac{\partial}{\partial x_s}\boldsymbol{f}_s(\boldsymbol{w})\right)\mathrm{d}x\,\mathrm{d}t \\ &\quad\approx |L|\left[\boldsymbol{w}_L^{j+1} - \boldsymbol{w}_L^j\right] + \tau^j \sum_{\substack{\Gamma\in\mathcal{E}(L)\cap\mathcal{F}_h^I \\ \Gamma = L\cap R, \\ R\in\mathcal{T}_h}} |\Gamma|\mathbf{H}_{conv}\left(\boldsymbol{w}_L^j,\boldsymbol{w}_R^j,z_L,z_R,\boldsymbol{n}_{L,\Gamma}\right) \\ &\quad + \tau^j \sum_{\Gamma\in\mathcal{E}(L)\cap\mathcal{F}_h^{\partial\Omega}} |\Gamma|\mathbf{H}_{conv}\left(\boldsymbol{w}_L^j,\mathcal{B}\left(x_\Gamma,\boldsymbol{w}_L^j\right),z_L,z_L,\boldsymbol{n}_{L,\Gamma}\right),\end{aligned} \tag{3.51}$$

where $\mathbf{H}_{conv}\left(\boldsymbol{w}_L,\boldsymbol{w}_R,z_L,z_R,\boldsymbol{n}\right) = \mathbb{Q}^{-1}\mathbf{g}_{conv}\left(\mathbb{Q}\boldsymbol{w}_L,\mathbb{Q}\boldsymbol{w}_R,z_L,z_R\right)$, with $\mathbb{Q}$ given by (3.26). On the right hand side, we approximate

$$\begin{aligned}\int_{t^j}^{t^{j+1}}\int_L \boldsymbol{s}(x,\boldsymbol{w})\,\mathrm{d}x\,\mathrm{d}t &\approx -\tau^j g h_L^j \int_L \begin{pmatrix}0\\ \nabla z(x)\end{pmatrix}\mathrm{d}x = -\tau^j g h_L^j \int_{\partial L} z(x)\begin{pmatrix}0\\ \boldsymbol{n}\end{pmatrix}\mathrm{dS} \\ &\approx -\tau^j g h_L^j \sum_{\Gamma\in\mathcal{E}(L)} |\Gamma| z_\star(z_L,z_R)\begin{pmatrix}0\\ \boldsymbol{n}_{L,\Gamma}\end{pmatrix},\end{aligned} \tag{3.52}$$

where $z_\star(z_L,z_R) = \frac{1}{2}\left(z_L + z_R\right)$ is an approximation of z on the face Γ. Using the fact $\mathbb{Q}\begin{pmatrix}0\\ \boldsymbol{n}\end{pmatrix} = \boldsymbol{e}_2$, we can put the terms from (3.51) and (3.51) together and rewrite the FV method as follows

$$\boldsymbol{w}_L^{j+1} = \boldsymbol{w}_L^j - \tau^j \sum_{\substack{\Gamma\in\mathcal{E}(L)\cap\mathcal{F}_h^I \\ \Gamma = L\cap R, \\ R\in\mathcal{T}_h}} \frac{|\Gamma|}{|L|}\mathbf{H}_{total}\left(\boldsymbol{w}_L^j,\boldsymbol{w}_R^j,z_L,z_R,\boldsymbol{n}_{L,\Gamma}\right) \tag{3.53}$$

$$- \tau^j \sum_{\Gamma \in \mathcal{E}(L) \cap \mathcal{F}_h^{\partial\Omega}} \frac{|\Gamma|}{|L|} \mathbf{H}_{\mathit{total}} \left(\boldsymbol{w}_L^j, \mathcal{B}\left(x_\Gamma, \boldsymbol{w}_L^j\right), z_L, z_L, \boldsymbol{n}_{L,\Gamma} \right), \quad L \in \mathcal{T}_h, \; j = 0, 1, \dots, N_T - 1,$$

where the "total" flux is defined by

$$\mathbf{H}_{\mathit{total}}\left(\boldsymbol{w}_L, \boldsymbol{w}_R, z_L, z_R, \boldsymbol{n}\right) = \mathbb{Q}^{-1} \mathbf{g}_{\mathit{total}}\left(\mathbb{Q}\boldsymbol{w}_L, \mathbb{Q}\boldsymbol{w}_R, z_L, z_R\right), \tag{3.54}$$

$$\mathbf{g}_{\mathit{total}}\left(\boldsymbol{w}_L, \boldsymbol{w}_R, z_L, z_R\right) = \mathbf{g}_{\mathit{conv}}\left(\boldsymbol{w}_L, \boldsymbol{w}_R, z_L, z_R\right) + g h_L z_\star \boldsymbol{e}_2. \tag{3.55}$$

So far, we did not specify the value $h_{\star\star}$. In (3.34), we used the arithmetic mean $h_{\star\star} = \frac{1}{2}\left(h_L + h_R\right)$. In the following, we show that this choice is no longer sufficient in the case $z \neq const.$

Theorem 3.7. *Assume*

$$h_{\star\star}(\boldsymbol{w}_L, \boldsymbol{w}_R) = \sqrt{\frac{h_L^2 + h_R^2}{2}}, \;\; \textit{whenever } \boldsymbol{w}_L = (h_L, 0, 0)^T, \boldsymbol{w}_R = (h_R, 0, 0)^T. \tag{3.56}$$

Consider the boundary condition (3.18) *with* $q^D(x) \equiv 0$*, or the boundary condition* (3.19)*. Then* (3.45) *is the discrete solution of* (3.53)*.*

Proof. We first analyze the numerical fluxes $\mathbf{H}_{\mathit{total}}$ on the interior interfaces. Let $\Gamma = L \cap R \in \mathcal{F}_h^I$ and $j = 0, 1, \dots, N_T - 1$. We put

$$\begin{aligned}
\boldsymbol{w}_\star &= (h_\star, 0, 0)^T = \frac{1}{2}\left(\boldsymbol{w}_L^j + \boldsymbol{w}_R^j\right), \\
c &= \sqrt{g h_\star}, \\
h_{\star\star} &= \sqrt{\frac{\left(h_L^j\right)^2 + \left(h_R^j\right)^2}{2}}, \\
z_\star &= \frac{1}{2}\left(z_L + z_R\right).
\end{aligned}$$

By (3.9)-(3.13), we have $\mathbb{A}_1(\boldsymbol{w}_\star) = \mathbb{T}\Lambda\backslash\mathbb{T}^{-1}$, where

$$\Lambda\backslash = \begin{pmatrix} -c & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & c \end{pmatrix}, \qquad \mathbb{T} = \begin{pmatrix} 1 & 0 & 1 \\ -c & 0 & c \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbb{T}^{-1} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2c} & 0 \\ 0 & 0 & 1 \\ -\frac{1}{2} & \frac{1}{2c} & 0 \end{pmatrix},$$

Direct computation gives

$$\mathbb{A}^+(\boldsymbol{w}_\star) = \frac{1}{2}\begin{pmatrix} c & 1 & 0 \\ c^2 & c & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbb{A}^-(\boldsymbol{w}_\star) = \frac{1}{2}\begin{pmatrix} -c & 1 & 0 \\ c^2 & -c & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \operatorname{sgn}\mathbb{A}^+(\boldsymbol{w}_\star) = \begin{pmatrix} 0 & c^{-1} & 0 \\ c & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

For the sake of brevity, we omit the time index j, $h_L = h_L^j$ and $h_R = h_R^j$. We get

$$\mathbf{g}_{\mathit{total}}\left(\boldsymbol{w}_L, \boldsymbol{w}_R, z_L, z_R\right) = \frac{1}{2}\begin{pmatrix} c\left(h_L - h_R\right) - g h_\star c^{-1}\left(z_R - z_L\right) \\ c^2\left(h_L + h_R\right) - g h_{\star\star}^2 + 2 g h_L z_\star \\ 0 \end{pmatrix} = \frac{g}{2}\left(2h_\star^2 - h_{\star\star}^2 + 2h_L z_\star\right)\boldsymbol{e}_2.$$

Moreover,

$$\begin{aligned}2h_\star^2 - h_{\star\star}^2 + 2h_L z_\star &= 2\left(\frac{h_L + h_R}{2}\right)^2 - \left(\frac{h_L^2 + h_R^2}{2}\right) + 2h_L \frac{z_L + z_R}{2} \\ &= h_L h_R + h_L\left(z_L + z_R\right) \\ &= h_L(h_R + z_R) + h_L z_L. \\ &= h_L H_0 + h_L z_L.\end{aligned}$$

Then,

$$\mathbf{H}_{total}\left(\boldsymbol{w}_L^j, \boldsymbol{w}_R^j, z_L, z_R, \boldsymbol{n}_{L,\Gamma}\right) = \frac{g}{2}\left(h_L H_0 + h_L z_L\right)\boldsymbol{n}_{L,\Gamma}. \tag{3.57}$$

Due to the choice of boundary conditions, we get the same result for the boundary faces $\Gamma = L \cap \partial\Omega \in \mathcal{F}_h^{\partial\Omega}$,

$$\mathbf{H}_{total}\left(\boldsymbol{w}_L^j, \mathcal{B}\left(x_\Gamma, \boldsymbol{w}_L^j\right), z_L, z_L, \boldsymbol{n}_{L,\Gamma}\right) = \frac{g}{2}\left(h_L H_0 + h_L z_L\right)\boldsymbol{n}_{L,\Gamma}. \tag{3.58}$$

Now,

$$\begin{aligned}&\sum_{\substack{\Gamma\in\mathcal{E}(L)\cap\mathcal{F}_h^I\\ \Gamma=L\cap R,\\ R\in\mathcal{T}_h}} \frac{|\Gamma|}{|L|}\mathbf{H}_{total}\left(\boldsymbol{w}_L^j, \boldsymbol{w}_R^j, z_L, z_R, \boldsymbol{n}_{L,\Gamma}\right) \\ &\qquad + \sum_{\Gamma\in\mathcal{E}(L)\cap\mathcal{F}_h^{\partial\Omega}} \frac{|\Gamma|}{|L|}\mathbf{H}_{total}\left(\boldsymbol{w}_L^j, \mathcal{B}\left(x_\Gamma, \boldsymbol{w}_L^j\right), z_L, z_L, \boldsymbol{n}_{L,\Gamma}\right) \\ &\qquad\qquad = \frac{g}{2|L|}\left(h_L H_0 + h_L z_L\right)\sum_{\Gamma\in\mathcal{E}(E)} |\Gamma|\boldsymbol{n}_{L,\Gamma} = 0.\end{aligned}$$

Since $\boldsymbol{w}_L^j = \boldsymbol{w}_L^{j+1}$ by (3.45), the equation (3.53) holds, and $\{\boldsymbol{w}_L^j\}_{L\in\mathcal{T}_h, j=0,\dots,N_T}$ is indeed the discrete solution. □

We set

$$h_{\star\star}(\boldsymbol{w}_L, \boldsymbol{w}_R) = \begin{cases} h_R, & \mu < -1, \\ \sqrt{\frac{1+\mu}{2}h_L^2 + \frac{1-\mu}{2}h_R^2}, & -1 < \mu < 1, \\ h_L, & \mu > 1, \end{cases} \qquad \begin{aligned}\text{where}\ \ \mu &= \frac{v_{1,\star}}{c_\star}, \\ v_{1\star} &= \frac{v_{1L} + v_{1R}}{2}, \\ c_\star &= \sqrt{gh_\star}, \\ h_\star &= \frac{h_L + h_R}{2}.\end{aligned} \tag{3.59}$$

The choice (3.59) satisfies (3.56). The derivation presented above was motivated by our requirement that the stationary solution (3.43) is preserved. The stationary solution (3.43) is an example of sub-critical ($|\boldsymbol{v}| < c$) flow. Numerical experiments (see appendix B) suggest that the scheme works well also in trans-critical modes ($|\boldsymbol{v}| > c$).

Conclusions

In the first chapter, we have shown that many of the well-known properties of classical Sobolev spaces $W^{1,p}(\Omega)$, such as trace theorems and imbedding theorems, remain valid for the broken Sobolev spaces $W^{1,p}(\Omega, \mathcal{T}_h)$. Obviously, the theory is not complete. It is well known that the traces of functions from $W^{1,p}(\Omega)$ belong to certain fractional-order Sobolev spaces on $\partial\Omega$. The fractional-order Sobolev spaces (the Sobolev-Slobodetskii spaces, and more generally, the Besov Spaces) can be also obtained from the Sobolev Spaces by interpolation.

Concerning the broken Sobolev spaces $W^{k,p}(\Omega, \mathcal{T}_h)$, we only considered the interpolation with respect to the exponent $p \in [1, \infty]$. Let us now discuss the interpolation with respect to the derivative order k. Naturally, we can define the broken Besov Space by interpolation

$$B^{s;p,q}(\Omega, \mathcal{T}_h) = \big(L^p(\Omega), W^{1,p}(\Omega, \mathcal{T}_h)\big)_{s,q}, \qquad s \in (0,1),\ p, q \in [1, \infty].$$

It is not clear to us, whether it is possible to give a direct characterization of this space, i.e. to give a formula for the norm. For simplicity, let us consider the case $p = q = 2$. Then, $B^{s;p,q}(\Omega)$ is the Sobolev-Slobodetskii space $H^s(\Omega)$, with the norm

$$\|v\|^2_{H^s(\Omega)} = \|v\|^2_{L^2(\Omega)} + \iint\limits_{\Omega\times\Omega} \frac{|v(x) - v(y)|^2}{|x-y|^{d+2s}} \,\mathrm{d}x\,\mathrm{d}y. \tag{$\star$}$$

How can one generalize the Sobolev-Slobodetskii norm $(\star)$ to the context of broken spaces? Motivated by (1.104), we are tempted to replace the singular factor $|x - y|$ by $|x - y| + h$. This simple idea itself certainly does not work, but we can use it at least for the "jump penalization" part of the norm. Analogously to (1.64) we express the "jump penalization" part of the norm with the help of the piecewise-constant projection P_h. We hypothesize that the norm in the space $H^s(\Omega, \mathcal{T}_h)$, $s \in (0, 1)$, is equivalent to the following norm (at least if the mesh is globally quasi-uniform):

$$\|v\|^2_{H^s(\Omega,\mathcal{T}_h)} = \sum_{K\in\mathcal{T}_h} \|v\|^2_{H^s(K)} + \iint\limits_{\Omega\times\Omega} \frac{|(P_h v)(x) - (P_h v)(y)|^2}{(|x-y| + h)^{d+2s}} \,\mathrm{d}x\,\mathrm{d}y.$$

However, we have no proof that it is actually the case.

The second open question of the theory of broken Sobolev spaces we consider important is, whether the traces of functions of $W^{1,p}(\Omega, \mathcal{T}_h)$ belong to a suitable fractional-order broken space on $\partial\Omega$. Again, we have no proof that the analogy between classical and broken spaces goes so far.

In the second chapter, we have analyzed the convergence of one-dimensional incomplete interior penalty Galerkin (IIPG) method, with the error measured in the L^2-norm. We have shown that the order of convergence is optimal, if and only if the degree of polynomial approximation is an odd number, and the penalization parameters are chosen in suitable way. We were not able to prove similar result for the NIPG method.

Also, the theory presented here does not extend to multidimensional discretizations. The proof of the one-dimensional Theorem 2.6 is based on two key ingredients: the test functions $w^{\star}_{h,p,k}$ (see (2.42) and below), and the summation by parts (2.64). Note that the test function $w^{\star}_{h,p,k}$ is continuous at all faces (mesh points in 1D) except one. It can be verified that such functions do not exist in 2D or 3D. These considerations suggest that in order to obtain optimal convergence in L^2, the penalization term (2.11) should be modified, as in (Burman and Stamm, 2008). Even with such modification, the analysis of the multidimensional interior penalty methods on nonuniform meshes with higher-order polynomial approximation appears to be plagued with technical obstacles which are not easy to overcome.

In the third chapter, we have constructed a numerical flux of Vijayasundaram type for shallow water equations. Regarding the accuracy, the numerical experiments show that the Vijayasundaram flux is less diffusive than the Lax-Friedrichs flux and the results are comparable with the Osher-Solomon flux. We have proved that the resulting finite volume scheme preserves the stationary solution of type "lake at rest". We did not consider other stationary states. Moreover, it would be interesting to analyze the stability of the discrete stationary state. Certainly, higher order methods are needed. The Vijayasundaram numerical flux can be also employed in the discontinuous Galerkin method.

Appendix A

Numerical evidence for Chapter 2

Within this section we present numerical examples of 1D Poisson problem with Dirichlet boundary conditions, for simplicity, namely

$$-u''(x) = x^{10} \quad x \in (0,1), \qquad x(0) = x(1) = 0. \tag{A.1}$$

For completeness, we employ IIPG, NIPG and SIPG variants of discontinuous Galerkin method although only the IIPG method was analyzed in section 2.2. We carried out computations using p-th degree polynomial approximations, $p = 1, \dots, 5$, on two types of partitions:

- *uniform* with $h = 1/N$, where $N = 24 \cdot 2^k,\ k = 0, \dots, 11$,
- *non-uniform* with $h_{3i} = \frac{1}{16}\frac{3}{N}$, $h_{3i+1} = \frac{5}{16}\frac{3}{N}$, $h_{3i+2} = \frac{10}{16}\frac{3}{N}$, $i = 1, \dots, N/3$, where $N = 24 \cdot 2^k,\ k = 0, \dots, 11$, similar grids were used in (Guzmán and Rivière, 2009).

We investigate the computational errors and the *experimental orders of convergence* (EOC) measured in the $\|\cdot\|_{L^2(\Omega)}$-norm. We employ the several settings of the penalization parameters H_k,

(i) uniform grids, $H_k = h,\ k = 1, \dots, N$, Table A.1,

(ii) non-uniform grids, $H_k = \max(h_{k-1}, h_k),\ k = 1, \dots, N-1$, Table A.2,

(iii) non-uniform grids, $H_k = h_{k-1} + h_k,\ k = 1, \dots, N-1$, Table A.3,

(iv) non-uniform grids, $H_k,\ k = 1, \dots, N-1$ given by (2.39), Table A.4.

The case **(i)** is presented for completeness, the cases **(ii)**, **(iii)** exhibit the usual choices of the penalty parameters and the case **(iv)** corresponds to the optimal choice of H_k (for odd p) introduced in Theorem 2.6.

All computations were carried out with the aid of Objective Caml language (Leroy et al., 2008) and the GNU MP library (Granlund et al., 2007; Monniaux and Filliatre, 2002), in extended precision where the floating-point numbers have at least 300 bits (standard double precision uses 64 bits).

We observe that

- SIPG method has always optimal order of convergence which is in agreement with theoretical results,
- NIPG and IIPG methods have optimal order of convergence for odd p on uniform grids, which is in agreement with results of other authors, see, e.g., (Rivière, 2008, Section 1.5, Table 1.2).
- IIPG method has optimal order of convergence for odd p on non-uniform grids only if the penalty parameters H_k are chosen according (2.39), which is in agreement with Theorem 2.6.
- IIPG method with $H_k = h_{k-1} + h_k$ gives optimal order of convergence for $p = 1$. This is caused by the fact that cases **(iii)** and **(iv)** are identical for $p = 1$.
- NIPG method has only suboptimal order of convergence for odd p on non-uniform grids for all three tested choices of the penalty parameters. The optimal choice of H_k for NIPG (similarly as (2.39) for IIPG) remains an open problem.

p= 1	**NIPG**		**IIPG**		**SIPG**	
N	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC
1536	6.956E-08	2.000	1.410E-08	1.997	8.406E-09	1.994
3072	1.739E-08	2.000	3.529E-09	1.999	2.106E-09	1.997
6144	4.347E-09	2.000	8.826E-10	1.999	5.271E-10	1.998
12288	1.087E-09	2.000	2.207E-10	2.000	1.319E-10	1.999
24576	2.717E-10	2.000	5.518E-11	2.000	3.297E-11	2.000
49152	6.792E-11	2.000	1.380E-11	2.000	8.244E-12	2.000
p= 2						
1536	2.182E-08	2.001	8.206E-10	2.005	2.999E-12	2.998
3072	5.452E-09	2.001	2.048E-10	2.003	3.750E-13	2.999
6144	1.363E-09	2.000	5.114E-11	2.001	4.689E-14	3.000
12288	3.407E-10	2.000	1.278E-11	2.001	5.863E-15	3.000
24576	8.516E-11	2.000	3.194E-12	2.000	7.329E-16	3.000
49152	2.129E-11	2.000	7.985E-13	2.000	9.161E-17	3.000
p= 3						
768	1.025E-13	4.009	2.861E-14	3.998	2.078E-14	3.997
1536	6.383E-15	4.005	1.789E-15	3.999	1.300E-15	3.998
3072	3.983E-16	4.002	1.119E-16	3.999	8.133E-17	3.999
6144	2.488E-17	4.001	6.994E-18	4.000	5.084E-18	4.000
12288	1.554E-18	4.001	4.372E-19	4.000	3.178E-19	4.000
24576	9.711E-20	4.000	2.732E-20	4.000	1.987E-20	4.000
49152	6.069E-21	4.000	1.708E-21	4.000	1.242E-21	4.000
p= 4						
768	2.077E-14	4.000	8.492E-16	4.009	1.031E-17	4.999
1536	1.298E-15	4.000	5.290E-17	4.005	3.223E-19	4.999
3072	8.114E-17	4.000	3.301E-18	4.002	1.008E-20	5.000
6144	5.071E-18	4.000	2.062E-19	4.001	3.149E-22	5.000
12288	3.169E-19	4.000	1.288E-20	4.001	9.840E-24	5.000
24576	1.981E-20	4.000	8.048E-22	4.000	3.075E-25	5.000
49152	1.238E-21	4.000	5.030E-23	4.000	9.610E-27	5.000
p= 5						
768	1.663E-20	6.005	5.211E-21	5.999	4.436E-21	5.999
1536	2.594E-22	6.002	8.145E-23	6.000	6.934E-23	5.999
3072	4.050E-24	6.001	1.273E-24	6.000	1.084E-24	6.000
6144	6.326E-26	6.001	1.989E-26	6.000	1.694E-26	6.000
12288	9.882E-28	6.000	3.108E-28	6.000	2.646E-28	6.000
24576	1.544E-29	6.000	4.856E-30	6.000	4.135E-30	6.000
49152	2.412E-31	6.000	7.588E-32	6.000	6.461E-32	6.000

Table A.1: Computational error and EOC for uniform partitions, $H_k = h,\ k = 1, \ldots, N$

p= 1	**NIPG**		**IIPG**		**SIPG**	
N	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC
1536	1.462E-06	0.760	1.830E-07	1.149	2.144E-08	1.995
3072	7.887E-07	0.890	8.663E-08	1.079	5.370E-09	1.998
6144	4.090E-07	0.947	4.212E-08	1.040	1.344E-09	1.999
12288	2.082E-07	0.974	2.076E-08	1.020	3.360E-10	1.999
24576	1.050E-07	0.987	1.031E-08	1.010	8.403E-11	2.000
49152	5.275E-08	0.994	5.136E-09	1.005	2.101E-11	2.000
p= 2						
1536	2.136E-08	2.026	1.086E-09	2.018	1.794E-11	3.002
3072	5.290E-09	2.013	2.700E-10	2.009	2.241E-12	3.001
6144	1.316E-09	2.007	6.728E-11	2.004	2.801E-13	3.000
12288	3.283E-10	2.003	1.680E-11	2.002	3.501E-14	3.000
24576	8.198E-11	2.002	4.196E-12	2.001	4.375E-15	3.000
49152	2.048E-11	2.001	1.049E-12	2.001	5.469E-16	3.000
p= 3						
768	2.546E-11	2.886	2.992E-13	3.806	1.988E-13	4.005
1536	3.307E-12	2.944	2.446E-14	3.613	1.240E-14	4.003
3072	4.213E-13	2.973	2.357E-15	3.375	7.744E-16	4.001
6144	5.317E-14	2.986	2.584E-16	3.189	4.838E-17	4.001
12288	6.677E-15	2.993	3.040E-17	3.088	3.023E-18	4.000
24576	8.366E-16	2.997	3.694E-18	3.040	1.889E-19	4.000
49152	1.047E-16	2.998	4.557E-19	3.019	1.181E-20	4.000
p= 4						
768	6.598E-14	4.055	3.501E-15	4.041	2.003E-16	5.006
1536	4.043E-15	4.029	2.159E-16	4.019	6.246E-18	5.003
3072	2.501E-16	4.015	1.341E-17	4.009	1.950E-19	5.001
6144	1.555E-17	4.007	8.356E-19	4.005	6.091E-21	5.001
12288	9.695E-19	4.004	5.214E-20	4.002	1.903E-22	5.000
24576	6.052E-20	4.002	3.256E-21	4.001	5.946E-24	5.000
49152	3.780E-21	4.001	2.034E-22	4.001	1.858E-25	5.000
p= 5						
768	1.835E-17	4.916	1.675E-19	5.980	1.507E-19	6.005
1536	5.902E-19	4.959	2.722E-21	5.944	2.351E-21	6.003
3072	1.871E-20	4.980	4.682E-23	5.861	3.670E-23	6.001
6144	5.887E-22	4.990	9.080E-25	5.688	5.731E-25	6.001
12288	1.846E-23	4.995	2.101E-26	5.434	8.953E-27	6.000
24576	5.780E-25	4.997	5.667E-28	5.212	1.399E-28	6.000
49152	1.808E-26	4.999	1.663E-29	5.091	2.185E-30	6.000

Table A.2: Computational error and EOC for non-uniform partitions, $H_k = \max(h_{k-1}, h_k)$, $k = 1, \dots, N-1$

p= 1	**NIPG**		**IIPG**		**SIPG**	
N	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC
1536	2.265E-06	0.791	5.846E-08	1.994	2.082E-08	1.989
3072	1.211E-06	0.903	1.464E-08	1.997	5.224E-09	1.995
6144	6.256E-07	0.953	3.665E-09	1.999	1.308E-09	1.997
12288	3.178E-07	0.977	9.167E-10	1.999	3.274E-10	1.999
24576	1.602E-07	0.989	2.292E-10	2.000	8.189E-11	1.999
49152	8.040E-08	0.994	5.732E-11	2.000	2.048E-11	2.000
p= 2						
1536	2.312E-08	2.029	1.444E-09	2.032	1.751E-11	2.998
3072	5.720E-09	2.015	3.569E-10	2.016	2.191E-12	2.999
6144	1.423E-09	2.008	8.870E-11	2.008	2.740E-13	2.999
12288	3.547E-10	2.004	2.211E-11	2.004	3.426E-14	3.000
24576	8.856E-11	2.002	5.520E-12	2.002	4.282E-15	3.000
49152	2.212E-11	2.001	1.379E-12	2.001	5.353E-16	3.000
p= 3						
768	2.701E-11	2.887	3.878E-13	3.306	1.964E-13	4.001
1536	3.508E-12	2.945	5.540E-14	2.807	1.227E-14	4.000
3072	4.469E-13	2.973	7.672E-15	2.852	7.671E-16	4.000
6144	5.639E-14	2.986	1.013E-15	2.921	4.794E-17	4.000
12288	7.082E-15	2.993	1.301E-16	2.960	2.996E-18	4.000
24576	8.873E-16	2.997	1.649E-17	2.980	1.873E-19	4.000
49152	1.110E-16	2.998	2.076E-18	2.990	1.170E-20	4.000
p= 4						
768	6.746E-14	4.056	4.660E-15	4.065	1.984E-16	5.003
1536	4.131E-15	4.029	2.846E-16	4.033	6.195E-18	5.001
3072	2.556E-16	4.015	1.759E-17	4.017	1.935E-19	5.001
6144	1.589E-17	4.008	1.093E-18	4.008	6.046E-21	5.000
12288	9.905E-19	4.004	6.811E-20	4.004	1.889E-22	5.000
24576	6.182E-20	4.002	4.251E-21	4.002	5.903E-24	5.000
49152	3.861E-21	4.001	2.655E-22	4.001	1.845E-25	5.000
p= 5						
768	1.877E-17	4.916	3.349E-19	5.172	1.498E-19	6.003
1536	6.036E-19	4.959	1.117E-20	4.906	2.337E-21	6.002
3072	1.913E-20	4.980	3.696E-22	4.918	3.650E-23	6.001
6144	6.021E-22	4.990	1.193E-23	4.953	5.702E-25	6.000
12288	1.888E-23	4.995	3.793E-25	4.975	8.908E-27	6.000
24576	5.911E-25	4.997	1.196E-26	4.987	1.392E-28	6.000
49152	1.849E-26	4.999	3.753E-28	4.994	2.175E-30	6.000

Table A.3: Computational error and EOC for non-uniform partitions, $H_k = h_{k-1} + h_k$, $k = 1, \dots, N-1$

p= 1	**NIPG**		**IIPG**		**SIPG**	
N	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC	$\|e_h\|_{L^2(\Omega)}$	EOC
1536	2.265E-06	0.791	5.846E-08	1.994	2.082E-08	1.989
3072	1.211E-06	0.903	1.464E-08	1.997	5.224E-09	1.995
6144	6.256E-07	0.953	3.665E-09	1.999	1.308E-09	1.997
12288	3.178E-07	0.977	9.167E-10	1.999	3.274E-10	1.999
24576	1.602E-07	0.989	2.292E-10	2.000	8.189E-11	1.999
49152	8.040E-08	0.994	5.732E-11	2.000	2.048E-11	2.000
p= 2						
1536	2.202E-08	2.029	1.196E-09	2.028	1.781E-11	3.000
3072	5.451E-09	2.015	2.961E-10	2.014	2.226E-12	3.000
6144	1.356E-09	2.007	7.367E-11	2.007	2.783E-13	3.000
12288	3.380E-10	2.004	1.837E-11	2.004	3.479E-14	3.000
24576	8.440E-11	2.002	4.587E-12	2.002	4.349E-15	3.000
49152	2.109E-11	2.001	1.146E-12	2.001	5.437E-16	3.000
p= 3						
768	2.572E-11	2.884	2.620E-13	4.002	1.984E-13	4.004
1536	3.344E-12	2.943	1.637E-14	4.001	1.238E-14	4.002
3072	4.262E-13	2.972	1.023E-15	4.000	7.735E-16	4.001
6144	5.379E-14	2.986	6.390E-17	4.000	4.833E-17	4.000
12288	6.755E-15	2.993	3.993E-18	4.000	3.020E-18	4.000
24576	8.465E-16	2.997	2.496E-19	4.000	1.887E-19	4.000
49152	1.059E-16	2.998	1.560E-20	4.000	1.180E-20	4.000
p= 4						
768	6.614E-14	4.056	3.590E-15	4.051	2.001E-16	5.005
1536	4.051E-15	4.029	2.205E-16	4.025	6.242E-18	5.002
3072	2.506E-16	4.015	1.367E-17	4.012	1.949E-19	5.001
6144	1.558E-17	4.007	8.507E-19	4.006	6.088E-21	5.001
12288	9.714E-19	4.004	5.306E-20	4.003	1.902E-22	5.000
24576	6.063E-20	4.002	3.313E-21	4.001	5.944E-24	5.000
49152	3.787E-21	4.001	2.069E-22	4.001	1.857E-25	5.000
p= 5						
768	1.837E-17	4.916	1.679E-19	6.005	1.507E-19	6.005
1536	5.907E-19	4.958	2.620E-21	6.002	2.350E-21	6.002
3072	1.873E-20	4.979	4.090E-23	6.001	3.669E-23	6.001
6144	5.893E-22	4.990	6.389E-25	6.001	5.731E-25	6.001
12288	1.848E-23	4.995	9.980E-27	6.000	8.952E-27	6.000
24576	5.786E-25	4.997	1.559E-28	6.000	1.399E-28	6.000
49152	1.810E-26	4.999	2.436E-30	6.000	2.185E-30	6.000

Table A.4: Computational error and EOC for non-uniform partitions, $H_k,\ k = 1, \ldots, N-1$ given by (2.39)

Appendix B

Numerical evidence for Chapter 3

Test problems with $z = 0$

We compare the Vijayasundaram flux proposed in section 3.2 with the Lax-Friedrichs flux and the Osher-Solomon flux (Zhao et al., 1996). We employ the Riemann problem for one-dimensional Shallow Water equations (3.4), restricted to the spatial domain $\Omega = (-1, 1)$

$$\frac{\partial \boldsymbol{w}}{\partial t} + \frac{\partial}{\partial x}\boldsymbol{f}_1(\boldsymbol{w}) = 0, \qquad x \in (-1,1),\ t > 0, \tag{B.1}$$

$$\boldsymbol{w}(x,0) = \begin{cases} (h_L,\, h_L v_{1L},\, h_L v_{2L})^T, & x \in (-1,0), \\ (h_R,\, h_R v_{1R},\, h_R v_{2R})^T, & x \in (0,1), \end{cases} \tag{B.2}$$

with the boundary conditions (3.19). We present numerical solution for two sets of initial datums. In both cases, we compare the discrete solutions computed with the aid of the above mentioned fluxes with analytical solution. The computations are performed on uniform mesh with $N_{ELEM} = 100, 200, 400, 800, 1600, 3200$ elements, and $CFL = 0.9$.

Test problem with continuous solution. The initial datum for the first test problem is given by

$$\begin{aligned} h_L &= 1, & h_R &= 2, \\ v_{1L} &= 0, & v_{1R} &= 2\left(\sqrt{gh_R} - \sqrt{gh_L}\right), \\ v_{2L} &= 0, & v_{2R} &= 0. \end{aligned}$$

The solution of the Riemann problem (B.1),(B.1) reads

$$h(x,t) = \begin{cases} h_L, & \frac{x}{t} < \lambda_{3L}, \\ \frac{1}{g}\left[\sqrt{gh_L} + \frac{1}{3}\left(\frac{x}{t} - \lambda_{3L}\right)\right]^2, & \lambda_{3L} < \frac{x}{t} < \lambda_{3R}, \\ h_R, & \frac{x}{t} > \lambda_{3R}, \end{cases}$$

$$
\begin{aligned}
v_1(x,t) &= \begin{cases} 0, & \frac{x}{t} < \lambda_{3L}, \\ \frac{2}{3}\left(\frac{x}{t} - \lambda_{3L}\right), & \lambda_{3L} < \frac{x}{t} < \lambda_{3R}, \\ v_{1R}, & \frac{x}{t} > \lambda_{3R}, \end{cases} \\
v_2(x,t) &= 0,
\end{aligned}
$$

where $\lambda_{3L} = v_L + \sqrt{gh_L}$ and $\lambda_{3R} = v_L + \sqrt{gh_R}$. In this case, the solution is 3-rarefaction wave. The solution is a continuous function.

The numerical results computed on a mesh with $N_{ELEM} = 100$ elements are shown in Fig. B.1. The numerical results for finer meshes are summarized in Tab. B.1.

Test problem with discontinuous solution . The initial data and the exact solution of the second test problem are given by

$$
\begin{aligned}
&h_L = 2, && h_R = 1, && h(x,t) = \begin{cases} h_L, & \frac{x}{t} < \sqrt{3g}, \\ h_R, & \frac{x}{t} > \sqrt{3g}, \end{cases} \\
&v_{1L} = \frac{1}{2}\sqrt{3g}, && v_{1R} = 0, && v_1(x,t) = \begin{cases} v_{1L}, & \frac{x}{t} < \sqrt{3g}, \\ v_{1R}, & \frac{x}{t} > \sqrt{3g}, \end{cases} \\
&v_{2L} = 1, && v_{2R} = 1, && v_2(x,t) = 1.
\end{aligned}
$$

In this case, the solution is 3-shock wave, with a discontinuity at the line $x - \sqrt{3g}t = 0$. The numerical results computed on a mesh with $N_{ELEM} = 100$ elements are shown in Fig. B.2. The numerical results for finer meshes are summarized in Tab. B.2.

Test problem with $z \neq 0$

The numerical method based on Vijayasundaram flux proposed in section 3.3 was tested on a problem with topography from (Gallouët et al., 2003). The computation domain, topography function, boundary conditions and initial condition were

$$
\begin{aligned}
\Omega &= (0, 25), \\
z(x) &= \begin{cases} 0.2 - 0.05(x - 10.0)^2, & x \in (2.0, 8.0), \\ 0, & \text{otherwise} \end{cases} \\
hv_1(x = 0, t) &= Q_{in} = 1.53, \\
h(x = 25, t) &= h_{out} = 0.66, \\
(h, hv_1, hv_2)^T(x, t = 0) &= (h_{out}, 0, 0)^T.
\end{aligned}
$$

The computation was carried out using an uniform mesh of $N_{ELEM} = 1000$ elements, with $CFL = 0.9$. The stationary solution is depicted in Fig. B.3.

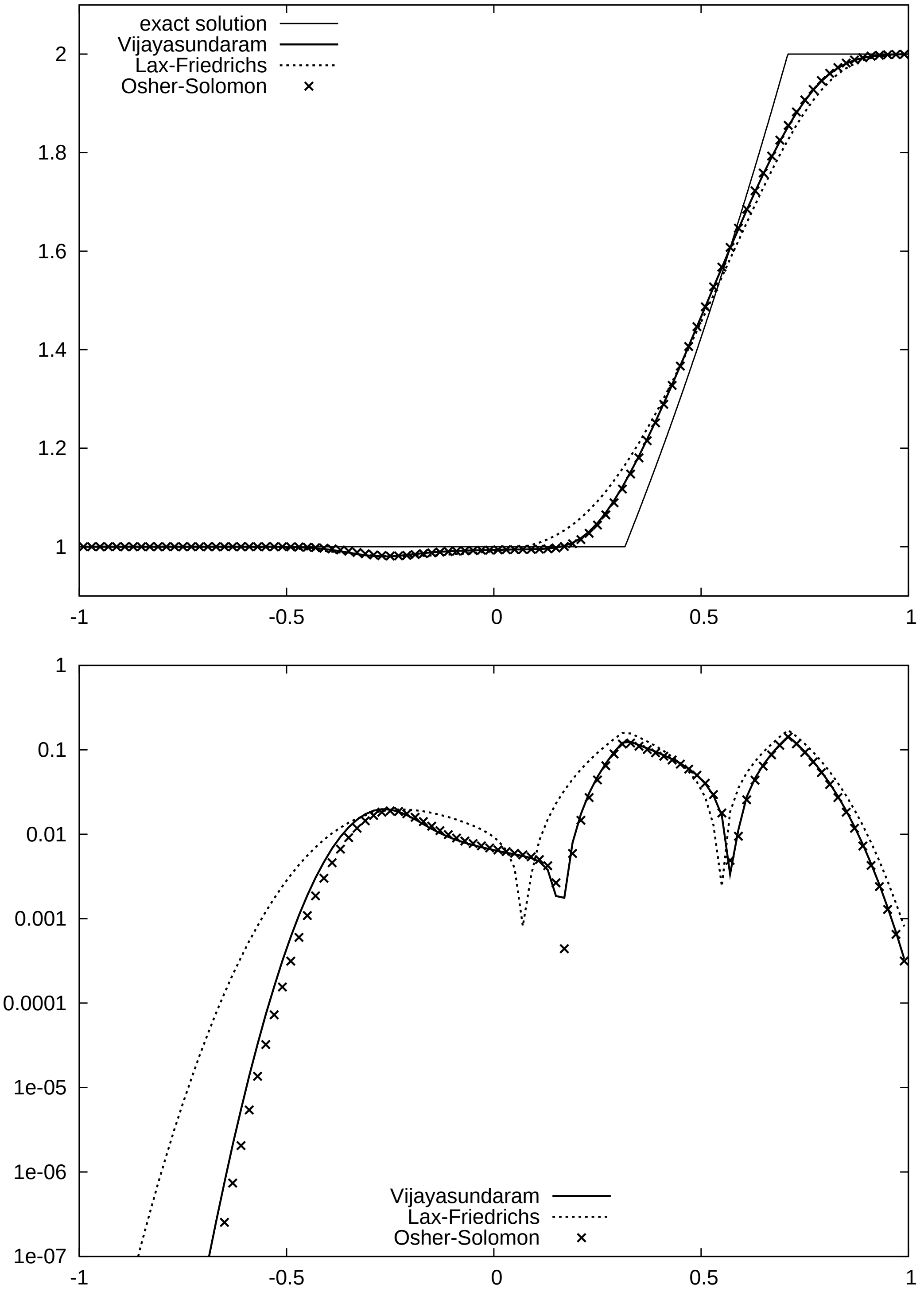

Figure B.1: Rarefaction wave test problem. Top: water height h at $t = 0.1$. Bottom: Discretization error.

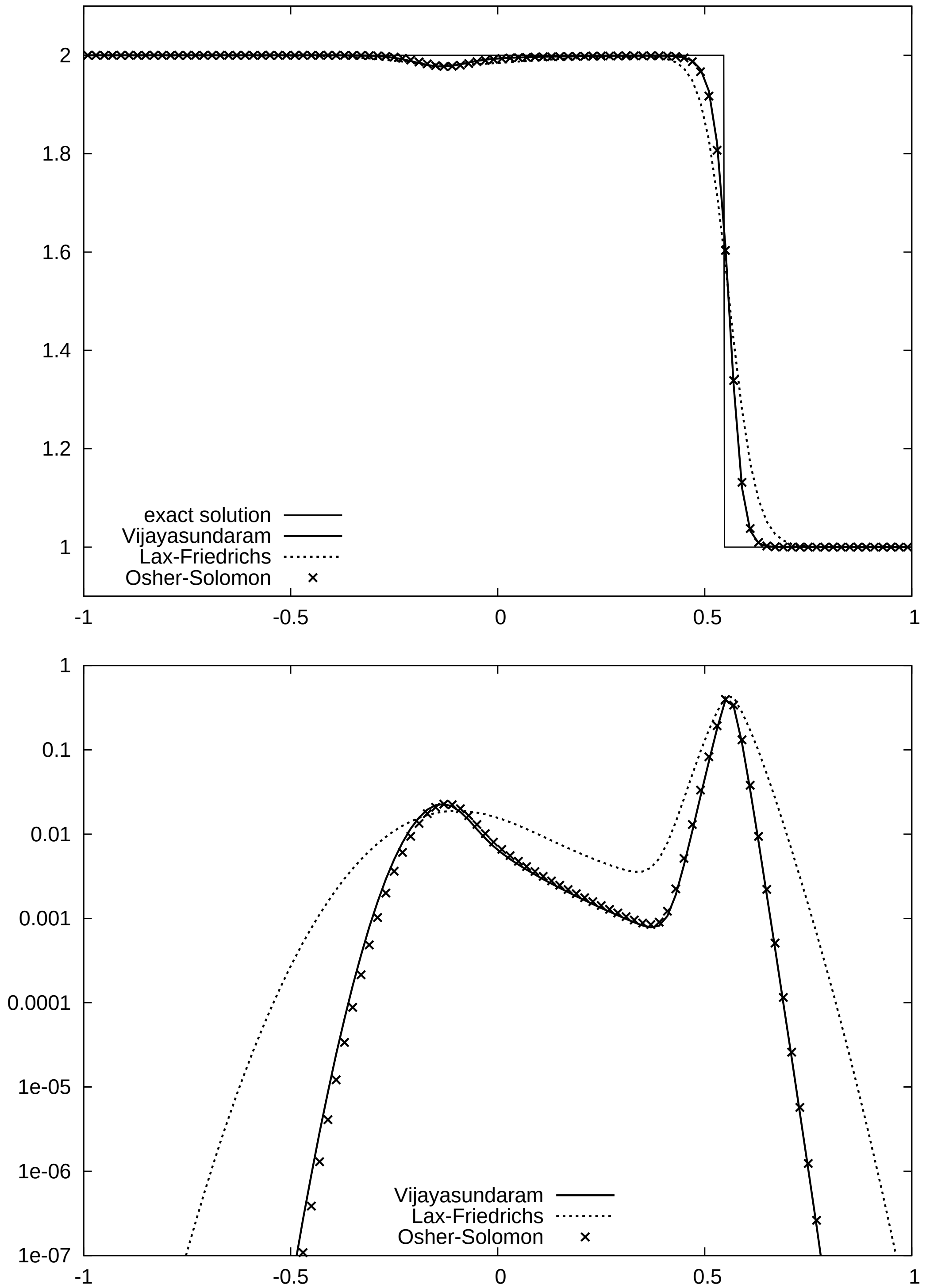

Figure B.2: 3-shock wave test problem. Top: water height h at $t = 0.1$. Bottom: Discretiazation error.

N_{ELEM}	Lax-Friedrichs	Osher-Solomon	Vijayasundaram
100	3.443e-01	2.495e-01	2.569e-01
200	2.196e-01	1.570e-01	1.607e-01
400	1.362e-01	9.639e-02	9.822e-02
800	8.241e-02	5.787e-02	5.876e-02
1600	4.873e-02	3.406e-02	3.449e-02
3200	2.825e-02	1.971e-02	1.992e-02
EOC	0.79	0.79	0.79

Table B.1: 3-rarefaction wave test problem. Discretization error at $t = 0.1$ measured in the L^1-norm, and the corresponding experimental order of convergence (EOC).

N_{ELEM}	Lax-Friedrichs	Osher-Solomon	Vijayasundaram
100	2.407e-01	1.481e-01	1.407e-01
200	1.279e-01	7.612e-02	7.248e-02
400	6.412e-02	3.784e-02	3.600e-02
800	3.168e-02	1.810e-02	1.711e-02
1600	1.604e-02	9.060e-03	8.567e-03
3200	7.916e-03	4.718e-03	4.485e-03
EOC	1.01	0.94	0.93

Table B.2: 3-shock wave test problem. Discretization error at $t = 0.1$ measured in the L^1-norm, and the corresponding experimental order of convergence (EOC).

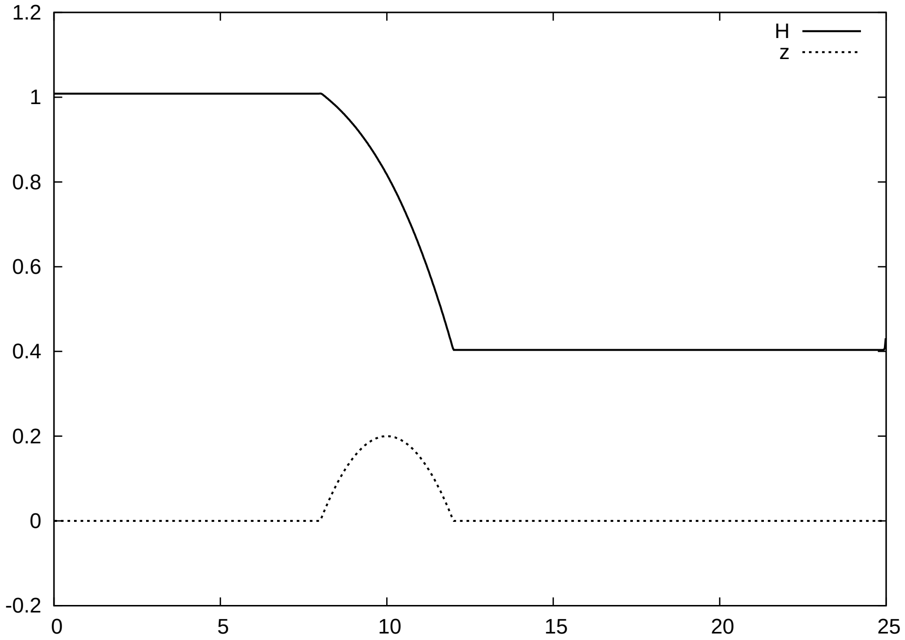

Figure B.3: The free surface H about the bottom z.

Bibliography

Adams, R. A. and Fournier, J. J. F. (2003), *Sobolev Spaces*, second edn, Academic Press, Oxford.

Arnold, D. N. (1982), 'An interior penalty finite element method with discontinuous elements', *SIAM Journal on Numerical Analysis* **19**(4), 742–760.

Arnold, D. N., Brezzi, F., Cockburn, B. and Marini, L. D. (2002), 'Unified analysis of discontinuous Galerkin methods for elliptic problems', *SIAM J. Numer. Anal.* **39**(5), 1749–1779.

Babuška, I., Baumann, C. E. and Oden., J. T. (1999), 'A discontinuous *hp* finite element method for diffusion problems: 1-d analysis.', *Computers and Mathematics with Applications* (37), 103–122.

Bergh, J. and Löfström, J. (1976), *Interpolation Spaces, An Introduction*, Springer, Berlin.

Bernetti, R., Titarev, V. A. and Toro, E. F. (2008), 'Exact solution of the riemann problem for the shallow water equations with discontinuous bottom geometry', *J. Comput. Phys.* **227**(6), 3212–3243.

Brenner, S. C. (2003), 'Poincaré-Friedrichs inequalities for piecewise h^1 functions', *SIAM J. Numer. Anal.* **41**(1), 306–324.

Brenner, S. C., Owens, L. and Sung, L.-Y. (2008), 'A weakly over-penalized symmetric interior penalty method', *Electron. Trans. Num. Anal.* **30**, 107–127.

Brenner, S. C. and Scott, L. R. (2002), *The Mathematical Theory of Finite Element Methods*, second edn, Springer, Berlin.

Buffa, A. and Ortner, C. (2009), 'Compact embeddings of broken Sobolev spaces and applications', *IMA Journal of Numerical Analysis* **29**(4), 827–855.

Burman, E. and Stamm, B. (2008), 'Low order discontinuous Galerkin methods for second order elliptic problems', *SIAM Journal on Numerical Analysis* **47**(1), 508–533.

Canestrelli, A., Siviglia, A., Dumbser, M. and Toro, E. F. (2009), 'Well-balanced high-order centred schemes for non-conservative hyperbolic systems. Applications to shallow water equations with fixed and mobile bed', *Advances in water resources* **32**(6), 834–844.

Chen., H. (2006), ‘Superconvergence properties of discontinuous Galerkin methods for two-point boundary value problems.’, *Int. J. Numer. Anal. Model.* **3**(2), 163–185.

Ciarlet, P. G. (1978), *The finite element method for elliptic problems*, North-Holland.

Cockburn, B., Karniadakis, G. and Shu, C.-W. (2000), The development of discontinuous Galerkin methods, *in* B. Cockburn, G. Karniadakis and C.-W. Shu, eds, ‘Discontinuous Galerkin Methods: Theory, Computation and Applications’, Springer, pp. 3–50.

Dawson, C. N., Sun, S. and Wheeler, M. F. (2004), ‘Compatible algorithms for coupled flow and transport’, *Comput. Meth. Appl. Mech. Engng.* **193**(23-26), 2565–2580.

DeVore, R. and Scherer, K. (1979), ‘Interpolation of linear operators on sobolev spaces’, *Annals of Mathematics* **109**(3), 583–599.

Dolejší, V., Feistauer, M. and Schwab, C. (2002), ‘A finite volume discontinuous Galerkin scheme for nonlinear convection-diffusion problems’, *Calcolo* **39**(1), 1–40.

Dolejší, V., Feistauer, M. and Havle, O. (2009), ‘DGFEM for nonlinear convection-diffusion problems with mixed Dirichlet-Neumann boundary conditions’, *Application of Mathematics* . accepted for publication.

Dolejší, V. and Havle, O. (2010), ‘The L2-optimality of the IIPG method for odd degrees of polynomial approximation in 1D’, *J. Sci. Comput.* **42**(1), 122–143.

Dubois, F. and Floch, P. L. (1988), ‘Boundary conditions for nonlinear hyperbolic systems of conservation laws’, *Journal of Differential Equations* **71**(1), 93–122.

Evans, L. C. (1998), *Partial Differential Equations*, AMS, Providence.

Eymard, R., Gallouët, T. and Herbin, R. (2000), Finite volume methods, *in* P. Ciarlet and J. Lions, eds, ‘Handbook of Numerical Analysis’, Vol. VII, North-Holland, pp. 713–1020.

Feistauer, M., Felcman, J. and Straškraba, I. (2003), *Mathematical and computational methods for compressible flow*, Clarendon Press, Oxford.

Felcman, J. and Havle, O. (2010), ‘On a numerical flux for the shallow water equations’, *Applied Mathematics and Computation* **1**, 1–22. accepted for publication.

Feng, X. and Karakashian, O. A. (2001), ‘Two-level additive Schwarz methods for a discontinuous Galerkin approximation of second order elliptic problems’, *SIAM J. Numer. Anal.* **39**, 1343–1365.

Gallouët, T., Hérard, J.-M. and Seguin, N. (2003), ‘Some approximate godunov schemes to compute shallow-water equations with topography’, *Comput. Fluids* **32**(4), 479–513.

Giusti, E. (1984), *Minimal surfaces and functions of bounded variation*, Birkhäuser.

Granlund, T. et al. (2007), 'The GNU Multiple Precision Arithmetic Library, version 4.2.2'.
URL: *http://gmplib.org/*

Guzmán, J. and Rivière, B. (2009), 'Sub-optimal convergence of non-symmetric discontinuous Galerkin method for odd polynomial approximations', *J. Sci. Comput.* (40), 273–280.

Larson, M. G. and Niklasson, A. J. (2004), 'Analysis of a family of discontinuous Galerkin methods for elliptic problems: the one dimensional case.', *Numer. Math.* **99**(1), 113–130.

Leroy, X. et al. (2008), 'OCAML language, version 3.10.2'.
URL: *http://caml.inria.fr*

LeVeque, R. J. (1990), *Numerical Methods for Conservation Laws*, Birkhäuser, Basel, Boston, Berlin.

Monniaux, D. and Filliatre, J.-C. (2002), 'ML GMP'.
URL: *http://www-verimag.imag.fr/ monniaux/programmes.html.en*

Pietro, D. A. D. and Ern, A. (2010), 'Discrete functional analysis tools for discontinuous Galerkin methods with application to the incompressible navier-stokes equations', *Math. Comp.* **79**, 1303–1330.

Powell, M. J. D. (1981), *Approximation Theory and Methods*, Cambridge University Press.

Rivière, B. (2008), *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation*, SIAM, Philadelphia.

Rivière, B., Wheeler, M. F. and Girault, V. (1999), 'Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems I', *Comput. Geosci.* **3**(3-4), 337–360.

Tartar, L. (2007), *An Introduction to Sobolev Spaces and Interpolation Spaces*, Springer, Berlin.

Toro, E. F. (1997), *Riemann Solvers and Numerical Methods for Fluid Dynamics*, Springer, Berlin.

Toro, E. F., Hidalgo, A. and Dumbser, M. (2009), 'FORCE schemes on unstructured meshes I: Conservative hyperbolic systems', *Journal of Computational Physics* **228**(9), 3368–3389.

Vijayasundaram, G. (1982), Resolution numérique des équations d'Euler pour des écoulements transsoniques avec un schéma de Godunov en éléments finis, PhD thesis, Paris IV.

Wang, K., Wang, H., Sun, S. and Wheeler, M. F. (2009), ‘An optimal-order L^2-error estimate for nonsymmetric discontinuous Galerkin methods for a parabolic equation in multiple space dimensions’, *Computer Methods in Applied Mechanics and Engineering* **198**(27-29), 2190–2197.

Zhao, Shen, Lai and Tabios (1996), ‘Approximate Riemann solvers in FVM for 2D hydraulic shock wave modelling’, *J. Hydraulic Engineering, ASCE* **12**(122), 692–702.