FINITE VOLUME AND FINITE VOLUME ELEMENT METHODS FOR NONSYMMETRIC PROBLEMS

A Dissertation

by

ILIA DIMITROV MICHEV

Submitted to the Office of the Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 1996

Major Subject: Mathematics

FINITE VOLUME AND FINITE VOLUME ELEMENT METHODS FOR NONSYMMETRIC PROBLEMS

A Dissertation

by

ILIA DIMITROV MICHEV

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Raytcho D. Lazarov
(Chair of Committee)

James H. Bramble
(Member)

Marvin L. Adams
(Member)

Richard E. Ewing
(Member)

William Rundell
(Head of Department)

May 1996
Major Subject: Mathematics

ABSTRACT

Mathematical models of many physical processes are desribed with elliptic boundary value problems. An interesting and still not well understood is the class of nonsymmetric problems. Important examples are steady-state convection dominated flows and Navier–Stakes flows with small viscosity. Our goal in this dissertation is to construct and study stable numerical approximations of some nonsymmetric boundary value problems that preserve the important characteristics of the continuous problem such as local mass conservation and monotonicity.

The standard finite element and finite difference methods for convection-dominated problems are stable only for sufficiently small mesh sizes. On Voronoi and curcumscribed cell-centered finite volume meshes we develop three upwind finite difference schemes: upwind, modified upwind and Il'in's. All of the considered schemes are locally mass conservative, unconditionally stable, and satisfy the discrete maximum principle. We show that the upwind scheme has first order of accuracy. All the other schemes are second order accurate in the discrete H^1–norm. We also provide L^2–error estimates utilizing a discrete variant of the Aubin–Nitsche trick.

We derive and study two upwind cell-centered finite difference schemes on locally patch refined meshes. Special attention is paid on the accurate interpolation of the interface between the coarse and fine regions It is shown that the reduction of their accuracy due to interface interpolation is only half an order.

The finite volume element method is another conservative discretization technique for elliptic partial differential equations. We develop a theory for both diffusion dominated and convection dominated problems on 3-D tetrahedral meshes.

Upwind approximations are applied for the discretization of the saturation equation in the total velocity model of two-phase flow in porous media. The linearization strategy is proposed and tested for some model problems.

ACKNOWLEDGMENTS

I am grateful to my advisor Raytcho Lazarov for all the inspiration and encouragement he gave me during the four years I spent in Texas A&M University. He was one of the people that supported my study as a graduate student at Sofia University in Bulgaria and helped me to continue my education in the U.S. He generously shared his experience and knowledge in our discussions and joint work, but even more valuable for me was his consideration and thoughtfulness. He was always ready to help in difficult moments and share the happy ones.

I am deeply indebted to Richard Ewing for his guidance through the years of my graduate studies in Texas A&M University and his support for my research via a research assistantship for almost entire time of my graduate education. He is the person that introduced me to a class of very interesting problems and made it possible for me to take part in one of the grand challenge projects in numerical analysis and scientific computing which enrich considerably my experience and knowledge.

I wish to thank Panayot Vassilevski for his support and care during my graduate study. I was fortunate to begin my graduate education in numerical analysis under his supervision and he provided the opportunity to continue my education in the U.S. His collaboration played a crucial role for the success of the research presented in this thesis. I am grateful for his consideration and the motivation he provided for me.

I had the privilege to work with Joseph Pasciak during my two fruitful visits to Brookhaven National Laboratory in New York and at Texas A&M University. I respect him as a brilliant scientist and a person one always can rely on.

It was pleasure to be around James Bramble in the last two years, in the Numerical Analysis Seminar, or in his class on Multigrid Methods. He was always open to questions and discussions and that definitely contributed to the informal and very inspiring environment in Blocker 505.

I would like to thank Marvin Adams for serving on my graduate committee and being extremely helpful to me in this regard.

My family in Bulgaria always believed in my capabilities, and their love, sacrifice and encouragement make this thesis possible.

My research as a graduate student was supported in part by the US Department of Energy under grant # DE–FG05–92ER25143. Although DOE provided valuable funding, the research in this thesis was not a subject to the Department review and therefore does not necessarily reflect their views and no official endorsement should be inferred.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

A very general, and as the history of science shows, successful way to study natural phenomena is through describing them as models with certain structures that reflect only some of the properties of the originals. With a slight abuse of the language we call these models "physical". Many physical models are written as partial differential equations with some added conditions–initial and/or boundary value problems. The next step is to investigate the qualitative behavior of the mathematical models–existence and uniqueness of the solution, smoothness and so on. Frequently, for further understanding of the properties of a given model, detailed quantitative information for the solution is necessary. In order to obtain this information the continuous problem is approximated with a simpler one, usually a discrete model; the properties of the discrete model are studied and the discrete problem is solved. Usually computations are performed on a particular computer, and therefore computer algorithms that utilize the potential of this computer's architecture must be developed. In order to draw conclusions scientists process the solution, for example, visualize some of its features. The results are compared with the natural phenomenon, and if they are not satisfactory the process is properly modified and repeated.

In this chain of mutually connected objects we distinguish the physical model, the continuous mathematical model, the discrete mathematical model, the computational model and the interpretation block.

In this dissertation we understand *scientific computing* (sometimes called *computer simulation*) as the part of the chain that begins with the physical model and finishes with the interpretation of the discrete solution. Scientific computing necessarily includes branches of the sciences that investigate the natural phenomenon, such as physics, geology, etc., parts of mathematics that study continuous and discrete models, and certain disciplines of computer science that are concerned with the development of computer algorithms and the interpretation of the solution. Frequently the physical model and the corresponding mathematical models are very complicated and the most realistic way to investigate them is through performing physical and numerical experiments. It is a tendency in natural science to reduce the number physical experiments by using numerical computations which are cheaper and much more easily performed on different sets of data. It is commonly accepted that the numerical experiments have to be considered as new tools that help extract the most from physical experiments, but not replace them.

We understand *numerical analysis* as the part of *mathematics* that constructs and studies discrete models of given continuous problems, investigates their properties, in particular how "close" the discrete models are to the continuous, and develops methods for solving the discrete models.

This dissertation is devoted to the study of some problems in *numerical analysis* and the numerical investigation of one class of very complicated practical problems. Thus the subject of this thesis is *scientific computing*.

We construct finite volume methods for a class of nonsymmetric problems and study both theoretically and computationally their properties. These problems describe physical processes that show both diffusion and transportation effects and therefore exhibit features of elliptic and hyperbolic problems. It is essential for the understanding of the mathematical problem to recognize that the transition from an elliptic to a hyperbolic problem is singular, causing the solution of the problem to change rapidly in a very small subdomain. This localized behavior of the solution makes the construction of good numerical methods a challenging problem.

Some examples of such processes are heat or contaminant transport problems with small diffusion effects. Other examples are Navier-Stokes equations with high Reynolds numbers. In these problems the analysis of the convection-diffusion equation is essential. Heat and contaminant transport are described by the equation itself. Navier–Stokes equations are frequently reduced to a coupled system of convection–diffusion and elliptic problems. The nonlinear convection term is treated via iteratively solving a linear nonsymmetric problem [125, 51].

Groundwater is one of the most important sources of drinking, irrigation and industrial process water. However, groundwater supplies are threatened by organic, inorganic and radioactive contaminants introduced by improper disposal or accidental release. Therefore, protection of the quality of groundwater supplies and their remediation is a problem with both economical and social significance.

Remediation methods are usually very expensive. Such strategies as pump-and-treat and in-situ vitrification require accurate knowledge of the location and extent of the contaminant plum. It is prohibitively expensive to monitor the contamination through physical observation. Another alternative is computer simulation of groundwater contaminant transport. As we have already discussed, physical and mathematical models of the underground processes have to be formulated and studied and efficient numerical methods for their solution developed.

Groundwater contamination problems typically involve a broad and complex range of physical, chemical, geochemical, nuclear, and biological processes. Still there is no available rigorous theory (physical model) that explains the behavior of such interconnected phenomena. Thus, *scientific computing* is an indispensable tool for studying such problems, and even defining good physical models.

Many physical models in the engineering practice have a natural mathematical formulation as variational principles, and therefore discrete models based on different formulations of the finite element method will perform very well. In groundwater fluid flow theory mathematical models are derived from the conservation of mass and the application of Darcy's law to each phase. It is desirable that the discrete model inherits the conservation of mass in both a local and global sense, i.e., the discrete model has local conservation of mass. Cell–centered finite difference schemes have local conservation properties and therefore are some of the most used methods in reservoir simulations [9]. Mixed finite element methods on regular meshes also perform very well for such problems.

The convection-diffusion problems we study in this thesis are coercive and satisfy the maximum principle. The maximum principle guarantees that the mathematical model produces physically meaningful solutions. The solvability of the continuous problem follows from the coercivity. Many numerical methods that do not satisfy the discrete maximum principle exhibit non-physical oscillations. Thus, we would like to derive methods that satisfy the discrete maximum principle and produce positive definite matrices.

For groundwater applications it is very common that the coefficients of the equations are discontinuous. Cell–centered finite difference schemes and related mixed finite element methods treat discontinuities in a very successful way, using harmonicly averaged coefficients when the mesh lines are aligned with the discontinuities. Therefore, another requirement for any successful method is that it works for distorted or even arbitrary meshes.

We summarize the desirable properties of numerical methods for nonsymmetric problems in the following list:

(i) stability,

(ii) "good" approximation,

(iii) local conservation,

(iv) satisfy the discrete maximum principle,

(v) produce positive definite matrices,

(vi) work for general domains and arbitrary grids.

In this dissertation we are interested in constructing cell–centered finite volume difference methods and vertex–centered finite volume element methods for model convection–diffusion problems that have the properties $(i) - (vi)$.

First we consider cell–centered finite volume methods. There are many results for cell–centered finite difference approximations of symmetric problems ([115, 43, 132, 135, 123]). For nonsymmetric problems Samarskii [114] has shown convergence in maximum norm for nonsymmetric problems with solutions from $C^4(\Omega)$. Spalding [121] and Runchal [111] have proposed upwind cell–centered methods on uniform meshes, but have not proved stability and error estimates. Herbin [59] has considered cell–centered finite difference schemes on special triangular meshes and has shown first order convergence in the discrete L^2–norm. There are many publications in the engineering and mathematical literature concerning various upwind methods, but we are not aware of works that contain rigorous theory for the schemes that satisfy the properties we have listed above. In fact, Kershaw [70] has shown that it is impossible to derive finite difference schemes that are second-order accurate and satisfy (iv) and (v) on arbitrary quadrilateral meshes. In order to overcome this difficulty we impose some restriction on the meshes. We consider two general classes of meshes – Voronoi and circumscribed grids that can be introduced in general domains. If the grid lines have to be aligned with a given coarse grid triangulation we use constrained Voronoi meshes.

Thus, the *first objective* of this thesis is to provide a theoretical framework for construction and study of conservative, H^1–coercive, monotone, and accurate approximations of convection–diffusion problems on Voronoi and circumscribed grids. Special emphasis is put on developing stability analysis and error estimates for problems with generalized solution from the Sobolev spaces H^s, $\frac{3}{2} < s \leq 3$.

For Voronoi and circumscribed meshes we construct and study three cell–centered finite difference schemes, upwind finite difference scheme (**UDS**), modified upwind difference scheme (**MUDS**), and Il'in's difference scheme (**IDS**), and show that they satisfy (i), (iii), (iv) and (v). We also prove that **MUDS** and **IDS** have second-order of convergence in discrete H^1 and L^2–norms and **UDS** is only first-order accurate. We point out that even for the Laplace equation results for such general meshes are not available in the literature. Our theory also covers the case of symmetric operators. To handle tensor coefficients we propose a generalization of the cell–centered finite difference scheme. We also extend the results of Ewing, Lazarov and Vassilevski [43] for grids with local refinement to nonsymmetric problems.

The *second objective* in this dissertation is to construct and study vertex-centered finite volume element methods for convection-diffusion problems. We note that although cell-centered and vertex-centered grids are dual, the stability and convergence of vertex-centered finite volume methods do not follow from the theory for cell-centered finite volume methods. The theory of finite volume element methods for 2-D symmetric problems has been developed by Bank and Rose [16], Hackbusch [55], Cai and McCormick [87, 28, 26], and Jianguo and Shitong [66]. The only available result for 2–D nonsymmetric problems is due to Hackbusch [55]. He has considered diffusion-dominated problems with convection terms in non-divergence form. We generalize the results for 2–D symmetric problems of Cai and McCormick [26, 28] and Jianguo and Shitong [66] to 2–D(3–D) nonsymmetric equations with convection terms in divergence form. We prove the stability and error estimates for both diffusion and convection dominated cases.

The *third objective* of this thesis is to develop stable numerical methods for the saturation equation in the global pressure/total velocity formulation of the mathematical model for two-phase fluid flow in porous media. This is a part of the Partnership in Computational Science Consortium (PICS) project in Groundwater Contaminant Transport (GCT) directed by Dr. Richard E. Ewing. First, following of Espedal and Ewing results [40] we develop linearization for the saturation equation which is based on a physically meaningful approximation. Then for the linear equation, we propose two different finite element discretizations and implement them in the PICS GCT 1.3 code. Numerical experiments are presented.

This dissertation is organized as follows. In Chapter II we consider elliptic boundary value problems. The contemporary theory of partial differential equations investigates their weak (or generalized) solutions in particular Hilbert spaces that are named after a prominent Soviet mathematician, Sobolev. In Section 2.1 we introduce the necessary notation and state the Sobolev imbedding theorem and the related trace theorem. These results are used as tools in analyzing the proposed numerical methods. Section 2.1.3 contains several theorems that estimate linear and bilinear functionals in Sobolev spaces. These results are usually related to the first such theorem due to Bramble and Hilbert [22].

Section 2.2.1 collects necessary results for abstract variational problems. In Section 2.2.2 we apply these theorems in order to establish the existence and uniqueness results for the nonsymmetric boundary value problems of interest.

In Section 2.3 we discuss the most important properties of the solution of elliptic boundary value problems. We pay special attention to the maximum principle and the conservation properties. In the next chapters we will construct numerical methods that satisfy the discrete analogs of such properties.

Chapter III is devoted to the general definition and discussion of properties of several finite volume methods. To define a particular finite volume method we have to specify two things: the control volumes and the grids associated with them, and the approximation of the fluxes. On the basis of these distinctive features we provide a classification of finite volume methods and an extensive discussion of the literature on this subject (see references). In Section 3.1 we introduce two grids, primary and secondary, and consider the control volumes that can be defined on such grids. Voronoi meshes and their dual Delaunay triangulations are interesting examples . We define various discrete inner products and norms and show that they are equivalent under certain conditions. These discrete norms will be used to estimate the error of approximation of numerical methods in Chapters IV, V and VI.

In Section 3.2 we define the discrete maximum principle and discuss the related notions of monotone and **M**–matrices. The necessary assertions that connect these terms are stated. Next we consider the discrete conservation property and the conditions under which every finite volume method is discretely conservative.

Section 3.3 contains several cell–centered finite volume methods. The basic finite difference scheme for problems with a scalar diffusion coefficient is derived in Section 3.3.1.1 and possible extensions for tensor coefficients are outlined in Section 3.3.1.2. The relations of cell–centered finite difference schemes with mixed finite element methods are discussed in Section 3.3.2. In Section 3.5 we introduce vertex–centered finite volume element methods.

In Chapter IV all basic types of approximation methods for strongly nonsymmetric problems are discussed and compared with respect to the conditions $(i) - (vi)$. In Section 4.1 we state the necessary conditions that finite volume meshes have to fulfill (FV regular triangulations). This condition is a natural extension of the regularity condition for finite element meshes (3.3) to more general finite volume grids. We also state the geometrical conditions, collected in the so called "symmetry assumption", that are sufficient for the higher convergence rate of properly designed finite volume methods.

In Section 4.1 we consider four finite difference schemes: central difference scheme (**CDS**),

upwind finite difference scheme (**UDS**), modified upwind difference scheme (**MUDS**), and Il'in's difference scheme (**IDS**). It is known that **CDS** is not stable for a cell Peclet number larger than 1. We show that **UDS**, **MUDS** and **IDS** satisfy the discrete maximum principle, produce **M**–matrices, and are unconditionally stable. In Section 4.2 we derive estimates for the error of approximation in discrete H^1 and L^2 norms. Section 4.3 contains the results of extensive numerical experiments performed with the **UDS**, **MUDS** and **IDS** schemes. Our theory also provides error estimates for **CDS** when this scheme is stable.

Chapter V is devoted to the analysis of cell–centered finite volume difference schemes with local patch refinement. In Section 5.1 we consider constant and linear interpolation along the interface between refined and non-refined domains and derive the corresponding finite difference schemes, **UDS** and **MUDS**, on the composite grid. In Section 5.2 we prove that these schemes are well defined and produce positive definite matrices. The error estimates are provided in Section 5.3 and the numerical experiments are presented in Section 5.4.

Finite volume element methods for nonsymmetric problems in 3–D are considered in Chapter VI. For diffusion-dominated methods we reformulate the FVE method as the Petrov–Galerkin method and apply the general theory (check the inf–sup condition [11]). The diffusion and convection parts of the FVE method are compared with those of the FE method and the estimates of the difference between them are derived. The inf–sup condition follows from these estimates. For the convection-dominated case we propose the upwind FVE method and with the tools developed in Chapter IV we show that this scheme is stable. We also prove error estimates in the discrete H^1–norm.

In Chapter VII we briefly discuss the conservation laws and the constitutive relations that govern the two phase fluid flow in porous media. We reformulate this model as a global pressure/total velocity model and sketch the derivation of the coupled system of nonlinear parabolic partial differential equations. We discuss two different approaches to handle the compressibility of the air.

Our main goal in this chapter is to develop numerical methods for the saturation equation. We introduce a new macro-dispersion term in the saturation equation that can be considered a result of the up-scaling and heterogeneity of the porous media. Because of this term the saturation equation does not degenerate. The saturation equation is nonlinear and convection dominated, and therefore the most important problems are related to handling the nonlinearity and producing stable approximation. We consider a linearization that takes into account the two separate regimes exhibited by the modeled physical process and is also well suited for sharp fronts. For the linear equation we propose two different finite element discretizations. The first one is based on trilinear finite elements with added artificial diffusion in order to obtain a stable method. The numerical results show that this method produces smearing. The second attempt is an upwind finite element method on tetrahedral meshes, constructed with the finite volume approach developed in the previous chapters.

Finally, in Chapter VIII conclusions and possibilities for future research are provided.

CHAPTER II

ELLIPTIC BOUNDARY VALUE PROBLEMS

Mathematical models of many physical processes are described with elliptic boundary value problems. Interesting and still not well understood is the class of nonsymmetric problems. Important examples include steady state convection dominated flows, certain classes of heat convection problems and the Navier–Stokes flows with small viscosity.

In this chapter we introduce the class of nonsymmetric boundary value problems we will solve numerically in this dissertation. The necessary definitions and theorems from functional analysis are collected in Section 2.1. The abstract variational problems are investigated in Section 2.2 and the results are applied for elliptic problems. We finish the chapter with a short discussion of the properties of the elliptic problems of interest, in particular, maximum principle and conservation properties.

2.1 Sobolev spaces

Many problems in the theory of partial differential equations are naturally formulated and studied in certain functional spaces associated with the name of the Soviet mathematician S. L. Sobolev because of his major contributions to their development in the late 1930s (cf. [120]). Below we state some of their basic properties.

2.1.1 Notation and basic properties

We use the term domain, and usually denote it by Ω, to refer to an open set in d–dimensional, real Euclidean space $\mathbb{R}^d$. A point in $\mathbb{R}^d$ is denoted by $\mathrm{x} = (x_1, x_2, \ldots, x_d)$, its norm is $|\mathrm{x}| = \left(\sum_{i=1}^d x_i^2\right)^{1/2}$ and the inner product of x and y is $(\mathrm{x}, \mathrm{y}) = \sum_{i=1}^d x_i y_i$. If $G \subset \mathbb{R}^d$, we denote by $\bar{G}$ the closure of G in $\mathbb{R}^d$ and by $\overset{\circ}{G}$ the interior of G in $\mathbb{R}^d$. We reserve the symbol ∂G for the boundary of G, i.e., $\partial G = \bar{G} \cap \left(\mathbb{R}^d \setminus G\right)$.

If $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ is an d–tuple of nonnegative integers α_i, we call α a *multi–index* and denote by D^α the differential operator of order $|\alpha|$

$$D^\alpha = D_1^{\alpha_1} \cdots D_d^{\alpha_d},$$

where $D_i^{\alpha_i} = \partial^{\alpha_i} / \partial x_i^{\alpha_i}$ and $|\alpha| = \sum_{i=1}^d \alpha_i$.

Let Ω be a domain in $\mathbb{R}^d$. For any nonnegative integer m let $C^m(\Omega)$ be the vector space consisting of all functions $u(\mathrm{x})$ which, together with all their partial derivatives $D^\alpha u(\mathrm{x})$ of order $|\alpha| \le m$, are continuous on Ω. Clearly, $C^0(\Omega) = C(\Omega)$. Let $C^\infty(\Omega) = \bigcap_{m=0}^\infty C^m(\Omega)$. The subspace $C_0^\infty(\Omega)$ consists of all those functions in $C^\infty(\Omega)$ which have compact support in Ω. We denote with $C_c^\infty(\bar{\Omega})$ the restriction of $C_0^\infty(\mathbb{R}^d)$ to Ω.

We often use $L^p(\Omega)$ spaces defined as classes of all Lebesgue measurable functions u on Ω, for which

$$\int_\Omega |u(\mathrm{x})|^p \, d\mathrm{x} < \infty, \qquad 1 \le p < \infty,$$
$$\operatorname*{ess\,sup}_{x \in \Omega} |u(\mathrm{x})| < \infty. \qquad p = \infty.$$

It is well known fact that $L^p(\Omega), 1 \le p \le \infty$ are Banach spaces equipped with the norms

$$\|u\|_{p.\Omega} = \left\{ \int_\Omega |u(\mathrm{x})|^p \, d\mathrm{x} \right\}^{1/p}, \qquad 1 \le p < \infty,$$
$$\|u\|_{\infty,\Omega} = \operatorname*{ess\,sup}_{x \in \Omega} |u(\mathrm{x})|, \qquad p = \infty.$$

Moreover, $L^2(\Omega)$ is a Hilbert space with an inner product

$$(u, v)_0 = \int_\Omega u(\mathrm{x}) v(\mathrm{x}) \, d\mathrm{x}, \quad u, v \in L^2(\Omega).$$

A function u, defined almost everywhere on Ω, is said to be locally integrable on Ω ($u \in L^1_{\mathrm{loc}}(\Omega)$) provided $u \in L^1(\Omega_1)$ for every $\Omega_1 \subset \Omega$ and $\bar{\Omega}_1$ compact. Let $u \in L^1_{\mathrm{loc}}(\Omega)$. We say that $v_\alpha \in L^1_{\mathrm{loc}}(\Omega)$ is a weak (or distributional) partial derivative of u provided v_α satisfies

$$\int_\Omega u(\mathrm{x}) D^\alpha \phi(\mathrm{x}) \, d\mathrm{x} = (-1)^{|\alpha|} \int_\Omega v_\alpha(\mathrm{x}) \phi(\mathrm{x}) \, d\mathrm{x} \qquad \forall \, \phi \in C_0^\infty(\Omega).$$

We denote v_α by $D^\alpha u$.

For any positive integer m and $1 \le p \le \infty$ Sobolev space $W^{m,p}$ is defined by

$$W^{m,p}(\Omega) = \{ u \in L^p(\Omega) \, : \, D^\alpha u \in L^p(\Omega), \text{ for } 0 \le |\alpha| \le m \}.$$

The norm in $W^{m,p}(\Omega)$ is denoted $\|.\|_{m,p,\Omega}$ and defined by

$$\|u\|_{m,p,\Omega} = \left(\sum_{0 \le |\alpha| \le m} \|D^\alpha u\|^p_{p,\Omega} \right)^{1/p}, \qquad 1 \le p < \infty,$$
$$\|u\|_{m,\infty,\Omega} = \max_{0 \le |\alpha| \le m} \|D^\alpha u\|_{\infty,\Omega}, \qquad p = \infty.$$

In the above definition $D^\alpha u$ is the weak partial derivative of u. We also make frequent use of the seminorms $|.|_{k,p,\Omega}$, $0 \le k \le m$, $1 \le p < \infty$

$$|u|_{k,p,\Omega} = \left(\sum_{|\alpha| = k} \|D^\alpha u\|^p_{p,\Omega} \right)^{1/p}.$$

For the spaces $W^{m,2}(\Omega)$ we use the special notation $H^m(\Omega)$. An equivalent way to construct Sobolev spaces $W^{m,p}(\Omega)$ is by completion of $C^\infty(\Omega)$ with respect to the norms $\|.\|_{m,p,\Omega}$. We define $W_0^{m,p}(\Omega)$ as the closure of $C_0^\infty(\Omega)$ in $W^{m,p}(\Omega)$.

For any integer $m \ge 0$ and any number $0 < \alpha \le 1$ we denote with $C^{m,\alpha}(\bar{\Omega})$ the space of all functions in $C^m(\bar{\Omega})$ whose m-th derivatives satisfy a Hölder condition with exponent α, i.e.,

$$|D^\beta v(\mathrm{x}) - D^\beta v(\mathrm{y})| \le C_\beta \|\mathrm{x} - \mathrm{y}\|^\alpha,$$

where $|\beta| = m$. Equipped with the norm

$$\|v\|_{C^{m,\alpha}(\bar{\Omega})} = \|v\|_{m,\infty,\Omega} + \max_{|\beta| = m} \sup_{\substack{\mathrm{x}, \mathrm{y} \in \bar{\Omega} \\ \mathrm{x} \ne \mathrm{y}}} \frac{|D^\beta v(\mathrm{x}) - D^\beta v(\mathrm{y})|}{\|\mathrm{x} - \mathrm{y}\|^\alpha},$$

the space $C^{m,\alpha}(\bar{\Omega})$ is a Banach space.

The following two results are well known.

Theorem 2.1 *$W^{m,p}(\Omega)$ $(W_0^{m,p}(\Omega))$ is a Banach space for p, $1 \le p \le \infty$.*

Theorem 2.2 *$H^m(\Omega)$ $(H_0^m((\Omega))$ is a Hilbert space with an inner product*

$$(u,v)_m = \sum_{0 \le |\alpha| \le m} (D^\alpha u, D^\alpha v)_0 \,.$$

We denote by $W^{-m,p'}(\Omega)$ the dual space of $W_0^{m,p}(\Omega)$, where $1/p + 1/p' = 1$. The norm in $W^{-m,p'}(\Omega)$ is defined by

$$\|u\|_{-m,p',\Omega} = \sup_{\substack{v \in W_0^{m,p}(\Omega) \\ v \neq 0}} \frac{|<u,v>|}{\|v\|_{m,p,\Omega}} \,.$$

Here we accepted the notation $<u,v>$ to designate the value of the continuous linear functional $u \in W^{-m,p'}(\Omega)$ on the element $v \in W_0^{m,p}(\Omega)$. In future references we will call it duality pairing for the corresponding spaces.

A standard way to introduce Sobolev spaces with real index $s > 0$ is by the so–called real method of interpolation of Lions and Peetre [80], [81]. We accept an alternative approach (cf. [5], [119], [1], [52]) to define $W^{s,p}(\Omega)$ as a space of of all function $u \in W^{[s],p}(\Omega), s = [s] + \sigma$, $[s]$ is the integral part of s and $0 < \sigma < 1$, such that

$$\int_{\Omega \times \Omega} \frac{|D^\alpha u(x) - D^\alpha u(y)|^p}{|x-y|^{d+\sigma p}} \, dx \, dy < \infty$$

for $|\alpha| = [s]$. The norm is defined in $W^{s,p}(\Omega)$

$$\|u\|_{s,p,\Omega} = \left\{ \|u\|^p_{[s],p,\Omega} + \sum_{|\alpha|=[s]} \int_{\Omega \times \Omega} \frac{|D^\alpha u(x) - D^\alpha u(y)|^p}{|x-y|^{d+\sigma p}} \, dx \, dy \right\}^{1/p} .$$

For the spaces $W^{s,p}(\Omega)$ we have the following density result.

Theorem 2.3 ([52]) *Let Ω be a domain in $\mathbb{R}^d$ with a continuous boundary. Then $C_c^\infty(\bar{\Omega})$ is dense in $W^{s,p}(\Omega)$ for all $s > 0$.*

2.1.2 Sobolev imbedding theorems

The most useful properties of spaces $W^{s,p}(\Omega)$, especially in studying differential operators are their imbedding characteristics. The imbeddings of Sobolev spaces depend upon the regularity properties of Ω. In many cases it is enough to characterize the boundary $\partial\Omega$. The next definition is sufficient for most of our subsequent purposes whenever some smoothness of the boundary is required (cf. [94]).

Definition 2.1 (Lipschitz–continuous boundary [31]) We say that a domain Ω has a Lipschitz–continuous boundary $\partial\Omega$ if the following conditions are fulfilled: there exist positive constants α and β, and a finite number of local coordinate systems and local maps a_r, $1 \le r \le R$, which are Lipschitz–continuous on their respective domains of definitions $\{\hat{x}^r \in \mathbb{R}^{d-1} : |\hat{x}^r| \le \alpha\}$, such that

$$\begin{aligned}
&\partial\Omega = \bigcup_{r=1}^{R} \{(x_1^r, \hat{x}^r);\ x_1^r = a_r(\hat{x}^r), \quad |\hat{x}^r| < \alpha\} \,, \\
&\{(x_1^r, \hat{x}^r);\ a_r(\hat{x}^r) < x_1^r < a_r(\hat{x}^r) + \beta; \quad |\hat{x}^r| < \alpha\} \subset \Omega \,, \quad 1 \le r \le R \,, \\
&\{(x_1^r, \hat{x}^r);\ a_r(\hat{x}^r) - \beta < x_1^r < a_r(\hat{x}^r); \quad |\hat{x}^r| < \alpha\} \subset (\mathbf{R} \backslash \Omega) \,,
\end{aligned}$$

where $\hat{x}^r = (x_2^r, \ldots, x_d^r)$, and $|\hat{x}^r| < \alpha$ stands for $|x_i^r| < \alpha$, $2 \le i \le d-1$.

If the functions a_r are of class C^m in their domain of definition we say that $\partial\Omega$ is of class C^m.

Definition 2.2 (Continuous imbedding) We say that the normed space X is continuously imbedded in the norm space Y, and write $X \hookrightarrow Y$ to designate this imbedding, provided that X is contained in Y with continuous injection; i.e., there is a positive constant C such that

$$\|x\|_Y \leq C\|x\|_X\,.$$

We frequently use the following fundamental result.

Theorem 2.4 (Sobolev Imbedding theorem) *Let Ω be a bounded domain with Lipschitz–continuous boundary in $\mathbb{R}^d$. Let $s > 0$ and $1 < p \leq d$.*

$$\text{If } d > sp \text{ then } W^{s,p}(\Omega) \hookrightarrow L^r(\Omega) \text{ for } p \leq r \leq dp/(d-sp). \tag{2.1a}$$

$$\text{If } d = sp \text{ then } W^{s,p}(\Omega) \hookrightarrow L^r(\Omega) \text{ for } p \leq r < \infty. \tag{2.1b}$$

$$\text{If } j < s - d/p < j+1 \text{ for some nonnegative integer } j \text{ then } W^{s,p}(\Omega) \hookrightarrow C^{j,\alpha}(\bar{\Omega}),\ \text{where } \alpha = s - j - d \tag{2.1c}$$

We recall that the elements in Sobolev spaces are equivalence classes, and therefore, the relation (2.1c) means that each equivalence class u in $W^{s,p}(\Omega)$, $s > d/p$ contains a continuous member in the corresponding space. Therefore, u has well defined values on each subset of Ω. We consider also values (traces) of functions in Sobolev spaces in the following weak sense.

Let Ω^k be a k–dimensional domain, $1 \leq k \leq d$, $\Omega^k \subset \Omega$ and let $\gamma : W^{s,p} \to W^{j,q}(\Omega^k)$ be a linear operator with the property that if $\lim_{n\to\infty} \|u - u_n\|_{m,p,\Omega} = 0$, $u_n \in C_c^\infty(\bar{\Omega})$ then $\lim_{n\to\infty} \|\gamma u - u_n\|_{j,q,\Omega^k} = 0$ and

$$\|\gamma u\|_{j,q,\Omega^k} \leq K\|u\|_{s,p,\Omega}$$

with a constant K independent of u. Then $\gamma u \in W^{j,q}(\Omega^k)$ is called trace of $u \in W^{s,p}(\Omega)$. In fact, γ is an unique continuous extension of the mapping $u(\Omega) \to u(\Omega^k)$ defined for smooth functions. In the following theorem we state the conditions when such mapping exists.

Theorem 2.5 (Trace theorem [1]) *Let Ω be a sufficiently regular domain in $\mathbb{R}^d$ and let Ω^k be the k–dimensional domain obtained by intersecting Ω with a k–dimensional plane in $\mathbb{R}^d$, $1 \leq k \leq d$. Let $s > 0$, $1 < p \leq q < \infty$, and $\chi = s - (d/p) + (k/q)$. If*

$$(i)\quad \chi \geq 0 \text{ and } p < q, \text{ or}$$
$$(ii)\quad \chi > 0 \text{ and } \chi \text{ is not an integer, or}$$
$$(iii)\quad \chi > 0 \text{ and } 1 < p \leq 2,$$

then (direct imbedding theorem)

$$W^{s,p}(\Omega) \hookrightarrow W^{\chi,q}(\Omega^k)\,. \tag{2.2a}$$

Imbedding (2.2a) *does not necessarily hold for $p = q > 2$ and χ nonnegative integer. Conversely, if $p = q$ and if either*

$$(iv)\quad \chi = s - (d-k)/p > 0 \text{ and is not an integer, or}$$
$$(v)\quad \chi \geq 0 \text{ and } p \geq 2,$$

then we have the reverse imbedding

$$W^{\chi,q}(\Omega^k) \hookrightarrow W^{s,p}(\Omega) \tag{2.2b}$$

in the sense that each $u \in W^{\chi,q}(\Omega^k)$ *is the trace on* Ω^k *of a function* $w \in W^{s,p}(\Omega)$ *satisfying*

$$\|w\|_{s,p,\Omega} \leq K\|u\|_{\chi,p,\Omega^k}$$

with K *independent of* u.

Remark 2.1 Suppose that Ω is a sufficiently regular domain and the domain $\Omega_1 \subset \Omega$ has piecewise linear boundary, i.e., each segment is of the type of Ω^k described above. As a simple corollary of the "trace theorem" we obtain that if $\mathbf{W} = (W_1, \ldots, W_d) \in (H^s(\Omega))^d$, $s > 1/2$ then $(\mathbf{W}, \mathbf{n}) \in H^{s-1/2}(\partial\Omega_1) \subset L^2(\partial\Omega_1)$, where $\mathbf{n}$ is the outward normal unit vector to $\partial\Omega_1$. If $\mathbf{W} = \nabla u$, then u has to be in the space $H^{s+1}(\Omega)$.

Another approach to work with vector functions is outlined below. The space $H_{div}(\Omega)$ is defined via

$$H_{div}(\Omega) = \{\mathbf{v} = (v_1, \ldots, v_d) : v_i \in L^2(\Omega), \operatorname{div}(\mathbf{v}) \in L^2(\Omega)\}$$

and is a Hilbert space with the norm

$$\|\mathbf{v}\|_{H_{div}(\Omega)} = \left(\|\mathbf{v}\|^2_{0,\Omega} + \|\operatorname{div}(\mathbf{v})\|^2_{0,\Omega}\right)^{1/2} .$$

The following theorem due to Thomas [126] answers the questions about the traces of functions from $H_{div}(\Omega)$.

Theorem 2.6 *The mapping* $\mathbf{v} \to (\mathbf{v}, \mathbf{n})$ *defined a priori from* $\left(H^1(\Omega)\right)^d$ *into* $L^2(\partial\Omega)$ *can be extended to a continuous linear mapping from* $H_{div}(\Omega)$ *onto* $H^{-1/2}(\partial\Omega)$. *Further we have the following characterization of the norm on* $H^{-1/2}(\partial\Omega)$*:*

$$\|\mu\|_{-1/2,\partial\Omega} = \inf_{\substack{\mathbf{v}\in H_{div}(\Omega) \\ (\mathbf{v},\mathbf{n})=\mu}} \|\mathbf{v}\|_{H_{div}(\Omega)} .$$

Therefore, instead of the integral $\int_{\partial\Omega}(\mathbf{v}, \mathbf{n})\, ds$ we can consider the duality pairing $<(\mathbf{v}, \mathbf{n}), 1>$ between the spaces $H^{-1/2}(\partial\Omega)$ and $H^{1/2}(\partial\Omega)$. In general, we prefer to work with functions in $L^2(\partial\Omega)$ and define the Hilbert space

$$\mathcal{H}_{div}(\Omega) = \{\mathbf{v} \in H_{div}(\Omega) : (\mathbf{v}, \mathbf{n}) \in L^2(\partial\Omega)\}$$

with the norm

$$\|\mathbf{v}\|_{\mathcal{H}_{div}(\Omega)} = \left(\|\mathbf{v}\|^2_{H_{div}(\Omega)} + \|(\mathbf{v}, \mathbf{n})\|^2_{0,\partial\Omega}\right)^{1/2} .$$

2.1.3 Bramble–Hilbert theorems

In our analysis of certain numerical methods we represent the error as a functional in some Sobolev spaces and estimate that functional using the well known results from functional analysis.

We denote with $P_k(\Omega)$ the space of all polynomials of degree $\leq k$ in each variable. We say that a linear form (functional) $f(.) : \mathcal{V} \to \mathbb{R}$ is continuous if the following inequality holds for every $v \in \mathcal{V}$

$$|f(v)| \leq \|f\|_{\mathcal{V}'}\|v\|_{\mathcal{V}} .$$

where $\|.\|_{\mathcal{V}'}$ is the norm in the dual space $\mathcal{V}'$ of $\mathcal{V}$. Similarly, we say that a bilinear form $\mathcal{A}(.,.) : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ is continuous if there exists a constant C such that for every $u \in \mathcal{U}$ and $v \in \mathcal{V}$ the following inequality holds

$$|\mathcal{A}(u,v)| \leq C\|u\|_{\mathcal{U}}\|v\|_{\mathcal{V}} .$$

The smallest constant C is the norm $\|\mathcal{A}\|$ of the bilinear form $\mathcal{A}$ in the space $\mathcal{L}_2(\mathcal{U} \times \mathcal{V}; \mathbb{R})$.

Theorem 2.7 (Bramble–Hilbert lemma [22]) *Let Ω be a domain in $\mathbb{R}^d$ with Lipschitz–continuous boundary.*

(i) For some integer $k \geq 0$ and some number $p \in [0, \infty]$, let $f(.)$ be a continuous linear form on the space $W^{k+1,p}(\Omega)$.

(ii) Let the linear form $f(.)$ satisfy

$$\forall p \in P_k(\Omega) , \qquad f(p) = 0 .$$

Then there is a constant $C(\Omega)$ such that

$$|f(v)| \leq C(\Omega)\|f\|^*_{k+1,p,\Omega}|v|_{k+1,p,\Omega} ,$$

where $\|.\|^*_{k+1,p,\Omega}$ is the norm in the dual space of $W^{k+1,p}(\Omega)$.

If the linear form $f(.)$ vanishes only for polynomials of degree $\leq k-1$ then the following modification of Theorem 2.2 holds [57].

Theorem 2.8 (Modified Bramble–Hilbert lemma) *Let Ω be a domain in $\mathbb{R}^d$ with Lipschitz–continuous boundary.*

(i) Let the assumption (i) of Theorem 2.8 be fulfilled, with $k \geq 1$.

(ii) Let the linear form $f(.)$ satisfy

$$\forall p \in P_{k-1}(\Omega) , \qquad f(p) = 0 .$$

Then there is a constant $C(\Omega)$ such that

$$|f(v)| \leq C(\Omega)\|f\|^*_{k+1,p,\Omega} \left(|v|_{k,p,\Omega} + |v|_{k+1,p,\Omega}\right) .$$

A similar result for bilinear forms is stated in the following theorem.

Theorem 2.9 (The bilinear lemma [31]) *Let Ω be a domain in $\mathbb{R}^d$ with Lipschitz continuous boundary. Let $\mathcal{A}(.,.)$ be a continuous bilinear form over the space $W^{k+1,p}(\Omega) \times W$, where the space W satisfies the inclusions*

$$P_l(\Omega) \subset W \subset W^{l+1,q}(\Omega),$$

and is equipped with the norm $\|.\|_{l+1,q,\Omega}$. We assume that

$$\begin{array}{llll} \forall p \in P_k(\Omega) , & \forall w \in W , & \mathcal{A}(p,w) = 0 , \\ \forall v \in W^{k+1,p}(\Omega) , & \forall q \in P_l(\Omega) , & \mathcal{A}(v,q) = 0 . \end{array}$$

Then there exists a constant $C(\Omega)$ such that

$$|\mathcal{A}(u,w)| \leq C(\Omega)\|\mathcal{A}\||v|_{k+1,p,\Omega}|w|_{l+1,q,\Omega} , \quad \forall v \in W^{k+1,p}(\Omega) , \forall w \in W ,$$

where $\|\mathcal{A}\|$ is the norm of the bilinear form $\mathcal{A}(,.)$ in the space $\mathcal{L}_2(W^{k+1,p}(\Omega) \times W; \mathbb{R})$.

2.2 Abstract variational problems

We consider several abstract variational problems that are closely connected with the non-symmetric boundary value problems we investigate in this thesis. We formulate the general theorems for existence and uniqueness in Hilbert space framework and state the conditions that spaces and bilinear form should satisfy. These results are applied to investigate solvability of particular partial differential equations.

Let $\mathcal{U}$ and $\mathcal{V}$ be two real Hilbert spaces with norms $\|.\|_{\mathcal{U}}$ and $\|.\|_{\mathcal{V}}$ respectively, and let $\mathcal{A} : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ be a bilinear form. We define the following variational problem:

Find an element $u \in \mathcal{U}$ such that

$$\mathcal{A}(u, v) = f(v)\,, \qquad \forall\, v \in \mathcal{V}\,. \tag{2.3}$$

We also consider generalized saddle point problems. Assume that $\mathcal{U}_i$, $\mathcal{V}_i$, $i = 1, 2$ are real Hilbert spaces and the following bilinear forms

$$\mathcal{A} : \mathcal{U}_1 \times \mathcal{U}_2 \to \mathbb{R}, \qquad \mathcal{B}_1 : \mathcal{V}_1 \times \mathcal{U}_2 \to \mathbb{R}\,, \qquad \mathcal{B}_2 : \mathcal{V}_2 \times \mathcal{U}_1 \to \mathbb{R} \tag{2.4}$$

and the linear form $f : \mathcal{U}_2 \to \mathbb{R}$ are defined. Consider the problem:

Find a pair $(u, v) \in \mathcal{U}_1 \times \mathcal{V}_1$ such that

$$\mathcal{A}(u, u_2) + \mathcal{B}_1(v, u_2) = f(v_2) \quad \forall\, u_2 \in \mathcal{U}_2 \tag{2.5a}$$

$$\mathcal{B}_2(v_2, u) = 0 \qquad \forall\, v_2 \in \mathcal{V}_2 \tag{2.5b}$$

In the following section we state a few theorems that answer the question whether the problems (2.3) and (2.5) can be solved.

2.2.1 Existence and uniqueness theorems

First we consider the case when $\mathcal{U} \equiv \mathcal{V}$. We say that a bilinear form $\mathcal{A}(.,.) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is $\mathcal{V}$–elliptic (coercive) if there exists a positive number α such that

$$\alpha \|u\|_{\mathcal{V}}^2 \le \mathcal{A}(u, u) \qquad \forall\, u \in \mathcal{V}\,.$$

Theorem 2.10 (Lax–Milgram lemma [31]) *Let $\mathcal{V}$ be a Hilbert space, let $\mathcal{A}(.,.) : \mathcal{V} \times \mathcal{V} \to R$ is a continuous $\mathcal{V}$–elliptic bilinear form, and let $f(.) : \mathcal{V} \to \mathbb{R}$ be a continuous linear form. Then the abstract variational problem* (2.3) *has one and only one solution and the following stability estimate holds:*

$$\|u\|_{\mathcal{V}} \le \frac{1}{\alpha} \|f\|_{\mathcal{V}'}\,.$$

We also need the generalization of Lax–Milgram lemma due to Nečas [94] and modified by Babuška and Aziz [11].

Theorem 2.11 *Let $\mathcal{U}$ and $\mathcal{V}$ be two real Hilbert spaces with norms $\|.\|_{\mathcal{U}}$ and $\|.\|_{\mathcal{V}}$, respectively. Assume that there exist a positive constants α such that the continuous bilinear form $\mathcal{A} : \mathcal{U} \times \mathcal{V} \to R$ satisfies*

$$\sup_{\substack{v \in \mathcal{V} \\ v \neq 0}} \frac{|\mathcal{A}(u, v)|}{\|v\|_{\mathcal{V}}} \ge \alpha \|u\|_{\mathcal{U}} \quad \forall\, u \in \mathcal{U}\,, \tag{2.6a}$$

$$\sup_{u \in \mathcal{U}} |\mathcal{A}(u, v)| > 0 \quad \forall\, v \in \mathcal{V},\, v \neq 0\,, \tag{2.6b}$$

and assume that $f(.) : \mathcal{V} \to \mathbb{R}$ *is a continuous linear form. Then the abstract variational problem* (2.3) *has one and only one solution and the following stability estimate holds:*

$$\|u\|_{\mathcal{U}} \leq \frac{1}{\alpha}\|f\|_{\mathcal{V}'}\,.$$

Remark 2.2 The most difficult of the conditions (2.6) is (2.6a). This condition is frequently represented in the following equivalent form:

$$\inf_{\substack{u\in\mathcal{U}\\ u\neq 0}} \sup_{\substack{v\in\mathcal{V}\\ v\neq 0}} \frac{|\mathcal{A}(u,v)|}{\|u\|_{\mathcal{U}}\|v\|_{\mathcal{V}}} \geq \alpha > 0$$

and is called *Ladyzhenskaya–Babuŝka–Brezzi* or *inf–sup* condition. The general scheme to check (2.6a) is to construct a mapping $T_1 : \mathcal{U} \to \mathcal{V}$ and such that

$$|\mathcal{A}(u, T_1u)| \geq \alpha\|u\|_{\mathcal{U}}\|T_1u\|_{\mathcal{V}} \tag{2.7a}$$

with $\alpha > 0$. Similarly, we construct a mapping $T_2 : \mathcal{V} \to \mathcal{U}$ and such that

$$|\mathcal{A}(T_2u, u)| > 0\,. \tag{2.7b}$$

Clearly, conditions (2.7a) and (2.7b) imply the conditions (2.6a) and (2.6b).

Now we consider some applications of Theorem 2.10 to the generalized saddle point problem (2.5) due to Nicolaides [95] (see also [21]). Consider the problem (2.3) with and additional restriction

$$\mathcal{A}(u, v) = 0 \qquad \forall\, v \in \mathcal{V}\,,\ u \in \mathcal{Z}\,, \tag{2.8}$$

where $\mathcal{U} = \mathcal{Z} \oplus \mathcal{W}$ and $\mathcal{W} = \mathcal{Z}^{\perp}$. We call $\mathcal{Z}$ a *null* space of the bilinear form $\mathcal{A}$.

Corollary 2.1 *Let* $\mathcal{A} : \mathcal{U} \times \mathcal{V} \to R$ *be continuous bilinear form that satisfy the conditions* (2.6) *for the spaces* $\mathcal{W}$ *and* $\mathcal{V}$ *and assume that* $f(.) : \mathcal{V} \to \mathbb{R}$ *is a continuous linear form. Then the abstract variational problem* (2.3), (2.8) *has one and only one solution in* $\mathcal{W}$ *and the following stability estimate holds:*

$$\|u\|_{\mathcal{U}} \leq \frac{1}{\alpha}\|f\|_{\mathcal{V}'}\,.$$

Theorem 2.12 ([95]) *Assume that* $\mathcal{U}_i$, $\mathcal{V}_i$, $i = 1, 2$ *are real Hilbert spaces and the continuous bilinear forms* $\mathcal{A}$, $\mathcal{B}_1$ *and* $\mathcal{B}_2$ *are defined in the corresponding spaces* (2.4). *Let* $\mathcal{Z}_i$ *be the null spaces of* $\mathcal{B}_i$, $i = 1, 2$. *Moreover, suppose that* $\mathcal{B}_1$ *and* $\mathcal{B}_2$ *satisfy the conditions* (2.6) *for the spaces* $\mathcal{V}_1, \mathcal{U}_2$ *and* $\mathcal{V}_2, \mathcal{U}_1$ *with constants* β_1 *and* β_2, *respectively, and* $\mathcal{A}$ *fulfills the conditions* (2.6) *for the spaces* $\mathcal{Z}_1$ *and* $\mathcal{Z}_2$ *with a constant* α *and assume that* $f(.) : \mathcal{U}_2 \to \mathbb{R}$ *is a continuous linear form. Then the abstract variational problem* (2.5) *has one and only one solution and the following stability estimate holds:*

$$\|u\|_{\mathcal{U}_1} \leq \frac{1}{\alpha}\|f\|_{\mathcal{U}_2'}\,, \qquad \|v\|_{\mathcal{U}_2} \leq \left(\frac{\alpha + \|\mathcal{A}\|}{\alpha\beta_1}\right)\|f\|_{\mathcal{U}_2'}\,. \tag{2.9}$$

2.2.2 Applications to elliptic boundary value problems

Let $\Omega \in \mathbb{R}^d$ be a bounded domain. We consider the operator of the form

$$\mathcal{L}u = \ \mathrm{div}(-A(\mathrm{x})\nabla u(\mathrm{x}) + \mathbf{b}(\mathrm{x})u(\mathrm{x})) \tag{2.10}$$

and the corresponding boundary value problem:
Find a function $u(x)$ which satisfies the following differential equation and boundary condition:

$$\mathcal{L}u(\mathrm{x}) = f(\mathrm{x}) \ \text{ in } \Omega \tag{2.11a}$$

$$u(\mathrm{x}) = 0 \qquad \text{on } \partial\Omega \tag{2.11b}$$

where a symmetric $d \times d$ matrix $A(\mathrm{x}) = \{a_{ij}(\mathrm{x})\}_{i,j=1}^{d}$, $a_{ij}(\mathrm{x}) \in L^{\infty}(\Omega)$, a vector $\mathbf{b}(\mathrm{x}) = (b_1(\mathrm{x}), \ldots, b_d(\mathrm{x}))$, $b_i(\mathrm{x}) \in L^{\infty}(\Omega)$, and right hand side $f(\mathrm{x}) \in H^{-1}(\Omega)$ are given functions in Ω. We say that $\mathcal{L}$ is *elliptic* in Ω if $A(\mathrm{x})$ is positive definite for almost every $\mathrm{x} \in \Omega$, and *uniformly elliptic* in Ω if there exist positive constants C and M such that

$$C^{-1}|\xi|^2 \le (A(x)\xi, \xi) \le C|\xi|^2 \,, \tag{2.12a}$$

$$|a_{ij}(\mathrm{x})| \le M, \ |b_i(\mathrm{x})| \le M \,, \quad i, j = 1, \ldots, d \tag{2.12b}$$

for any $\xi \in \mathbb{R}^d$ and for almost every $\mathrm{x} \in \Omega$.

We note that less restrictive assumptions on the coefficients a_{ij}, b_i could be made (cf. [73]). In general, the problem (2.11) has no classical solution $u \in C^2(\Omega) \bigcap C(\bar{\Omega})$ even for bounded coefficients. We introduce the bilinear form associated with the problem (2.11)

$$\mathcal{A}(u, v) = \int_{\Omega} (A(\mathrm{x})\nabla u(\mathrm{x}), \nabla v(\mathrm{x}))\, d\mathrm{x} - \int_{\Omega} (\mathbf{b}(\mathrm{x}).\nabla v(\mathrm{x}))u(\mathrm{x})\, d\mathrm{x} \tag{2.13a}$$

and the linear form

$$f(v) = < f(\mathrm{x}), v(\mathrm{x}) > \ . \tag{2.13b}$$

The problem (2.11) can also be formulated in the following weak form:

Find $u \in H_0^1(\Omega)$ such that

$$\mathcal{A}(u, v) = f(v) \quad \text{for all } v \in H_0^1(\Omega) \,. \tag{2.14}$$

The solution u of (2.14) is called weak (or generalized) solution of (2.11) in $H_0^1(\Omega)$..

In order to assure the global solvability of the problem (2.14) we impose the following assumptions.

Assumption 2.1 *The operator $\mathcal{L}$ is uniformly elliptic.*

Assumption 2.2 $\mathbf{b}(x) \in \left(W^{1,\infty}(\Omega)\right)^d$ *and* $div(\mathbf{b}(x)) \ge 0$ *for almost every* $x \in \Omega$.

We consider also a weaker version of Assumption 2.2.

Assumption 2.3

$$\int_{\Omega} (\mathbf{b}, \nabla v)\, dx \le 0 \quad \forall\, v \in C^{\infty}(\Omega) \,,\ v \ge 0 \,.$$

Here we apply results from the previous section to the problem (2.14). Suppose that Assumptions 2.1 and 2.2 are satisfied. From

$$\int_{\Omega} (\mathbf{b}.\nabla u)u\, d\mathrm{x} = -\int_{\Omega} \mathrm{div}(\mathbf{b}u)u\, d\mathrm{x}$$

$$= -\int_{\Omega} \mathrm{div}(\mathbf{b})u^2\, d\mathrm{x} - \int_{\Omega} (\mathbf{b}.\nabla u)u\, d\mathrm{x}$$

we obtain

$$\int_\Omega (\mathbf{b}.\nabla u)u \, d\mathrm{x} = -\frac{1}{2}\int_\Omega \mathrm{div}(\mathbf{b})u^2 \, d\mathrm{x}$$

and hence

$$\mathcal{A}(u,u) = \int_\Omega (A\nabla u, \nabla v)\, d\mathrm{x} + \frac{1}{2}\int_\Omega \mathrm{div}(\mathbf{b})u^2 \, d\mathrm{x} \tag{2.15}$$

The Assumptions 2.1 and 2.2 guarantee that $\mathcal{A}(,.)$ is $H_0^1(\Omega)$–elliptic, i.e.,

$$C\|u\|_{1,\Omega}^2 \le \mathcal{A}(u,u)\,.$$

We also show that $\mathcal{A}(,.)$ is continuous,

$$|\mathcal{A}(u,v)| \le \|A\|_{0,\infty,\Omega}\|u\|_{1,\Omega}\|v\|_{1,\Omega} + \|\mathbf{b}\|_{0,\infty,\Omega}\|v\|_{1,\Omega}\|u\|_{0,\Omega} \le C\|u\|_{1,\Omega}\|v\|_{1,\Omega}\,.$$

Clearly,

$$|f(v)| \le \|f\|_{-1,\Omega}\|v\|_{1,\Omega}\,.$$

Therefore, from Lax-Milgram lemma follows that the problem (2.14) has an unique solution in $H_0^1(\Omega)$. The Lax–Milgram lemma also guarantees that the solution is stable with respect to the right hand side, i.e.,

$$\|u\|_{1,\Omega} \le C^{-1}\|f\|_{-1,\Omega}.$$

We also prove that $\mathcal{A}(.,.)$ is $H_0^1(\Omega)$–coercive if Assumption 2.2 is replaced by Assumption 2.3. This follows from the equality

$$-\int_\Omega (\mathbf{b}(\mathrm{x}).\nabla u(\mathrm{x}))u(\mathrm{x})\, d\mathrm{x} = -\frac{1}{2}\int_\Omega (\mathbf{b}(\mathrm{x}).\nabla u^2)\, d\mathrm{x}$$

and the density of $C_0^\infty(\Omega)$ in $H_0^1(\Omega)$.

Remark 2.3 We can relax the Assumption 2.2 in the following sense [105]. Let C_Ω be the constant in the Poincaré inequality

$$\|v\|_{0,\Omega} \le C_\Omega |v|_{1,\Omega} \qquad \forall\, v \in H_0^1(\Omega)\,.$$

Then for the $H_0^1(\Omega)$–coercivity of $\mathcal{A}(.,.)$ it suffices that

$$\mathrm{div}(\mathbf{b}) \ge \eta\,, \qquad -\frac{2}{CC_\Omega} < \eta < \infty\,,$$

where C is the constant in the condition (2.12a) of the Assumption 2.1.

We consistently use the notion of flux $\mathbf{q}$ defined below. Then the equation (2.11a) can be written in the "flux" form

$$\mathbf{q} = -A\nabla u + \mathbf{b}u\,, \tag{2.16a}$$

$$\mathrm{div}(\mathbf{q}) = f\,. \tag{2.16b}$$

We briefly discuss the dual weak form of the problem (2.11). First we rewrite (2.16) in the following form:

$$K\mathbf{q} + \nabla u - \beta u = 0\,,$$
$$\mathrm{div}(\mathbf{q}) = f\,,$$

where $K = A^{-1}$ and $\beta = K\mathbf{b}$. The dual weak form is obtain by testing the equation with functions from the appropriate spaces and applying the Green's formulas

Find the pair $(\mathbf{q}, u) \in H_{div}(\Omega) \times L^2(\Omega)$ such that

$$(K\mathbf{q}, \mathbf{v})_0 - (\mathrm{div}(\mathbf{v}), u)_0 - (\beta u, \mathbf{v})_0 = 0 \qquad \forall\, \mathbf{v} \in H(div; \Omega)\,, \tag{2.17a}$$

$$(\mathrm{div}(\mathbf{q}), w)_0 = (f, w)_0 \qquad \forall\, w \in L^2(\Omega)\,. \tag{2.17b}$$

Clearly, the spaces here are $\mathcal{U}_1 = \mathcal{U}_2 = H(div; \Omega)$, $\mathcal{V}_1 = \mathcal{V}_2 = L^2(\Omega)$ and the bilinear forms are $\mathcal{A}(\mathbf{q}, \mathbf{v}) = (K\mathbf{q}, \mathbf{v})_0$, $\mathcal{B}_1(u, \mathbf{q}) = -(u, \mathrm{div}(\mathbf{q}))_0 - (\beta u, \mathbf{v})_0$ and
$\mathcal{B}_2(\mathbf{q}, w) = (\mathrm{div}(\mathbf{q}), w)_0$.

The finite volume weak formulation is define as follows:

Find $u \in H^s(\Omega) \cap H_0^1(\Omega)$, $s > 3/2$ such that for any volume $V \subset \bar{\Omega}$ with Lipschitz–continuous boundary

$$\int_{\partial V} (\mathbf{q}, \mathbf{n})\, ds = \int_V f \,\mathrm{x}\,. \tag{2.18}$$

2.3 Properties of the solutions of elliptic problems

In this section we consider two properties of the solutions of elliptic problems. These characteristics of the solution are very important for the applications. In next chapter we construct numerical methods that satisfy discrete versions of these properties.

2.3.1 Maximum principle

One of the most important properties of the elliptic problems is that they satisfy the maximum principle under certain conditions. In order to state the maximum principle for the weak solution of (2.11) due to Stampacchia [122] we introduce some definitions.

Definition 2.3 We call a function $u(\mathrm{x}) \in H^1(\Omega)$ superelliptic (subelliptic) if

$$\mathcal{A}(u, v) \le 0\ (\ge 0) \qquad \forall\, v \in C_0^\infty(\Omega)\,, \quad v \ge 0$$

and elliptic if it is both superelliptic and subelliptic.

Definition 2.4 We say that $u \in H^1(\Omega)$ satisfies the inequality $u \ge 0$ on $\partial\Omega$, if its negative part $u^- = \min(u, 0) \in H_0^1(\Omega)$.

Theorem 2.13 (Maximum principle [49]) *Let $u(x) \in H^1(\Omega)$ be a subelliptic function, let Assumptions 2.1 and 2.3 be satisfied. Then if $u(x) \ge k$ on $\partial\Omega$,*

$$\operatorname*{ess\,inf}_{x \in \Omega} u(x) \ge \min(0, k)\,.$$

Suppose the problem (2.11) is a mathematical model of some physical process, for example $u(\mathrm{x})$ can be a concentration of some substance. Then since the concentration on $\partial\Omega$ is greater or equal to zero, so is the concentration inside Ω. This shows that our model produces physically meaningful solutions.

The maximum principle is a powerful tool for the investigating the solvability of partial differential equations. We cite some results that can be obtain with application of the maximum principle for the more general problem defined below. Consider the differential operator

$$\mathcal{L}u = \ \mathrm{div}(-A(\mathrm{x})\nabla u(\mathrm{x}) + \mathbf{b}(\mathrm{x})u(\mathrm{x})) + \mathbf{c}(\mathrm{x})\nabla u(\mathrm{x}) + d(\mathrm{x})u(\mathrm{x}) \tag{2.19}$$

where the coefficients $A = \{a_{ij}\}_{i,j=1}^d$, $\mathbf{b}(\mathrm{x}) = \{b_i\}_{i=1}^d$, $\mathbf{c}(\mathrm{x}) = \{c_i\}_{i=1}^d$ and $d(\mathrm{x})$ are functions in $L^\infty(\Omega)$. The corresponding boundary value problem is defined by (2.11) and the bilinear form is

$$\begin{aligned}\mathcal{A}(u,v) &= \int_\Omega (A(\mathrm{x})\nabla u(\mathrm{x}), \nabla v(\mathrm{x}))\, d\mathrm{x} - \int_\Omega (\mathbf{b}(\mathrm{x}).\nabla v(\mathrm{x}))u(\mathrm{x})\, d\mathrm{x} \\ &\quad + \int_\Omega (\mathbf{c}(\mathrm{x}).\nabla u(\mathrm{x}))v(\mathrm{x})\, d\mathrm{x} + \int_\Omega d(\mathrm{x})u(\mathrm{x})v(\mathrm{x})\, d\mathrm{x}\end{aligned}$$

We add to the conditions (2.12) assumptions that the coefficients $\mathbf{c}$ and d are also uniformly bounded, i.e.,

$$|c_i(\mathrm{x})| \le M\,, \quad |d(\mathrm{x})| \le M\,, \quad i = 1, \ldots, d \tag{2.20}$$

where M is the same constant as in (2.12b). Therefore, we say that $\mathcal{L}$ is uniformly elliptic if conditions (2.12a), (2.12b) and (2.20) are satisfied. We also modify the Assumption 2.3. in the following way.

Assumption 2.4

$$\int_\Omega [dv - (\mathbf{b}, \nabla v)]\, dx \ge 0 \quad \forall\, v \in C^\infty(\Omega)\,,\ v \ge 0\,.$$

The uniqueness of the generalized solution of the problem (2.11) where differential operator $\mathcal{L}$ is defined in (2.19) is immediate consequence of the maximum principle.

Corollary 2.2 *If $u \in H_0^1(\Omega)$ is an elliptic function in Ω, then $u = 0$ almost everywhere in Ω.*

The existence result is stated in the following theorem [49].

Theorem 2.14 *If Assumptions 2.1 and 2.4 are satisfied, then the problem* (2.11) *has a generalized solution.*

Remark 2.4 Note that the bilinear form $\mathcal{A}(.,.)$ is not $H_0^1(\Omega)$–coercive under the Assumptions 2.1 and 2.4 . We have only the Gärding's inequality satisfied:

$$\mathcal{A}(u,u) \ge C_1 |u|_{1,\Omega}^2 - C_0 \|u\|_{0,\Omega}^2\,,$$

where $C_1 = C^{-1}/2$ and $C_0 = (2Cd - 1)M$.

This can be seen by the following simple computations:

$$\begin{aligned}\mathcal{A}(u,u) &= C^{-1}|u|_{1,\Omega}^2 - \int_\Omega ((\mathbf{b}-\mathbf{c}), \nabla u)u\, d\mathrm{x} + \int_\Omega du^2\, d\mathrm{x} \\ -\int_\Omega ((\mathbf{b}-\mathbf{c}), \nabla u)u\, d\mathrm{x} &\ge -\frac{C^{-1}}{2}|u|_{1,\Omega}^2 - C\left(\sum_{i=1}^d |b_i| + |c_i|\right)\|u\|_{0,\Omega}^2 \\ &\ge -\frac{C^{-1}}{2}|u|_{1,\Omega}^2 - 2CMd\|u\|_{0,\Omega}^2\end{aligned}$$

In general we cannot show that the constant C_0 is positive, but if the function $d(\mathrm{x})$ is big enough or the domain Ω is small enough (recall that in the Poancaré inequality the constant C_Ω is proportional to $diam(\Omega)$), then we can prove that $\mathcal{A}(.,.)$ is $H_0^1(\Omega)$–elliptic.

2.3.2 General solvability and regularity results

Here we present a few comments on the solvability and regularity of the solution of the problem (2.11) where $\mathcal{L}$ is defined by (2.19). Consider the problem

$$(\mathcal{L} - \lambda E)\, u = f(\mathrm{x}) \ \text{ in } \Omega \tag{2.21a}$$

$$u(\mathrm{x}) = 0 \qquad \text{on } \partial\Omega\,, \tag{2.21b}$$

where λ is a complex number. From the remark above it is clear that if λ is real and $|\lambda| \le \lambda_0$ for some positive number λ_0, then the problem (2.21) has a unique generalized solution. The general result is stated below in a theorem due to Ladyzhenskaya and Uralt'tseva [73]. We collect all the conditions of the theorem in the following assumption.

Assumption 2.5 *Suppose that the condition* (2.12a) *is satisfied. Moreover,*

$$\left\| \sum_{i=1}^{d} b_i^2,\ \sum_{i=1}^{d} c_i^2 \right\|_{0,q/2,\Omega}, \quad \|d(x)\|_{0,q/2,\Omega} \le C \qquad \text{for } q > d\,.$$

$$\frac{1}{\mathrm{meas}(\Omega)} \int_\Omega d(x)\, x \ge C_d > 0\,.$$

$$f = f_0 + div(\mathbf{f}), \quad \mathbf{f} = \{f_i\}_{i=1}^{d}, \ \text{where } f \in L^m(\Omega), \mathbf{f} \in L^2(\Omega) \text{ and}$$

$$m = \frac{2\hat{n}}{\hat{n}+2}\,, \qquad \hat{n} = \begin{cases} d\,, & d > 2 \\ 2+\varepsilon\,, & d = 2\,, \varepsilon > 0 \end{cases}$$

Theorem 2.15 *Suppose that Assumption 2.5 is satisfied. Then the problem* (2.21) *has a unique generalized solution in* $H_0^1(\Omega)$ *for every* $\lambda \in \mathcal{C}$ *except at a countable set of values* $\lambda = \lambda_k$, $k = 1, 2, \ldots$ *such that* $|\lambda_k| \to \infty$ *as* $k \to \infty$. *To every* λ_k *there corresponds a finite number of linearly independent generalized solutions* $u_k^{(i)}$ *of* (2.21) *in* $H_0^1(\Omega)$ *if and only if* $u_k^{(i)}$ *are solutions of the problem*

$$\left(\mathcal{L}^* - \bar{\lambda} E\right) u = 0 \ \text{ in } \Omega$$

$$u(x) = 0 \ \text{ on } \partial\Omega\,,$$

where $\mathcal{L}^*$ *is the conjugate operator of* $\mathcal{L}$ *and* $\bar{\lambda}$ *is the complex conjugate of* λ.

To prove convergence of the numerical methods considered in this dissertation we will need higher regularity of the weak solution of the problem (2.11) than just membership in $H^1(\Omega)$. The regularity result is usually formulated in the following form:

$$\|u\|_{1+\alpha} \le C \|f\|_{-1+\alpha} \qquad \alpha \in (0, 1]\,. \tag{2.22}$$

The case $\alpha = 1$ is called full elliptic regularity.

There are three factors that influence the regularity: the regularity of the domain Ω, the smoothness of the coefficients of (2.10) and the regularity of the right hand side. For thorough discussion of the first factor we refer books by Grisvard [52, 53] and Dauge [34]. Results concerning the other two factors are presented in the monographs by Ladyzhenskaya and Ural'tseva [73] and Lions and Magenes [80] for example.

2.3.3 Conservation properties

We start with a short discussion of the Gauss divergence formulas. Below we derive it from the Green's formulas [31].

Given functions u, $v_i \in H^1(\Omega)$, $i = 1, \ldots, d$, the following fundamental formulas

$$\int_\Omega u \partial_i v_i \, d\mathrm{x} = -\int_\Omega \partial_i u \, v_i \, d\mathrm{x} + \int_{\partial\Omega} u \, v_i n_i \, ds \tag{2.23}$$

holds for any $i = 1, \ldots, d$. Here $\mathbf{n} = (n_1, \ldots, n_d)$ is the unit outward normal vector defined on the boundary $\partial\Omega$. By letting $u = 1$ and summing from 1 to d, we get the *Gauss divergence formulae*

$$\int_\Omega \mathrm{div}(\mathbf{v}) \, d\mathrm{x} = \int_{\partial\Omega} (\mathbf{v}, \mathbf{n}) \, ds \,, \tag{2.24}$$

where $\mathbf{v} \in \left(H^1(\Omega)\right)^d$.

For $\mathbf{v} \in \mathcal{H}_{div}(\Omega)$, $u \in H^1(\Omega)$ the following version of Green's formulas holds (cf. [108]):

$$\int_\Omega u \, \mathrm{div}(\mathbf{v}) \, d\mathrm{x} = -\int_\Omega (\nabla u, \mathbf{v}) \, d\mathrm{x} + \int_{\partial\Omega} u \, (\mathbf{v}, \mathbf{n}) \, ds \,.$$

Again by letting $u = 1$ we obtain (2.24).

The following simple fact in Lebesgue integration theory (cf. [110]) is frequently called a *localization theorem.*

Proposition 2.1 *Let $f(x) \in L^1(\Omega)$ and be such that for any measurable subset $G \subset \Omega$*

$$\int_G f(x) \, dx = 0 \,.$$

Then $f(x) = 0$ for almost every $x \in \Omega$.

With an obvious argument this fact can be proven for more reasonable domains, say, domains with Lipschitz–continuous boundary and for functions in $L^2(\Omega)$.

Most of the PDEs used for modeling of some processes of interest are derived from physical laws of conservation. Without formalization, these laws state that the net change of a physical quantity by way of fluxes through the boundary of a given region equals the net contribution to this quantity from the source or sink inside the region. Typical examples are conservation of mass, conservation of momentum and conservation of energy laws. These laws usually are augmented by constitutive laws like Fourier's law of heat conduction, Darcy's law for porous media flow, or Ohm's law of electric conduction.

Let $\mathbf{q}$ be the flux defined by (2.16a) and $f(\mathrm{x})$ be the density of the source/sink. Then, one particular conservation law can be written as

$$\int_{\partial V} (\mathbf{q}, \mathbf{n}) \, ds = \int_V f(\mathrm{x}) \, d\mathrm{x} \,, \tag{2.25}$$

for every sufficiently regular domain $V \subset \Omega$. If the flux is smooth enough, i.e., $\mathbf{q} \in \left(H^1(\Omega)\right)^d$ ($\mathbf{q} \in \mathcal{H}_{div}(\Omega)$), then we can apply the Gauss divergence formulas and obtain

$$\int_V \mathrm{div}(\mathbf{q}) \, d\mathrm{x} = \int_V f(\mathrm{x}) \, d\mathrm{x} \,. \tag{2.26}$$

Using the localization theorem we get exactly the equation (2.16b) in the definition of the boundary value problem (2.16).

On the other hand if the equation (2.16b) is integrated over a given volume V and the Gauss divergence formulas is applied, the result is (2.25).

Previous remarks showed that the problem (2.11) is equivalent to the problem (2.26) with condition (2.16b). The equation (2.25) with condition (2.16b) can be considered as a

"weak" formulation of the problem (2.11) in the corresponding spaces. If the flux $\mathbf{q}$ and u are approximated separately, this leads naturally to an analogy of the mixed method formulation (cf. [24], [108]) in the couple of spaces $\left(\mathcal{H}_{div}(\Omega), H_0^1(\Omega)\right)$. Our approach is to represent the flux $\mathbf{q}$ via u, and consequently, because of the Theorem 2.4 we need higher regularity for the flux $\mathbf{q} \in \left(H^s(\Omega)\right)^d$, $s > 1/2$, which results for smooth coefficients A and $\mathbf{b}$ in the requirement $u \in H^{s+1}(\Omega)$, $s > 1/2$.

CHAPTER III

FINITE VOLUME DISCRETIZATIONS OF ELLIPTIC PROBLEMS

Finite volume methods are one of the most popular approximation techniques for partial differential equations in the engineering calculation and computational physics. Their distinct conservation property that stems from the approximation of integral conservation laws is very important for the accurate simulation of complicated physical processes on relatively coarse grids. For some problems using conservative methods is almost a necessity (recall the famous Lax–Wendroff theorem for conservation laws [75] and many known non–conservative methods that have local approximation properties, but do not converge globally). In other cases, as in the simulation of fluid flow in porous media, finite volume methods produce more accurate solutions compared to non–conservative ones because of the proper treatment of the discontinuous coefficients through so called "harmonic average transmissibilities".

In this chapter we introduce a class of cell–centered and vertex–centered methods for second order elliptic problems and necessary notations and technical tools to analyze them.

The basic idea of all finite volume methods is to use a finite set of control volumes to describe the equations and restrict the unknowns to be in a finite–dimensional space. We use the flux form (2.16) of the elliptic boundary problem (2.11) and integrate over specially chosen volumes, called boxes, cells, or control volumes, i.e.,

$$\int_V \operatorname{div}(\mathbf{q})\, d\mathrm{x} = \int_V f\, d\mathrm{x}\,.$$

After applying the *Gauss divergence formulas* for the integral on the left, we get

$$\int_{\partial V} (\mathbf{q}, \mathbf{n})\, ds = \int_V f\, d\mathrm{x}\,, \tag{3.1}$$

where $\mathbf{n}$ is the outward unit normal vector to ∂V. We accept the term *control volume* for the volume where the integration is performed.

The solution $u(\mathrm{x})$ of (2.11) is approximated at a set of points called a grid. The values of $u(\mathrm{x})$ in the the grid points are called *degrees of freedom*. We introduce finite element triangulations and two grids related to them: *primary* and *secondary*, denoted by ω_P and ω_S, respectively. The cardinality of the corresponding grids are denoted by n_I, $I = P, S$. The primary grid consists of all vertexes of finite elements, therefore, is determined by the FE triangulations. The secondary grid is designed by choosing one point in the interior of every finite element - "the cell center". We consider two cases: circumcenters (centers of circumscribed circle of the corresponding finite element) and barycenters (centers of gravity).

We choose n_I control volumes V_i, $i = 1, \ldots, n_I$ in a such way that in each control volume there is only one point of the associated grid. The control volumes of cell–centered finite volume methods coincide with finite elements and the degrees of freedom are at the vertexes of the secondary grid. The construction of the control volumes of vertex–centered finite volume methods is more complicated and is outlined in Section 3.1. The degrees of freedom are in the vertexes of the primary grid.

On each control volume we approximate the integral of the normal component of the flux and the integral of the source (sink) function

$$\int_{\partial V_i} (\mathbf{q}, \mathrm{n})\, ds \approx \sum_{j=1}^{k_i} q_{ij}\,, \qquad \int_{V_i} f\, d\mathrm{x} \approx \phi_i\,,$$

or the equation (3.1) on the control volumes V_i, $i = 1, \ldots, n_I$ is replaced by

$$\sum_{j=1}^{k_i} q_{ij} = \phi_i \,, \qquad i = 1, \ldots, n_I, \; I = P \text{ or } S \,. \tag{3.2}$$

We call q_{ij} approximate fluxes. For the approximate fluxes we impose the natural assumption that the approximate fluxes on a common face sum up to zero (cf. Assumption 3.3 for a detailed description.) This condition will ensure that finite volume discretizations have certain discrete conservation properties introduced in Definition 3.3.

The finite volume discretizations are based on (3.2) and the relation between the flux and the scalar variable expressed in the equation (2.16a). If (2.16a) is considered as a separate equation, then the resulting approximations are mixed methods. In the standard finite volume methods (2.16a) is directly incorporated into the equation (3.2).

Remark 3.1 Note that in order to apply the finite volume approach we need some regularity of the flux (cf. Remark 2.1 and the discussion after).

A particular finite volume method is uniquely determined by specifying

(i) the control volumes and the associated grid,

(ii) the approximation of the flux surface integrals.

With respect to the control volumes, we distinguish cell–centered and vertex–centered finite volume methods. Approximation of flux surface integrals can be done in the framework of finite difference methods, or with the tools of finite element methods. Therefore, the consistent names for such methods are: finite volume difference methods and finite volume element methods. In this dissertation we consider:

(i) cell–centered finite volume difference methods,

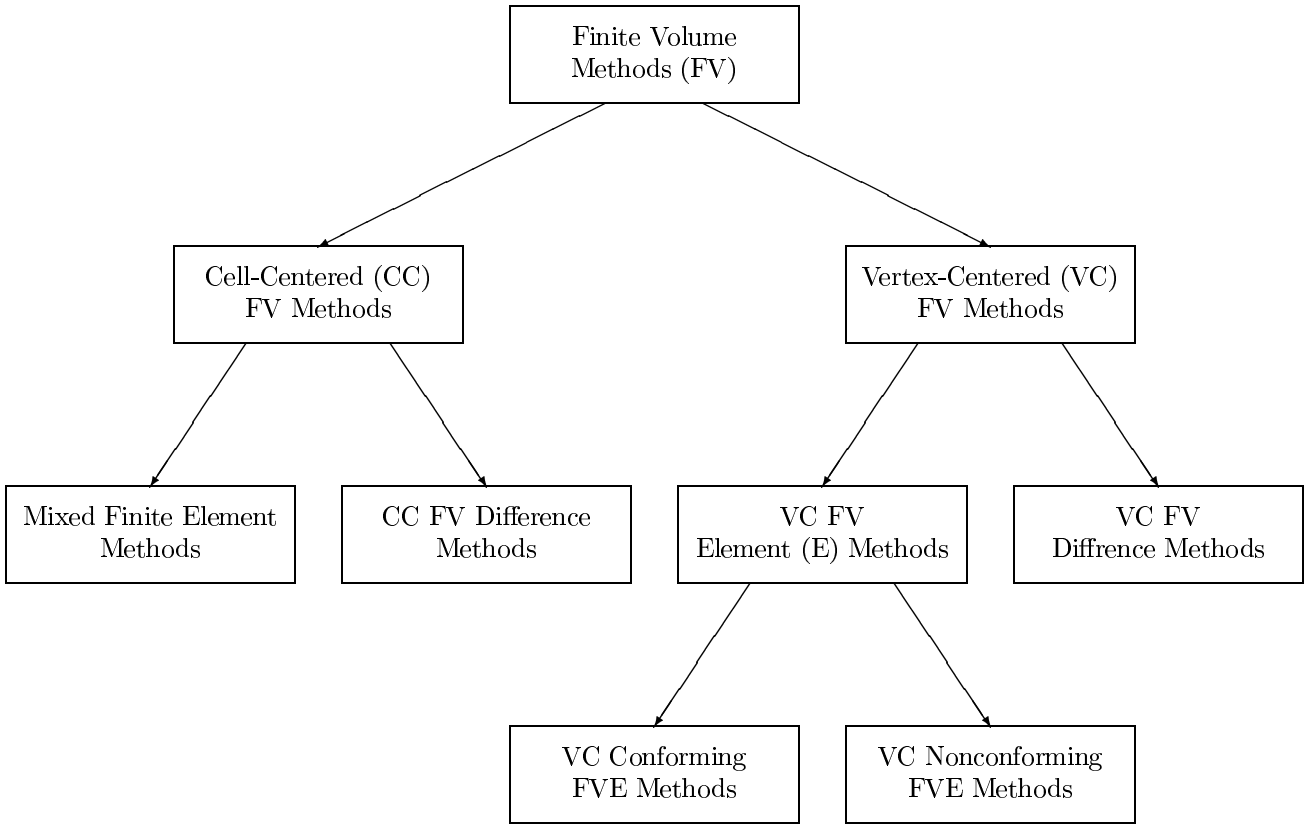
(ii) vertex–centered finite volume element methods.

Finite volume element methods on their own right can be divided in two groups: conforming and nonconforming with respect to the finite element spaces used for the approximation. We sketch a simple classification in Fig. 3.1 and briefly discuss some of the methods.

Finite volume difference methods were the first conservative approximations. They have been used as a systematic approach for effective discretization of conservation law equations (cf., e.g. Patankar and Spalding [101] and Hirsch [60] for fluid flow). Pioneering work in this area for one-dimensional elliptic and parabolic equations with piece-wise smooth coefficients has been done by Samarskii in the early 60-s (for a comprehensive presentation see e.g. Samarskii [114]). Among the characterizations, Tikhonov and Samarskii [129] have proved that the conservation property is a necessary condition for the convergence of finite difference solutions for problems with discontinuous coefficients. For 1–D problems on uniform grids and also multidimensional problems on tensor product uniform grids cell–centered and vertex–centered methods coincides ignoring the boundary conditions. The theory for such methods have been extended for problems with generalized solution by Samarskii, Lazarov and Makarov [115]. For nonuniform meshes second order convergence in the maximum norm has been proven in the papers by Manteuffel and White [85] and Kreiss et. al. [72].

The theory of vertex–centered FV difference methods on triangular meshes has been developed by Heinrich [57] using similar technical tools as in the book by Samarskii, Lazarov and Makarov [115]. Interesting results have been reported for quadrilateral vertex–centered FV difference methods by Morton and Süli [93], Mackenzie and Morton [84] and Süli [124], but the consistent theory for such meshes is still not available.

Figure 3.1: Finite volume methods classification



In late seventies and early eighties the mixed finite element theory was a subject of intensive study (cf. [24, 108]). Russell and Wheeler [112] have shown that certain cell–centered FV difference methods on regular meshes can be obtained from the mixed method system via specially chosen quadrature rules. Weiser and Wheeler [135] have used this relation to prove superconvergence results for cell–centered FV difference methods on rectangular meshes. Recently Arbogast, Wheeler and Yotov [4] have generalized the results in [135] for diffusion problems with tensor coefficients and have derived and analyzed new cell–centered FV difference schemes. Different mixed finite element methods have been proposed and investigated by Thomas and Trujillo [127, 128].

A few terms are used for the finite volume element methods. In the engineering literature they are popular as Control Volume FEM. Bank, Rose and Hackbusch call them box methods. We believe that McCormick promoted the name Finite Volume Element Methods. We accept this name throughout this dissertation because it relatively well indicates the main ideas of the methods.

First attempts to apply finite element ideas in the finite element context were made in the late sixties by engineers and physicists (cf. Winslow [136]). At that time it was noticed that FEM and FVEM were identical for certain grids and coefficients. In the early eighties Patankar and his coworkers applied FVEM for convection–diffusion problems [13, 104, 103, 61]. Exponentially fitted finite element spaces have been utilized in order to align the approximating function to streamlines of the convection term. The resulting approximate solution is continuous only in the nodal points. We consider this as a nonconforming FVEM. Bank and Rose [16] have derived the estimates for the difference between FE and conforming FVE solutions of the Poisson equation. Hackbusch [55] has considered the case when the stiffness matrices of FEM and FVEM coincide and estimated the difference of the right hand sides. Cai with collaboration of McCormick and Mandel has derived the estimates for the more general diffusion equation [27, 26, 28] including also local refinement. Under certain geometrical conditions some superconvergence results have been obtained. Later Jianguo and Shitong [66] have extended the results of Cai to more general meshes. They also have shown with a counterexample that an L^2 lifting is not possible in general for FVE methods. Süli [123] has considered FVEM on arbitrary rectangular meshes. He has derived superconvergence estimates for bilinear and linear finite element spaces. Schmidt [116] has proven a first order rate of convergence for FVEM on quasi-uniform quadrilateral meshes with bilinear spaces and superconvergence on rectangular grids.

Exponentially fitted nonconforming FVEM on a a Voronoi mesh and its dual Delaunay triangulation have been analyzed by Miller and Wang [88].

FVEM are well suited for nonregular domains and meshes and handle successfully different boundary conditions [87]. Because of their conservation properties they are frequently used for approximation of practical problems like fluid flow and heat transfer [14, 82], the Navier–Stokes equations [71] and the equations of the fluid flow in porous media [45].

The remainder of this chapter is organized as follows. In Section 3.1 we consider various types of control volumes introduced using finite element triangulations, Voronoi or circumscribed meshes. The necessary notations, discrete inner products and norms are defined and some simple results that are used in the next chapters proven. In Section 3.2 we state the conditions for the discrete maximum principle and discrete conservation. Section 3.3 is devoted to cell–centered finite volume methods. We derive the basic difference scheme for scalar coefficients and propose a new method for problems with tensor coefficients. The schemes for problems with scalar coefficients are extended and studied in Chapters IV and V. We discuss the relation of mixed finite element methods with cell–centered finite difference schemes in Section 3.3.2. The definition and some simple observations for finite volume element methods are collected in Section 3.4

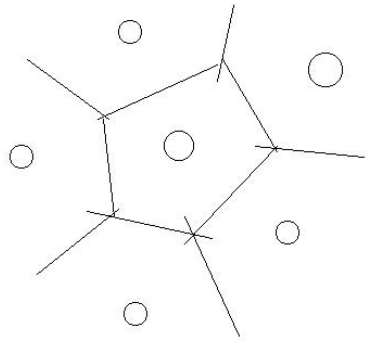


Figure 3.2: Voronoi diagram

3.1 Control volumes

In the beginning of this chapter we outlined the general idea of constructing primary and secondary triangulations. Frequently the names Voronoi, Delaunay and Dirichlet are associated with this process. We introduce some definitions and historical references in order to explain the ideas behind these names. We use substantially the survey papers by Aurenhammer [7] and Fortune [47].

Let P denote a set of n_P points (or sites) in the plane. For two different points p, $q \in P$, the *set of dominance* of p over q is the set

$$dom(p,q) = \{x \in \mathbb{R}^2 \ : d(x,p) \leq d(x,q)\}\,.$$

Clearly $dom(p,q)$ is a closed half plane bounded by the perpendicular bisector of p and q. The *Voronoi region* of p is defined by

$$reg(p) = \bigcap_{q \in P \setminus \{p\}} dom(p,q)\,.$$

By the definition follows that $reg(p)$ is a closed convex polygon as intersection of $n_P - 1$ half planes. It is also easy to see that the interior of $reg(p)$ is the set of points that are closer to p than to any other point of $P\setminus\{p\}$. This process generates a partition of $\mathbb{R}^2$ called *Voronoi diagram*. Restricting this partition to a finite subdomain Ω we obtain *Voronoi mesh*. One simple Voronoi diagram is shown on Fig. 3.2. The vertexes of the Voronoi diagram are called *Voronoi vertexes*.

It seems that the earliest motivation for the study of Voronoi diagrams stemmed from their application in the theory of quadratic forms observed by Gauss and further exploited by Dirichlet. Voronoi [133] generalized the results of Gauss and Dirichlet to higher dimensions and called these triangulations *Dirichlet tessellations*. Closely connected with Voronoi diagrams are *Delaunay triangulations*. Delaunay triangulations contains an edge connecting two points of P if and only if their Voronoi regions share a common edge. This construction was introduced by Voronoi . Delaunay [35] extended it to irregular domains by considering all triangles formed by points of P and such that the circumcircle of each triangle is empty of other points of P.

The planar Voronoi diagram and the Delaunay triangulation are dual in a graph-theoretical sense. Voronoi vertexes correspond to Delaunay triangles. It is clear that Delaunay edges are orthogonal to Voronoi edges. These constructions extend also to n–dimensional case.

From the discussion above is clear that given a domain Ω and a primary grid inside Ω, we can construct the Voronoi mesh and the corresponding Delaunay triangulation. We choose Voronoi regions as vertex–centered control volumes and the Delaunay triangulation as a finite element triangulation. We note that the Delaunay triangulation is one of the most popular in computational mesh generation [48] especially for finite element computations because of its optimal properties [19]. For example the Delaunay triangulation maximize the minimum angle of the triangles.

On the other hand we can choose first the secondary mesh as Voronoi sites and construct the Voronoi mesh. Then the primary mesh consist of Voronoi vertexes and cell–centered control volumes coincides with Voronoi regions.

3.1.1 Finite element triangulations

For the FVE methods we will need some of the basic constructions of the finite element methods. Let Ω be a polyhedral domain in $\mathbb{R}^d$, $d = 2, 3$. We consider a family of triangulations $\mathcal{F}_h$ of Ω into finite elements K. By a triangulation, $\mathcal{F}_h$, we mean a set of polygonal (polyhedral) elements such that the intersection of any two distinct elements in $\mathcal{F}_h$ either consists of a common face, common side or common vertex, or is empty, and $\bar{\Omega} = \bigcup_{K\in\mathcal{F}_h} K$.

For any $K \in \mathcal{F}_h$, let

$$h_K = \operatorname{diam}(K)\,, \qquad h = \max_{K\in\mathcal{F}_h} h_K\,,$$

and

$$\rho_K = \sup\{\operatorname{diam}(B) : B \text{ is a ball contained in } K\}\,.$$

We assume that $\mathcal{F}_h$ is regular in sense of Ciarlet [31, p. 132], i.e., the following two conditions are satisfied:

(i) There exists a positive constant σ such that , for all $K \in \mathcal{F}_h$, $h \in \mathcal{R}^+$

$$h_K \le \sigma \rho_K\,. \tag{3.3}$$

(ii) The parameter h approaches zero.

Later we will specify the admissible finite elements.

3.1.2 Affine mappings

We need some facts about affine mappings and estimates of the seminorms. We say that two domains Ω and $\hat{\Omega}$ in $\mathbb{R}^d$ are *affine–equivalent* if there exists an invertible affine mapping $F : \hat{\Omega} \to \Omega$, $x = F(\hat{\mathrm{x}}) = B_\Omega \hat{\mathrm{x}} + \mathbf{b}$ such that $\Omega = F(\hat{\Omega})$. We use the correspondence

$$v : \Omega \to \mathbb{R},\ \hat{v} : \hat{\Omega} \to \mathbb{R}, \quad v = \hat{v} \circ F^{-1}\,, \quad v(\mathrm{x}) = \hat{v}(\hat{\mathrm{x}})\,.$$

Theorem 3.1 ([31]) *Let Ω and $\hat{\Omega}$ be two affine–equivalent domains in $\mathbb{R}^d$. If a function v belongs to the space $W^{m,p}(\Omega)$ for some integer $m \ge 0$ and some number p, $1 \le p \le \infty$, the function $\hat{v} = v \circ F$ belongs to the space $W^{m,p}(\hat{\Omega})$, and in addition, there exists a constant $C = C(m, d)$ such that*

$$|\hat{v}|_{m,p,\hat{\Omega}} \le C\|B_\Omega\|^m |\det(B_\Omega)|^{-1/p} |v|_{m,p,\Omega}\,. \tag{3.4a}$$

Analogously, one has

$$|v|_{m,p,\Omega} \le C\|B_\Omega^{-1}\|^m |\det(B_\Omega)|^{1/p} |\hat{v}|_{m,p,\hat{\Omega}}\,. \tag{3.4b}$$

The norms $\|B_\Omega\|$ and $\|B_\Omega^{-1}\|$ are evaluated in terms of the following geometric quantities:

$$\begin{aligned} h &= diam(\Omega)\,, \qquad \hat{h} = diam(\hat{\Omega}), \\ \rho &= \sup\{diam(S) : S \text{ is the ball contained in } \Omega\}, \\ \hat{\rho} &= \sup\{diam(\hat{S}) : \hat{S} \text{ is the ball contained in } \hat{\Omega}\}. \end{aligned}$$

Theorem 3.2 ([31]) *Let Ω and $\hat{\Omega}$ be two affine–equivalent domains in $\mathbb{R}^d$, where $F : \hat{\Omega} \to \Omega$, $x = F(\hat{x}) = B_\Omega \hat{x} + \mathbf{b}$ is an invertible affine mapping. Then the upper bounds*

$$\|B_\Omega\| \le \frac{h}{\hat{\rho}}\,, \tag{3.5a}$$

$$\|B_\Omega^{-1}\| \le \frac{\hat{h}}{\rho}\,. \tag{3.5b}$$

hold. Moreover,

$$|\det(B_\Omega)| = \frac{\text{meas}(\Omega)}{\text{meas}(\hat{\Omega})} \tag{3.5c}$$

We introduce the notion of reference element $\hat{K}$ and affine families of finite elements. Let $\hat{K}$ be a given finite element. We say that the triangulation $\mathcal{F}_h$ is *affine family* if any finite element $K \in \mathcal{F}_h$ is affine–equivalent to the reference finite element.

We use a simple corollary of Theorem 3.2 stated below.

Proposition 3.1 *Assume that $\mathcal{F}_h$ is an affine family of finite elements that satisfies the regularity assumption (3.3). Then, there exist positive constants C_1 and C_2 such that for any $\mathbf{u} \in \mathbb{R}^4$ the following inequalities hold:*

$$\frac{C_1}{h}\|\mathbf{u}\| \le \|B_K^{-1}\mathbf{u}\| \le \frac{C_2}{h}\|\mathbf{u}\|\,, \quad \forall K \in \mathcal{F}_h\,. \tag{3.6}$$

Proof: Note that by (3.5b) we have $\|B_K^{-1}u\| \le (\hat{h}/\rho_K)\|u\|$. The chain of inequality

$$1 = \|B_K B_K^{-1}\| \le \|B_K\|\|B_K^{-1}\|$$

combined with (3.5a) gives $(\hat{\rho}/h_K)\|u\| \le \|B_K^{-1}u\|$. Application of the regularity condition (3.3) completes the proof. □

We consider some *polynomial preserving operators.* i.e., which satisfy a relation of the form (3.7) for some integer $k \ge 0$.

Theorem 3.3 ([31]) *For some integers $k \ge 0$ and $m \ge 0$ and some numbers p, $q \in [1, \infty]$, let $W^{k+1,p}(\Omega)$ and $W^{m,q}(\Omega)$ be Sobolev spaces satisfying the inclusion*

$$W^{k+1,p}(\Omega) \hookrightarrow W^{m,q}(\Omega)$$

and let $\Pi : W^{k+1,p}(\Omega) \to W^{m,q}(\Omega)$ be a continuous linear mapping such that

$$\forall\, p \in P_k(\Omega), \qquad \Pi p = p\,. \tag{3.7}$$

Then there exists a constant $C(\Pi, \Omega)$ such that

$$|v - \Pi v|_{m,q,\Omega} \le C(\Pi, \Omega)\,(\text{meas}(\Omega))^{1/q-1/p}\, h^{k+1-m}|v|_{k+1,p,\Omega}\,. \tag{3.8}$$

Later we will use this general theorem for linear and constant interpolants in $H^{k+1}(\Omega)$, $k = 0$, 1 and and $m = 0$, 1.

3.1.3 Primary and secondary grids

Here we describe another general way to construct grids starting from a FE triangulation. We assume that we are given a primary grid and some finite element triangulation with vertexes at the points of the primary grid. Suppose that the vertexes of the finite element triangulation are numbered in a unique way, i.e., $\{x_i : i = 1, \ldots, n_P\}$. The finite element triangulation uniquely determines a primary grid $\overline{\omega}_P$,

$$\overline{\omega}_P = \left\{ x_i \in \overline{\Omega} : x_i \text{ is a vertex in a finite element } K \right\} .$$

We also need the set of interior grid points ω_P and the boundary grid points γ_P;

$$\omega_P = \overline{\omega}_P \cap \Omega, \quad \gamma_P = \overline{\omega}_P \backslash \omega_P .$$

Consider a particular finite element K with vertexes $x_{i_1}, \ldots, x_{i_k}$ and let I_K be the index set $\{i_1, \ldots, i_k\}$. Denote by $\{Z_{K,ij}\}_{i,j \in I_K}$ the edges and by $\{Z_{K,j_1 \ldots j_l}\}_{j_1,\ldots,j_l \in I_K}$ the faces of a given finite element (the polygons with vertexes $x_{j_1}, \ldots, x_{j_l} \in K$), $K \in \mathcal{F}_h$, i.e.,

$$\partial K = \left(\bigcup_{\substack{i,j \in I_K \\ x_i, x_j \text{ edge in } K}} Z_{K,ij} \right) \bigcup \left(\bigcup_{\substack{j_1,\ldots,j_l \in I_K \\ x_{j_1},\ldots,x_{j_l} \text{ face in } K}} Z_{K,j_1 \ldots j_l} \right) .$$

We define the secondary grid in the following way. Choose one interior point $S_K \in \overset{\circ}{K}$ in every finite element $K \in \mathcal{F}_h$. Then

$$\omega_S = \{S_K, \, K \in \mathcal{F}_h\} .$$

The cell–centered control volumes coincide with the finite elements and there is one–to–one correspondence of finite elements and nodes of the secondary grid. We assume that the vertexes of the secondary grid are also numbered in unique way, i.e. $\{x_i : i = 1, \ldots, n_S\}$. If $x_i \in \omega_S$ we denote with K_i the corresponding finite element. Whether we use nodes from a primary or secondary grid will be clear from the context.

Given a primary grid vertex x_i we define by $\Pi(i)$ the index set of all neighbors of x_i in ω_P, i.e.,

$$\Pi(i) = \{j : \text{ there is an edge between } x_i \text{ and } x_j \text{ in } \mathcal{F}_h\} . \tag{3.9}$$

We denote with $\Sigma(i)$ the index set of all neighbors of $x_i \in \omega_S$

$$\Sigma(i) = \{j : \text{ finite elements } K_i \text{ and } K_j \text{ have common face}\} . \tag{3.10}$$

To describe vertex–centered control volumes we select one interior point on each face of every finite element K_i, $M_{K_i,j_1 \ldots j_l} \in Z_{K_i,j_1 \ldots j_l}$ such that if $Z_{K_i,j_1 \ldots j_l} \equiv Z_{K_p,j_1 \ldots j_l}$, $i \neq p$ then $M_{K_i,j_1 \ldots j_l} \equiv M_{K_p,i_1 \ldots j_l}$, i.e., on each face only one point is chosen. The points on the edges are selected in the same manner. Connect a given point from the secondary grid x_i, $K_i \in \mathcal{F}_h$ with $M_{K,j_1 j_2}$, $j_1, j_2 \in I_{K_i}$ and $M_{K,i_1 \ldots i_l}$, $i_1, \ldots, i_l \in I_{K_i}$. These lines and the planes that they span form a polygonal (polyhedral) domain around each vertex of the primary grid and are called vertex–centered control volumes. There is one–to–one correspondence of nodes in primary grid with vertex–centered control volumes. If $x_i \in \omega_P$ we denote the corresponding vertex–centered control volume with V_i and with

$$\gamma_{ij} = V_i \cap V_j , \quad j \in \Pi(i) .$$

To specify a particular primary and secondary grids we have to choose the finite elements, secondary grid points and points $M_{K_i,j_1 j_2}$ on the edges, $M_{K_i,j_1 \ldots j_l}$ on the faces.

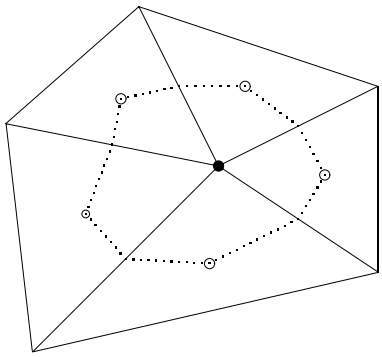


Figure 3.3: Vertex–centered control volume

First we define grids for the finite volume element methods we will investigate in Chapter VI. We choose finite elements to be triangles in 2–D and tetrahedra in 3–D. The "cell–centers" are the barycenters of the finite elements and points M are barycenters of the edges and faces, correspondingly. A specific 2–D example is shown on Fig. 3.3, where the primary node is displayed with a filled circle and the secondary nodes are shown with empty circles. The control volume corresponding to the primary node is depicted by a dotted line. Note that in general γ_{ij} is not a straight line. We also consider Delaunay triangulations and corresponding Voronoi vertex–centered control volumes.

For cell–centered difference methods we require that finite elements be chosen in a such way that there exists a circumscribed circle around each finite element or cell–centered control volumes are Voronoi regions. The cell–centers are chosen in the centers of the circumscribed circles for the first case. We call the former one a *circumscribed cell–centered* grid and the latter one a *Voronoi cell–centered* grid. Note that in both choices the line connecting two neighboring cell–centers is perpendicular to face between them. Therefore, γ_{ij} is a straight line now. An example of cell–centered circumscribed grid is shown in Fig 3.4.

We do not impose any restriction that the finite elements have the same shape.

3.1.4 Finite element spaces. Discrete inner products and norms

We introduce a piecewise linear finite element space for the simplex triangulation

$$\mathcal{V}^h = \{v \in C^0(\Omega) : v_{|K} \text{ is linear for all } K \in \mathcal{F}_h\}$$

where $v_{|K}$ is the restriction of v to K. The finite element space $\mathcal{V}_0^h$ is defined by

$$\mathcal{V}_0^h = \{v \in \mathcal{V}^h : v_{|\Gamma} = 0\}.$$

Functions defined for $x \in \omega_I$, $I = P, S$ are called vertex–centered (cell–centered) grid functions. To emphasize their dependence of the triangulation we use the subscript h, for example $u_h(x)$, $x \in \omega_P$ is a vertex–centered grid function. Denote with χ_i the characteristic functions that corresponds to the vertex–centered control volume V_i and with $\mathcal{W}^h$ the space spanned on $\{\chi_i\}_{x_i \in \omega_P}$. Let $\{\varphi_i\}_{x_i \in \omega_P}$ be the basis of $\mathcal{V}_0^h$. We define a few operators. The linear

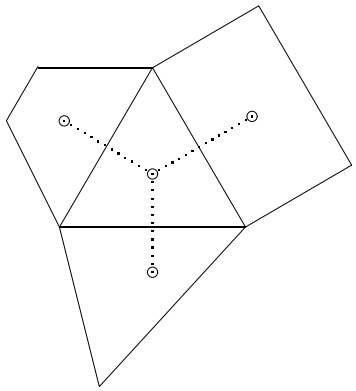


Figure 3.4: Cell–centered control volume

interpolant $I_h^l : \omega_P \to \mathcal{V}_0^h$ is defined by

$$I_h^l u_h(x) = \sum_{x_i \in \omega_P} u_h(x_i)\varphi_i(x)$$

and the "inverse" mapping $R_h^l := (I_h^l)^{-1}$, $R_h^l : \mathcal{V}_0^h \to \omega_P$ is simply the restriction of elements of $\mathcal{V}_0^h$ on ω_P. The "box" interpolant (constant interpolant) $I_h^c : \omega_P \to \mathcal{W}^h$ is given by

$$I_h^c u_h(x) = \sum_{x_i \in \omega_P} u_h(x_i)\chi_i(x)$$

and $R_h^c : \mathcal{W}^h \to \omega_S$, the restriction of elements of $\mathcal{W}^h$ on ω_S. We define also the mapping $\bar{I}_h^c$ between spaces $\mathcal{V}_0^h$ and $\mathcal{W}^h$ via $\bar{I}_h^c := R_h^l I_h^c$ and the mapping $\bar{I}_h^l : \mathcal{W} \to \mathcal{V}$ by $\bar{I}_h^l := R_h^c I_h^l$. Let $H^s(\Omega)$ be a Sobolev space with $s > 3/2$. Define $\tilde{I}_h^l : H^s(\Omega) \to \mathcal{V}_0^h$ and $\tilde{I}_h^c : H^s(\Omega) \to \mathcal{W}^h$ by

$$\tilde{I}_h^l u(x) = \sum_{x_i \in \omega_P} u(x_i)\varphi_i(x)$$
$$\tilde{I}_h^c u(x) = \sum_{x_i \in \omega_P} u(x_i)\chi_i(x)$$

When there is no danger of ambiguity we will skip the bars and tildes.

We use Theorem 3.3 to estimate the error of interpolation.

Corollary 3.1 *For every function $v \in H^{3/2+\alpha}(\Omega)$, $0 < \alpha \leq \frac{1}{2}$ the following estimates hold:*

$$|v - \tilde{I}_h^c v|_{0,\Omega} \leq Ch|v|_{1,\Omega} \tag{3.11}$$
$$|v - \tilde{I}_h^l v|_{1,\Omega} \leq Ch^{1/2+\alpha}|v|_{3/2+\alpha,\Omega} \tag{3.12}$$

Given the cell–centered grid functions $u_h(x)$, $v_h(x)$, $x \in \overline{\omega}_S$ we define the following discrete inner products and norms:

$$(u_h, v_h)_S = \sum_{x_i \in \omega_S} \mathrm{meas}(K_i)u_h(x_i)v_h(x_i)\,, \quad \|u_h\|^2_{0,\omega_S} = (u_h, u_h)_S;$$

$$|u_h|^2_{1,\omega_S} = \frac{1}{2}\sum_{x_i\in\overline{\omega}_S}\sum_{y\in\Sigma(x_i)} \mathrm{meas}(K_i)\left(\frac{u_h(x_i)-u_h(y)}{d(x_i,y)}\right)^2$$

where $d(x,y)$ is the Euclidean distance between x and y. The discrete H^1–norm is defined by

$$\|u_h\|^2_{1,\omega_S} = \|u_h\|^2_{0,\omega_S} + |u_h|^2_{1,\omega_S}\,.$$

We define vertex–centered inner products and norms in the following way:

$$(u_h,v_h)_P = (I_h^l u_h, I_h^l v_h)_{L^2}\,, \qquad \|u_h\|^2_{0,\omega_P} = (u_h,u_h)_P\,,$$

$$|u_h|_{1,\omega_P} = |I_h^l u_h|_{1,\Omega}\,, \qquad \|u_h\|^2_{1,\omega_P} = \|u_h\|^2_{0,\omega_P} + |u_h|^2_{1,\omega_P}\,.$$

We also use the "box" norms and seminorms

$$\|u_h\|^2_{0,B} = \sum_{x\in\omega_P}\mathrm{meas}(V_x)u_h(x)v_h(x)\,,$$

$$|u_h|^2_{1,B} = \frac{1}{2}\sum_{K\in\mathcal{F}_h}\mathrm{meas}(K)\sum_{x,y\in K}\left(\frac{u_h(x)-u_h(y)}{d(x,y)}\right)^2\,.$$

Note that

$$\|u\|_{0,B} = \|I_h^c u_h\|_{L^2}\,.$$

The following result is well known (see for example [98] for the 2–D case and regular geometry, [16] for the 2–D case and general geometry, and [57] for the finite difference case discussion), but we include it for the sake of completeness. There are some small differences in 3–D.

Lemma 3.1 *The norms $\|.\|_{0,\omega_P}$, $\|.\|_{0,B}$ and $\|.\|_{1,\omega_P}$, $\|.\|_{1,B}$ are equivalent, i.e., there exist positive constants C_1, C_2, C_3 and C_4 independent of h such that*

$$C_1\|u_h\|_{0,\omega_P} \le \|u_h\|_{0,B} \le C_2\|u_h\|_{0,\omega_P}\,, \tag{3.13}$$

$$C_3\|u_h\|_{1,\omega_P} \le \|u_h\|_{1,B} \le C_4\|u_h\|_{1,\omega_P}\,. \tag{3.14}$$

Proof: We prove the equivalence on the element level. Let $\hat{K}$ be the reference tetrahedron and K be the current tetrahedron. Consider $|u_h|_{1,\omega_P}$ restricted on K, i.e., $|I_h^l u_h|_{1,K}$. Denote $w := I_h^l u_h$.

$$|w|^2_{1,K} = \int_K |\nabla w|^2\,dx = \int_{\hat{K}} |B_K^{-T}\nabla\hat{w}|^2|\det(B_K)|\,d\hat{x} = \frac{\det(B_K)}{6}\|B_K^{-T}\nabla\hat{w}\|^2\,.$$

Using the equality (3.5c) and the fact that $\mathrm{meas}(\hat{K}) = 1/6$ we see that

$$\frac{\det(B_K)}{6} = \mathrm{meas}(K)\,.$$

Let the values of u in the vertexes of K be u_0, u_1, u_2, u_3. Then $\nabla\hat{w} = (u_1-u_0, u_2-u_0, u_3-u_0)^T$. Taking into account the inequalities

$$C_1 d(u_i,u_j) \le h_K \le C_2 d(u_i,u_j)\,,$$

$$C_3\sum_{j=2}^{d+1}(u_j-u_1)^2 \le \sum_{i=1}^{d}\sum_{j>i}^{d+1}(u_j-u_i)^2 \le C_4\sum_{j=2}^{d+1}(u_j-u_1)^2$$

and (3.6) we get that

$$C_1|w|_{1,K} \le (|u_h|_{1,B})_{|K} \le C_2|w|_{1,K}\,. \tag{3.15}$$

The contribution in $\|u_h\|_{0,B}$ from a particular element K is

$$\frac{\mathrm{meas}(K)}{d+1}\left(\sum_{i=1}^{d+1} u_i^2\right)\,. \tag{3.16}$$

On the other hand

$$\|u_h\|_{0,\omega_P,K} = \int_K |I_h^l u_h|^2\,dx = |\det(B_K)| \int_{\hat K} |I_h^l u_h|^2\,d\hat x \tag{3.17}$$

and there exist positive constants C_1 and C_2 such that

$$C_1 \sum_{i=1}^{d+1} u_i^2 \le \int_{\hat K} |I_h^l u_h|^2\,d\hat x \le C_2 \sum_{i=1}^{d+1} u_i^2\,. \tag{3.18}$$

Combining (3.17) and (3.18) we obtain the estimates in (3.13). The inequalities in (3.14) follow from (3.13) and (3.15). □

Remark 3.2 In the proof of Lemma 3.1 we used that the secondary grid is formed by barycenters in the expression (3.16). If the secondary grid is arbitrary the norms $\|.\|_{0,\omega_P}$ and $\|.\|_{0,B}$ are not equivalent. This is seen by the following simple example. Consider one control volume V_i, such that $\mathrm{meas}(V_i) \to 0$, i.e., the secondary points around x_i go to x_i. Pick a function $u_h = (0,\dots,1,\dots,0)$, where the only nonzero element is on the i^{th} position. Then $\|u_h\|_{0,B} \to 0$, but $\|u_h\|_{0,\omega_P}$ is bounded from below.

If the finite element triangulation is regular, then

$$|u_h|_{1,B}^2 \sim \frac{1}{2}\sum_{x_i\in\omega_P} \mathrm{meas}(V_i) \sum_{j\in\Pi(i)} \left(\frac{u_h(x_i)-u_h(x_j)}{d(x_i,x_j)}\right)$$

as can be seen easily by comparing

$$\left(\frac{u_h(x)-u_h(y)}{d(x,y)}\right)^2 [\mathrm{meas}(V_x)+\mathrm{meas}(V_y)] \text{ and } \left(\frac{u_h(x)-u_h(y)}{d(x,y)}\right)^2 [\mathrm{meas}(K_1)+\mathrm{meas}(K_2)]\,,$$

where $(x_i,x_j) = K_1\cap K_2$. We will use this definition in Chapter VI.

The seminorms $|.|_{1,B}$ and $|.|_{1,\omega_P}$ are equivalent without any restriction on the secondary grid.

We pay special attention to rectangular cell–centered meshes. In this case we denote the points in S by $x = (x_1,x_2) = (x_{1,i},x_{2,j}) = (ih,jh)$, where $i = 0,1,\dots,N_x$, $j = 0,1,\dots,N_y$ are integer indices. Therefore

$$\overline{\omega}_S = \left\{(x_{1,i},x_{2,j}) \in \overline{\Omega} : i = 0,1,\dots,N_x,\ j = 0,1,\dots,N_y\right\};$$

and $\omega_S = \overline{\omega}_S \cap \Omega$. We also use subgrids of ω_S

$$\omega_{S,i}^{\pm} = \omega_S \cup \gamma_i^{\pm}\ \text{, where }\ \gamma_i^{\pm} = \{x\in\gamma : cos(x_i,\mathbf{n}) = \pm 1\}\ ,\ i = 1,\,2\,.$$

We consistently use the dual notation for the value of the function y at the grid point $x = (x_{1,i},x_{2,j})$; $y(x) = y(x_{1,i},x_{2,j}) = y_{i,j}$ and in the points $(x_{1,i},x_{2,j}\pm h/2) = (x_{1,i},x_{2,j\pm 1/2})$ and $(x_{1,i}\pm h/2,x_{2,j}) = (x_{1,i\pm 1/2},x_{2,j})$, $y_{i,j\pm 1/2} = y(x_{1,i},x_{2,j\pm 1/2}), y_{i\pm 1/2,j} = y(x_{1,i\pm 1/2},x_{2,j})$.

Let the (i,j) cell have length $h_{1,i}$ and height $h_{2,j}$ For rectangular meshes we can rewrite cell–centered inner products and norms in the following way:

$$(y,v)_S = \sum_{x_{i,j}\in\omega_S} h_{1,i}h_{2,j}y(x_{i,j})v(x_{i,j}), \quad \|y\|_{0,\omega} = (y,y)^{\frac{1}{2}};$$
$$(y,v]_s = \sum_{x_{i,j}\in\omega^+_{S,i}} h_{1,i}h_{2,j}y(x_{i,j})v(x_{i,j}), \quad \|y]\|_i = (y,y]_i^{\frac{1}{2}}, \quad i=1,2.$$

We introduce the following finite differences for grid functions $y(x)$:

(i) forward difference $\Delta_1 y_{i,j} = y_{i+1,j} - y_{i,j}$ and divided forward difference $y_{x_1,i,j} = \Delta_1 y_{i,j}/dist(x_{i+1,j}, x_{i,j})$;

(ii) backward difference $\overline{\Delta}_1 y_{i,j} = y_{i,j} - y_{i-1,j}$ and divided backward difference $y_{\overline{x}_1,i,j} = \overline{\Delta}_1 y_{i,j}/dist(x_{i,j}, x_{i-1,j})$;

(iii) divided central difference of second order
(for uniform meshes with $h_{1,i} = h_{2,j} = h$)

$$y_{\overline{x}_1 x_1} = \frac{\Delta_1 y_{i,j} - \overline{\Delta}_1 y_{i,j}}{h^2}.$$

Similarly, the differences are defined in x_2 and in combination of x_1 and x_2 coordinate directions.

We also introduce the discrete analogs of H^2 -norms:

$$|y|^2_{2,\omega_S} = |y_{\overline{x}_1 x_1}|^2 + 2|y_{\overline{x}_1\overline{x}_2}|^2 + |y_{\overline{x}_2 x_2}|^2,$$
$$\|y\|^2_{2,\omega_S} = |y|^2_{2,\omega_S} + \|y\|^2_{1,\omega_S}.$$

We will also need the negative norm:

$$\|y\|_{-1,\omega_I} = \sup_{v\neq 0} \frac{|(y,v)_I|}{\|v\|_{1,\omega_I}}, \quad I = P, S.$$

3.2 Properties of finite volume methods

In Chapter II we have discussed the properties of the continuous problem. Here we define the corresponding discrete analogs and briefly discuss them.

Suppose the approximate fluxes q_{ij} have been expressed through the values of u_h. Denote by $\mathcal{S}(i)$ the index set of all points x_j that enter the approximation of the integral $\int_{\partial K_i}(\mathbf{q},\mathbf{n})\,ds$. This set is called the *stencil* of the i^{th} vertex. Then the system of equations (3.2) is equivalent to the following system

$$a_{ii}u_{h,i} + \sum_{j\in\mathcal{S}(i)} a_{ij}u_{h,j} = \phi_i, \qquad i = 1,\dots,n_I,\ I = P \text{ or } S. \tag{3.19}$$

Usually the stencil $\mathcal{S}(i)$ coincides with the index sets $\Pi(i)$ or $\Sigma(i)$ (cf. (3.9), (3.10)), correspondingly.

Any grid function $y_h(x)$ can be considered as an element of a vector space of dimension equal to n_I, the number of the grid points in ω_I, $I = P, S$. In this case, we denote $y_h(x)$ by $\mathbf{y}\in\mathbb{R}^{n_I}$ and consider it as an n_I-dimensional column vector. Then $\mathbf{y}^T$ will be the row vector transpose of $\mathbf{y}$.

We can rewrite (3.19) in the following form:

$$\mathcal{L}_h u_h = \phi, \tag{3.20}$$

where $\mathcal{L}_h : \mathbb{R}^{n_I} \to \mathbb{R}^{n_I}$ is a linear operator.

We say that $\mathcal{L}_h$ satisfies the discrete maximum principle if the following conditions hold [57]:

(*i*) $a_{ii} > 0\,, \quad a_{ij} \le 0\,, \quad j \ne i$

(*ii*) For each $x_i \in \bar{\omega}_I\,, I = P, S$, the inequality

$$a_{ii} + \sum_{j \in \Lambda(i)} a_{ij} \ge 0 \tag{3.21}$$

holds.

(*iii*) At least at one point $x_i \in \bar{\omega}_I\,, I = P, S$ the inequality (3.21) is strict, i.e., instead of $\ge$ we have $>$.

(*iv*) If (3.21) is equality at a point $x_i \in \bar{\omega}_I\,, I = P, S$, then there is a finite sequence of grid points $x^{(0)}$, $x^{(1)}, \ldots, x^{(m)}$ from $\bar{\omega}_I$ with $x_i^{(0)} := x_i$, $x^{(j)} \in \Lambda(j-1)$ for $j = 1, 2, \ldots, , m$ $(m \ge)$ and $(\mathcal{L}_h(1))(x^{(m)}) > 0$, $(\mathcal{L}_h(1))(x^{(j)}) = 0$ for $j = 0, 1, \ldots, m-1$.

Theorem 3.4 (Discrete maximum principle) *Let y be a grid function defined on a connected grid $\bar{\omega}$ and the conditions (i)–(iv) be satisfied. If $\mathcal{L}_h(y)(x_i) \ge 0$ and $y_{|(\bar{\omega}\setminus\omega)} \ge 0$, then $y(x_i) \ge 0, \quad x_i \in \omega$.*

Remark 3.3 If the condition (*ii*) is replaced by:
(*ii*)′ For each $x_i \in \bar{\omega}_I\,, I = P, S$, the following inequality holds

$$a_{ii} + \sum_{j \in \Lambda(i)} a_{ij} > 0\,. \tag{3.22}$$

Then the the discrete maximum principle follows from (*i*) and (*ii*)′. The inequality (3.22) means that the matrix is strictly diagonally dominant.

Closely related with discrete maximum principle are monotone matrices [131].

Definition 3.1 ([137]) A matrix A is a monotone matrix , if A is nonsingular and $A^{-1} \ge 0$.

We use the notation $B \ge 0$ for a matrix $B = \{b_{ij}\}_{i,j=1}^{n}$ to denote that for every entry $b_{ij} \ge 0$ holds.

Theorem 3.5 *A real matrix A is monotone if and only if $A\mathbf{x} \ge 0$ implies $\mathbf{x} \ge 0$.*

The notion of an **M**–matrix is stronger condition than monotonicity.

Definition 3.2 ([137]) A real matrix A is an **M**–matrix if A is nonsingular and the following conditions hold:

$$\begin{aligned} &(i) \quad a_{ij} \le 0\,, \qquad i \ne j\,, \\ &(ii) \quad A^{-1} \ge 0\,. \end{aligned}$$

Theorem 3.6 ([137]) *If A is an **M**–matrix, then A is a monotone matrix. On the other hand , if A is a monotone matrix, such that* (3.23) *holds, then A is an M–matrix.*

Theorems 3.4 and 3.6 show that the conditions (*i*)–(*iv*) guarantee that the matrix $\{a_{ij}\}$ is an M–matrix.

We distinguish global and local conservation properties. It will become clear from the Definition 3.3 that the local conservation property and an appropriate treatment of the boundary conditions imply a global conservation property. Examples show that a global conservation property does not imply a local conservation property (standard finite element methods are classical examples).

Definition 3.3 (Discrete conservation) We say that a particular discretization method is discrete conservative, if for any connected volume V that is a union of control volumes the sum of discrete fluxes on ∂V is equal to the sum of discrete sources/sinks inside V.

We impose two natural conditions on the control volumes (cf. [60]).

Assumption 3.1 *The union of all control volumes in Ω is equal to the domain Ω.*

We call two volumes V_i and V_j "neighbors", if they share a common face γ_{ij}.

Assumption 3.2 *Control volumes may overlap, but every control volume has to have a "neighbor" on each of its faces.*

Examples of overlapping control volumes have been investigated by Schmidt [116]. All the control volumes considered in this chapter satisfy the Assumptions 3.1 and 3.2.

We impose also the following condition on the approximate fluxes:

Assumption 3.3 *Suppose that the control volumes V_i and V_j are "neighbors" and $\gamma_{ij} = V_i \cap V_j$. Denote by q_{ij} the approximation of $\int_{\gamma_{ij}} (\mathbf{q}, \mathbf{n}_{V_i})\, ds$ and by q_{ji} the approximation of $\int_{\gamma_{ij}} (\mathbf{q}, \mathbf{n}_{V_j})\, ds$. Then we require that the following equality be satisfied*

$$q_{ij} + q_{ji} = 0\,. \tag{3.23}$$

Here $\mathbf{n}_{V_i}$ is the outward normal to ∂V_i and $\mathbf{n}_{V_j}$ is the outward normal to ∂V_j restricted to γ_{ij}.

The following result is self–evident.

Proposition 3.2 *Every finite volume method defined by* (3.2) *and satisfying Assumptions 3.1, 3.2 and 3.3 is discrete conservative.*

3.3 Cell–centered finite volume methods

In this section we define the classical 5–pont cell–centered FV difference schemes and investigate one possible generalization. The second subsection is devoted to a short introduction of mixed finite element methods and their relations with cell–centered FV difference schemes.

3.3.1 Difference methods

All discussed difference schemes are considered either on circumscribed cell–centered grids or on Voronoi cell–centered grids. In both cases the straight line connecting two neighboring vertexes is perpendicular to the face that separates them.

3.3.1.1 Scalar diffusion coefficient

We consider first the case of a scalar diffusion coefficient, i.e., $A(\mathrm{x}) = a(\mathrm{x})I$, where I is the identity $d \times d$ matrix. Uniform ellipticity guarantees that $a(\mathrm{x}) \geq C^{-1} > 0$. We split the flux $\mathbf{Q}$ into a diffusive part $\mathbf{W}$ and a convective part $\mathbf{V}$, $\mathbf{Q} = \mathbf{W} + \mathbf{V}$. We want to approximate

the integrals $\int_{\gamma_{ij}} (\mathbf{W}, \mathbf{n})\, ds$ and $\int_{\gamma_{ij}} (\mathbf{V}, \mathbf{n})\, ds$. Pick a point $x_{ij} \in \gamma_{ij}$ and apply the simplest one point quadrature formula

$$\int_{\gamma_{ij}} (\mathbf{W}, \mathbf{n})\, ds \approx \text{meas}(\gamma_{ij})(\mathbf{W}, \mathbf{n})(x_{ij})\,, \tag{3.24a}$$

$$\int_{\gamma_{ij}} (\mathbf{V}, \mathbf{n})\, ds \approx \text{meas}(\gamma_{ij})(\mathbf{W}, \mathbf{n})(x_{ij})\,. \tag{3.24b}$$

We choose x_{ij} to be the intersection point of γ_{ij} and the straight line through x_i and x_j. We can rewrite $(\mathbf{W}, \mathbf{n}) = (-a(\mathrm{x})\nabla u, \mathbf{n})$ as

$$\frac{\partial u}{\partial \mathbf{n}} = -\frac{(\mathbf{W}, \mathbf{n})}{a(\mathrm{x})}$$

and integrate along the interval with end points x_i and x_j. Since the integration is performed along the normal vector $\mathbf{n}$ we have

$$u_i - u_j = -\int_{x_i}^{x_j} \frac{(\mathbf{W}, \mathbf{n})}{a(s)}\, ds \approx -(\mathbf{W}, \mathbf{n})(x_{ij}) \int_{x_i}^{x_j} \frac{ds}{a(s)}\,. \tag{3.25}$$

Combining (3.24a) and (3.25) we can write the following approximate relation for the diffusive flux

$$\int_{\gamma_{ij}} (\mathbf{W}, \mathbf{n})\, ds \approx -meas(\gamma_{ij}) \left(\frac{1}{dist(x_i, x_j)} \int_{x_i}^{x_j} \frac{ds}{a(s)} \right)^{-1} \frac{[u_i - u_j]}{dist(x_i, x_j)}\,.$$

This approximate relation allows us to define the approximate diffusive flux w_{ij} on γ_{ij} with the the following formulas

$$w_{ij}(\mathrm{x}) \equiv -\,\text{meas}(\gamma_{ij}) k_{ij} \frac{[u_{h,j} - u_{h,i}]}{dist(x_i, x_j)}\,, \tag{3.26}$$

where

$$k_{ij} = \left(\frac{1}{dist(x_i, x_j)} \int_{x_i}^{x_j} \frac{ds}{a(s)} \right)^{-1} \tag{3.27}$$

and u_h is the approximate solution. Usually the coefficients k_{ij} are called harmonic average transmissibilities.

The integral $\int_{\gamma_{ij}} (\mathbf{V}, \mathbf{n})\, ds$ can be approximated as follows

$$\begin{aligned} \int_{\gamma_{ij}} (\mathbf{V}, \mathbf{n})\, ds &\approx \text{meas}(\gamma_{ij})(\mathbf{b}, \mathbf{n})(x_{ij}) u(x_{ij}) \\ &\approx \text{meas}(\gamma_{ij})(\mathbf{b}, \mathbf{n})(x_{ij}) \left[\frac{dist(x_j, x_{ij})}{dist(x_i, x_j)} u_i + \frac{dist(x_i, x_{ij})}{dist(x_i, x_j)} u_j \right]. \end{aligned}$$

And thus we can define the approximation:

$$v_{ij}(\mathrm{x}) = \text{meas}(\gamma_{ij})(\mathbf{b}, \mathbf{n})(x_{ij}) \left[\frac{dist(x_j, x_{ij})}{dist(x_i, x_j)} u_{h,i} + \frac{dist(x_i, x_{ij})}{dist(x_i, x_j)} u_{h,j} \right]. \tag{3.28}$$

In the next chapter we will consider other approximations of the convective flux that are especially suited for convection dominated problems.

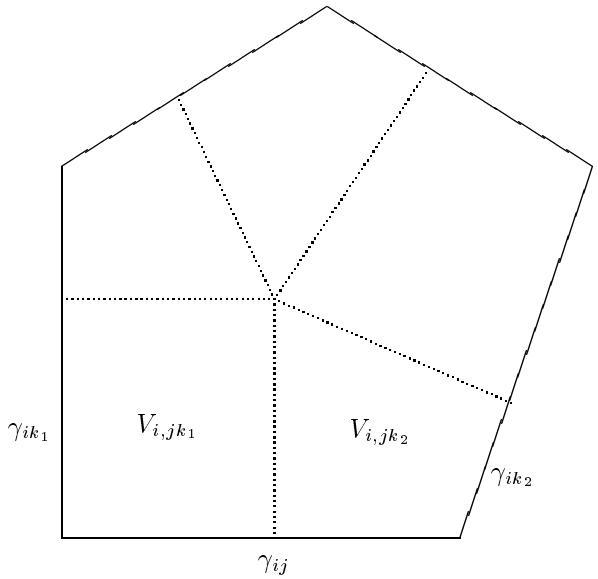


Figure 3.5: Quadrilateral parts $V_{i,jk}$ in the volume V_i

3.3.1.2 Tensor diffusion coefficient

Approximating diffusive fluxes with a full tensor $A(\mathrm{x})$ on distorted meshes is still a "hot" problem. There are many "ad hoc" algorithms in the engineering literature, especially in Computational Fluid Dynamics. We sketch one scheme that can be considered as a generalization of the classical 5–point cell–centered schemes. This method is an extension of the work of Ware, Parrott and Rogers [134] where only rectangular meshes have been used.

We assume that the grid is Voronoi and that the intervals (x_i, x_j) intersect the faces γ_{ij}. The intervals (x_i, x_j), $j \in \Sigma(i)$ divide the control volume V_i into quadrilateral parts $V_{i,jk}$, $j, k \in \Sigma(i)$, $j \neq k$. A simple example is shown on Fig. 3.5. We approximate the flux $\mathbf{W}$ in $V_{i,jk}$ with a constant vector $\mathbf{w}_{i,jk}$. Therefore, the conservation law takes the form

$$\sum_{j \in \Sigma(i)} \left[(\mathbf{w}_{i,jk_1}, \mathbf{n}_{ij}) \operatorname{meas}(\gamma_{ij} \cap V_{i,jk_1}) + (\mathbf{w}_{i,jk_2}, \mathbf{n}_{ij}) \operatorname{meas}(\gamma_{ij} \cap V_{i,jk_2}) \right] = \phi_i \tag{3.29}$$

Here γ_{ik_1} and γ_{ik_2} are two neighbors of γ_{ij} and $\mathbf{n}_{ij}$ is the outward normal vector to γ_{ij}. We have to express $\mathbf{w}_{i,jk}$ through the values of the approximate solution $u_{h,i}$, $u_{h,j}$ and $u_{h,k}$ (k is k_1 and k_2 correspondingly).

We split the integral of the flux into two parts

$$\int_{\gamma_{ij}} (\mathbf{W}, \mathbf{n}_{ij})\, ds = \int_{\gamma_{ij} \cap V_{i,jk_1}} (\mathbf{W}, \mathbf{n}_{ij})\, ds + \int_{\gamma_{ij} \cap V_{i,jk_2}} (\mathbf{W}, \mathbf{n}_{ij})\, ds\,.$$

and approximate each of them

$$\int_{\gamma_{ij} \cap V_{i,jk_1}} (\mathbf{W}, \mathbf{n}_{ij})\, ds \approx \int_{\gamma_{ij} \cap V_{i,jk_1}} (\mathbf{w}_{i,jk}, \mathbf{n}_{ij})\, ds\,.$$

Suppose that $A(\mathrm{x})$ is a piecewise constant matrix in each control volume, i.e., $A(\mathrm{x}) = A_i$ for

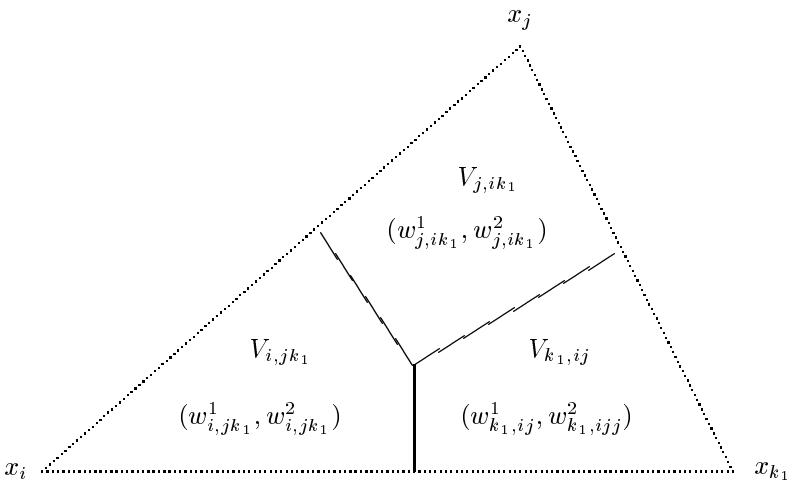


Figure 3.6: Three quadrilaterals

x $\in V_i$ and $A_i^{-1} = L_i$. By the definition of the flux $\mathbf{W}$ and L_i

$$\mathbf{W} = -A_i \nabla u\,, \qquad L_i \mathbf{W} = -\nabla u\,.$$

and taking the inner product with $\mathbf{n}_{ij}$ we have

$$(L_i \mathbf{W}, \mathbf{n}_{ij}) = (\mathbf{W}, L_i \mathbf{n}_{ij}) = (\mathbf{W}, \mathbf{l}_{i,ij}) = -(\nabla u, \mathbf{n}_{ij})$$

We denote the vector $L_i \mathbf{n}_j$ by $\mathbf{l}_{i,ij}$.

For 2–D Voronoi meshes the degree of every Voronoi vertex is either 3 or 4, i.e., either three or four control volumes intersect in this point. First we consider one particular Voronoi vertex with degree three. Let V_{i,j,k_1}, V_{j,ik_1} and $V_{k_1,ij}$ be three neighboring quadrilaterals sketched on the Fig. 3.6. Integrating from x_i to x_j and approximating $\mathbf{W}$ with $\mathbf{w}_{i,jk_1}$ in V_{i,jk_1} and with $\mathbf{w}_{j,ik_1}$ in V_{j,ik_1} we get

$$\begin{aligned} u_i - u_j = -\int_{x_i}^{x_j} (\nabla u, \mathbf{n}_{ij})\, ds &= \int_{x_i}^{x_{ij}} (\mathbf{W}, L_i \mathbf{n}_{ij})\, ds + \int_{x_{ij}}^{x_j} (\mathbf{W}, L_j \mathbf{n}_{ij})\, ds \\ &\approx \int_{x_i}^{x_{ij}} (\mathbf{w}_{i,jk}, \mathbf{l}_{i,ij})\, ds + \int_{x_{ij}}^{x_j} (\mathbf{w}_{j,ik_1}, \mathbf{l}_{j,ij})\, ds \\ &= dist(x_i, x_{i,j}) \left(w^1_{i,jk_1} l^1_{i,ij} + w^2_{i,jk_1} l^2_{i,ij} \right) \\ &\quad + dist(x_{i,j}, x_j) \left(w^1_{j,ik_1} l^1_{j,ij} + w^2_{j,ik_1} l^2_{j,ij} \right)\,. \end{aligned}$$

The continuity of the flux across γ_{ij} gives the relation

$$w^1_{i,jk_1} n^1_{ij} + w^2_{i,jk_1} n^2_{ij} = w^1_{j,ik_1} n^1_{ij} + w^2_{j,ik_1} n^2_{ij}\,.$$

In the same way we derive four more equations integrating along (x_i, x_{k_1}) and (x_j, x_{k_1}) and using the assumption for the continuity of normal flux on the control volume faces. Note that as long as we consider only one triangle we can use the notation $\mathbf{w}_i$ for $\mathbf{w}_{i,jk_1}$, $\mathbf{w}_j$ for $\mathbf{w}_{j,ik_1}$ and $\mathbf{w}_{k_1}$ for $\mathbf{w}_{k_1,ij}$. We also denote in the system below $dist(.,.)$ with $d(.,.)$. These

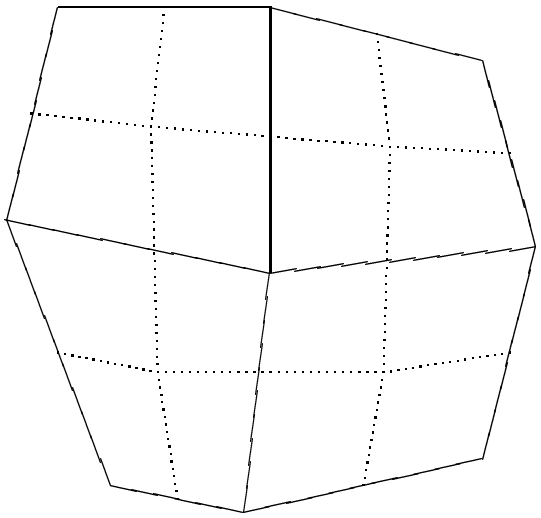


Figure 3.7: Degree four – four quadrilaterals

six equations form a linear system for the unknown fluxes $\mathbf{w}_i$, $\mathbf{w}_j$ and $\mathbf{w}_{k_1}$.

$$
\begin{aligned}
d(x_i, x_{i,j})\left(w_i^1 l_{i,ij}^1 + w_i^2 l_{i,ij}^2\right) + d(x_{i,j}, x_j)\left(w_j^1 l_{j,ij}^1 + w_j^2 l_{j,ij}^2\right) &= u_i - u_j\,, \\
w_i^1 n_{ij}^1 + w_i^2 n_{ij}^2 - w_j^1 n_{ij}^1 - w_j^2 n_{ij}^2 &= 0\,, \\
d(x_i, x_{i,k_1})\left(w_i^1 l_{i,ik_1}^1 + w_i^2 l_{i,ik_1}^2\right) + d(x_{i,k_1}, x_{k_1})\left(w_{k_1}^1 l_{k_1,ik_1}^1 + w_{k_1}^2 l_{k_1,ik_1}^2\right) &= u_i - u_{k_1}\,, \\
w_i^1 n_{ik_1}^1 + w_i^2 n_{ik_1}^2 - w_{k_1}^1 n_{ik_1}^1 - w_{k_1}^2 n_{ik_1}^2 &= 0\,, \\
d(x_{k_1}, x_{j,k_1})\left(w_{k_1}^1 l_{k_1,jk_1}^1 + w_{k_1}^2 l_{k_1,jk_1}^2\right) + d(x_{j,k_1}, x_j)\left(w_j^1 l_{j,jk_1}^1 + w_j^2 l_{j,jk_1}^2\right) &= u_{k_1} - u_j\,, \\
w_{k_1}^1 n_{jk_1}^1 + w_{k_1}^2 n_{jk_1}^2 - w_j^1 n_{jk_1}^1 - w_j^2 n_{jk_1}^2 &= 0\,.
\end{aligned}
$$

We can solve this 6×6 linear system for the fluxes and substitute the result into the discrete conservation law (3.29).

If the considered vertex has degree four then the system is 8×8. This case is shown on Fig. 3.7.

This scheme shows striking similarity with mixed finite element methods and probably in some cases can be considered as a MFEM method with a properly chosen quadrature formulae. We will not attempt to investigate the properties of the derived cell–centered finite volume method for tensor coefficients in this dissertation.

3.3.2 Mixed finite element methods

The theory of mixed finite element methods for nonsymmetric problems has been developed by Douglas and Roberts in a series of papers [37, 38]. Here we define the discrete problem and briefly discuss its relation with cell–centered FV difference methods.

Let $\mathbf{V}_h \subset H_{div}(\Omega)$ and $W_h \subset L^2(\Omega)$. We consider the mixed finite element method define as:

Find the pair $(\mathbf{q}_h, u_h) \in \mathbf{V}_h \times W_h$ such that

$$(K\mathbf{q}_h, \mathbf{v})_0 - (\mathrm{div}(\mathbf{v}), u_h)_0 - (\beta u_h, \mathbf{v})_0 = 0 \qquad \forall\, \mathbf{v} \in \mathbf{V}_h\,, \tag{3.30a}$$

$$(\mathrm{div}(\mathbf{q}), w)_0 = (f, w)_0 \forall\, w \in W_h\,. \tag{3.30b}$$

Suppose we can express $\mathbf{q}_h$ through u_h from (3.30a) and substitute this relation in (3.30b). This will lead to a particular cell–centered finite volume difference method. Usually this is achieved by using a quadrature formulae to approximate the inner products in (3.30) (cf. [112, 4]).

3.4 Finite volume element methods

The discrete FVE problem is defined as follows:

Find $u_h \in \mathcal{V}_0^h$ such that, for all vertex–centered control volumes V_i, $i = 1, \ldots, n_P$

$$\int_{\partial V_i} (-A\nabla u_h + \mathbf{b}u_h, \mathbf{n})\, ds = \int_{V_i} f\, dx\,. \tag{3.31}$$

In order to investigate the problem (3.31) we introduce the following bilinear forms:

$$B_h(u_h, v_h) = -\sum_{i=1}^{n_P} \int_{\partial V_i} (A\nabla u_h, \mathbf{n})\, ds\; v_h(x_i) + \sum_{i=1}^{n_P} \int_{\partial V_i} (\mathbf{b}, \mathbf{n}) u_h\, ds\; v_h(x_i)\,. \tag{3.32}$$

Note that $\int_{\partial V_i} (-A\nabla u_h + \mathbf{b}u_h, \mathbf{n}) v_h\, ds$ is not well defined for functions $v_h \in \mathcal{W}$. To overcome this little discrepancy we define

$$\int_{\partial V_i} (-A\nabla u_h + \mathbf{b}u_h, \mathbf{n}) v_h\, ds = \lim_{n\to\infty} \int_{\partial V_i^{(n)}} (-A\nabla u_h + \mathbf{b}u_h, \mathbf{n}) v_h\, ds\,, \tag{3.33}$$

where $V_i^{(n)} \subset\subset V_i$ and $dist(\partial V_i, \partial V_i^{(n)}) \le 1/n$. Since V_i is a domain with Lipschitz–continuous boundary such a sequence of domains $V_i^{(n)}$ exists. Using the Lebesgue dominated convergence theorem we see that the limit on the right hand side exists and is equal to $\int_{\partial V_i} (-A\nabla u_h + \mathbf{b}u_h\mathbf{n})\, ds\, v_h(x_i)$.

Therefore, we can use the notation

$$B(u_h, v_h) = \sum_{i=1}^{n_P} \int_{\partial V_i} (-A\nabla u_h + \mathbf{b}u_h, \mathbf{n}) v_h\, ds\,.$$

and the problem (3.31) is equivalent to

Find $u_h \in \mathcal{V}_0^h$ such that

$$B_h(u_h, v_h) = f(v_h) \qquad \forall\, v_h \in \mathcal{W}^h\,,$$

where

$$f(v_h) = \sum_{i=1}^{n_P} \int_{V_i} f\, \mathrm{dx}\, v(x_i)\,.$$

In fact, we have formulated the problem (3.31) as a Petrov–Galerkin method.

In Chapter VI we will investigate the properties of the bilinear form $B_h(.,.)$.

CHAPTER IV

FINITE VOLUME METHODS FOR NONSYMMETRIC PROBLEMS

In this chapter we construct cell–centered FV difference schemes for the continuous problem discussed in Chapter II (cf. (2.10), (2.11)). We pay special attention to the convection dominated case, i.e., when the ratio $\|A\|/\|\mathbf{b}\| \ll 1$. A typical example of such an equation is the model singularly perturbed problem:

$$\begin{aligned} -\varepsilon\Delta u + \operatorname{div}(\mathbf{b}u) &= f \quad \text{in} \quad \Omega\,, \\ u &= 0 \quad \text{on} \quad \partial\Omega \end{aligned} \tag{4.1}$$

with $0 < \varepsilon \ll 1$ and $\|\mathbf{b}\| = O(1)$. The problem (4.1) exhibits both hyperbolic and elliptic features and, moreover, its solution possesses boundary (and in some cases interior) layers in general, i.e., the derivatives of u are of order $O(\varepsilon^{-p})$ for some positive number p in certain subdomains of Ω with measure of the order of ε.

Our goal is to develop approximation methods that have the following properties:

(i) stability,

(ii) "good" approximation,

(iii) local conservation,

(iv) satisfy the discrete maximum principle,

(v) produce positive definite matrices,

(vi) work for general domains and suitable grids introduced into them.

We briefly discuss the existent methods for the solution of the problem (4.1) and point out what are our objectives.

The approximation methods for the solution of convection dominated problems can be separated into three large groups: *characteristics* methods, methods that require special types of *refinement* into the boundary layer regions, and methods on *uniform or regular* grids.

The characteristics numerical methods use special meshes that are aligned with the characteristics of the hyperbolic part of (4.1) (for parabolic problems see [40], [44]). Because of their problem dependence, we will not consider such methods in this dissertation. The methods in the second group are based on a simple idea due to the Russian mathematician Bakhvalov [12] to construct an 1–D mesh in a such way that on each interval the error of the approximation is almost the same. Clearly, these constructions can be extended to tensor product grids, but not to general meshes (cf. Shishkin [118, 56] for multidimensional generalizations). We will discuss only methods from the third group.

For problems like (4.1) the standard finite element and finite difference methods (central finite difference approximation of the convection term) are conditionally stable, i.e., only for sufficiently small h. For the model problem (4.1), $h \approx \varepsilon$, and this requirement makes using standard methods prohibitively expensive for 1–D problems and impossible for multidimensional ones. The upwind finite difference schemes [111, 121] and corresponding finite element methods with modified basis functions, also called Petrov–Galerkin methods

[58, 17], overcome the stability restrictions, but reduce the accuracy to first order and introduce considerable smearing. Upwind finite element methods have been extensively studied by Japanese mathematicians Baba, Tabata [10], Ikeda et.al (see the monograph by Ikeda [63] and references there). Their methods handle the convection term in a way very similar to finite volume methods. Some modified upwind schemes have been proposed by Samarskii [113] and later by Axelsson and Gustafsson [8] . It has been shown that they have second order of accuracy (cf. also [90, 78]). The work of Il'in [64] established the exponentially fitted methods as another alternative to obtain stable approximations. More resent developments in the finite element theory are the streamline-diffusion methods ([62, 25, 68]) and strongly consistent stabilization methods (see [105] and works cited there). Instead of adding artificial viscosity in all direction as upwind methods, streamline-diffusion methods try to stabilize the problem via increasing the diffusion only in the directions of the streamlines. Unfortunately, this can be done only for a constant vector $\mathbf{b}$. For varying convection coefficients $\mathbf{b}$ these schemes get closer to upwind methods and keep only the advantage that there are strongly consistent in sense that the exact solution satisfies the difference scheme. An uniform framework to describe different upwind methods including also streamline-diffusion case have been proposed by Bank et. al. [15]. Discontinuous Galerkin methods [106, 107, 46], sometimes called explicit methods, are also stable for $\varepsilon \leq h$ and seems to be very efficient for the constant coefficient case.

The problem of obtaining a discretization scheme with good approximation properties is considerably more difficult than that of satisfying the stability requirements. For elliptic problems the classical error estimates (for finite element theory see Ciarlet [31]) and for finite difference schemes see Samarskii, Lazarov and Makarov [115]) are of the following type:

$$\|u - u_h\|_{k,\Omega} \leq Ch^{m-k}|u|_{m,\Omega}\,.$$

Since the m^{th} derivative of the solution u is not bounded as $\varepsilon \to 0$ these estimates are not appropriate for convection–dominated problems. Another definition of "good approximation properties" can be

$$\|u - u_h\|_{*,\Omega} \leq Ch^{\alpha}, \quad \alpha > 0\,, \tag{4.2}$$

where C does not depend on ε and h and $\|.\|_{*,\Omega}$ is a suitable norm. Discretization methods that produce approximations satisfying (4.2) are called *uniformly convergent* (with respect to ε). Note that the estimate (4.2) requires global convergence. Allen and Southwell [2] proposed the first scheme that later was proven to converge uniformly for 1–D problems, but their work was not widely noticed. In the late sixties a Russian mathematician A. Il'in independently rediscovered exponentially fitted schemes. His work initiated the development of many methods (at least ten according to Roos [109]) and as a result comprehensive theory for 1–D problems crystallized (cf. [36]). The attempts to generalize exponentially fitted schemes for multidimensional problems do not produce the same results. The best known global estimates of $O(h^{1/2})$ convergence rate are proven by O'Riordan and Stynes [99] for a special exponentially fitted discretization. The same accuracy can be achieved by solving only the hyperbolic part of (4.1). Our numerical experiments show that upwind schemes also asymptoticly converge with half an order. There are no available better global results for streamline-diffusion methods either. If we consider only the maximum error away from the boundary layers streamline-diffusion methods for problems with constant convection coefficients are definitely superior to the upwind methods. Johnson, Shatz and Walbin [67] have proven an estimate later improved by Nijima [96] to $h^{11/8}|\log h|$ and recently Zhou and Rannacher [138] have shown optimal error in maximum norm for special grids.

Although the construction of uniformly convergent approximations of singularly perturbed problems is a notoriously difficult problem that is still not solved, the explanation for this is

relatively simple. The 1–D problems like (4.1) have solutions that are essentially exponential functions and therefore, every method which can approximate exponentials well has good properties. The behavior of the solution dramatically changes for multidimensional problems. New types of layers like ordinary differential, parabolic, elliptic and corner layers have been studied for 2–D problems (cf. the survey papers by Eckhaus [39] and Shih and Kellogg [117]) and probably the solutions are more complicated for higher dimension (still not classified). The different types of layers just mentioned cannot be approximated easily by exponential like functions. For example, parabolic and elliptic boundary layers are solution of properly constructed parabolic or elliptic partial differential equations, correspondingly. In fact, the approximation rates of piecewise exponential function is not better than that of the piecewise linear functions.

Only a few of the methods of finite volume type as those proposed by Spalding [121] and Runchal [111] possess local conservation properties. In resent years the first attempts have been made to construct mixed finite element methods for convection dominated problems by Junping Wang et. al. [83] and van Nooyen [130]. For related problems that consider models of semiconductor devices, mixed approximations have been proposed by Miller and Wang [89] and and Brezzi, Marini and Pietra [23].

In summary: there are still no reliable and robust discretization methods that outperform the upwind and exponentially fitted schemes in terms of global accuracy and have the properties (i), $(iii) - (vi)$. Therefore, methods of upwind type that satisfy $(i) - (v)$ will be superior to the known ones.

This chapter is devoted to the construction of monotone cell-centered FV difference schemes for convection-diffusion equations on Voronoi and curcumscribed grids that are unconditionally stable and have second-order accuracy in a discrete H^1–norm for grids that satisfy some additional assumptions. From the discussion above is clear that we can hope for $O(h^2)$ convergence rate only in the diffusion dominated case. For such problems Samarskii [114] has proven such an estimate in the discrete maximum norm under rather demanding assumptions on the solution (to have four continuous derivatives). A general approach for cell-centered finite difference schemes on triangles including local refinement was considered in Vassilevski, Petrova and Lazarov [132]. The error estimates derived in [132] are in a discrete H^1-norm including some superconvergence type estimates on uniform triangulations, namely, $O(h^2)$ error estimate on uniform triangulations. Our results generalize the results in [132] in two directions: they are proven for convection–diffusion problems and are valid also for more general grids in 2–D and 3–D.

Cell-centered discretizations on tensor-product nonuniform meshes were considered by Weiser and Wheeler [135] and superconvergence type error estimates derived. Similar results for the Poisson equation were proved in Süli [123], i.e., H^1-estimates of order $O(h^{1+\alpha})$, $\frac{1}{2} < \alpha \leq 1$. Cai [26], Cai, Mandel and McCormick [27] also have shown some superconvergence results for finite volume element methods. In Hackbusch [55] second order error estimates in H^1-norm on uniform mesh has been proved.

We also provide error estimates in L^2-norm elaborating the discrete "Aubin-Nitsche trick" of duality argument proposed in Samarskii, Lazarov, and Makarov [115] and used in the case of finite difference schemes for general self-adjoint elliptic equations in Lazarov, Makarov and Weinelt [76]. For the original duality technique in the finite element method, cf., Aubin [6], Nitsche [97], which can also be found in Ciarlet [31]. For another approach for L^2 error estimates see Herbin [59].

The remainder of the chapter is organized as follows. The discretization schemes are presented in Section 4.1. It is shown that they satisfy the discrete maximum principle and the discrete operators are positive definite. The stability (a priori estimates) and error estimates in H^1-norm are derived in Section 4.2.1. The error estimates in L^2-norm are proved in Section

4.2.2. Finally, in Section 4,3 some computational results that illustrate the developed theory are presented.

4.1 Discretization schemes

We assume that the triangulation is given by circumscribed or Voronoi cell–centered grid ω (we skip in this chapter the subindex S). For such triangulations we add an extra regularity condition to condition (3.3) in order to accommodate more general elements (control volumes).

Assumption 4.1 (FV regular triangulations) *We say that a cell–centered triangulation* $\{V_i\}_{i=1}^{n_S}$ *is* finite volume regular *if every control volume satisfies* (3.3) *and, moreover, there exist two positive constants* C_1 *and* C_2 *such that the following inequalities hold*

$$C_1 \,\mathrm{meas}(\gamma_{ij}) dist(x_i, x_j) \le \mathrm{meas}(V_i) \le C_2 \,\mathrm{meas}(\gamma_{ij}) dist(x_i, x_j) \quad i = 1, \ldots, n_S,\ j \in \Sigma(i).$$

The rate of convergence of cell–centered finite volume methods depends on the geometric properties of the triangulation. For some special triangulations we will prove that higher rates of convergence can be achieved for properly designed finite volume methods (cf. Theorem 4.1). Such triangulations usually exhibit some special symmetries of the position of the point $x_{ij} = (x_i, x_j) \cup \gamma_{ij}$ with respect to the points x_i and x_j and to the face γ_{ij}. The exact conditions are formulated in the following assumption.

Assumption 4.2 (The symmetry assumption) *We say that the finite volume triangulation* $\{V_i\}_{i=1}^{n_S}$ *satisfies the symmetry assumption if the following conditions hold:*

(i) x_{ij} *is the middle point of the interval* (x_i, x_j)*;*

(ii) for triangular faces γ_{ij}*,* x_{ij} *is the barycenter of* γ_{ij}*. Otherwise, we require that* γ_{ij} *has two perpendicular axes of symmetry and* x_{ij} *is their intersection point.*

We point out that the symmetry assumption is only a sufficient condition.

We recall the derivation of finite volume approximation of the equation (2.11a). We integrate (2.11a) over each cell–centered control volume V_i, $i = 1, \ldots n_S$

$$\int_{V_i} \mathrm{div}(-a(x)\nabla u(x) + \mathbf{b}(x)u(x))\, dx = \int_{V_i} f(x)\, dx$$

and then using the Green's formula and dividing by $meas(V_i)$ we get

$$\frac{1}{meas(V_i)} \int_{\partial V_i} (-a\nabla u + \mathbf{b}u, \mathbf{n})\, ds = \frac{1}{meas(V_i)} \int_{V_i} f(x)\, dx \tag{4.3}$$

where $\mathbf{n}$ is the unit outward vector normal to the boundary of V_i. Denote

$$\mathbf{W} = -a(\mathrm{x})\nabla u(\mathrm{x}) \quad \text{and} \quad \mathbf{V} = \mathbf{b}(\mathrm{x})u(\mathrm{x}).$$

Splitting $\partial V_i = \cup_{j\in\Sigma(i)} \gamma_{ij}$ (see Fig. 4.1) the left-hand side of this identity is written in the form:

$$\frac{1}{\mathrm{meas}(V_i)} \left[\int_{\partial V_i} (\mathbf{W}, \mathbf{n})\, ds + \int_{\partial V_i} (\mathbf{V}, \mathbf{n})\, ds \right] = \frac{1}{meas(V_i)} \left[\sum_{j\in\Sigma(i)} \int_{\gamma_{ij}} (\mathbf{W}, \mathbf{n})\, ds + \sum_{j\in\Sigma(i)} \int_{\gamma_{ij}} (\mathbf{V}, \mathbf{n})\, ds \right] \tag{4.4}$$

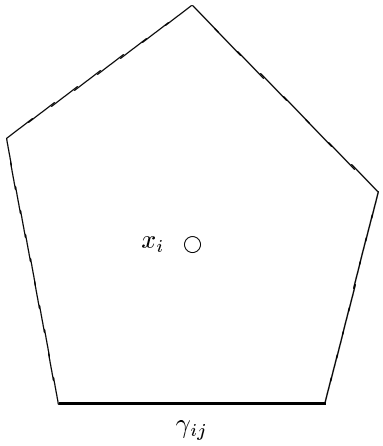


Figure 4.1: General control volume V_i

In order to construct the finite difference scheme we approximate the balance equation (4.4). We split the approximation of the balance equation (4.4) in two parts

$$A^{(2)}u_h + A^{(1)}u_h \tag{4.5}$$

where $A^{(2)}$ is the part arising from the approximation of the second derivatives, and $A^{(1)}$ comes from the approximation of the first derivatives; u_h is an approximation to the exact solution solution u. We have the expressions

$$\begin{aligned} A^{(2)}u_h &= \sum_{j\in\Sigma(i)} w_{i,j}\,, \quad x_i \in \omega\,, \\ A^{(1)}u_h &= \sum_{j\in\Sigma(i)} v_{i,j}\,, \quad x_i \in \omega\,. \end{aligned} \tag{4.6}$$

In these formulae w_{ij} and v_{ij} are some approximations of the corresponding integrals $\int_{\gamma_{ij}}(\mathbf{W},\mathbf{n})\,ds$ and $\int_{\gamma_{ij}}(\mathbf{V},\mathbf{n})\,ds$. Now, in order to complete the finite difference scheme we have to express the approximate fluxes w_{ij} and v_{ij} by the approximate values $u_h(x)$ of the solution $u(x)$ at the grid points. We consider the following approximations:

1. central difference scheme (**CDS**);
2. upwind difference scheme (**UDS**);
3. modified upwind difference scheme (**MUDS**);
4. Il'in's difference scheme (**IDS**).

We denote by $\beta_{i,j}$ an approximation of the integral $\int_{\gamma_{ij}}(\mathbf{b},\mathbf{n})\,ds$ with the properties:

$$(i) \quad \beta_{i,j} + \beta_{j,i} = 0\,. \tag{4.7a}$$

$$(ii) \quad |\beta_{i,j}| \le C meas(\gamma_{ij})\|\mathbf{b}\|_{d/2+\alpha,\infty,\Omega}\,, \tag{4.7b}$$

$$(iii) \quad \left|\int_{\gamma_{ij}} (\mathbf{b},\mathbf{n})\; ds - \beta_{i,j}\right| \le C h^{d+\alpha}|\mathbf{b}|_{1+\alpha,\infty,\Omega}\,, \tag{4.7c}$$

where C is a positive constant and $\alpha > 0$. We consider some examples of quadrature formulas that satisfy conditions (4.7). Let γ_{ij} be an interval with end points a_1 and a_2 and a middle point a_{12}. The following well known quadrature formulas clearly satisfy the conditions (4.7)

$$\beta_{ij} = (\mathbf{b}, \mathbf{n})(a_{12}) meas(\gamma_{ij}),$$
$$\beta_{ij} = \frac{meas(\gamma_{ij})}{2} \left[(\mathbf{b}, \mathbf{n})(a_1) + (\mathbf{b}, \mathbf{n})(a_2) \right].$$

For triangular and rectangular faces γ_{ij} the quadrature formulae

$$\beta_{ij} = (\mathbf{b}, \mathbf{n})(a_{bary}) meas(\gamma_{ij})$$

fulfills (4.7) with a_{bary} the barycenter of γ_{ij}.

4.1.1 Central difference scheme (CDS)

We call this scheme "central" because of the analogy of $A^{(1)}$ and a central difference approximation of the first derivatives. We recall the formulas derived in Chapter III

$$w_{ij}(\mathrm{x}) = -\frac{\mathrm{meas}(\gamma_{ij})}{meas(V_i)} k_{ij} \frac{[u_{h,j} - u_{h,i}]}{dist(x_i, x_j)}, \tag{4.8}$$

$$v_{ij}(\mathrm{x}) = \frac{\beta_{i,j}}{meas(V_i)} \left[\frac{dist(x_j, x_{ij})}{dist(x_i, x_j)} u_{h,i} + \frac{dist(x_i, x_{ij})}{dist(x_i, x_j)} u_{h,j} \right] \tag{4.9}$$

with $k_{i,j}$ defined by

$$k_{i,j} = \left(\frac{1}{dist(x_i, x_j)} \int_{x_i}^{x_j} \frac{ds}{a(s)} \right)^{-1}.$$

An application of the discrete maximum principle shows that **CDS** is stable if the following inequalities are satisfied

$$P_i = \max_{j \in \Sigma(i)} \frac{|\beta_{i,j}|}{meas(\gamma_{ij})} . \frac{dist(x_i, x_{ij})}{k_{i,j}} \leq 1, \qquad x_i = 1, \ldots, n_S. \tag{4.10}$$

In some application P_i is called a local (cell) Peclet number (cf. [58], [114]). Note that the quantity $|\beta_{i,j}|/meas(\gamma_{ij})$ does not depend upon h and therefore the inequalities (4.10) are satisfied only for sufficiently small h. We will not further consider the **CDS** because of its conditional stability.

4.1.2 Upwind difference scheme (UDS)

One of the ways to find stable finite difference approximation for convection-diffusion boundary value problem is to use upwind approximation for the first derivatives. In this case, $A^{(2)}$ is defined as in **CDS** and the terms $v_{i,j}$ in $A^{(1)}$ are approximated in the following way:

$$v_{i,j} = \beta^+_{i,j} u_{h,i} + \beta^-_{i,j} u_{h,j} \tag{4.11a}$$

where $\beta^+_{i,j}$ and $\beta^-_{i,j}$ are defined via the formulas

$$\beta^+_{i,j} = \frac{1}{meas(V_i)} . \frac{(\beta_{i,j} + |\beta_{i,j}|)}{2}, \qquad \beta^-_{i,j} = \frac{1}{meas(V_i)} . \frac{(\beta_{i,j} - |\beta_{i,j}|)}{2}. \tag{4.11b}$$

In order to investigate the properties of the **UDS** we need the following auxiliary result.

Proposition 4.1 *Let* $\mathbf{b}(x) \in (W^{d/2x+\alpha,\infty}(\Omega))^d$, $\alpha > 0$ *and there exists a positive constant* β_0 *such that*

$$\int_{\partial V} (\mathbf{b}(x), \mathbf{n})\, ds \geq \beta_0 meas(V) \tag{4.12}$$

for any volume $V \subset \Omega$ *with Lipschitz–continuous boundary* $\partial V = \cup_{j\in S(i)} \gamma_{ij}$. *Suppose that* $\beta_{i,j}$ *satisfies the condition* (4.7c). *Then there exists* h_0 *such that for* $h \in (0, h_0)$ *the following inequality holds:*

$$\sum_{j\in\Sigma(i)} \beta_{i,j} \geq c_0 meas(V)\,, \tag{4.13}$$

where $c_0 = \beta_0 - O(h^\alpha)$.

Proof: It follows from the FV regularity of the control volume V and the condition (4.7c). □

We replace the condition (4.12) with the stronger one.

Assumption 4.3 $\mathbf{b}(x) \in \left(W^{d/2+\alpha,\infty}(\Omega)\right)^d$, $\alpha > 0$ *and* $div(\mathbf{b}(x)) \geq \beta_0 > 0$ *for almost every* $x \in \Omega$.

In fact, this is a stronger version of the Assumption 2.2.

Remark 4.1 We can consider the left hand side of (4.13) as a definition of the discrete divergence operator. Then the above proposition means that, if the divergence of the vector $\mathbf{b}$ is greater than $\beta_0 > 0$, the discrete analogy of $div(\mathbf{b})$ is also positive for sufficiently small h.

First we will prove that the considered scheme is monotone.

Proposition 4.2 *Let the Assumption 4.3 be satisfied, the discrete fluxes* $w_{i,j}$ *and* $v_{i,j}$ *be defined by the formulas* (4.8) *and* (4.11), *respectively, and the approximations* $\beta_{i,j}$ *fulfill the condition* (4.7c). *Then* **UDS** *satisfies the discrete maximum principle and the corresponding matrix* A *is an* **M***–matrix.*

Proof: Let $a_{i,j}$ be the coefficients in front of $u_{h,j}$ in the i^{th} equation. Then it is enough to check the conditions [57]:

1. $a_{i,i} > 0 \qquad a_{i,j} \leq 0 \quad j \neq i$;
2. $a_{i,i} + \sum_{j\in\Sigma(i)} a_{i,j} > 0$, i.e., A is strictly diagonally dominant.

We have

1.

$$a_{i,i} = \frac{1}{meas(V_i)} \sum_{j\in\Sigma(i)} \left[\frac{meas(\gamma_{ij})}{dist(x_i, x_j)} k_{i,j} + \beta_{i,j} + |\beta_{i,j}|\right] > 0\,,$$

$$a_{i,j} = \frac{1}{meas(V_i)} \left[-\frac{meas(\gamma_{ij})}{dist(x_i, x_j)} k_{i,j} + \beta_{i,j} - |\beta_{i,j}|\right] < 0\,,$$

2.

$$a_{i,i} + \sum_{j\in\Sigma(i)} a_{i,j} = \frac{1}{meas(V_i)} \sum_{j\in\Sigma(i)} \beta_{i,j} \geq c_0 > 0\,.$$

The last inequality follows from the Proposition 4.1. □

Note that to prove Proposition 4.2 we used only that $k_{ij} > 0$, the Assumption 4.3 and (4.7c).

Now we discuss on the positive definiteness of the operator A_h and the matrix A. In Chapter II we have shown that the bilinear form, corresponding to the continuous problem (2.11) is H_0^1–elliptic. In the following proposition we establish that the discrete analog of the bilinear form inherits this property.

Proposition 4.3 *Let the Assumptions 4.1 and 4.3 be satisfied, the discrete fluxes $w_{i,j}$ and $v_{i,j}$ be defined by the formulas* (4.8) *and* (4.11)*, respectively, and the approximations $\beta_{i,j}$ fulfill the conditions* (4.7)*. Then the matrix A of* **UDS** *is a positive real matrix and there exists a constant C such that the following inequality is true:*

$$(A_h y, y) \geq C\|y\|^2_{1,\omega} \ \text{, for all } y \in D^0 = \{y,\, y_{|\gamma} = 0\}.$$

The constant C depends only on the ratio $a(x)/|\mathbf{b}(x)|$.

Proof: Let $z(x)$ and $y(x)$ be grid functions from D^0. Then

$$\begin{aligned}(A_h y, z)_S &= -\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} meas(V_i) w_{i,j} z_j + \sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} meas(V_i) v_{i,j} z_j \\ &= I + J\,.\end{aligned} \tag{4.14}$$

We transform the sums in formulae (4.14)

$$\begin{aligned} I &= -\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} meas(V_i)\left[\frac{meas(\gamma_{ij})}{meas(V_i)} k_{i,j} \frac{[y_j - y_i]}{dist(x_i,x_j)}\right] z_i \\ &= \frac{1}{2}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} \frac{meas(\gamma_{ij})}{dist(x_i,x_j)} k_{i,j}\left([y_j - y_i] z_j - [y_j - y_i] z_i\right) \\ &= \frac{1}{2}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} meas(V_i) k_{i,j} \frac{[y_j - y_i]}{dist(x_i,x_j)}\left[\left(\frac{meas(\gamma_{ij})}{meas(V_i)}\right)[z_j - z_i]\right]\,. \end{aligned}$$

Using (4.11) we rewrite J in the following way

$$\begin{aligned} J &= \frac{1}{2}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} [(\beta_{i,j} + |\beta_{i,j}|) y_i + (\beta_{i,j} - |\beta_{i,j}|) y_j] \\ &= \frac{1}{2}\sum_{x_i\in\omega}\left[\sum_{j\in\Sigma(i)} \beta_{i,j} y_i z_i + \sum_{j\in\Sigma(i)} |\beta_{i,j}|(y_i - y_j) + \sum_{j\in\Sigma(i)} \beta_{i,j} y_j z_i\right] \\ &= J_1 + J_2 + J_3\,. \end{aligned} \tag{4.15}$$

We now transform the second term in (4.15)

$$J_2 = \frac{1}{2}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} |\beta_{i,j}|(y_i - y_j) z_i = \frac{1}{4}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} |\beta_{i,j}|(y_i - y_j)(z_i - z_j)$$

and the third term in (4.15)

$$J_3 = \frac{1}{2}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)}\beta_{i,j}y_jz_i = \frac{1}{4}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)}\beta_{i,j}y_jz_i+\beta_{j,i}y_iz_j$$
$$= \frac{1}{4}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)}\beta_{i,j}(y_jz_i - y_iz_j)\,.$$

Finally we get

$$\begin{aligned}(A_hy,z) &= \frac{1}{2}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} meas(V_i)k_{i,j}\frac{[y_j-y_i]}{dist(x_i,x_j)}\left[\left(\frac{meas(\gamma_{ij})}{meas(V_i)}\right)[z_j-z_i]\right]\\ &+\frac{1}{2}\sum_{x_i\in\omega}\left(\sum_{j\in\Sigma(i)}\beta_{i,j}\right)y_iz_i+\frac{1}{4}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)}|\beta_{i,j}|(y_i-y_j)(z_i-z_j)\\ &+\frac{1}{4}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)}\beta_{i,j}(y_jz_i-y_iz_j)\,.\end{aligned}$$

Letting $z=y$ in the above formula the desired result follows using Proposition 4.1. and the FV regularity of the control volumes. □

4.1.3 Modified upwind difference scheme (MUDS)

As we will later show the **UDS** is only $O(h)$ accurate. We would like to exploit the symmetry of some special triangulations in order to obtain higher order convergence and still have a diagonally dominant matrix. Such triangulations for example are Voronoi meshes. We sketch the derivation of **MUDS** only for 2–D mesh for simplicity. We assume that $dist(x_j,x_{ij}) = dist(x_i,x_{ij})$ and that the ratio $meas(\gamma_{ij})/dist(x_i,x_j)$ is bounded by constants independent of h. Without loss of generality we suppose that this ratio is equal to one. Then we modify the upwind scheme in the following way [8], (see also [114])

$$\int_{\gamma_{ij}}(\mathbf{b},\mathbf{n})u\,ds = \frac{\beta_{i,j}+|\beta_{i,j}|}{2}u_i+\frac{\beta_{i,j}-|\beta_{i,j}|}{2}u_j+O(h) = I_1+O(h)\,,$$
$$\int_{\gamma_{ij}}(\mathbf{b},\mathbf{n})u\,ds = \frac{\beta_{i,j}}{2}u_i+\frac{\beta_{i,j}}{2}u_j+O(h^2) \qquad = I_2+O(h^2)\,,$$

$$\begin{aligned}\int_{\gamma_{ij}}(-a\nabla u+\mathbf{b}u,\mathbf{n})\,ds &= -k_{i,j}[u_{h,j}-u_{h,i}]+I_2+O(h^2)\\ &= -\left(k_{i,j}-\frac{|\beta_{i,j}|}{2}\right)[u_{h,j}-u_{h,i}]+I_1+O(h^2)\\ &= -\frac{k_{i,j}}{1+|\beta_{i,j}|/2k_{i,j}}[u_{h,j}-u_{h,i}]\\ &\quad -\left(k_{i,j}-\frac{|\beta_{i,j}|}{2}-\frac{k_{i,j}^2}{k_{i,j}+|\beta_{i,j}|/2}\right)[u_{h,j}-u_{h,i}]\\ &\quad +I_1+O(h^2)\\ &= -\frac{k_{i,j}}{1+|\beta_{i,j}|/2k_{i,j}}[u_{h,j}-u_{h,i}]\\ &\quad +\frac{\beta_{i,j}^2/4}{k_{i,j}+|\beta_{i,j}|/2}[u_{h,j}-u_{h,i}]+I_1+O(h^2)\\ &= -\frac{k_{i,j}}{1+|\beta_{i,j}|/2k_{1,i,j}}[u_{h,j}-u_{h,i}]+I_1+O(h^2).\end{aligned}$$

In the last step we have taken into account that $\beta_{i,j} = O(h)$. This heuristic formulae show that if we want to get a second-order finite difference scheme we should choose $w_{i,j}$ and $v_{i,j}$ in such a way that they satisfy the following condition:

$$w_{i,j} + v_{1,i,j} = \frac{1}{meas(V_i)} \left[-meas(\gamma_{ij}) \tilde{k}_{i,j} \frac{[u_{h,j} - u_{h,i}]}{dist(x_i, x_j)} + \frac{\beta_{i,j} + |\beta_{i,j}|}{2} u_{h,i} + + \frac{\beta_{i,j} - |\beta_{i,j}|}{2} u_{h,j} \right] .$$

We remark here that we split the scheme into two parts only for convenience of the error analysis. Then we define **MUDS** as follows: $A^{(1)}$ is the same as in **CDS** and the expressions $w_{i,j}$ in $A^{(2)}$ are defined by

$$w_{i,j} = -\frac{1}{meas(V_i)} \left[meas(\gamma_{ij}) \tilde{k}_{i,j} + dist(x_i, x_j) \frac{|\beta_{i,j}|}{2} \right] \frac{[u_{h,j} - u_{h,i}]}{dist(x_i, x_j)} , \tag{4.16}$$

where

$$\tilde{k}_{i,j} = \frac{k_{i,j}}{1 + |B_{i,j}|/2k_{1,i,j}} , \qquad \text{with } B_{i,j} = \frac{\beta_{i,j} dist(x_i, x_j)}{meas(\gamma_{ij})} . \tag{4.17}$$

Using similar argument as in Proposition 4.2 and Proposition 4.3 we can prove the following.

Proposition 4.4 *Let the Assumption 4.3 be satisfied, the discrete fluxes $w_{i,j}$ and $v_{i,j}$ be defined by the formulas* (4.16) *and* (4.9)*, respectively, and the approximations $\beta_{i,j}$ fulfill the condition* (4.7c)*. Then* **MUDS** *satisfies the discrete maximum principle and the corresponding matrix A is an* **M***–matrix.*

Proposition 4.5 *Let the Assumptions 4.1 and 4.3 be satisfied, the discrete fluxes $w_{i,j}$ and $v_{i,j}$ be defined by the formulas* (4.16) *and* (4.9)*, respectively, and the approximations $\beta_{i,j}$ fulfill the conditions* (4.7)*. Then the matrix A of the* **MUDS** *is a positive real matrix and there exists a constant C such that the following inequality is true:*

$$(A_h y, y) \geq C \|y\|^2_{1,\omega} \text{ , for all } y \in D^0 = \{y,\, y_{|\gamma} = 0\} .$$

The constant C depends only on the ratio $a(x)/|\mathbf{b}(x)|$.

4.1.4 Il'in's difference scheme (IDS)

Another approximation we derive in a similar way as in [64]

$$\int_{\gamma_{ij}} (-a\nabla u + \mathbf{b}u, \mathbf{n})\, ds \approx -meas(\gamma_{ij}) \hat{k}_{i,j} \frac{[u_{h,j} - u_{h,i}]}{dist(x_i, x_j)} + \frac{\beta_{i,j}}{2} u_{h,i} + \frac{\beta_{i,j}}{2} u_{h,j}$$

or

$$w_{i,j} = -\frac{meas(\gamma_{ij})}{meas(V_i)} \hat{k}_{i,j} \frac{[u_{h,j} - u_{h,i}]}{dist(x_i, x_j)} \tag{4.18}$$

and $v_{i,j}$ are defined as in **CDS**. We choose the coefficient $\hat{k}_{i,j}$ such that the above approximate relation is exact for $u = e^{(\mathbf{b},\mathbf{n})t/a}$ when $a(x)$ and $\mathbf{b}$ are constants and t is a variable on the line x_i, x_j. Simple computations show that $\hat{k}_{i,j}$ have to be defined through the equality

$$\hat{k}_{i,j} = B_{i,j} \coth\left(\frac{B_{i,j}}{k_{i,j}}\right) , \quad \text{where } B_{i,j} = \frac{\beta_{i,j} dist(x_i, x_j)}{2meas(\gamma_{ij})} . \tag{4.19}$$

It is easy to see that $\hat{k}_{i,j}$ are positive regardless of the sign of $(\mathbf{b}, \mathbf{n})$. From $|\coth(x)| > 1$ we have $\hat{k}_{i,j} > |B_{i,j}|$. In order to investigate the properties of the **IDS** we rewrite $w_{i,j} + v_{i,j}$ with $w_{i,j}$ defined by (4.18) in the following way:

$$w_{i,j} + v_{i,j} = -\frac{meas(\gamma_{ij})}{meas(V_i)} \left[\hat{k}_{i,j} - B_{i,j}\right] \frac{[u_{h,j} - u_{h,i}]}{dist(x_i, x_j)} + \frac{(\beta_{i,j} + |\beta_{i,j}|)}{2} u_{h,i} + \frac{(\beta_{i,j} - |\beta_{i,j}|)}{2} u_{h,j} \,.$$

Using the same technique as in previous propositions we have:

Proposition 4.6 *Let the Assumption 4.3 be satisfied, the discrete fluxes $w_{i,j}$ and $v_{i,j}$ be defined by the formulas* (4.18) *and* (4.9)*, respectively, and the approximations $\beta_{i,j}$ fulfill the condition* (4.7c)*. Then* **IDS** *satisfies the discrete maximum principle and the corresponding matrix A is an* **M**–*matrix.*

Proposition 4.7 *Let the Assumptions 4.1 and 4.3 be satisfied, the discrete fluxes $w_{i,j}$ and $v_{i,j}$ be defined by the formulas* (4.18) *and* (4.9)*, respectively, and the approximations $\beta_{i,j}$ fulfill the conditions* (4.7)*. Then the matrix A of the* **IDS** *is a positive real matrix and there exists a constant C such that the following inequality is true:*

$$(A_h y, y) \geq C \|y\|^2_{1,\omega} \text{ , for all } y \in D^0 = \{y,\, y_{|\gamma} = 0\} \,.$$

The constant C depends only on the ratio $a(x)/|\mathbf{b}(x)|$.

Remark 4.2 If $\beta_0 = 0$ the **UDS**, **MUDS** and **IDS** does not satisfy the discrete maximum principle, but for sufficiently small h the considered finite difference operators are coercive. This means that the error estimates which we prove in the next sections hold in this case with one more restriction.

Summarizing these approximations we formulate the following discrete problem for (2.11): Find a grid function $u_h(x)$, which satisfies the finite difference equations:

$$\sum_{j \in \Sigma(i)} w_{i,j} + \sum_{j \in \Sigma(i)} v_{i,j} = \phi_i \quad \text{in} \quad \omega \,,\ i = 1, \ldots n_S \,,$$
$$u_h(\mathrm{x}) = 0 \quad \text{on} \quad \Gamma \,,$$

where $w_{i,j}$, $v_{i,j}$ are defined by (4.8), (4.16), (4.18), (4.9) and (4.11), respectively, and $\phi_i = \frac{1}{meas(V_i)} \int_{V_i} f(\mathrm{x})\, d\mathrm{x}$. These schemes can be written as systems of linear algebraic equations

$$A\mathbf{u}_h = \phi \,. \tag{4.20}$$

4.2 Stability and error analysis

The stability of problem (4.20) is a simple consequence of the positive definiteness of the matrix A. Namely, we prove the following lemma.

Lemma 4.1 *Let the Assumptions 4.1 and 4.3 be satisfied. Then for all considered difference schemes the following a priori estimate is valid:*

$$\|u_h\|_{1,\omega} \leq C \|\phi\|_{-1,\omega} \,,$$

where u_h is the discrete solution and ϕ is the right-hand side of (4.20). The constant C in this estimate does not depend on h or ϕ.

Proof: The proof follows from the inequalities based on the the coercivity of the operator A and on the definition of the norm $||.||_{-1,\omega}$

$$\|u_h\|^2_{1,\omega} \leq C(A_h\, u_h, u_h) = C(\phi, u_h)_S \leq C\|\phi\|_{-1,\omega}\|u_h\|_{1,\omega}\,.$$

□

Remark 4.3 Since $\|\phi\|_{-1,\omega} \leq \|\phi\|_{0,\omega}$ and $\|u_h\|_{0,\omega} \leq \|u_h\|_{1,\omega}$ we also can obtain the following estimate:

$$\|u_h\|_{0,\omega} \leq C\|\phi\|_{0,\omega}\,,.$$

4.2.1 Error estimates in discrete H^1–norm

The error analysis presented here is done in the general framework of the methods developed in [115] and [43]. We consider only the case when $a(x) \equiv 1$. Let

$$z(x) = u_h(x) - u(x)\,,\ x \in \omega$$

be the error of the finite difference method. Substituting $u_h = z + u$ in (4.20) we obtain

$$Az = \phi - Au \equiv \psi\,. \tag{4.21}$$

Then using (4.4)–(4.20) we transform ψ in the following form

$$\sum_{j\in\Sigma(i)}\left[\frac{1}{meas(V_i)}\int_{\gamma_{ij}}(\mathbf{W},\mathbf{n})\,ds - w_{i,j}\right] + \sum_{j\in\Sigma(i)}\left[\frac{1}{meas(V_i)}\int_{\gamma_{ij}}(\mathbf{V},\mathbf{n})\,ds - v_{i,j}\right] \equiv \psi_{1,i} + \psi_{2,i} = \psi_i\,.$$

We define the local truncation error in the following way:

$$\eta_{i,j} = \frac{1}{meas(\gamma_{ij})}\int_{\gamma_{ij}}(\mathbf{W},\mathbf{n})\,ds - \frac{meas(V_i)}{meas(\gamma_{ij})}w_{i,j}\,, \tag{4.22a}$$

$$\mu_{i,j} = \frac{1}{meas(\gamma_{ij})}\int_{\gamma_{ij}}(\mathbf{V},\mathbf{n})\,ds - \frac{meas(V_i)}{meas(\gamma_{ij})}v_{i,j}\,. \tag{4.22b}$$

First we consider the term $(\phi_2, z)_S$. By the definition of the discrete inner product and $\phi_{2,i}$ we have

$$\begin{aligned}(\phi_2, z)_S &= \sum_{x_i\in\omega} meas(V_i)\phi_{2,i}z_i \\ &= \sum_{x_i\in\omega}\sum_{j\in\Sigma(i)}\left[\int_{\gamma_{ij}}(\mathbf{W},\mathbf{n})\,ds + meas(\gamma_{ij})k_{ij}\frac{[u_j - u_i]}{dist(x_i,x_j)}\right]z_i\,.\end{aligned}$$

We can regroup the terms (we call this nonuniform summation by parts) to get

$$\begin{aligned}
(\phi_2, z)_S &= \frac{1}{2}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)}\left\{\left[\int_{\gamma_{ij}}(\mathbf{W},\mathbf{n}) + meas(\gamma_{ij})k_{ij}\frac{[u_j-u_i]}{dist(x_i,x_j)}\right]z_i\right. \\
&\qquad \left. + \left[\int_{\gamma_{ji}}(\mathbf{W},\mathbf{n}) + meas(\gamma_{ji})k_{ji}\frac{[u_i-u_j]}{dist(x_i,x_j)}\right]z_j\right\} \\
&= -\frac{1}{2}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)}\left[\int_{\gamma_{ij}}(\mathbf{W},\mathbf{n}) + meas(\gamma_{ij})k_{ij}\frac{[u_j-u_i]}{dist(x_i,x_j)}\right][z_j-z_i] \\
&= -\frac{1}{2}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} dist(x_i,x_j)meas(\gamma_{ij}) \\
&\qquad \left[\frac{1}{meas(\gamma_{ij})}\int_{\gamma_{ij}}(\mathbf{W},\mathbf{n}) - \frac{meas(V_i)}{meas(\gamma_{ij})}w_{i,j}\right]\frac{[z_j-z_i]}{dist(x_i,x_j)} \\
&= -\frac{1}{2}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} dist(x_i,x_j)meas(\gamma_{ij})\eta_{i,j}\frac{[z_j-z_i]}{dist(x_i,x_j)}\,.
\end{aligned}$$

By the FV regularity of the grid and Cauchy–Schwartz inequality follows

$$\begin{aligned}
(\phi_2, z)_S &\le \left(\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} meas(V_i)\eta_{i,j}^2\right)^{1/2}\left(\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} meas(V_i)\left[\frac{[z_j-z_i]}{dist(x_i,x_j)}\right]^2\right)^{1/2} \\
&\le \|\eta\|_{*,\omega}\|z\|_{1,\omega}\,.
\end{aligned}$$

Here for convenience we denote with $\|\eta\|_{*,\omega}$ the first sum above.

Likewise

$$\begin{aligned}
(\phi_1, z)_S &= \sum_{x_i\in\omega} meas(V_i)\phi_{2,i}z_i \\
&= \sum_{x_i\in\omega}\sum_{j\in\Sigma(i)}\left[\int_{\gamma_{ij}}(\mathbf{W},\mathbf{n})\,ds - \left(\frac{\beta_{i,j}+|\beta_{i,j}}{2}u_i + \frac{\beta_{i,j}-|\beta_{i,j}}{2}u_j\right)\right]z_i \\
&= -\frac{1}{2}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} dist(x_i,x_j)meas(\gamma_{ij}) \\
&\qquad \left[\frac{1}{meas(\gamma_{ij})}\int_{\gamma_{ij}}(\mathbf{V},\mathbf{n}) - \frac{meas(V_i)}{meas(\gamma_{ij})}v_{i,j}\right]\frac{[z_j-z_i]}{dist(x_i,x_j)} \\
&= -\frac{1}{2}\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} dist(x_i,x_j)meas(\gamma_{ij})\mu_{i,j}\frac{[z_j-z_i]}{dist(x_i,x_j)} \\
&\le \left(\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} meas(V_i)\mu_{i,j}^2\right)^{1/2}\left(\sum_{x_i\in\omega}\sum_{j\in\Sigma(i)} meas(V_i)\left[\frac{[z_j-z_i]}{dist(x_i,x_j)}\right]^2\right)^{1/2} \\
&\le \|\mu\|_{*,\omega}\|z\|_{1,\omega}\,.
\end{aligned}$$

Summarizing these results and using Propositions 4.3, 4.5, 4.7 we obtain the following main result.

Lemma 4.2 *Let the Assumptions 4.1 and 4.3 be satisfied. The error* $z(x) = u_h(x) - u(x)\,, x \in \omega$ *of all considered finite difference schemes satisfies the a priori estimate*

$$\|z\|_{1,\omega} \le C\left(\|\eta\|_{*,\omega} + \|\mu\|_{*,\omega}\right) \tag{4.23}$$

where the components $\eta_{i,j}$ and $\mu_{i,j}$ of the local truncation error are defined by (4.22) with approximate fluxes $w_{i,j}$ and $v_{i,j}$ determined by (4.8), (4.11), (4.16) and (4.18) for the **UDS**, **MUDS** *and* **IDS**, *correspondingly. The constant C in this estimate does not depend on h or z.*

In order to use the estimate (4.23) of Lemma 4.2 we have to bound the corresponding norms of the local truncation error components $\eta_{i,j}$ and $\mu_{i,j}$ defined by (4.22). These estimates are provided in the lemmas given below.

Consider one fixed face γ_{ij} and the prism e_{ij} with two faces through x_i and other faces are parallel to the straight line (x_i, x_j) and go through the boundary of γ_{ij}. Note that γ_{ij} is a convex polygon by construction.

Lemma 4.3 *Let the solution of the problem (2.11) be H^s-regular, $\frac{3}{2} < s$, and the component of the local truncation error $\eta_{i,j}$ be defined by (4.22a) with the approximate flux $w_{i,j}$ determined by (4.8), (4.16) and (4.18). Then the following estimate holds:*

$$|\eta_{i,j}| \le Ch^{s-d/2-1}|u|_{s,e_{ij}}\,, \quad \frac{3}{2} < s \le 2 + sym, \tag{4.24}$$

where $sym = 0$ for a general triangulation, and equals 1 if the symmetry assumption is satisfied.

Proof: We transform e_{ij} into a $\bar{e}_{ij}$ with a linear transformation $(x_1, x_2, x_3) \to (\xi_1, \xi_2, \xi_3)$ such that x_j is mapped into $(0,0,0)$, x_i into $(-1,0,0)$ and $meas(\bar{\gamma}_{ij}) = 1$. Let $u(x_1, x_2, x_3) = \bar{u}(\xi_1, \xi_2, \xi_3)$. Consider first the component $\eta_{i,j}(u)$ for the **UDS**. Denote $h_1 = dist(x_i, x_j)$. Then

$$\begin{aligned}
\eta_{i,j}(u) &= -\frac{meas(V_i)}{meas(\gamma_{ij})} w_{i,j} + \frac{1}{meas(\gamma_{ij})} \int_{\gamma_{ij}} (\mathbf{W}, \mathbf{n})\, ds \\
&= \frac{[u_j - u_i]}{dist(x_i, x_j)} - \frac{1}{meas(\gamma_{ij})} \int_{\gamma_{ij}} \frac{\partial u}{\partial n}\, ds \\
\eta_{i,j}(u) = \eta_{i,j}(\bar{u}) &= \frac{1}{h_1} \left[\bar{u}_j - \bar{u}_i - \int_{\bar{\gamma}_{ij}} \frac{\partial \bar{u}}{\partial n}(\bar{x}_{ij}, \xi_2, \xi_3)\, d\xi_2 d\xi_3 \right] .
\end{aligned}$$

Using the imbedding of Sobolev spaces $H^s(\Omega) \hookrightarrow C^{0,\alpha}(\Omega)$, $\alpha = s - d/2 > 0$ (cf. (2.1c)) and $H^s(\Omega) \hookrightarrow H^\chi(\gamma_{ij})$, $\chi = s - 1/2 > 1$ (cf. (2.2a)) we have

$$|\eta_{i,j}(\bar{u})| \le \frac{1}{h_1} \left[2|\bar{u}|_{C^{0,\alpha}(\bar{e}_{ij})} + |\bar{u}|_{1,\gamma_{ij}} \right] \le \frac{C}{h_1} \|\bar{u}\|_{s,\bar{e}_{ij}} .$$

It is easy to check that $\eta_{i,j}(\bar{u})$ vanishes if $\bar{u}$ is a polynomial of first degree. Therefore, by the Bramble-Hilbert lemma (Theorem 2.7) we get that

$$|\eta_{i,j}(\bar{u})| \le \frac{C}{h_1} |\bar{u}|_{s,\bar{e}_{ij}}\,, \quad \frac{3}{2} < s \le 2$$

and with inverse transformation and the inequality (3.4a) we get

$$|u|_{s,\bar{e}_{ij}} \le Ch^{s-d/2} |u|_{s,e_{ij}}\,, \quad \frac{3}{2} < s \le 2\,.$$

Therefore,

$$|\eta_{i,j}(u)| \le Ch^{s-d/2-1} |u|_{s,e_{ij}}\,, \quad \frac{3}{2} < s \le 2\,. \tag{4.25}$$

Suppose that the symmetry assumption is satisfied. Then $\eta_{i,j}(\bar{u})$ vanishes if $\bar{u}$ is a polynomial of second degree. In this case the estimate (4.25) holds for $\frac{3}{2} < s \le 3$.

Now we consider $\eta_{i,j}$ for the **MUDS**. By construction

$$\begin{aligned}\eta_{i,j}(u) &= \left[\frac{1}{1+|B_{i,j}|/2} + \frac{|B_{i,j}|}{2}\right] \cdot \frac{[u_j - u_i]}{dist(x_i, x_j)} - \frac{1}{meas(\gamma_{ij})} \int_{\gamma_{ij}} \frac{\partial u}{\partial n}\, ds \\ &= \left\{ \frac{[u_j - u_i]}{dist(x_i, x_j)} - \frac{1}{meas(\gamma_{ij})} \int_{\gamma_{ij}} \frac{\partial u}{\partial n}\, ds \right\} + \left[C_1(x)(dist(x_i, x_j))^2 \frac{[u_j - u_i]}{dist(x_i, x_j)} \right]\end{aligned}$$

since

$$\left(\frac{1}{1+|B_{i,j}|/2} + \frac{|B_{i,j}|}{2}\right) = 1 + C_1(x)(dist(x_i, x_j))^2\,, \quad C_1(x) \sim |\mathbf{b}(\mathrm{x})|^2.$$

Here $B_{i,j}$ is defined via the formulae (4.17).

We consider $[u_j - u_i]/h_1$ as a linear functional of u. With the same argument as for the **UDS** we show

$$\left|\frac{u_j - u_i}{h_1}\right| \le \frac{2}{h_1} |u|_{C^{0,\alpha}(\bar{e}_{ij})} \le \frac{C}{h_1} \|u\|_{s,\bar{e}_{ij}}\,, \qquad \frac{d}{2} < s.$$

This functional vanishes for constants. Therefore, if follows from the modified Bramble–Hilbert lemma (Theorem 2.8) that

$$dist(x_i, x_j)^2 \left|\frac{u_j - u_i}{h_1}\right| \le C h^{3-d/2-1} \|u\|_{s,e_{ij}}\,, \qquad \frac{3}{2} < s\,.$$

Hence the estimate (4.25) is valid in this case as well and if the symmetry assumption is satisfied s can reach 3.

Finally for the **IDS** the result follows from the fact

$$B_{i,j} coth(B_{i,j}) = 1 + \tilde{C}_1(x) dist(x_i, x_j)^2\,, \quad \tilde{C}_1(x) \sim |\mathbf{b}(\mathrm{x})|^2,$$

and the same reasons as in the case for the **MUDS**. Here $B_{i,j}$ is defined by (4.19). □

Lemma 4.4 *Let the solution of the problem (2.11) be H^s-regular, $\frac{3}{2} < s$, and the component of the local truncation error $\mu_{i,j}$ be defined by (4.22b) with the approximate flux $v_{i,j}$ determined by* (4.9) *and* (4.11). *Then the following estimate holds:*

$$|\mu_{i,j}| \le \begin{cases} C h^{s-d/2} \|\mathbf{b}\|_{d/2+\alpha,\infty,\Omega} \|u\|_{s,e_{ij}} & \text{for } \mathbf{MUDS} \text{ and } \mathbf{IDS}, \\ C h^{1-d/2} \left[|\mathbf{b}|_{0,\infty,\Omega} |u|_{1,e_{ij}} + h^{s-1} \|\mathbf{b}\|_{d/2+\alpha,\infty,\Omega} \|u\|_{s,e_{ij}} \right] & \text{for } \mathbf{UDS}, \end{cases} \tag{4.26}$$

where $\frac{d}{2} < s \le 2$.

Proof: We consider two cases. For **UDS** suppose $\beta_{i,j} > 0$. Then

$$\begin{aligned}\mu_{i,j}(u) &= \frac{1}{meas(\gamma_{ij})} \int_{\gamma_{ij}} (\mathbf{V}, \mathbf{n})\, ds - \frac{meas(V_i)}{meas(\gamma_{ij})} v_{i,j} \\ &= \frac{1}{meas(\gamma_{ij})} \int_{\gamma_{ij}} (\mathbf{b}, \mathbf{n}) u\, ds - \frac{1}{meas(\gamma_{ij})} \beta_{i,j} u_i.\end{aligned}$$

We again use the linear transformation of coordinates from the proof of Lemma 4.3. After the linear transformation is performed, the truncation error $\mu_{i,j}$ simplifies to

$$\mu_{i,j}(\bar{u}) = \int_{\bar{\gamma}_{ij}} (\bar{\mathbf{b}}, \mathbf{n}) \bar{u}\, ds - \bar{\beta}_{i,j} \bar{u}_i.$$

Denote $l_1(\bar{\mathbf{b}}, \bar{u}) = -\mu_{i,j}(\bar{u})$. We represent l_1 in the following way:

$$\begin{aligned} l_1(\bar{\mathbf{b}}, \bar{u}) &= \bar{\beta}_{i,j}\bar{u}_i - \int_{\bar{\gamma}_{ij}} (\bar{\mathbf{b}}, \mathbf{n})\bar{u}\, ds \\ &= \bar{\beta}_{i,j}\left[\bar{u}_i - \int_{\bar{\gamma}_{ij}} \bar{u}\, ds\right] + \left[\int_{\bar{\gamma}_{ij}} [\bar{\beta}_{i,j} - (\bar{\mathbf{b}}, \mathbf{n})]\bar{u}\, ds\right] \\ &= \bar{\beta}_{i,j}\left[\bar{u}_i - \int_{\bar{\gamma}_{ij}} \bar{u}\, ds\right] + \int_{\bar{\gamma}_{ij}} [\bar{\beta}_{i,j} - (\bar{\mathbf{b}}, \mathbf{n})][\bar{u} - \bar{u}_i]\, ds \\ &\quad + \bar{u}_i \int_{\bar{\gamma}_{ij}} [\beta_{i,j} - (\mathbf{b}, \mathbf{n})]\, ds \\ &= \bar{\beta}_{i,j} p_1(\bar{u}) + c(\bar{\mathbf{b}}, \bar{u}) + \bar{u}_i q(\mathbf{b})\,, \end{aligned}$$

where the linear functionals $p_1(\bar{u})$, $q(\mathbf{b})$ and the bilinear functional $c(\bar{\mathbf{b}}, \bar{u})$ are defined by

$$\begin{aligned} p_1(\bar{u}) &= \bar{u}_i - \int_{\bar{\gamma}_{ij}} \bar{u}\, ds, \\ c(\bar{\mathbf{b}}, \bar{u}) &= \int_{\bar{\gamma}_{ij}} [\bar{\beta}_{i,j} - (\bar{\mathbf{b}}, \mathbf{n})][\bar{u} - \bar{u}_i]\, ds, \\ q(\mathbf{b}) &= \int_{\bar{\gamma}_{ij}} [\beta_{i,j} - (\mathbf{b}, \mathbf{n})]\, ds\,. \end{aligned}$$

Therefore, using (4.7b) we have

$$|l_1(\bar{\mathbf{b}}, \bar{u})| \le |\mathbf{b}|_{0,\infty,\bar{e}_{ij}} |p_1(\bar{u})| + |c(\bar{\mathbf{b}}, \bar{u})| + |\bar{u}|_{0,\infty,\bar{e}_{ij}} |q(\mathbf{b})|\,.$$

First we consider $p_1(\bar{u})$. The Sobolev imbedding theorem (Theorem 2.4) and the trace theorem (Theorem 2.5) imply that $p_1(\bar{u})$ is bounded for $\bar{u} \in H^s(\bar{e}_{ij})$:

$$|p_1(\bar{u})| \le |\bar{u}|_{0,\infty,\bar{e}_{ij}} + |u|_{0,\bar{\gamma}_{ij}} \le C\|\bar{u}\|_{s,\bar{e}_{ij}}\,, \quad \frac{d}{2} < s\,.$$

Moreover, $p_1(.)$ vanishes for constants. By the modified Bramble–Hilbert lemma (Theorem 2.9) follows

$$|p_1(\bar{u})| \le C(|\bar{u}|_{1,\bar{e}_{ij}} + |\bar{u}|_{s,\bar{e}_{ij}})\,, \quad \frac{d}{2} < s \le 2\,.$$

The inverse transformation (cf. (3.4a)) produces the estimate

$$|p_1(u)| \le Ch^{1-d/2}(|u|_{1,e_{ij}} + h^{s-1}|\bar{u}|_{s,\bar{e}_{ij}})\,, \quad \frac{d}{2} < s \le 2.$$

Obviously $c(\bar{\mathbf{b}}, \bar{u})$ is a bilinear functional bounded for $(\bar{\mathbf{b}}, \bar{u}) \in \left(W^{1,\infty}(\bar{e}_{ij})\right)^d \times H^1(\bar{e}_{ij})$ and vanishes for $r,\ s$ polynomials of zero degree, i.e., $c(r, \bar{u}) = 0$ for $\bar{u} \in H^1(\bar{e}_{ij})$ and $c(\bar{\mathbf{b}}, s) = 0$ for $\bar{\mathbf{b}} \in (W^{1,\infty}(\bar{e}_{ij}))^d$. Then by the bilinear variant of the Bramble-Hilbert lemma (Theorem 2.9) and the inverse transformation we have

$$|c(\mathbf{b}, u)| \le Ch^{2-d/2}|\mathbf{b}|_{1,\infty,e_{ij}}|u|_{1,e_{ij}}\,.$$

And finally the linear functional is estimated by the assumption (4.7c)

$$|q(\mathbf{b})| \le Ch^{d+\alpha}\|\mathbf{b}\|_{d/2+\alpha,\infty,\Omega}.$$

Combining the estimates for $p(.)$, $c(.,.)$ and $q(.)$ we get the desired assertion.

Now we consider $\mu_{i,j}$ for **MUDS** and **IDS**

$$\begin{aligned}\mu_{i,j}(u) &= \frac{1}{meas(\gamma_{ij})}\int_{\gamma_{ij}}(\mathbf{V},\mathbf{n})\,ds - \frac{meas(V_i)}{meas(\gamma_{ij})}v_{i,j}\\ &= \frac{1}{meas(\gamma_{ij})}\int_{\gamma_{ij}}(\mathbf{b},\mathbf{n})u\,ds - \frac{1}{meas(\gamma_{ij})}\beta_{i,j}[\alpha_j u_i + \alpha_i u_j]\,,\end{aligned}$$

where

$$\alpha_j = \frac{dist(x_j,x_{ij})}{dist(x_i,x_j)}\,,\quad \alpha_i = \frac{dist(x_i,x_{ij})}{dist(x_i,x_j)}\,.$$

With the usual linear transformation of coordinates we get

$$\mu_{i,j}(\bar{u}) = \int_{\bar{\gamma}_{ij}}(\bar{\mathbf{b}},\mathbf{n})\bar{u}\,ds - \bar{\beta}_{i,j}[\alpha_j\bar{u}_i + \alpha_i\bar{u}_j].$$

Define $l_2(\bar{\mathbf{b}},\bar{u}) = -\mu_{i,j}(\bar{u})$. With the similar argument as in the case for **UDS** we rewrite l_2 is the following form:

$$l_2(\bar{\mathbf{b}},\bar{u}) = \bar{\beta}_{i,j}p_2(\bar{u}) + c(\bar{\mathbf{b}},\bar{u}) + \bar{u}_i q(\mathbf{b})\,,$$

where the linear functional $p_2(.)$ is given by the formulae

$$p_2(\bar{u}) = [\alpha_j\bar{u}_i + \alpha_j\bar{u}_j] - \int_{\bar{\gamma}_{ij}}\bar{u}\,ds$$

and $q(.)$ and $c(.,.)$ are the same as above.

$p_2(\bar{u})$ is bounded for $\bar{u}\in H^s(\bar{e}_{ij})$, $\frac{d}{2} < s$ and vanishes for all polynomials of first degree. Hence

$$|p_2(u)| \le Ch^{s-d/2}|u|_{s,e_{ij}}\,,\quad \frac{d}{2} < s \le 2.$$

□

We point out that the symmetry condition is used only to estimate the truncation error $\eta_{i,j}$ of the diffusion term.

Now we are ready to prove the main result of this subsection.

Theorem 4.1 *If the solution $u(x)$ of the problem (2.11) is H^s-regular, with $\frac{3}{2} < s \le 3$ and the Assumptions 4.1 and 4.3 are satisfied then: (i) if the Assumption 4.2 is satisfied, the* **MUDS** *and the* **IDS** *defined by (4.16), (4.9), (4.18) and (4.9) have $O(h^{s-1})$ rate of convergence in the H^1- discrete norm, and*

$$\|u_h - u\|_{1,\omega} \le Ch^{s-1}\left(1 + h^{\delta}\|\mathbf{b}\|_{d/2+\alpha,\infty,\Omega}\right)\|u\|_{s,\Omega}\,,$$

(ii) the **UDS** *defined by (4.8) and (4.11) has at most first order of convergence in the H^1- discrete norm, and*

$$\|u_h - u\|_{1,\omega} \le Ch|\mathbf{b}|_{0,\infty,\Omega}|u|_{1,\Omega} + Ch^{s-1}\left(1 + h^{\delta}\|\mathbf{b}\|_{d/2+\alpha,\infty,\Omega}\right)\|u\|_{s,\Omega}.$$

Here

$$\delta = \begin{cases} 1 & \frac{3}{2} < s \le 2,\\ 3-s & 2 \le s \le 3.\end{cases}$$

Proof: In Lemmas 4.3 and 4.4 we have proved the estimates for the components $\eta_{i,j}$ and $\mu_{i,j}$ of the local truncation error. Hence

$$\begin{aligned}\|\eta\|_{*,\omega} &= \left(\sum_{x_i\in\omega} meas(V_i)\sum_{j\in\Sigma(i)}\eta_{i,j}^2\right)^{1/2}\\ &\le C\left(\sum_{x_i\in\omega} h^d\sum_{j\in\Sigma(i)} h^{2s-d-2}|u|^2_{s,e_{ij}}\right)^{1/2}\\ &\le C_1 h^{s-1}|u|_{s,\Omega}\,, \qquad \frac{d}{2} < s \le 2 + sym\,.\end{aligned}$$

In the same way we show that

$$\|\mu\|_{*,\omega} \le Ch^s\|\mathbf{b}\|_{d/2+\alpha,\infty,\Omega}\|u\|_{s,\Omega}$$

when **MUDS** or **IDS** are used, and

$$\|\mu\|_{*,\omega} \le C\left(h^s\|\mathbf{b}\|_{d/2+\alpha,\infty,\Omega}\|u\|_{s,\Omega} + h|\mathbf{b}|_{0,\infty,\Omega}|u|_{1,\Omega}\right)$$

otherwise. This completes the proof. □

4.2.2 Error estimates in discrete L^2-norm

Here we elaborate the discrete Aubin-Nitsche "trick" for finite difference operators introduced in Section 4.1. (see also [115]). We consider only 2–D square meshes An example of such a grid is shown on Fig. 4.2. Since the issue of constructing and studying of monotone approximation to convection-diffusion operators is our main goal we disregard the differences that may occur from the approximation of the right hand side. Thus we consider the following homogeneous problem

$$\left\{\begin{array}{rcll}\mathrm{div}(-a(x)\nabla u(x) + \mathbf{b}(x)u(x)) &=& 0 & \text{in } \Omega\,,\\ u(x) &=& g(x) & \text{on } \Gamma\,,\end{array}\right. \tag{4.27}$$

where $g(x) \in L^2(\Gamma)$. In order to simplify our presentation we consider only the case $a(x) \equiv 1$. We use the fact that the mesh is aligned with the coordinate axes and define the approximate fluxes in each direction. A typical volume V is shown on Fig. 4.3. The boundary ∂V is split $\partial V = s_1^+ \cup s_2^+ \cup s_1 \cup s_2$. If this subsection we denote the approximate solution u_h with y in order to reduce the subindices. The operators $A^{(1)}$ and $A^{(2)}$ are defined by

$$\begin{aligned}A^{(2)}y &= w^+_{1,i,j} - w_{1,i,j} + w^+_{2,i,j} - w_{2,i,j}\,, \quad x\in\omega\,,\\ A^{(1)}y &= v^+_{1,i,j} - v_{1,i,j} + v^+_{2,i,j} - v_{2,i,j}\,, \quad x\in\omega\end{aligned} \tag{4.28}$$

and the approximate fluxes are given via

$$w_l^+(x) \equiv w^+_{l,i,j} = -\frac{k^+_{l,i,j}}{h}y_{x_l,i,j}\,, \qquad l = 1,\,2, \tag{4.29}$$

$$w_l(x) \equiv w_{l,i,j} = -\frac{k_{l,i,j}}{h}y_{\overline{x}_l,i,j}\,, \qquad l = 1,\,2\,,$$

and

$$v^+_{1,i,j} = B^+_{1,i,j}(y_{i+1,j} + y_{i,j})\,, \quad B^+_{1,i,j} = \frac{b_{1,i+1/2,j}}{2h},$$

$$v_{1,i,j} = B_{1,i,j}(y_{i,j} + y_{i-1,j})\,, \quad B_{1,i,j} = \frac{b_{1,i-1/2,j}}{2h},$$

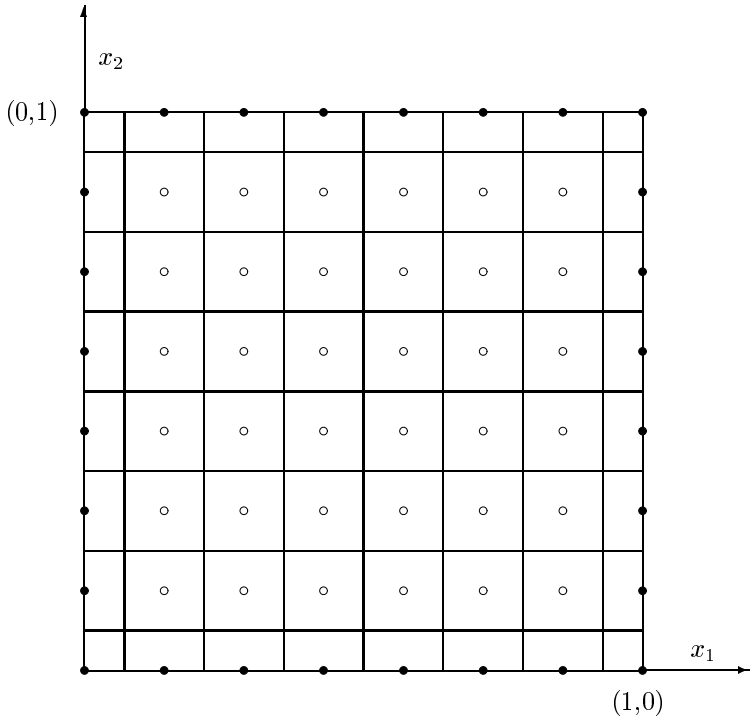


Figure 4.2: Cell-centered mesh

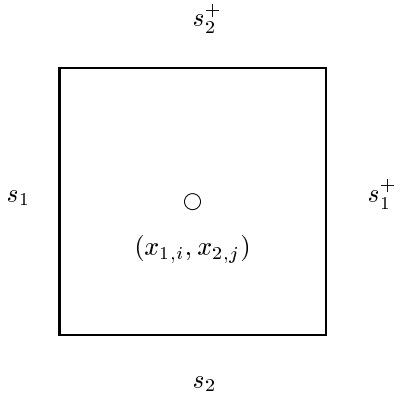


Figure 4.3: Control volume $V(x)$

$$v^+_{2,i,j} = B^+_{2,i,j}(y_{i,j+1} + y_{i,j})\,, \quad B^+_{2,i,j} = \frac{b_{2,i,j+1/2}}{2h}, \tag{4.30}$$

$$v_{2,i,j} = B_{2,i,j}(y_{i,j} + y_{i,j-1})\,, \quad B_{2,i,j} = \frac{b_{2,i,j-1/2}}{2h},$$

where

$$k_{1,i,j} = \left(\frac{1}{h}\int_{x_{1,i-1}}^{x_{1,i}} \frac{ds}{a(s,x_{2,j})}\right)^{-1}, \quad k^+_{1,i,j} = k_{1,i+1,j} \tag{4.31}$$

$$k_{2,i,j} = \left(\frac{1}{h}\int_{x_{2,j-1}}^{x_{2,j}} \frac{ds}{a(x_{1,i},s)}\right)^{-1}, \quad k^+_{2,i,j} = k_{2,i,j+1}\,.$$

For the **UDS** we define the approximate fluxes $v_{l,i,j}$ with the formulas

$$\begin{aligned} v^+_{1,i,j} &= (B^+_{1,i,j} - |B^+_{1,i,j}|)y_{i+1,j} + (B^+_{1,i,j} + |B^+_{1,i,j}|)y_{i,j}, \\ v_{1,i,j} &= (B_{1,i,j} - |B_{1,i,j}|)y_{i,j} + (B_{1,i,j} + |B_{1,i,j}|)y_{i-1,j}. \end{aligned} \tag{4.32}$$

The approximate fluxes for **MUDS** are introduced by

$$w^+_{l,i,j} = -\frac{1}{h}\left(\tilde{k}^+_{l,i,j} + |B^+_{l,i,j}h^2|\right) y_{x_l,i,j}\,, \quad l = 1,\,2\,, \tag{4.33}$$

$$w_{l,i,j} = -\frac{1}{h}\left(k_{l,i,j} + |B_{l,i,j}h^2|\right) y_{\overline{x}_l i,j}\,, \quad l = 1,\,2\,,$$

where

$$\tilde{k}_{1,i,j} = \frac{k_{1,i,j}}{1 + |B_{1,i,j}h^2|/k_{1,i,j}}\,, \quad \tilde{k}^+_{1,i,j} = \tilde{k}_{1,i+1,j}\,, \tag{4.34}$$

$$\tilde{k}_{2,i,j} = \frac{k_{2,i,j}}{1 + |B_{2,i,j}h^2|/k_{2,i,j}}\,, \quad \tilde{k}^+_{2,i,j} = \tilde{k}_{2,i,j+1}\,,$$

and for **IDS**

$$w^+_{l,i,j} = -\frac{\gamma^+_{l,i,j}}{h} y_{x_l,i,j}\,, \quad w_{l,i,j} = -\frac{\gamma_{l,i,j}}{h} y_{\overline{x}_l,i,j}\,, \quad l = 1,\,2\,, \tag{4.35}$$

$$\begin{aligned} \gamma^+_{l,i,j} &= B^+_{l,i,j}h^2 \coth\left(\frac{B^+_{l,i,j}h^2}{k^+_{l,i,j}}\right), \\ \gamma_{l,i,j} &= B_{l,i,j}h^2 \coth\left(\frac{B_{l,i,j}h^2}{k_{l,i,j}}\right). \end{aligned} \tag{4.36}$$

First, we introduce the following averaging operators [115]:

$$S_i u = \frac{1}{h}\int_{x_{i-h/2}}^{x_{i+h/2}} u(x_1,\ldots,\xi_i,\ldots,x_n)\,d\xi_i\,,$$

$$S^+_i u = \frac{1}{h}\int_{x_i}^{x_i+h} u(x_1,\ldots,\xi_i,\ldots,x_n)\,d\xi_i\,,$$

$$S^-_i u = \frac{1}{h}\int_{x_i-h}^{x_i} u(x_1,\ldots,\xi_i,\ldots,x_n)\,d\xi_i\,,$$

$$T_i = S_i^2 = S^+_i S^-_i\,,\ T = T_1 T_2\,.$$

Then applying T to the differential equation (4.27) at any grid point $x \in \omega$ and using the properties:

$$T_i\left(\frac{\partial^2 u}{\partial x_i^2}\right)(x) = u_{\overline{x}_i, x_i}\,, \quad S_i^+\left(\frac{\partial u}{\partial x_i}\right)(x) = u_{x_i}$$

we get

$$-(T_2 u)_{\overline{x}_1, x_1} - (T_1 u)_{\overline{x}_2, x_2} + T_2 S_1^-(b_1 u)_{x_1} + T_1 S_2^-(b_2 u)_{x_2} = 0\,. \tag{4.37}$$

We express the operator A_h in the form

$$\begin{aligned} h w_{1,x_1} + h w_{2,x_2} + A^{(1)} y &= 0\,, \quad x \in \omega\,, && (4.38)\\ y(x) &= T_{3-l} g(x)\,,\ x \in \gamma_l^{\pm},\ l = 1,2\,. && (4.39) \end{aligned}$$

Let $z(x) = y(x) - u(x)\,,\ x \in \overline{\omega}_S$ be the error of the finite difference method. We define

$$\overline{u} = \begin{cases} u(x)\,, & x \in \omega \\ T_{3-l} g(x)\,, & x \in \gamma_l^{\pm},\ l = 1,2\,. \end{cases}$$

Then $z = (y - \overline{u}) + (\overline{u} - u) = \overline{z} + \overline{u} - u$. Note that $\overline{z} = 0$ on γ. Substituting $y = \overline{z} + \overline{u}$ in (4.38) we obtain

$$A_h \overline{z} = A_h y - A_h \overline{u}\,. \tag{4.40}$$

The right-hand side of (4.40) is the local truncation error. In order to obtain a priori estimate we represent the local truncation error in a divergence or almost divergence form (depending upon the choice of the difference scheme). Next, we rewrite (4.40) as

$$\begin{aligned} A_h \overline{z} &= \sum_{l=1}^{2} \left[h w_l + (T_{3-l} u)_{\overline{x}_l}\right]_{x_l} + \sum_{l=1}^{2} \left[h v_l - T_{3-l} S_l^-(b_l u)\right]_{x_l} \\ &= \sum_{l=1}^{2} (T_{3-l} u - u)_{\overline{x}_l x_l} + \sum_{l=1}^{2} \left[h v_l - T_{3-l} S_l^-(b_l u)\right]_{x_l} \\ &\quad + \sum_{l=1}^{2} \left[-(k_l - 1) u_{\overline{x}_l}\right]_{x_l}\,. \end{aligned}$$

Finally, we find the expression for the local truncation error

$$A_h \overline{z} = \eta_{1,\overline{x}_1 x_1} + \eta_{2,\overline{x}_2 x_2} + \mu_{1,x_1} + \mu_{2,x_2} + \xi_{1,x_1} + \xi_{2,x_2} \tag{4.41}$$

where

$$\eta_l = T_{3-l} u - \overline{u}\,,\ x \in \omega_l^{\pm}\,, \tag{4.42}$$

$$\mu_l = h \overline{v}_l - T_{3-l} S_l^-(b_l u)\,, \quad \xi_l = -(k_l - 1)\overline{u}_{\overline{x}_l}\,,\ x \in \omega_l^+\,, \tag{4.43}$$

where $\overline{v}$ is gotten by replacing u with $\overline{u}$ in the formulas (4.30) and (4.32).

Let us introduce the solution of the following auxiliary discrete problem

$$\begin{aligned} A_h^T w &= \overline{z} \quad \text{in } \omega_S\,, \\ w &= 0 \quad \text{on } \gamma\,. \end{aligned} \tag{4.44}$$

Note that similarly to the Aubin-Nitsche "trick" w is a solution of a discrete second order problem with a right-hand side the error $\overline{z}(x)$ of the method. Obviously,

$$(A_h\overline{z}, w) = (A_h^T w, \overline{z})_S = (\overline{z}, \overline{z})_S = \|\overline{z}\|_{0,\omega}^2 . \tag{4.45}$$

On the other hand from (4.41) we get

$$\begin{aligned}
(A_h\overline{z}, w)_S &= \sum_{l=1}^{2} [(\eta_{l,\overline{x}_l x_l}, w)_S + (\mu_{l,x_l}, w)_S + (\xi_{l,x_l}, w)_S] \\
&= \sum_{l=1}^{2} (\eta_l, w_{\overline{x}_l x_l})_S - \sum_{l=1}^{2} \{(\mu_l, w_{\overline{x}_l}]_l + (\xi_l, w_{\overline{x}_l}]_l\} \qquad (4.46)\\
&\le \sum_{l=1}^{2} (\|\eta_l\|_{0,\omega} + \|\mu_l\rceil|_l + \|\xi_l\rceil|_l)(\|w_{\overline{x}_l x_l}\|_{0,\omega} + \|w_{\overline{x}_l}\rceil|_l) .
\end{aligned}$$

To complete the proof of the a priori estimate we need the following lemma.

Lemma 4.5 *Let* $\mathbf{b}(x) \in \left(W^{1,\infty}(\Omega)\right)^2$. *Then for the error* $\overline{z}(x) = y(x) - \overline{u}(x)$, $x \in \omega$ *of all considered schemes and the solution* w *of the problem (4.44) the inequalities are valid:*

$$\|w\|_{2,\omega} \le C_1 \|A_h^{(2)} w\|_{0,\omega} \le C_2 \|\overline{z}\|_{0,\omega} \tag{4.47}$$

for sufficiently small h.

Proof: Using the definition of $A_h^{(2)}$ and the triangle inequality we get

$$\begin{aligned}
\|A_h^{(2)} w\|_{0,\omega} &= \|[k_1 w_{\overline{x}_1}]_{x_1} + [k_2 w_{\overline{x}_2}]_{x_2}\|_{0,\omega} \\
&= \|[(1 + C_1(x)h^2) w_{\overline{x}_1}]_{x_1} + [(1 + C_2(x)) w_{\overline{x}_2}]_{x_2}\|_{0,\omega} \\
&\ge \|w_{\overline{x}_1 x_1} + w_{\overline{x}_2 x_2}\|_{0,\omega} \\
&\quad -h^2 \|C_{1,x_1} w_{x_1} + C_1 w_{\overline{x}_1 x_1} + C_{2,x_2} w_{x_2} + C_2 w_{\overline{x}_2 x_2}\|_{0,\omega} \\
&\ge \|w_{\overline{x}_1 x_1} + w_{\overline{x}_2 x_2}\|_{0,\omega} - D_2 h^2 \|w\|_{2,\omega} .
\end{aligned}$$

Here $k_l = 1, C_l = 0$, $l = 1, 2$ for the **UDS** and

$$k_l = 1 + C_l(x)h^2 , \quad C_l(x) \sim b_l^2(x) , \ l = 1, 2$$

otherwise. We use also that C_1, C_2 and C_{1,x_1}, C_{2,x_2} are bounded.

Finally using the equivalence of $\|w_{\overline{x}_1 x_1} + w_{\overline{x}_2 x_2}\|_{0,\omega}$ and $\|w\|_{2,\omega}$ in the space D^0 we obtain

$$\|A_h^{(2)}\|_{0,\omega} \ge \left(D_1 - D_2 h^2\right) \|w\|_{2,\omega} ,$$

where D_1 and D_2 are positive constants. Hence for sufficiently small h the lower bound in (4.47) is proved.

An upper estimate for $\|A_h^{(2)}\|_{0,\omega}$ is derived by using the standard a priori estimate in $W_2^1(\omega)$, $\|w\|_{1,\omega} \le C\|\overline{z}\|_{0,\omega}$. Then

$$\begin{aligned}
\|A_h^{(2)} w\|_{0,\omega} &= \|A_h^{(2)^T} w\|_{0,\omega} = \|(A_h - A_h^{(1)})^T w\|_{0,\omega} \\
&\le \|A_h^T w\|_{0,\omega} + \|A_h^{(1)^T} w\|_{0,\omega} \\
&\le \|\overline{z}\|_{0,\omega} + C\|w\|_{1,\omega} \\
&\le C\|\overline{z}\|_{0,\omega} .
\end{aligned}$$

□

Remark 4.4 Lemma 4.5 is actually a discrete regularity result in $W^{2,2}(\omega)$ (cf., Hackbusch [54])

$$\|w\|_{2,\omega} \le C\|\overline{z}\|_{0,\omega}\,.$$

Then (4.45) and (4.46) yield

$$\|\overline{z}\|^2_{0,\omega} = (A_h\overline{z}, w) \le C\sum_{l=1}^{2}(\|\eta_l\|_{0,\omega} + \|\mu_l\rceil|_l + \|\xi_l\rceil|_l)\|\overline{z}\|_{0,\omega}.$$

Thus, we have proved the following a priori estimate

Lemma 4.6 *The error $\overline{z}(x) = y(x) - \overline{u}(x)$, $x \in \omega$ of all considered finite difference schemes satisfies the a priori estimate:*

$$\|\overline{z}\|_{0,\omega} \le C\sum_{l=1}^{2}\left(\|\eta_l\|_{0,\omega} + \|\mu_l\rceil|_l + \|\xi_l\rceil|_l\right),$$

where the components η_l, μ_l and ξ_l, $l = 1, 2$ of the local truncation error are defined by (4.42) and (4.43). The constant C does not depend on h or $\overline{z}$.

Now we are ready to prove the following basic lemma.

Lemma 4.7 *If the solution u of the problem (4.27) with constant coefficient $a(x)$ is $H^s(\Omega)$–regular, $1 < s \le 2$ then the components of the local truncation error η_l and μ_l, $l = 1, 2$, defined by (4.42) and (4.43), respectively, satisfy the following estimates:*

(i)

$$\|\eta_l\|_{0,\omega} \le Ch^s\|u\|_{s,\Omega}\,,$$

(ii)

$$\|\mu_l\rceil|_l \le \begin{cases} Ch^s\|b_l\|_{1,\infty,\Omega}\|u\|_{s,\Omega} & \text{for } \textbf{MUDS} \text{ and } \textbf{IDS} \\ C\left(h|b_l|_{0,\infty,\Omega}|u|_{1,\Omega} + h^s\|b_l\|_{1,\infty,\Omega}\|u\|_{s,\Omega}\right) & \text{for } \textbf{UDS}, \end{cases}$$

(iii)

$$\|\xi_l\rceil|_l \le \begin{cases} Ch^2\|u\|_{s,\Omega} & \text{for } \textbf{MUDS} \text{ and } \textbf{IDS} \\ 0 & \text{for } \textbf{UDS}. \end{cases}$$

Proof: Consider $e_{i,j} = \{(x_1, x_2) : x_{1,i-1} \le x_1 \le x_{1,i+1}\,,\ x_{2,j-1} \le x_2 \le x_{2,j+1}\}$. We begin with **UDS**. To obtain (i) we rewrite (4.42) in the form

$$\eta_1 = u(x_{1,i}, x_{2,j}) - \int_{-1}^{1}(1 - |s_2|)u(x_{1,i}, x_{2,j} + s_2h)\,ds_2.$$

It suffices to prove the estimate for $x \in \omega$ because by construction $\eta_l = 0$ on $\omega_l^{\pm}$. We have that η_1 is a linear functional of $u(x)$, bounded for $u \in H^s(\Omega)$, $1 < s \le 2$. This functional vanishes for all polynomial of first degree. Therefore, by a Bramble-Hilbert lemma argument we get

$$|\eta_1(x)| \le Ch^{s-1}|u|_{s,e}\,,\ 1 < s \le 2. \tag{4.48}$$

$$\|\eta_1\|_{0,\omega} = \left(\sum_{x\in\omega}\eta_1^2(x)h^2\right)^{\frac{1}{2}} \le Ch^s|u|_{s,e}.$$

We note that in this case $\xi_1(x) \equiv 0$. Now, let us take the component $\eta_1(x)$ for the **MUDS** and the **IDS**. In both schemes the coefficients $\tilde{k}_1(x)$ and $\gamma_1(x)$ are perturbations of the coefficient $k_1(x) \equiv 1$ of the **UDS** with a term of order $O(h^2)$. More precisely,

$$\tilde{k}_1(x) = \frac{1}{1+|b_1(x)h/2|} + \frac{|b_1(x)|h}{2} = 1 + C_1 h^2 \qquad (\textbf{MUDS})$$

and

$$\gamma_1(x) = \frac{b_1(x)h}{2} \coth\left(\frac{b_1(x)h}{2}\right) = 1 + \tilde{C}_1 h^2 \qquad (\textbf{IDS}).$$

Since

$$\xi_1(x) = -(k_1(x)-1)\overline{u}_{\overline{x}_1} = -C_1 h^2 \frac{[\overline{u}_{i,j} - \overline{u}_{i,j-1}]}{h}$$

we have

$$|\xi(x)| \le Ch(|u|_{1,e} + h^{s-1}|u|_{s,e}), \ 1 < s \le 2$$

for the interior points and hence

$$\|\xi_1]|_1 \le Ch^2 \|u\|_{s,\Omega}, 1 < s \le 2.$$

For the boundary points we have $\xi_1(x) = -Ch[u_{i,,j} - u_{i-1,j}] + Ch[u_{i,j} - \overline{u}_{i,j}]$ and the second term is estimated with the approach used in the proof of Theorem 4.1.

(ii) For the second component $\mu_1(x)$ we proceed in the same way as in Lemma 3.3. First, we need the equality (see [115]):

$$T_2 S_1^-(b_1 u)(x_{1,i}, x_{2,j}) = \int_{-1}^{1} (1-|s_2|) \left[\int_{-1}^{0} b_1(x_{1,i}+s_1 h, x_{2,j}+s_2 h) u(x_{1,i}+s_1 h, x_{2,j}+s_2 h)\, ds_1\right] ds_2.$$

Now, let us consider the component for the **MUDS** and **IDS**

$$\mu_1(x) = \frac{b_{1,i-1/2,j}}{2}[\overline{u}_{i,j} + \overline{u}_{i-1,j}] - T_2 S_1^-(b_1 u)(x_{1,i}, x_{2,j}).$$

We can represent μ_1 in the following way

$$\mu_1(x) = b_{1,i-1/2,j} p(u) - c(b_1, u) + u_{i,j} q(b_1)$$

where

$$\begin{aligned} p(u) &= \frac{[\overline{u}_{i,j} + \overline{u}_{i-1,j}]}{2} \\ &\quad - \int_{-1}^{1} (1-|s_2|) \left[\int_{-1}^{0} u(x_{1,i}+s_1 h, x_{2,j}+s_2 h)\, ds_1\right] ds_2, \end{aligned}$$

$$\begin{aligned} c(b_1, u) &= \int_{-1}^{1} (1-|s_2|) \left[\int_{-1}^{0} u(x_{1,i}+s_1 h, x_{2,j}+s_2 h) - \overline{u}_{i,j}\right] \\ &\quad \times \left[b_1(x_{1,i}+s_1 h, x_{2,j}+s_2 h) - b_{1,i-1/2,j}\right] ds_1\, ds_2 \end{aligned}$$

and

$$\begin{aligned} q(b_1) &= b_{1,i-1/2,j} \\ &\quad - \int_{-1}^{1} (1-|s_2|) \left[\int_{-1}^{0} b_1(x_{1,i}+s_1 h, x_{2,j}+s_2 h) - b_{1,i-1/2,j}\, ds_1\right] ds_2. \end{aligned}$$

We have the estimates:

$$\begin{aligned} |p(u)| &\leq Ch^{s-1}|u|_{s,e}\,,\ 1<s\leq 2,\\ |c(b_1,u)| &\leq Ch|b_1|_{1,\infty,e}|u|_{1,e},\\ |q(u)| &\leq Ch|b_1|_{1,\infty,e}. \end{aligned}$$

Hence

$$|\mu_1(x)| \leq Ch^{s-1}\|b_1\|(|u|_{s,e}+h^{2-s}(|u|_{1,e}+|u|_{0,\infty,e})).$$

For **UDS** we have to add the error of the term $-|b_{1,i,j}|u_{\overline{x}_1}$ which is $h\left(|b_1|_{0,\infty,e}(|u|_{1,e}+h^{s-1}|u|_{s,e})\right]$.
Combining the above results we obtain the assertions of the lemma. □

Now we can prove the main result in this subsection.

Theorem 4.2 *If the solution of problem (2.11) is H^s-regular, $1<s\leq 2$ then:*

(i) the **MUDS** *and* **IDS** *defined by (4.33), (4.30), (4.35) and (4.30) have $O(h^m)$ rate of convergence in the L^2- discrete norm, i.e.,*

$$\|y-u\|_{0,\omega} \leq Ch^s[(1+\|b_1\|_{1,\infty,\Omega}+\|b_2\|_{1,\infty,\Omega})\,\|u\|_{s,\Omega}+\|g\|_{s-\frac{1}{2},\Gamma}]$$

(ii) the **UDS** *defined by (4.29) and (4.32) has at most first order of convergence in the L^2–discrete norm, i.e.,*

$$\begin{aligned} \|y-u\|_{0,\omega} &\leq Ch\left(|b_1|_{0,\infty,\Omega}+|b_2|_{0,\infty,\Omega}\right)|u|_{1,\Omega}\\ &\quad +Ch^s[(1+\|b_1\|_{1,\infty,\Omega}+\|b_2\|_{1,\infty,\Omega})\,\|u\|_{s,\Omega}+\|g\|_{s-\frac{1}{2},\Gamma}]. \end{aligned}$$

Proof: We have $\|y-u\|_{0,\omega}\leq\|y-\overline{u}\|_{0,\omega}+\|u-\overline{u}\|_{0,\omega}$. From Lemma 4.6 and Lemma 4.7 we get immediately the estimate for $\|y-\overline{u}\|_{0,\omega}$. To find the upper bound of the second term

$$\|u-\overline{u}\|_{0,\omega}^2=\sum_{l=1}^{2}\sum_{\gamma_l^{\pm}}h^2\left(T_{3-l}g(x)-g(x)\right)^2\,,$$

we observe that we can consider $T_{3-l}g-g$ as a linear functional of g which is bounded in $H^{m-\frac{1}{2}}(\Gamma)$ and vanishes for all polynomials of first degree. Then
$|T_{3-l}g-g|\leq Ch^{m-1}\|g\|_{m-\frac{1}{2},e_\gamma}$ where $e_\gamma=(x_l-h,x_l+h)$ which shows that
$\|u-\overline{u}\|_{0,\omega}\leq Ch^m\|g\|_{m-\frac{1}{2},\Gamma}$. □

Remark 4.5 The technique used in Subsections 4.2.1 and 4.2.2 directly gives the same estimates for the **CDS** as for **MUDS** and **IDS**, when this scheme is stable, i.e., when (4.10) holds.

4.3 Numerical results

In this section we study the error behavior of our three schemes (**UDS**, **MUDS**, and **IDS**) in both H^1 and L^2 discrete norms on model test examples.

We consider

$$\begin{cases} div(-\varepsilon\nabla u(x,y)+\mathbf{b}(x,y)u(x,y)) &= f(x,y)\,, & \text{in } \Omega\,,\\ u(x,y) &= 0\,, & \text{on } \Gamma\,, \end{cases} \tag{4.49}$$

and for velocity vector $\mathbf{b}$ we choose

$$b_1=-(1-x\cos\alpha)\cos\alpha\,,\ b_2=-(1-y\sin\alpha)\sin\alpha\,,$$

where the angle is $\alpha=15^0$.

Table 4.1: **UDS** , $\alpha = 15^0$, $d = 0$

$\epsilon \setminus N$		16	32	64	128	256
1	L^2	$0.389.10^{-3}$	$0.198.10^{-3}$	$0.100.10^{-3}$	$0.503.10^{-4}$	$0.252.10^{-4}$
	β	0.947	0.974	0.986	0.991	0.997
	H^1	$0.154.10^{-2}$	$0.859.10^{-3}$	$0.454.10^{-3}$	$0.233.10^{-3}$	$0.118.10^{-3}$
	β	0.699	0.842	0.920	0.962	0.982
10^{-2}	L^2	$0.149.10^{-1}$	$0.811.10^{-2}$	$0.425.10^{-2}$	$0.218.10^{-2}$	$0.110.10^{-2}$
	β	0.780	0.878	0.932	0.963	0.987
	H^1	$0.633.10^{-1}$	$0.462.10^{-1}$	$0.288.10^{-1}$	$0.163.10^{-1}$	$0.868.10^{-2}$
	β	0.298	0.454	0.682	0.821	0.909
10^{-5}	L^2	$0.233.10^{-1}$	$0.135.10^{-1}$	$0.737.10^{-2}$	$0.388.10^{-2}$	$0.200.10^{-2}$
	β	0.667	0.787	0.873	0.926	0.956
	H^1	$0.110.10^{0}$	$0.779.10^{-1}$	$0.505.10^{-1}$	$0.305.10^{-1}$	$0.180.10^{-1}$
	β	0.338	0.498	0.625	0.727	0.761

Problem 4.1 *$f(x, y)$ is chosen such that the solution is*

$$u(x, y) = x(1 - x)y(1 - y)e^{d(x+2y)} \text{ for } d = 0 \text{ or } d = 1.$$

In Tables 4.1-4.6 we display the error for smooth solutions without boundary layer behavior. In the first and the second rows we show the $L^2(\omega)$ and $H^1(\omega)$-norms of the error $z = y - u$ and "numerical" rate of convergence β, i.e., h^β. Our computational experiments clearly show that **MUDS** and **IDS** exhibit second order of convergence both in L^2 and H^1-norms for problems with moderate convection (i.e., not too small $\varepsilon > 0$) ; the factor β is in the range of 1.822–1.995, correspondingly. For these problems **UDS** is only first order accurate: β is between 0.947-1.260. For highly dominating convection all schemes show about first order of accuracy. The results for $\varepsilon = 10^{-2}$, 10^{-5} show that all considered schemes are stable.

Problem 4.2

$$f(x, y) = \nabla \cdot (\mathbf{b}u_0) \, , \; u_0(x, y) = x^2 y(1 - y) \, .$$

Here u_0 is the solution of the equation (4.49) when $\varepsilon = 0$. In Tables 4.7-4.9 we show $\|y - u_0\|_{0,\overline{\omega}}$, where $\overline{\omega}$ is a grid in $\overline{\Omega} = [0, 7/8] \times [0, 1]$, i.e., away from the boundary layer. This gives us a reasonable information since for small ε the function u_0 is close to the exact solution of problem 2, except within the boundary layer. In fact we have an estimate $\|u - u_0\|_{0,\overline{\Omega}} \leq C\varepsilon$, and when ε is significantly less than h we may use u_0 instead of the unknown exact solution u in $\overline{\Omega}$. In case that h and ε are of the same order this is inappropriate as is shown by Tables 4.7 - 4.9 $h = 1/256$ and $\varepsilon = 10^{-3}$. Our experiments show very weak dependence of the numerical solution with respect to $\varepsilon \to 0$ in $\overline{\Omega}$. This means that if we use a more sophisticated method near the boundary layer, e.g., local refinement, or defect–correction, in combination with the proposed schemes outside the layer, we can get better results.

Table 4.2: **MUDS** , $\alpha = 15^0$, $d = 0$

$\epsilon \setminus N$		16	32	64	128	256
1	L^2	$0.213.10^{-4}$	$0.567.10^{-5}$	$0.146.10^{-5}$	$0.372.10^{-6}$	$0.940.10^{-7}$
	β	1.822	1.909	1.957	1.973	1.985
	H^1	$0.818.10^{-4}$	$0.239.10^{-4}$	$0.649.10^{-5}$	$0.169.10^{-5}$	$0.431.10^{-6}$
	β	1.559	1.775	1.881	1.941	1.971
10^{-2}	L^2	$0.102.10^{-1}$	$0.416.10^{-2}$	$0.148.10^{-2}$	$0.468.10^{-3}$	$0.134.10^{-3}$
	β	1.100	1.294	1.491	1.661	1.804
	H^1	$0.436.10^{-1}$	$0.240.10^{-1}$	$0.101.10^{-1}$	$0.347.10^{-2}$	$0.104.10^{-2}$
	β	0.609	0.861	1.249	1.541	1.738
10^{-5}	L^2	$0.233.10^{-1}$	$0.135.10^{-1}$	$0.736.10^{-2}$	$0.387.10^{-2}$	$0.198.10^{-2}$
	β	0.667	0.787	0.875	0.927	0.967
	H^1	$0.110.10^{0}$	$0.784.10^{-1}$	$0.511.10^{-1}$	$0.309.10^{-1}$	$0.174.10^{-1}$
	β	0.338	0.489	0.618	0.728	0.820

Table 4.3: **IDS** , $\alpha = 15^0$, $d = 0$

$\epsilon \setminus N$		16	32	64	128	256
1	L^2	$0.169.10^{-4}$	$0.451.10^{-5}$	$0.116.10^{-5}$	$0.295.10^{-6}$	$0.740.10^{-7}$
	β	1.840	1.906	1.959	1.975	1.995
	H^1	$0.650.10^{-4}$	$0.189.10^{-4}$	$0.511.10^{-5}$	$0.133.10^{-5}$	$0.338.10^{-6}$
	β	1.578	1.782	1.887	1.942	1.976
10^{-2}	L^2	$0.860.10^{-2}$	$0.288.10^{-2}$	$0.816.10^{-3}$	$0.213.10^{-3}$	$0.540.10^{-4}$
	β	1.253	1.578	1.819	1.937	1.980
	H^1	$0.366.10^{-1}$	$0.166.10^{-1}$	$0.557.10^{-2}$	$0.158.10^{-2}$	$0.420.10^{-3}$
	β	0.786	1.141	1.575	1.818	1.911
10^{-5}	L^2	$0.233.10^{-1}$	$0.133.10^{-1}$	$0.736.10^{-2}$	$0.387.10^{-2}$	$0.198.10^{-2}$
	β	0.667	0.787	0.875	0.927	0.967
	H^1	$0.110.10^{0}$	$0.770.10^{-1}$	$0.511.10^{-1}$	$0.309.10^{-1}$	$0.175.10^{-1}$
	β	0.338	0.515	0.592	0.728	0.820

Table 4.4: **UDS** , $\alpha = 15^0$, $d = 1$

$\epsilon \setminus N$		16	32	64	128	256
1	L^2	$0.232.10^{-2}$	$0.102.10^{-2}$	$0.470.10^{-3}$	$0.223.10^{-3}$	$0.113.10^{-3}$
	β	1.260	1.186	1.118	1.040	0.981
	H^1	$0.930.10^{-2}$	$0.451.10^{-2}$	$0.218.10^{-2}$	$0.106.10^{-2}$	$0.545.10^{-3}$
	β	0.984	1.044	1.049	1.040	0.960
10^{-2}	L^2	$0.486.10^{-1}$	$0.267.10^{-1}$	$0.141.10^{-1}$	$0.725.10^{-2}$	$0.368.10^{-2}$
	β	0.769	0.864	0.921	0.960	0.978
	H^1	$0.228.10^{0}$	$0.170.10^{0}$	$0.110.10^{0}$	$0.637.10^{-1}$	$0.345.10^{-1}$
	β	0.291	0.423	0.628	0.788	0.885
10^{-5}	L^2	$0.719.10^{-1}$	$0.408.10^{-1}$	$0.219.10^{-1}$	$0.114.10^{-1}$	$0.585.10^{-2}$
	β	0.690	0.817	0.898	0.942	0.963
	H^1	$0.329.10^{0}$	$0.215.10^{0}$	$0.138.10^{0}$	$0.847.10^{-1}$	$0.496.10^{-1}$
	β	0.536	0.614	0.640	0.704	0.772

Table 4.5: **MUDS** , $\alpha = 15^0$, $d = 1$

$\epsilon \setminus N$		16	32	64	128	256
1	L^2	$0.768.10^{-3}$	$0.204.10^{-3}$	$0.526.10^{-4}$	$0.134.10^{-4}$	$0.337.10^{-5}$
	β	1.830	1.911	1.956	1.973	1.991
	H^1	$0.302.10^{-2}$	$0.900.10^{-3}$	$0.246.10^{-3}$	$0.646.10^{-4}$	$0.165.10^{-4}$
	β	1.531	1.747	1.871	1.929	1.969
10^{-2}	L^2	$0.291.10^{-1}$	$0.117.10^{-1}$	$0.415.10^{-2}$	$0.130.10^{-2}$	$0.374.10^{-3}$
	β	1.101	1.315	1.495	1.675	1.797
	H^1	$0.132.10^{0}$	$0.721.10^{-1}$	$0.306.10^{-1}$	$0.106.10^{-1}$	$0.319.10^{-2}$
	β	0.677	0.872	1.236	1.529	1.732
10^{-5}	L^2	$0.719.10^{-1}$	$0.407.10^{-1}$	$0.218.10^{-1}$	$0.114.10^{-1}$	$0.576.10^{-2}$
	β	0.690	0.821	0.901	0.935	0.985
	H^1	$0.329.10^{0}$	$0.216.10^{0}$	$0.138.10^{0}$	$0.840.10^{-1}$	$0.432.10^{-1}$
	β	0.536	0.607	0.646	0.716	0.822

Table 4.6: **IDS** , $\alpha = 15^0$, $d = 1$

$\epsilon \setminus N$		16	32	64	128	256
1	L^2	$0.752.10^{-3}$	$0.200.10^{-3}$	$0.515.10^{-4}$	$0.131.10^{-4}$	$0.330.10^{-5}$
	β	1.828	1.911	1.957	1.975	1.989
	H^1	$0.296.10^{-2}$	$0.883.10^{-3}$	$0.242.10^{-3}$	$0.634.10^{-4}$	$0.162.10^{-4}$
	β	1.532	1.745	1.867	1.932	1.968
10^{-2}	L^2	$0.227.10^{-1}$	$0.755.10^{-2}$	$0.212.10^{-2}$	$0.553.10^{-3}$	$0.138.10^{-3}$
	β	1.277	1.588	1.832	1.939	2.002
	H^1	$0.100.10^{0}$	$0.454.10^{-1}$	$0.153.10^{-1}$	$0.443.10^{-2}$	$0.116.10^{-2}$
	β	0.880	1.139	1.569	1.788	1.933
10^{-5}	L^2	$0.718.10^{-1}$	$0.407.10^{-1}$	$0.218.10^{-1}$	$0.114.10^{-1}$	$0.576.10^{-2}$
	β	0.690	0.819	0.901	0.935	0.985
	H^1	$0.329.10^{0}$	$0.216.10^{0}$	$0.138.10^{0}$	$0.840.10^{-1}$	$0.475.10^{-1}$
	β	0.536	0.607	0.646	0.716	0.822

Table 4.7: **UDS** , $\alpha = 15^0$, **boundary layer**

$\epsilon \setminus N$		16	32	64	128	256
10^{-3}	L^2	$0.427.10^{-2}$	$0.252.10^{-2}$	$0.147.10^{-2}$	$0.894.10^{-3}$	$0.594.10^{-3}$
	β	0.622	0.761	0.778	0.717	0.590
	H^1	$0.414.10^{-1}$	$0.276.10^{-1}$	$0.172.10^{-1}$	$0.109.10^{-1}$	$0.744.10^{-2}$
	β	0.332	0.585	0.682	0.658	0.551
10^{-4}	L^2	$0.393.10^{-2}$	$0.225.10^{-2}$	$0.122.10^{-2}$	$0.645.10^{-3}$	$0.342.10^{-3}$
	β	0.641	0.805	0.883	0.920	0.920
	H^1	$0.365.10^{-1}$	$0.233.10^{-1}$	$0.134.10^{-1}$	$0.733.10^{-2}$	$0.396.10^{-2}$
	β	0.380	0.648	0.798	0.870	0.888
10^{-5}	L^2	$0.391.10^{-2}$	$0.223.10^{-2}$	$0.119.10^{-2}$	$0.621.10^{-3}$	$0.318.10^{-3}$
	β	0.642	0.810	0.906	0.938	0.965
	H^1	$0.361.10^{-1}$	$0.229.10^{-1}$	$0.130.10^{-1}$	$0.700.10^{-2}$	$0.364.10^{-2}$
	β	0.387	0.657	0.817	0.893	0.943

Table 4.8: **MUDS** , $\alpha = 15^0$, **boundary layer**

$\epsilon \setminus N$		16	32	64	128	256
10^{-3}	L^2	$0.392.10^{-2}$	$0.224.10^{-2}$	$0.122.10^{-2}$	$0.682.10^{-3}$	$0.340.10^{-3}$
	β	0.640	0.807	0.877	0.839	0.652
	H^1	$0.364.10^{-1}$	$0.233.10^{-1}$	$0.135.10^{-1}$	$0.790.10^{-2}$	$0.525.10^{-2}$
	β	0.381	0.644	0.787	0.773	0.415
10^{-4}	L^2	$0.390.10^{-2}$	$0.223.10^{-2}$	$0.119.10^{-2}$	$0.618.10^{-3}$	$0.315.10^{-3}$
	β	0.643	0.806	0.906	0.945	0.972
	H^1	$0.361.10^{-1}$	$0.228.10^{-1}$	$0.130.10^{-1}$	$0.696.10^{-2}$	$0.361.10^{-2}$
	β	0.384	0.663	0.811	0.901	0.947
10^{-5}	L^2	$0.390.10^{-2}$	$0.223.10^{-2}$	$0.119.10^{-2}$	$0.618.10^{-3}$	$0.315.10^{-3}$
	β	0.643	0.806	0.906	0.945	0.972
	H^1	$0.360.10^{-1}$	$0.229.10^{-1}$	$0.130.10^{-1}$	$0.696.10^{-2}$	$0.361.10^{-2}$
	β	0.388	0.653	0.806	0.901	0.947

Table 4.9: **IDS** , $\alpha = 15^0$, **boundary layer**

$\epsilon \setminus N$		16	32	64	128	256
10^{-3}	L^2	$0.390.10^{-2}$	$0.222.10^{-2}$	$0.116.10^{-2}$	$0.594.10^{-3}$	$0.376.10^{-3}$
	β	0.643	0.813	0.936	0.966	0.660
	H^1	$0.361.10^{-1}$	$0.227.10^{-1}$	$0.124.10^{-1}$	$0.665.10^{-2}$	$0.454.10^{-2}$
	β	0.384	0.669	0.872	0.899	0.551
10^{-4}	L^2	$0.390.10^{-2}$	$0.223.10^{-2}$	$0.119.10^{-2}$	$0.619.10^{-3}$	$0.314.10^{-3}$
	β	0.643	0.806	0.906	0.943	0.979
	H^1	$0.360.10^{-1}$	$0.229.10^{-1}$	$0.130.10^{-1}$	$0.697.10^{-2}$	$0.360.10^{-2}$
	β	0.384	0.653	0.817	0.899	0.953
10^{-5}	L^2	$0.390.10^{-2}$	$0.223.10^{-2}$	$0.119.10^{-2}$	$0.619.10^{-3}$	$0.315.10^{-3}$
	β	0.643	0.806	0.906	0.943	0.975
	H^1	$0.360.10^{-1}$	$0.229.10^{-1}$	$0.130.10^{-1}$	$0.697.10^{-2}$	$0.361.10^{-2}$
	β	0.384	0.653	0.817	0.899	0.949

CHAPTER V

LOCAL REFINEMENT FOR FV PROBLEMS

The goal of every numerical simulation is to capture accurately the behavior of the modeled quantities. In this chapter we will concentrate on the error due to the approximation of a differential model with a discrete one. The classical error estimates state that the approximation error is proportional to some norm of the solution u and to some degree of the mesh step h , i.e., the error is proportional to $h^p \|u\|_{s,\Omega}$ for some real numbers p and s, $p \leq s$. We say that the solution u has singularities if it is not smooth enough (s is small), or the norm $\|u\|_{s,\Omega}$ is very large (usually called large gradients).

If the solution has singularities, in order to achieve good accuracy, the mesh size h has to be very small. Even with the most powerful computers available this cannot be done in a uniform manner for many multidimensional problems. However, if the singularities of the solution are localized, a substantial reduction of the computational cost can be achieved via local refinement. This means that where the gradient is large we refine the mesh so that the overall error is not big.

The most common way to refine the grid is via smooth variation of the mesh size. Practical applications of such algorithms are usually very complicated. Grid refinement procedures that consist of underlying coarse grid and patches of locally refined grids (possibly in more than one level), have been used and discussed by many authors. This approach has been also widely used in reservoir modeling (see, e.g., Pedrosa [102] and references there).

The local patch refinement procedure requires accurate treatment of the interface between the coarse and fine regions. This issue has been investigated for symmetric problems by Ewing, Lazarov and Vassilevski [43]. They have developed various interpolation procedures for cell–centered finite volume difference schemes on uniform rectangular meshes and have derived the corresponding error estimates. An extension to triangular meshes has been considered by Vassilevski, Petrova and Lazarov [132]. The convergence theory for finite volume element methods has been provided by Cai, Mandel and McCormick [28, 27] and the analysis of mixed finite element methods on locally refined grids has been considered by Ewing and J. Wang [41].

In recent years a more general approach to combine different meshes and approximation techniques, the so called "mortar" element methods has been popularized by Maday, Le Tallec and their coworkers [79, 20]. The idea of mortar element methods is to construct some matching condition between different domains and elements.

In this chapter we construct conservative cell-centered approximations on locally-refined grids for convection-diffusion second-order elliptic equations that have optimal order of convergence and satisfy the discrete maximum principle.

This chapter is organized as follows. In Section 5.1 the finite difference schemes are described and studied. In Section 5.1.1 and Section 5.1.2 the constant and linear interpolation on the interface are derived. Section 5.2 deals with the main properties of the discrete problems. The error analysis is presented in Section 5.3. In Section 5.4 extensive computer experiments are provided for a variety of convection- diffusion problems, including convection dominated ones. These tests support our theoretical results and assess the applicability of the derived schemes and error bounds. Some technical details are given in Appendices A and B.

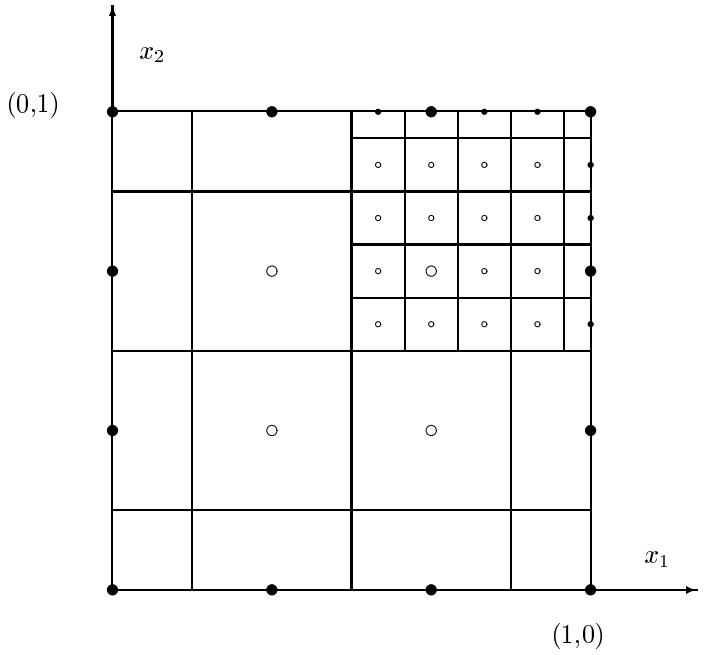


Figure 5.1: Composite cell-centered mesh

5.1 Finite difference schemes

We investigate cell–centered finite difference schemes with local refinement on 2–D uniform meshes. The schemes for such meshes are given in Section 4.2.2. We will derive error estimates for **UDS** and **MUDS**.

Now we consider the case with local refinement, where some of the cells are refined into a number of fine grid cells and introduced as grid points the centers of the new finer cells (see Fig. 5.1). The subregion covered by the refined grid is denoted by Ω_2 and the remaining part of Ω is denoted by Ω_1, i.e., $\bar{\Omega} = \bar{\Omega}_1 \bigcup \bar{\Omega}_2$. We assume that the cells are squares. There are cells of two different sizes: coarse grid cells of size h_c and fine grid cells of size $h_f = \frac{1}{m} h_c$, where m is a given positive integer.

The centers of the coarse grid cells contained in Ω define the coarse grid, which is denoted by $\tilde{\omega}$. The set of coarse grid points in Ω_2 is designated by $\tilde{\omega}_2$, i.e., $\tilde{\omega}_2 = \tilde{\omega} \bigcap \Omega_2$. The coarse grid points in Ω_1 and the fine grid points in Ω_2 define the composite grid denoted by ω. The grid points of the composite grid next to the boundary between Ω_1 and Ω_2 we will call irregular. All remaining grid points will be called regular.

From now on we will consider only the terms of the difference schemes in the x_1–direction. In the other direction the corresponding expressions are derived similarly.

Conservation of mass implies that output flux through the side of a coarse cell ($s^+_{1,i-1,j+1}$) is equal to the sum of input fluxes through the sides of neighboring fine cells, ($s_{1,i,j}$, $s_{1,i,j+1}$

and $s_{1,i,j+2}$ for the particular mesh shown on Fig. 5.1, $m = 3$), i.e.,

$$\int_{s^+_{1,i-1,j+1}} (W_1 + V_1)\, d\gamma = \int_{s_{1,i,j}} (W_1 + V_1)\, d\gamma + \int_{s_{1,i,j+1}} (W_1 + V_1)\, d\gamma + \int_{s_{1,i,j+2}} (W_1 + V_1)\, d\gamma\,.$$

We require that the finite difference schemes fulfill a conservation law at the irregular points as well. We define $w^+_{1,i-1,j+1}$ and $v^+_{1,i-1,j+1}$ via the relation

$$w^+_{1,i-1,j+1} + v^+_{1,i-1,j+1} = w_{1,i,j} + v_{1,i,j} + w_{1,i,j+1} + v_{1,i,j+1} + w_{1,i,j+2} + v_{1,i,j+2}\,.$$

There exist various ways to approximate the fluxes $w_{1,i,j+l}$ and $v_{1,i,j+l}$, $l = 0,\ 1,\ 2$. Next we consider two simple ways based on constant and linear interpolation.

5.1.1 Constant approximation

We suppose that the grid function $y(x)$, $x \in \omega$ is extended in Ω as a constant over each cell $e(x)$, $x \in \omega$. We have to consider the modification of our finite volume schemes that have to be made along the interface between Ω_1 and Ω_2, i.e., at the irregular points. We use the following formulae for non uniform mesh (see Fig. 5.2) where for definiteness we assume $h_c = 3h_f$

$$w_{1,i,j+l} = -\frac{2h_f}{h_c + h_f} k_{1,i,j+l} \overline{\triangle}_l y_{i,j+l}\,,\ l = 0,\ 1,\ 2,$$

here

$$k_{1,i,j+l} = \left(\frac{2}{h_c + h_f} \int_{x_{1,i-1}}^{x_{1,i}} \frac{ds}{a(s, x_{2,j})} \right)^{-1}\,,\ k^+_{1,i,j} = k_{1,i+1,j}$$

and

$$\overline{\triangle}_1 y_{i,j+l} = y_{i,j+l} - y_{i-1,j+l} = y_{i,j+l} - y_{i-1,j+1}\,.$$

Note that $(x_{1,i-1}, x_{2,j+l})$, $l = 0,\ 2$ are "slave" nodes and $y_{i-1,j+1} = y_{i-1,j+l}$, $l = 0,\ 2$ because we use constant interpolation. Since $h_c = 3h_f$ we get

$$w_{1,i,j+l} = -\frac{1}{2} k_{1,i,j+l} \overline{\triangle}_1 y_{i,j+l}\,, \tag{5.1}$$

$$v_{1,i,j+l} = (B_{1,i,j+l} - |B_{1,i,j}|) y_{i,j+l} + (B_{1,i,j+l} + |B_{1,i,j}|) y_{i-1,j+1}\,. \tag{5.2}$$

Because of the poor approximation properties we do not consider constant approximation for **MUDS**.

5.1.2 Linear approximation

We use this approximations for **MUDS** in irregular points. In this case is supposed that $y(x)$, $x \in \omega$ is interpolated linearly between any two neighboring coarse grid nodes. For simplicity of presentation we confine again only with the case $h_c = 3h_f$. We will need values of y at the points $(x_{1,i-1}, x_{2,j})$ and $(x_{1,i-1}, x_{2,j+2})$ which are not grid ones (see Fig. 5.2). To get them we use the following linear interpolation

$$\begin{aligned} y_{i-1,j} &= \frac{2}{3} y_{i-1,j+1} + \frac{1}{3} y_{i-1,j-1}\,, \\ y_{i-1,j+2} &= \frac{2}{3} y_{i-1,j+1} + \frac{1}{3} y_{i-1,j+4}\,. \end{aligned} \tag{5.3}$$

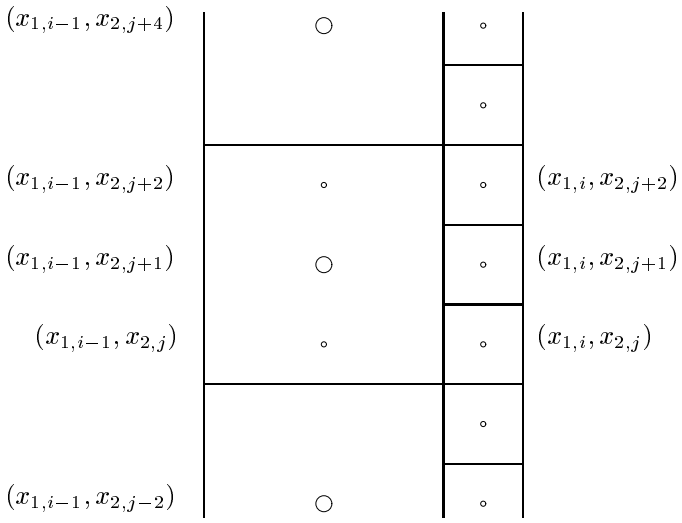


Figure 5.2: Irregular cell $e(x_{1,i-1}, x_{2,j+1})$

We sketch the derivation of **MUDS** at the irregular points (see Sections 4.1.3 for a detailed derivation of **MUDS** at the regular points and 4.2.2 for the formulas on a uniform mesh). First we write the standard central finite difference scheme, i.e., $\int_s V$ is approximated by the analog of central differences

$$\begin{aligned}\int_{s_{1,i,j}} (W+V)\, ds &= -\frac{2h_f}{h_c+h_f} k_{1,i,j}\left[y_{i,j} - y_{i-1,j}\right] \\ &\quad + b_{1,i-1/2,j} h_f \left[\frac{h_f y_{i-1,j} + h_c y_{i,j}}{h_f + h_c}\right] + O(h^2) \\ &= -\frac{1}{2} k_{1,i,j}\left[y_{i,j} - y_{i-1,j}\right] + \frac{B_{1,i,j}}{2}\left[y_{i-1,j} + 3y_{i,j}\right] + O(h^2)\,.\end{aligned}$$

Next we substitute $y_{i-1,j}$ from (5.3) and represent the terms approximating $\int_s V$ in an upwind manner

$$\begin{aligned}\int_{s_{1,i,j}} (W+V)\, ds &= -\frac{1}{2} k_{1,i,j}\left(y_{i,j} - \frac{2}{3} y_{i-1,j+1} - \frac{1}{3} y_{i-1,j-1}\right) \\ &\quad + \frac{B_{1,i,j}}{2}\left(\frac{2}{3} y_{i-1,j+1} + \frac{1}{3} y_{i-1,j-1} + 3y_{i,j}\right) + O(h^2) \\ &= -\frac{1}{2} k_{1,i,j}\left[\left(y_{i,j} - y_{i-1,j+1}\right) + \frac{1}{3}\left(y_{i-1,j+1} - y_{i-1,j-1}\right)\right] \\ &\quad + \left(B_{1,i,j} - |B_{1,i,j}|\right) y_{i,j} \\ &\quad + \left(B_{1,i,j} + |B_{1,i,j}|\right)\left(\frac{2}{3} y_{i-1,j+1} + \frac{1}{3} y_{i-1,j-1}\right) \\ &\quad + \left(\frac{B_{1,i,j}}{2} + |B_{1,i,j}|\right)\left(y_{i,j} - y_{i-1,j+1}\right) \\ &\quad + \frac{1}{3}\left(\frac{B_{1,i,j}}{2} + |B_{1,i,j}|\right)\left(y_{i-1,j+1} - y_{i-1,j-1}\right) + O(h^2)\,.\end{aligned}$$

Finally, we get

$$\begin{aligned}\int_{s_{1,i,j}} (W+V)\, ds &= -\frac{1}{2}\left[k_{1,i,j} - (2|B_{1,i,j}| + B_{1,i,j})\right] \overline{\triangle}_1 y_{i,j} \\ &\quad -\frac{1}{6}\left[k_{1,i,j} - (2|B_{1,i,j}| + B_{1,i,j})\right] \overline{\triangle}_2 y_{i-1,j+1} \\ &\quad + (B_{1,i,j} - |B_{1,i,j}|)\, y_{i,j} \\ &\quad + (B_{1,i,j} + |B_{1,i,j}|) \left(\frac{2}{3} y_{i-1,j+1} + \frac{1}{3} y_{i-1,j-1}\right) + O(h^2)\,.\end{aligned}$$

In order to obtain an upwind scheme we approximate the first term in the above formulae

$$\begin{aligned}k_{1,i,j} - (2|B_{1,i,j}| + B_{1,i,j}) &= \frac{k_{1,i,j}}{1 + (2|B_{1,i,j}| + B_{1,i,j})/k_{1,i,j}} \\ &\quad - \frac{(2|B_{1,i,j}| + B_{1,i,j})^2}{k_{1,i,j} + 2|B_{1,i,j}| + B_{1,i,j}} \\ &= \frac{k_{1,i,j}}{1 + (2|B_{1,i,j}| + B_{1,i,j})/k_{1,i,j}} + O(h^2)\,.\end{aligned}$$

In the last step we have taken into account that $B_1 = O(h)$. In this way we define the approximate fluxes w and v as follows:

$$\begin{aligned}w_{1,i,j} &= -\frac{1}{2}.\frac{k_{1,i,j}}{1 + (2|B_{1,i,j}| + B_{1,i,j})/k_{1,i,j}} \overline{\triangle}_1 y_{i,j} \\ &\quad -\frac{1}{6}\left(\frac{k_{1,i,j}}{1 + (2|B_{1,i,j}| + B_{1,i,j})/k_{1,i,j}}\right) \overline{\triangle}_2 y_{i-1,j+1}\,, \\ w_{1,i,j+1} &= -\frac{1}{2}.\frac{k_{1,i,j+1}}{1 + (2|B_{1,i,j+1}| + B_{1,i,j+1})/k_{1,i,j+1}} \overline{\triangle}_1 y_{i,j+1}\,, \\ w_{1,i,j+2} &= -\frac{1}{2}.\frac{k_{1,i,j+2}}{1 + (2|B_{1,i,j+2}| + B_{1,i,j+2})/k_{1,i,j+2}} \overline{\triangle}_1 y_{i,j+2} \\ &\quad -\frac{1}{6}.\frac{k_{1,i,j+2}}{1 + (2|B_{1,i,j+2}| + B_{1,i,j+2})/k_{1,i,j+2}} \triangle_2 y_{i-1,j+1}\,,\end{aligned} \tag{5.4}$$

and

$$\begin{aligned}v_{1,i,j} &= (B_{1,i,j} - |B_{1,i,j}|)\, y_{i,j} + (B_{1,i,j} + |B_{1,i,j}|)\, y_{i-1,j+1} \\ &\quad -\frac{1}{3}(B_{1,i,j} + |B_{1,i,j}|) \overline{\triangle}_2 y_{i-1,j+1}\,, \\ v_{1,i,j+1} &= (B_{1,i,j+1} - |B_{1,i,j+1}|)\, y_{i,j+1} + (B_{1,i,j+1} + |B_{1,i,j+1}|)\, y_{i-1,j+1}\,, \\ v_{1,i,j+2} &= (B_{1,i,j+2} - |B_{1,i,j+2}|)\, y_{i,j+2} + (B_{1,i,j+2} + |B_{1,i,j+2}|)\, y_{i-1,j+1} \\ &\quad +\frac{1}{3}(B_{1,i,j+2} + |B_{1,i,j+2}|) \triangle_2 y_{i-1,j+1}\,.\end{aligned} \tag{5.5}$$

5.2 Formulation of the discrete problems

Two difference schemes derived in Sections 5.1.1 and 5.1.2 can be written in the general form

$$\begin{cases} \sum_{x\in\omega} \sum_{l=1}^{2} \left(w_l^+(x) - w_l(x)\right) + \left(u_l^+(x) - u_l(x)\right) = \int_e f(x)\, ds & \text{in } \Omega, \\ y(x) = g(x) & \text{on } \Gamma. \end{cases}$$

For **MUDS** the approximate fluxes $w_l^+(x)$, $w_l(x)$, $v_l^+(x)$ and $v_l(x)$ are defined by (4.33), (4.30) at the regular and by (5.4) and (5.5) at the irregular points. In matrix terms we write

$$A\mathbf{y} = \mathbf{f}. \tag{5.6}$$

where in the right-hand side $\mathbf{f}$ we have taken into account the boundary conditions. We will denote A_0 the matrix of constant approximation (5.1), (5.2) in irregular points and (4.29), (4.32) in regular points; for this scheme we denote

$$A_0\mathbf{y} = \mathbf{f}. \tag{5.7}$$

Consider $x = (x_{1,i-1}, x_{2,j+1})$ (see Fig. 5.2). We let

$$B^+_{1,i-1,j+1} = B_{1,i,j} + B_{1,i,j+1} + B_{1,i,j+2}\,.$$

We will use the following auxiliary result.

Lemma 5.1 *Let* $\mathbf{b}(x) \in (W^{1+\alpha,\infty}(\Omega))^2$, $\alpha > 0$ *and* $(div(\mathbf{b}(x)) \geq 0$. *Then there exists a positive constant* C_0 *such that the inequality holds*

$$\left[(B^+_{1,i,j} - B_{1,i,j}) + (B^+_{2,i,j} - B_{2,i,j})\right] \geq -C_0 h^{2+\alpha}\,, \quad 0 < \alpha \leq 1\,.$$

Proof: We first show the result for regular points. Consider the linear functional:

$$l(b_1) := \frac{b_{1,i+1/2,j} - b_{1,i-1/2,j}}{h} - \frac{\partial b_{1,i,j}}{\partial x_1}\,.$$

This functional is bounded for $b_1 \in W^{1+\alpha,\infty}(\Omega)$, $0 \leq \alpha \leq 2$ and vanishes for all polynomials of second degree. Therefore, by the Bramble-Hilbert lemma argument we get

$$|l(b_1)| \leq Ch^{\alpha}|b_1|_{1+\alpha,\infty,e}\,.$$

Similar inequality holds for b_2. Using the triangle inequality and the assumption $\mathrm{div}(\mathbf{b}(\mathrm{x})) \geq 0$, the desired inequality is obtained.

For the irregular point $x = (x_{1,i-1}, x_{2,j+1})$ and the adjacent points $(x_{1,i}, x_{2,j})$, $(x_{1,i}, x_{2,j+1})$ and $(x_{1,i}, x_{2,j+2})$ in the refined region we consider the linear functional l

$$\begin{aligned} l(b_1) \quad = \quad & \frac{b_{1,i-1/2,j} + b_{1,i-1/2,j+1} + b_{1,i-1/2,j+2} - 3b_{1,i-3/2,j+1}}{3h_c} \\ & - \frac{\partial b_1(x_{1,i-1}, x_{2,j+1})}{\partial x_1}\,. \end{aligned}$$

The functional l is bounded for $b_1 \in W^1_\infty(e(x_{1,i}, x_{2,j}))$ and vanishing for all polynomials of first degree. Hence

$$|l(b_1)| \leq Ch^{\alpha}|b_1|_{1+\alpha,\infty,e}\,,\ 0 < \alpha \leq 1\,.$$

A similar inequality holds for b_2. Using the triangle inequality and the assumption $\mathrm{div}(\mathbf{b}) \geq 0$ the result follows. □

Our goal now is to show that both schemes have unique solutions. First we investigate some properties of **UDS** in the following lemma.

Lemma 5.2 *Let* $p(x)$ *and* $q(x)$ *be grid functions. If* A_0 *is the matrix defined by (4.29), (4.32), (5.1) and (5.2) then the following formulae holds*

$$\begin{aligned} \mathbf{p}^T A_0 \mathbf{q} \quad = \quad & -\sum_{x\in\omega}\sum_{l=1}^{2} w_l(x)\overline{\triangle}_l p(x) \\ & + \sum_{x\in\omega}\sum_{l=1}^{2} |B_l(x)|\overline{\triangle}_l p(x)\overline{\triangle}_l q(x) \\ & + \sum_{x\in\omega}\sum_{l=1}^{2} \left(B^+_l(x) - B_l(x)\right) p(x)q(x) \\ & + \sum_{x\in\omega}\sum_{l=1}^{2} B_l(x)\left(p(x)\overline{\triangle}_l q(x) - q(x)\overline{\triangle}_l p(x)\right)\,, \end{aligned} \tag{5.8}$$

for all $\mathbf{p}, \mathbf{q} \in D^0 = \{p : p_{|\Gamma} = 0\}$*, where the approximate fluxes* w_l *are defined by the values of* $q(x)$*.*

(The proof is provided in Appendix A.)

Corollary 5.1 *If* $\mathbf{b} \in \left(W^{1+\alpha,\infty}(\Omega)\right)^2$, $\alpha > 0$ *then there exists an* h_0 *such that for* $h < h_0$ *the matrix* A_0 *is positive real, i.e. iits symmetric part is a positive definite matrix, and hence the* **UDS** *defined by (4.29), (4.32), (5.1) and (5.2) has unique solution. Moreover, we have the discrete* H_0^1*-coercivity estimate*

$$\mathbf{q}^T A_0 \mathbf{q} \geq C \|q\|_{1,\omega}^2, \quad \mathbf{q} \in D^0 .$$

Proof: Consider a one-dimensional grid function $q(x_{1,i}, x_{2,0})$, $i = 0, \ldots, N$, where $x_{2,0}$ is fixed and $q(x_{1,N}, x_{2,0}) = 0$. The following inequality holds [114]

$$\sum_{i=1}^{N} \overline{\triangle}_1^2 q_{i,0} \geq C \sum_{i=0}^{N} h^2 q_{i,0}^2 .$$

Combining similar inequalities in the x_1 and x_2 directions, we obtain $|q|_{1,\omega} \geq C\|q\|_{0,\omega}$ for $\mathbf{q} \in D^0$. To conclude the proof we set $\mathbf{p} = \mathbf{q}$ in (5.8) and apply Lemma 5.1. □

We establish the uniqueness of the solution for **MUDS** in the following theorem.

Theorem 5.1 *If* $\mathbf{b}(x) \in \left(W^{1+\alpha,\infty}(\Omega)\right)^2$, $\alpha > 0$ *then there exists an* h_0 *such that for* $h < h_0$ **MUDS** *defined by (4.33), (4.30), (5.4) and (5.5) has a unique solution. Moreover, the following inequalities hold*

$$\gamma_1 \mathbf{q}^T A_0 \mathbf{q} \leq \mathbf{q}^T A \mathbf{q} \leq \gamma_2 \mathbf{q}^T A_0 \mathbf{q}, \quad \mathbf{q} \in D^0,$$

$$|\mathbf{p}^T A \mathbf{q}| \leq \gamma_2 (\mathbf{p}^T A_0 \mathbf{p})^{1/2} (\mathbf{q}^T A_0 \mathbf{q})^{1/2}, \quad \mathbf{p}, \mathbf{q} \in D^0,$$

where A_0 *is the matrix of constant approximation, and* A *is the matrix of linear approximation.*

Proof: We can write matrix A_0 in the following form

$$A_0 = A_0^{(2)} + A_0^{(1)},$$

where $A_0^{(2)}$ corresponds to the diffusion part and $A_0^{(1)}$ corresponds to the remaining convection part. For $A_0^{(2)}$ we have

$$\begin{aligned} \mathbf{p}^T A_0^{(2)} \mathbf{q} &= -\sum_{x \in \omega} w_1(x) \overline{\triangle}_1 p(x) + w_2(x) \overline{\triangle}_2 p(x) \\ &= \sum_{x \in \omega} \left(\alpha_1 \overline{\triangle}_1 q \overline{\triangle}_1 p + \alpha_2 \overline{\triangle}_2 q \overline{\triangle}_2 p \right) , \end{aligned} \tag{5.9}$$

where

$$\alpha_1 = \alpha_{1,i,j} = \begin{cases} k_{1,i,j}/2 & \text{for } j \geq 0,\ i = 0, \\ k_{1,i,j} & \text{for the remaining indices,} \end{cases}$$

$$\alpha_2 = \alpha_{2,i,j} = \begin{cases} k_{2,i,j}/2 & \text{for } i \geq 0,\ j = 0, \\ k_{2,i,j} & \text{for the remaining indices} \end{cases}$$

(See Fig. 5.3). In the same way we split the matrix A arising from linear approximation into two parts

$$A = A^{(2)} + A^{(1)}.$$

For $A^{(2)}$ we get

$$\begin{aligned}\mathbf{p}^T A^{(2)}\mathbf{q} &= \sum_{x\in\omega}\left(\beta_1\overline{\triangle}_l q\overline{\triangle}_l p + \beta_2\overline{\triangle}_2 q\overline{\triangle}_2 p\right)\\ &+\frac{1}{6}\sum_{j=1,4,7,\ldots}\left[\beta_{1,0,j-1}\overline{\triangle}_2 q_{-1,j}\overline{\triangle}_1 p_{0,j-1} - \beta_{1,0,j+1}\triangle_2 q_{-1,j}\overline{\triangle}_1 p_{0,j+1}\right]\\ &+\frac{1}{6}\sum_{i=1,4,7,\ldots}\left[\beta_{2,i-1,0}\overline{\triangle}_1 q_{i,-1}\overline{\triangle}_2 p_{i-1,0} - \beta_{2,i+1,0}\triangle_1 q_{i,-1}\overline{\triangle}_2 p_{i+1,0}\right],\end{aligned}$$

where

$$\beta_1 = \beta_{1,i,j} = \begin{cases}\tilde{k}_{1,i,j}/2 & \text{for } j\geq 0,\ i=0,\\ \hat{k}_{1,i,j} & \text{for the remaining indices,}\end{cases}$$

$$\beta_2 = \beta_{2,i,j} = \begin{cases}\tilde{k}_{2,i,j}/2 & \text{for } i\geq 0,\ j=0,\\ \hat{k}_{2,i,j} & \text{for the remaining indices,}\end{cases}$$

and

$$\tilde{k}_{l,i,j} = \frac{k_{l,i,j}}{1+\left(2|B_l|+B_l\right)/k_{l,i,j}}, \quad \hat{k}_{l,i,j} = \frac{k_{l,i,j}}{1+|B_l|/k_{l,i,j}}.$$

Applying the Cauchy inequality to the $\mathbf{p}^T A^{(2)}\mathbf{q}$ and taking into account that $\tilde{k}_{l,i,j}$ and $\hat{k}_{l,i,j}$ are less than $k_{l,i,j}$ we get

$$|\mathbf{p}^T A^{(2)}\mathbf{q}| \leq \left(\frac{7}{6}+C_2 h\right)\left(\mathbf{p}^T A_0^{(2)}\mathbf{p}\right)^{1/2}\left(\mathbf{q}^T A_0^{(2)}\mathbf{q}\right)^{1/2},$$

where the constant C_2 depends on the values of the coefficient $a(x)$ only locally, i.e., cell by cell. To derive a lower bound we need the inequality

$$Pk_{l,i,j} - k_{l,i,j} - |B_{l,i,j}| > 0, \quad P := 1+\sup_{\substack{x\in\omega\\ l=1,2}}\frac{|b_l(x)|h_c}{2k_l(x)}.$$

Consider auxiliary matrix $A_*^{(2)}$ obtained by replacing in (5.9) the coefficients α_1, α_2 with β_1, β_2. For $\mathbf{p}=\mathbf{q}$ combining

$$\left(\frac{5}{6}-C_1 h\right)\mathbf{q}^T A_*^{(2)}\mathbf{q} \leq \mathbf{q}^T A^{(2)}\mathbf{q} \quad \text{and} \quad P^{-1}\mathbf{q}^T A_0^{(2)}\mathbf{q} \leq \mathbf{q}^T A_*^{(2)}\mathbf{q}$$

we get

$$P^{-1}\left(\frac{5}{6}-C_1 h\right)\mathbf{q}^T A_0^{(2)}\mathbf{q} \leq \mathbf{q}^T A^{(2)}\mathbf{q} \leq \left(\frac{7}{6}+C_2 h\right)\mathbf{q}^T A_0^{(1)}\mathbf{q}. \tag{5.10}$$

The remark above is also valid for the constant C_1. The derivation of Lemma 5.2, (5.8) and (5.5) gives us

$$\begin{aligned}\mathbf{p}^T A^{(1)}\mathbf{q} &= \mathbf{p}^T A_0^{(1)}\mathbf{q} + \frac{1}{3}\sum_{j=1,4,7,\ldots}\left[\left(B_{1,0,j}+|B_{1,0,j}|\right)\overline{\triangle}_2 q_{-1,j}\overline{\triangle}_1 p_{0,j}\right.\\ &\quad\left.-\left(B_{1,0,j+2}+|B_{1,0,j+2}|\right)\triangle_2 q_{-1,j}\overline{\triangle}_1 p_{0,j+2}\right]\\ &+\frac{1}{3}\sum_{i=1,4,7,\ldots}\left[\left(B_{2,i,0}+|B_{2,i,0}|\right)\overline{\triangle}_1 q_{i,-1}\overline{\triangle}_2 p_{i,0}\right.\\ &\quad\left.-\left(B_{2,i+2,0}+|B_{2,i+2,0}|\right)\triangle_1 q_{i,-1}\overline{\triangle}_2 p_{i+2,0}\right].\end{aligned}$$

Similarly as (5.10) was derived we find

$$\left(\frac{1}{3} - C_3 h\right) \mathbf{q}^T A_0^{(1)} \mathbf{q} \le \mathbf{q}^T A^{(1)} \mathbf{q} \le \left(\frac{5}{3} + C_4 h\right) \mathbf{q}^T A_0^{(1)} \mathbf{q}.$$

The above inequalities show the desired result. □

Remark 5.1 P is in fact local Peclet number plus 1 and γ_1 depends on P. This shows that condition number of the matrix $A_0^{-1}A$ can become very large when P is a large number.

Corollary 5.2 *If* $\mathbf{b}(x) \in \left(W^{1+\alpha,\infty}(\Omega)\right)^2$, $\alpha > 0$ *then there exists* h_0 *such that for* $h < h_0$ *the matrix* A *is positive real. Moreover, the discrete* H_0^1*-coercivity holds*

$$\mathbf{q}^T A \mathbf{q} \ge C \|q\|_{1,\omega}^2, \quad \mathbf{q} \in D^0.$$

Remark 5.2 The corollaries 3.1 and 3.2 asserts that for sufficiently small step-size h A and A^0 are **M**-matrices.

Remark 5.3 If the equation (2.11) is singularly perturbed we still have stability, but the constant C in the corollaries 5.1 and 5.2 depends on ε. In this case error estimates derived in §5.3 deteriorate. However, for fixed ε the asymptotic behavior of the error is correctly predicted by our estimates for $h \approx \varepsilon$.

5.3 Error estimates

The error analysis presented here is done in the general framework of the methods developed in [115] and [43]. We consider only the case when $a(x) \equiv 1$. Let

$$z(x) = y(x) - u(x), \ x \in \omega$$

be the error of the finite difference method. Substituting $y = z + u$ in (5.6) (5.7) we obtain

$$A\mathrm{z} = \mathrm{f} - Au \equiv \psi. \tag{5.11}$$

Then using (4.4)–(5.11) we transform ψ in the following form

$$\sum_{l=1}^{2} \left\{ \left[\int_{s_l^+} -\frac{\partial u}{\partial x_l} \, d\gamma - w_l^+ \right] - \left[\int_{s_l} -\frac{\partial u}{\partial x_l} \, d\gamma - w_l \right] \right\}$$
$$+ \sum_{l=1}^{2} \left\{ \left[\int_{s_l^+} b_l u \, d\gamma - v_l^+ \right] - \left[\int_{s_l} b_l u \, d\gamma - v_l \right] \right\} \equiv \psi_1 + \psi_2 = \psi.$$

where the local truncation error ψ has been split up into two terms

$$\psi_2 \equiv \sum_{l=1}^{2} \left(\eta_l^+(x) - \eta_l(x)\right), \ \psi_1 \equiv \sum_{l=1}^{2} \left(\mu_l^+(x) - \mu_l(x)\right),$$

$$\eta_l = \int_{s_l} -\frac{\partial u}{\partial x_l} \, d\gamma - w_l, \ \mu_l = \int_{s_l} b_l u \, d\gamma - v_l. \tag{5.12}$$

Here ψ_1 is the error of approximation of first derivatives, and ψ_2 is the error of approximation of the second derivatives.

Note that the approximate fluxes w_l^+, w_l, v_l^+, v_l are defined by the values of $u(x)$ at the grid points, and the components of the local truncation error η_l and μ_l are determined on the shifted grids ω_l^+, $l = 1, 2$. Using summation by parts and the Schwartz inequality, we get

$$\begin{aligned}
(\psi_2, z) &= \sum_{l=1}^{2} \sum_{x \in \omega} \left[\eta_l^+(x) - \eta_l(x)\right] z(x) \\
&= -\sum_{l=1}^{2} \sum_{x \in \omega_l^+} \eta_l(x) \overline{\triangle}_l z(x) \\
&\leq \left(\sum_{l=1}^{2} \sum_{x \in \omega_l^+} \eta_l^2(x) \right)^{1/2} \left(\sum_{l=1}^{2} \sum_{x \in \omega_l^+} \overline{\triangle}_l^2 z(x) \right)^{1/2} \\
&\leq \left(\|\eta_1\rceil|_1 + \|\eta_2\rceil|_2 \right) \|z\|_{1,\omega} \, .
\end{aligned}$$

Likewise,

$$(\psi_2, z) \leq \left(\|\mu_1\rceil|_1 + \|\mu_2\rceil|_2 \right) \|z\|_{1,\omega} \, .$$

Summarizing these results and using the corollaries 1 and 2 we obtain the following main result.

Lemma 5.3 *The error $z(x) = y(x) - u(x)$, $x \in \omega$ of all considered finite difference schemes satisfies the a priori estimate*

$$\|z\|_{1,\omega} \leq C \sum_{l=1}^{2} \left(\|\eta_l\rceil|_l + \|\mu_l\rceil|_l \right) \, , \tag{5.13}$$

where the components η_l, μ_l, $l = 1, 2$ of the local truncation error are defined by (5.12) with approximate fluxes w_l^+, w_l, v_l^+, v_l, $l = 1, 2$ determined by (4.29), (4.32), (5.1) and (5.2) for **UDS** *and (4.33), (4.30), (5.4) and (5.5) for* **MUDS**. *(The constant C does not depend on h or z.)*

In order to use the estimate (5.13) of Lemma 5.3 we have to bound the norms of η_l, μ_l, $l = 1, 2$ defined by (5.12). We state the estimates for the local truncation error components in regular points, proved in Ewing, Lazarov and Vassilevski [43]

$$|\eta_l(x)| \leq C h^{m-1} |u|_{m,\overline{e}}, \; \frac{3}{2} < m \leq 3, \tag{5.14}$$

and in Lazarov, Mishev and Vassilevski [78]

$$|\mu_l(x)| \leq \begin{cases} C h^m \|b_l\|_{1,\infty,\Omega} |u|_{m,\overline{e}} & \text{for } \mathbf{MUDS}, \\ C \left[h |b_l|_{0,\infty,\Omega} |u|_{1,\overline{e}} + h^m \|b_l\|_{1,\infty,\Omega} |u|_{m,\overline{e}} \right] & \text{for } \mathbf{UDS}, \end{cases} \tag{5.15}$$

where $1 < m \leq 2$; $\overline{e} = e_{i-1,j} \bigcup e_{i,j}$ for $l = 1$ and $\overline{e} = e_{i,j-1} \bigcup e_{i,j}$ for $l = 2$.

Now we consider the components of the local truncation error for the **MUDS** at the irregular points $(x_{1,i}, x_{2,j+l})$, $l = 0, 1, 2$. We remark here that we split the schemes into two

parts only for convenience of the analysis. We replace (5.4) by

$$\begin{aligned}
w_{1,i,j} &= -\frac{1}{2}\left(\frac{1}{1+2|B_{1,i,j}|+B_{1,i,j}}+2|B_{1,i,j}|+B_{1,i,j}\right)\\
&\quad\times\left[u_{i,j}-\frac{2}{3}u_{i-1,j+l}-\frac{1}{3}u_{i-1,j-1}\right],\\
w_{1,i,j+1} &= -\frac{1}{2}\left(\frac{1}{1+2|B_{1,i,j+1}|+B_{1,i,j+1}}+2|B_{1,i,j+1}|+B_{1,i,j+1}\right)\\
&\quad\times\left[u_{i,j+1}-u_{i-1,j+1}\right],\\
w_{1,i,j+2} &= -\frac{1}{2}\left(\frac{1}{1+2|B_{1,i,j+2}|+B_{1,i,j+2}}+2|B_{1,i,j+2}|+B_{1,i,j+2}\right)\\
&\quad\times\left[u_{i,j+2}-\frac{2}{3}u_{i-1,j+1}-\frac{1}{3}u_{i-1,j+4}\right],
\end{aligned}$$

and (5.5) by

$$\begin{aligned}
v_{1,i,j} &= \frac{B_{1,i,j}}{2}\left[\frac{2}{3}u_{i-1,j+1}+\frac{1}{3}u_{i-1,j-1}+3y_{i,j}\right],\\
v_{1,i,j+1} &= \frac{B_{1,i,j+1}}{2}\left[u_{i-1,j+1}+3u_{i,j+l}\right],\\
v_{1,i,j+2} &= \frac{B_{1,i,j+2}}{2}\left[\frac{2}{3}u_{i-1,j+1}+\frac{1}{3}u_{i-1,j+4}+3u_{i,j+2}\right].
\end{aligned}$$

Note that $w_{1,i,j+l}+v_{1,i,j+l}$, $l=0,\,1,\,2$ is not changed. Consider η_1. By construction

$$\left(\frac{1}{1+2|B_{1,i,j+l}|+B_{1,i,j+l}}+2|B_{1,i,j+l}|+B_{1,i,j+l}\right)=1+C_1(x)h^2,$$

where $C_1(x)\sim b_1^2(x)$. Then in the point $(x_{1,i},x_{2,j})$ we have

$$\begin{aligned}
\eta_1(x_{1,i},x_{2,j}) &= -\int_{s(i,j)}\frac{\partial u}{\partial x_1}\,d\gamma+w_1(x)\\
&= -\int_{s(i,j)}\frac{\partial u}{\partial x_1}\,d\gamma+\frac{1}{2}(1+C_1h^2)\left[u_{i,j}-\frac{2}{3}u_{i-1,j+1}-\frac{1}{3}u_{i-1,j-1}\right].
\end{aligned}$$

Taking into account

$$\left|u_{i,j}-\frac{2}{3}u_{i-1,j+1}-\frac{1}{3}u_{i-1,j-1}\right|\le C(|u|_{1,\overline{e}}+h^{m-1}|u|_{m,\overline{e}}),\ 1<m\le 2$$

and the estimate (see [43])

$$\left|\int_{s(i,j)}\frac{\partial u}{\partial x_1}\,d\gamma-\frac{1}{2}\left[u_{i,j}-\frac{2}{3}u_{i-1,j+1}-\frac{1}{3}u_{i-1,j-1}\right]\right|\le Ch^{m-1}|u|_{m,\overline{e}},\ \frac{3}{2}<m\le 2,$$

we get

$$|\eta_1(x_{1,i},x_{2,j})|\le Ch^{m-1}|u|_{m,\overline{e}},\ \frac{3}{2}<m\le 2. \tag{5.16}$$

With the similar argument we obtain the estimate (5.16) for $\eta(x_{1,i},x_{2,j+l})$,$l=1,\,2$. The inequalities (5.14) and (5.16) imply

$$\begin{aligned}
\sum_{x\in\omega}\eta_1^2(x) &\le Ch^{2m-2}\left(\sum_{x\in\Omega_h}|u|^2_{m,\overline{e}(x)}+\sum_{x\in\omega}h^{2\alpha}|u|^2_{m+\alpha,\overline{e}(x)}\right)\\
&\le Ch^{2m-2}(|u|^2_{m,\Omega_h}+h^{2\alpha}|u|^2_{m+\alpha,\Omega}),
\end{aligned}$$

here Ω_h is a strip with a width $4h$ around the interface between Ω_1 and Ω_2 (coarse and fine grid regions) and $\frac{3}{2} < m \leq 2\,,\ 0 \leq \alpha \leq 1$.

The first term in the right is estimated by the well-known Il'in's inequality [115], [43]

$$\|u\|_{0,\Omega_\delta} \leq C\delta^\alpha \|u\|_{\alpha,\Omega}\,, \quad 0 \leq \alpha < \frac{1}{2}\,,$$

where Ω_δ is a strip in Ω with a width δ. Therefore, we have

$$\|\eta_1\rceil|_1 = \left(\sum_{x\in\omega} \eta_1^2(x)\right)^{1/2} \leq Ch^{m-1}\|u\|_{m,\Omega}\,, \quad \frac{3}{2} < m < \frac{5}{2}\,. \tag{5.17}$$

In a similar way we can estimate $\eta_2(x)$.

For the component $\mu_1(x)$ at the irregular points we prove in the Appendix B the upper bound

$$\|\mu_1\rceil|_1 \leq Ch^m \|b_1\|_{1,\infty,\Omega}\|u\|_{m,\Omega}\,, \quad 1 < m \leq 2. \tag{5.18}$$

Summarizing these results we get

Theorem 5.2 *If the solution of the problem (2.11) is H^m–regular, $\frac{3}{2} < m < \frac{5}{2}$ then for the* **MUDS** *is valid*

$$\|y-u\|_{1,\omega} \leq Ch^{m-1}\left[1 + h^\delta\left(\|b_1\|_{1,\infty,\Omega} + \|b_2\|_{1,\infty,\Omega}\right)\right]\|u\|_{m,\Omega}\,.$$

Here

$$\delta = \begin{cases} 1 & \frac{3}{2} < m \leq 2, \\ 3-m & 2 \leq m < \frac{5}{2}. \end{cases}$$

With the same approach one can prove the following result for **UDS**.

Theorem 5.3 *If the solution $u(x)$ of the problem (2.11) is H^m–regular, $\frac{3}{2} < m \leq 3$ then for the* **UDS** *is valid*

$$\|y-u\|_{1,\omega} \leq Ch^{1/2}\left[1 + h^{1/2}(\|b_1\|_{1,\infty,\Omega} + \|b_2\|_{1,\infty,\Omega})\right]\|u\|_{m,\Omega}$$

5.4 Numerical results

In this section on the basis of model examples we study the error behavior of all considered schemes. We consider three test problems. In first two examples we solved (2.11) with the velocity field

$$b_1 = (1 + x\cos(\alpha))\cos(\alpha)\,, \quad b_2 = (1 + y\sin(\alpha))\sin(\alpha)\,, \tag{5.19}$$

where the angle was $\alpha = 15^0$.

Problem 5.1 *Consider a smooth solution with a diffusion coefficient $a(x) = 1$*

$$u(x) = \begin{cases} 10\,\exp(-\frac{c^2}{c^2-r^2}), & r < c, \\ 0, & r \geq c, \end{cases}$$

where $c = 0.125\,,\ r^2 = (x-x_0)^2 + (y-y_0)^2\,,\ x_0 = 0.8\,,\ y_0 = 0.7$.

Table 5.1: Problem 5.1, **MUDS**

n_c	h_c/h_f	*error* (1)	*order*	*error* (2)	*order*
10	1	$0.284.10^{+1}$		$0.284.10^{+1}$	
	3	$0.870.10^{0}$		$0.285.10^{+1}$	
	5	$0.365.10^{0}$		$0.291.10^{+1}$	
	7	$0.250.10^{0}$		$0.296.10^{+1}$	
20	1	$0.132.10^{+1}$	1.105	$0.132.10^{+1}$	1.105
	3	$0.398.10^{0}$	1.128	$0.135.10^{+1}$	1.078
	5	$0.172.10^{0}$	1.085	$0.136.10^{+1}$	1.094
	7	$0.825.10^{-1}$	1.599	$0.138.10^{+1}$	1.100
40	1	$0.745.10^{0}$	0.825	$0.745.10^{0}$	0.825
	3	$0.119.10^{0}$	1.742	$0.681.10^{0}$	0.987
	5	$0.459.10^{-1}$	1.906	$0.694.10^{0}$	0.971
	7	$0.236.10^{-1}$	1.806	$0.700.10^{0}$	0.979
80	1	$0.232.10^{0}$	1.683	$0.232.10^{0}$	1.683
	3	$0.328.10^{-1}$	1.859	$0.243.10^{0}$	1.487
	5	$0.119.10^{-1}$	1.948	$0.249.10^{0}$	1.479
	7	$0.610.10^{-2}$	1.952	$0.252.10^{0}$	1.474
160	1	$0.691.10^{-1}$	1.747	$0.691.10^{-1}$	1.747

We choose two different domains $\Omega = \Omega_2^{(l)}$, $l = 1, 2$ for local refinement to investigate the influence of the interpolation along the boundary of Ω_2. When the support of $u(x)$ is in $\Omega_2^{(1)} = \{0.5 \le x \le 1\,,\ 0.5 \le y \le 1\}$, the error caused by the interpolation is eliminated and we get approximately second order of convergence. This shows that when $|u|_{1,\Omega_1}$ is comparatively small we can expect good results using schemes with local refinement. The worst possible case is when the solution $u(x)$ has a large gradient along the boundary of Ω_2. We tested this case for a subdomain $\Omega_2^{(2)} = \{0.7 \le x \le 1\,,\ 0.7 \le y \le 1\}$. The results in Table 5.1 show $O(h^{3/2})$ convergence rate in the discrete H^1–norm, i.e., we lose half of order of accuracy which is in agreement with Theorem 5.2.

Problem 5.2 *Consider a solution $u \in H^m(\Omega)$, $m < \frac{5}{2}$ which support is in $\Omega_2 = \{0.5 \le x \le 1\,,\ 0.5 \le y \le 1\}$ and a smooth coefficient $a(x)$,*

$$a(x) = \left[1 + 10(x^2 + y^2)\right]^{-1}\,, \quad u(x) = \phi(x)\psi(y)\,,$$

$$\phi(x) = \begin{cases} \sin^2\left(\pi \frac{x-d_1}{1-d_1}\right), & x \in (d_1, 1), \\ 0, & \textit{otherwise}, \end{cases}$$

$$\psi(y) = \begin{cases} \sin^2\left(\pi \frac{y-d_2}{1-d_2}\right), & y \in (d_2, 1), \\ 0, & \textit{otherwise}, \end{cases}$$

where $d_1 = d_2 = 0.875$.

We compare the H^1 error for both schemes, **UDS** and **MUDS**. In the last column of Table 5.2 the number of unknowns N is shown. It is clear from the results in Table 5.2 that **MUDS** is superior to **UDS** and it is also seen that a prescribed accuracy can be achieved for less unknowns when local refinement is used.

Problem 5.3 *Consider a smooth solution u with a boundary layer along line $x = 1$,*

$$u(x) = 4xy(1-y)\left(1 - \frac{\exp(x/\varepsilon) - 1}{\exp(1/\varepsilon) - 1}\right)\,,$$

Table 5.2: Problem 5.2

n_c	h_c/h_f	UDS	MUDS	N
10	1	$0.128.10^{0}$	$0.101.10^{0}$	100
	3	$0.855.10^{-1}$	$0.494.10^{-1}$	331
	5	$0.606.10^{-1}$	$0.193.10^{-1}$	804
	7	$0.470.10^{-1}$	$0.120.10^{-1}$	1519
20	1	$0.138.10^{0}$	$0.174.10^{0}$	400
	3	$0.545.10^{-1}$	$0.149.10^{-1}$	1261
	5	$0.355.10^{-1}$	$0.569.10^{-2}$	3004
	7	$0.265.10^{-1}$	$0.447.10^{-2}$	5629
40	1	$0.743.10^{-1}$	$0.836.10^{-1}$	1600
	3	$0.311.10^{-1}$	$0.599.10^{-2}$	4921
	5	$0.194.10^{-1}$	$0.248.10^{-2}$	11604
	7	$0.141.10^{-1}$	$0.786.10^{-3}$	21649
80	1	$0.443.10^{-1}$	$0.466.10^{-1}$	6400
	3	$0.165.10^{-1}$	$0.197.10^{-2}$	19441
	5	$0.101.10^{-1}$	$0.850.10^{-3}$	45604
	7	$0.293.10^{-2}$	$0.207.10^{-3}$	84889

Table 5.3: Problem 5.3, $\mathbf{b}(x)$ defined by (5.19)

n_c	h_c/h_f	$norm$	$\varepsilon = 1$	$\varepsilon = 10^{-2}$	$\varepsilon = 10^{-3}$
40	1	L^∞	$0.214.10^{-4}$	$0.207.10^{0}$	$0.623.10^{-2}$
		L^2	$0.840.10^{-5}$	$0.241.10^{-1}$	$0.148.10^{-2}$
		H^1	$0.489.10^{-4}$	$0.842.10^{0}$	$0.106.10^{-1}$
	3	L^∞	$0.263.10^{-2}$	$0.580.10^{-1}$	$0.252.10^{-1}$
		L^2	$0.111.10^{-2}$	$0.807.10^{-2}$	$0.722.10^{-2}$
		H^1	$0.633.10^{-2}$	$0.478.10^{0}$	$0.506.10^{0}$
80	1	L^∞	$0.565.10^{-5}$	$0.215.10^{0}$	$0.624.10^{-2}$
		L^2	$0.222.10^{-5}$	$0.195.10^{-1}$	$0.941.10^{-3}$
		H^1	$0.138.10^{-4}$	$0.893.10^{0}$	$0.360.10^{-1}$

a coefficient $a(x) = \varepsilon$ and two different velocity fields. First is (5.19) and second is

$$b_1 = 2y(1-x^2) + 0.1x\,, \quad b_2 = -2x(1-y^2) + 0.1y\,. \tag{5.20}$$

We refine in the strip along the boundary layer $\Omega_2 = \{0.7 \le x \le 1\,,\ 0 \le y \le 1\}$. The objective is to compare the behavior of the finite difference scheme (**MUDS**) with and without refinement. We report the discrete L^∞, L^2 and H^1 norm in the first, second and third row in Tables 5.3 and 5.4 correspondingly. For mildly dominated convection ($\varepsilon = 10^{-2}$) the scheme with local refinement shows better accuracy for both velocity fields.

5.5 Appendix A

We prove Lemma 5.2 for the case shown on Fig. 5.3.

Table 5.4: Problem 5.3, $\mathbf{b}(x)$ defined by (5.20)

n_c	h_c/h_f	*norm*	$\varepsilon = 1$	$\varepsilon = 10^{-2}$	$\varepsilon = 10^{-3}$
40	1	L^∞	$0.244.10^{-2}$	$0.166.10^{0}$	$0.125.10^{0}$
		L^2	$0.121.10^{-2}$	$0.311.10^{-1}$	$0.211.10^{-1}$
		H^1	$0.536.10^{-2}$	$0.377.10^{0}$	$0.363.10^{0}$
	3	L^∞	$0.216.10^{-2}$	$0.417.10^{-1}$	$0.316.10^{0}$
		L^2	$0.111.10^{-2}$	$0.115.10^{-1}$	$0.281.10^{-1}$
		H^1	$0.600.10^{-2}$	$0.110.10^{0}$	$0.179.10^{+1}$
80	1	L^∞	$0.123.10^{-2}$	$0.698.10^{-1}$	$0.247.10^{-1}$
		L^2	$0.615.10^{-3}$	$0.143.10^{-1}$	$0.232.10^{-1}$
		H^1	$0.280.10^{-2}$	$0.175.10^{0}$	$0.132.10^{+1}$

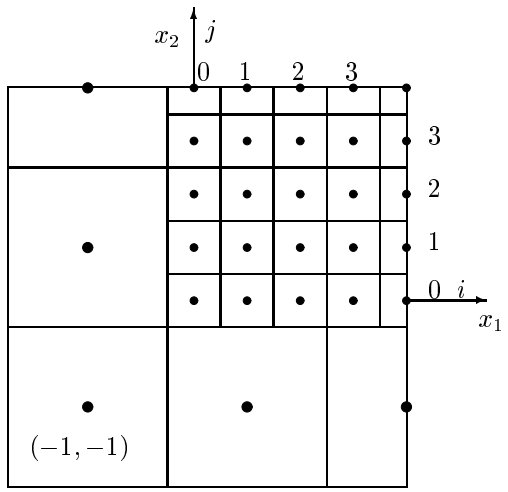


Figure 5.3: Example of a composite cell-centered mesh

Consider the following inner product

$$\begin{aligned} p^T A_0 q &= \sum_{x\in\omega} p(x) \sum_{l=1}^{2} \left[\left(w_l^+(x) - w_l(x)\right) + \left(v_l^+(x) - v_l(x)\right) \right] \\ &= \sum_{l=1}^{2} \sum_{x\in\omega} p(x) \left[\left(w_l^+(x) - w_l(x)\right) + \left(v_l^+(x) - v_l(x)\right) \right] \\ &= \sum_{l=1}^{2} I_l. \end{aligned}$$

We represent the term I_1 for the case of Figure 5.3 in the form

$$I_1 = \sum_{j<0} \sum_{\forall i} + \sum_{j>0} \sum_{i<0} + \sum_{j\geq 0} \sum_{i\geq 0}$$

or

$$I_1 = A_1 + A_2 + B_1 + B_2 + C_1 + C_2$$

$$(.)_1 = \sum_{l=1}^{2} w_l^+ - w_l \,,\ (.)_2 = \sum_{l=1}^{2} v_l^+ - v_l \,.$$

Expressions for A_1, B_1 and C_1 were derived in [43]:

$$\begin{aligned} A_1 &= \sum_{j<0} \sum_{\forall i} \left(w_{1,i,j}^+ - w_{1,i,j}\right) p_{i,j} = -\sum_{j<0} \sum_{\forall i} w_{1,i,j} \overline{\triangle}_1 p_{i,j} \,, \\ B_1 &= \sum_{j>0} \sum_{i<0} \left(w_{1,i,j}^+ - w_{1,i,j}\right) p_{i,j} = \sum_{j>0} \left[\sum_{i<0} w_{1,i,j} \overline{\triangle}_1 p_{i,j} - w_{1,-1,j}^+ p_{-1,j} \right] , \\ C_1 &= \sum_{j\geq 0} \sum_{i\geq 0} \left(w_{1,i,j}^+ - w_{1,i,j}\right) p_{i,j} = \sum_{j\geq 0} \left[\sum_{i\geq 0} w_{1,i,j} \overline{\triangle}_1 p_{i,j} - w_{1,0,j} p_{0,j} \right] . \end{aligned}$$

For A_2 we have

$$\begin{aligned} A_2 &= \sum_{j<0} \sum_{\forall i} \left[\left(B_{1,i,j}^+ - |B_{1,i,j}^+|\right) q_{i+1,j} + \left(B_{1,i,j}^+ + |B_{1,i,j}^+|\right) q_{i,j} \right] p_{i,j} \\ &\quad - \sum_{j<0} \sum_{\forall i} \left[\left(B_{1,i,j} - |B_{1,i,j}|\right) q_{i,j} + \left(B_{1,i,j} + |B_{1,i,j}|\right) q_{i-1,j} \right] p_{i,j} \end{aligned}$$

and after using partial summation we get [78]

$$\begin{aligned} A_2 &= \sum_{j<0} \sum_{\forall i} |B_1(i,j)| \overline{\triangle}_1 q_{i,j} \overline{\triangle}_1 p_{i,j} \\ &\quad + \sum_{j<0} \sum_{\forall i} B_{1,i,j} \left(p_{i,j} \overline{\triangle}_1 q_{i,j} - q_{i,j} \overline{\triangle}_1 p_{i,j} \right) + \sum_{j<0} \sum_{\forall i} \left(B_{1,i,j}^+ - B_{1,i,j}\right) q_{i,j} p_{i,j} \,. \end{aligned}$$

In the same way

$$\begin{aligned} B_2 &= \sum_{j>0} \sum_{i<0} \left[v_{1,i,j}^+ - v_{1,i,j} \right] p_{i,j} = \sum_{j>0} \left[v_{1,-1,j}^+ - v_{1,-1,j} \right] p_{-1,j} \\ &\quad + \sum_{j>0} \sum_{i<-1} |B_{1,i,j}| \overline{\triangle}_1 q_{i,j} \overline{\triangle}_1 p_{i,j} - \sum_{j>0} |B_{1,-2,j}^+| \left(q_{-1,j} - q_{-2,j}\right) p_{-2,j} \\ &\quad + \sum_{j>0} \sum_{i<-1} B_{1,i,j} \left(p_{i,j} \overline{\triangle}_1 q_{i,j} - q_{i,j} \overline{\triangle}_1 p_{i,j} \right) + \sum_{j>0} B_{1,-2,j}^+ q_{-1,j} p_{-2,j} \\ &\quad + \sum_{j>0} \sum_{i<-1} \left(B_{1,i,j}^+ - B_{1,i,j}\right) q_{i,j} p_{i,j} \,. \end{aligned}$$

Using the fact that $B_{1,-1,j} = B^+_{1,-2,j}$ we finally get

$$\begin{aligned} B_2 &= \sum_{j>0}\sum_{i<0} |B_{1,i,j}|\overline{\triangle}_1 q_{i,j}\overline{\triangle}_1 p_{i,j} + \sum_{j>0}\sum_{i<0} B_{1,i,j}\left(p_{i,j}\overline{\triangle}_1 q_{i,j} - q_{i,j}\overline{\triangle}_1 p_{i,j}\right) \\ &\quad + \sum_{j>0}\sum_{i<-1}\left(B^+_{1,i,j} - B_{1,i,j}\right) q_{i,j}p_{i,j} + \sum_{j>0} v^+_{1,-1,j}p_{-1,j} - \sum_{j>0} B_{1,-1,j}q_{-1,j}p_{-1,j}\,. \end{aligned}$$

Expression for C_2 is derived similarly

$$\begin{aligned} C_2 &= \sum_{j\geq 0}\sum_{i\geq 0} |B_{1,i,j}\overline{\triangle}_1 q_{i,j}\overline{\triangle}_1 p_{i,j} + \sum_{j\geq 0}\sum_{i\geq 0} B_{1,i,j}\left(p_{i,j}\overline{\triangle}_1 q_{i,j} - q_{i,j}\overline{\triangle}_1 p_{i,j}\right) \\ &\quad + \sum_{j\geq 0}\sum_{i\geq 0}\left(B^+_{1,i,j} - B_{1,i,j}\right) q_{i,j}p_{i,j} + \sum_{j\geq 0}\left(B^+_{1,0,j}q_{0,j}p_{0,j} - v_{1,0,j}p_{0,j}\right) \end{aligned}$$

Summarizing these results and taking into account the equalities

$$\begin{aligned} v^+_{1,-1,j+1} &= v_{1,0,j} + v_{1,0,j+1} + v_{1,0,j+2} \\ B^+_{1,-1,j+1} &= B_{1,0,j} + B_{1,0,j+1} + B_{1,0,j+2} \end{aligned}$$

we get the assertion of the lemma.

5.6 Appendix B

Here we investigate the local truncation errors μ_1, μ_2 in the irregular points and prove the inequality (5.18). For the component $\mu_1(x)$ we have

$$\begin{aligned} \mu_1(x) &= \int_{x_{2,j}+(l-0.5)h}^{x_{2,j}+(l+0.5)h} b_1(x_{1,i-1/2},s)u(x_{1,i-1/2},s)\,ds \\ &\quad - \left(\frac{b_{1,i-1/2,j+l}h_f}{2} - \frac{|b_{1,i-1/2,j+l}|h_f}{2}\right) u_{i,j+l} \\ &\quad - \left(\frac{b_{1,i-1/2,j+l}h_f}{2} + \frac{|b_{1,i-1/2,j+l}|h_f}{2}\right) u_{i-1,j+1} \end{aligned} \tag{5.21}$$

Using the equality

$$\begin{aligned} &\left(\frac{b_{1,i-1/2,j+l}}{2} - \frac{|b_{1,i-1/2,j+l}|}{2}\right) u_{i,j+l} + \left(\frac{b_{1,i-1/2,j+l}}{2} + \frac{|b_{1,i-1/2,j+l}|}{2}\right) u_{i-1,j+1} \\ &\quad = b_{1,i-1/2,j+l}\left(\frac{3u_{i,j+l}+u_{i-1,j+1}}{4}\right) - \left(\frac{b_{1,i-1/2,j+l}}{4} + \frac{|b_{1,i-1/2,j+l}|}{2}\right)\overline{\triangle}_1 u_{i,j+l} \\ &\quad = b_{1,i-1/2,j+l}\left(\frac{3u_{i,j+l}+u_{i-1,j+l}}{4}\right) + b_{1,i-1/2,j+l}\left(\frac{u_{i-1,j+1}-u_{i-1,j+l}}{4}\right) \\ &\quad - \left(\frac{b_{1,i-1/2,j+l}}{4} + \frac{|b_{1,i-1/2,j+l}|}{2}\right)\overline{\triangle}_1 u_{i,j+l} \end{aligned}$$

we represent formulae (5.21) in the form

$$\begin{aligned} \mu_1(x) &= \left[\int_{x_{2,j}+(l-0.5)h}^{x_{2,j}+(l+0.5)h} b_1(x_{1,i-1/2},s)u(x_{1,i-1/2},s)\,ds\right. \\ &\quad \left. - b_{1,i-1/2,j+l}h_f\left(\frac{3u_{i,j+l}+u_{i-1,j+l}}{4}\right)\right] \\ &\quad - b_{1,i-1/2,j+l}h_f\left(\frac{u_{i-1,j+1}-u_{i-1,j+l}}{4}\right) \\ &\quad + \left(\frac{b_{1,i-1/2,j+l}h_f}{4} + \frac{|b_{1,i-1/2,j+l}|h_f}{2}\right)\overline{\triangle}_1 u_{i,j+l}\,. \end{aligned} \tag{5.22}$$

Thus yields

$$\begin{aligned} |\mu_1(x)| &\leq |l(b_1,u)| + \frac{3h_f}{4}|b_{1,i-1/2,j+l}||\overline{\triangle}_1 u_{i,j+l}| \\ &\quad + \frac{h_f}{4}|b_{1,i-1/2,j+l}|\,|u_{i-1,j+1} - u_{i-1,j+l}|\ , \end{aligned} \tag{5.23}$$

where the bilinear functional $l(b_1,u)$ is defined by

$$\begin{aligned} l(b_1,u) &= \int_{x_{2,j}+(l-0.5)h}^{x_{2,j}+(l+0.5)h} b_1(x_{1,i-1/2},s)u(x_{1,i-1/2},s)\,ds \\ &\quad - b_1(x_{1,i-1/2},x_{2,j+l})h_f\left(\frac{3u_{i,j+l}+u_{i-1,j+l}}{4}\right)\ . \end{aligned} \tag{5.24}$$

We consider $u_{i,j+l} - u_{i-1,j+1}$ as a linear functional of u for a fixed $x \in \omega^+$. This functional is bounded in $H^m(\overline{e})$, $1 < m \leq 3$ and vanishes for all polynomials of zero degree. Therefore, by the corollary of the Bramble-Hilbert lemma we get

$$|u_{i,j+l} - u_{i-1,j+1}| \leq C(|u|_{1,\overline{e}} + h^{m-1}|u|_{m,\overline{e}})\,,\ 1 < m \leq 3. \tag{5.25}$$

Hence for the second term in the inequality (5.23) we get

$$\frac{3h_f}{4}|b_{1,i-1/2,j+l}||\overline{\triangle}_1 u_{i,j+l}| \leq Ch|b_1|_{0,\infty,\Omega}(|u|_{1,\overline{e}} + h^{m-1}|u|_{m,\overline{e}})\,,\ 1 < m \leq 3.$$

Similarly we estimate the third term in (5.23) by

$$\begin{aligned} &\frac{h_f}{4}|b_{1,i-1/2,j+l}|\,|u_{i-1,j+1} - u_{i-1,j+l}| \\ &\quad \leq Ch|b_1|_{0,\infty,\Omega}(|u|_{1,\overline{e}} + h^{m-1}|u|_{m,\overline{e}})\,,\ 1 < m \leq 3. \end{aligned}$$

The functional $l(b_1,u)$ is estimated in the following lemma, proved in Lazarov, Mishev and Vassilevski [78].

Lemma 5.4 *If the solution of problem (2.11) is H^m-regular, $1 < m$, then for the bilinear functional $l(b_1,u)$ defined by (5.24) the following estimate is valid:*

$$|l(b_1,u)| \leq Ch^m\|b_1\|_{1,\infty,\Omega}\|u\|_{m,\overline{e}},\ 1 < m \leq 2.$$

Above remarks give us the upper bound for $|\mu_1(x)|$ which coincides with the estimates (5.15) for the regular points.

CHAPTER VI

FINITE VOLUME ELEMENT METHODS FOR NONSYMMETRIC PROBLEMS

In this chapter we generalize the results by Cai and McCormick [26, 28] and Jianguo and Shitong [66] for 2–D symmetric problems to 2–D(3–D) nonsymmetric ones. We prove the stability and error estimates for both diffusion and convection dominated cases. For the diffusion dominated case we show that the *inf–sup* condition is satisfied. The upwind finite volume element method is analyzed in the framework of the theory developed in Chapter IV. We note that the error estimates for the diffusion dominated case can be proven with the same technique, but for us the inf–sup condition approach seems more elegant (or at least a little different).

Our theory assumes barycentric control volumes. All the necessary notations are introduced in Chapter III.

We recall the formulation of the discrete finite volume element method:

Find $u_h \in \mathcal{V}_0^h$ such that, for all vertex–centered control volumes V_i, $i = 1, \ldots, n_P$

$$\int_{\partial V_i} (-A\nabla u_h + \mathbf{b}u_h, \mathbf{n})\, ds = \int_{V_i} f\, dx\,, \tag{6.1}$$

or as a Petrov–Galerkin method:

Find $u_h \in \mathcal{V}_0^h$ such that

$$B_h(u_h, v_h) = f(v_h) \qquad \forall\, v_h \in \mathcal{W}^h\,, \tag{6.2}$$

where

$$f(v_h) = \sum_{x_i \in \omega_P} \int_{V_i} f\, \mathrm{dx}\, v(x_i)\,.$$

and $B_h(.,.)$, $B_h^{(2)}(.,.)$ and $B_h^{(1)}(.,.)$ are bilinear form in $\mathcal{V}_0^h \times \mathcal{W}^h$

$$B_h(u,v) = B_h^{(2)}(u,v) + B_h^{(1)}(u,v)\,, \tag{6.3a}$$

$$B_h^{(2)}(u,v) = -\sum_{x_i \in \omega_P} \int_{\partial V_i} (A\nabla u, \mathbf{n})\, ds\; v(x_i)\,, \tag{6.3b}$$

$$B_h^{(1)}(u,v) = \sum_{x_i \in \omega_P} \int_{\partial V_i} (\mathbf{b}, \mathbf{n}) u\, ds\; v(x_i)\,. \tag{6.3c}$$

We also need the finite element bilinear forms in $\mathcal{V}_0^h \times \mathcal{V}_0^h$

$$A(u,v) = A^{(2)}(u,v) + A^{(1)}(u,v)\,, \tag{6.4a}$$

$$A^{(2)}(u,v) = \int_\Omega (A\nabla u, \nabla v)\, dx\,, \tag{6.4b}$$

$$A^{(1)}(u,v) = -\int_\Omega (\mathbf{b}, \nabla v) u\, dx\,. \tag{6.4c}$$

6.1 Diffusion dominated problem

First we elaborate the finite volume element theory for the compact perturbation of symmetric problem. We use Theorem 2.11 to prove uniqueness and existence of the solution of (6.2). In

order to use Remark 2.2 we have to show that the following inequalities hold:

$$(i) \quad |B_h(u_h, v_h)| \quad \leq C\|u_h\|_{1,\omega_P}\|v_h\|_{1,B}\,, \tag{6.5a}$$

$$(ii) \quad |B_h(u_h, I_h^c u_h)| \geq \alpha\|u_h\|_{1,\omega_P}\|I_h^c u_h\|_{1,B}\,, \tag{6.5b}$$

$$(iii) \quad |B_h(I_h^l u_h, u_h)| > 0\,. \tag{6.5c}$$

We prove (6.5) via comparing with the bilinear forms for the finite element method (6.4b) and (6.4c). First we prove some auxiliary results.

Proposition 6.1 *Suppose that the matrix A has piecewise constant entries. Then the following equality holds:*

$$A^{(2)}(v_h, v_h) = B_h^{(2)}(v_h, I_h^c v_h) \qquad \forall\, v \in \mathcal{V}_0^h\,.$$

Proof: We have to show

$$\sum_{x_i \in \omega_P} \int_{\partial V_i} -(A\nabla v, \mathbf{n})\, ds\, v(x_i) = \int_\Omega (A\nabla v, \nabla v)\, dx\,,$$

or

$$\sum_{x_i \in \omega_P} v_i \sum_{x_j \in \omega_P} v_j \left[\int_{\partial V_i} -(A\nabla\varphi_j, \mathbf{n})\, ds\right] = \sum_{x_i \in \omega_P} v_i \sum_{x_j \in \omega_P} v_j \left[\int_\Omega (A\nabla\varphi_j, \nabla\varphi_i)\, dx\right]\,, \tag{6.6}$$

where φ_j and φ_i are linear basis functions. Hence it suffices to prove only that

$$\int_{\partial V_i} -(A\nabla\varphi_j, \mathbf{n})\, ds = \int_\Omega (A\nabla\varphi_j, \nabla\varphi_i)\, dx\,.$$

Consider one tetrahedral finite element K with vertexes x_i, x_j, x_k and x_l and the points on edges M_{ij}, M_{ik}, M_{il} and points on the faces M_{ijk}, M_{ikl}, M_{ijl} and the barycenter of the tetrahedron S_K (see Fig. 6.1). The equality (6.6) will follow from

$$\int_{K\cap\partial V_i} -(A\nabla\varphi_j, \mathbf{n})\, ds = \int_K (A\nabla\varphi_j, \nabla\varphi_i)\, d\mathrm{x}\,. \tag{6.7}$$

We apply the Green's formulae for the left integral in (6.7) and get

$$\begin{aligned}\int_{K\cap\partial V_i} -(A\nabla\varphi_j, \mathbf{n})\, ds &= \int_{K\cap V_i} -\nabla.(A\nabla\varphi_j)\, d\mathrm{x} \\ &\quad + \int_{\mathcal{P}_{ijk}} (A\nabla\varphi_j, \mathbf{n})\, ds + \int_{\mathcal{P}_{ikl}} (A\nabla\varphi_j, \mathbf{n})\, ds + \int_{\mathcal{P}_{ijl}} (A\nabla\varphi_j, \mathbf{n})\, ds \\ &= \mathrm{meas}(\mathcal{P}_{ijk})(A\nabla\varphi_j, \mathbf{n}_{\mathcal{P}_{ijk}}) + \mathrm{meas}(\mathcal{P}_{ikl})(A\nabla\varphi_j, \mathbf{n}_{\mathcal{P}_{ikl}}) \\ &\quad + \mathrm{meas}(\mathcal{P}_{ijl})(A\nabla\varphi_j, \mathbf{n}_{\mathcal{P}_{ijl}})\,.\end{aligned}$$

Here $\mathcal{P}_{ijk} = x_i M_{ij} M_{ijk} M_{ik}$, $\mathcal{P}_{ikl} = x_i M_{ik} M_{ikl} M_{il}$ and $\mathcal{P}_{ijl} = x_i M_{ij} M_{ijl} M_{il}$. We used that A is a constant in the first line $\left(\int_{K\cap V_i} -\nabla.(A\nabla\varphi_j)\, d\mathrm{x} = 0\right)$ and in the third line of the last chain of equalities.

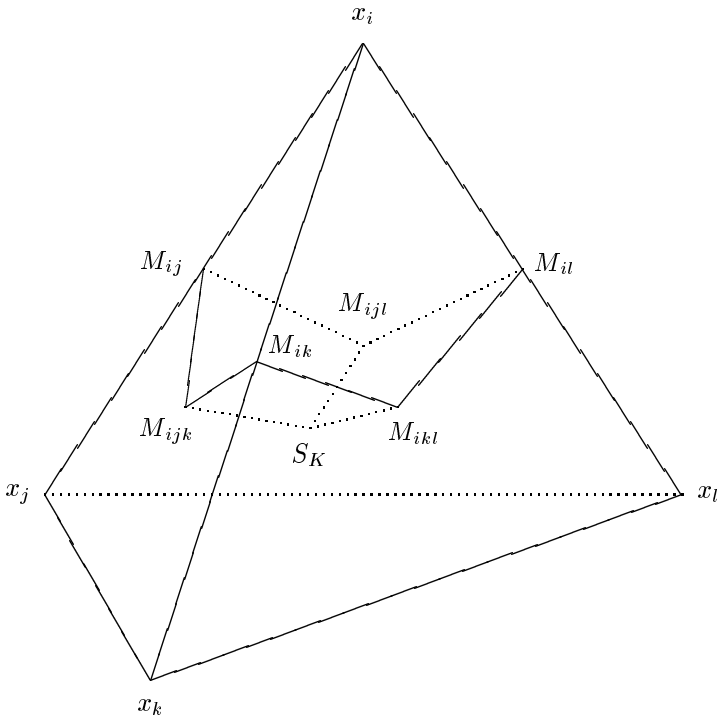


Figure 6.1: Finite element K

We now apply Green's formulae to the right integral in (6.7)

$$\begin{aligned}\int_K (A\nabla\varphi_j, \nabla\varphi_i)\, d\mathrm{x} &= \int_K -\nabla.(A\nabla\varphi_j)\varphi_i\, d\mathrm{x} \\ &\quad + \int_{\partial K} (A\nabla\varphi_j, \mathbf{n})\varphi_i\, ds \\ &= (A\nabla\varphi_j, \mathbf{n}_{\mathcal{P}_{ijk}}) \int_{x_i x_j x_k} \varphi_i\, ds + (A\nabla\varphi_j, \mathbf{n}_{\mathcal{P}_{ikl}}) \int_{x_i x_k x_l} \varphi_i\, ds \\ &\quad + (A\nabla\varphi_j, \mathbf{n}_{\mathcal{P}_{ijl}}) \int_{x_i x_j x_l} \varphi_i\, ds + (A\nabla\varphi_j, \mathbf{n}_{\mathcal{P}_{jkl}}) \int_{x_j x_k x_l} \varphi_i\, ds\end{aligned}$$

Note that $\int_{x_j x_k x_l} \varphi_i\, ds = 0$ and

$$\begin{aligned}\int_{x_i x_j x_k} \varphi_i\, ds &= \frac{1}{3}\,\mathrm{meas}(x_i x_j x_k) = \mathrm{meas}(\mathcal{P}_{ijk})\,, \\ \int_{x_i x_k x_l} \varphi_i\, ds &= \frac{1}{3}\,\mathrm{meas}(x_i x_k x_l) = \mathrm{meas}(\mathcal{P}_{ikl})\,, \\ \int_{x_i x_j x_l} \varphi_i\, ds &= \frac{1}{3}\,\mathrm{meas}(x_i x_j x_l) = \mathrm{meas}(\mathcal{P}_{ijl})\end{aligned}$$

since the points M_{ijk}, M_{ikl} and M_{ijl} are barycenters of the corresponding faces. □

The fact that the linear finite element method and finite volume element method coincide for constant coefficient tensor A was extensively used by Hackbusch [55].

For completeness we prove the following lemma due to H. Jianguo and X. Shitong [66].

Lemma 6.1 *For every $v \in \mathcal{V}_0^h$ the following estimate holds:*

$$|B_h^{(2)}(v, I_h^c v) - A^{(2)}(v, v)| \le Ch\|A\|_{1,\infty.\Omega}|v|_{1,\Omega}^2\,.$$

Proof: Define the L^2 projection $\hat{A}$ of A

$$\hat{A}_{ij|K} = \frac{1}{\mathrm{meas}(K)} \int_K A_{ij}(x)\, dx \quad 1 \le i,\, j \le 3, K \in T_h\,.$$

Then

$$\begin{aligned}|B_h^{(2)}(v, I_h^c v) - A^{(2)}(v, v)| &= \left| \sum_{x_i \in \omega_P} \int_{\partial V_i} ((A - \hat{A})\nabla v, \mathbf{n})\, ds\, v_i - \int_\Omega ((A - \hat{A})\nabla v, \nabla v)\, dx \right| \\ &\le \left| \sum_{x_i \in \omega_P} \int_{\partial V_i} ((A - \hat{A})\nabla v, \mathbf{n}) ds\, v_i \right| + \left| \int_\Omega ((A - \hat{A})\nabla v, \nabla v) dx \right| \\ &= I_1 + I_2\end{aligned}$$

For I_2 we immediately obtain

$$I_2 \le Ch\|A\|_{1,\infty,\Omega}|v|_{1,\Omega}^2\,.$$

The first term is written as follows:

$$\begin{aligned} I_1 &= \left| \sum_{x_i \in \omega_P} \sum_{j \in \Pi(i)} \int_{\gamma_{ij}} ((A - \hat{A})\nabla v, \mathbf{n})\, ds\, v_i \right| \\ &= \frac{1}{2} \left| \sum_{x_i \in \omega_P} \sum_{j \in \Pi(i)} \left[\int_{\gamma_{ij}} ((A - \hat{A})\nabla v, \mathbf{n})\, ds\, v_i + \int_{\gamma_{ji}} ((A - \hat{A})\nabla v, \mathbf{n})\, ds\, v_j \right] \right| \\ &= \frac{1}{2} \left| \sum_{x_i \in \omega_P} \sum_{j \in \Pi(i)} \int_{\gamma_{ij}} ((A - \hat{A})\nabla v, \mathbf{n})\, ds\, (v_i - v_j) \right| \\ &\leq \frac{1}{2} \sum_{x_i \in \omega_P} \sum_{j \in \Pi(i)} \int_{\gamma_{ij}} \left| (A - \hat{A})\nabla v \right| \, ds\, |v_i - v_j| \,. \end{aligned}$$

The integral on γ_{ij} is estimated on each finite element K that has nonempty intersection with $\gamma_{i,j}$

$$\int_{\gamma_{ij}} \left| (A - \hat{A})\nabla v \right| \, ds = \sum_K \int_{\gamma_{ij} \cap K} \left| (A - \hat{A})\nabla v \right| \, ds$$

and

$$\begin{aligned} \int_{\gamma_{ij} \cap K} \left| (A - \hat{A})\nabla v \right| \, ds &\leq \|A - \hat{A}\|_{0,\infty,\gamma_{ij} \cap K} \int_{\gamma_{ij} \cap K} |\nabla v| \, ds \\ &\leq C_1 h_K \|A\|_{1,\infty,K} (\mathrm{meas}(\gamma_{ij} \cap K))^{1/2} |v|_{1,\gamma_{ij}} \\ &\leq C_2 h \|A\|_{1,\infty,\Omega} h_K^{(d-1)/2} |v|_{3/2,K} \end{aligned}$$

Since v is a linear polynomial in K we have that $|v|_{3/2,K} = |v|_{1,K}$. Therefore,

$$\int_{\gamma_{ij}} \left| (A - \hat{A})\nabla v \right| \, ds \leq C h^{3/2} \|A\|_{1,\infty,\Omega} h^{(d-2)/2} |v|_{1,K} \,.$$

We have used the trace theorem (Theorem 2.5) and the fact that $\mathrm{meas}(\gamma_{ij} \cap K) = O(h^{(d-1)/2})$.

Applying Cauchy–Schwartz inequality we get

$$\begin{aligned} I_1 &\leq C h^{3/2} \|A\|_{1,\infty,\Omega} |v|_{1,\Omega} \left[\sum_{x_i \in \omega_P} h_K^d \sum_{j \in \Pi(i)} \left(\frac{v_i - v_j}{h_K} \right)^2 \right]^{1/2} \\ &\leq C_1 h^{3/2} \|A\|_{1,\infty,\Omega} |v|_{1,\Omega} \left[\sum_{x_i \in \omega_P} \mathrm{meas}(V_i) \sum_{j \in \Pi(i)} \left(\frac{v_i - v_j}{dist(x_i, x_j)} \right)^2 \right]^{1/2} \\ &\leq C h \|A\|_{1,\infty,\Omega} |v|_{1,\Omega} |v|_{1,\omega_P} \,. \end{aligned}$$

Now the result follows from Lemma 3.1. □

We wrote the proof of the last estimate in details in order to point out the importance of the regularity of the finite element triangulation. We compare $B_h^{(1)}(v, I_h^c v)$ and $A^{(1)}(v, v)$ in the following lemma.

Lemma 6.2 *For every $v \in \mathcal{V}_0^h$ the following estimate holds:*

$$|B_h^{(1)}(v, I_h^c v) - A^{(1)}(v, v)| \leq C h \|\mathbf{b}\|_{1,\infty,\Omega} |v|_{1,\Omega}^2 \,.$$

Proof: Consider the contribution of one particular element K in the computation of $B_h^{(1)}(v, I_h^c v)$ corresponding to the i^{th} node

$$\begin{aligned}\int_{\partial V_i \cap K} (\mathbf{b}.\mathbf{n}) v \, ds \, v_i &= \left[\int_{(\partial V_i \cap K) \cup M_i} (\mathbf{b}.\mathbf{n}) v \, ds - \int_{M_i} (\mathbf{b}.\mathbf{n}) v \, ds \right] v_i \\ &= \int_{V_i \cap K} \operatorname{div}(\mathbf{b}v) \, d\mathrm{x} \, v_i - \int_{M_i} (\mathbf{b}.\mathbf{n}) v \, ds \, v_i \\ &= \int_K \operatorname{div}(\mathbf{b}v) v_i \hat{\varphi}_i \, d\mathrm{x} - \int_{\partial K} (\mathbf{b}.\mathbf{n}) v v_i \hat{\varphi}_i \, ds \,,\end{aligned}$$

where $M_i = \partial K \cap V_i$ and $\hat{\varphi}_i$ is a basis function in $\mathcal{W}^h$ corresponding to the i^{th} node. Then, the contribution of the element K is equal to

$$B_h^{(1)}(v, I_h^c v)_{|K} = \int_K \operatorname{div}(\mathbf{b}v) I_h^c v \, d\mathrm{x} - \int_{\partial K} (\mathbf{b}.\mathbf{n}) v I_h^c v \, ds$$

and

$$B_h^{(1)}(v, I_h^c v) = \sum_{K \in T_h} \int_K \operatorname{div}(\mathbf{b}v) I_h^c v \, d\mathrm{x} \,.$$

because the surface integrals vanish. Therefore,

$$\begin{aligned}|B_h^{(1)}(v, I_h^c v) - A^{(1)}(v, v)| &\leq \sum_{K \in T_h} \left| \int_K \operatorname{div}(\mathbf{b}v)(I_h^c v - v) \, d\mathrm{x} \right| \\ &\leq \|\mathbf{b}\|_{1,\infty,\Omega} \sum_{K \in T_h} |v|_{1,K} \|v - I_h^c v\|_{0,K} \\ &\leq Ch \|\mathbf{b}\|_{1,\infty,\Omega} |v|_{1,\Omega}^2\end{aligned}$$

by Corollary 3.1. □

Using Lemmas 6.1 and 6.2 we easily prove the following theorem.

Theorem 6.1 *There exists h_0 such that for any $h < h_0$ the problem* (6.2) *has one and only one solution and the following stability estimates holds:*

$$|u_h|_{1,\omega_P} \leq C \|f\|_{-1,B} \,.$$

Proof: From the continuity and coercivity of the bilinear form $A(.,.)$ (cf. Chapter II) and Lemmas 6.1 and 6.2 follows that there exists h_0 and positive constants C_0 and C_1 such that for $h < h_0$ the inequalities hold

$$C_0 A(v, v) \leq B(v, I_h^c v) \leq C_1 A(v, v) \,.$$

The continuity of $B_h(.,.)$ (6.5a) and inf–sup condition (6.5b) are consequence of the equivalence of the norms $|.|_{1,B}$ and $|.|_{1,\omega_P}$ (Lemma 3.1) and the fact that $|.|_{1,\omega_P}$ coincides with $|.|_{1,\Omega}$ for piecewise linear functions. The inequality (6.5c) follows from the observation $B_h(I_h^l v_h, v_h) = B_h(I_h^l v_h, I_h^c(I_h^l v_h))$. □

Note that we have shown the inequality

$$C|v|_{1,\Omega}^2 \leq B_h(v, I_h^c v) \qquad \forall \, v \in \mathcal{V}_0^h \,. \tag{6.8}$$

Now, we are ready to prove our main result.

Theorem 6.2 *Let u denote the solution of (2.11) and u_h be the solution of FVE (6.1). Then we have the following estimate*

$$\|u - u_h\|_{1,\Omega} \leq Ch(\|A\|_{0,\infty,\Omega} + h\|\mathbf{b}\|_{0,\infty,\Omega})|u|_{2,\Omega}\,.$$

Proof: First, we establish an a priori estimate

$$|I_h^l u - u_h|_{1,\Omega} \leq \frac{1}{C}\left(\sum_{x_i\in\omega_P}\sum_{j\in\Pi(i)} k_{i,j}^2\left(l_{ij}(I_h^l u - u)\right)^2\right)^{1/2}, \tag{6.9}$$

where

$$l_{i,j}(v) = \int_{\gamma_{ij}} (-A\nabla v + \mathbf{b}v).\mathbf{n}\,ds \quad \text{and} \quad k_{i,j}^2 = \frac{dist(x_i,x_j)^2}{\text{meas}(V_i)}\,.$$

We estimate $w = I_h^l u - u_h$. Note that

$$B_h(u, v_h) = B_h(u_h, v_h)\,, \tag{6.10}$$

where u is the solution of (2.11) and u_h is the solution of (6.2) and therefore,

$$B_h(I_h^l u - u, v_h) = B_h(I_h^l u - u_h, v_h)\,. \tag{6.11}$$

Combining (6.8) and (6.11) we have

$$\begin{aligned}
C|w|_{1,\Omega}^2 &\leq B_h(w, I_h^c w) = B_h(I_h^l u - u, I_h^c w)\\
&= \sum_{x_i\in\omega_P}\sum_{j\in\Pi(i)}\int_{\partial V_i}(-A\nabla(I_h^l u - u) + \mathbf{b}(I_h^l u - u)).\mathbf{n}\,ds\, w_i\\
&= \frac{1}{2}\sum_{x_i\in\omega_P}\sum_{j\in\Pi(i)}\left[l_{ij}(I_h^l u - u)w_i + l_{ji}(I_h^l u - u)w_j\right]\\
&= \frac{1}{2}\sum_{x_i\in\omega_P}\sum_{j\in\Pi(i)} l_{ij}(I_h^l u - u)(w_i - w_j)\\
&\leq C\left(\sum_{x_i\in\omega_P}\sum_{j\in\Pi(i)} k_{i,j}^2\,|l_{i,j}|^2\right)^{1/2}\left(\sum_{x_i\in\omega_P}\text{meas}(V_i)\sum_{j\in\Pi(i)}\left(\frac{w_j - w_i}{dist(x_i,x_j)}\right)^2\right)^{1/2}.
\end{aligned}$$

Now the a priori estimate (6.9) follows from the equivalence of $|.|_{1,\omega_P}$ and $|.|_{1,\Omega}$ for linear polynomials.

We have to estimate the functional $|l_{ij}|$. Define the linear functionals

$$\begin{aligned}
f_a(u) &= -\int_{\gamma_{ij}} A\nabla(I_h^l u - u).\mathbf{n}\,ds\,,\\
f_b(u) &= \int_{\gamma_{ij}} \mathbf{b}(I_h^l u - u).\mathbf{n}\,ds\,.
\end{aligned}$$

First, we estimate $|f_a()|$

$$\begin{aligned}
|f_a(u)| = |f_{\tilde{a}}(\tilde{u})| &= \left|\int_{\tilde{\gamma}_{ij}} |\det J|\left(\tilde{A}J^{-T}\nabla(I_h^l\tilde{u} - \tilde{u}).J^{-T}\tilde{n}\right)\,d\tilde{s}\right|\\
&\leq \|A\|_{0,\infty,\gamma_{ij}}.\|J^{-1}\|^2.|\det J|.|\tilde{\gamma}_{ij}|.\|\tilde{u}\|_{2,\bar{K}}\\
&\leq C\|A\|_{0,\infty,\Omega}\|J\|^2\|J^{-1}\|^2.|\det J|^{1/2}.|u|_{2,K}\\
&\leq Ch^{d/2}\|A\|_{0,\infty,\Omega}|u|_{2,K}\,.
\end{aligned}$$

Similarly , for $|f_b(u)|$ we have

$$\begin{aligned} |f_b(u)| = |f_{\tilde{b}}(\tilde{u})| &= \left| \int_{\tilde{\gamma}_{ij}} |\det J| \left(\tilde{\mathbf{b}}(I_h^l \tilde{u} - \tilde{u}).J^{-T}\tilde{n} \right) d\tilde{s} \right| \\ &\leq \|J^{-1}\|.|\det J|.\|\mathbf{b}\|_{0,\infty,\bar{K}}.|\tilde{\gamma}_{ij}|.\|\tilde{u}\|_{2,\bar{K}} \\ &\leq C\|J\|^2\|J^{-1}\||\det J|^{1/2}\|\mathbf{b}\|_{0,\infty,\Omega}|u|_{2,K} \\ &\leq Ch^{d/2+1}\|\mathbf{b}\|_{0,\infty,\Omega}|u|_{2,K} \end{aligned}$$

Taking into account that $k_{i,j} = O(h^{2-d})$ we find that

$$\begin{aligned} |I_h^l u - u_h|_{1,\omega_P} &\leq C \left(\sum_{x_i \in \omega_P} \sum_{j \in \Pi(i)} k_{i,j}^2 |l_{i,j}|^2 \right)^{1/2} \\ &\leq C_1 h^{1-d/2} h^{d/2} \left(\sum_{x_i \in \omega_P} \sum_{j \in \Pi(i)} \|A\|_{0,\infty,\Omega}^2 |u|_{2,K}^2 + h^2 \|\mathbf{b}\|_{0,\infty,\Omega}^2 |u|_{2,K}^2 \right)^{1/2} \\ &\leq Ch(\|A\|_{0,\infty,\Omega} + h\|\mathbf{b}\|_{0,\infty,\Omega})|u|_{2,\Omega}\,. \end{aligned}$$

Finally the result follows from the triangle inequality

$$|u - u_h|_{1,\Omega} \leq |u - I_h^l u|_{1,\Omega} + |I_h^l u - u_h|_{1,\Omega}$$

and the estimate (3.12) for the linear interpolant. □

6.2 Upwind finite volume element method

In this section we modify the definition of $B_h(.,.)$ (6.3c) in order to obtain a stable approximation. This means that the equality (6.10) will not be satisfied anymore. The upwind approximation of the convection term can be considered as a quadrature formulae applied to (6.3c) and estimated in a similar way as in the paper by Cai [26]. We will use the technique developed in Chapter IV in order to prove stability and convergence estimate.

We redefine the problem (6.2) into a matrix form

$$\mathcal{B}_h u_h = \phi\,, \tag{6.12}$$

where the entries ϕ_i of the right hand side are defined by $\phi_i = \frac{1}{\text{meas}(V_i)} \int_{V_i} f(\mathrm{x})\, d\mathrm{x}$. The matrix $\mathcal{B}_h$ is defined in the following way:

$$\mathcal{B}_h = \mathcal{B}_h^{(2)} + \mathcal{B}_h^{(1)}\,, \tag{6.13a}$$

where

$$\mathcal{B}_h^{(2)} u_h = \frac{1}{\text{meas}(V_i)} \sum_{j \in \Pi(i)} \int_{\gamma_{ij}} -(A\nabla u_h, \mathbf{n})\, ds\,, \quad i = 1, \ldots, n_P \tag{6.13b}$$

$$\mathcal{B}_h^{(1)} u_h = \frac{1}{\text{meas}(V_i)} \sum_{j \in \Pi(i)} (\beta_{ij}^+ u_i + \beta_{ij}^- u_j)\,, \quad i = 1, \ldots, n_P\,. \tag{6.13c}$$

and the "upwind" approximation are given via

$$\beta_{ij}^+ = \frac{\beta_{ij} + |\beta_{ij}|}{2}\,, \qquad \beta_{ij}^- = \frac{\beta_{ij} - |\beta_{ij}|}{2}\,. \tag{6.13d}$$

Let β_{ij} be an approximation of $\int_{\gamma_{ij}} (\mathbf{b}, \mathbf{n})\, ds$ with the properties

$$(i) \quad \beta_{i,j} + \beta_{j,i} = 0\,. \tag{6.14a}$$

$$(ii) \quad |\beta_{i,j}| \le C \operatorname{meas}(\gamma_{ij}) \|\mathbf{b}\|_{d/2+\alpha,\infty,\Omega}\,, \tag{6.14b}$$

$$(iii) \quad \left| \int_{\gamma_{ij}} (\mathbf{b}, \mathbf{n})\; ds - \beta_{i,j} \right| \le C h^{d+\alpha} |\mathbf{b}|_{1+\alpha,\infty,\Omega}\,, \tag{6.14c}$$

where C is a positive constant and $\alpha > 0$.

Remark 6.1 Note that these are the same conditions as (4.7), but on different control volumes. The approximation of the diffusion term is not changed.

Proposition 6.2 *Let the Assumption 4.3 be satisfied, the upwind FVE method be defined by* (6.13) *and the approximations $\beta_{i,j}$ fulfill the conditions* (6.14). *Then the matrix $\mathcal{B}$ of the upwind FVE is a positive real matrix and there exists a constant C such that the following inequality is true:*

$$(\mathcal{B}_h u_h, I_h^c u_h)_B \ge C \|I_h^c u_h\|_{1,B}^2 \;, \text{ for all } u_h \in \mathcal{V}_0^h\,.$$

The constant C depends only on the matrix A and the vector $\mathbf{b}$.

Proof: We point out that by the construction of the inner product $(.,.)_B$ and the matrix $\mathcal{B}_h^{(2)}$

$$(\mathcal{B}_h^{(2)} v, I_h^c v)_B = B_h^{(2)}(v, I_h^c v) \qquad \forall\, v \in \mathcal{V}_0^h\,.$$

Therefore,

$$(\mathcal{B}^{(2)} v, I_h^c v)_B \ge C |v|_{1,\omega_P}^2 \ge C_1 |I_h^c v|_{1,B} \qquad \forall\, v \in \mathcal{V}_0^h\,.$$

For the convection term we have

$$\begin{aligned}
(\mathcal{B}_h^{(1)} u_h, v) &= \frac{1}{2} \sum_{x_i \in \omega_P} \left[\sum_{j \in \Pi(i)} (\beta_{ij} + |\beta_{ij}|) u_i + (\beta_{ij} - |\beta_{ij}|) u_j \right] v_i \\
&= \frac{1}{2} \sum_{x_i \in \omega_P} \left(\sum_{j \in \Pi(i)} \beta_{ij} \right) u_i v_i + \frac{1}{2} \sum_{x_i \in \omega_P} \left(\sum_{j \in \Pi(i)} |\beta_{ij}| (u_i - u_j) \right) v_i \\
&\qquad + \frac{1}{2} \sum_{x_i \in \omega_P} \left(\sum_{j \in \Pi(i)} \beta_{ij} u_j \right) v_i \\
&= I_1 + I_2 + I_3\,.
\end{aligned}$$

We transform the second term

$$\begin{aligned}
I_2 &= \frac{1}{2} \sum_{x_i \in \omega_P} \left(\sum_{j \in \Pi(i)} |\beta_{ij}| (u_i - u_j) \right) v_i \\
&= \frac{1}{4} \sum_{x_i \in \omega_P} \sum_{j \in \Pi(i)} \left[|\beta_{ij}| (u_i - u_j) v_i + |\beta_{ji}| (u_j - u_i) v_j \right] \\
&= \frac{1}{4} \sum_{x_i \in \omega_P} \sum_{j \in \Pi(i)} |\beta_{ij}| (u_i - u_j)(v_i - v_j)\,.
\end{aligned}$$

And the third term equals

$$I_3 = \frac{1}{2} \sum_{x_i \in \omega_P} \sum_{j \in \Pi(i)} \beta_{ij} \left[u_j v_i - u_i v_j \right] .$$

Letting $v = I_h^c u_h$ and using the Proposition 4.1. we obtain the desired estimate. □

In the same way as in Chapter IV we prove the stability estimate.

Corollary 6.1 *For the upwind FVEM the following a priori estimate is valid:*

$$\|u_h\|_{1,\Omega} \le \|f\|_{-1,\omega} .$$

Let $z(\mathrm{x}) = u_h(\mathrm{x}) - u(\mathrm{x}) , \quad x \in \omega_P$. Substituting $u_h = z + u$ in (6.12) we obtain

$$\mathcal{B}_h z = \phi - \mathcal{B}_h u \equiv \psi . \tag{6.15}$$

Then using (6.12)–(6.13) we transform ψ in the following form

$$\sum_{j \in Pi(i)} \left[\frac{1}{\mathrm{meas}(V_i)} \int_{\gamma_{ij}} (-A \nabla u, \mathbf{n})\, ds - \int_{\gamma_{ij}} (-A \nabla I_h^l u, \mathbf{n})\, ds \right]$$
$$+ \sum_{j \in Pi(i)} \left[\frac{1}{\mathrm{meas}(V_i)} \int_{\gamma_{ij}} (\mathbf{b}, \mathbf{n}) u \, ds - \left[\beta_{i,j}^{+} u_{h,i} + \beta^{-} u_{h,j} \right] \right] \equiv \psi_{1,i} + \psi_{2,i} = \psi_i .$$

We define the local truncation error in the following way:

$$\eta_{i,j} = \frac{1}{\mathrm{meas}(\gamma_{ij})} \int_{\gamma_{ij}} (-A(\nabla(u - I_h^l u), \mathbf{n})\, ds , \tag{6.16a}$$

$$\mu_{i,j} = \frac{1}{\mathrm{meas}(\gamma_{ij})} \int_{\gamma_{ij}} (\mathbf{b}, \mathbf{n}) u \, ds - \frac{\mathrm{meas}(V_i)}{\mathrm{meas}(\gamma_{ij})} \left[\beta_{i,j}^{+} u_{h,i} + \beta^{-} u_{h,j} \right] . \tag{6.16b}$$

First we consider the term $(\phi_2, z)_B$. By the definition of the discrete inner product and $\phi_{2,i}$ we have

$$\begin{aligned}
(\phi_2, z)_B &= \sum_{x_i \in \omega} \mathrm{meas}(V_i) \phi_{2,i} z_i \\
&= \sum_{x_i \in \omega} \sum_{j \in \Pi(i)} \int_{\gamma_{ij}} (-A(\nabla(u - I_h^l u), \mathbf{n})\, ds\, z_i \\
&= -\frac{1}{2} \sum_{x_i \in \omega} \sum_{j \in \Pi(i)} dist(x_i, x_j)\, \mathrm{meas}(\gamma_{ij}) \eta_{i,j} \frac{[z_j - z_i]}{dist(x_i, x_j)} \\
&\le C \left(\sum_{x_i \in \omega} \sum_{j \in \Pi(i)} \mathrm{meas}(V_i) \eta_{i,j}^2 \right)^{1/2} \left(\sum_{x_i \in \omega} \sum_{j \in \Pi(i)} \mathrm{meas}(V_i) \left[\frac{[z_j - z_i]}{dist(x_i, x_j)} \right]^2 \right)^{1/2} \\
&\le C \|\eta\|_{*,\omega} \|z\|_{1,B} .
\end{aligned}$$

Likewise

$$
\begin{aligned}
(\phi_1, z)_B &= \sum_{x_i \in \omega} \mathrm{meas}(V_i)\phi_{2,i} z_i \\
&= \sum_{x_i \in \omega} \sum_{j \in \Pi(i)} \left[\int_{\gamma_{ij}} (\mathbf{b}, \mathbf{n}) u \, ds - \left(\frac{\beta_{i,j} + |\beta_{i,j}}{2} u_i + \frac{\beta_{i,j} - |\beta_{i,j}}{2} u_j \right) \right] z_i \\
&= -\frac{1}{2} \sum_{x_i \in \omega} \sum_{j \in \Pi(i)} dist(x_i, x_j)\, \mathrm{meas}(\gamma_{ij}) \\
&\qquad \left[\frac{1}{\mathrm{meas}(\gamma_{ij})} \int_{\gamma_{ij}} (\mathbf{V}, \mathbf{n}) - \frac{\mathrm{meas}(V_i)}{\mathrm{meas}(\gamma_{ij})} v_{i,j} \right] \frac{[z_j - z_i]}{dist(x_i, x_j)} \\
&= -\frac{1}{2} \sum_{x_i \in \omega} \sum_{j \in \Pi(i)} dist(x_i, x_j)\, \mathrm{meas}(\gamma_{ij}) \mu_{i,j} \frac{[z_j - z_i]}{dist(x_i, x_j)} \\
&\leq \left(\sum_{x_i \in \omega} \sum_{j \in \Pi(i)} \mathrm{meas}(V_i) \mu_{i,j}^2 \right)^{1/2} \left(\sum_{x_i \in \omega} \sum_{j \in \Pi(i)} \mathrm{meas}(V_i) \left[\frac{[z_j - z_i]}{dist(x_i, x_j)} \right]^2 \right)^{1/2} \\
&\leq \|\mu\|_{*,\omega} \|z\|_{1,\omega} \, .
\end{aligned}
$$

Summarizing these results and using Proposition 6.2 we obtain the following main result.

Lemma 6.3 *Let the Assumptions 4.1 and 4.3 be satisfied. The error* $z(x) = u_h(x) - u(x)\,, x \in \omega$ *of the upwind finite volume element method satisfies the a priori estimate*

$$\|z\|_{1,B} \leq C \left(\|\eta\|_{*,\omega} + \|\mu\|_{*,,\omega} \right) \tag{6.17}$$

where the components $\eta_{i,j}$ *and* $\mu_{i,j}$ *of the local truncation error are defined by (6.16). The constant* C *does not depend on* h *or* z.

In order to use the estimate (6.17) of Lemma 6.3 we have to bound the corresponding norms of the local truncation error components $\eta_{i,j}$ and $\mu_{i,j}$ defined by (4.22). Note that the first term has been taken care of in the diffusion dominated case. The estimate for $\mu_{i,j}$ is provided in the lemma given below. The proof of the lemma is manor modification of the result in Chapter IV and we skip it.

Lemma 6.4 *Let the solution of the problem (2.11) be* H^s*-regular,* $\frac{3}{2} < s$*, and the component of the local truncation error* $\mu_{i,j}$ *be defined by (6.16b). Then the following estimate holds:*

$$|\mu_{i,j}| \leq C h^{1-d/2} \left[|\mathbf{b}|_{0,\infty,\Omega} |u|_{1,e_{ij}} + h^{s-1} \|\mathbf{b}\|_{d/2+\alpha,\infty,\Omega} \|u\|_{s,e_{ij}} \right] \tag{6.18}$$

where $\frac{d}{2} < s \leq 2$.

Theorem 6.3 *If the solution* $u(x)$ *of the problem (2.11) is* H^s*-regular, with* $\frac{3}{2} < s \leq 3$ *and the Assumptions 4.1 and 4.3 are satisfied then the upwind finite volume element method has at most first order of convergence in the* H^1*–discrete norm, and*

$$\|u_h - u\|_{1,\omega} \leq C h |\mathbf{b}|_{0,\infty,\Omega} |u|_{1,\Omega} + C h^{s-1} \left(1 + h^{\delta} (\|\mathbf{b}\|_{d/2+\alpha,\infty,\Omega}) \right) \|u\|_{s,\Omega}.$$

CHAPTER VII

APPLICATIONS TO GROUNDWATER FLOW MODELS

Groundwater aquifers are one of the basic sources of drinking and industrial water supply. The quality of the water is of utmost importance for many users. It is very expensive to monitor the water contamination through physical observations. In many cases computer simulations are preferable because they can be run many times with different data for a small portion of the cost of digging and maintaining wells.

Many mathematical models are proposed in the literature for modeling of groundwater flow (see [18], [9], [30], [3], [32]). We consider the two phase total velocity/global pressure model introduced by Chavent and Jaffre [30]. The model is described by a nonlinear system of PDEs.

We concentrate on the equation that governs the saturation of the wetting phase. The saturation equation is strongly nonlinear and convection dominated. Although many papers are devoted to construction and study of numerical methods for such problems (see [40], [33]), a comprehensive theory is still not available.

Our goal is to investigate numerically some linearization techniques for the saturation equation that utilize ideas of operator splitting introduced by Espedal and Ewing [40]. We consider two stabilization techniques. The first one is stabilizing the discrete method by adding artificial diffusion, i.e., we make the main diagonal of the matrix dominant. This method is applied for trilinear finite elements on a distorted cubical mesh. The second discretization is upwind finite element methods on tetrahedral meshes.

This chapter is organized as follows. In Section 7.1 we state the conservation laws and in Section 7.2 we outline the constitutive equations. The mathematical model is described in Section 7.3. We consider the first linearization of the saturation equation and apply the stabilized with artificial diffusion trilinear finite element discretization in Section 7.4. In Section 7.5 we present an improved linearization and an upwind finite element method on tetrahedral meshes.

7.1 Conservation laws

Let Ω be a bounded domain in $\mathbb{R}^3$. We consider the displacement of two immiscible compressible fluids in Ω. In particular fluid 1 is water and fluid 2 is air. We will call them more often phases and will use a notation: α–phase, $\alpha = water, air$.

We consider the following unknowns:

S_α – saturation of phase α,
p_α – pressure of phase α,
$\mathbf{v}_\alpha$ – volumetric flux (or Darcy flux) of phase α,

and the physical parameters

$$
\begin{aligned}
\mathrm{x} &- \text{coordinates of a given point in } \Omega,\ \mathrm{x} = (x_1, x_2, x_3)\,,\\
\mathbf{g} &- \text{gravity acceleration },\ \mathbf{g} = g\nabla z\,,\\
\phi &- \text{porosity of the porous media},\ \phi = \phi(\mathrm{x}, p_a, p_w)\,,\\
\mathbf{K} &- \text{absolute permeability},\ \mathbf{K} = \mathbf{K}(\mathrm{x})\,,\\
\rho_\alpha &- \text{density of phase } \alpha,\ \rho_\alpha = \rho_\alpha(p_\alpha)\,,\\
\mu_\alpha &- \text{fluid viscosity of phase } \alpha,\ \mu_\alpha = \mu_\alpha(p_\alpha)\,,\\
k_{r\alpha} &- \text{relative permeability of phase } \alpha,\ k_{r\alpha} = k_{r\alpha}(S_w, \mathrm{x}, p_\alpha)\,,\\
F_\alpha &- \text{source or sink of phase } \alpha,\ F_\alpha = F_\alpha(\mathrm{x}, t)\,,
\end{aligned}
$$

The governing equations for fluid flow motion through porous media are the mass conservation laws (mass balance equations) for each phase α [9, 30],

$$\frac{\partial(\phi\rho_\alpha S_\alpha)}{\partial t} + \nabla.(\rho_\alpha \mathbf{v}_\alpha) = F_\alpha \tag{7.1}$$

and Darcy's law for two phase flow

$$\mathbf{v}_\alpha = -\frac{\mathbf{K}k_{r\alpha}}{\mu_\alpha}\left(\nabla p_\alpha - \rho_\alpha \mathbf{g}\right). \tag{7.2}$$

We suppose that fluids fill the volume, i.e., volume balance equation

$$S_w + S_a = 1\,.$$

7.2 Constitutive equations

In order to get a closed system of equations we need one more equation: the capillary pressure law

$$p_a - p_w = p_c(S_w, \mathrm{x})\,.$$

We assume that we can compute the capillary pressure p_c as a function only of the saturation of water S_w and spatial coordinates, which is a very rough approximation, but sufficient for our model.

For air density we propose the functional relation

$$\rho_\alpha = \rho_{\alpha,ref}\left(1 + \frac{p_a}{p_{a,ref}}\right).$$

7.3 Global pressure / total velocity formulations

Here we briefly sketch the derivation of global pressure/total velocity–saturation equations. For the detailed discussion of the notion of total pressure and global velocity we refer the book by Chavent and Jaffre [30].

7.3.1 Assumptions for relative permeability and capillary pressure functions

The residual saturation $S_{\alpha r}$ for the fluid α is the value of saturation S_α below that the fluid α cannot be replaced. In general $S_{\alpha r}$ is a function of x and possibly p_α. We suppose that we know the values of residual saturations at a point $x_0 \in \Omega$, i.e., $S_{wr}(x_0)$ and $S_{ar}(x_0)$. We

assume that the relative permeability functions are uniquely determined by the saturation of the wetting phase S_w, i. e.,

$$k_{rw}(x, S_w, p_w) = k^0_{rw}(S_w)\,, \quad k_{ra}(x, S_w, p_w) = k^0_{ra}(S_w)\,.$$

For the functions $k_{r\alpha}$ we assume that

$$k_{rw} = 0 \text{ for } S_w \in [0, S_{wr}]\,, \quad \text{increasing and smooth}\,,$$

$$k_{ra} = 0 \text{ for } S_w \in [S_{ar}, 1]\,, \quad \text{decreasing and smooth}\,.$$

Define the volume factors B_α and mobility factors d_α

$$B_\alpha(p_w) = \frac{\rho_\alpha(p_w)}{\rho_{\alpha,ref}}\,, \quad d_\alpha(p_w) = \frac{B_\alpha}{\mu_\alpha}\,, \quad \alpha = air,\, water\,, \tag{7.3}$$

global mobility d and fractional flow function f_w

$$d(S_w, p_w) = d_w(p_w)k^0_{rw}(S_w) + d_a(p_w)k^0_{ra}(S_w)\,,$$

$$f_w(S_w, p_w) = \frac{d_w(p_w)k^0_{rw}(S_w)}{d_w(p_w)k^0_{rw}(S_w) + d_a(p_w)k^0_{ra}(S_w)}\,.$$

Clearly, given mobility factors, global mobility and fractional flow we can find relative permeability functions via

$$k_{rw} = f_w(S_w, p_w)\frac{d(S_w, p_w)}{d_w(p_w)}\,, \quad k_{ra} = (1 - f_w(S_w, p_w))\frac{d(S_w, p_w)}{d_a(p_w)}\,. \tag{7.4}$$

We suppose that the capillary pressure, as a function of the saturation S_w, is independent of x up to a scaling factor $p_{CM}(\mathrm{x})$, i.e,

$$p_c(S_w, \mathrm{x}) = p_{CM}(\mathrm{x})p_c(S_w)\,, \quad -1 \le p_c(S_w) \le 1\,.$$

where p_c is a decreasing function defined in $[S_{wr}, S_{ar}]$ with $p_c(S_c) = 0$.

7.3.2 Assumptions for pressure dependent coefficients

We assume that functions $\phi(\mathrm{x}, p_w)$, $\rho_\alpha(p_w)$, $B_\alpha(p_w)$, $\mu_\alpha(p_w)$, $k_{r\alpha}(x, p_w)$ vary very slowly with p_w. We replace p_w with a some intermediate value p, i.e., $p \in [p_w, p_a]$. Then

$$\phi(\mathrm{x}, p_w) = \tilde{\phi}(\mathrm{x}, p)\,, \quad \rho_\alpha(p_w) = \tilde{\rho}_\alpha(p)\,, \quad B_\alpha(p_w) = \tilde{B}_\alpha(p)\,,$$

$$\mu_\alpha(p_w) = \tilde{\mu}_\alpha(p)\,, \quad d_\alpha(p_w) = \tilde{d}_\alpha(p)\,.$$

From now on we will work only with functions depending on p and will skip the tilde.

7.3.3 Derivation of the equations

We define global pressure by

$$p = \frac{p_w + p_a}{2} - p_{CM}(x)\int_{S_c}^{S}\left(f_w(s, p) - \frac{1}{2}\right)\frac{dp_c(s)}{ds}\,ds. \tag{7.5}$$

The definition (7.5) is meaningful since (7.5) defines a contraction mapping from $[p_w, p_a]$ into itself [30].

The total velocity is defined by

$$\mathbf{v} = B_w\mathbf{v}_w + B_a\mathbf{v}_a\,. \tag{7.6}$$

The pressure equation is derived by summing (7.1) for $\alpha = a,\ w$

$$\frac{\partial}{\partial t}\left(\phi\left(B_w S_w + B_a(1-S_w)\right)\right) + \nabla.\mathbf{v} = \frac{F_w}{\rho_{w,ref}} + \frac{F_a}{\rho_{a,ref}}\,. \tag{7.7}$$

Denote

$$\gamma = \int_{S_c}^{S}\left(f_w - \frac{1}{2}\right)\frac{dp_c(s)}{ds}\,ds\,, \quad \gamma_1 = \int_{S_c}^{S}\frac{\partial f_w}{\partial s}p_c(s)\,ds\,.$$

We differentiate (7.5) and obtain

$$\begin{aligned}
\nabla p &= \nabla\left(\frac{p_w+p_a}{2}\right) - \gamma\nabla p_{CM} - p_{CM}\nabla\gamma \\
&= \nabla\left(\frac{p_w+p_a}{2}\right) - \gamma\nabla p_{CM} - p_{CM}\frac{\partial\gamma}{\partial S}\nabla S - p_{CM}\frac{\partial\gamma}{\partial P}\nabla P\,, \\
\left(1+p_{CM}\frac{\partial\gamma}{\partial p}\right)\nabla p &= \nabla\left(\frac{p_w+p_a}{2}\right) - \gamma\nabla p_{CM} - p_{CM}\frac{\partial\gamma}{\partial S}\nabla S \\
&= \nabla\left(\frac{p_w+p_a}{2}\right) - \gamma\nabla p_{CM} - p_{CM}\left(f_w-\frac{1}{2}\right)\frac{dp_c}{dS}\nabla S \\
&= \nabla\left(\frac{p_w+p_a}{2}\right) - \gamma\nabla p_{CM} - \left(f_w-\frac{1}{2}\right)\left[\nabla(p_a-p_w) - p_c\nabla p_{CM}\right] \\
&= f_w\nabla p_w + (1-f_w)\nabla p_a + \gamma_1\nabla p_{CM}\,.
\end{aligned} \tag{7.8}$$

We multiply the fluxes given by the phase Darcy laws with the corresponding volume factors and using the definition of the volume and mobility factors (7.3) and the formulas (7.4) we get

$$\begin{aligned}
B_w\mathbf{v}_w &= -B_w\frac{\mathbf{K}k_{rw}}{\mu_w}(\nabla p_w - \rho_w\mathbf{g}) \\
&= -d_w\mathbf{K}k_{rw}(\nabla p_w - \rho_w\mathbf{g}) \\
&= -f_w\left[\mathbf{K}d(\nabla p_w - \rho_w\mathbf{g})\right]
\end{aligned} \tag{7.9}$$

and

$$\begin{aligned}
B_a\mathbf{v}_a &= -B_a\frac{\mathbf{K}k_{ra}}{\mu_a}(\nabla p_a - \rho_a\mathbf{g}) \\
&= -d_a\mathbf{K}k_{ra}(\nabla p_a - \rho_a\mathbf{g}) \\
&= -(1-f_w)\left[\mathbf{K}d(\nabla p_a - \rho_a\mathbf{g})\right]\,.
\end{aligned} \tag{7.10}$$

Now, for the velocity equation we get from (7.6), (7.8), (7.9) and (7.10)

$$\mathbf{v} = -\mathbf{K}d\left(\left(1+p_{CM}\frac{\partial\gamma}{\partial p}\right)\nabla p - [f_w\rho_w + (1-f_w)\rho_a]\mathbf{g} - \gamma_1\nabla p_{CM}\right)\,.$$

We can recover the phase pressure from the global pressure by the formulas :

$$\begin{aligned}
p_w &= p + \left[\gamma(S_w,p) - \frac{1}{2}p_c(S_w)\right]p_{CM}\,, \\
p_a &= p + \left[\gamma(S_w,p) + \frac{1}{2}p_c(S_w)\right]p_{CM}\,.
\end{aligned}$$

The relations between total velocity $\mathbf{v}$ and and phase velocities $\mathbf{v}_a$ and $\mathbf{v}_w$ are

$$\begin{aligned}\mathbf{v}_w &= \frac{f_w}{B_w}\mathbf{v} + D_w(1-f_w)p_{CM}\nabla p_c - D_w(1-f_w)\delta\rho\mathbf{g} \\ &\quad -D_w\left(\gamma_1+\gamma-\frac{1}{2}p_c\right)\nabla p_{CM}\,, \\ \mathbf{v}_a &= \frac{1-f_w}{B_a}\mathbf{v} - D_a f_w p_{CM}\nabla p_c + D_a f_w\delta\rho\mathbf{g} \\ &\quad -D_a\left(\gamma_1+\gamma+\frac{1}{2}p_c\right)\nabla p_{CM}\,,\end{aligned} \tag{7.11}$$

where $D_\alpha = \mathbf{K}k_{r\alpha}/\mu_\alpha = \mathbf{k}f_w d/B_\alpha$. Change of the variables in the definition of the global pressure (7.7) gives

$$p = p_a - p_{CM}(x)\int_{S_c}^{S} f_w(s,p)\frac{dp_c(s)}{ds}\,ds = p_a - p_{CM}(x)\int_0^{p_c} f_w(p_c^{-1},p)\,d\,\xi\,,$$

and therefore

$$\frac{\partial p_a}{\partial t} = \frac{\partial p}{\partial t} + f_w p_{CM}\frac{\partial p_c}{\partial t}\,.$$

After substitution of $\partial\rho_a/\partial t = d\rho_a/dp_a\,\partial p_a/\partial t$ in the pressure equation we get

$$A(S_w,p,x)\frac{\partial p}{\partial t} + \nabla.\mathbf{v} + \left(B(S_w,p)\frac{\partial\phi}{\partial t} + C(S_w,p,x)\frac{\partial S_w}{\partial t}\right) = \frac{F_w}{\rho_{w,ref}} + \frac{F_a}{\rho_{a,ref}}\,,$$

where

$$A = \phi\frac{(1-S_w)}{\rho_{a,ref}}\frac{d\rho_a}{dp_a}\,,\quad B = S_w + \frac{\rho_a}{\rho_{a,ref}}(1-S_w)\,,\quad C = \phi\left(1-\frac{\rho_a}{\rho_{a,ref}}\right) + Af_w p_{CM}\frac{dp_c}{dS_w}\,.$$

7.4 Saturation equation. Artificial diffusion approach

We replace $\mathbf{v}_w$ with the right hand side of (7.11) in the equation (7.1) for $\alpha = w$ (water) and taking into account that ρ_w does not depend on t (incompressibility of water) we get the equation for the saturation of water S_w:

$$\rho_w\phi\frac{\partial S_w}{\partial t} + \nabla.\rho_w(f_w\mathbf{v} + f_g\mathbf{g}) - \nabla.\rho_w((\mathbf{D}+\mathbf{D}_1)\nabla S_w) + \rho_w\frac{\partial\phi}{\partial t}S_w = Q(\mathrm{x},t), \tag{7.12}$$

with initial and boundary conditions

$$\begin{aligned} S_w(\mathrm{x},0) &= S_w^0(\mathrm{x})\,,\quad \mathrm{x}\in\Omega\,, \\ S_w(\mathrm{x},t) &= S_D(\mathrm{x},t)\,,\quad \mathrm{x}\in\Gamma_{s,1}\,,\ t>0\,, \\ \rho_w(f_w\mathbf{v} + f_g\mathbf{g} - (\mathbf{D}+\mathbf{D}_1)\nabla S_w).\mathbf{n} &= g_w(\mathrm{x},t)\,,\quad \mathrm{x}\in\Gamma_{s,2}\,,\ t>0\,, \end{aligned} \tag{7.13}$$

where

$$\mathbf{D} = -\mathbf{K}df_w(1-f_w)p_{CM}\frac{dp_c}{dS_w}\ \text{ and }\ f_g = -\mathbf{K}df_w(1-f_w)p_{CM}(\rho_w-\rho_a)\,.$$

Note that we add a new macro-dispersion term $\mathbf{D}_1$ which is a result of up-scaling of the saturation equation. The statistical theory for the concentration equation is developed by

Dagan [32]. Macro-dispersion and heterogeneous models are also discussed in [50] and [74]. This term is defined by

$$\mathbf{D}_1 = \begin{bmatrix} d_{xx}^{(1)} & d_{xy}^{(1)} & d_{xz}^{(1)} \\ d_{xy}^{(1)} & d_{yy}^{(1)} & d_{yz}^{(1)} \\ d_{xz}^{(1)} & d_{yz}^{(1)} & d_{zz}^{(1)} \end{bmatrix}, \tag{7.14}$$

where the entries are computed by the formulas

$$\begin{aligned} d_{xx}^{(1)} &= (d_l v_1^2 + d_{trh} v_2^2 + d_{trv} v_3^2)/|\mathbf{v}| + d_m\,, \\ d_{xy}^{(1)} &= (d_l - d_{trh}) v_1 v_2/|\mathbf{v}|\,, \\ d_{xz}^{(1)} &= (d_l - d_{thv}) v_1 v_3/|\mathbf{v}|\,, \\ d_{yy}^{(1)} &= (d_{trh} v_1^2 + d_l v_2^2 + d_{trv} v_3^2)/|\mathbf{v}| + d_m\,, \\ d_{yz}^{(1)} &= (d_{trh} - d_{trv}) v_2 v_3/|\mathbf{v}|\,, \\ d_{zz}^{(1)} &= (d_{trh} v_1^2 + d_{trh} v_2^2 + d_l v_3^2)/|\mathbf{v}| + d_m\,. \end{aligned}$$

Here d_l is the coefficient of longitudinal dispersion, d_{trh} is the coefficient of transversal horizontal dispersion, d_{trv} is the coefficient of transversal vertical dispersion, d_m is molecular diffusion and $\mathbf{v} = (v_1, v_2, v_3)$.

The problem (7.12), (7.13) is nonlinear and convection dominated. Therefore, the most important decisions are how to resolve the nonlinearity of the convection term and how to choose stable discretization scheme. The modeled physical process exhibits two separate regimes:

(i) build up of the saturation front,

(ii) movement of the front.

We assume that if the saturation S_w is less than some critical value S_0 the front is not established yet. We find the maximal value $S_{w,max}^n$ of the approximate solution on the n^{th} time step and compare with S_0. If $S_{w,max}^n < S_0$ we approximate fractional flow function in each element by a piecewise linear function

$$f_w(S_w^{n+1}) \approx \begin{cases} 0\,, & \text{if } S_w^{n+1} \le S_{wr}\,, \\ f_w(S_w^n)\,.\,(S_w^{n+1} - S_{wr})/(S_w^n - S_{wr})\,, & \text{otherwise}\,. \end{cases} \tag{7.15}$$

We time lag diffusion and gravity terms, i.e.,

$$\mathbf{D}(S_w^{n+1}, \mathbf{v}^{n+1}) \approx \mathbf{D}(S_w^n, \mathbf{v}^{n+1})\,, \quad f_g(S_w^{n+1})\mathbf{g} \approx f_g(S_w^n)\mathbf{g}. \tag{7.16}$$

Because of the property of air–water system we believe that when the saturation front is established the following simple splitting will produce reasonable results:

$$f_w(S_w^{n+1}) \approx \begin{cases} F(S_w^{n+1})\,, & \text{if } \mathbf{v}^{n+1}.\nabla S_w^n \le 0\,, \\ f_w(S_w^n)\,, & \text{if } \mathbf{v}^{n+1}.\nabla S_w^n > 0 \end{cases}, \tag{7.17}$$

where $F(S_w^{n+1})$ is defined by (7.15) with S_w^n replaced by the point S_0 where F is tangent to f_w. Note that first approximation is meaningful if S_0 is close to 1, i.e., we use $f_w(S_w^n)$ instead of $f_w(S_w^{n+1})$ on a small interval. We again time lag for the terms corresponding to the diffusion and gravity, i.e., we use the approximation (7.16).

The resulting linear equation is:

$$\phi \frac{\partial S_w}{\partial t} + \nabla.\rho_w(b\mathbf{v}S_w) - \nabla.\rho_w(\bar{\mathbf{D}}\nabla S_w) + \frac{\partial \phi}{\partial t} S_w = F(\mathrm{x}, t), \tag{7.18}$$

with initial and boundary conditions

$$\begin{aligned} S_w(\mathrm{x},0) &= S_w^0(\mathrm{x})\,, \quad \mathrm{x}\in\Omega\,, \\ S_w(\mathrm{x},t) &= S_D(\mathrm{x},t)\,, \quad \mathrm{x}\in\Gamma_{s,1}\,,\ t>0\,, \\ \rho_w(b\mathbf{v}S_w - \bar{\mathbf{D}}\nabla S_w).\mathbf{n} &= g(\mathrm{x},t)\,, \quad \mathrm{x}\in\Gamma_{s,2}\,,\ t>0\,, \end{aligned} \tag{7.19}$$

The coefficients of (7.18), (7.19) are connected with the coefficients of (7.12), (7.13) by the relations:

$$\begin{aligned} b(\mathrm{x}) &= cS_w^{n+1}, \\ c &= \begin{cases} f(S_w^n)/S_w^n & \text{if the front is not built up,} \\ f(S_0)/S_0 & \text{if there is a front and } \mathbf{v}^{n+1}.\nabla S_w^n \le 0 \\ 0 & \text{if there is a front and } \mathbf{v}^{n+1}.\nabla S_w^n > 0 \end{cases} \\ F &= F_w - \nabla.\rho_w f_g(S_w^n)\mathbf{g} - A\,, \\ A &= \begin{cases} 0 & \text{if the front is not built up,} \\ 0 & \text{if there is a front and } \mathbf{v}^{n+1}.\nabla S_w^n \le 0 \\ \nabla.\rho_w f_w(S_w^n)\mathbf{v}^{n+1} & \text{if there is a front and } \mathbf{v}^{n+1}.\nabla S_w^n > 0 \end{cases} \\ \bar{\mathbf{D}} &= \mathbf{D}(S_w^n,\mathbf{v}) + \mathbf{D}_1(\mathbf{v})\,. \end{aligned}$$

The above algorithm is summarized in the following "program" like style:

```
Given functions $k_{rw}(S_w)$ and $k_{ra}(S_w)$ find S_0
do time loop
   if (front is already built) then
      apply splitting (8) and time lag  (6), (7)
   else
      if (the front is built on the last time step) then
          apply splitting (8) and time lag (6), (7)
      else
          in each element approximate
          fractional flow function by (5)
      endif
   endif
   solve linear elliptic problem
end do
```

7.4.1 Finite element approximation for the linearized equation

We use trilinear finite element method for the linearized equation. In order to make the discretization scheme stable we use simple upstream weighting, i.e., if d_{xx} , d_{yy} and d_{zz} are the diagonal entries of the matrix $D + D_1$ we add artificial diffusion by making them bigger:

$$d_{xx} = \max(d_{xx}, dx/2)\,, \quad d_{yy} = \max(d_{yy}, dy/2)\,, \quad d_{zz} = \max(d_{zz}, dz/2)\,.$$

A similar approach is described in [29].

7.4.2 Parallelization

The parallelization of the saturation code for the message passing distributed memory computers Intel iPSC/860 and PARAGON is based on the parallel flow code developed by J.E. Pasciak and A.T. Vassilev [42]. We parallelize the most time consuming tasks: matrix assembly, operator evaluation and inner products in the conjugate gradient square algorithm.

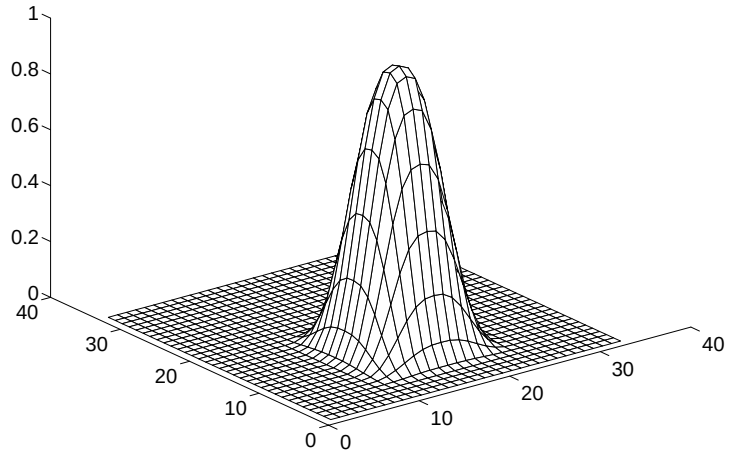


Figure 7.1: Exact solution $t = 0$

For the implementation we used the software developed at Brookhaven National Laboratory [86], which provides the user with tools for remote procedure calls in addition to the standard message passing libraries provided by Intel. This package helps avoid the inconveniences of explicit message passing and it also greatly simplifies the development and the debugging of rather complex user codes.

7.4.3 Numerical experiments

We report the computed results for a simple model problem. The exact solution $S_w(x, y, z, t)$ is given by

$$S_w(x, y, z, t) = f(x, t)f(y, t_0)f(z, t_0)\,, \quad t_0 = 0.5\,,$$

and the function f is defined via the formulas

$$z = 4(x - g(t)) - 2\,, \quad g(t) = \frac{6}{20}t + \frac{3}{20}\,,$$

$$f(x,t) = \begin{cases} 0\,, & 0 \le x \le 0.25\,, \\ (z+1)^3(6z^2 - 3z + 1)\,, & 0.25 + g(t) \le x \le 0.5 + g(t)\,, \\ (1-z)^3(6z^2 + 3z + 1)\,, & 0.5 + g(t) \le x \le 0.75 + g(t)\,, \\ 0\,, & 0.75 \le x \le 1\,. \end{cases}$$

Notice that the "bell" moves only in x–direction. The projection $z = 0.5$ of the exact solution is plotted on Fig. 7.1 and Fig. 7.2.

Relative permeability functions of water k_{rw} and air k_{ra} are (see Fig. 7.3)

$$k_{rw}(s) = \begin{cases} 0\,, & 0 \le s \le 0.4\,, \\ (5/3s - 2/3)^3\,, & 0.4 \le s \le 1\,, \end{cases}$$

$$k_r(s) = 3(-6s^4 + 13s^3 - 8s^2 + 1)(1 - s)^3(6s^2 + 3s + 1)\,.$$

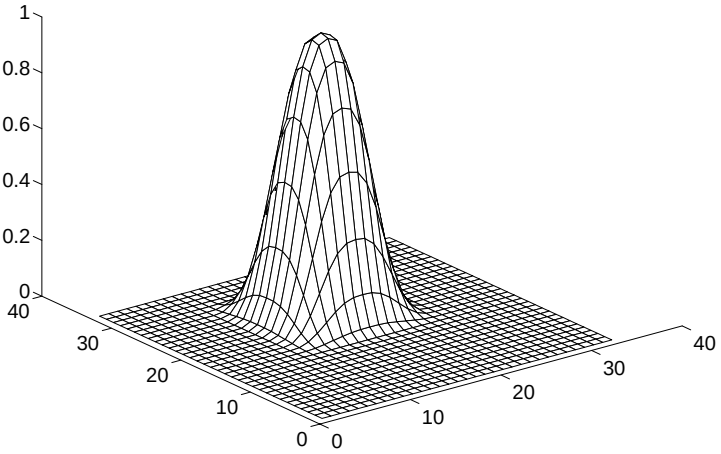


Figure 7.2: Exact solution $t = 1$

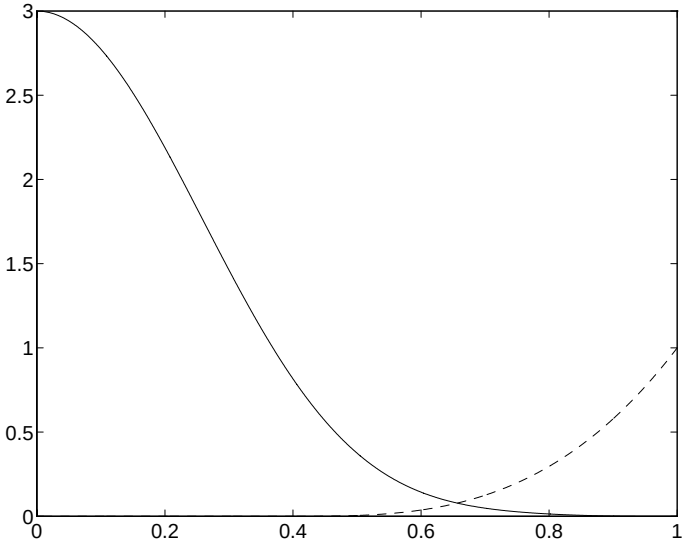


Figure 7.3: Relative permeability functions of air and water

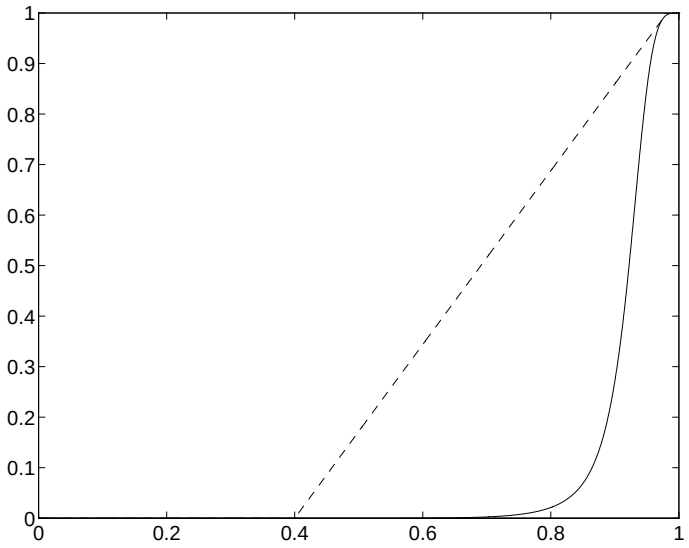


Figure 7.4: Fractional flow function and its approximation

Fractional flow function f_w is plotted on Fig. 7.4.

Capillary pressure is defined by (see Fig. 7.5)

$$p_c(s) = \begin{cases} 1 - \sqrt{1+z}\,, & 0.4 \le s \le 0.7\,, \\ \sqrt{1-z} - 1\,, & 0.7 \le s \le 1\,, \end{cases} \qquad z = \frac{10}{3}s - \frac{7}{3}\,.$$

We neglect gravity. For viscosity we choose $\mu_a = 10$ and $\mu_w = 10000$. We assume that $p_{CM} = 1$ and $\frac{\partial \gamma}{\partial p} = 0$. The constitutive law for air density ρ_a is

$$\rho_a = \rho_{a,ref}\left(1 + \frac{p}{p_{ref}}\right)\,,$$

where $p_{ref} = 1$, $\rho_{a,ref} = 0.001$. The global pressure is given by

$$p = 1 - \frac{x(2-t)}{2}\,.$$

Then for the component of the total velocity we have

$$v_1 = \frac{\mathbf{K}\lambda(2-t)}{2}\,, \quad v_2 = 0\,, \quad v_3 = \mathbf{K}\lambda[f_w + (1-f_w)\rho_a]\,.$$

where

$$\lambda = \frac{k_{rw}}{\mu_w} + \frac{k_{ra}}{\mu_a}\,.$$

Numerical results reported in Tables 7.1–7.4 show that the linearization technique is reliable for the whole range of the diffusion coefficient. We also observe that upstream weighting introduce significant smearing for small ε. The finite element method is not conservative and occasionally we compute negative saturations.

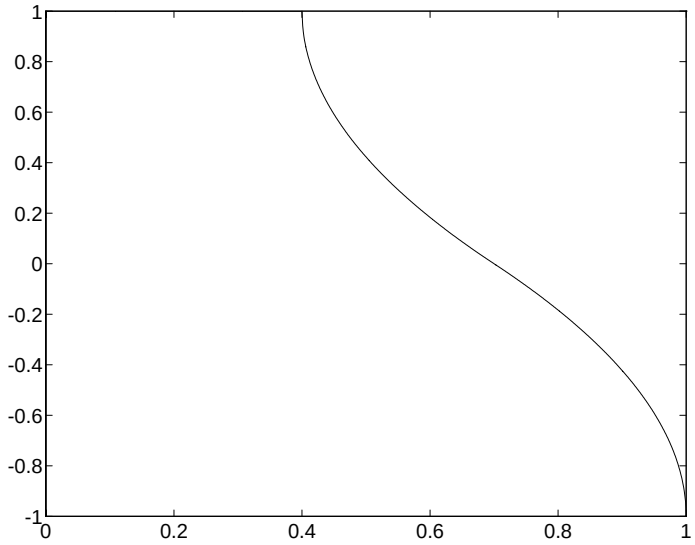


Figure 7.5: Capillary pressure

Table 7.1: Problem 1, $\varepsilon = 1$

$1/h$	$1/\Delta t$	L^2 error	max. error
4	10	0.16961	0.76230
8	20	0.04446	0.30927
16	40	0.00603	0.11507
32	80	0.00289	0.04095

Table 7.2: Problem 1, $\varepsilon = 0.1$

$1/h$	$1/\Delta t$	L^2 error	max. error
4	10	0.14293	0.78093
8	20	0.04474	0.32130
16	40	0.01273	0.12677
32	80	0.01071	0.07822

Table 7.3: Problem 1, $\varepsilon = 0.01$

$1/h$	$1/\Delta t$	L^2 error	max. error
4	10	0.10576	0.84208
8	20	0.07378	0.68442
16	40	0.04528	0.52926
32	80	0.02692	0.26734

Table 7.4: Problem 1, $\varepsilon = 0.001$

$1/h$	$1/\Delta t$	L^2 error	max. error
4	10	0.10749	0.85371
8	20	0.08223	0.76043
16	40	0.07026	0.84775
32	80	0.05675	0.72625

7.5 Saturation equation. Upwind discretization

In this section we consider an alternative global pressure/total velocity mathematical model based on two phase incompressible fluid flow in porous media. The compressibility of the air is introduced via the compressibility coefficient C_a (see (7.22)). For this model we apply an improved linearization and an upwind finite element method based on tetrahedral linear elements. For the full description we refer to User's Guide to GCT [100].

7.5.1 Alternative global pressure / total velocity formulation

We introduce the phase mobilities λ_α and the total mobility λ

$$\lambda_\alpha = \frac{k_{r\alpha}}{\mu_\alpha}\,, \quad \alpha = air,\, water\,, \qquad \lambda = \lambda_a + \lambda_w\,.$$

The global pressure is defined by

$$p = \frac{1}{2}(p_a + p_w) + \frac{1}{2}\int_{s_c}^{s} \frac{\lambda_a - \lambda_w}{\lambda}\frac{dp_c}{d\xi}d\xi\,,$$

and the total velocity satisfies the equations

$$\begin{aligned} &\mathbf{v} = \mathbf{v}_a + \mathbf{v}_w\,, \\ &\mathbf{v} = -\mathbf{K}\lambda(\nabla p - G(S_w, p))\,, \end{aligned} \tag{7.20}$$

where,

$$G(S_w, p) = \frac{\lambda_a \rho_a + \lambda_w \rho_w}{\lambda}\mathbf{g}\,. \tag{7.21}$$

Note that the definitions of the total velocities in the two models (7.6) and (7.20) correspondingly are different. The "incompressible" approach leads to many new terms in the right hand side (7.23).

In a similar way as in Section 7.3 we derive the equations of the model:

$$\begin{aligned} &C(p, S_w)\frac{\partial p}{\partial t} + \nabla.\mathbf{u} = f(p, S_w)\,, \\ &\mathbf{u} = -\mathbf{K}\lambda(\nabla p - G(S_w, p))\,, \\ &\frac{\partial(\phi\rho_w S_w)}{\partial t} + \nabla.\rho_w(f_w\mathbf{u} + f_g\mathbf{g}) - \nabla.((\mathbf{D}(S_w) + \mathbf{D}_1(\mathbf{v}))\nabla S_w) = F_w\,. \end{aligned}$$

Here

$$\begin{aligned} f_w = \frac{\lambda_w}{\lambda}\,, \quad f_g = -\mathbf{K}\lambda f_w(\rho_a - \rho_w)\,, \\ C(p, S_w) = \phi S_w C_a\,, \quad C_a = \frac{1}{\rho_a}\frac{d\rho_a}{dp_a}\,, \quad \mathbf{D}(S_w) = -\mathbf{K}\lambda_a f_w \frac{dp_c}{dS_w}\,. \end{aligned} \tag{7.22}$$

C_a is called compressibility. The term $\mathbf{D}_1$ is defined by (7.14).

The right hand side is given by

$$f(p,S_w) = \left(F_a - \mathbf{u}_a.\nabla\rho_a - \phi(1-S_w)\frac{\partial\rho_a}{\partial t}\right)/\rho_a + \left(F_w - \mathbf{u}_w.\nabla\rho_w - \phi S_w\frac{\partial\rho_w}{\partial t}\right)/\rho_w \quad (7.23)$$

To close the model we need two more constitutive laws:

$$\rho_a = \rho_{0a}\left(1+\frac{p_a}{p_{0a}}\right) \quad \text{and} \quad p_c = p_c(S_w)\,,$$

ρ_{0a} is the air density at reference pressure p_{0a}.

7.5.2 Linearization

The linearization in this section is an extension and refinement of the procedure described in Section 7.4. In each time step we apply inner iteration in order to resolve the nonlinear convection term. We denote the current solution by S^i and set

$$S^0 = S_w^n\,, \quad S_w^{n+1} = S^{last}\,.$$

We assume that the nonlinearities at gravity and diffusion terms are non-crucial and we time lag them, i.e.,

$$\mathbf{D}(S^i,\mathbf{v}^{n+1}) \approx \mathbf{D}(S^{i-1},\mathbf{v}^{n+1})\,, \quad f_g(S^i) \approx f_g(S^{i-1})\,.$$

First, we compute the shock saturation S_0 for the Riemann problem as an asymptotic guide. S_0 is the solution of the equation

$$f_w^{'}(S_0) = \frac{f_w(S_0)}{S_0}$$

We assume that if the maximal value $S_{w,max}^n$ of the approximate solution on the n^{th} time step is less than S_0, then the saturation front is not established yet. We split f_w in the following way

$$f_w(S)\mathbf{v} = \bar{f}_w(S)\mathbf{v} + b(S,\mathbf{v})S\,,$$

with function $\bar{f}_w(S)$ defined by

$$\bar{f}_w(S) = \begin{cases} 0\,, & \text{if } S \le S_{wr}\,, \\ f_w(S_{\max})\cdot\frac{(S-S_{wr})}{(S_{\max}-S_{wr})}\,. & \text{if } S_{wr} \le S \le S_{\max}\,, \\ f_w(S)\,, & \text{otherwise} \end{cases}$$

Note that $\bar{f}_w(S)$ is still a nonlinear function because we do not know $S_{\max}$. We approximate $\bar{f}_w(S)$ with a piecewise linear function $\tilde{f}_w(S)$

$$\tilde{f}_w(S^i) = \begin{cases} 0\,, & \text{if } S^{i-1} \le S_{wr}\,, \\ f_w(S^{i-1})\cdot\frac{(S^i-S_{wr})}{(S_{\max}^{i-1}-S_{wr})}\,. & \text{if } S_{wr} \le S^{i-1} < S_{\max}^{i-1}\,, \\ f_w(S_{\max}^{i-1})\,, & \text{otherwise}\,. \end{cases}$$

When the front is established we use the same splitting, but now we try to predict the movement of the front by using the shock velocity $\mathbf{v}_f$ defined by

$$\mathbf{v}_f = \frac{f_w(S_0)}{S_0 - S_{wr}}\mathbf{v}\,.$$

The first term in the splitting is given by

$$\bar{f}_w(S)\mathbf{v} = \begin{cases} 0\,, & \text{if } S \le S_{wr}\,, \\ \mathbf{v}_f(S - S_{wr})\,, & \text{if } S_{wr} \le S \le S_0\,, \\ f_w(S)\mathbf{v}\,, & \text{if } S_0 \le S \le 1\,. \end{cases}$$

This is still a nonlinear function and we approximate it with

$$\tilde{f}_w(S^i)\mathbf{v}^{n+1} = \begin{cases} 0\,, & \text{if } S^{i-1} \le S_{wr}\,, \\ \mathbf{v}_f(S^i - S_{wr})\,, & \text{if } S_{wr} \le S^{i-1} \le S_0\,, \\ f_w(S^{i-1})\mathbf{v}^{n+1}\,, & \text{if } S_0 \le S^{i-1} \le 1 \end{cases}$$

The described algorithm is implemented as follows:

(i) Predict S^0. ($S^0 = S_w^n$ or use an explicit scheme).
If ($nofront$) predict $S^0_{\max}$.

(ii) $for(i = 1;\, i \le imax;\, i++)$
Solve implicitly for S^i

$$\phi(x,p)\frac{\partial S^i}{\partial t} + \nabla.\left(\tilde{f}_w(S^i)\mathbf{v}^{n+1}\right) - \nabla.(D(S^{i-1},\mathbf{v}^{n+1})\nabla S^i) + \frac{\partial \phi}{\partial t}S^i$$
$$= Q(x,t) - \nabla.\left(b(S^{i-1},\mathbf{v}^{n+1})S^{i-1} + f_g(S^{i-1})(\mathbf{K}\nabla z)\right)\,. \quad (7.24)$$

If $\|S^i - S^{i-1}\|_0 < \varepsilon$
$S_w^{n+1} = S^i$;
Exit the loop;
else
$S^{i-1} = S^i$;
goto (i);

7.5.3 Upwind finite element method

We suppose that in the domain $\Omega \subset \mathbb{R}^d$ is introduced structured grid with elements distorted cubes. We devide each cube into five tetrahedra. In order to get a conforming finite element triagulation we have to devide the cubes in some order (see Fig. 7.6 and refer for details to [100]).

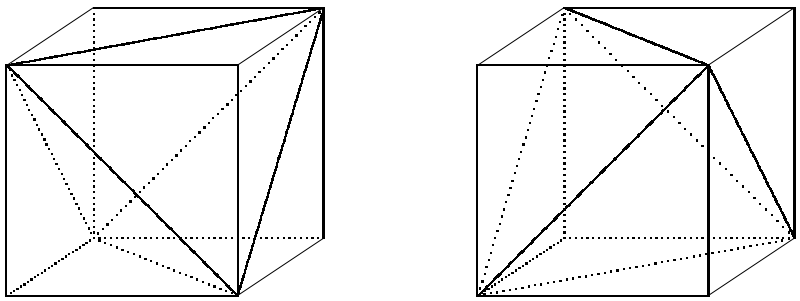


Figure 7.6: Partition of odd and even cells into five tetrahedra

We discretize the linear equation (7.24) using backward Euler scheme for the time derivative and trilinear finite elements on tetrahedral mesh.

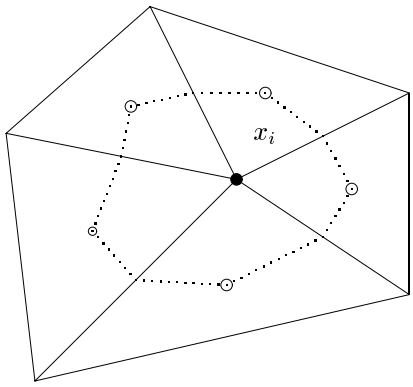


Figure 7.7: Baricentric region V_i

We develop an upwind approximation of the convection term $\int_\Omega \nabla.(\mathbf{b}u_h)v\,d\mathrm{x}$ using the finite volume approach. Here u_h is the approximate solution of the equation (7.24). For similar approach see works by Baba and Tabata [10] and Ikeda [63]. Let φ_i be the linear basis function corresponding to the i^{th} node and $\hat{\varphi}_i$ be a characteristic function of the barycentric region V_i of x_i (see Fig. 7.7).

We approximate φ_i with $\hat{\varphi}_i$, i.e.,

$$\int_\Omega \nabla.(\mathbf{b}u_h)\varphi_i\,d\mathrm{x} \approx \int_\Omega \nabla.(\mathbf{b}u_h)\hat{\varphi}_i\,d\mathrm{x}\,,$$

and using Gauss divergence formulas we get

$$\int_\Omega \nabla.(\mathbf{b}u_h)\hat{\varphi}_i\,d\mathrm{x} = \int_{\partial V_i} (\mathbf{b},\mathbf{n})u_h\,ds\,.$$

Let γ_{ij} be one of the faces of ∂V_i. We use the approximation

$$\int_{\gamma_{ij}} (\mathbf{b},\mathbf{n})u\,ds \approx \beta^+ u_i + \beta^- u_j\,,$$

where

$$\beta^+ = \frac{\beta_{i,j} + |\beta_{i,j}}{2}\,, \qquad \beta^- = \frac{\beta_{i,j} - |\beta_{i,j}}{2}$$

and $\beta_{i,j}$ is defined by formulas (6.14) or (4.7). Note that we used the same approach to derive upwind discretizations in Chapters IV, V and VI.

The method we described in this section still have to be tested numerically.

CHAPTER VIII

CONCLUSIONS

In this dissertation we derived several new finite volume and finite volume element methods for nonsymmetric elliptic boundary value problems. We studied theoretically their propertiies and tested computationally their efficiency on real–life problems. Our new results and future directions of research are briefly outlined in this chapter.

In Chapter III we provided a general framework for finite volume methods and discussed several ways to construct discretization schemes. This general approach allowed us to formulate a novel cell–centered finite difference scheme on Voronoi or circumscribed meshes for problems with tensor coefficients. It is an interesting problem to explore the properties of this method and apply it to mathematical models of physical phenomena of interest. Now it is well understood how to derive finite volume schemes from mixed finite element methods on uniform meshes for problems with scalar coefficients. However, there are many open problems when employing these methods on irregular meshes or solving equations with tensor coefficients. Construction of stable and accurate mixed finite element approximations for convection–diffusion problems is another important and interesting problem.

Stable accurate and locally mass conservative methods for strongly nonsymmetric problems that satisfy the discrete maximum principle on general meshes were discussed in Chapter IV. We proposed three different cell–centered finite difference schemes, **UDS**, **MUDS** and **IDS**, and proved that they satisfy the discrete maximum principle on Voronoi or circumscribed meshes under some conditions that are natural analogs of the conditions for the differential problem. Consequently, the corresponding discrete problems are well defined even for extremely irregular meshes. Such results are not valid for the classical finite element methods.

We introduced the so called Finite Volume regular meshes, and for this class of meshes we showed that the discrete problems were coercive, stable and convergent and derived in discrete H^1–norm. We elaborated the discrete Aubin–Nitsche "trick" and proved that under some conditions the convergence in discrete L^2–norm is with one order higher than in discrete H^1–norm. We specified some geometric assumptions (the symmetry assumptions) under which we showed that **MUDS** and **IDS** were superconvergent in discrete H^1–norm. We note that the symmetry assumptions are only sufficient and conjecture that for all "reasonably" regular Voronoi meshes such superconvergence estimates in discrete H^1–norm are valid. Another possible direction of research is to investigate the superconvergence of finite difference schemes on Voronoi meshes locally in discrete maximum norm.

It would be an interesting problem to extend our theory for elliptic problems to parabolic equations. Consistent theory for finite volume discretization of parabolic problems is still not available.

We presented extensive numerical experiments of the proposed methods that illustrated our theoretical estimates. The properties of finite volume methods on nonregular meshes have to be investigated numerically.

In Chapter V we studied cell–centered finite difference schemes with local patch refinement. We investigate two different interpolations along the interface between the coarse and fine regions, constant and linear, and constructed two conservative cell–centered finite difference methods, **UDS** and **MUDS**, that employed these interpolations. We proved that these schemes satisfy the discrete maximum principle and showed that the discrete problems were coercive. We provided the stability and error estimates in discrete H^1–norm with loss of accuracy half of order due to the interface interpolation. It would be an interesting oppor-

tunity to apply a more general approach, the so called mortar element methods, to the finite volume methods and reduce the error along the interface.

A general way for discretization of elliptic boundary value problems in divergence form is the finite volume element methods. This methods combines the advantages of both finite volume and finite element methods. FVE methods give locally mass conservative discretization, work for tensor coefficients without any special modifications, and can handle discontinuous coefficients similarly to the way the finite volume methods do. On the other hand in FVE methods the approximation of the fluxes on the faces of the control volumes is produced by using specified finite element spaces, and consequently they have the flexibility of finite element methods. We generalized the known results for 2–D symmetric problems to 2–D (3–D) nonsymmetric ones for barycentric control volumes. We proved the stability and error estimates for diffusion dominated cases using classical technique from finite element theory for Petrov–Galerkin methods (inf–sup condition). A new upwind finite volume element method was derived for convection dominated problems.The stability and error estimates were provided with the technique developed in Chapter IV.

We plan to investigate all considered methods on different types of meshes, for example Voronoi grids, and derive streamline diffusion finite volume element methods. These discretizations have to be tested numerically. It would be interesting to extend our theory to discretization of parabolic equations.

In Chapter VII we considered application of finite volume methods to groundwater flow simulations. We stated the conservation laws that govern the fluid flows in porous media and augmented them with constitutive relations in order to get a closed system of partial differential equations. For this mathematical model we discussed two different implementations of global pressure/total velocity formulations developed by Chavent and Jaffre [30]. The first one was based on the two phase compressible fluid flows. We developed a linearization procedure for the saturation equation and discretized the resulting linear problem with trilinear finite elements stabilized by adding artificial viscosity. This discretization was numerically tested and our results showed that the proposed linearization technique is reliable for the wide range of the diffusion coefficients. We also observed that artificial viscosity stabilization introduced significant smearing for small diffusion coefficients. We parallelized our computer code using the tools developed by Joseph Pasciak and Apostol Vassilev [42].

This parallelization was not compatible with the parallelization of the pressure code in PICS project (for details refer to User's Guide to GCT [100]), so we implemented tetrahedral meshes. We considered an alternative global pressure/total velocity formulation based on two phase incompressible fluid flow in porous media. The compressibility of the air is introduced via the compressibility coefficient. We extended and refined the linearization procedure for the saturation equation. We made it consistent in sense that when the time step goes to zero or two consecutive approximations get closer, the linear equation converges to the nonlinear one. Moreover, we developed an inner iteration of Picard type, each step of which includes solving of the linear problem. We constructed a new upwind finite element method using the finite volume approach to discretize the convection term. This discretization was tested for linear problems and implemented for the saturation equation. Now we are in a process of testing our algorithm for various practical problems.

In order to improve the overall performance of the numerical method we plan to concentrate on the following tasks:

1. Add residual control and eventually time step control in the nonlinear iteration.

2. Consider streamline diffusion FEM and their nonlinear performance.

3. Investigate local mass conservative (streamline) upwind finite volume element methods.

4. Implement fractional step method with explicit solve for the hyperbolic part of the saturation equation discretized using Godunov's methods and implicit solve on the diffusion part discretized by mixed finite element methods. This algorithm will be locally mass conservative.

5. Testing of our upwind discretization for problems with reaction terms.

Extension of the proposed algorithms has to be developed for three phase fluid flow models, which will allow modeling of more complex physical processes.

REFERENCES

[1] R. A. Adams. *Sobolev spaces.* Academic Press, New York, 1975.

[2] D. Allen and R. Southwell. Relaxation method applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder. *J. Mech. Appl. Math.*, 8:129–145, 1955.

[3] M. B. Allen III, G. A. Behie, and J. A. Trangenstein. *Multiphase Flow in Porous Media.* Number 34 in Lecture Notes in Engineering. Springer–Verlag, Berlin, 1988.

[4] Todd Arbogast, Mary F. Wheeler, and Ivan Yotov. Mixed finite elements for elliptic problems with tensor coefficients as cell–centered finite differences. *SIAM J.Numer. Anal.*, to appear.

[5] H. Aronszajn. Boundary value of functions with finite Dirichlet integral. In *Proc. 14th Conference on Partial Diff. Equat., University of Kansas, Lawrence*, 1955.

[6] J. P. Aubin. Behavior of the error of the approximate solutions of boundary value problems for linear elliptic operators by Galerkin and finite difference methods. *Ann. Scuola Norm. Sup. Pisa*, 21:599–637, 1967.

[7] Franz Aurenhammer. Voronoi diagrams – A survey of a fundamental geometric data structure. *ACM Comput. Surveys*, 23:345–405, 1991.

[8] O. Axelsson and I. Gustafson. A modified upwind scheme for convective transport equations and the use of conjugate gradient method for the solution of non-symmetric systems of equations. *J. Int. Math. Appl.*, 23(11):867–889, 1979.

[9] K. Aziz and A. Settari. *Petroleum Reservoir Simulation.* Applied Science, New York, 1979.

[10] Kinji Baba and Masahisa Tabata. On a conservative upwind finite element scheme for convection diffusion equations. *R.A.I.R.O. Analyse numerique/Numerical Analysis*, 15(1):3–25, 1981.

[11] Ivo Babuška and A. K. Aziz. Foundations of the finite element method. In A. K. Aziz, editor, *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, pages 3–362. Academic Press, New York, 1972.

[12] A. S. Bakhvalov. On the optimization of methods for solving boundary value problems with boundary layers. *Zh. Vychisl. Mat. i Mat. Fiz.*, 9:841–859, 1969. (Russian).

[13] B. R. Baliga and S. V. Patankar. A new finite-element formulation for convection-diffusion problems. *Num. Heat Transfer*, 3:393–409, 1980.

[14] B. R. Baliga and S. V. Patankar. A control volume finite element method for two–dimensional fluid flow and heat transfer. *Num. Heat Transfer*, 6:245–261, 1983.

[15] R. E. Bank, J. F. Bürler, W. Fichtner, and R. K. Smith. Some upwinding techniques for finite element approximations of convection–diffusion equations. *Numer. Math.*, 58:185–202, 1990.

[16] R. E. Bank and D. J. Rose. Some error estimates for the box method. *SIAM J.Numer. Anal.*, 24:777–787, 1987.

[17] J. W. Barret and K. W. Morton. Approximate symmetrization and Petrov–Galerkin methods for diffusion–convection problems. *Comput. Methods Appl. Mech. and Engrg.*, 45:97–122, 1984.

[18] Jacob Bear. *Dynamics of fluids in Porous Media.* Dover Publications, Inc., New York, 1972.

[19] Marshall Bern and David Eppstein. Mesh generation and optimal triangulation. In D. Z. Du and F. K. Hwang, editors, *Computing in Euclidean Geometry*, volume 1 of *Lecture Notes Series of Computing*, pages 23–90. World Scientific, 1992.

[20] C. Bernardi, Y. Maday, and A.T. Patera. A new nonconforming approach to domain decomposition: the mortar element method. In H. Brezis and J.L. Lions, editors, *Nonlinear Partial Differential Equations and Their Applications*, number 290 in Pitman Research Notes on Mathematics, pages 13–51, Harlow, UK, 1989. Longman Scientific&Technical.

[21] Christine Bernardi, Claudio Canuto, and Yvon Maday. Generalized inf–sub conditions for Chebyshev spectral approximation of the Stokes problem. *SIAM J.Numer. Anal.*, 25(6):1237–1271, 1988.

[22] J. H. Bramble and S. R. Hilbert. Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation. *SIAM J.Numer. Anal.*, 7:113–124, 1970.

[23] F. Brezzi, L. D. Marini, and P. Pietra. Two–dimensional exponential fitting and application to drift–diffusion models. *SIAM J. Numer. Anal.*, 26:1342–1355, 1989.

[24] Franco Brezzi and Michel Fortin. *Mixed and Hybrid Finite Element Methods.* Springer–Verlag, New York, 1991.

[25] A. N. Brooks and T. J. R. Hughes. Streamline upwind/petrov–galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations. *Comput. Methods Appl. Mech. Engrg*, 32:199–259, 1982.

[26] Z. Cai. On the finite volume element method. *Numer. Math.*, 58:713–735, 1991.

[27] Z. Cai, J. Mandel, and S. F. McCormick. The finite volume element method for diffusion equations on composite grids. *SIAM J.Numer. Anal.*, 28:392–402, 1991.

[28] Z. Cai and S. F. McCormick. On the accuracy of the finite volume element method for diffusion equations on composite grids. *SIAM J.Numer. Anal.*, 27(3):636–655, 1990.

[29] Michael A. Celia and William G. Gray. *Numerical Methods for Differential Equations: Fundamental Concepts for Scientific and Engineering Applications.* Prentice Hall, Englewood Cliffs, New Jersey, 1992.

[30] Guy Chavent and Jerome Jaffre. *Mathematical Models and Finite Elements for Reservoir Simulation.* North-Holland, Amsterdam, 1986.

[31] Phillipe G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North-Holland Publishers, Amsterdam, 1978.

[32] Gedeon Dagan. *Flow and Transport in Porous Formations.* Springer–Verlag, Berlin, 1989.

[33] H. K. Dahle, M. S. Espedal, and R. E. Ewing. Characteristic petrov–galerkin subdomain methods for convection diffusion problems. In *Numerical Simulation in Oil Recovery (M.F. Wheeler, ed), IMA*, volume 11, pages 77–88, Berlin, 1988. Springer–Verlag.

[34] M. Dauge. *Elliptic boundary value problems on corner domains.* Number 1341 in Lecture Notes in Mathematics. Springer–Verlag, Berlin, 1988.

[35] B. N. Delaunay. Sur la sphere vide. *Bull. Acad. Science USSR VII: Class. Sci. Math.*, pages 793–800, 1934.

[36] E. P. Doolan, J. J. H. Miller, and W. H. A. Schilders. *Uniform numerical methods for problems with initial and boundary layers.* Boole press, Dublin, 1980.

[37] Jim Douglas, Jr. and Jean E. Roberts. Mixed finite element methods for second order elliptic equations. *Mat. Aplic. Comp.*, 1(1):91–103, 1982.

[38] Jim Douglas, Jr. and Jean E. Roberts. Global estimates for mixed methods for second order elliptic equations. *Math. Comp.*, 44(169):39–52, 1985.

[39] Wiktor Eckhaus. Boundary layers in linear elliptic singular perturbation problems. *SIAM Review*, 14(2):225–270, 1972.

[40] M. S. Espedal and R. E. Ewing. Characteristic Petrov–Galerkin subdomain methods for twophase immiscible flow. *Comput. Methods Appl. Mech. and Engrg.*, 64:113–135, 1987.

[41] R. E. Ewing and J. Wang. Analysis of mixed finite element methods on locally refined grids. *Numer. Math.*, 63:183–194, 1992.

[42] R.E. Ewing, M.A. Celia, P. O'Leary, J.E. Pasciak, and A.T. Vassilev. Parallelization of multiphase models for contaminant transport in porous media. In R. Sincovec, D. Keyes, M. Leuze, L. Petzold, and D. Reed, editors, *Parallel Processing for Scientific Computing*, volume 1, pages 83–91, Philadelphia, Pennsylvania, 1993. SIAM.

[43] Richard Ewing, Raytcho Lazarov, and Panayot Vassilevski. Local refinement techniques for elliptic problems on cell-centered grids I. Error analysis. *Math. Comp.*, 56(194):437–461, 1991.

[44] Richard E. Ewing and Hong Wang. An optimal order estimates for Eulerian–Lagrangian adjoint methods for variable coefficient advection–reaction problems. *SIAM J.Numer. Anal.*, 33(1):318–348, 1996.

[45] Robert Eymard and Femand Sonier. Mathematical and numerical properties of control volume finite element scheme for reservoir simulation. *SPE Reservoir Engineering*, pages 283–289, November 1994.

[46] R. S. Falk and G. R. Richter. Local error estimates for a finite element method for hyperbolic and convection–dominated equations. *SIAM J.Numer. Anal.*, 29(3):730–754, 1992.

[47] Steven Fortune. Voronoi diagrams and Delaunay triangulations. In D. Z. Du and F. K. Hwang, editors, *Computing in Euclidean Geometry*, volume 1 of *Lecture Notes Series of Computing*, pages 193–233. World Scientific, 1992.

[48] P. L. George. *Automatic mesh generation. Application to finite element methods.* John Wiley and Sons, Masson, France, 1991.

[49] David Gilbarg and Nail Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer–Verlag, Berlin, 1983.

[50] J. Glimm and B. Lindquist. Scaling laws for macrodispersion. In T.F. Russell, R.E. Ewing, C.A. Brebbia, W.G. Gray, and G.F. Pinder, editors, *Mathematical Modeling in Water Resources, Computational Methods in Water Resources IX*, volume 2, pages 35–50, London, 1992. Elsevier Applied Sciences.

[51] Ronald Glowinski and Olivier Pironneau. Finite element methods for Navier-Stokes equations. *Annu. Rev. Fluid Mech.*, 24:167–204, 1992.

[52] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman Advance Publishing Program, Boston, 1985.

[53] P. Grisvard. *Singularities in Boundary Value Problems*. Springer–Verlag, Berlin, 1992.

[54] W. Hackbusch. On the regularity of difference schemes. *Ark. Mat.*, 19:71–95, 1981.

[55] W. Hackbusch. On first and second order box schemes. *Computing*, 41:277–296, 1989.

[56] A. F. Hegarty, J. J. H. Miller, E. O'Riordan, and G. I. Shishkin. Special meshes for finite difference approximations to an advection diffusion equation with parabolic layers. *J. Comp. Physics*, 117(1):47–54, 1995.

[57] B. Heinrich. *Finite Difference Methods on Irregular Networks*. Akademie-Verlag, Berlin, 1987.

[58] J. C. Heinrich, P. S. Hyakorn, O. C. Zienkiewicz, and A. R. Mitchell. An "upwind" finite element scheme for two-dimensional convective transport equations. *Int. J. Num. Meth. Engin.*, 11:131–143, 1977.

[59] R. Herbin. An error estimate for a finite volume scheme for a diffusion-convection problem on a triangular mesh. *Numer. Meth. for Partial Diff. Equat.*, 11(2):165–174, 1995.

[60] Charles Hirsch. *Numerical Computation of Internal and External flows, vol. 1 and 2*. John Wiley and Sons, Chichester, 1988.

[61] N. A. Hookey, B. R. Baliga, and C. Prakash. Evaluation and enhancements of some control volume finite–element methods. part 1. Convection–diffusion problems. *Num. Heat Transfer*, 14:255–272, 1988.

[62] T. J. R. Hughes and A. N. Brooks. A multidimensional upwind scheme with no cross-wind diffusion. In T.J.R. Hughes, editor, *Finite Element Methods for Convection Dominated Problems*, pages 19–35. The American Society of Mechanical Engineers, New York, 1979.

[63] Tsutomu Ikeda. *Maximum Principle in Finite Element Models for Convection–Diffusion Phenomena*. North-Holland, Amsterdam, 1983.

[64] A. M. Il'in. Differencing scheme for a differential equation with a small parameter affecting the highest derivative. *Math. Notes, Academy of Sciences, U.S.S.R.*, 6:592–602, 1969.

[65] Jonathan Istok. *Groundwater Modeling by the Finite Element Method*. American Geophysical Union, Washington, District of Columbia, 1989.

[66] Huang Jianguo and Xi Shitong. On the finite volume element method for general self–adjoint elliptic problem. *SIAM J.Numer. Anal.*, to appear.

[67] C. Johnson, A. H. Schatz, and L. B. Walbin. Crosswind smear and pointwise errors in streamline diffusion element methods. *Math. Comp.*, 49(179):25–38, 1987.

[68] Claes Johnson. *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Cambridge University Press, University of Cambridge, 1987.

[69] Ivo Kazda. *Finite Element Techniques in Groundwater Flow Studies*. Elsevier, Amsterdam, 1990.

[70] D. S. Kershaw. Differencing of the diffusion equation in lagrangian hydrodynamic codes. *J. Comput. Phys*, 39:375–395, 1987.

[71] M. S. Krakov. Control volume finite element method for Navier–Stokes equations in vortex–streamfunction formulation. *Num. Heat Transfer, Part B*, 21:125–145, 1992.

[72] H. O. Kreiss, T. A. Manteuffel, B. Swartz, B. Wendroff, and A. B. White, Jr. Supra–convergent schemes on irregular grids. *Math. Comp.*, 47(176):537–554, 1986.

[73] Olga A. Ladyzhenskaya and Nina N. Ural'tseva. *Linear and Quasilinear Elliptic Equations*. Academic Press, New York, 1968.

[74] O. Langlo and M. Espedal. Heterogeneous reservoir models, two phase immiscible flow in 2–d. In T.F. Russell, R.E. Ewing, C.A. Brebbia, W.G. Gray, and G.F. Pinder, editors, *Mathematical Modeling in Water Resources, Computational Methods in Water Resources IX*, volume 2, pages 71–80, London, 1992. Elsevier Applied Sciences.

[75] P. D. Lax and B. Wendroff. Systems of conservation laws. *Comm. Pure and Applied Mathematics*, 13:217–237, 1960.

[76] R. D. Lazarov, V. L. Makarov, and W. Weinelt. On the convergence of difference schemes for the approximation of solutions $u \in W_2^m$ $(m > 0.5)$ of elliptic equations with mixed derivatives. *Numer. Math.*, 44:223–232, 1984.

[77] R. D. Lazarov, I. D. Mishev, and P. S. Vassilevski. Finite volume methods with local refinement for convection-diffusion problems. *Computing*, 53:33–57, 1994.

[78] R. D. Lazarov, I. D. Mishev, and P. S. Vassilevski. Finite volume methods for convection-diffusion problems. *SIAM J.Numer. Anal.*, 33(1):31–55, 1996.

[79] P Le Tallec, T. Sassi, and M. Vidrascu. Three–dimensional domain decomposition methods with nonmatching grids and unstructured coarse solvers. In D. E. Keyes and J. Xu, editors, *Proc. 7th Internat. Symposium on Domain Decomposition*, volume 190 of *Contemporary Mathematics*, pages 61–74, Providence, Rhode Island, 1994. AMS.

[80] J. L. Lions and E. Magenes. *Non–Homogeneous Boundary Value Problems and Applications*. Springer–Verlag, New York, 1972.

[81] J. L. Lions and J. Peetre. Sur une classe d'espaces d'interpolation. In *Publ. Math.*, pages 5–68. Institut des Hautes Etudes Scientifique, Paris, France, 1964.

[82] C. Liu and S. F. McCormick. The finite volume element method (FVE) for planar cavity flow. In *Proc. 11th Internat. Conf. on CFD, Williamsburg, Virginia, June 28–July 2*, 1988.

[83] Mingjin Liu, Junping Wang, and Ning Yan. A discretization for convection dominated diffusion problems based combining the mixed method with the discontinuous Galerkin procedure. 1995. University of Wyoming, in progress.

[84] J. A. Mackenzie and K. W. Morton. Finite volume solutions of convection–diffusion test problems. *Math. Comp.*, 60(201):189–220, 1992.

[85] T.A. Manteuffel and A. B. White, Jr. The numerical solution of second order boundary value problems on nonuniform meshes. *Math. Comp.*, 47(176):511–535, 1986.

[86] R. B. Marr, J. E. Pasciak, and R. F. Peierls. Parallel computation with remote procedure requests using structured messages. Technical Report 47917, Brookhaven National Laboratory, Upton, New York, 1994.

[87] S. F. McCormick. *Multilevel Adaptive Methods for Partial Differential equations*. SIAM, Philadelphia, 1989.

[88] J. J. H. Miller and S. Wang. A new non-conforming Petrov–Galerkin finite element method with triangular elements for a singularly perturbed advection–diffusion problem. *IMA J.Numer. Anal.*, 14:257–276, 1994.

[89] J. J. H. Miller and S. Wang. A tetrahedral mixed finite element method for the stationary semiconductor continuity equations. *SIAM J.Numer. Anal.*, 31:196–216, 1994.

[90] Ilya D. Mishev. Some error estimates for convection–diffusion problems. *Comptes rendus de l´Académie bulgare des Sciences*, 45:17–20, 1992.

[91] Ilya D. Mishev. Preconditioning cell–centered finite difference equations on grids with local refinement. In D. E. Keyes and J. Xu, editors, *Proc. 7th Internat. Symposium on Domain Decomposition*, volume 190 of *Contemporary Mathematics*, pages 283–288, Providence, Rhode Island, 1994. AMS.

[92] Ilya D. Mishev, V. Austel, T. F. Chan, and P. S. Vassilevski. Experiments with algebraic multilevel preconditioners on connection machine. Technical Report CAM 45, UCLA, Los Angeles, California, 1993.

[93] K. W. Morton and E. Süli. Finite volume methods and their analysis. *IMA J.Numer. Anal.*, 11:241–260, 1991.

[94] J. Nečas. *Les méthodes directes dans la théorie des equations elliptiques*. Academia, Prague, 1967.

[95] R. A. Nicolaides. Existence, uniqueness and approximation for generalized saddle appoint problems. *SIAM J.Numer. Anal.*, 19(2):349–357, 1982.

[96] K. Nijima. Pointwise error estimates for streamline diffusion finite element schemes. *Numer. Math.*, 56:707–719, 1990.

[97] J. A. Nitsche. Ein kriterium fur die quasi-optimalitat des ritzchen verfahrens. *Numer. Math.*, 11:346–348, 1968.

[98] L. A. Oganesyan and L. A. Rukhovets. *Variational–difference methods for solving elliptic equations*. Izdatelstvo Akad. Nauk Arm. SSR, Jerevan, 1979. (Russian).

[99] E. O'Riordan and M. Stynes. A globally uniformly convergent finite element method for a singularly perturbed elliptic problem in two dimension. *Math. Comp.*, 57(195):47–62, 1991.

[100] Partnership in Computational Science Consortium. User's guide to GCT: the groundwater contaminant transport simulator. (in preparation).

[101] S. V. Patankar and D. B. Spalding. *Heat and mass transfer in boundary layers.* Morgan-Grampian, London, 1967.

[102] O. A. Pedrosa, Jr. *Use of hybrid grid in reservoir simulation.* PhD thesis, Stanford University, California, 1984.

[103] C. Prakash. Examination of the upwind (donor–cell) formulation in control volume finite–element methods for fluid flow and heat transfer. *Num. Heat Transfer*, 11:401–416, 1986.

[104] C. Prakash. An improved control volume finite–element method for heat and mass transfer and for fluid flow using equal–order velocity–pressure interpolation. *Num. Heat Transfer*, 9:253–276, 1986.

[105] Alfio Quarteroni and Alberto Valli. *Numerical Approximation of Partial Differential Equations.* Springer–Verlag, Berlin, 1994.

[106] G. R. Richter. An explicit finite element method for convection-dominated steady state convection-diffusion equations. *SIAM J.Numer. Anal.*, 28(3):744–759, 1991.

[107] G. R. Richter. The discontinuous Galerkin method with diffusion. *Math. Comp.*, 58(198):509–521, 1992.

[108] J. E. Roberts and J. M.. Thomas. Mixed and hybrid methods. In P. G. Ciarlet and J. L. Lions, editors, *Handbook of Numerical Analysis*, volume 2, pages 524–639. Elsevier Science Publisher B.V., 1991.

[109] H. G. Roos. Ten ways to generate the Il'in and related schemes. *J. Comp. Appl. Math.*, 53:43–59, 1994.

[110] H. L. Royden. *Real Analysis.* Macmillan Publishing Company, New York, 3 edition, 1988.

[111] A. K. Runchal. Convergence and accuracy of three finite difference schemes for a two-dimensional convection and conduction problem. *Int. J. Num. Meth. Engin.*, 4:541–550, 1972.

[112] T. F. Russell and M. F. Wheeler. Finite element and finite difference methods for continuous flows in porous media. In R. E. Ewing, editor, *The Mathematics of Reservoir Simulation*, pages 35–106. SIAM, Philadelphia, 1983.

[113] A. A. Samarskii. On monotone difference scheme for elliptic and parabolic equations in the case of a non-selfadjoint elliptic operator. *Zh. Vychisl. Mat. i Mat. Fiz.*, 5:548–551, 1965. (Russian).

[114] A.A. Samarskii. *Theory of Difference Schemes.* Nauka, Moscow, 1983. (Russian).

[115] Alexander A. Samarskii, Raytcho D. Lazarov, and Vladimir L. Makarov. *Difference Schemes for Differential Equations having Generalized Solutions.* Vysshaya Shkola Publishers, Moscow, 1987. (Russian).

[116] T. Schmidt. Box schemes on quadrilateral meshes. *Computing*, 51:271–292, 1993.

[117] Shagi Di Shih and R. Bruce Kellogg. Asymptotic analysis of a singular perturbation problems. *SIAM J. Math. Anal.*, 14(5):1467–1511, 1987.

[118] G. I. Shishkin. Grid approximation of singularly perturbed boundary value problems with convective terms. *Soviet J. Numer. Anal. Math. Model.*, 5(2):173–187, 1990.

[119] M. I. Sloboditskii. Generalized Sobolev spaces and their applications to boundary value problems for partial differential equations. *Amer. Math. Soc., Tran.*, 57(2):207–275, 1966.

[120] S. L. Sobolev. *Some Applications of Functional Analysis in Mathematical Physics, 3 ed.* American Mathematical Society, Providence, Rhode Island, 1991.

[121] D. B. Spalding. A novel finite difference formulation for differential expression involving both first and second derivatives. *Int. J. Num. Meth. Engin.*, 4:551–559, 1972.

[122] G. Stampacchia. Le probleme de Dirichlet pour les equations elliptiques du second ordre a coefficients discontinus. *Ann. Inst. Fourier*, 15:189–258, 1965.

[123] E. Süli. Convergence of finite volume schemes for Poisson equation on nonuniform meshes. *SIAM J.Numer. Anal.*, 28:1419–1430, 1991.

[124] E. Süli. The accuracy of cell vertex finite volume methods on quadrilateral meshes. *Math. Comp.*, 59(200):359–382, 1992.

[125] Masahita Tabata. A theoretical and computational study of upwind–type finite element methods. In T. Nishida, M. Mimura, and H. Fujii, editors, *Patterns and Waves– Qualitative Analysis of Nonlinear Differential Equations*, pages 319–356. Kinokuniya Company Ltd., Tokyo, 1986.

[126] J. M. Thomas. *Sur l'analyse numerique des methodes d'elements finis hybrides et mixtes.* PhD thesis, Universite Pierre et Marie Curie, Paris, France, 1977.

[127] J. M. Thomas and D. Trujillo. Analysis of finite volume methods. Technical Report 19, Universite de Pau et des Pays de L'adour, Pau, France, 1995.

[128] J. M. Thomas and D. Trujillo. Convergence of finite volume methods. Technical Report 20, Universite de Pau et des Pays de L'adour, Pau, France, 1995.

[129] A. N. Tikhonov and A. A. Samarskii. Homogeneous difference schemes. *Zh. Vychisl. Mat. i Mat. Fiz,*, 2:812–832, 1962. (Russian).

[130] R. R. P. van Nooyen. A Petrov–Galerkin mixed finite element method with exponential fitting. *Numer. Meth. for Partial Diff. Equat.*, 11(5):501–524, 1995.

[131] Richard S. Varga. *Matrix Iterative Analysis.* Prentice–Hall, Inc., Englewoods Cliffs, New Jersey, 1962.

[132] P. S. Vassilevski, S. I. Petrova, and R. D. Lazarov. Finite difference schemes on triangular cell-centered grids with local refinement. *SIAM J. Sci. Stat. Comput.*, 13:1287–1313, 1992.

[133] M. G. Voronoi. Nouvelles applications des parametres continus a la theorie des formes quadratiquues. *J. Reine Angew. Math.*, 134:198–287, 1908.

[134] A. F. Ware, A. K. Parrott, and C. Rogers. A finite volume discretization for porous media flows governed by non-diagonal permeability tensors. In P. A. Thibault and D. M. Bergeron, editors, *Proc. CFD95, Third Annual Conference of the CFD Society of Canada, 25–27 June*, Banff, Alberta, Canada, 1995.

[135] A. Weiser and M. F. Wheeler. On convergence of block-centered finite differences for elliptic problems. *SIAM J. Numer. Anal.*, 25:351–375, 1988.

[136] A. M. Winslow. Numerical solutions of the quasilinear Poisson equation in a nonuniform triangular mesh. *J. Comput. Phys*, 2:149–172, 1967.

[137] David M. Young. *Iterative Solution of Large Linear Systems.* Academic Press, New York, 1971.

[138] G. Zhou and R. Rannacher. Pointwise superconvergence of the streamline diffusion finite element method. *Numer. Meth. for Partial Diff. Equat.*, 12(1):99–122, 1996.