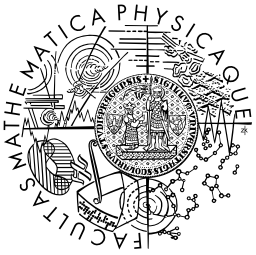

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Jan Lamač

Adaptive methods for singularly perturbed partial differential equations

Department of Numerical Mathematics

Supervisor of the doctoral thesis:	doc. Mgr. Petr Knobloch, Dr.
Study programme:	Mathematics
Study branch:	Scientific and Technical Calculations

Prague 2017

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

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

In Prague June 30, 2017 signature of the author

Title: Adaptive methods for singularly perturbed partial differential equations

Author: Jan Lamač

Department: Department of Numerical Mathematics

Supervisor: doc. Mgr. Petr Knobloch, Dr., Department of Numerical Mathematics

Abstract: This thesis deals with solving singularly perturbed convection-diffusion equations. Firstly, we construct a matched asymptotic expansion of the solution of the singularly perturbed convection-diffusion equation in 1D and derive a formula for the zeroth-order asymptotic expansion in several two-dimensional polygonal domains. Further, we present a set of stabilization methods for solving singularly perturbed problems and prove the uniform convergence of the Il'in-Allen-Southwell scheme in 1D. Finally, we introduce a modification of the streamline upwind Petrov/Galerkin (SUPG) method on convection-oriented meshes. This new method enjoys several profitable properties such as the fulfilment of the discrete maximum principle. Besides the analysis of the method and derivation of a priori error estimates in respective energy norms we also carry out several numerical experiments verifying the theoretical results.

Keywords: asymptotic expansion, singularly perturbed, convection, diffusion, partial differential equations, finite elements, oriented mesh, SUPG, discrete maximum principle

Název práce: Adaptivní metody pro singulárně porušené parciální diferenciální rovnice

Autor: Jan Lamač

Katedra: Katedra numerické matematiky

Vedoucí disertační práce: doc. Mgr. Petr Knobloch, Dr., Katedra numerické matematiky

Abstrakt: V práci se zabýváme řešením singulárně porušených rovnic konvekce-difúze. Nejprve zkonstruujeme sdruženou asymptotickou expanzi řešení singulárně porušené rovnice konvekce-difúze v 1D a odvodíme vzorec pro asymptotickou expanzi nultého řádu v několika dvoudimenzionálních polygonálních oblastech. Následně prezentujeme soubor stabilizačních metod pro řešení singulárně porušených problémů a dokážeme stejnoměrnou konvergenci Il'inova-Allenova-Southwellova schématu v 1D. Nakonec představíme obměnu metody proudnicové difúze (SUPG) na orientovaných sítích. Tato nová metoda s sebou přináší řadu výhodných vlastností, jako například splnění diskrétního principu maxima. Kromě analýzy metody a odvození apriorních odhadů chyby v odpovídajících energetických normách provedeme rovněž několik numerických experimentů potvrzujících teoretické výsledky.

Klíčová slova: asymptotická expanze, singulárně porušené, konvekce, difúze, parciální diferenciální rovnice, konečné prvky, orintovaná síť, SUPG, diskrétní princip maxima

I would like to thank all those who supported me in my long doctoral study and the work on my thesis. I very appreciate the help and guidance received from my supervisor Petr Knobloch and I am grateful for numerous corrections, remarks and advices he gave me throughout my work.

I also wish to express my deepest gratitude to my parents, whose support, patience and understanding made this work possible.

Last but not least, I would like to thank all my friends and colleagues for many inspiring suggestions, remarks and stimulating conversations. I am especially grateful to Petr Milanov, Tomáš Flodrman, Miroslav Zezula, Miloslav Vlasák and Václav Kučera.

My thanks also go to the Grant Agency of the Czech Republic, which provided partial financial support for my research through their grant No. P201/11/1304.

Contents

List of Symbols

$(u, v)_\Omega$	$L^2(\Omega)$-inner product, $(u, v)_\Omega = \int_\Omega u(x)v(x)\,\mathrm{d}x$, page 124
$\|\cdot\|_{k,p,\Omega}$	Norm on the space $W^{k,p}(\Omega)$, $\|u\|_{k,p,\Omega} = \left(\sum_{\|\alpha\|\le k}\|D^\alpha u\|^p_{0,p,\Omega}\right)^{1/p}$, $p \neq \infty$, page 124
$\|\cdot\|_{k,p,\Omega}$	Seminorm on the space $W^{k,p}(\Omega)$, $\|u\|_{k,p,\Omega} = \left(\sum_{\|\alpha\|=k}\|D^\alpha u\|^p_{0,p,\Omega}\right)^{1/p}$, $p \neq \infty$, page 125
$\|\cdot\|_{\infty,d}$	Discrete maximum norm, $\|v_h\|_{\infty,d} = \max\limits_{1\le i\le N}\|v_i\|$, page 37
$\|\|\|\cdot\|\|\|_b$	Energy norm, $\|\|\|v\|\|\|_b^2 = \varepsilon\|v\|^2_{1,\Omega} + \frac{\omega\kappa}{2}\|v\|^2_{0,\Omega} + \sum\limits_{K\in\mathcal{T}_h}\frac{\|\boldsymbol{d}_{K,1}\|}{2\|\boldsymbol{b}_K\|}\|\boldsymbol{b}_K\cdot\nabla v\|^2_{0,K}$, page 68
$\|\|\|\cdot\|\|\|_{b,*}$	Energy norm, $\|\|\|v\|\|\|_{b,*}^2 = \varepsilon\|v\|^2_{1,\Omega} + C_2^*\sum\limits_{K\in\mathcal{T}_h} h_K\|v\|^2_{0,K} + C_b^*\sum\limits_{K\in\mathcal{T}_h}\frac{\|\boldsymbol{d}_{K,1}\|}{2\|\boldsymbol{b}_K\|}\|\boldsymbol{b}_K\cdot\nabla v\|^2_{0,K}$, page 70
a_1	Bilinear form, $a_1(u,\varphi) = \varepsilon(u',\varphi')_\Omega + (bu',\varphi)_\Omega$, page 34
a_h, F_h	Bilinear form and functional used in the definition of the modified SUPG method, page 57
$a_h^{(2)}$, $F_h^{(2)}$	Bilinear form and functional used in the definition of the modified SUPG method, using second order finite elements, page 98
a_Γ, a_h^∞	Bilinear forms used for the L^∞-convergence improvement, page 86
$\boldsymbol{b}_K$	Piecewise constant approximation of $\boldsymbol{b}$, $\boldsymbol{b}_K = -\frac{1}{\|K\|}\left(\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,1}\,\mathrm{d}\boldsymbol{x}\right)\boldsymbol{d}_{K,1}$, page 56
$\boldsymbol{b}_K^I$	Orthogonal L^2-projection of the vector $\boldsymbol{b}$ on a given polynomial space, page 75
$\mathbb{B}_s$	Matrix resulting from the discretization of the convective term, page 104
β_j^s	Weighted average value of $\|\boldsymbol{b}_K\|$ on $\mathcal{C}_j^s$, $\beta_j^s = \frac{1}{\|\mathcal{C}_j^s\|}\sum_{K\subset\mathcal{C}_j^s}\|\boldsymbol{b}_K\|\|K\|$, page 60
β	Minimum of weighted average values of $\|\boldsymbol{b}_K\|$ on $\mathcal{C}_j^s$, $\beta = \min\limits_{j,s}\{\beta_j^s\}$, page 70
C_2^*	Constant appearing in the definition of $\|\|\|\cdot\|\|\|_{b,*}$, $C_2^* = \frac{(4-\delta)\kappa\beta}{2L^2R}(n+1)$, page 70
C_b^*	Constant appearing in the definition of $\|\|\|\cdot\|\|\|_{b,*}$, $C_b^* = \frac{4-\delta}{4}$, page 70
C_{inv}	Constant from Theorem 4.2.2 (inverse inequality), page 127

$\mathcal{C}_j^s$	Cluster surrounding the edge $P_{j-1}^s P_j^s$, $\mathcal{C}_j^s = \cup_{P_{j-1}^s, P_j^s \subset \overline{K}} K$, page 59
C_K	Barycentre of the element K, page 55
C_Π	Constant from Theorem 4.2.3 (approximation property), page 128
C_S, C_E	Constants used for estimating the derivatives of S-decomposition components, page 13
$C_\sigma, C_T, C_z, C_\alpha$	Constants used in the proof of the uniform convergence of the Il'in-Allen-Southwell scheme, page 46
C_X	Constant from Theorem 4.2.1 (interpolation inequality), page 127
$\boldsymbol{d}_{K,j}$	Oriented edge of the n-simplex K, $\boldsymbol{d}_{K,j} = P_{K,n+1} - P_{K,j}$, page 99
g_j	Local Green's function, page 41
Γ_+, Γ_0, Γ_-	Parts of $\partial\Omega$ determined by the sign of $\boldsymbol{b} \cdot \boldsymbol{n}$, page 54
h_j^s	Length of the cluster $\mathcal{C}_j^s$ in the streamline direction, $h_j^s = \lvert\boldsymbol{d}_{K,1}\rvert$, page 60
κ	Mesh structure parameter, $\kappa = \min\limits_{j,s} \left\{ \frac{\lvert\mathcal{C}_j^s\rvert}{\lvert\Omega_j^s\rvert}, \frac{\lvert\mathcal{C}_{j+1}^s\rvert}{\lvert\Omega_j^s\rvert} \right\}$, page 68
κ_{ij}^K, κ_{pq}^j	Coefficients used for the definition of $R_K^{(2)}$ and $\mathbb{B}_s$, page 103
L	Upper bound for the length of the discrete streamline, $L = \max_{j,s}\{N_s h_j^s\}$, page 70
L_h	Method matrix, $(L_h)_{ki} = a_h(\lambda_i, \lambda_k)$, page 83
$\tilde{L}_h$	Scaled method matrix, $(\tilde{L}_h)_{ki} = \frac{(L_h)_{ki}}{\lvert\operatorname{supp}\{\lambda_k\}\rvert}$, page 83
L_h^*	Matrix generated by the Il'in-Allen-Southwell scheme, page 46
$\lambda_{K,j}$	Barycentric coordinates of K satisfying $\lambda_{K,j}(P_{K,i}) = \delta_{ij}$, page 55
μ, ν, μ_K, ν_K	Coefficients used for the L^∞-convergence improvement, page 86
N_s	Number of edges forming the s-th discrete streamline, page 60
Ω_j^s	Patch surrounding the node P_j^s, $\Omega_j^s = \cup_{P_j^s \subset \overline{K}} K$, page 59
$\Omega_{0,j}^s$	Complementary set, $\Omega_{0,j}^s = \Omega_j^s \backslash \left(\mathcal{C}_j^s \cup \mathcal{C}_{j+1}^s \right)$, page 59
P_j^s	j-th mesh node lying on the s-th discrete streamline, page 59
$P_{K,j}$	j-th vertex of the element K, page 55
$P_{K,j}^{(2)}$	Nodes used for a construction of basis functions of $P_2(K)$, page 98
$\mathcal{P}$	Number of discrete streamlines, page 60

Pe, $\mathrm{Pe}(x)$	Péclet number, in 1D there holds $\mathrm{Pe} = \frac{bh}{2\varepsilon}$ and $\mathrm{Pe}(x) = \frac{b(x)h}{2\varepsilon}$, page 35
Pe_Γ, Pe_K	Péclet number, in 2D it is $\mathrm{Pe}_\Gamma = \frac{h(\boldsymbol{b}\cdot\boldsymbol{n}_\Gamma)^2}{2\varepsilon\lvert\boldsymbol{b}\rvert}$ and $\mathrm{Pe}_K = \frac{\lvert\boldsymbol{d}_{K,1}\rvert(\boldsymbol{b}_K\cdot\boldsymbol{n}_K)^2}{2\varepsilon\lvert\boldsymbol{b}_K\rvert}$, in the SUPG method there holds $\mathrm{Pe}_K = \frac{\lVert\boldsymbol{b}\rVert_{0,\infty,K}h_K}{2\varepsilon}$, page 87
$\Pi^{(2)}_{b,K}$	Linear mapping satisfying $\left(\boldsymbol{b}_K^{(1)}\cdot\nabla\varphi^{(2)}_{K,i}, \Pi^{(2)}_{b,K}\left(\varphi^{(2)}_{K,j}\right) - \varphi^{(2)}_{K,j}\right)_K = 0$ for all $i,j = 1,2,\ldots,6$, page 101
$\varphi^{(2)}_{K,i}$	Basis functions of $P_2(K)$ satisfying $\varphi^{(2)}_{K,i}\left(P^{(2)}_{K,j}\right) = \delta_{ij}$ for all $i,j \in \{1,2,\ldots,6\}$, page 98
q_j^s	Flux through $\mathcal{C}_j^s$, $q_j^s = -\sum_{K\subset\mathcal{C}_j^s}\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,1}\,\mathrm{d}\boldsymbol{x}$, page 60
R	Mesh structure parameter, $R = \max\limits_{j,s}\left\{\frac{\max_{K\subset\Omega_j^s}h_K}{h_j^s}\right\}$, page 70
$\mathcal{R}_\Gamma$, $\mathcal{R}_K$	Mesh parameters, $\mathcal{R}_\Gamma = \mathrm{Pe}_\Gamma\big(\coth(\mathrm{Pe}_\Gamma) - 1\big) - 1$, page 88
$R_K^{(2)}$	Linear mapping enabling the fulfilment of the discrete maximum principle, page 102
$S + E$	S-decomposition of the function $u = S + E$, page 13
$\overline{\sigma}$, $\overline{\nu}_K$	Quantities used for proving coercivity of the bilinear form a_h, page 69
τ, τ_j, τ_{upw}, τ_*	Stabilization parameters used in the SUPG method, page 35
$\mathcal{T}_h$	Triangulation of the domain Ω, page 55
θ_K	Mesh parameters, $\theta_K = \frac{1}{\lvert K\rvert}\max\left\{\max\limits_{2\le i\le n}\left\lvert\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,i}\,\mathrm{d}\boldsymbol{x}\right\rvert, \left\lvert\sum\limits_{i=2}^{n}\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,i}\,\mathrm{d}\boldsymbol{x}\right\rvert\right\}$, page 60
u_0	Reduced solution, page 9
$u_{as,m}$	m-th order matched asymptotic expansion, page 11
$u_{g,m}$	m-th order global expansion, page 9
$v_{loc,m}$, $V_{loc}^{k,m}$	m-th order local expansion, page 10
$W^{k,p}(\Omega)$	Sobolev space of functions v for which $\lVert v\rVert_{k,p,\Omega}$ exists and is finite, page 124
$\boldsymbol{w}_K$	Stabilization vector, $\boldsymbol{w}_K = (P_{K,n+1} - C_K) + \frac{\varepsilon\mu_K}{(\boldsymbol{b}_K\cdot\boldsymbol{n}_K)^2}\boldsymbol{b}_K + \frac{\varepsilon\nu_K}{\boldsymbol{b}_K\cdot\boldsymbol{n}_K}\boldsymbol{n}_K$, page 91
ξ	Local variable, in 1D it is $\xi(x) = \frac{1-x}{\varepsilon}$, page 9
$Z_{cor}^{k,m}$	m-th order corner expansion, page 21

Introduction

"God does not play dice with the world," said many times Albert Einstein believing there must be some fundamental laws of nature that could make possible to calculate the speed and position of any particle (Hermanns and Einstein (1983)). Contemporary science formulates these laws in terms of partial differential equations and use them for describing a wide spectrum of phenomena such as fluid dynamics, quantum mechanics, elasticity, heat transfer, electrostatics, electrodynamics, but also dynamics of flocking, pricing of options, crystal growth or gene propagation.

From the theoretical point of view one can be concerned with proving the existence, uniqueness or regularity of the solution of these equations. However, due to the high complexity of partial differential equations it is often impossible to solve them analytically and a numerical approach is typically required.

Within most of this thesis we deal with the numerical solution of a singularly perturbed convection-diffusion equation using the finite element method. It describes the flow of particles, heat, or other physical quantities and since it is singularly perturbed it contains small diffusivity constant (i.e. convection dominates).

The solution of a convection-dominated convection-diffusion equation possesses, in general, interior or boundary layers. These are narrow regions where the solution changes rapidly. If the mesh size is much larger then the width of these regions, the layers cannot be resolved properly, and thus spurious (non-physical) oscillations occur in the numerical solution. In order to remove them, one can use some *adaptive* mesh-refinement algorithm and refine the mesh along layers. However, it does not always bring the desired effect since the mesh width in layer regions should be extremely small.

The second possibility is to *adapt* the numerical method and enhance its stability. Various stabilization strategies have been developed during the last decades. The pioneer contribution to this development was made in the seventies of the last century by Christie et al. (1976) and Heinrich et al. (1977). Christie et al. (1976) used nonsymmetric test functions in the one-dimensional case to achieve the method stability while Heinrich et al. (1977) derived the two-dimensional upwind finite elements.

Many nonconsistent methods were developed until Brooks and Hughes (1982) came with the streamline upwind/Petrov-Galerkin (SUPG) method. It introduces artificial diffusion along streamlines only and the stability is obtained without the loss of accuracy. Since it is consistent one may also derive convergence results. Hughes et al. (1989) then also introduced the Galerkin/least-squares finite element method which represents a conceptual simplification of the SUPG method.

The SUPG method produces oscillation-free solutions in regions, where the solution of the respective partial differential equation is smooth enough and does not change abruptly. However, it is neither monotone nor monotonicity preserving and the spurious oscillations unfortunately persist in narrow regions along sharp (boundary, characteristic) layers. Hence, various (often nonlinear) methods adding further stabilizing terms to the original SUPG method have been proposed. John and Knobloch (2007, 2008) call these methods spurious oscilla-

tions at layers diminishing (SOLD) methods and find Franca et al. (1992) and the modification of Dutra do Carmo and Galeão (1991), Codina (1993) to be the only reasonably promising approaches among the SOLD methods they studied. Nevertheless, they conclude with the result that obtaining oscillation-free solutions is still completely open problem.

The thesis is composed of three chapters and their content is following. We start with the construction of a matched asymptotic expansion of the solution of the convection-diffusion equation in 1D and derive a formula for the zeroth-order asymptotic expansion in several two-dimensional triangular domains (Chapter 1). In Chapter 2 we present a set of stabilization methods, employ them on simple one-dimensional examples and prove the uniform convergence of the Il'in-Allen-Southwell scheme in 1D. Finally, we introduce a modification of the SUPG method on convection-oriented meshes. This new method enjoys several profitable properties such as linearity, fulfillment of the discrete maximum principle or possibility to derive valuable convergence results. Besides the analysis of the method and derivation of a priori error estimates we also carry out several numerical experiments verifying the theoretical results (Chapter 3).

1. Asymptotic expansion

1.1 Introduction

While solving singularly perturbed problems, such as convection-diffusion equation or convection-diffusion-reaction equation, we would like to have some test solution of the respective differential equation (equipped with some simple boundary data) which can confirm or disprove our analyses or methods. This solution can be either exact or asymptotically exact. We can also have the same demand while constructing anisotropic and adaptively refined meshes.

In this context, finding the asymptotically exact solution of the respective differential equation is more convenient. Although it seems that we loose the accuracy of the solution it is not the case, since we can choose the accuracy of the solution ourselves. The construction of the asymptotically exact solutions for differential equations – the method of matched asymptotic expansions – is well described for one-dimensional cases and several two-dimensional cases, see e.g. Eckhaus (1979) or Roos et al. (2008) and the references cited therein. However, for multidimensional cases the construction of the asymptotic expansions of the solutions of partial differential equations is more complicated and in fact treated mostly on simple domains - squares and rectangles in 2D. Moreover, the analysis of the singularly perturbed problems is performed on these rectangular domains, as well. Therefore, the main goal of this chapter is to extend the type of these domains to other convex polygons and enable a generalization of the above mentioned analysis of these problems.

1.2 Asymptotic expansion in 1D

We would like to apply a one-dimensional theory to higher dimensions, and hence we start with a one-dimensional issue. It is well described in Roos et al. (2008) and therefore we proceed according to the theory employed therein.

Let us investigate the singularly perturbed problem

$$\begin{aligned} Lu \;:=\; -\varepsilon u'' + b(x)u' \;&=\; f \quad \text{in } \Omega = (0,1), && (1.1)\\ u(0) \;=\; u(1) \;&=\; 0, && (1.2) \end{aligned}$$

with $b(x) > \underline{\beta} > 0$ and $1 \gg \varepsilon > 0$. Since we are going to use the derivatives of functions b and f, let us also assume that b and f are sufficiently smooth on $[0,1]$.

It is sometimes difficult to compute the exact solution u of the boundary value problem (1.1)–(1.2), therefore we would like to find some approximation of it. We use the fact that ε is considered to be a very small positive number and use an asymptotic expansion to approximate the solution u.

Definition 1.2.1. *The function v is an* asymptotic expansion *of order m of the function u (in the maximum norm) if there exists a constant C independent of ε such that*

$$|u(x) - v(x)| \;\le\; C\,\varepsilon^{m+1} \quad \textit{for all } x \in [0,1] \textit{ and } \varepsilon \textit{ sufficiently small.} \qquad (1.3)$$

The previous definition implies that if v is the asymptotic expansion of order m of the function u and $c \in \mathbb{R}$ is any constant, then $v + c\varepsilon^{m+1}$ is also an asymptotic expansion of order m of the function u. Thus, v is not a uniquely defined function and our aim is to find any v satisfying (1.3). One possibility is to construct the *matched asymptotic expansion* $u_{as,m}$ which we will describe now.

Firstly, we formally set $\varepsilon = 0$ in the equation (1.1) and obtain the so-called *reduced problem*

$$L_0 u_0 := b(x)u_0'(x) = f(x) \quad \text{in} \quad \Omega, \tag{1.4}$$

$$u(0) = 0. \tag{1.5}$$

The reduced solution u_0 is, in fact, the first term of the so-called *global (or regular) expansion* of the solution u, which is a good approximation of u away from the boundary layers.

Definition 1.2.2. *We call the function $u_{g,m}$ the m-th order global expansion of the function u when $u_{g,m} = \sum_{j=0}^{m} \varepsilon^j u_j$, where u_0 is the reduced solution and u_j, $j \in \{1, 2, \ldots, m\}$, satisfy*

$$L_0 u_j = u_{j-1}'', \qquad u_j(0) = 0. \tag{1.6}$$

This definition immediately implies that

$$L(u - u_{g,m}) = f + \varepsilon u_{g,m}'' - L_0 u_{g,m} = f + \varepsilon \sum_{j=0}^{m} \varepsilon^j u_j'' - f - \sum_{j=1}^{m} \varepsilon^j L_0 u_j =$$

$$= \sum_{j=0}^{m} \varepsilon^{j+1} u_j'' - \sum_{j=1}^{m} \varepsilon^j u_{j-1}'' = \varepsilon^{m+1} u_m'', \tag{1.7}$$

$$(u - u_{g,m})(0) = 0 \quad \text{and} \tag{1.8}$$

$$(u - u_{g,m})(1) = -\sum_{j=0}^{m} \varepsilon^j u_j(1). \tag{1.9}$$

Since $u_j(1)$ generally does not vanish for all $j = 0, 1, \ldots, m$, local correction terms at $x = 1$ must be added. Therefore, we introduce the local variable

$$\xi(x) = \frac{1 - x}{\varepsilon} \qquad \Rightarrow \qquad x(\xi) = 1 - \varepsilon\xi, \tag{1.10}$$

and consider the following Taylor's expansion of the function b at $x = 1$

$$b(x) = \sum_{j=0}^{\infty} b_j (1 - x)^j. \tag{1.11}$$

If we now define a new function U of the variable ξ by setting $U(\xi) = u(x(\xi))$ (i.e. $u(x) = U(\xi(x))$), then there holds

$$\frac{\mathrm{d}u}{\mathrm{d}x} = \frac{\mathrm{d}U}{\mathrm{d}\xi}\frac{\mathrm{d}\xi}{\mathrm{d}x} = -\frac{1}{\varepsilon}\frac{\mathrm{d}U}{\mathrm{d}\xi} \qquad \text{and} \qquad \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} = \frac{1}{\varepsilon^2}\frac{\mathrm{d}^2 U}{\mathrm{d}\xi^2}. \tag{1.12}$$

Consequently, we can express the differential operator L in terms of ξ as

$$Lu = \frac{1}{\varepsilon}\left(-\frac{\mathrm{d}^2 U}{\mathrm{d}\xi^2} - b(1 - \varepsilon\xi)\frac{\mathrm{d}U}{\mathrm{d}\xi}\right) = \tag{1.13}$$

$$= \frac{1}{\varepsilon}\left(-\frac{\mathrm{d}^2 U}{\mathrm{d}\xi^2} - b_0\frac{\mathrm{d}U}{\mathrm{d}\xi}\right) - \frac{1}{\varepsilon}\sum_{j=1}^{\infty} b_j \varepsilon^j \xi^j \frac{\mathrm{d}U}{\mathrm{d}\xi} =: \mathcal{L}U. \tag{1.14}$$

We observe that the stretching factor $1/\varepsilon$ in the definition of ξ has been chosen in such a way that the coefficients at the first and the second derivative in (1.13) are of the same order with respect to ε.

In the expression (1.14), let us denote

$$L_0^* \;:=\; -\frac{\mathrm{d}^2}{\mathrm{d}\xi^2} - b_0\frac{\mathrm{d}}{\mathrm{d}\xi} \qquad \text{and} \qquad L_j^* \;:=\; -b_j\xi^j\frac{\mathrm{d}}{\mathrm{d}\xi},\ j\geq 1, \tag{1.15}$$

and let us introduce the *local expansion* of the m-th order

$$v_{loc,m}(\xi) \;=\; \sum_{j=0}^{m}\varepsilon^j v_j(\xi), \tag{1.16}$$

where v_j, $j=0,1,\ldots,m$, are functions independent of ε which will be defined in what follows. Combining this definition with (1.14) then yields (we denote $s=j+k$, $q=s-m-2$ and $t=j-s+m+1=j-q-1$)

$$\begin{aligned}
\mathcal{L}v_{loc,m+1} \;&=\; \frac{1}{\varepsilon}\sum_{j=0}^{\infty}\varepsilon^j L_j^* v_{loc,m+1} \;=\; \frac{1}{\varepsilon}\sum_{j=0}^{\infty}\varepsilon^j L_j^*\sum_{k=0}^{m+1}\varepsilon^k v_k \;=\; \frac{1}{\varepsilon}\sum_{j=0}^{\infty}\sum_{k=0}^{m+1}\varepsilon^{j+k}L_j^* v_k \;=\\
&=\; \frac{1}{\varepsilon}\sum_{j=0}^{\infty}\sum_{s=j}^{m+1+j}\varepsilon^s L_j^* v_{s-j} \;=\; \frac{1}{\varepsilon}\sum_{s=0}^{m+1}\varepsilon^s\sum_{j=0}^{s}L_j^* v_{s-j} + \frac{1}{\varepsilon}\sum_{s=m+2}^{\infty}\varepsilon^s\sum_{j=s-m-1}^{s}L_j^* v_{s-j} \;=\\
&=\; \frac{1}{\varepsilon}\sum_{s=0}^{m+1}\varepsilon^s\sum_{j=0}^{s}L_j^* v_{s-j} + \varepsilon^{m+1}\sum_{q=0}^{\infty}\varepsilon^q\sum_{t=0}^{m+1}L_{t+q+1}^* v_{m+1-t},
\end{aligned} \tag{1.17}$$

where the sum equality $\sum_{j=0}^{\infty}\sum_{s=j}^{m+1+j} = \sum_{s=0}^{m+1}\sum_{j=0}^{s} + \sum_{s=m+2}^{\infty}\sum_{j=s-m-1}^{s}$ results from the discrete Fubini theorem (see Remark 1.2.1, page 10).

Thus, in order to obtain $\mathcal{L}v_{loc,m+1} = \mathcal{O}(\varepsilon^{m+1})$ we require $\sum_{j=0}^{s}L_j^* v_{s-j} = 0$ for all $s=0,1,\ldots,m+1$. This is a system of ordinary differential equations, so-called *boundary layer equations*, which can be solved recursively, i.e. for each $s=0,1,\ldots,m+1$ we solve the differential equation

$$L_0^* v_s \;=\; -\sum_{j=1}^{s}L_j^* v_{s-j}\ \left(=\ -\sum_{j=0}^{s-1}L_{s-j}^* v_j\right), \quad \text{in } (0,\infty), \tag{1.18}$$

equipped with the boundary conditions $v_s(0) = -u_s(1)$ and $\lim_{\xi\to\infty} v_s(\xi) = 0$ (when $s=0$ we consider $-\sum_{j=1}^{s}L_j^* v_{s-j} = 0$). For instance, the first two solutions of this system (the first-order correction and the second-order correction) have the following form

$$v_0(\xi) \;=\; -u_0(1)\,\mathrm{e}^{-b_0\xi} \;=\; -u_0(1)\,\mathrm{e}^{-b(1)\xi}, \tag{1.19}$$

$$\begin{aligned}
v_1(\xi) \;&=\; \left(-u_1(1) + \frac{b_1\,u_0(1)\,\xi}{b_0} + \frac{b_1\,u_0(1)\,\xi^2}{2}\right)\mathrm{e}^{-b_0\xi} \;=\\
&=\; -\left(u_1(1) + \frac{b'(1)\,u_0(1)\,\xi}{b(1)} + \frac{b'(1)\,u_0(1)\,\xi^2}{2}\right)\mathrm{e}^{-b(1)\xi}.
\end{aligned} \tag{1.20}$$

Remark 1.2.1. The above applied equality

$$\sum_{j=0}^{\infty}\sum_{s=j}^{m+1+j}a_{js} \;=\; \sum_{s=0}^{m+1}\sum_{j=0}^{s}a_{js} + \sum_{s=m+2}^{\infty}\sum_{j=s-m-1}^{s}a_{js} \tag{1.21}$$

does not hold for arbitrary functions (or numbers) a_{js}. For instance, if $a_{js} = \delta_{j,s} - \delta_{j+1,s}$, then $\sum_{j=0}^{\infty}\sum_{s=j}^{m+1+j} a_{js} = \sum_{j=0}^{\infty}(1-1) = 0$. On the other hand, there holds $\sum_{s=0}^{m+1}\sum_{j=0}^{s} a_{js} = a_{00} + \sum_{s=1}^{m+1}(-1+1) = 1+0 = 1$ and $\sum_{s=m+2}^{\infty}\sum_{j=s-m-1}^{s} a_{js} = \sum_{s=m+2}^{\infty}(-1+1) = 0$. Hence, some conditions on a_{js} are necessary in the case of infinite sums.

If $a_{js} \geq 0$ for all considered j and s or if $\sum_{j=0}^{\infty}\sum_{s=j}^{m+1+j} |a_{js}| < \infty$, then the equality (1.21) results from the Fubini and the Tonelli theorems (see, e.g., Wheeden and Zygmund (2015), Chapter 6). However, since the terms in (1.17) with $s \geq m+2$ are all $\mathcal{O}(\varepsilon^{m+1})$, the interchange of sums in (1.17) is always correct, up to some $\mathcal{O}(\varepsilon^{m+1})$-term.

Summing global expansion and local expansion together we obtain the following theorem. Since the proof in Roos et al. (2008) does not go into details we present the full proof here.

Theorem 1.2.1. *For sufficiently smooth data and $b(x) > \underline{\beta} > 0$ in $[0,1]$ the solution of the boundary value problem (1.1)–(1.2) has a matched asymptotic expansion of the m-th order of the form*

$$u_{as,m}(x) \;=\; \sum_{j=0}^{m} \varepsilon^j u_j(x) + \sum_{j=0}^{m} \varepsilon^j v_j\left(\frac{1-x}{\varepsilon}\right), \tag{1.22}$$

such that for any sufficiently small fixed constant ε_0 there holds

$$|u(x) - u_{as,m}(x)| \;\leq\; C\varepsilon^{m+1} \qquad \text{for } x \in [0,1] \text{ and } \varepsilon \leq \varepsilon_0. \tag{1.23}$$

The constant C is independent of x and ε.

Proof. Let us consider $u^*_{as,m}(x) = u_{as,m}(x) + \varepsilon^{m+1} v_{m+1}\left(\frac{1-x}{\varepsilon}\right)$, then

$$\begin{aligned} L\Big(u(x) - u^*_{as,m}(x)\Big) \;&=\; L\Big(u(x) - u_{g,m}(x)\Big) + L\Big(u_{g,m}(x) - u^*_{as,m}(x)\Big) \;= \\ &=\; \varepsilon^{m+1} u''_m(x) - L\left(v_{loc,m+1}\left(\frac{1-x}{\varepsilon}\right)\right) \;= \\ &=\; \varepsilon^{m+1} u''_m(x) - \mathcal{L} v_{loc,m+1}(\xi) \;=\; \mathcal{O}(\varepsilon^{m+1}) \end{aligned} \tag{1.24}$$

and

$$\Big(u - u^*_{as,m}\Big)(0) \;=\; 0 - \sum_{j=0}^{m} \varepsilon^j v_j(1/\varepsilon) \;=\; \mathcal{O}(\varepsilon^{\mu}) \quad \text{for any } \mu > 0, \tag{1.25}$$

$$\Big(u - u^*_{as,m}\Big)(1) \;=\; -\sum_{j=0}^{m} \varepsilon^j \Big(u_j(1) + v_j(0)\Big) - \varepsilon^{m+1} v_{m+1}(0) \;=\; \varepsilon^{m+1} u_{m+1}(1). \tag{1.26}$$

Hence, one may find a positive constant C_λ independent of ε such that there holds $\Big|L(u(x) - u^*_{as,m}(x))\Big| \leq C_\lambda \varepsilon^{m+1}$ for all $x \in (0,1)$ and $\Big|(u - u^*_{as,m})(0)\Big| \leq C_\lambda \varepsilon^{m+1}$, $\Big|(u - u^*_{as,m})(1)\Big| \leq C_\lambda \varepsilon^{m+1}$.

If we now denote $w(x) = \max\left\{1, \frac{1}{\underline{\beta}}\right\} C_\lambda \varepsilon^{m+1}(x+1)$, we obtain the following

inequalities

$$Lw \;\geq\; \frac{b(x)}{\underline{\beta}} C_\lambda \varepsilon^{m+1} \;\geq\; L\Big(u - u^*_{as,m}\Big) \;\geq\; -\frac{b(x)}{\underline{\beta}} C_\lambda \varepsilon^{m+1} \;\geq\; -Lw, \quad (1.27)$$

$$w(0) \;\geq\; C_\lambda \varepsilon^{m+1} \;\geq\; \Big(u - u^*_{as,m}\Big)(0) \;\geq\; -C_\lambda \varepsilon^{m+1} \;\geq\; -w(0), \quad (1.28)$$

$$w(1) \;\geq\; 2C_\lambda \varepsilon^{m+1} \;\geq\; \Big(u - u^*_{as,m}\Big)(1) \;\geq\; -2C_\lambda \varepsilon^{m+1} \;\geq\; -w(1). \quad (1.29)$$

The comparison principle (Theorem 4.1.6, page 126) then implies that

$$w(x) \;\geq\; u(x) - u^*_{as,m}(x) \;\geq\; -w(x) \qquad \text{for all } x \in [0,1]. \quad (1.30)$$

It means that

$$\begin{aligned} |u(x) - u_{as,m}(x)| \;&\leq\; |u(x) - u^*_{as,m}(x)| + \varepsilon^{m+1} \left| v_{m+1}\left(\frac{1-x}{\varepsilon}\right)\right| \;\leq \\ &\leq\; \|w\|_{0,\infty,\Omega} + \varepsilon^{m+1} \|v_{m+1}\|_{0,\infty,(0,\infty)} \;\leq \\ &\leq\; \left(2C_\lambda \max\left\{1, \frac{1}{\underline{\beta}}\right\} + \|v_{m+1}\|_{0,\infty,(0,\infty)}\right) \varepsilon^{m+1}. \end{aligned} \quad (1.31)$$

Since both C_λ and v_{m+1} do not depend on ε we obtain (1.23). □

Remark 1.2.2. Using the above derived expression (1.19) the matched asymptotic expansion of the zeroth order of the solution u of the boundary value problem (1.1)–(1.2) has the form

$$u_{as,0}(x) \;=\; u_0(x) + v_0\left(\frac{1-x}{\varepsilon}\right) \;=\; u_0(x) - u_0(1)\,\exp\left(-b(1)\frac{1-x}{\varepsilon}\right). \quad (1.32)$$

However, sometimes it is convenient to use another version of the zeroth-order asymptotic expansion

$$\tilde{u}_{as,0}(x) \;=\; u_0(x)\left(1 - \exp\left(-b(1)\frac{1-x}{\varepsilon}\right)\right). \quad (1.33)$$

To see that this is also an asymptotic expansion of the zeroth order of the solution u of the boundary value problem (1.1)–(1.2) we estimate the difference $\tilde{u}_{as,0}(x) - u_{as,0}(x)$. Therefore, let us denote

$$e_0(x) \;:=\; \tilde{u}_{as,0}(x) - u_{as,0}(x) \;=\; \Big(u_0(x) - u_0(1)\Big)\,\exp\left(-b(1)\frac{1-x}{\varepsilon}\right), \quad (1.34)$$

$$\overline{x} \;=\; \arg\max_{x\in[0,1]} |e_0(x)|. \quad (1.35)$$

Since there holds $e_0(1) = 0$ and $e_0(0) = -u_0(1)\,\exp\left(-b(1)/\varepsilon\right)$ then either $|e_0(x)| \leq |u_0(1)|\,\exp\left(-b(1)/\varepsilon\right) \leq \frac{|u_0(1)|}{b(1)}\,\varepsilon$ for $x \in [0,1]$ or $e_0'(\overline{x}) = 0$, i.e. $u_0'(\overline{x}) + \frac{b(1)}{\varepsilon}\Big(u_0(\overline{x}) - u_0(1)\Big) = 0$. In the latter case, using $bu_0' = f$, we can estimate

$$|e_0(x)| \;\leq\; |e_0(\overline{x})| \;=\; \varepsilon\,\frac{|f(\overline{x})|}{b(1)b(\overline{x})}\,\exp\left(-b(1)\frac{1-\overline{x}}{\varepsilon}\right) \;\leq\; \varepsilon\frac{\|f\|_{0,\infty,\Omega}}{b(1)\underline{\beta}}. \quad (1.36)$$

Thus $|u(x) - \tilde{u}_{as,0}(x)| \leq |u(x) - u_{as,0}(x)| + |e_0(x)| \leq C(b,f)\varepsilon$.

Using the matched asymptotic expansion of the solution u we may also construct the so-called *S-decomposition* of the solution u. It decomposes the solution of the boundary value problem (1.1)–(1.2) into a smooth part S (with derivatives bounded uniformly in ε) satisfying $LS = f$ and a layer part E with a property $LE = 0$.

Definition 1.2.3. *Let u by the solution of the boundary value problem (1.1)–(1.2), then $u = S + E$ is an S-decomposition of u, if*

$$S(x) = \sum_{j=0}^{m} \varepsilon^j u_j(x) + \varepsilon^{m+1} u^*_{m+1}(x), \tag{1.37}$$

$$E(x) = \sum_{j=0}^{m} \varepsilon^j v_j\left(\frac{1-x}{\varepsilon}\right) + \varepsilon^{m+1} v^*_{m+1}(x), \tag{1.38}$$

where u_j,v_j, $j = 0, 1, \ldots, m$, are standard terms of the matched asymptotic expansion, whereas u^*_{m+1} and v^*_{m+1} are solutions of the differential equations

$$Lu^*_{m+1} = u''_m \qquad \text{in } (0,1), \tag{1.39}$$

$$Lv^*_{m+1} = -\varepsilon^{-(m+1)} L\left(\sum_{j=0}^{m} \varepsilon^j v_j\left(\frac{1-x}{\varepsilon}\right)\right) \qquad \text{in } (0,1), \tag{1.40}$$

equipped with the boundary conditions

$$\begin{array}{rcl} u^*_{m+1}(0) &=& 0 \\ u^*_{m+1}(1) &=& 0 \end{array} \qquad \text{and} \qquad \begin{array}{rcl} v^*_{m+1}(0) &=& -\varepsilon^{-(m+1)} \sum_{j=0}^{m} \varepsilon^j v_j(0) \\ v^*_{m+1}(1) &=& 0 \end{array}. \tag{1.41}$$

Remark 1.2.3. If the data of the boundary value problem (1.1)–(1.2) are constant, then choosing $m = 0$ leads to the S-decomposition $S(x) = u_0(x) + \varepsilon u^*_1(x)$ and $E(x) = v_0\left(\frac{1-x}{\varepsilon}\right) + \varepsilon v^*_1(x)$, where

$$-\varepsilon(u^*_1)'' + b(u^*_1)' = u''_0 = \left(\frac{f}{b}x\right)'' = 0 \quad \text{in } (0,1), \tag{1.42}$$

$$-\varepsilon(v^*_1)'' + b(v^*_1)' = -\frac{1}{\varepsilon} L\left(-\frac{f}{b}\exp\left(-\frac{b}{\varepsilon}(1-x)\right)\right) = 0 \quad \text{in } (0,1), \tag{1.43}$$

with $u^*_1(0) = u^*_1(1) = 0$ and $v^*_1(0) = \frac{f}{b\varepsilon}\exp(-b/\varepsilon)$, $v^*_1(1) = 0$. It means that $u^*_1 \equiv 0$ and

$$v^*_1(x) = \frac{f}{b\varepsilon}\,\frac{\exp\left(-\frac{b}{\varepsilon}\right) - \exp\left(-\frac{b}{\varepsilon}(2-x)\right)}{1 - \exp\left(-\frac{b}{\varepsilon}\right)}. \tag{1.44}$$

Therefore, the S-decomposition takes the form

$$S(x) = \frac{f}{b}\,x \tag{1.45}$$

$$E(x) = -\frac{f}{b}\exp\left(-\frac{b}{\varepsilon}(1-x)\right) + \varepsilon\,\frac{f}{b\varepsilon}\,\frac{\exp\left(-\frac{b}{\varepsilon}\right) - \exp\left(-\frac{b}{\varepsilon}(2-x)\right)}{1 - \exp\left(-\frac{b}{\varepsilon}\right)}. \tag{1.46}$$

Let us also mention that despite the presence of the factor $1/\varepsilon$ the function v^*_1 for all $x \in [0,1]$ satisfies $|v^*_1| \le |v^*_1(0)| = \frac{|f|}{b\varepsilon}\exp(-b/\varepsilon) \le \frac{|f|}{b\varepsilon}\,\frac{\varepsilon}{\mathrm{e}b} = \frac{|f|}{b^2\mathrm{e}}$, where we have used the inequality $\exp(-x) \le \frac{1}{\mathrm{e}x}$, which is valid for all $x > 0$.

Using the S-decomposition of the solution u and setting $q = m+1$ in Definition 1.2.3 we may prove the following lemma.

Lemma 1.2.1 (S-decomposition)**.** *Let q be some positive integer. Consider the boundary value problem (1.1)–(1.2) with $b(x) > \underline{\beta} > 0$ and sufficiently smooth data. Its solution u can be decomposed as $u = S + E$, where the smooth part S satisfies $LS = f$ and*

$$|S^{(j)}(x)| \;\le\; C_S \qquad \text{for } 0 \le j \le q, \tag{1.47}$$

while the layer part E satisfies $LE = 0$ and

$$|E^{(j)}(x)| \;\le\; C_E\, \varepsilon^{-j} \exp\left(-\frac{\underline{\beta}(1-x)}{\varepsilon}\right) \qquad \text{for } 0 \le j \le q. \tag{1.48}$$

Here C_S and C_E are positive constants independent of ε.

Proof. One can find the proof, e.g., in (Roos et al., 2008, pages 23–24). □

1.3 Asymptotic expansion in two dimensions

Analogously to the one-dimensional case we construct the asymptotic expansion using global and local expansions in the two-dimensional case. The main difference will be the presence of multiple boundary layers (caused by the presence of multiple boundary edges). Moreover, in some cases the inner (parabolic, characteristic) layers occur in the solution which causes complications while constructing the asymptotic expansion.

1.3.1 Model equation and reduced problem

As in the one-dimensional case, the model equation for our purposes will be a scalar convection-diffusion equation

$$\begin{aligned} Lu := -\varepsilon\Delta u(x,y) + \boldsymbol{b}(x,y)\cdot\nabla u(x,y) &= f(x,y) \quad \text{in} \quad \Omega \subset \mathbb{R}^2, & (1.49)\\ u(x,y) &= 0 \quad \text{on} \quad \partial\Omega, & (1.50) \end{aligned}$$

where Ω is a *convex* polygonal domain with boundary $\partial\Omega$ satisfying

$$\overline{\partial\Omega} = \overline{\Gamma}_+ \cup \overline{\Gamma}_0 \cup \overline{\Gamma}_- \quad \text{and} \quad \Gamma_+ \cap \Gamma_0 = \Gamma_0 \cap \Gamma_- = \Gamma_- \cap \Gamma_+ = \emptyset \tag{1.51}$$

with Γ_+, Γ_0 and Γ_- defined as follows:

$$\begin{aligned} \Gamma_+ &= \{(x,y) \in \partial\Omega, \boldsymbol{b}(x,y)\cdot\boldsymbol{n}(x,y) > 0\},\\ \Gamma_0 &= \{(x,y) \in \partial\Omega, \boldsymbol{b}(x,y)\cdot\boldsymbol{n}(x,y) = 0\},\\ \Gamma_- &= \{(x,y) \in \partial\Omega, \boldsymbol{b}(x,y)\cdot\boldsymbol{n}(x,y) < 0\}. \end{aligned} \tag{1.52}$$

Here $\boldsymbol{n}(x,y)$ denotes a unit vector at $(x,y) \in \partial\Omega$ orthogonal to the boundary $\partial\Omega$.

Since we are not interested in solving the equation (1.49) for general data but in finding some test solution for given domain, we can confine ourselves to sufficiently smooth data, namely $\boldsymbol{b} \in C^1(\overline{\Omega})^2$ and $f \in L^2(\Omega)$. In what follows

we shall also consider that the vector $\boldsymbol{b}$ possesses the Taylor expansion in $\overline{\Omega}$, particularly in the neighborhood of $\partial\Omega$.

As $\varepsilon \to 0+$, the equation (1.49) becomes singularly perturbed and near the boundary Γ_+ it is usually difficult to compute the solution numerically. Thus we would like to determine the asymptotic expansion of the solution of the equation (1.49) near the boundary Γ_+. At first we again formally set $\varepsilon = 0$ in the equation (1.49) and obtain the *reduced problem*

$$\begin{aligned} \boldsymbol{b}(x,y)\cdot\nabla u_0(x,y) &= f(x,y) \quad \text{in} \quad \Omega \subset \mathbb{R}^2, && (1.53)\\ u_0(x,y) &= 0 \quad \text{on} \quad \Gamma_-, && (1.54) \end{aligned}$$

where we have to consider only the boundary condition on Γ_- due to the cancellation law (see Roos et al. (2008), p. 12 and 35, for details). The problem (1.53)–(1.54) is a hyperbolic problem and we assume that the solution of this problem is known, more specifically, we consider only problems with a sufficiently smooth and (analytically) computable reduced solution u_0. Some basic results on existence, uniqueness and regularity of the solution of (1.53)–(1.54) can be found in Goering et al. (1983). As we already know, the reduced solution u_0 is the first term of the *global (or regular) expansion* of the solution u, which is a good approximation of u away from the layers.

Definition 1.3.1. *We call the function $u_{g,m}$ the m-th order global expansion of the function u when $u_{g,m} = \sum_{j=0}^{m} \varepsilon^j u_j$, where u_0 is the reduced solution and u_j, $j \in \{1, 2, \ldots, m\}$ satisfy*

$$\begin{aligned} [L_0 u_j](x,y) := \boldsymbol{b}(x,y)\cdot\nabla u_j(x,y) &= \Delta u_{j-1}(x,y) \quad \textit{in} \quad \Omega \subset \mathbb{R}^2, && (1.55)\\ u_j(x,y) &= 0 \quad \textit{on} \quad \Gamma_-. && (1.56) \end{aligned}$$

From this definition it follows that

$$\begin{aligned} L(u-u_{g,m}) &= f + \varepsilon\Delta u_{g,m} - L_0 u_{g,m} = \varepsilon\sum_{j=0}^{m}\varepsilon^j\Delta u_j - \sum_{j=1}^{m}\varepsilon^j L_0 u_j = \\ &= \sum_{j=0}^{m}\varepsilon^{j+1}\Delta u_j - \sum_{j=1}^{m}\varepsilon^j\Delta u_{j-1} = \varepsilon^{m+1}\Delta u_m, && (1.57)\\ u-u_{g,m} &= 0 \qquad \text{on } \Gamma_- \qquad \text{and} && (1.58)\\ u-u_{g,m} &= -u_{g,m} \quad \text{on } \Gamma_+\cup\Gamma_0. && (1.59) \end{aligned}$$

Due to the last property, considering $\varepsilon \ll \|u_{g,m}\|_{\infty,\Omega}$, the comparison principle (Theorem 4.1.6, page 126) yields only $\|u-u_{g,m}\|_{\infty,\Omega} \le \|u_{g,m}\|_{\infty,\Omega}$. This is reason why the local correction terms must be introduced. (See Definition 4.1.1, page 124, for the definition of norms.)

1.3.2 Local expansion in exponential layers

In order to construct the local correction terms we introduce new coordinates in the neighborhood of Γ_+. For this purpose let us assume that $\Gamma_0 = \emptyset$ and that

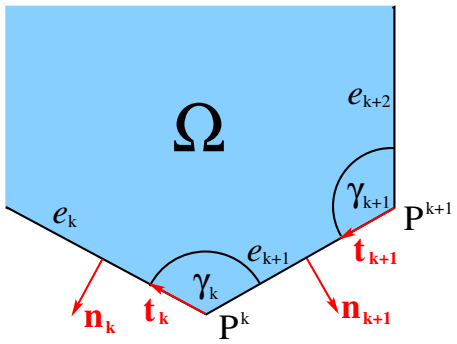

Figure 1.1: A part of the general convex domain Ω.

there are only two vertices $\{P^0, P^H\} = \overline{\Gamma}_- \cap \overline{\Gamma}_+$ and consider $\overline{\Gamma}_+ = \cup_{k=1}^H e_k$, where e_k are the edges of $\overline{\Gamma}_+$. Then $P^0 \in e_1$, $P^H \in e_H$ and the remaining vertices of $\overline{\Gamma}_+$ satisfy $P^k = e_k \cap e_{k+1}, k = 1, \ldots, H-1$.

The transformation of coordinates Ψ_k corresponding to the edge e_k, $k = 1, 2, \ldots, H$, is now defined as $\Psi_k : (x, y) \to (\xi_k, \eta_k)$, where

$$\xi_k(x,y) = (P_y^{k-1} - y)\cos\alpha_k - (P_x^{k-1} - x)\sin\alpha_k = (P^{k-1} - X)\cdot \boldsymbol{n}_k, \quad (1.60)$$

$$\eta_k(x,y) = (P_x^{k-1} - x)\cos\alpha_k + (P_y^{k-1} - y)\sin\alpha_k = (P^{k-1} - X)\cdot \boldsymbol{t}_k. \quad (1.61)$$

Here $P^{k-1} = [P_x^{k-1}, P_y^{k-1}]$, $\boldsymbol{t}_k = (\cos\alpha_k, \sin\alpha_k)^T$ is the unit (tangent) vector parallel to the edge e_k and $\boldsymbol{n}_k = (-\sin\alpha_k, \cos\alpha_k)^T$ is the normal vector, orthogonal to the edge e_k, see Figure 1.1. Further, let us denote $d_k = \eta_k(P^k)$ and due to the convexity of Ω, we may for simplicity assume that the domain Ω is oriented in such a way that $\alpha_k \in [0, 2\pi)$ and $\alpha_k < \alpha_{k+1}$ for all $k = 1, 2, \ldots, H-1$. This notation also implies that the angle corresponding to the vertex P^k is equal to $\gamma_k = \pi + \alpha_k - \alpha_{k+1}$.

Next we stretch the scale in the ξ_k direction and define the transformation $\Psi_k^\varepsilon : (x, y) \to \left(\frac{\xi_k}{\varepsilon}, \eta_k\right)$ using the same ξ_k, η_k, i.e. $(\xi_k, \eta_k) = \Psi_k(x, y)$.

Under the transformation Ψ_k^ε the differential operator L (cf. (1.49)) changes into

$$L_k^\Psi := \frac{1}{\varepsilon}\left(-\frac{\partial^2}{\partial\xi_k^2} + B_1^k(\varepsilon\xi_k, \eta_k)\frac{\partial}{\partial\xi_k}\right) - \varepsilon\frac{\partial^2}{\partial\eta_k^2} + B_2^k(\varepsilon\xi_k, \eta_k)\frac{\partial}{\partial\eta_k}, \quad (1.62)$$

where

$$B_1^k(\xi_k, \eta_k) = -\boldsymbol{b}\left(\Psi_k^{-1}(\xi_k, \eta_k)\right)\cdot\boldsymbol{n}_k = -\boldsymbol{b}\left(P^{k-1} - \xi_k\boldsymbol{n}_k - \eta_k\boldsymbol{t}_k\right)\cdot\boldsymbol{n}_k, \quad (1.63)$$

$$B_2^k(\xi_k, \eta_k) = -\boldsymbol{b}\left(\Psi_k^{-1}(\xi_k, \eta_k)\right)\cdot\boldsymbol{t}_k = -\boldsymbol{b}\left(P^{k-1} - \xi_k\boldsymbol{n}_k - \eta_k\boldsymbol{t}_k\right)\cdot\boldsymbol{t}_k. \quad (1.64)$$

Now we assume that both functions B_1^k and B_2^k possess a Taylor expansion in the variable ξ_k and write

$$B_1^k(\xi_k, \eta_k) = \sum_{j=0}^\infty \frac{\xi_k^j}{j!}\frac{\partial^j B_1^k(0, \eta_k)}{\partial\xi_k^j} = \sum_{j=0}^\infty \frac{\xi_k^j}{j!}B_{1,j}^k(0, \eta_k) \quad \text{and} \quad (1.65)$$

$$B_2^k(\xi_k, \eta_k) = \sum_{j=0}^\infty \frac{\xi_k^j}{j!}\frac{\partial^j B_2^k(0, \eta_k)}{\partial\xi_k^j} = \sum_{j=0}^\infty \frac{\xi_k^j}{j!}B_{2,j}^k(0, \eta_k), \quad (1.66)$$

where naturally $B^k_{1,0}(0,\eta_k) = B^k_1(0,\eta_k)$ and $B^k_{2,0}(0,\eta_k) = B^k_2(0,\eta_k)$ are negative functions defined on the edge e_k.

Using this expansion, we can express the differential operator L^Ψ_k in the local coordinates as

$$L^\Psi_k \;=\; \frac{1}{\varepsilon}\sum_{j=0}^{\infty}\varepsilon^j L^{(j)}_k, \tag{1.67}$$

where

$$\begin{aligned}
L^{(0)}_k &= -\tfrac{\partial^2}{\partial\xi_k^2} + B^k_1(0,\eta_k)\tfrac{\partial}{\partial\xi_k},\\
L^{(1)}_k &= \xi_k B^k_{1,1}(0,\eta_k)\tfrac{\partial}{\partial\xi_k} + B^k_2(0,\eta_k)\tfrac{\partial}{\partial\eta_k},\\
L^{(2)}_k &= -\tfrac{\partial^2}{\partial\eta_k^2} + \tfrac{1}{2}\xi_k^2 B^k_{1,2}(0,\eta_k)\tfrac{\partial}{\partial\xi_k} + \xi_k B^k_{2,1}(0,\eta_k)\tfrac{\partial}{\partial\eta_k},\\
L^{(j)}_k &= \tfrac{1}{j!}\xi_k^j B^k_{1,j}(0,\eta_k)\tfrac{\partial}{\partial\xi_k} + \tfrac{1}{(j-1)!}\xi_k^{j-1} B^k_{2,j-1}(0,\eta_k)\tfrac{\partial}{\partial\eta_k}, \quad \text{for } j\geq 3.
\end{aligned} \tag{1.68}$$

Expressing the differential operator L^Ψ_k in the local coordinates allows us to introduce the *local expansion* of the m-th order

$$V^{k,m}_{loc}(\xi_k,\eta_k) \;=\; \sum_{j=0}^{m}\varepsilon^j V^k_j(\xi_k,\eta_k). \tag{1.69}$$

Analogously as in the one-dimensional case the local corrections V^k_j have to satisfy the *boundary layer equations* in $\mathbb{R}^+\times(0,d_k)$

$$L^{(0)}_k V^k_0 \;=\; 0, \tag{1.70}$$

$$L^{(0)}_k V^k_j \;=\; -\sum_{i=1}^{j} L^{(i)}_k V^k_{j-i}, \qquad \text{for } j=1,2,\ldots,m, \tag{1.71}$$

equipped for all $j=0,1,\ldots,m$ with the boundary conditions

$$V^k_j(0,\eta_k) \;=\; -u_j\left(\Psi_k^{-1}(0,\eta_k)\right), \quad \text{and} \tag{1.72}$$

$$\lim_{\xi_k\to+\infty} V^k_j(\xi_k,\eta_k) \;=\; 0, \quad \forall\eta_k\in(0,d_k). \tag{1.73}$$

While the first condition ensures the fulfilment of the boundary condition on the edge e_k, the latter condition provides the local character of the local correction.

The ordinary differential equations (1.70)–(1.71) are then uniquely solvable, for instance the zeroth-order local correction has the form

$$V^k_0(\xi_k,\eta_k) \;=\; -u_0\left(\Psi_k^{-1}(0,\eta_k)\right)\exp\left(B^k_1(0,\eta_k)\xi_k\right). \tag{1.74}$$

We use the following lemma and corollary for the estimate of the difference of the values of the function V^k_0.

Lemma 1.3.1. *Let $s_{max}\in\mathbb{R}^+$ and let $\rho\in\mathcal{C}^1[0,s_{max}]$ and $g\in\mathcal{C}^2[0,s_{max}]$ be arbitrary functions. If $g(s)\leq -\underline{g}<0$ for all $s\in[0,s_{max}]$, then there exists a constant $C>0$ independent of ε such that*

$$\rho(s)\exp\left(g(s)\frac{s}{\varepsilon}\right) - \rho(0)\exp\left(g(0)\frac{s}{\varepsilon}\right) \;\leq\; C\varepsilon \qquad \textit{for all } s\in[0,s_{max}]. \tag{1.75}$$

Proof. The proof is analogous to the proof in Remark 1.2.2. Firstly, we show that there exists a constant C_1 independent of ε such that $|(\rho(s)-\rho(0))\exp\left(g(0)\frac{s}{\varepsilon}\right)| \leq C_1\varepsilon$ for all $s \in [0, s_{max}]$. Let us therefore denote

$$e_1(s) \;=\; (\rho(s)-\rho(0))\exp\left(g(0)\frac{s}{\varepsilon}\right). \tag{1.76}$$

Since $e_1(0) = 0$, then either $|e_1(s)| \leq |e_1(s_{max})| \leq \frac{2|\rho|_{0,\infty,(0,s_{max})}}{\underline{g}s_{max}}\varepsilon$ for all $s \in [0, s_{max}]$ or there exists $s_1 \in (0, s_{max})$ such that $s_1 = \arg\max\limits_{s\in(0,s_{max})} |e_1(s)|$. Consequently, there holds

$$e_1'(s_1) \;=\; \left(\rho'(s_1) + \frac{g(0)}{\varepsilon}\left(\rho(s_1)-\rho(0)\right)\right)\exp\left(g(0)\frac{s_1}{\varepsilon}\right) \;=\; 0, \tag{1.77}$$

which results into the inequality

$$|e_1(s)| \;\leq\; |e_1(s_1)| \;=\; \left|\frac{\rho'(s_1)}{g(0)}\right|\varepsilon\exp\left(g(0)\frac{s_1}{\varepsilon}\right) \;\leq\; \frac{|\rho|_{1,\infty,(0,s_{max})}}{\underline{g}}\,\varepsilon. \tag{1.78}$$

Thus, we take $C_1 = \max\left\{\frac{2|\rho|_{0,\infty,(0,s_{max})}}{\underline{g}s_{max}}, \frac{|\rho|_{1,\infty,(0,s_{max})}}{\underline{g}}\right\}$.

It remains to estimate the expression $\left|\rho(s)\left(\exp\left(g(s)\frac{s}{\varepsilon}\right) - \exp\left(g(0)\frac{s}{\varepsilon}\right)\right)\right|$. Let us therefore denote

$$e_2(s) \;=\; \exp\left(g(s)\frac{s}{\varepsilon}\right) - \exp\left(g(0)\frac{s}{\varepsilon}\right). \tag{1.79}$$

Since $e_2(0) = 0$, then either $|e_2(s)| \leq |e_2(s_{max})| \leq \frac{2}{\overline{g}s_{max}}\varepsilon$ for all $s \in [0, s_{max}]$ or there exists $s_2 \in (0, s_{max})$ such that $s_2 = \arg\max\limits_{s\in(0,s_{max})} |e_2(s)|$. Consequently, there holds

$$e_2'(s_2) \;=\; \frac{g'(s_2)s_2 + g(s_2)}{\varepsilon}\exp\left(g(s_2)\frac{s_2}{\varepsilon}\right) - \frac{g(0)}{\varepsilon}\exp\left(g(0)\frac{s_2}{\varepsilon}\right) \;=\; 0, \tag{1.80}$$

which results into the inequality

$$|e_2(s)| \;\leq\; |e_2(s_2)| \;=\; \left|\frac{g(0)}{g'(s_2)s_2 + g(s_2)} - 1\right|\exp\left(g(0)\frac{s_2}{\varepsilon}\right). \tag{1.81}$$

Denoting $r(s) = \frac{g(0)}{g'(s)s+g(s)}$ we find out that $r(0) = 1$, hence from the first part of this prove it follows that there exists a constant C_2 independent of ε such that $|(r(s_2) - r(0))\exp\left(g(0)\frac{s_2}{\varepsilon}\right)| \leq C_2\varepsilon$. Combining all estimates we get

$$\begin{aligned}
&\left|\rho(s)\exp\left(g(s)\frac{s}{\varepsilon}\right) - \rho(0)\exp\left(g(0)\frac{s}{\varepsilon}\right)\right| \;= \\
&\quad= \;\left|\rho(s)\left(\exp\left(g(s)\frac{s}{\varepsilon}\right) - \exp\left(g(0)\frac{s}{\varepsilon}\right)\right) + (\rho(s)-\rho(0))\exp\left(g(0)\frac{s}{\varepsilon}\right)\right| \;\leq \\
&\quad\leq \;\|\rho\|_{0,\infty,(0,s_{max})}\,C_2\varepsilon + C_1\varepsilon \;\leq\; C\varepsilon.
\end{aligned} \tag{1.82}$$

□

As we already mentioned, the following corollary of Lemma 1.3.1 is applied in the proof of Theorem 1.3.1 (page 22) where it enables us to estimate the difference $u - u_{as,0}$ on Γ_+.

Corollary 1.3.1. Let the functions V_0^{k-1} and V_0^{k+1} be defined using the expression (1.74). Then there exist constants C_A and C_B independent of ε such that for each $\eta_k \in [0, d_k]$ there holds

$$\begin{aligned}\left|V_0^{k-1}\left(\tfrac{1}{\varepsilon}\eta_k \sin\gamma_{k-1}, d_{k-1} - \eta_k\cos\gamma_{k-1}\right) - V_0^{k-1}\left(\tfrac{1}{\varepsilon}\eta_k\sin\gamma_{k-1}, d_{k-1}\right)\right| &\leq C_A\varepsilon,\\ \left|V_0^{k+1}\left(\tfrac{1}{\varepsilon}(d_k-\eta_k)\sin\gamma_k, (d_k-\eta_k)\cos\gamma_k\right) - V_0^{k+1}\left(\tfrac{1}{\varepsilon}(d_k-\eta_k)\sin\gamma_k, 0\right)\right| &\leq C_B\varepsilon.\end{aligned}$$

Proof. In the first case we use a substitution $s = \eta_k \sin\gamma_{k-1} \in [0, d_k \sin\gamma_{k-1}]$. Then choosing

$$\rho_1(s) = -u_0\left(\Psi_{k-1}^{-1}(0, d_{k-1} - s\cot\gamma_{k-1})\right) \quad \text{and} \quad g_1(s) = B_1^k(0, d_{k-1} - s\cot\gamma_{k-1}) \tag{1.83}$$

leads (together with an application of the previous lemma) to the desired estimate.

In the second case one uses a substitution $s = (d_k - \eta_k)\sin\gamma_k \in [0, d_k\sin\gamma_k]$. Consequently, the functions

$$\rho_2(s) = -u_0\left(\Psi_{k+1}^{-1}(0, s\cot\gamma_k)\right) \quad \text{and} \quad g_2(s) = B_1^k(0, s\cot\gamma_k) \tag{1.84}$$

and the previous lemma provide the desired estimate. □

Remark 1.3.1. The previous estimates are valid if the functions u_0 and $\boldsymbol{b}$ are sufficiently smooth and if there holds $d_{k-1} - \eta_k\cos\gamma_{k-1} \in [0, d_{k-1}]$ and $(d_k - \eta_k)\cos\gamma_k \in [0, d_{k+1}]$. It means that we should consider only $\cos\gamma_k \geq 0$ for all $k = 0, 1, \ldots, H$. If we want to use these estimates for obtuse angles γ_k, $k = 0, 1, \ldots, H$, the problem data (and consequently the function u_0) have to be defined also in some neighborhood of Ω.

1.3.3 Corner correction

Unlike the one-dimensional case the considered two-dimensional domain contains corners and if for any order $j \in \{0, 1, \ldots, m\}$ and any $k \in \{1, 2, \ldots, H-1\}$ we sum

$$\begin{aligned}u_j(P^k) + V_j^k(\Psi_k^\varepsilon(P^k)) + V_j^{k+1}(\Psi_{k+1}^\varepsilon(P^k)) &= u_j(P^k) + V_j^k(0, d_k) + V_j^{k+1}(0,0) =\\ &= u_j(P^k) - u_j\left(P^k\right) - u_j\left(P^k\right) = -u_j(P^k),\end{aligned} \tag{1.85}$$

we find out, that the boundary condition at the corner corresponding to the vertex P^k is not satisfied. Thus, we have to add some *corner correction terms*.

Firstly, for each $k = 1, 2, \ldots, H-1$ we define the transformation of the coordinates $\Phi_k : (x, y) \to (\xi_k, \xi_{k+1})$ corresponding to the vertex P^k as

$$\xi_k(x,y) = (P_y^{k-1} - y)\cos\alpha_k - (P_x^{k-1} - x)\sin\alpha_k, \tag{1.86}$$

$$= (P_y^k - y)\cos\alpha_k - (P_x^k - x)\sin\alpha_k, \tag{1.87}$$

$$\xi_{k+1}(x,y) = (P_y^k - y)\cos\alpha_{k+1} - (P_x^k - x)\sin\alpha_{k+1}, \tag{1.88}$$

where we used the fact, that the vector $P^k - P^{k-1}$ is perpendicular to the normal vector $\boldsymbol{n}_k = (-\sin\alpha_k, \cos\alpha_k)$.

In order to define the corner correction terms we stretch the scale in both ξ_k and ξ_{k+1} direction using the transformation $\Phi_k^{\varepsilon,\varepsilon} : (x, y) \to \left(\frac{\xi_k}{\varepsilon}, \frac{\xi_{k+1}}{\varepsilon}\right)$ with ξ_k and ξ_{k+1} defined in (1.86)–(1.88).

Under the transformation $\Phi_k^{\varepsilon,\varepsilon}$ the differential operator L (cf. (1.49)) changes into

$$\mathcal{L}_k^{\varepsilon,\varepsilon} = \frac{1}{\varepsilon}\left(-\frac{\partial^2}{\partial\xi_k^2} + 2\cos\gamma_k\frac{\partial^2}{\partial\xi_k\partial\xi_{k+1}} - \frac{\partial^2}{\partial\xi_{k+1}^2} + \right.$$
$$\left. + \mathcal{B}_1^k(\varepsilon\xi_k, \varepsilon\xi_{k+1})\frac{\partial}{\partial\xi_k} + \mathcal{B}_2^k(\varepsilon\xi_k, \varepsilon\xi_{k+1})\frac{\partial}{\partial\xi_{k+1}}\right) \quad (1.89)$$

with

$$\begin{aligned}
\mathcal{B}_1^k(\xi_k, \xi_{k+1}) &= -\boldsymbol{b}\left(\Phi_k^{-1}(\xi_k, \xi_{k+1})\right)\cdot\boldsymbol{n}_k = \\
&= -\boldsymbol{b}\left(P^k + \frac{1}{\sin\gamma_k}(\xi_{k+1}\boldsymbol{t}_k - \xi_k\boldsymbol{t}_{k+1})\right)\cdot\boldsymbol{n}_k, \quad (1.90)\\
\mathcal{B}_2^k(\xi_k, \xi_{k+1}) &= -\boldsymbol{b}\left(\Phi_k^{-1}(\xi_k, \xi_{k+1})\right)\cdot\boldsymbol{n}_{k+1} = \\
&= -\boldsymbol{b}\left(P^k + \frac{1}{\sin\gamma_k}(\xi_{k+1}\boldsymbol{t}_k - \xi_k\boldsymbol{t}_{k+1})\right)\cdot\boldsymbol{n}_{k+1}. \quad (1.91)
\end{aligned}$$

We again assume that both functions $\mathcal{B}_1^k$ and $\mathcal{B}_2^k$ possess Taylor's expansion in the form

$$\mathcal{B}_1^k(\xi_k, \xi_{k+1}) = \sum_{i,j=0}^{\infty}\frac{\xi_k^i\xi_{k+1}^j}{i!j!}\frac{\partial^{i+j}\mathcal{B}_1^k(0,0)}{\partial^i\xi_k\partial^j\xi_{k+1}} = \sum_{i,j=0}^{\infty}\frac{\xi_k^i\xi_{k+1}^j}{i!j!}\mathcal{B}_{1,ij}^k \quad \text{and} \quad (1.92)$$

$$\mathcal{B}_2^k(\xi_k, \xi_{k+1}) = \sum_{i,j=0}^{\infty}\frac{\xi_k^i\xi_{k+1}^j}{i!j!}\frac{\partial^{i+j}\mathcal{B}_2^k(0,0)}{\partial^i\xi_k\partial^j\xi_{k+1}} = \sum_{i,j=0}^{\infty}\frac{\xi_k^i\xi_{k+1}^j}{i!j!}\mathcal{B}_{2,ij}^k, \quad (1.93)$$

where it obviously holds $\mathcal{B}_{1,00}^k = \mathcal{B}_1^k(0,0) = -\boldsymbol{b}\left(P^k\right)\cdot\boldsymbol{n}_k = B_1^k(0, d_k)$ and $\mathcal{B}_{2,00}^k = \mathcal{B}_2^k(0,0) = -\boldsymbol{b}\left(P^k\right)\cdot\boldsymbol{n}_{k+1} = B_1^{k+1}(0,0)$. Since we are going to use the (negative) values $\mathcal{B}_{1,00}^k$ and $\mathcal{B}_{2,00}^k$ frequently, we also denote $\beta_1^k = \mathcal{B}_{1,00}^k$ and $\beta_2^k = \mathcal{B}_{2,00}^k$.

Using Taylor's expansions (1.92) and (1.93) of the functions $\mathcal{B}_1^k$ and $\mathcal{B}_2^k$ we can express the differential operator $\mathcal{L}_k^{\varepsilon,\varepsilon}$ in the local coordinates as

$$\mathcal{L}_k^{\varepsilon,\varepsilon} = \frac{1}{\varepsilon}\sum_{j=0}^{\infty}\varepsilon^j\mathcal{L}_k^{(j)}, \quad (1.94)$$

where

$$\mathcal{L}_k^{(0)} = -\frac{\partial^2}{\partial\xi_k^2} + 2\cos\gamma_k\frac{\partial^2}{\partial\xi_k\partial\xi_{k+1}} - \frac{\partial^2}{\partial\xi_{k+1}^2} + \beta_1^k\frac{\partial}{\partial\xi_k} + \beta_2^k\frac{\partial}{\partial\xi_{k+1}}, \quad (1.95)$$

$$\mathcal{L}_k^{(r)} = \left(\sum_{i+j=r}\frac{\xi_k^i\xi_{k+1}^j}{i!j!}\mathcal{B}_{1,ij}^k\right)\frac{\partial}{\partial\xi_k} + \left(\sum_{i+j=r}\frac{\xi_k^i\xi_{k+1}^j}{i!j!}\mathcal{B}_{2,ij}^k\right)\frac{\partial}{\partial\xi_{k+1}}, \text{ for } r \geq 1. \quad (1.96)$$

Expressing the operator L in the local coordinates (equality (1.94)) allows us to introduce the two-dimensional *corner expansion*

$$Z_{cor}^{k,m}(\xi_k,\xi_{k+1}) \;=\; \sum_{i=0}^{m}\varepsilon^i Z_i^k(\xi_k,\xi_{k+1}). \tag{1.97}$$

Therefore, we can evaluate (cf. (1.17))

$$\mathcal{L}_k^{\varepsilon,\varepsilon} Z_{cor}^{k,m+1}(\xi_k,\xi_{k+1}) \;=\; \frac{1}{\varepsilon}\sum_{j=0}^{\infty}\sum_{i=0}^{m+1}\varepsilon^{j+i}\mathcal{L}_k^{(j)} Z_i^k(\xi_k,\xi_{k+1}) \;= \tag{1.98}$$

$$= \;\frac{1}{\varepsilon}\sum_{s=0}^{m+1}\varepsilon^s\sum_{r=0}^{s}\mathcal{L}_k^{(r)} Z_{s-r}^k(\xi_k,\xi_{k+1}) + \varepsilon^{m+1}\sum_{q=0}^{\infty}\varepsilon^q\sum_{t=0}^{m+1}\mathcal{L}_k^{(t+q+1)} Z_{m+1-t}^k(\xi_k,\xi_{k+1})$$

and thus, the corner corrections Z_j^k are the solutions of the system of partial differential equations in $\mathbb{R}^+\times\mathbb{R}^+$

$$\mathcal{L}_k^{(0)} Z_0^k \;=\; 0, \tag{1.99}$$

$$\mathcal{L}_k^{(0)} Z_j^k \;=\; -\sum_{i=1}^{j}\mathcal{L}_k^{(i)} Z_{j-i}^k, \qquad \text{for } j=1,2,\ldots,m, \tag{1.100}$$

equipped for all $j=0,1,\ldots,m$ with the boundary conditions

$$Z_j^k(\xi_k,0) \;=\; -V_j^k\left(\xi_k,d_k\right), \tag{1.101}$$

$$Z_j^k(0,\xi_{k+1}) \;=\; -V_j^{k+1}\left(\xi_{k+1},0\right) \quad \text{and} \tag{1.102}$$

$$\lim_{\xi_k,\xi_{k+1}\to+\infty} Z_j^k(\xi_k,\xi_{k+1}) \;=\; 0. \tag{1.103}$$

The boundary conditions (1.101)–(1.102) are formulated in such a way that we obtain a simple form of the zeroth-order matched asymptotic expansion (from Remark 1.2.2 we know that asymptotic expansion is not uniquely defined). Therefore, we consider only $m=0$ in Theorem 1.3.1.

In the derivation of the zeroth-order matched asymptotic expansion we shall use the mapping compositions which can be rewritten using the following lemma.

Lemma 1.3.2. *For the mappings $\Psi_{k-1},\Psi_k,\Psi_{k+1}$ and Φ_k,Φ_{k-1} there holds*

$$\begin{aligned}
\Psi_{k-1}\Psi_k^{-1}(\xi_k,\eta_k) &= [\eta_k\sin\gamma_{k-1}-\xi_k\cos\gamma_{k-1}, d_{k-1}-\eta_k\cos\gamma_{k-1}-\xi_k\sin\gamma_{k-1}],\\
\Psi_{k+1}\Psi_k^{-1}(\xi_k,\eta_k) &= [(d_k-\eta_k)\sin\gamma_k-\xi_k\cos\gamma_k, (d_k-\eta_k)\cos\gamma_k+\xi_k\sin\gamma_k],\\
\Phi_{k-1}\Psi_k^{-1}(\xi_k,\eta_k) &= [\eta_k\sin\gamma_{k-1}-\xi_k\cos\gamma_{k-1}, \xi_k],\\
\Phi_k\Psi_k^{-1}(\xi_k,\eta_k) &= [\xi_k, (d_k-\eta_k)\sin\gamma_k-\xi_k\cos\gamma_k],
\end{aligned} \tag{1.104}$$

where ξ_k,ξ_{k+1} and η_k all belong to the domains of respective mappings.

Proof. Since the proof is in all cases analogous, we prove only the first equality. From (1.60)–(1.61) it follows that

$$P^{k-2}-\Psi_k^{-1}(\xi_k,\eta_k) \;=\; P^{k-2}-X \;=\; P^{k-2}-P^{k-1}+\xi_k(X)\boldsymbol{n}_k+\eta_k(X)\boldsymbol{t}_k. \tag{1.105}$$

Hence, using the equalities $\boldsymbol{n}_k \cdot \boldsymbol{n}_{k-1} = \boldsymbol{t}_k \cdot \boldsymbol{t}_{k-1} = -\cos\gamma_{k-1}$, $\boldsymbol{n}_k \cdot \boldsymbol{t}_{k-1} = -\sin\gamma_{k-1}$, $\boldsymbol{t}_k \cdot \boldsymbol{n}_{k-1} = \sin\gamma_{k-1}$ and $P^{k-2} - P^{k-1} = d_{k-1}\boldsymbol{t}_{k-1}$ together with the definition of Ψ_{k-1} we obtain

$$\begin{aligned} (P^{k-2} - \Psi_k^{-1}(\xi_k, \eta_k)) \cdot \boldsymbol{n}_{k-1} &= \eta_k \sin\gamma_{k-1} - \xi_k \cos\gamma_{k-1} \qquad \text{and} \quad (1.106) \\ (P^{k-2} - \Psi_k^{-1}(\xi_k, \eta_k)) \cdot \boldsymbol{t}_{k-1} &= d_{k-1} - \eta_k \cos\gamma_{k-1} - \xi_k \sin\gamma_{k-1}. \quad (1.107) \end{aligned}$$

□

The behavior of the mapping compositions on an edge $e_k \subset \Gamma_+$ immediately results from the previous lemma.

Corollary 1.3.2. For the mappings $\Psi_{k-1}, \Psi_k, \Psi_{k+1}$ and Φ_k, Φ_{k-1} there holds

$$\begin{aligned} \Psi_{k-1}\Psi_k^{-1}(0, \eta_k) &= [\eta_k \sin\gamma_{k-1}, d_{k-1} - \eta_k \cos\gamma_{k-1}], \quad (1.108) \\ \Psi_{k+1}\Psi_k^{-1}(0, \eta_k) &= [(d_k - \eta_k)\sin\gamma_k, (d_k - \eta_k)\cos\gamma_k], \quad (1.109) \\ \Phi_{k-1}\Psi_k^{-1}(0, \eta_k) &= [\eta_k \sin\gamma_{k-1}, 0], \quad (1.110) \\ \Phi_k\Psi_k^{-1}(0, \eta_k) &= [0, (d_k - \eta_k)\sin\gamma_k], \quad (1.111) \end{aligned}$$

where $\eta_k \in [0, d_k]$.

Now we have all necessary ingredients for a construction of a zeroth-order matched asymptotic expansion.

Theorem 1.3.1. *Let $\Gamma_0 = \emptyset$, let f and $\boldsymbol{b}$ are sufficiently smooth functions and let all the characteristics through points of $\overline{\Omega}$ leave $\overline{\Omega}$ at points of Γ_+ in finite time. Then the solution of the problem (1.49)–(1.50) has a zeroth-order matched asymptotic expansion of the form*

$$\begin{aligned} u_{as,0}(x,y) &= \quad (1.112) \\ &= u_0(x,y) + \sum_{k=1}^{H} V_{loc}^{k,0}\left(\frac{\xi_k(x,y)}{\varepsilon}, \eta_k(x,y)\right) + \sum_{k=1}^{H-1} Z_{cor}^{k,0}\left(\frac{\xi_k(x,y)}{\varepsilon}, \frac{\xi_{k+1}(x,y)}{\varepsilon}\right), \end{aligned}$$

where $V_{loc}^{k,0}$ and $Z_{cor}^{k,0}$ are defined in (1.69) and (1.97), respectively. Moreover, there exists a constant C independent of x, y and ε such that

$$|u(x,y) - u_{as,0}(x,y)| \le C\varepsilon \qquad \text{for } [x,y] \in \overline{\Omega} \text{ and } \varepsilon \le \varepsilon_0. \quad (1.113)$$

Here ε_0 is any sufficiently small fixed positive constant.

Proof. Instead of $u_{as,0}$, we firstly prove the theorem considering the function

$$\begin{aligned} u_{as,0}^*(x,y) &= u_0(x,y) + \quad (1.114) \\ &+ \sum_{k=1}^{H} V_{loc}^{k,1}\left(\frac{\xi_k(x,y)}{\varepsilon}, \eta_k(x,y)\right) + \sum_{k=1}^{H-1} Z_{cor}^{k,1}\left(\frac{\xi_k(x,y)}{\varepsilon}, \frac{\xi_{k+1}(x,y)}{\varepsilon}\right), \end{aligned}$$

which has additional local correction and corner correction terms. Using the definitions and equalities (1.67)–(1.71) together with the interchange of sums analogous to the one-dimensional case (cf. (1.17)) we obtain

$$\begin{aligned} L_k^{\Psi} V_{loc}^{k,1} &= \frac{1}{\varepsilon}\sum_{j=0}^{\infty}\sum_{r=0}^{1} \varepsilon^{j+r} L_k^{(j)} V_r^k = \quad (1.115) \\ &= \frac{1}{\varepsilon}\sum_{s=0}^{1}\varepsilon^s \sum_{j=0}^{s} L_k^{(j)} V_{s-j}^k + \frac{1}{\varepsilon}\sum_{s=2}^{\infty}\varepsilon^s \sum_{j=s-1}^{s} L_k^{(j)} V_{s-j}^k = \frac{1}{\varepsilon}\sum_{s=2}^{\infty}\varepsilon^s \sum_{j=s-1}^{s} L_k^{(j)} V_{s-j}^k. \end{aligned}$$

Consequently, there exists a positive constant C_V (independent of ε) such that for all $(\xi_k, \eta_k) \in \mathbb{R}^+ \times (0, d_k)$ there holds

$$\left|L_k^{\Psi} V_{loc}^{k,1}\right| \;=\; \left|\frac{1}{\varepsilon}\sum_{s=2}^{\infty}\varepsilon^s \sum_{j=s-1}^{s} L_k^{(j)} V_{s-j}^{k}\right| \;\le\; C_V\varepsilon. \tag{1.116}$$

Similarly, there exists a positive constant C_Z (independent of ε) such that for all $(\xi_k, \xi_{k+1}) \in \mathbb{R}^+ \times \mathbb{R}^+$ it holds (cf. (1.94)–(1.100))

$$\left|\mathcal{L}_k^{\varepsilon,\varepsilon} Z_{cor}^{k,1}\right| \;=\; \left|\frac{1}{\varepsilon}\sum_{s=2}^{\infty}\varepsilon^s \sum_{j=s-1}^{s} \mathcal{L}_k^{(j)} Z_{s-j}^{k}\right| \;\le\; C_Z\varepsilon. \tag{1.117}$$

Consequently, the function $u_{as,0}^*$ in Ω satisfies

$$\left|L\left(u - u_{as,0}^*\right)\right| \;\le\; \left(2|u_0|_{2,\infty,\Omega} + C_V + C_Z\right)\,\varepsilon \;=\; C_0^*\,\varepsilon, \tag{1.118}$$

where we employed the equality (1.57).

Further, since the functions $V_{loc}^{k,1}$ and $Z_{cor}^{k,1}$ have an exponential decay away from the boundary Γ_+, then for arbitrary $\kappa > 0$ there exists a constant $C_\kappa^- > 0$ such that for every $\varepsilon \le \varepsilon_0$ there holds

$$\left|(u - u_{as,0}^*)|_{\Gamma_-}\right| \;=\; \left|-\sum_{k=1}^{H} V_{loc}^{k,1}|_{\Gamma_-} - \sum_{k=1}^{H-1} Z_{cor}^{k,1}|_{\Gamma_-}\right| \;\le\; C_\kappa^-\,\varepsilon^\kappa. \tag{1.119}$$

Finally, let $e_k \subset \Gamma_+$ be an arbitrary edge and let $X = \Psi_k^{-1}(0, \eta_k) \in e_k$ be any point laying on this edge. Then $u(X) = 0$ and the value $u_{as,0}^*(X)$ is given by the global expansion, the local corrections corresponding to the edges e_{k-1}, e_k, e_{k+1} and the corner corrections in the corners P^{k-1} and P^k. All the remaining correction terms are $\mathcal{O}(\varepsilon)$ due to the presence of exponential functions (exponential decay). Thus, using the boundary conditions (1.101)–(1.102) results in the estimate

$$\begin{aligned}
&(u - u_{as,0}^*)(\Psi_k^{-1}(0,\eta_k)) \;= \\
&=\; \mathcal{O}(\varepsilon) - u_0(\Psi_k^{-1}(0,\eta_k)) - V_{loc}^{k,1}(0,\eta_k) - V_{loc}^{k-1,1}(\Psi_{k-1}^{\varepsilon}\Psi_k^{-1}(0,\eta_k)) - \\
&\quad - Z_{cor}^{k-1,1}(\Phi_{k-1}^{\varepsilon,\varepsilon}\Psi_k^{-1}(0,\eta_k)) - V_{loc}^{k+1,1}(\Psi_{k+1}^{\varepsilon}\Psi_k^{-1}(0,\eta_k)) - Z_{cor}^{k,1}(\Phi_k^{\varepsilon,\varepsilon}\Psi_k^{-1}(0,\eta_k)) = \\
&=\; \mathcal{O}(\varepsilon) - \Big\{u_0(\Psi_k^{-1}(0,\eta_k)) + V_0^k(0,\eta_k)\Big\} - \varepsilon V_1^k(\Psi_k^{\varepsilon}\Psi_k^{-1}(0,\eta_k)) - \\
&\quad - \sum_{j=0}^{1}\varepsilon^j\Big\{V_j^{k-1}(\Psi_{k-1}^{\varepsilon}\Psi_k^{-1}(0,\eta_k)) + Z_j^{k-1}(\Phi_{k-1}^{\varepsilon,\varepsilon}\Psi_k^{-1}(0,\eta_k))\Big\} - \\
&\quad - \sum_{j=0}^{1}\varepsilon^j\Big\{V_j^{k+1}(\Psi_{k+1}^{\varepsilon}\Psi_k^{-1}(0,\eta_k)) + Z_j^{k}(\Phi_k^{\varepsilon,\varepsilon}\Psi_k^{-1}(0,\eta_k))\Big\} \;= \\
&=\; \mathcal{O}(\varepsilon) - \Big\{V_0^{k-1}\left(\tfrac{1}{\varepsilon}\eta_k\sin\gamma_{k-1}, d_{k-1} - \eta_k\cos\gamma_{k-1}\right) + Z_0^{k-1}\left(\tfrac{1}{\varepsilon}\eta_k\sin\gamma_{k-1}, 0\right)\Big\} - \\
&\quad - \Big\{V_0^{k+1}\Big(\tfrac{1}{\varepsilon}(d_k - \eta_k)\sin\gamma_k, (d_k - \eta_k)\cos\gamma_k\Big) + Z_0^{k}\Big(0, \tfrac{1}{\varepsilon}(d_k - \eta_k)\sin\gamma_k\Big)\Big\} \;= \\
&=\; \mathcal{O}(\varepsilon) - \Big\{V_0^{k-1}\Big(\tfrac{1}{\varepsilon}\eta_k\sin\gamma_{k-1}, d_{k-1} - \eta_k\cos\gamma_{k-1}\Big) - V_0^{k-1}\Big(\tfrac{1}{\varepsilon}\eta_k\sin\gamma_{k-1}, d_{k-1}\Big)\Big\} - \\
&\quad - \Big\{V_0^{k+1}\Big(\tfrac{1}{\varepsilon}(d_k - \eta_k)\sin\gamma_k, (d_k - \eta_k)\cos\gamma_k\Big) - V_0^{k+1}\Big(\tfrac{1}{\varepsilon}(d_k - \eta_k)\sin\gamma_k, 0\Big)\Big\}.
\end{aligned}$$

Hence, applying the estimate of Corollary 1.3.1 one can find a constant $C_0^+ > 0$ such that $\Big|(u - u^*_{as,0})|_{\Gamma_+}\Big| \le C_0^+ \varepsilon$.

One may be interested in the situation in some neighborhood $\mathcal{U}_\varepsilon(P)$ of a node $P \in \overline{\Gamma}_+ \cap \overline{\Gamma}_-$. Since there holds $u_0 = 0$ on Γ_- and we consider u_0 being sufficiently smooth, then it follows that $|u_0| = \mathcal{O}(\varepsilon)$ in $\mathcal{U}_\varepsilon(P)$ and consequently, we indeed obtain $|u - u^*_{as,0}| = \mathcal{O}(\varepsilon)$ in $\mathcal{U}_\varepsilon(P)$.

From the assumptions of the theorem and from Lemma 4.1.1 (page 124) it follows that there exists a function ϕ such that $L_0\phi = \boldsymbol{b} \cdot \nabla\phi \ge \phi_0 > 0$ in $\overline{\Omega}$. Since any function $\phi_c = \phi + c$, $c \in \mathbb{R}$, satisfies $L_0\phi_c = L_0\phi$, we can choose ϕ in such a way that $\phi > 0$ in Ω and $\|\phi\|_{0,\infty,\Omega}$ is the smallest possible. Consequently, for $\varepsilon \le \frac{1}{4}\,\phi_0/\|\phi\|_{2,\infty,\Omega}$ there holds

$$L\phi \;\ge\; \phi_0 - 2\varepsilon\|\phi\|_{2,\infty,\Omega} \;\ge\; \frac{1}{2}\,\phi_0. \tag{1.120}$$

Therefore, if we define the function W by the relation

$$W = \max\left\{\frac{2}{\phi_0}\,C_0^*, C_0^-, C_0^+\right\}\Big(\phi + 1\Big)\,\varepsilon, \tag{1.121}$$

then employing the inequality (1.118) we may estimate

$$LW \;\ge\; C_0^*\,\varepsilon \;\ge\; L\left(u - u^*_{as,0}\right) \;\ge\; -C_0^*\,\varepsilon \;\ge\; -LW. \tag{1.122}$$

Moreover, from the inequalities (1.119) and (1.120) it follows that

$$\begin{aligned} W \;\ge\; \max\Big\{C_0^-, C_0^+\Big\}\,\varepsilon \;\ge\; \left(u - u^*_{as,0}\right)|_{\partial\Omega} \;\ge\; \\ \ge\; -\max\Big\{C_0^-, C_0^+\Big\}\,\varepsilon \;\ge\; -W. \end{aligned} \tag{1.123}$$

Applying the comparison principle (Theorem 4.1.6, page 126) then gives

$$W \;\ge\; \left(u - u^*_{as,0}\right) \;\ge\; -W \qquad \text{in } \Omega, \tag{1.124}$$

which for all $[x, y] \subset \Omega$ implies

$$\begin{aligned} \Big|(u - u_{as,0})(x,y)\Big| \;&\le \Big|(u - u^*_{as,0})(x,y)\Big| + \varepsilon \sum_{k=1}^{H} \left|V_1^k\left(\frac{\xi_k(x,y)}{\varepsilon}, \eta_k(x,y)\right)\right| + \\ &\qquad + \varepsilon \sum_{k=1}^{H-1} \left|Z_1^k\left(\frac{\xi_k(x,y)}{\varepsilon}, \frac{\xi_{k+1}(x,y)}{\varepsilon}\right)\right| \le \\ \le\;& W + \varepsilon\left(\sum_{k=1}^{H} \|V_1^k\|_{0,\infty,\mathbb{R}^+\times(0,d_k)} + \sum_{k=1}^{H-1} \|Z_1^k\|_{0,\infty,\mathbb{R}^+\times\mathbb{R}^+}\right) \le \\ \le\;& \varepsilon\left(\max\left\{\frac{2}{\phi_0}\,C_0^*, C_0^-, C_0^+\right\}\Big(1 + \|\phi\|_{0,\infty,\Omega}\Big) + \right. \\ &\qquad \left. + \sum_{k=1}^{H} \|V_1^k\|_{0,\infty,\mathbb{R}^+\times(0,d_k)} + \sum_{k=1}^{H-1} \|Z_1^k\|_{0,\infty,\mathbb{R}^+\times\mathbb{R}^+}\right). \end{aligned} \tag{1.125}$$

□

Let us try to find a solution of the differential equation (1.99) equipped with the boundary conditions (1.101)–(1.103) for $j = 0$. We shall seek the solution of the equation (1.99) in the form

$$Z_0^k(\xi_k, \xi_{k+1}) \;=\; u_0\left(P^k\right)\left\{\sum_{j=0}^{\rho_k}\exp\left(p_j^k\xi_k + q_j^k\xi_{k+1}\right) - \sum_{j=0}^{\rho_k-1}\exp\left(p_{j+1}^k\xi_k + q_j^k\xi_{k+1}\right)\right\}, \tag{1.126}$$

where $\rho_k \in \mathbb{N}, p_j^k$ and q_j^k have yet to be defined.

From (1.126) it follows that there holds $Z_0^k(0, \xi_{k+1}) = u_0\left(P^k\right)\exp\left(q_{\rho_k}^k\xi_{k+1}\right)$ and $Z_0^k(\xi_k, 0) = u_0\left(P^k\right)\exp\left(p_0^k\xi_k\right)$. Thus, if we choose $q_{\rho_k}^k = \beta_2^k$ and $p_0^k = \beta_1^k$ the boundary conditions (1.101)–(1.102) are fulfilled.

On the other hand fulfilment of the boundary conditions (1.103) is guaranteed only if $p_j^k < 0$ and $q_j^k < 0$ for all $j \in \{0, 1, \ldots, \rho_k\}$. If we, for instance, choose $p_m^k > 0$, then the difference $\exp\left(p_m^k\xi_{k+1} + q_m^k\xi_k\right) - \exp\left(p_m^k\xi_{k+1} + q_{m-1}^k\xi_k\right)$ has to tend to zero as $\xi_{k+1} \to +\infty$ or $\xi_k \to +\infty$. However, this is not possible since for fixed ξ_k one has

$$\lim_{\xi_{k+1}\to+\infty}\left|\exp\left(p_m^k\xi_{k+1}\right)\left(\exp\left(q_m^k\xi_k\right) - \exp\left(q_{m-1}^k\xi_k\right)\right)\right| \;=\; +\infty. \tag{1.127}$$

In order to fulfil the equation (1.99) one also requires the fulfilment of (1.99) for each function from sums (1.126). Consequently, for any $j \in \{0, 1, \ldots, \rho_k - 1\}$ we obtain the following set of equations

$$\begin{array}{rcrcrcrcrcl}
-\left(p_j^k\right)^2 & + & 2p_j^kq_j^k\cos\gamma_k & - & \left(q_j^k\right)^2 & + & \beta_1^kp_j^k & + & \beta_2^kq_j^k & = & 0,\\
-\left(p_{j+1}^k\right)^2 & + & 2p_{j+1}^kq_j^k\cos\gamma_k & - & \left(q_j^k\right)^2 & + & \beta_1^kp_{j+1}^k & + & \beta_2^kq_j^k & = & 0,\\
-\left(p_{j+1}^k\right)^2 & + & 2p_{j+1}^kq_{j+1}^k\cos\gamma_k & - & \left(q_{j+1}^k\right)^2 & + & \beta_1^kp_{j+1}^k & + & \beta_2^kq_{j+1}^k & = & 0.
\end{array} \tag{1.128}$$

Subtracting the second equation from the first one and the third one from the second one yields

$$\left(p_{j+1}^k - p_j^k\right)\left(p_{j+1}^k + p_j^k - 2q_j^k\cos\gamma_k - \beta_1^k\right) \;=\; 0, \tag{1.129}$$

$$\left(q_{j+1}^k - q_j^k\right)\left(q_{j+1}^k + q_j^k - 2p_{j+1}^k\cos\gamma_k - \beta_2^k\right) \;=\; 0. \tag{1.130}$$

If $p_{j+1}^k = p_j^k$ for some $j \in \{0, 1, \ldots, \rho_k - 1\}$, then two exponential functions in (1.126) cancel each other, hence one can omit them and set $\rho_k := \rho_k - 1$. The same argument holds for q_j^k. Therefore we consider $p_{j+1}^k \neq p_j^k$ and $q_{j+1}^k \neq q_j^k$ for all $j \in \{0, 1, \ldots, \rho_k - 1\}$. Consequently, from (1.129) it follows that

$$p_{j+1}^k \;=\; -p_j^k + 2q_j^k\cos\gamma_k + \beta_1^k \tag{1.131}$$

and using this equality together with (1.130) we get

$$q_{j+1}^k \;=\; -q_j^k + 2p_{j+1}^k\cos\gamma_k + \beta_2^k \;= \tag{1.132}$$

$$\;=\; -2p_j^k\cos\gamma_k + \left(4\cos^2\gamma_k - 1\right)q_j^k + 2\beta_1^k\cos\gamma_k + \beta_2^k. \tag{1.133}$$

In order to simplify further calculations we denote $\boldsymbol{w}_j^k = (p_j^k, q_j^k)^T$ for all $j \in \{0, 1, \ldots, \rho_k\}$. Then for all $j \in \{1, 2, \ldots, \rho_k\}$ it holds

$$\boldsymbol{w}_j^k \;=\; \mathbb{A}\boldsymbol{w}_{j-1}^k + \boldsymbol{r}^k, \tag{1.134}$$

where

$$\mathbb{A} \;=\; \begin{pmatrix} -1 & 2\cos\gamma_k \\ -2\cos\gamma_k & 4\cos^2\gamma_k - 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{r}^k \;=\; \begin{pmatrix} \beta_1^k \\ \beta_2^k + 2\beta_1^k\cos\gamma_k \end{pmatrix}. \tag{1.135}$$

Since $p_0^k = \beta_1^k$ and $q_0^k \neq 0$ then from the first equation in (1.128) it follows that $q_0^k = \beta_2^k + 2\beta_1^k\cos\gamma_k$. Thus, there holds $\boldsymbol{w}_0^k = \boldsymbol{r}^k$. Let us further define the fixed point $\widetilde{\boldsymbol{w}}^k$ of the iterations given by (1.134)

$$\widetilde{\boldsymbol{w}}^k \;=\; \mathbb{A}\widetilde{\boldsymbol{w}}^k + \boldsymbol{r}^k. \tag{1.136}$$

Its value is $\widetilde{\boldsymbol{w}}^k = \frac{1}{2\sin^2\gamma_k}\left(\beta_1^k + \beta_2^k\cos\gamma_k, \beta_2^k + \beta_1^k\cos\gamma_k\right)^T$. Subtracting (1.136) from (1.134) then yields

$$\begin{aligned} \boldsymbol{w}_j^k - \widetilde{\boldsymbol{w}}^k \;&=\; \mathbb{A}\left(\boldsymbol{w}_{j-1}^k - \widetilde{\boldsymbol{w}}^k\right) \;=\; \mathbb{A}^j\left(\boldsymbol{w}_0^k - \widetilde{\boldsymbol{w}}^k\right) \;=\; \mathbb{A}^j\left(\boldsymbol{r}^k - \widetilde{\boldsymbol{w}}^k\right) \;= \\ &=\; \mathbb{A}^j\left(-\mathbb{A}\widetilde{\boldsymbol{w}}^k\right) \;=\; -\mathbb{A}^{j+1}\widetilde{\boldsymbol{w}}^k. \end{aligned} \tag{1.137}$$

Thus $\boldsymbol{w}_{\rho_k}^k = (\mathbb{I} - \mathbb{A}^{\rho_k+1})\,\widetilde{\boldsymbol{w}}^k$ and we would like to find such an index $\rho_k \in \mathbb{N}$ for which the second component of $\boldsymbol{w}_{\rho_k}^k$ is equal to β_2^k. Consequently, the first component of this vector has to be equal to $\beta_1^k + 2\beta_2^k\cos\gamma_k$. This follows directly from the first equation in (1.128).

Let us therefore denote $\widehat{\boldsymbol{w}}^k = (\beta_1^k + 2\beta_2^k\cos\gamma_k, \beta_2^k)^T$ and observe that the vector $\widehat{\boldsymbol{w}}^k$ satisfies $-\mathbb{A}\widehat{\boldsymbol{w}}^k = \boldsymbol{r}^k = (\mathbb{I} - \mathbb{A})\widetilde{\boldsymbol{w}}^k$ by (1.136). From this observation it follows that $\boldsymbol{w}_{\rho_k}^k = \widehat{\boldsymbol{w}}^k$ if and only if $(\mathbb{I} - \mathbb{A}^{\rho_k+1})\,\widetilde{\boldsymbol{w}}^k = (\mathbb{I} - \mathbb{A}^{-1})\widetilde{\boldsymbol{w}}^k$, i.e. if

$$\left(\mathbb{A}^{\rho_k+1} - \mathbb{A}^{-1}\right)\widetilde{\boldsymbol{w}}^k \;=\; \boldsymbol{0}. \tag{1.138}$$

Since the eigenvalues of $\mathbb{A}$ are $\exp(\pm 2\pi i)$, we can express the j-th power of $\mathbb{A}$ in the following way

$$\begin{aligned} \mathbb{A}^j \;&=\; \frac{-1}{2\sin\gamma_k}\begin{pmatrix} \mathrm{e}^{\frac{1}{2}\pi i} & \mathrm{e}^{-\frac{1}{2}\pi i} \\ \mathrm{e}^{-\frac{1}{2}\pi i} & \mathrm{e}^{\frac{1}{2}\pi i} \end{pmatrix}\begin{pmatrix} \mathrm{e}^{-2j\pi i} & 0 \\ 0 & \mathrm{e}^{2j\pi i} \end{pmatrix}\begin{pmatrix} \mathrm{e}^{\frac{1}{2}\pi i} & -\mathrm{e}^{-\frac{1}{2}\pi i} \\ -\mathrm{e}^{-\frac{1}{2}\pi i} & \mathrm{e}^{\frac{1}{2}\pi i} \end{pmatrix} = \\ &=\; \frac{1}{\sin\gamma_k}\begin{pmatrix} \sin(1-2j)\gamma_k & \sin 2j\gamma_k \\ -\sin 2j\gamma_k & \sin(1+2j)\gamma_k \end{pmatrix}. \end{aligned} \tag{1.139}$$

Consequently, the condition (1.138) can be rewritten in the form

$$\frac{-\sin(\rho_k+2)\gamma_k}{\sin^2\gamma_k}\begin{pmatrix} \beta_1^k\sin\rho_k\gamma_k + \beta_2^k\sin(\rho_k-1)\gamma_k \\ \beta_2^k\sin\rho_k\gamma_k + \beta_1^k\sin(\rho_k+1)\gamma_k \end{pmatrix} \;=\; \boldsymbol{0}. \tag{1.140}$$

From the definition of β_1^k and β_2^k it follows that both β_1^k and β_2^k are negative. Thus, the vector on the left-hand side of (1.140) is zero only in the trivial cases $\gamma_k = m\pi$, $m \in \{0, 1, 2\}$. Therefore, (excluding these trivial cases) the whole expression on the left-hand side of (1.140) vanishes if and only if $(\rho_k+2)\gamma_k = m\pi$, for some $m \in \mathbb{N}$.

This means that if $\gamma_k = \frac{\mu_k}{\nu_k}\pi$ for some incommensurable $\mu_k, \nu_k \in \mathbb{N}$, $\nu_k \geq 2$, $\frac{\mu_k}{\nu_k} \in (0, 1)$, then choosing $\rho_k = m_k\nu_k - 2$ for any $m_k \in \mathbb{N}$ and p_j^k, q_j^k defined recursively using (1.134) leads to the fulfilment of (1.99) together with the boundary conditions (1.101) and (1.102). It remains to verify the fulfilment of the boundary conditions (1.103).

Using (1.137) and (1.139) we find out that for $j = 0, 1, \ldots, \rho_k$ there holds

$$\begin{pmatrix} p_j^k \\ q_j^k \end{pmatrix} = \left(\mathbb{I} - \mathbb{A}^{j+1}\right) \widetilde{\boldsymbol{w}}^k = \frac{\sin^2(j+1)\gamma_k}{\sin^2\gamma_k} \begin{pmatrix} \beta_1^k + \beta_2^k \frac{\sin j\gamma_k}{\sin(j+1)\gamma_k} \\ \beta_2^k + \beta_1^k \frac{\sin(j+2)\gamma_k}{\sin(j+1)\gamma_k} \end{pmatrix}. \quad (1.141)$$

According to (1.127) all p_j^k and q_j^k, $j \in \{0, 1, \ldots, \rho_k\}$, have to be negative. Comparing p_j^k with q_{j-1}^k we deduce that both terms $\beta_1^k + \beta_2^k \frac{\sin j\gamma_k}{\sin(j+1)\gamma_k}$ and $\beta_2^k + \beta_1^k \frac{\sin(j+1)\gamma_k}{\sin j\gamma_k} = \frac{\sin(j+1)\gamma_k}{\sin j\gamma_k}\left(\beta_1^k + \beta_2^k \frac{\sin j\gamma_k}{\sin(j+1)\gamma_k}\right)$ have to be negative and that is why

$$\frac{\sin(j+1)\gamma_k}{\sin j\gamma_k} > 0, \qquad \text{for all } j \in \{1, 2, \ldots, \rho_k\}. \quad (1.142)$$

However, this property implies that one can consider only $\mu_k = 1$ and $m_k = 1$, i.e. $\gamma_k = \frac{\pi}{\nu_k}$ and $\rho_k = \nu_k - 2$. Indeed, if $m_k \geq 2$, then $\rho_k \geq 2\nu_k - 2 \geq 2$ and for $j = \left\lfloor \frac{\nu_k}{\mu_k} \right\rfloor \leq \nu_k \leq \frac{\rho_k+2}{2} \leq \rho_k$ it holds

$$2\pi > (j+1)\gamma_k = \left(\left\lfloor \frac{\nu_k}{\mu_k} \right\rfloor + 1\right) \frac{\mu_k}{\nu_k}\pi > \pi > \left\lfloor \frac{\nu_k}{\mu_k} \right\rfloor \frac{\mu_k}{\nu_k}\pi = j\gamma_k > 0, \quad (1.143)$$

which results in $\frac{\sin(j+1)\gamma_k}{\sin j\gamma_k} < 0$.

Similarly, if $\rho_k = \nu_k - 2$ and $\gamma_k = \frac{\mu_k}{\nu_k}\pi$ for some $\mu_k \geq 2$, then for $j = \left\lfloor \frac{\nu_k}{\mu_k} \right\rfloor = \left\lfloor \frac{\rho_k+2}{\mu_k} \right\rfloor \leq \left\lfloor \frac{\rho_k}{2} + 1 \right\rfloor \leq \rho_k$ the inequality (1.143) again causes the unfulfilment of the boundary condition (1.103). Here we considered only $\rho_k \geq 1$, since for $\rho_k = 0$ there is just one $j = 0$ and thus the condition (1.142) is pointless.

Hence, we construct a matched asymptotic expansion in two-dimensional polygonal domain containing only exponential boundary layers. Unfortunately, we were able to derive the exact formula only for the inner angles (i.e. angles included by two neighboring outflow boundary edges) of the form π/m, $m \in \mathbb{N}$, $m \geq 2$. Analogous approach can be used for derivation of a matched asymptotic expansion in 3D, see, for instance, López et al. (2007).

1.3.4 Parabolic boundary layers

In the previous sections we have considered that no parabolic boundary layers occur in the solution, i.e. $\Gamma_0 = \emptyset$. One can use an analogous approach as in the case of the exponential boundary layers and derive a matched asymptotic expansion in the parabolic boundary layer(s) (for more details see e.g. Eckhaus (1979) or Goering et al. (1983)). We shortly describe its construction.

Firstly, we employ the same global (regular) expansion. Further, we construct the local expansion by stretching the scale in the ξ_k direction. However, in this case we use the transformation $\Psi_k^{\sqrt{\varepsilon}} : (x, y) \to \left(\frac{\xi_k}{\sqrt{\varepsilon}}, \eta_k\right)$, where $(\xi_k, \eta_k) = \Psi_k(x, y)$.

Under the transformation $\Psi_k^{\sqrt{\varepsilon}}$ the differential operator L (cf. (1.49)) changes into

$$L_k^{\sqrt{\varepsilon}} = -\frac{\partial^2}{\partial\xi_k^2} - \varepsilon\frac{\partial^2}{\partial\eta_k^2} + \frac{B_1^k(\sqrt{\varepsilon}\xi_k, \eta_k)}{\sqrt{\varepsilon}}\frac{\partial}{\partial\xi_k} + B_2^k(\sqrt{\varepsilon}\xi_k, \eta_k)\frac{\partial}{\partial\eta_k}. \quad (1.144)$$

A boundary edge $e_k \subset \Gamma_0$ (where parabolic boundary layer occurs) is characterized by the condition $B_1^k(0, \eta_k) = -\boldsymbol{b}(\Psi_k^{-1}(0, \eta_k)) \cdot \boldsymbol{n}_k = 0$, for all $\eta_k \in [0, d_k]$.

Therefore, the Taylor expansion (1.65) of the function $B_1^k(\xi_k,\eta_k)$ in the variable ξ_k does not contain the first (absolute) term, hence

$$\frac{B_1^k(\sqrt{\varepsilon}\xi_k,\eta_k)}{\sqrt{\varepsilon}} \;=\; \sum_{j=1}^{\infty}\frac{\xi_k^j\,(\sqrt{\varepsilon})^{j-1}}{j!}\,\frac{\partial^j B_1^k(0,\eta_k)}{\partial\xi_k^j} \;=\; \sum_{j=1}^{\infty}\frac{\xi_k^j\,(\sqrt{\varepsilon})^{j-1}}{j!}B_{1,j}^k(0,\eta_k). \tag{1.145}$$

Using this expansion, we can express the differential operator $L_k^{\sqrt{\varepsilon}}$ in the local coordinates as

$$L_k^{\sqrt{\varepsilon}} \;=\; \sum_{j=0}^{\infty}\left(\sqrt{\varepsilon}\right)^j F_k^{(j)}, \tag{1.146}$$

where

$$\begin{array}{rclrcl}
F_k^{(0)} &=& -\frac{\partial^2}{\partial\xi_k^2} &+& \xi_k B_{1,1}^k(0,\eta_k)\,\frac{\partial}{\partial\xi_k} &+& B_2^k(0,\eta_k)\,\frac{\partial}{\partial\eta_k},\\
F_k^{(1)} &=& &+& \frac{1}{2}\xi_k^2 B_{1,2}^k(0,\eta_k)\,\frac{\partial}{\partial\xi_k} &+& \xi_k B_{2,1}^k(0,\eta_k)\,\frac{\partial}{\partial\eta_k},\\
F_k^{(2)} &=& -\frac{\partial^2}{\partial\eta_k^2} &+& \frac{1}{6}\xi_k^3 B_{1,3}^k(0,\eta_k)\,\frac{\partial}{\partial\xi_k} &+& \frac{1}{2}\xi_k^2 B_{2,2}^k(0,\eta_k)\,\frac{\partial}{\partial\eta_k},\\
F_k^{(j)} &=& &&\frac{1}{(j+1)!}\xi_k^{j+1} B_{1,j+1}^k(0,\eta_k)\,\frac{\partial}{\partial\xi_k} &+& \frac{1}{j!}\xi_k^j B_{2,j}^k(0,\eta_k)\,\frac{\partial}{\partial\eta_k}, \quad j\ge 3.
\end{array} \tag{1.147}$$

Consequently, the local expansion of the m-th order in the parabolic boundary layer has a form $W_{loc}^{k,m}(\xi_k,\eta_k)=\sum_{j=0}^{2m}(\sqrt{\varepsilon})^j\,W_j^k(\xi_k,\eta_k)$, where the local corrections W_j^k satisfy the boundary layer equations in $\mathbb{R}^+\times(0,d_k)$

$$F_k^{(0)}W_0^k \;=\; 0, \tag{1.148}$$

$$F_k^{(0)}W_j^k \;=\; -\sum_{i=1}^{j}F_k^{(i)}W_{j-i}^k, \qquad \text{for } j=1,2,\ldots,2m, \tag{1.149}$$

equipped for all $j=0,1,\ldots,2m$ with the boundary conditions (functions u_j, $j=0,1,\ldots,m$, and $u_{g,m}$ are defined in Definition 1.3.1, page 15)

$$W_j^k(0,\eta_k) \;=\; -u_{j/2}\left(\Psi_k^{-1}(0,\eta_k)\right), \quad \forall\eta_k\in(0,d_k), \quad j \text{ even}, \tag{1.150}$$

$$W_j^k(0,\eta_k) \;=\; 0, \quad \forall\eta_k\in(0,d_k), \quad j \text{ odd}, \tag{1.151}$$

$$\lim_{\xi_k\to+\infty} W_j^k(\xi_k,\eta_k) \;=\; 0, \quad \forall\eta_k\in(0,d_k), \tag{1.152}$$

$$W_j^k(\xi_k,0) \;=\; 0, \quad \forall\xi_k\in\mathbb{R}^+. \tag{1.153}$$

To see that $u_{g,m}(x,y)+W_{loc}^{k,m}\left(\Psi_k^{\sqrt{\varepsilon}}(x,y)\right)$ is a matched asymptotic expansion in the parabolic boundary layer in the vicinity of an edge e_k, consider

$$\begin{aligned}
&L\left(u(x,y)-u_{g,m}(x,y)-W_{loc}^{k,m}\left(\Psi_k^{\sqrt{\varepsilon}}(x,y)\right)\right) \;=\\
&\quad=\; \varepsilon^{m+1}\Delta u_m(x,y)-\sum_{i=2m+1}^{\infty}(\sqrt{\varepsilon})^i\sum_{j=0}^{2m}F_k^{(i-j)}W_j^k(\Psi_k(x,y)) \;=\; \mathcal{O}(\varepsilon^{m+1/2}),
\end{aligned} \tag{1.154}$$

$$\begin{aligned}
&u\left(\Psi_k^{-1}(0,\eta_k)\right)-u_{g,m}\left(\Psi_k^{-1}(0,\eta_k)\right)-W_{loc}^{k,m}\left(0,\eta_k\right) \;=\\
&\quad=\; 0-\sum_{j=0}^{m}\varepsilon^j u_j\left(\Psi_k^{-1}(0,\eta_k)\right)-\sum_{j=0}^{m}\varepsilon^j W_{2j}^k(0,\eta_k) \;=\; 0 \qquad \text{and}
\end{aligned} \tag{1.155}$$

$$\left(u(x,y)-u_{g,m}(x,y)-W_{loc}^{k,m}\left(\Psi_k^{\sqrt{\varepsilon}}(x,y)\right)\right)\Big|_{\Gamma_-} = \mathcal{O}(\varepsilon^{m+1/2}). \tag{1.156}$$

Hence, applying the comparison principle one can show that there exists a constant $C_P > 0$ independent of ε such that in the vicinity of the edge e_k there holds

$$\left|u(x,y) - u_{g,m}(x,y) - W^{k,m}_{loc}\left(\Psi^{\sqrt{\varepsilon}}_k(x,y)\right)\right| \;\leq\; C_P\,\varepsilon^{m+1/2}. \tag{1.157}$$

Let us again derive a particular form of the zeroth-order matched asymptotic expansion in the parabolic boundary layer. For simplicity, let us consider that $\boldsymbol{b}$ is constant in Ω and $\boldsymbol{b}\cdot\boldsymbol{n}_k = 0$. Then $B^k_1(\xi_k,\eta_k) = 0$ and $B^k_2(\xi_k,\eta_k) = -\boldsymbol{b}\cdot\boldsymbol{t}_k = |\boldsymbol{b}|$. Thus, function W^k_0 is a solution of the parabolic initial-boundary value problem

$$-\frac{\partial^2 W^k_0}{\partial \xi_k^2} + |\boldsymbol{b}|\,\frac{\partial W^k_0}{\partial \eta_k} \;=\; 0 \qquad \text{in } \mathbb{R}^+\times(0,d_k), \tag{1.158}$$

equipped with the initial condition $W^k_0(\xi_k,0) = 0$ for all $\xi_k \in \mathbb{R}^+$ and the boundary condition $W^k_0(0,\eta_k) = g(\eta_k) = -u_0(\Psi^{-1}_k(0,\eta_k)) = -u_0(P^{k-1} - \eta_k\boldsymbol{t}_k)$ for all $\eta_k \in (0,d_k)$. This is, in fact, the heat equation whose solution can be expressed in a form

$$W^k_0(\xi_k,\eta_k) \;=\; \int_0^{\eta_k} G(\xi_k,\eta_k - s)g(s)\,\mathrm{d}s, \tag{1.159}$$

where

$$G(\xi,\eta) \;=\; \frac{\xi\sqrt{|\boldsymbol{b}|}}{2\sqrt{\pi\eta^3}}\,\exp\left(-\frac{|\boldsymbol{b}|\xi^2}{4\eta}\right). \tag{1.160}$$

If we employ a substitution $s = \eta_k - \frac{|\boldsymbol{b}|\xi_k^2}{2r^2}$ the expression (1.159) takes a form

$$W^k_0(\xi_k,\eta_k) \;=\; \sqrt{\frac{2}{\pi}}\int_{\xi_k\sqrt{\frac{|\boldsymbol{b}|}{2\eta_k}}}^{+\infty} g\left(\eta_k - \frac{|\boldsymbol{b}|\xi_k^2}{2r^2}\right)\exp\left(-\frac{r^2}{2}\right)\,\mathrm{d}r. \tag{1.161}$$

Further, considering $g(\eta_k) = -\eta_k\frac{f}{|\boldsymbol{b}|}$, we can evaluate the previous integral and obtain

$$W^k_0(\xi_k,\eta_k) \;=\; \frac{f\eta_k}{|\boldsymbol{b}|}\left[\mathrm{erf}\left(\frac{\xi_k\sqrt{|\boldsymbol{b}|}}{2\sqrt{\eta_k}}\right) - 1 + \frac{\xi_k^2|\boldsymbol{b}|}{4\eta_k\sqrt{\pi}}\,\Gamma\left(-\frac{1}{2},\frac{\xi_k^2|\boldsymbol{b}|}{4\eta_k}\right)\right]. \tag{1.162}$$

where $\Gamma(s,x) = \int_x^\infty t^{s-1}\exp(-t)\,\mathrm{d}t$ is the upper incomplete gamma function and $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2/2)\,\mathrm{d}t = 1 - \frac{1}{\sqrt{\pi}}\Gamma(1/2,x^2)$ is the error function. Finally, the zeroth-order matched asymptotic expansion in the parabolic boundary layer has a form (see Figure 1.2 for an example with $\varepsilon = 0.01$, $f \equiv 1$ and $|\boldsymbol{b}| = 1$)

$$u_0\,(x,y) + W^k_0\left(\frac{\xi_k(x,y)}{\sqrt{\varepsilon}},\eta_k(x,y)\right). \tag{1.163}$$

Remark 1.3.2. Since $G(0,\eta) = 0$ for all $\eta \in (0,d_k)$ one may deduce from (1.159) that $W^k_0(0,\eta_k) = 0$ for all $\eta_k \in (0,d_k)$ and the boundary condition is not fulfilled. Let us therefore compute the limit $\lim_{\xi\to 0}\int_0^\eta G(\xi,\eta-s)g(s)$. Since for any $0 \leq A \leq B \leq \eta$ there holds

$$\int_A^B G(\xi,\eta-s)\,\mathrm{d}s \;=\; \mathrm{erf}\left(\frac{\xi\sqrt{|\boldsymbol{b}|}}{2\sqrt{\eta-B}}\right) - \mathrm{erf}\left(\frac{\xi\sqrt{|\boldsymbol{b}|}}{2\sqrt{\eta-A}}\right), \tag{1.164}$$

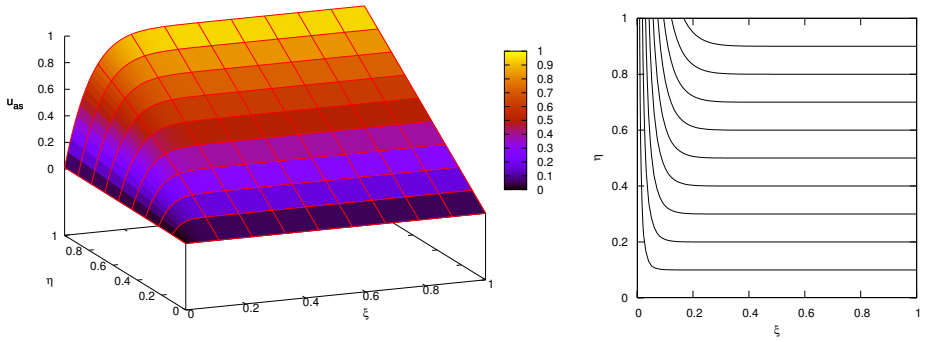

Figure 1.2: The zeroth-order matched asymptotic expansion in the parabolic boundary layer and its contours, $\varepsilon = 0.01$, $f = 1$, $|\boldsymbol{b}| = 1$.

we can split the integral (1.159) into

$$\int_0^{\eta} G(\xi, \eta - s) g(s)\,\mathrm{d}s \;=\; \int_0^{\eta - C\xi} G(\xi, \eta - s) g(s)\,\mathrm{d}s + \int_{\eta - C\xi}^{\eta} G(\xi, \eta - s) g(s)\,\mathrm{d}s \tag{1.165}$$

and estimate

$$\left[\operatorname{erf}\left(\frac{\sqrt{\xi|\boldsymbol{b}|}}{2\sqrt{C}}\right) - \operatorname{erf}\left(\frac{\xi\sqrt{|\boldsymbol{b}|}}{2\sqrt{\eta}}\right)\right] \min_{s\in[0,\eta-C\xi]} g(s) \;\le\; \int_0^{\eta - C\xi} G(\xi, \eta - s) g(s)\,\mathrm{d}s \;\le$$
$$\le \left[\operatorname{erf}\left(\frac{\sqrt{\xi|\boldsymbol{b}|}}{2\sqrt{C}}\right) - \operatorname{erf}\left(\frac{\xi\sqrt{|\boldsymbol{b}|}}{2\sqrt{\eta}}\right)\right] \max_{s\in[0,\eta-C\xi]} g(s) \tag{1.166}$$

and

$$\left[1 - \operatorname{erf}\left(\frac{\sqrt{\xi|\boldsymbol{b}|}}{2\sqrt{C}}\right)\right] \min_{s\in[\eta-C\xi,\eta]} g(s) \;\le\; \int_{\eta - C\xi}^{\eta} G(\xi, \eta - s) g(s)\,\mathrm{d}s \;\le$$
$$\le \left[1 - \operatorname{erf}\left(\frac{\sqrt{\xi|\boldsymbol{b}|}}{2\sqrt{C}}\right)\right] \max_{s\in[\eta-C\xi,\eta]} g(s), \tag{1.167}$$

where we employed the equality $\lim_{z\to+\infty} \operatorname{erf}(z) = 1$. Since there also holds $\operatorname{erf}(0) = 0$, taking the limit $\xi \to 0$ gives $\lim_{\xi\to 0} \int_0^{\eta} G(\xi, \eta - s) g(s)\,\mathrm{d}s = g(\eta)$, providing g is continuous and bounded function in $[0, \eta]$.

Using the equality $G(\sqrt{\eta}, \eta) = \frac{\sqrt{|\boldsymbol{b}|}}{2\eta\sqrt{\pi}} \exp\left(-\frac{|\boldsymbol{b}|}{4}\right)$ we realize that the function $G(\xi, \eta)$ has a singularity in $[0, 0]$. Moreover, considering $g \equiv 1$ in the above derived limit we find out that there holds $\lim_{\xi\to 0} \int_0^{\eta} G(\xi, \eta - s)\,\mathrm{d}s = 1$. Hence, $G(0, \eta)$ is the Dirac delta function.

Remark 1.3.3. If we consider that the function g possesses the Taylor expansion

of the form $g\left(\eta_k - \frac{|\boldsymbol{b}|\xi_k^2}{2r^2}\right) = \sum_{j=0}^{\infty} \frac{g^{(j)}(\eta_k)}{j!}\left(-\frac{|\boldsymbol{b}|\xi_k^2}{2r^2}\right)^j$, then from (1.161) it follows

$$\begin{aligned} W_0^k(\xi_k,\eta_k) &= \sqrt{\frac{2}{\pi}}\int_{\xi_k\sqrt{\frac{|\boldsymbol{b}|}{2\eta_k}}}^{+\infty}\sum_{j=0}^{\infty}\frac{g^{(j)}(\eta_k)}{j!}\left(-\frac{|\boldsymbol{b}|\xi_k^2}{4}\right)^j\left(\frac{2}{r^2}\right)^j\exp\left(-\frac{r^2}{2}\right)\mathrm{d}r = \\ &= \sqrt{\frac{2}{\pi}}\sum_{j=0}^{\infty}\frac{g^{(j)}(\eta_k)}{j!}\left(-\frac{|\boldsymbol{b}|\xi_k^2}{4}\right)^j\int_{\frac{\xi_k^2|\boldsymbol{b}|}{4\eta_k}}^{+\infty} t^{-j}\exp(-t)\frac{\mathrm{d}t}{\sqrt{2t}} = \\ &= \frac{1}{\sqrt{\pi}}\sum_{j=0}^{\infty}\frac{g^{(j)}(\eta_k)}{j!}\left(-\frac{|\boldsymbol{b}|\xi_k^2}{4}\right)^j\Gamma\left(\frac{1}{2}-j,\frac{|\boldsymbol{b}|\xi_k^2}{4\eta_k}\right), \end{aligned}$$

whenever the interchange of the sum and the integral is admissible. (One may again use the Fubini and the Tonelli theorems as in Remark 1.2.1.) This complicated structure of the parabolic boundary layer function causes difficulties in a derivation of the uniformly convergent numerical schemes (for details see Ainsworth and Dörfler (2001) or Shishkin (1997)).

1.4 Numerical experiments

Now we shall numerically verify the theoretical estimate (1.113) for the zeroth-order matched asymptotic expansion of the solution of the equation (1.49) with simple data $\boldsymbol{b} = (1,0)^T$ and $f \equiv 1$ on a triangle with vertices $P^0 = [0, -\tan\frac{\gamma}{2}]$, $P^1 = [1,0]$ and $P^2 = [0, \tan\frac{\gamma}{2}]$, where $\gamma = \gamma(r) = \frac{\pi}{r+2}$, $r \in \{0,1,2,4\}$ (see Figure 1.3).

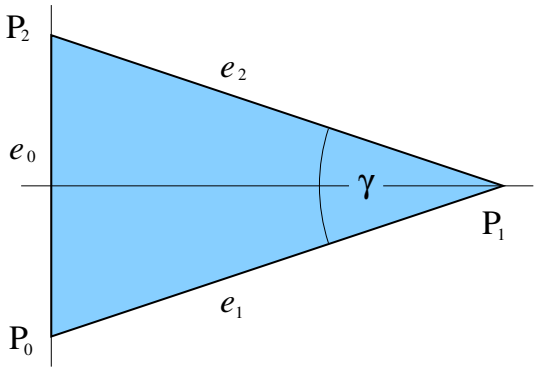

Figure 1.3: A simple triangular domain considered in the numerical experiment.

The mappings Ψ_1, Ψ_2 corresponding to the edges e_1, e_2 are then defined by

$$\Psi_1(x,y) = (\xi_1(x,y),\eta_1(x,y)) \qquad \text{and} \qquad \Psi_2(x,y) = (\xi_2(x,y),\eta_2(x,y)), \tag{1.168}$$

where

$$\xi_1(x,y) = (1-x)\sin\tfrac{\gamma}{2} + y\cos\tfrac{\gamma}{2}, \quad \eta_1(x,y) = x\cos\tfrac{\gamma}{2} + (y+\tan\tfrac{\gamma}{2})\sin\tfrac{\gamma}{2}, \tag{1.169}$$

$$\xi_2(x,y) = (1-x)\sin\tfrac{\gamma}{2} - y\cos\tfrac{\gamma}{2}, \quad \eta_2(x,y) = (1-x)\cos\tfrac{\gamma}{2} + y\sin\tfrac{\gamma}{2}. \tag{1.170}$$

Consequently, the inverse mappings satisfy

$$\Psi_1^{-1}(0,\eta_1(x,y)) = \left(\eta_1(x,y)\cos\tfrac{\gamma}{2}, \eta_1(x,y)\sin\tfrac{\gamma}{2} - \tan\tfrac{\gamma}{2}\right), \quad (1.171)$$

$$\Psi_2^{-1}(0,\eta_2(x,y)) = \left(1-\eta_2(x,y)\cos\tfrac{\gamma}{2}, \eta_2(x,y)\sin\tfrac{\gamma}{2}\right). \quad (1.172)$$

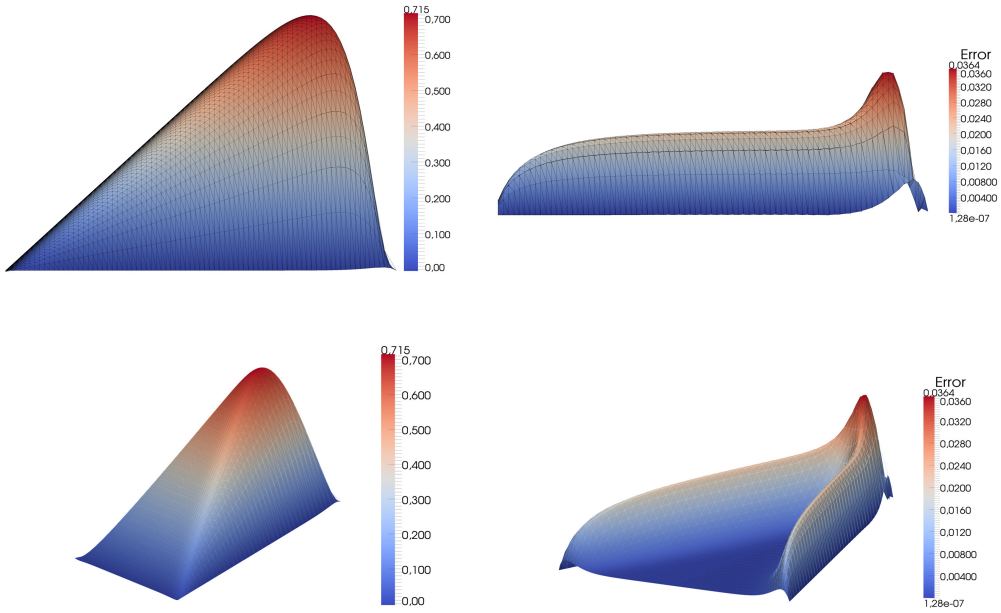

Figure 1.4: The three-dimensional plots of the zeroth-order matched asymptotic expansion (left) and the corresponding distribution of error (right) for the case $\gamma = \frac{\pi}{4}$ and $\varepsilon = 0.01$.

Figure 1.4 (left) shows the particular case of the zeroth-order asymptotic expansion $u_{as,0}$ ($\gamma = \frac{\pi}{4}$ and $\varepsilon = 0.01$). The general form of the function $u_{as,0}$ for this simple domain is

$$\begin{aligned} u_{as,0}(x,y) = {} & u_0(x,y) - u_0\left(\Psi_1^{-1}(0,\eta_1(x,y))\right)\exp\left(B_1^1(0,\eta_1(x,y))\,\frac{\xi_1(x,y)}{\varepsilon}\right) - \\ & - u_0\left(\Psi_2^{-1}(0,\eta_2(x,y))\right)\exp\left(B_1^2(0,\eta_2(x,y))\,\frac{\xi_2(x,y)}{\varepsilon}\right) + \\ & + u_0(P^1)\left\{\sum_{j=0}^{r}\exp\left(p_j^r\frac{\xi_1(x,y)}{\varepsilon} + q_j^r\frac{\xi_2(x,y)}{\varepsilon}\right) - \right. \\ & \left. - \sum_{j=0}^{r-1}\exp\left(p_{j+1}^r\frac{\xi_1(x,y)}{\varepsilon} + q_j^r\frac{\xi_2(x,y)}{\varepsilon}\right)\right\}, \quad (1.173) \end{aligned}$$

where $u_0(x,y)$ is the solution of the reduced problem given by (1.53)–(1.54) and

$$B_1^1 = B_1^1(0,d_1) = -\boldsymbol{b}(P^1)\cdot\boldsymbol{n}_1, \quad p_j^r = \frac{\sin^2((j+1)\gamma)}{\sin^2\gamma}\left(B_1^1 + B_1^2\frac{\sin(j\gamma)}{\sin((j+1)\gamma)}\right), \quad (1.174)$$

$$B_1^2 = B_1^2(0,0) = -\boldsymbol{b}(P^1)\cdot\boldsymbol{n}_2, \quad q_j^r = \frac{\sin^2((j+1)\gamma)}{\sin^2\gamma}\left(B_1^2 + B_1^1\frac{\sin((j+2)\gamma)}{\sin((j+1)\gamma)}\right), \quad (1.175)$$

with $\boldsymbol{n}_1 = \left(\sin\frac{\gamma}{2}, -\cos\frac{\gamma}{2}\right)$, $\boldsymbol{n}_2 = \left(\sin\frac{\gamma}{2}, \cos\frac{\gamma}{2}\right)$ and $j \in \{0,1,\ldots,r\}$.

Numerical experiments were carried out with the use of the discontinuous Galerkin method (see, e.g. Riviére (2008)) with piecewise linear approximations on uniformly refined meshes having approximately 5000 elements for several different values of γ and ε. The difference between the numerical solution u_h and asymptotic expansion $u_{as,0}$ is depicted in Figure 1.4 (right). Table 1.1 records the corresponding errors $u_h - u_{as,0}$ in $L^\infty(\Omega)$-norm together with the experimental order of convergence (EOC) with respect to ε. We observe that $EOC \approx 1$ for all considered angles γ, which is in a good agreement with derived theoretical results of order $O(\varepsilon)$ according to (1.113).

ε	$\gamma = \pi/2$	$\gamma = \pi/3$	$\gamma = \pi/4$	$\gamma = \pi/6$
0.04	8.0211E-02	9.5164E-02	1.4183E-01	2.6822E-01
0.02	3.3935E-02	4.6841E-02	7.8396E-02	1.4062E-01
0.01	1.6045E-02	2.1667E-02	3.8123E-02	7.7320E-02
0.005	9.0778E-03	1.1417E-02	2.0733E-02	4.2789E-02
EOC	1.051	1.029	0.936	0.881

Table 1.1: Computational errors in $L^\infty(\Omega)$-norm and experimental orders of convergence for different values of γ and ε (adopted from Lamač (2013)).

2. Stabilization and Upwind techniques

In this section we present several stabilization methods and demonstrate their behavior on simple one-dimensional examples. In the second part we prove the uniform convergence of the Il'in-Allen-Southwell scheme in 1D.

2.1 Stabilization in 1D

For an illustration let us consider a one-dimensional convection-diffusion equation

$$\begin{aligned} -\varepsilon u'' + b(x)u' &= f(x) \qquad \text{in } \Omega = (0,1), && (2.1)\\ u|_{\partial\Omega} &= 0, && (2.2) \end{aligned}$$

where $b(x) > \underline{\beta} > 0$, $-b' \geq 0$, $1 \gg \varepsilon > 0$ and $f \in L^2(\Omega)$.

For solving the equations (2.1)–(2.2) we would like to use the finite element method. Thus, we have to construct the weak formulation of (2.1)–(2.2). Multiplying (2.1) by any function $\varphi \in H_0^1(\Omega)$, integrating over Ω and using Green's theorem or integration by parts in 1D (Theorem 4.1.1, page 124) the weak formulation reads

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

$$a_1(u,\varphi) \;=\; (f,\varphi)_\Omega \qquad \forall \varphi \in H_0^1(\Omega), \tag{2.3}$$

where

$$a_1(u,\varphi) \;=\; \varepsilon(u',\varphi')_\Omega + (bu',\varphi)_\Omega. \tag{2.4}$$

Then $a_1(\varphi,\varphi) = \varepsilon|\varphi|_{1,\Omega}^2 - \frac{1}{2}(b',\varphi^2)_\Omega \geq \varepsilon|\varphi|_{1,\Omega}^2$ and since it is also

$$a_1(u,\varphi) \;\leq\; \varepsilon|u|_{1,\Omega}|\varphi|_{1,\Omega} + \|b\|_{\infty,\Omega}|u|_{1,\Omega}|\varphi|_{0,\Omega} \;\leq\; \left(\varepsilon + \frac{1}{\pi}\|b\|_{\infty,\Omega}\right)|u|_{1,\Omega}|\varphi|_{1,\Omega}, \tag{2.5}$$

it follows from the Lax-Milgram theorem (Theorem 4.1.2, page 125) that there exists a unique solution to this weak formulation. In estimates (2.5) we have applied the Cauchy-Schwarz-Bunyakovsky inequality (Theorem 4.1.4, page 126) and the one-dimensional Friedrichs' inequality (Theorem 4.1.3, page 125). (See Definition 4.1.1, page 124, for the definition of norms.)

To define the finite element discretization of (2.1)–(2.2) we introduce a partition $\mathcal{T}_h$ of the domain Ω consisting of a finite number of open intervals $I_j = (x_{j-1}, x_j)$, $j = 1, 2, \ldots, N$, $x_0 = 0$, $x_N = 1$. For all types of partitions the discretization parameter h in the notation $\mathcal{T}_h$ is a positive real number satisfying $|I_j| \leq h$ for all $j = 1, 2, \ldots, N$. Here $|I_j| = x_j - x_{j-1}$ denotes the length of the interval I_j.

We obtain Galerkin's finite element discretization of (2.1)–(2.2) simply by replacing the space $H_0^1(\Omega)$ by a finite element subspace $V_h = X_h \cap H_0^1(\Omega)$, where

$$X_h = \{\varphi_h \in C(\Omega), \varphi_h|_{I_j} \in P_1(I_j)\ \forall j = 1, 2, \ldots, N\}. \tag{2.6}$$

Then $u_h \in V_h$ is a discrete solution of (2.1)–(2.2) if

$$a_1(u_h, \varphi_h) \;=\; (f, \varphi_h)_\Omega \qquad \forall \varphi_h \in V_h. \tag{2.7}$$

Again, in the space V_h there exists a unique solution to this discrete problem.

2.1.1 Spurious oscillations

Let us now for simplicity consider the equation (2.1) with constant data $b = const.$, $f = const.$ and let the partition $\mathcal{T}_h$ of the domain Ω be equidistant with the mesh parameter h satisfying $h = 1/N$. If we denote $u_j = u_h(x_j)$, for $j = 0, 1, \ldots, N$, then the equation (2.7) can be rewritten in the form

$$\varepsilon \, \frac{-u_{j-1} + 2u_j - u_{j+1}}{h} + b \, \frac{u_{j+1} - u_{j-1}}{2} \;=\; fh \qquad \forall\, j = 1, 2, \ldots, N-1. \tag{2.8}$$

The solution of this difference equation has a form

$$u_j \;=\; \begin{cases} \frac{f}{b}\left(jh - \frac{r^j - 1}{r^N - 1}\right), & \text{for } r \;=\; \frac{1+\mathrm{Pe}}{1-\mathrm{Pe}} \text{ and } \mathrm{Pe} \neq 1, \\ \frac{f}{b}\, jh, & \text{for } \mathrm{Pe} = 1, \end{cases} \tag{2.9}$$

where $\mathrm{Pe} = \frac{bh}{2\varepsilon}$ is the so-called Péclet number. When b is nonconstant we also write $\mathrm{Pe}(x) = \frac{b(x)h}{2\varepsilon}$. For large Péclet numbers the value of r is approximately equal to -1 and consequently the discrete solution oscillates (see Figure 2.1). However the exact solution

$$u(x) \;=\; \frac{f}{b}\left(x - \frac{\exp\left(\frac{b}{\varepsilon}x\right) - 1}{\exp\left(\frac{b}{\varepsilon}\right) - 1}\right) \tag{2.10}$$

does not possess any oscillations. Thus, the oscillations in the discrete solution are spurious and we need to adjust the method in order to remove them.

There are two main possibilities as to how this may be done: We can change the discretization of the derivatives or we can refine the computational mesh. The first technique is called stabilization and in what follows we describe several (in some cases equivalent) methods that stabilize the discrete solution.

2.1.2 SUPG method

The streamline upwind Petrov/Galerkin (SUPG) method introduced by Brooks and Hughes (1982) adds weighted residuals $R(u) = -\varepsilon u'' + bu' - f$ to the usual Galerkin finite element method. Since $R(u)$ vanishes for the exact solution, we can add any multiple of $R(u)$ to the weak formulation and the method remains consistent, providing $u \in H_0^1(\Omega) \cap H^2(\Omega)$. Thus, for any $\tau_j \in \mathbb{R}$ the SUPG method reads

Find $u_h \in V_h$ such that

$$a_1(u_h, \varphi_h) + \sum_{j=1}^{N} \tau_j (R(u_h), b\,\varphi_h')_{I_j} \;=\; (f, \varphi_h)_\Omega, \qquad \text{for all } \varphi_h \in V_h. \tag{2.11}$$

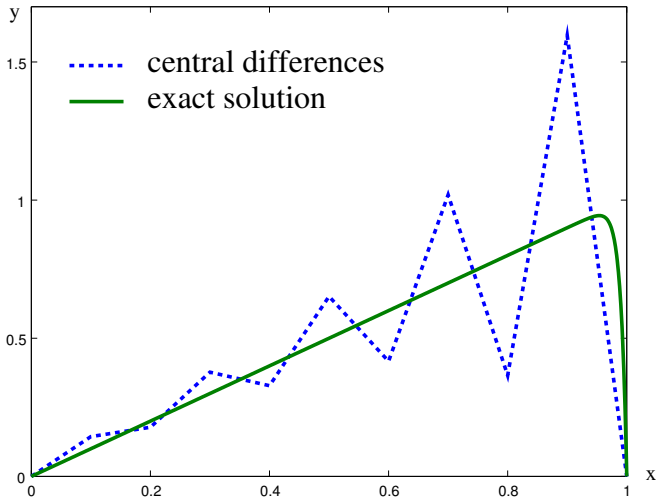

Figure 2.1: The solution obtained using the finite element method without any stabilization (or using the finite difference method with central differences) contains spurious oscillations. In this example we considered $\varepsilon = 0.01$, $b = f = 1$ and $h = 0.1$ (i.e. $\mathrm{Pe} = 5$).

In our case $u_h|_{I_j} \in P_1(I_j)$ which implies $R(u_h) = 0 + bu_h' - f$ and the equality (2.11) changes to

$$\varepsilon(u_h', \varphi_h')_\Omega + \sum_{j=1}^{N}(bu_h', \varphi_h + \tau_j b\,\varphi_h')_{I_j} \;=\; \sum_{j=1}^{N}(f, \varphi_h + \tau_j b\,\varphi_h')_{I_j}, \qquad \text{for all } \varphi_h \in V_h. \tag{2.12}$$

The stabilization parameter τ_j affects the quality of the stabilization. For $\tau_j = 0$, $j = 1, 2, \ldots, N$ we obtain the original Galerkin method. If we consider a special case when all τ_j are nonzero and equal to τ, the stencil of the method takes the form

$$\left(\varepsilon + b^2\tau\right)\frac{-u_{j-1} + 2u_j - u_{j+1}}{h} + b\,\frac{u_{j+1} - u_{j-1}}{2} \;=\; fh, \qquad \forall\, j = 1, 2, \ldots, N-1. \tag{2.13}$$

Now choosing $\tau = \tau_{upw} = \frac{h}{2b}$ leads to the simple upwind scheme

$$\varepsilon\,\frac{-u_{j-1} + 2u_j - u_{j+1}}{h} + b\left(u_j - u_{j-1}\right) \;=\; fh, \qquad \forall\, j = 1, 2, \ldots, N-1. \tag{2.14}$$

The solution of the respective difference equation is also easily computable

$$u_j^{upw} \;=\; \frac{f}{b}\left(jh - \frac{r^j - 1}{r^N - 1}\right), \qquad r \;=\; 1 + 2\mathrm{Pe}. \tag{2.15}$$

Unlike the central difference solution (2.9), the upwind solution (2.15) does no longer oscillate. However it does not converge uniformly (with respect to ε) since one can prove only that there exist positive constants $\widehat{C}_0, \widehat{C}_1, \widehat{C}_2$ and β^* such that (cf. e.g. (Roos et al., 2008, p. 49))

$$|u(x_j) - u_j^{upw}| \;\le\; \begin{cases} \widehat{C}_0 h\,[1 + \varepsilon^{-1}\exp(-\beta^*(1 - x_j)/\varepsilon)] & \text{for } h \le \varepsilon \\ \widehat{C}_1 h + \widehat{C}_2\exp(-\beta^*(1 - x_{j+1})/\varepsilon) & \text{for } h \ge \varepsilon \end{cases} \tag{2.16}$$

Thus, for $1 > h \geq \varepsilon$ and $x_{j+1} \in (1 - \frac{\varepsilon}{\beta^*}|\ln h|, 1)$ the order of convergence is lost. To improve the convergence order we have to adjust the stabilization parameter. The best adjustment provides the Il'in-Allen-Southwell scheme which uses the stabilization parameter $\tau_* = \frac{h}{2b}\coth \mathrm{Pe} - \frac{\varepsilon}{b^2}$. For constant data the solution to the respective difference equation is

$$u_j^* \;=\; \frac{f}{b}\left(jh - \frac{r^j - 1}{r^N - 1}\right), \qquad r \;=\; \exp(2\mathrm{Pe}), \tag{2.17}$$

which means that it is nodally exact.

In section 2.2 we prove the uniform convergence of the Il'in-Allen-Southwell scheme for nonconstant data in the discrete maximum norm $\|v_h\|_{\infty,d} = \max\limits_{1\leq i\leq N} |v(x_i)|$.

2.1.3 Changing test functions

If we turn back to the equality (2.12) we observe that since $\tau_j b\,\varphi_h'$ is constant on the element I_j we can obtain a relation equivalent to (2.12) if we change the test (weighting) function from φ_h to $\varphi_h + \tau_j b\,\varphi_h'$ in (2.7) (see Figure 2.2 left). Consequently, for all $\varphi_h \in V_h$ there holds

$$\sum_{j=1}^{N}\left\{\varepsilon\Big(u_h', (\varphi_h + \tau_j b\,\varphi_h')'\Big)_{I_j} + \Big(bu_h', \varphi_h + \tau_j b\,\varphi_h'\Big)_{I_j}\right\} \;=\; \sum_{j=1}^{N}\Big(f, \varphi_h + \tau_j b\,\varphi_h'\Big)_{I_j}. \tag{2.18}$$

A method characterized by the use of different shape and test function spaces is called a *Petrov-Galerkin method* (see, e.g., Heinrich et al. (1977) for one of the first publication about this topic).

One can achieve the same effect as in the SUPG method by using continuous Petrov-Galerkin test functions. Since they are continuous, they belong to $H_0^1(\Omega)$, which can be in many cases useful. The simplest way is to choose continuous piecewise quadratic test functions. They are for each $j = 1, 2, \ldots, N-1$ defined as (see Figure 2.2 right)

$$\tilde{\varphi}_j \;=\; \begin{cases} \dfrac{x - x_{j-1}}{h} - \dfrac{3\sigma}{h^2}(x - x_{j-1})(x - x_j), & \text{for } x \in [x_{j-1}, x_j], \\[2ex] \dfrac{x_{j+1} - x}{h} + \dfrac{3\sigma}{h^2}(x - x_{j+1})(x - x_j), & \text{for } x \in (x_j, x_{j+1}], \\[2ex] 0, & \text{otherwise.} \end{cases} \tag{2.19}$$

Consequently, the left-hand side of (2.7) changes to

$$\begin{aligned} \varepsilon\left(u_h', \tilde{\varphi}_j'\right)_\Omega + (bu_h', \tilde{\varphi}_j)_\Omega \;&= \\ &= \left(\varepsilon + \sigma\,\frac{bh}{2}\right)\frac{-u_{j-1} + 2u_j - u_{j+1}}{h} + b\,\frac{u_{j+1} - u_{j-1}}{2}. \end{aligned} \tag{2.20}$$

Comparing (2.20) with (2.13) we realize that considering $\sigma = 1$ leads to the simple upwind scheme, whereas taking $\sigma = \coth \mathrm{Pe} - 1/\mathrm{Pe}$ yields the Il'in-Allen-Southwell scheme.

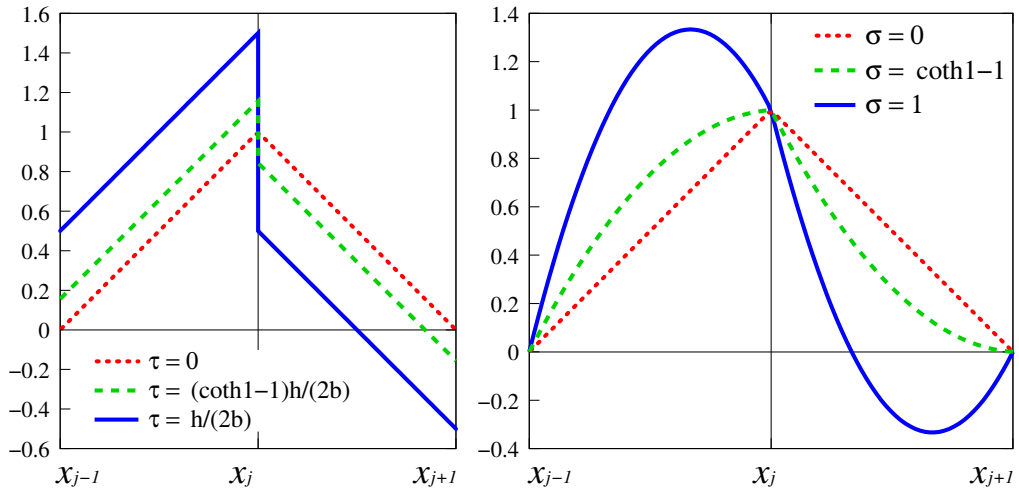

Figure 2.2: SUPG (left) and continuous Petrov-Galerkin (right) test functions for different choices of τ and σ.

In order to obtain the Il'in-Allen-Southwell scheme local Green's function of the adjoint operator of L can be also used as a test function. However, using this approach one cannot obtain the simple upwind scheme (cf. section 2.1.6).

2.1.4 Adding artificial diffusion

From the equality (2.13) it follows that one can obtain the stabilized solution also by adding an artificial diffusion to ε. Thus, instead of (2.7) we consider a discrete problem of the form

Find $u_h \in V_h$ such that

$$(\varepsilon + \widetilde{\varepsilon})\left(u_h', \varphi_h'\right)_\Omega + \left(bu_h', \varphi_h\right)_\Omega = \left(f, \varphi_h\right)_\Omega \qquad \forall \varphi_h \in V_h. \tag{2.21}$$

Comparing (2.21) with (2.13) we find out that we obtain the simple upwind scheme by choosing $\widetilde{\varepsilon} = \widetilde{\varepsilon}_{upw} = b^2\tau_{upw} = \frac{bh}{2}$. Similarly, taking $\widetilde{\varepsilon} = \widetilde{\varepsilon}_* = b^2\tau_* = \frac{bh}{2}\coth \mathrm{Pe} - \varepsilon$ yields the Il'in-Allen-Southwell scheme.

Since there holds $\widetilde{\varepsilon}_{upw} > \widetilde{\varepsilon}_*$ we say that the simple upwind scheme adds too much artificial diffusion to the original finite element (difference) method and the discrete solution is *overdiffusive* – adding greater amount of the artificial diffusion causes the smearing of layers (see Figure 2.3). On the other hand if the amount of the added artificial diffusion is to small it does not suppress all the spurious oscillations.

Thus, one has to choose the proper amount of the artificial diffusion. Let us, for instance, consider the equation (2.21) with $\varepsilon = 0.01$, $b = f = 1$ and $h = 1/N$, where $N \in \{10, 20, 30, 40, 50\}$. For each N and each $\widetilde{\varepsilon}$ we may compute the signed error in the last inner node of the computational domain, i.e. $u_{N-1} - u(x_{N-1})$.

We observe (see Figure 2.4) that despite the increasing number of nodes (i.e. decreasing h) the error of the simple upwind scheme evaluated at the last inner node of the computational domain can increase (cf. the example in the beginning of the section 2.2).

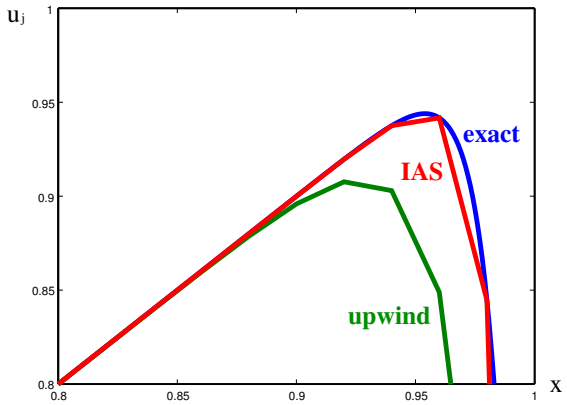

Figure 2.3: The discrete solution obtained by the simple upwind scheme is more smeared as compared with the Il'in-Allen-Southwell scheme (or the exact solution). The problem data are $\varepsilon = 0.01$, $b = f = 1$ and $h = 0.02$.

2.1.5 Adding bubble functions

Another way how we can stabilize the discrete solution is adding bubble functions to the space V_h (cf. Brezzi and Russo (1994)). For each element $I_j \in \mathcal{T}_h$ a bubble function $b_j \in L^2(\Omega)$ is any function satisfying $b_j|_{I_j} \in H_0^1(I_j)$ and $\operatorname{supp}\{b_j\} = \overline{I_j}$. Then the space of bubble functions is defined as $B = \operatorname{span}\{b_j, 1 \le j \le N\}$. Consequently, the space of shape and test functions is given by $W_h = V_h \oplus B$.

The finite element formulation then reads: Find $u_h = u_L + u_B \in W_h$ ($u_L \in V_h$, $u_B \in B$) such that

$$a_1(u_h, \varphi_L) = (f, \varphi_L)_\Omega \qquad \text{for all } \varphi_L \in V_h \qquad \text{and} \tag{2.22}$$

$$a_1(u_h, \varphi_B) = (f, \varphi_B)_\Omega \qquad \text{for all } \varphi_B \in B, \tag{2.23}$$

where we again for simplicity consider b, f to be constant functions.

Since u_L is linear function on each I_j and b_j vanishes on ∂I_j, we have

$$(u_L', b_j')_{I_j} = [u_L' b_j]_{x_{j-1}}^{x_j} - (u_L'', b_j)_{I_j} = 0 \qquad \text{and} \tag{2.24}$$

$$(b\,u_B', b_j)_{I_j} = \sum_{i=1}^{N} c_i (b\,b_i', b_j)_{I_j} = c_j (b\,b_j', b_j)_{I_j} = \frac{c_j}{2}\left[b\,b_j^2\right]_{x_{j-1}}^{x_j} = 0, \tag{2.25}$$

where we considered $u_B = \sum_{i=1}^{N} c_i b_i$.

Using these equalities the equation (2.23) written for one basis function $\varphi_B = b_j$ reduces to

$$\varepsilon (u_B', b_j')_{I_j} + (b u_L', b_j)_{I_j} = (f, b_j)_{I_j} \qquad \text{for all } j \in \{1, 2, \ldots, N\}. \tag{2.26}$$

Since $\operatorname{supp}\{b_j\} = \overline{I_j}$, we can write $u_B|_{I_j} = c_j b_j$, $c_j \in \mathbb{R}$. For the coefficients c_j then holds

$$c_j = \frac{(1, b_j)_{I_j}}{\varepsilon |b_j|_{1,I_j}^2} \left(f - b u_L'|_{I_j}\right). \tag{2.27}$$

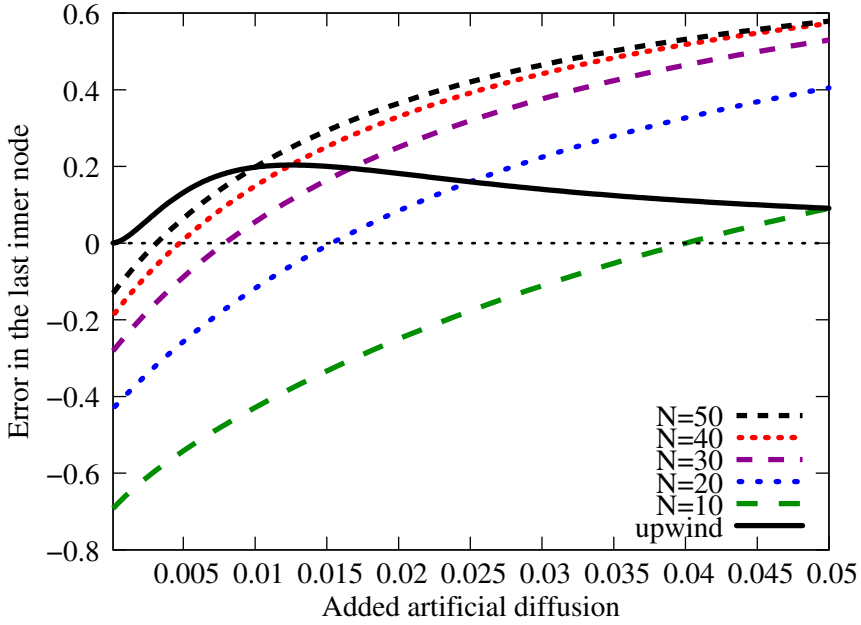

Figure 2.4: A comparison of the discrete solutions obtained by adding artificial diffusion. Each dashed curve corresponds to a different partition of Ω and the zero values of each dashed curve correspond to the artificial diffusion resulting in the Il'in-Allen-Southwell scheme. The intersection of the black solid curve with any dashed curve corresponds to the artificial diffusion providing the simple upwind scheme.

Further, for bubbles-containing terms in (2.22) we obtain

$$(u'_B, \varphi'_L)_\Omega \;=\; \sum_{j=1}^{N} c_j (b'_j, \varphi'_L)_{I_j} \;=\; 0 \qquad \text{and} \tag{2.28}$$

$$\begin{aligned}(b\,u'_B, \varphi_L)_\Omega \;&=\; \sum_{j=1}^{N} c_j b\,(b'_j, \varphi_L)_{I_j} \;=\; -\sum_{j=1}^{N} c_j b\,\varphi'_L|_{I_j}\,(b_j, 1)_{I_j} \;= \\ &=\; \sum_{j=1}^{N} \underbrace{\frac{1}{|I_j|}\frac{(b_j,1)^2_{I_j}}{\varepsilon |b_j|^2_{1,I_j}}}_{\tau_j^B} \Big(bu'_L - f, b\varphi'_L\Big)_{I_j} = \sum_{j=1}^{N} \tau_j^B \Big(bu'_L - f, b\varphi'_L\Big)_{I_j}. \end{aligned} \tag{2.29}$$

Consequently, the part $u_L \in V_h$ of the solution u_h for each $\varphi_L \in V_h$ satisfies

$$\varepsilon(u'_L, \varphi'_L)_\Omega + (bu'_L, \varphi_L)_\Omega + \sum_{j=1}^{N} \tau_j^B \Big(bu'_L - f, b\varphi'_L\Big)_{I_j} \;=\; (f, \varphi_L)_\Omega. \tag{2.30}$$

The formulation (2.30) is equivalent to the SUPG formulation (2.12) and the stabilization parameters τ_j^B depends only on the chosen bubbles. The simplest choice for the bubble function on the element I_j is the quadratic function $b_j(x) = (x - x_{j-1})(x_j - x)$. Then

$$\tau_j^{B_1} \;=\; \frac{1}{|I_j|}\frac{(b_j,1)^2_{I_j}}{\varepsilon |b_j|^2_{1,I_j}} \;=\; \frac{1}{h}\,\frac{\left(\frac{1}{6}h^3\right)^2}{\varepsilon\,\frac{1}{3}h^3} \;=\; \frac{h^2}{12\varepsilon}. \tag{2.31}$$

This choice is not suitable for $\varepsilon \to 0$. We can obtain the optimal stabilization parameter $\tau_j^{B_2} = \tau_*$ if we choose the bubble function b_j as the solution to the problem

$$\begin{aligned} -\varepsilon b_j'' + b\, b_j' &= 1 \qquad \text{in } I_j, && (2.32)\\ b_j &= 0 \qquad \text{on } \partial I_j. && (2.33) \end{aligned}$$

Then using Green's theorem (Theorem 4.1.1, page 124) and the equality (2.25) we obtain $-\varepsilon(b_j'', b_j)_{I_j} = \varepsilon |b_j|^2_{1,I_j} = (1, b_j)_{I_j}$ and since the solution of (2.32)–(2.33) has an exact form $b_j(x) = \frac{1}{b}\left(x - x_{j-1} - h\,\frac{\exp(-2\mathrm{Pe}(x_j - x)/h) - \exp(-2\mathrm{Pe})}{1-\exp(-2\mathrm{Pe})}\right)$, it holds

$$\begin{aligned} \tau_j^{B_2} &= \frac{1}{|I_j|}\frac{(b_j,1)^2_{I_j}}{\varepsilon |b_j|^2_{1,I_j}} = \frac{1}{h}(b_j,1)_{I_j} = \\ &= \frac{1}{h}\frac{1}{b}\left(\frac{h^2}{2} - h\,\frac{\frac{h}{2\mathrm{Pe}}(1-\exp(-2\mathrm{Pe})) - h\exp(-2\mathrm{Pe})}{1-\exp(-2\mathrm{Pe})}\right) = \\ &= \frac{h}{2b} - \frac{\varepsilon}{b^2} + \frac{h}{b}\frac{1}{\exp(2\mathrm{Pe})-1} = \frac{h}{2b}\frac{\exp(2\mathrm{Pe})+1}{\exp(2\mathrm{Pe})-1} - \frac{\varepsilon}{b^2} = \frac{h}{2b}\coth(\mathrm{Pe}) - \frac{\varepsilon}{b^2}. \end{aligned}$$

Moreover, in this case, the discrete solution not only is nodally exact, but also coincides with the exact solution everywhere in Ω (cf. Brezzi and Russo (1994)). Therefore, the functions b_j are called the *residual-free bubble functions*. In Russo (2006) one can find an extended comparison of the SUPG method and the residual-free bubbles method.

Remark 2.1.1. Since there holds

$$\tau_j^{B_2} = \frac{h}{2b}\left(\coth(\mathrm{Pe}) - \frac{1}{\mathrm{Pe}}\right) = \frac{h}{2b}\left(\frac{\mathrm{Pe}}{3} + \mathcal{O}\big(\mathrm{Pe}^3\big)\right) = \frac{h^2}{12\varepsilon}\Big(1 + \mathcal{O}\big(\mathrm{Pe}^2\big)\Big), \tag{2.34}$$

the stabilization parameter $\tau_j^{B_1}$ (defined in (2.31)) is optimal for $\mathrm{Pe} \to 0$.

2.1.6 Local Green's function method

An alternative method that provides the nodally exact solution for constant data can be derived by constructing the local Green's function of the adjoint operator of L (see Marchuk (1982)). Let us therefore introduce the formal adjoint operator L^* of $Lw = -\varepsilon w'' + bw'$

$$L^* w = -\varepsilon w'' - (bw)'. \tag{2.35}$$

Thus, if $v, w \in H_0^1(\Omega) \cap H^2(\Omega)$, the following identity holds

$$\int_\Omega (Lv) w \,\mathrm{d}x = \int_\Omega v(L^* w)\,\mathrm{d}x. \tag{2.36}$$

The local Green's functions g_j, $j = 1, 2, \ldots, N-1$, of the operator L^* with respect to the nodes x_j are then defined by the identities

$$\begin{aligned} L^* g_j &= 0 \quad \text{in } I_j \cup I_{j+1} && (2.37)\\ g_j(x_{j-1}) = g_j(x_{j+1}) &= 0 \qquad \text{and} && (2.38)\\ \varepsilon\big[g_j'(x_j^-) - g_j'(x_j^+)\big] &= 1. && (2.39) \end{aligned}$$

If we now multiply the identity $Lu = f$ by the local Green's function g_j corresponding to the node x_j, $j \in \{1, 2, \ldots, N-1\}$, and integrate, we obtain the equation

$$\int_{x_{j-1}}^{x_{j+1}} (Lu) g_j \,\mathrm{d}x \;=\; \int_{x_{j-1}}^{x_{j+1}} f g_j \,\mathrm{d}x. \tag{2.40}$$

Using the integration by parts the left-hand side of (2.40) can now be rewritten in the form

$$\begin{aligned}
\int_{x_{j-1}}^{x_{j+1}} (Lu) g_j \,\mathrm{d}x \;&=\; \int_{x_{j-1}}^{x_j} (-\varepsilon u'' + bu') g_j \,\mathrm{d}x + \int_{x_j}^{x_{j+1}} (-\varepsilon u'' + bu') g_j \,\mathrm{d}x \;= \\
&= \;\Big[-\varepsilon u' g_j + b u g_j \Big]_{x_{j-1}}^{x_j} - \int_{x_{j-1}}^{x_j} -\varepsilon u' g_j' + u (b g_j)' \,\mathrm{d}x + \\
&\qquad + \Big[-\varepsilon u' g_j + b u g_j \Big]_{x_j}^{x_{j+1}} - \int_{x_j}^{x_{j+1}} -\varepsilon u' g_j' + u (b g_j)' \,\mathrm{d}x.
\end{aligned}$$

The property (2.38) together with the continuity of $-\varepsilon u' g_j + b u g_j$ at the node x_j then yields

$$\Big[-\varepsilon u' g_j + b u g_j \Big]_{x_{j-1}}^{x_j} + \Big[-\varepsilon u' g_j + b u g_j \Big]_{x_j}^{x_{j+1}} \;=\; 0. \tag{2.41}$$

Thus, we obtain

$$\begin{aligned}
\int_{x_{j-1}}^{x_{j+1}} (Lu) g_j \,\mathrm{d}x \;&=\; \int_{x_{j-1}}^{x_j} \varepsilon u' g_j' - u (b g_j)' \,\mathrm{d}x + \int_{x_j}^{x_{j+1}} \varepsilon u' g_j' - u (b g_j)' \,\mathrm{d}x \;= \\
&= \;\Big[\varepsilon u g_j' \Big]_{x_{j-1}}^{x_j} - \int_{x_{j-1}}^{x_j} \varepsilon u g_j'' \,\mathrm{d}x - \int_{x_{j-1}}^{x_j} u (b g_j)' \,\mathrm{d}x + \\
&\quad + \Big[\varepsilon u g_j' \Big]_{x_j}^{x_{j+1}} - \int_{x_j}^{x_{j+1}} \varepsilon u g_j'' \,\mathrm{d}x - \int_{x_j}^{x_{j+1}} u (b g_j)' \,\mathrm{d}x \;= \\
&= \;-\varepsilon g_j'(x_{j-1}) u(x_{j-1}) + u(x_j) + \varepsilon g_j'(x_{j+1}) u(x_{j+1}),
\end{aligned} \tag{2.42}$$

where we used the property (2.39) and the fact that $-\int_{x_{j-1}}^{x_{j+1}} u(\varepsilon g_j'' + (b g_j)') \,\mathrm{d}x = \int_{x_{j-1}}^{x_{j+1}} u(L^* g_j) \,\mathrm{d}x = 0$ by (2.37).

Consequently, the equation (2.40) changes to

$$-\varepsilon g_j'(x_{j-1}) u_{j-1} + u_j + \varepsilon g_j'(x_{j+1}) u_{j+1} \;=\; \int_{x_{j-1}}^{x_{j+1}} f g_j \,\mathrm{d}x, \tag{2.43}$$

which is a finite difference scheme producing the nodally exact solution for all sufficiently smooth data. Indeed, for constant data we can solve the ordinary differential equation (2.37)–(2.39) and compute the exact form of the local Green's function g_j

$$g_j(x) \;=\; \begin{cases}
\dfrac{1}{b} \dfrac{1 - \exp\left(-\frac{b}{\varepsilon}(x - x_{j-1})\right)}{1 + \exp\left(-2\mathrm{Pe}\right)} & \text{for } x \in [x_{j-1}, x_j], \\[2ex]
\dfrac{1}{b} \dfrac{\exp\left(\frac{b}{\varepsilon}(x_{j+1} - x)\right) - 1}{\exp\left(2\mathrm{Pe}\right) + 1} & \text{for } x \in [x_j, x_{j+1}], \\[2ex]
0 & \text{otherwise.}
\end{cases} \tag{2.44}$$

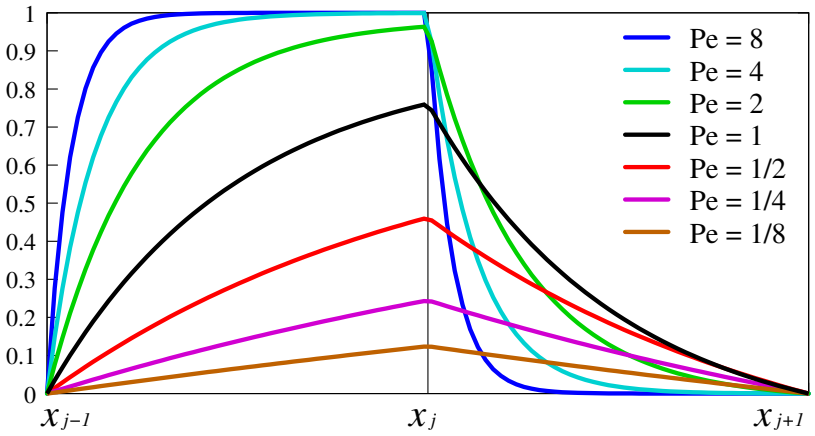

Figure 2.5: Local Green's functions for $b = 1$ and different choices of Pe.

Using this expression, we can evaluate all the terms containing g_j in the scheme (2.43) and obtain

$$-\frac{\mathrm{e}^{2\mathrm{Pe}}}{\mathrm{e}^{2\mathrm{Pe}}+1}\,u_{j-1} + u_j - \frac{1}{\mathrm{e}^{2\mathrm{Pe}}+1}\,u_{j+1} \;=\; \frac{h}{b}\,\frac{\mathrm{e}^{2\mathrm{Pe}}-1}{\mathrm{e}^{2\mathrm{Pe}}+1}\,f. \tag{2.45}$$

This can be rearranged to the equivalent form

$$-\frac{b}{h}\,\frac{\mathrm{e}^{2\mathrm{Pe}}}{\mathrm{e}^{2\mathrm{Pe}}-1}u_{j-1} + \frac{b}{h}\,\frac{\mathrm{e}^{2\mathrm{Pe}}+1}{\mathrm{e}^{2\mathrm{Pe}}-1}u_j - \frac{b}{h}\,\frac{1}{\mathrm{e}^{2\mathrm{Pe}}-1}u_{j+1} \;=\; f, \tag{2.46}$$

or to the form containing differences of u at the nodes x_{j-1}, x_j and x_{j+1}

$$\frac{b}{2h}\,\frac{\mathrm{e}^{2\mathrm{Pe}}+1}{\mathrm{e}^{2\mathrm{Pe}}-1}\,(-u_{j-1}+2u_j-u_{j+1}) + \frac{b}{2h}\,\frac{\mathrm{e}^{2\mathrm{Pe}}-1}{\mathrm{e}^{2\mathrm{Pe}}-1}\,(u_{j+1}-u_{j-1}) \;=\; f. \tag{2.47}$$

Since $\frac{\mathrm{e}^{2\mathrm{Pe}}+1}{\mathrm{e}^{2\mathrm{Pe}}-1} = \coth \mathrm{Pe}$, we find out that this is again the Il'in-Allen-Southwell scheme. For a multi-dimensional extension of this technique, see, e.g., Axelsson et al. (2009).

2.1.7 Exponentially fitted schemes

Since the behavior of the solution in the exponential layer is well known, one can also derive a method producing an oscillation-free discrete solution by requiring nodal exactness for functions from $\{1, x, \exp(bx/\varepsilon)\}$ (see Gartland (1987)). Thus, considering the equidistant partition of Ω we try to find the unknown coefficients in the scheme

$$p_i u_{i-1} + q_i u_i + r_i u_{i+1} \;=\; f_i \;=\; (Lu)_i. \tag{2.48}$$

These coefficients have to satisfy the equalities

$$\begin{array}{lcllcllcl}
p_i & + & q_i & + & r_i & = & (L(1))_i & = & 0, \\
p_i(x_i-h) & + & q_i x_i & + & r_i(x_i+h) & = & (L(x))_i & = & b, \\
p_i\left(\mathrm{e}^{\frac{bx_i}{\varepsilon}}\mathrm{e}^{-2\mathrm{Pe}}\right) & + & q_i\left(\mathrm{e}^{\frac{bx_i}{\varepsilon}}\right) & + & r_i\left(\mathrm{e}^{\frac{bx_i}{\varepsilon}}\mathrm{e}^{2\mathrm{Pe}}\right) & = & \left(L\left(\mathrm{e}^{\frac{bx}{\varepsilon}}\right)\right)_i & = & 0,
\end{array} \tag{2.49}$$

and can be rewritten in the form

$$\begin{array}{rcrcrcl}
p_i & + & q_i & + & r_i & = & 0, \\
-hp_i & + & 0 & + & hr_i & = & b, \\
\mathrm{e}^{-2\mathrm{Pe}}p_i & + & q_i & + & \mathrm{e}^{2\mathrm{Pe}}r_i & = & 0.
\end{array} \tag{2.50}$$

The solution of this system of linear equations takes the form

$$(p_i, q_i, r_i) \;=\; \left(-\frac{b}{h}\,\frac{\mathrm{e}^{2\mathrm{Pe}}}{\mathrm{e}^{2\mathrm{Pe}}-1}, \frac{b}{h}\,\frac{\mathrm{e}^{2\mathrm{Pe}}+1}{\mathrm{e}^{2\mathrm{Pe}}-1}, -\frac{b}{h}\,\frac{1}{\mathrm{e}^{2\mathrm{Pe}}-1}\right), \tag{2.51}$$

which leads to the scheme (2.46). As we can see, for constant data the coefficients p_i, q_i, r_i do not depend on i. In the non-constant case, we have to change b for b_i. Let us also mention that for $\mathrm{Pe} \to \infty$ we obtain $(p, q, r) \approx \frac{b}{h}(-1, 1, 0)$, which is the backward Euler method, whereas for $\mathrm{Pe} \to 0$ we obtain $(p, q, r) \approx \frac{\varepsilon}{h^2}(-1, 2, -1) + \frac{b}{2h}(-1, 0, 1)$, which is the central difference scheme.

The main advantage of this approach is that one can easily adjust it for derivation of schemes of higher order. For instance, considering the equidistant partition with mesh parameter h, a five-point scheme of order h^2 is constructed by requiring nodal exactness for functions from $\{1, x, x^2, \exp(bx/\varepsilon), x\exp(bx/\varepsilon)\}$. This results into the following quintuple of coefficients

$$\begin{aligned}
&(p, q, r, s, t) = \\
&= \quad \frac{b}{2h}\left(\tfrac{\mathrm{e}^{4\mathrm{Pe}}(\mathrm{e}^{2\mathrm{Pe}}+1)}{(\mathrm{e}^{2\mathrm{Pe}}-1)^3},\ \tfrac{-4\mathrm{e}^{2\mathrm{Pe}}(\mathrm{e}^{4\mathrm{Pe}}+1)}{(\mathrm{e}^{2\mathrm{Pe}}-1)^3},\ \tfrac{3(\mathrm{e}^{4\mathrm{Pe}}+1)\left(\mathrm{e}^{2\mathrm{Pe}}+1\right)}{(\mathrm{e}^{2\mathrm{Pe}}-1)^3},\ \tfrac{-4(\mathrm{e}^{4\mathrm{Pe}}+1)}{(\mathrm{e}^{2\mathrm{Pe}}-1)^3},\ \tfrac{\mathrm{e}^{2\mathrm{Pe}}+1}{(\mathrm{e}^{2\mathrm{Pe}}-1)^3}\right) + \\
&\quad + \frac{\varepsilon}{h^2}\left(\ \tfrac{-\mathrm{e}^{4\mathrm{Pe}}}{(\mathrm{e}^{2\mathrm{Pe}}-1)^2},\quad \tfrac{2\mathrm{e}^{2\mathrm{Pe}}(\mathrm{e}^{2\mathrm{Pe}}+1)}{(\mathrm{e}^{2\mathrm{Pe}}-1)^2},\quad -\tfrac{\mathrm{e}^{4\mathrm{Pe}}+4\mathrm{e}^{2\mathrm{Pe}}+1}{(\mathrm{e}^{2\mathrm{Pe}}-1)^2},\quad \tfrac{2(\mathrm{e}^{2\mathrm{Pe}}+1)}{(\mathrm{e}^{2\mathrm{Pe}}-1)^2},\quad \tfrac{-1}{(\mathrm{e}^{2\mathrm{Pe}}-1)^2}\right).
\end{aligned}$$

If we again compute the approximation for $\mathrm{Pe} \to \infty$, we obtain $(p, q, r, s, t) \approx \frac{b}{2h}(1, -4, 3, 0, 0) + \frac{\varepsilon}{h^2}(-1, 2, -1, 0, 0)$, which is the backward difference formula of the second order used on the convective term and the shifted central differences used on the diffusive term. Computing the limit for $\mathrm{Pe} \to 0$ yields $(p, q, r, s, t) \approx \frac{\varepsilon}{h^2}\left(\frac{1}{12}, -\frac{4}{3}, \frac{5}{2}, -\frac{4}{3}, \frac{1}{12}\right) + \frac{b}{h}\left(\frac{1}{12}, -\frac{2}{3}, 0, \frac{2}{3}, -\frac{1}{12}\right)$, which is the central difference formula for five-point stencil.

This technique can be also used for construction of schemes in higher dimensions (2D, 3D). For an arbitrary mesh in the vicinity of the outflow boundary, one adjusts the coefficients of the numerical method so that the created scheme is nodally exact for corresponding boundary layer function (cf. Section 3.6).

This list of stabilization methods clearly is not complete (see Roos (1994) for another interesting comparison). Other worth-mentioning methods are, e.g., the collocation methods (e.g. Surla and Stojanović (1988)), the local projection stabilization method (e.g. Matthies et al. (2007)), Galerkin least squares methods (e.g. Hughes et al. (1989)), the discontinuous Galerkin method (e.g. Riviére (2008) or Dolejší and Feistauer (2015)) or a suitable numerical quadrature (e.g. Hughes (1978) or Payre et al. (1982)).

2.2 Uniform convergence of classical Il'in-Allen-Southwell scheme

Prior to proving the uniform convergence result we demonstrate the difference in behavior of the simple upwind scheme and the Il'in-Allen-Southwell scheme on a simple nonconstant example ($\varepsilon = 10^{-6}$)

$$-\varepsilon u'' + u' \;=\; 2x \qquad \text{in } (0,1), \tag{2.52}$$

$$u(0) \;=\; u(1) \;=\; 0. \tag{2.53}$$

The exact solution of the problem (2.52)–(2.53), the solution obtained using the simple upwind scheme and the solution obtained by the Il'in-Allen-Southwell scheme have the following form

$$u(x) = x^2 + 2\varepsilon x - (1+2\varepsilon)\,\frac{\mathrm{e}^{x/\varepsilon}-1}{\mathrm{e}^{1/\varepsilon}-1}, \tag{2.54}$$

$$u_k^{upw} = (kh)^2 + (2\varepsilon+h)kh - (1+2\varepsilon+h)\,\frac{\left(1+\frac{h}{\varepsilon}\right)^k - 1}{\left(1+\frac{h}{\varepsilon}\right)^N - 1}, \tag{2.55}$$

$$u_k^{IAS} = (kh)^2 + kh^2\,\frac{\mathrm{e}^{h/\varepsilon}+1}{\mathrm{e}^{h/\varepsilon}-1} - \left(1 + h\,\frac{\mathrm{e}^{h/\varepsilon}+1}{\mathrm{e}^{h/\varepsilon}-1}\right)\frac{\mathrm{e}^{kh/\varepsilon}-1}{\mathrm{e}^{1/\varepsilon}-1}, \tag{2.56}$$

where we have used the equidistant partition of $(0,1) = (x_0, x_N)$ with a mesh step $h = 1/N$.

The error of both methods computed at the last five inner nodes laying in (the vicinity of) the exponential boundary layer is depicted in Figure 2.6. We observe that in contrast to the Il'in-Allen-Southwell scheme, the simple upwind scheme does not converge uniformly (with respect to ε), i.e. for fixed ε the error of the simple upwind scheme does not always decrease with decreasing h (increasing $N = 1/h$). In this case the error of the solution obtained by the simple upwind scheme at the node x_{N-j} for fixed j possesses local minimum for $N \approx \varepsilon^{-\frac{j}{j+1}}$, i.e. for $h \approx \varepsilon^{\frac{j}{j+1}}$.

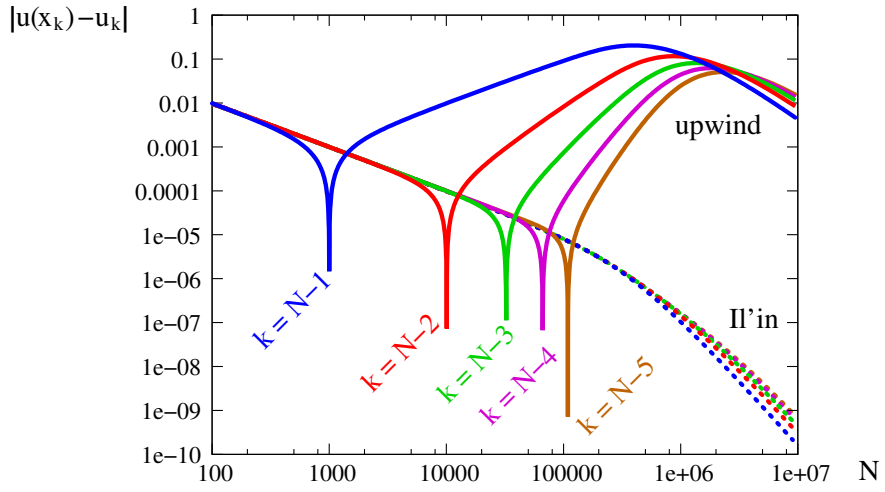

Figure 2.6: Comparison of convergence of the simple upwind scheme and the Il'in-Allen-Southwell scheme.

The proof of the uniform convergence of the Il'in-Allen-Southwell scheme can be found for example in Roos et al. (2008) or in more details in Kellogg and Tsan (1978). However, estimates resulting from these proofs contain unknown multiplicative constants which can in many cases make the estimates significantly worse. Thus, we derive all estimates with a concrete form of these constants.

We use the classical Il'in-Allen-Southwell scheme for solving model one-dimensional convection-diffusion equation (2.1)–(2.2), i.e.

$$-\varepsilon u'' + b(x)u' = f(x) \qquad \text{in } \Omega = (0,1), \tag{2.57}$$

$$u(0) = u(1) = 0. \tag{2.58}$$

Denoting $u_j = u_h(x_j)$, for $j = 0, 1, \ldots, N$, the stencil (and corresponding matrix L_h^*) generated by the Il'in-Allen-Southwell scheme has for all $j = 1, 2, \ldots, N-1$ form

$$(L_h^* u_h)_j = \varepsilon \mathrm{Pe}(x_j) \coth\left(\mathrm{Pe}(x_j)\right) \frac{-u_{j-1} + 2u_j - u_{j+1}}{h^2} + b(x_j) \frac{u_{j+1} - u_{j-1}}{2h} = f(x_j). \tag{2.59}$$

We divide the proof of the convergence of the Il'in-Allen-Southwell scheme into several lemmas. Firstly, we express the solution u as the sum of the layer function $v(x) = v_0\left(\frac{1-x}{\varepsilon}\right) = -u_0(1)\left(-b(1)\frac{1-x}{\varepsilon}\right)$ and the remainder part z. Then we prove both the consistency and stability of each part v and z.

The function v is the first term of the layer part E of the S-decomposition of the solution $u = S + E$ (cf. Definition 1.2.3, page 13) and thus

$$z \;=\; u - v \;=\; S + E - v. \tag{2.60}$$

Further, we define the splitting of the discrete solution $u_h = v_h + z_h$ into functions $v_h, z_h \in X_h$. They are solutions of the equations

$$L_h^* v_h \;=\; R_h(Lv), \qquad (v_h)_0 = v(0), (v_h)_N = v(1), \tag{2.61}$$
$$L_h^* z_h \;=\; R_h(Lz), \qquad (z_h)_0 = z(0), (z_h)_N = z(1), \tag{2.62}$$

where $R_h : \mathcal{C}(\overline{\Omega}) \to \mathbb{R}^{N+1}$ is the interpolation operator satisfying $[R_h v]_j = v(jh)$, $j = 0, 1, \ldots, N$.

2.2.1 Consistency

Before proving the consistency result for the function v, we prove one technical lemma.

Lemma 2.2.1. *The function* $\sinh(x)$ *satisfies following estimates*

$$\sinh(x) \;\geq\; \frac{x\mathrm{e}^x}{2(x+1)}, \quad \textit{for } x > 0, \tag{2.63}$$
$$|\sinh(x) - x| \;\leq\; \frac{|x|^3 \mathrm{e}^{|x|}}{6(1+x^2)}. \tag{2.64}$$

Proof. For the first inequality we use the estimate $\mathrm{e}^x \geq 1 + x$ which for $x > 0$ implies $\mathrm{e}^{-x} \leq \frac{1}{1+x} \leq \frac{\mathrm{e}^x}{1+x}$. Consequently

$$\sinh x \;=\; \frac{1}{2}(\mathrm{e}^x - \mathrm{e}^{-x}) \;\geq\; \frac{1}{2}\left(\mathrm{e}^x - \frac{\mathrm{e}^x}{1+x}\right) \;=\; \frac{x\mathrm{e}^x}{2(x+1)}. \tag{2.65}$$

Since both functions $|\sinh(x) - x|$ and $\frac{|x|^3\mathrm{e}^{|x|}}{6(1+x^2)}$ are even it suffices to prove the second inequality for $x \geq 0$. We firstly consider $x \geq 4$. Then

$$\frac{x^3\mathrm{e}^x}{6(1+x^2)} \;=\; \frac{x\mathrm{e}^x}{6}\,\frac{x^2}{1+x^2} \;\geq\; \frac{4\mathrm{e}^x}{6}\,\frac{16}{17} \;=\; \frac{32}{51}\,\mathrm{e}^x \;\geq\; \frac{1}{2}\,\mathrm{e}^x \;\geq\; \sinh(x) - x, \tag{2.66}$$

where we used the fact that the function $\frac{x^2}{1+x^2}$ is increasing on $(0, +\infty)$.

For the case $0 \le x < 4$ we use Taylor's polynomials. Since all derivatives of the function $\sinh x$ in 0 are nonnegative, the Taylor polynomial in 0 of the function $\sinh x$ has nonnegative coefficients. Thus, if we subtract several terms of this Taylor polynomial from $\sinh x$, we obtain a nondecreasing function (even if we divide it by the order of the resulting function). This idea leads us to the estimate

$$\frac{\sinh(x) - x - \frac{x^3}{6}}{x^5} \;\le\; \frac{\sinh(4) - 4 - \frac{4^3}{6}}{4^5} \;=\; C_\sigma \qquad \text{for } x \in [0,4]. \tag{2.67}$$

If we now take into account the estimate $\mathrm{e}^x \ge 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120}$, then

$$\sinh x - x \;\le\; \frac{x^3}{6} + C_\sigma x^5 \;\le\; \frac{x^3\left(1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120}\right)}{6(1+x^2)} \;\le\; \frac{x^3\mathrm{e}^x}{6(1+x^2)}. \tag{2.68}$$

The second inequality in (2.68) comes from the estimate

$$\begin{aligned} & x^3\left(1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120}\right) - 6\left(1+x^2\right)\left(\frac{x^3}{6} + C_\sigma x^5\right) \;= \\ &= \;\left[1 - \left(3C_\sigma + \frac{1}{4}\right)x\right]^2 x^4 + C_T x^6 + \left[30\left(6C_\sigma - \frac{1}{24}\right) - \frac{x}{2}\right]^2 \frac{x^6}{30} \;\ge\; 0, \end{aligned}$$

where $C_T = \frac{1}{6} - \left(3C_\sigma + \frac{1}{4}\right)^2 - 30\left(6C_\sigma - \frac{1}{24}\right)^2 = \frac{5}{96} + \frac{27}{2}C_\sigma - 1089C_\sigma^2 \;\approx\; 0.053.$ □

Now we use the obtained inequalities and prove the following lemma which provides the consistency of the method for the layer function v.

Lemma 2.2.2. *Let L_h^* be the matrix generated by the Il'in-Allen-Southwell scheme (2.59), let the function b be Lipschitz-continuous with a constant β_1 and let $h_0^* \in \left(0, \frac{b(1)-\beta}{\beta_1}\right)$ be a positive real number. Then there exists a positive constant C_z independent of h and ε such that for all $h < h_0^*$ there holds*

$$\left|R_h(Lv) - L_h^* R_h v\right| \;\le\; \frac{C_z\, h^2}{\varepsilon(\varepsilon + h)} R_h \exp\left(-\underline{\beta}\frac{|1-x|}{\varepsilon}\right). \tag{2.69}$$

Proof. At first we compute the exact form of both terms Lv and $L_h^* R_h v$.

$$(Lv)(x) \;=\; -\frac{b(1)}{\varepsilon}(b(1) - b(x))v(x), \tag{2.70}$$

$$(L_h^* R_h v)(x_j) \;=\; -\frac{2b(x_j)\sinh(\mathrm{Pe}(1))}{h\sinh(\mathrm{Pe}(x_j))}\Big(\sinh\big(\mathrm{Pe}(1) - \mathrm{Pe}(x_j)\big)\Big)v(x_j). \tag{2.71}$$

Further, let us denote $q = \mathrm{Pe}(1)$, $p = \mathrm{Pe}(x_j)$, $S(x) = \sinh x - x$ and $\psi(x) = \frac{x^2}{1+x^2}$. Then we observe that the function $\psi(x)$ is increasing for $x > 0$, whereas functions $S(x)$ and $\sinh(x)$ satisfy the estimates (2.63) and (2.64) from Lemma

2.2.1. We use these observations for estimating the consistency error

$$\begin{aligned}
&\Big|(Lv)(x_j) - (L_h^* R_h v)(x_j)\Big| \;= \\
&= \;\frac{4\varepsilon}{h^2}\left|-q(q-p) + p\frac{\sinh(q)}{\sinh(p)}\sinh(q-p)\right||v(x_j)| \;= \\
&= \;\frac{4\varepsilon}{h^2}\Big|p(q-p)S(q) - q(q-p)S(p) + pS(q)S(q-p) + pqS(q-p)\Big|\frac{|v(x_j)|}{\sinh(p)} \;\le \\
&\le \;\frac{4\varepsilon pq|q-p|}{6h^2\sinh(p)}\Big[\psi(q)\mathrm{e}^q + \psi(p)\mathrm{e}^p + \psi(|q-p|)\mathrm{e}^{|q-p|}\Big(1+\psi(q)\mathrm{e}^q\Big)\Big]|v(x_j)| \;\le \\
&\le \;\frac{2\varepsilon pq|q-p|}{3h^2}\,\frac{2(1+p)}{p\mathrm{e}^p}\,4\max_{s\in\{p,q\}}\{\psi(s)\}\,\mathrm{e}^{q+|q-p|}|v(x_j)|. \qquad (2.72)
\end{aligned}$$

Since the function b is Lipschitz-continuous with a constant β_1, we can estimate the difference $|q-p|$ by

$$|q-p| \;\le\; \frac{\beta_1 h}{2\varepsilon}\,|1-x_j|. \qquad (2.73)$$

Consequently, for the consistency error of the function v there holds

$$\begin{aligned}
&\Big|(Lv)(x_j) - (L_h^* R_h v)(x_j)\Big| \;\le \\
&\le \;\frac{2\varepsilon h^2\|b\|_\infty\beta_1|1-x_j|\;4(2\varepsilon+\|b\|_\infty h)\;\|b\|_\infty^2 h^2}{3h^2(2\varepsilon)^2\;\varepsilon\;(4\varepsilon^2+\|b\|_\infty^2 h^2)}\;\exp\left(\frac{h}{\varepsilon}\beta_1|1-x_j|\right)|v(x_j)| \;\le \\
&\le \;\frac{4\beta_1\|b\|_\infty^3 h^2|1-x_j|}{3\varepsilon^2(2\varepsilon+\|b\|_\infty h)}\;\exp\left(\frac{h}{\varepsilon}\beta_1|1-x_j|\right)|v(x_j)|. \qquad (2.74)
\end{aligned}$$

Now we rewrite the last two factors in the form

$$|u_0(1)|\exp\left(\Big(h\beta_1 - (b(1)-\underline{\beta})\Big)\frac{|1-x_j|}{\varepsilon}\right)\exp\left(-\underline{\beta}\frac{|1-x_j|}{\varepsilon}\right). \qquad (2.75)$$

If $h < h_0^* < \frac{b(1)-\underline{\beta}}{\beta_1}$ then using $\exp(-x) < \frac{1}{x}$ (for $x>0$) we have

$$\exp\left(\Big(h\beta_1 - (b(1)-\underline{\beta})\Big)\frac{|1-x_j|}{\varepsilon}\right) \;<\; \frac{\varepsilon}{|1-x_j|\Big(b(1)-\underline{\beta}-h_0^*\beta_1\Big)}. \qquad (2.76)$$

Consequently for the consistency error there holds

$$\Big|R_h(Lv) - L_h^* R_h v\Big| \;\le\; \frac{C_z\;h^2}{\varepsilon(\varepsilon+h)}R_h\exp\left(-\underline{\beta}\frac{|1-x|}{\varepsilon}\right), \qquad (2.77)$$

where $C_z = \frac{4\beta_1\|b\|_\infty^3|z(1)|}{3\min\{2,\|b\|_\infty\}}\,\frac{1}{b(1)-\underline{\beta}-h_0^*\beta_1}$. Here we used the fact that $u_0(1) = -v(1) = z(1) - u(1) = z(1)$. □

Let us notice that the quality of the above derived estimate depends on the bound $\underline{\beta}$. If it is too small, then the exponential decay is very slow. On the other hand, if $\underline{\beta}$ is close to $b(1)$, the constant C_z goes to infinity.

The consistency corresponding to the smooth part z is given by the next lemma.

Lemma 2.2.3. *Let L_h^* be the matrix generated by the Il'in-Allen-Southwell scheme (2.59), then for the consistency error corresponding to the function z it holds*

$$\begin{aligned}
&|(Lz)(x_j) - (L_h^* R_h z)_j| \ \leq \\
&\quad \leq \ C_S\,(\varepsilon + 2\|b\|_\infty)\,h + \frac{C_E}{\underline{\beta}}\,(1 + 2\|b\|_\infty)\sinh\left(\frac{\underline{\beta} h}{\varepsilon}\right)\exp\left(-\frac{\underline{\beta}}{\varepsilon}(1 - x_j)\right). \qquad (2.78)
\end{aligned}$$

Proof. At the beginning we express the consistency error as the sum of three terms and estimate them using Taylor polynomial with the remainder in the integral form.

$$\begin{aligned}
&|(Lz)(x_j) - (L_h^* R_h z)_j| \ = \\
&= \ \left|\varepsilon\big[D_h^2 z(x_j) - z''(x_j)\big] + \frac{b(x_j)h}{2}\xi(\mathrm{Pe}(x_j)) D_h^2 z(x_j) + b(x_j)\big[D_h^1 z(x_j) - z'(x_j)\big]\right| \\
&\leq \ \frac{1}{2}\int_{x-h}^{x+h} \varepsilon|z^{(3)}| + \|b\|_\infty|z^{(2)}| + \|b\|_\infty|z^{(2)}|\,\mathrm{d}t, \qquad (2.79)
\end{aligned}$$

where we used the expressions

$$\begin{aligned}
D_h^2 z(x) \ &= \ z''(x) + \frac{1}{2h^2}\left[\int_x^{x+h}(x + h - t)^2 z^{(3)}(t)\,\mathrm{d}t - \int_{x-h}^{x}(x - h - t)^2 z^{(3)}(t)\,\mathrm{d}t\right], \\
D_h^2 z(x) \ &= \ \frac{1}{h^2}\int_x^{x+h}\int_{t-h}^{t} u''(s)\,\mathrm{d}s\,\mathrm{d}t, \\
D_h^1 z(x) \ &= \ z'(x) + \frac{1}{2h}\left[\int_x^{x+h}(x + h - t) z^{(2)}(t)\,\mathrm{d}t + \int_{x-h}^{x}(x - h - t) z^{(2)}(t)\,\mathrm{d}t\right]
\end{aligned}$$

resulting in estimates

$$\begin{aligned}
|(D_h^2 z(x_j) - z''(x_j)| \ &\leq \ \frac{1}{2h^2}\left[h^2\int_x^{x+h}|z^{(3)}|\,\mathrm{d}t + h^2\int_{x-h}^{x}|z^{(3)}|\,\mathrm{d}t\right] = \frac{1}{2}\int_{x-h}^{x+h}|z^{(3)}|\,\mathrm{d}t, \\
|D_h^2 z(x)| \ &\leq \ \frac{1}{h^2}\,h\max_{t\in[x,x+h]}\left|\int_{t-h}^{t} z''(s)\,\mathrm{d}s\right| \ \leq \ \frac{1}{h}\int_{x-h}^{x+h}|z''(t)|\,\mathrm{d}t, \\
|(D_h^1 z(x_j) - z'(x_j)| \ &\leq \ \frac{1}{2h}\left[h\int_x^{x+h}|z^{(2)}|\,\mathrm{d}t + h\int_{x-h}^{x}|z^{(2)}|\,\mathrm{d}t\right] = \frac{1}{2}\int_{x-h}^{x+h}|z^{(2)}|\,\mathrm{d}t.
\end{aligned}$$

Since $z = S + (E - v)$, we can use Lemma 1.2.1 (page 14) and estimate the derivatives of z by

$$|z^{(j)}(t)| \leq C_S + C_E\varepsilon^{1-j}\exp\left(-\frac{\underline{\beta}}{\varepsilon}(1 - t)\right), \qquad (2.80)$$

where the constants C_S and C_E are independent from ε and h. This estimate contains a factor ε^{1-j} instead of ε^j which is caused by the fact that the layer part of the S-decomposition of the function z begins with $0.\varepsilon^0$. Consequently

$$\begin{aligned}
&|(Lz)(x_j) - (L_h^* R_h z)_j| \ \leq \\
&\quad \leq \ C_S\,(\varepsilon + 2\|b\|_\infty)\,h + C_E\left(\frac{1}{2} + \|b\|_\infty\right)\frac{1}{\varepsilon}\int_{x_j-h}^{x_j+h}\exp\left(-\frac{\underline{\beta}}{\varepsilon}(1 - t)\right)\mathrm{d}t \ = \\
&\quad = \ C_S\,(\varepsilon + 2\|b\|_\infty)\,h + \frac{C_E}{\underline{\beta}}\,(1 + 2\|b\|_\infty)\sinh\left(\frac{\underline{\beta} h}{\varepsilon}\right)\exp\left(-\frac{\underline{\beta}}{\varepsilon}(1 - x_j)\right).
\end{aligned}$$

□

2.2.2 Stability

Both consistency estimates contain exponential functions. This means that a stability result for exponential functions is necessary.

Lemma 2.2.4. *Let $b(x) > \underline{\beta} > 0$. Then for any function $\exp\left(\frac{\alpha x}{\varepsilon}\right)$, $0 < \alpha < \underline{\beta}$, there exists positive constant C_α (independent of h and ε) such that*

$$L_h^* R_h \exp\left(\frac{\alpha x}{\varepsilon}\right) \;\geq\; \frac{C_\alpha}{\max\{h,\varepsilon\}} R_h \exp\left(\frac{\alpha x}{\varepsilon}\right). \tag{2.81}$$

Proof. At first, exact computation gives

$$L_h^* R_h \exp\left(\frac{\alpha x}{\varepsilon}\right) \;=\; 2R_h \left\{ \exp\left(\frac{\alpha x}{\varepsilon}\right) \frac{b(x)}{h} \frac{\sinh\left(\frac{\alpha h}{2\varepsilon}\right)}{\sinh\left(\frac{b(x)h}{2\varepsilon}\right)} \sinh\left(\frac{\big(b(x)-\alpha\big)h}{2\varepsilon}\right) \right\}. \tag{2.82}$$

Now we distinguish two situations. In the case when $h \leq \varepsilon$, then for all $\kappa \geq 0$ it holds $\kappa\frac{h}{\varepsilon} \leq \sinh\left(\kappa\frac{h}{\varepsilon}\right) \leq \sinh(\kappa)\frac{h}{\varepsilon}$. Consequently we have

$$\begin{aligned} L_h^* R_h \exp\left(\frac{\alpha x}{\varepsilon}\right) \;&\geq\; 2R_h \left\{ \exp\left(\frac{\alpha x}{\varepsilon}\right) \frac{\underline{\beta}}{h} \frac{\frac{\alpha h}{2\varepsilon}}{\sinh\left(\frac{b(x)}{2}\right)\frac{h}{\varepsilon}} \frac{\big(b(x)-\alpha\big)h}{2\varepsilon} \right\} \;\geq \\ &\geq\; \frac{\underline{\beta}\alpha(\underline{\beta}-\alpha)}{2\varepsilon \sinh\left(\frac{1}{2}\|b\|_\infty\right)} R_h \exp\left(\frac{\alpha x}{\varepsilon}\right). \end{aligned} \tag{2.83}$$

On the other hand, if $h \geq \varepsilon$, then $\frac{\sinh\kappa}{\exp\kappa}\exp\left(\kappa\frac{h}{\varepsilon}\right) \leq \sinh\left(\kappa\frac{h}{\varepsilon}\right) \leq \frac{1}{2}\exp\left(\kappa\frac{h}{\varepsilon}\right)$ for all $\kappa \geq 0$ and there holds

$$\begin{aligned} &L_h^* R_h \exp\left(\frac{\alpha x}{\varepsilon}\right) \;\geq \\ &\geq\; 2R_h \left\{ \exp\left(\frac{\alpha x}{\varepsilon}\right) \frac{\underline{\beta}}{h} \frac{\frac{\sinh\left(\frac{\alpha}{2}\right)}{\exp\left(\frac{\alpha}{2}\right)} \exp\left(\frac{\alpha h}{2\varepsilon}\right)}{\frac{1}{2}\exp\left(\frac{b(x)h}{2\varepsilon}\right)} \frac{\sinh\left(\frac{b(x)-\alpha}{2}\right)}{\exp\left(\frac{b(x)-\alpha}{2}\right)} \exp\left(\frac{\big(b(x)-\alpha\big)h}{2\varepsilon}\right) \right\} \;\geq \\ &\geq\; \frac{4\underline{\beta}\sinh\left(\frac{\alpha}{2}\right)\sinh\left(\frac{\underline{\beta}-\alpha}{2}\right)}{h\exp\left(\frac{1}{2}\|b\|_\infty\right)} R_h \exp\left(\frac{\alpha x}{\varepsilon}\right). \end{aligned} \tag{2.84}$$

Thus the constant C_α is given by

$$C_\alpha \;=\; \min\left\{ \frac{\underline{\beta}\alpha(\underline{\beta}-\alpha)}{2\sinh\left(\frac{1}{2}\|b\|_\infty\right)}, \frac{4\underline{\beta}\sinh\left(\frac{\alpha}{2}\right)\sinh\left(\frac{\underline{\beta}-\alpha}{2}\right)}{\exp\left(\frac{1}{2}\|b\|_\infty\right)} \right\}. \tag{2.85}$$

□

Remark 2.2.1. The constant C_α from Lemma 2.2.4 vanishes for $\alpha = \underline{\beta}$ and thus the stability of the method is lost. In fact, the difference $\underline{\beta} - \alpha$ that occurs in the definition of the constant C_α is an estimate for the difference $b(x) - \alpha$, and thus if $b(x) > \underline{\beta}$, the stability preserves.

The last lemma results from the so-called M-criterion (Theorem 4.1.5, page 126) and provides an estimate on the norm of the matrix $(L_h^*)^{-1}$.

Lemma 2.2.5. *For the inverse matrix corresponding to the Il'in-Allen-Southwell scheme it holds* $\|(L_h^*)^{-1}\|_{\infty,d} \leq 1/\underline{\beta}$.

Proof. Since we would like to employ the M-criterion, we firstly examine the sign of the entries of the tridiagonal matrix L_h^*.

$$(L_h^*)_{j,j} = \frac{b(x_j)}{h}\coth(\mathrm{Pe}(x_j)) > 0, \tag{2.86}$$

$$(L_h^*)_{j,j+1} = \frac{b(x_j)}{2h}\Big(1-\coth(\mathrm{Pe}(x_j))\Big) < 0, \tag{2.87}$$

$$(L_h^*)_{j,j-1} = -\frac{b(x_j)}{2h}\Big(1+\coth(\mathrm{Pe}(x_j))\Big) < 0. \tag{2.88}$$

Secondly, the vector $e_h = R_h x$ is positive inside Ω and satisfies $L_h^* e_h = R_h b(x) > 0$. Thus the matrix L_h^* is an M-matrix and it holds

$$\|(L_h^*)^{-1}\|_{\infty,d} \leq \frac{\|e_h\|_{\infty,d}}{\min_k (L_h^* e_h)_k} \leq \frac{1}{\underline{\beta}}. \tag{2.89}$$

□

2.2.3 Convergence

Now we have all important ingredients for proving the uniform convergence of the classical Il'in-Allen-Southwell scheme.

Theorem 2.2.1. *There exists a positive constant* $\widehat{C}$ *(independent of* h *and* ε*) such that for the discrete solution* u_h^* *of the problem (2.1)–(2.2) obtained by the Il'in-Allen-Southwell scheme (2.59) there holds*

$$\|R_h u - u_h^*\|_{\infty,d} \leq \widehat{C} h. \tag{2.90}$$

Proof. The proof is standard - we use consistency and stability for proving the convergence. At first we decompose the consistency error

$$\begin{aligned} |L_h^*(R_h u - u_h^*)| &= |L_h^* R_h u - L_h^* u_h^*| = |L_h^* R_h (z+v) - L_h^*(z_h+v_h)| \leq \\ &\leq |L_h^* R_h z - L_h^* z_h| + |L_h^* R_h v - L_h^* v_h|. \end{aligned} \tag{2.91}$$

Then we choose arbitrary $\alpha \in (0, \underline{\beta})$ and use Lemmas 2.2.2 and 2.2.4 for estimation of the consistency and stability of the function v

$$\begin{aligned} L_h^*(R_h v - v_h) &= L_h^* R_h v - R_h(Lv) \leq \frac{C_z\, h^2}{\varepsilon(\varepsilon+h)} R_h \exp\left(-\underline{\beta}\frac{|1-x|}{\varepsilon}\right) \leq \\ &\leq \frac{C_z\, h^2}{\varepsilon(\varepsilon+h)} R_h \exp\left(-\alpha\frac{(1-x)}{\varepsilon}\right) = \frac{C_z\, h^2}{\varepsilon(\varepsilon+h)} \exp\left(-\frac{\alpha}{\varepsilon}\right) R_h \exp\left(\alpha\frac{x}{\varepsilon}\right) \leq \\ &\leq \frac{\max\{h,\varepsilon\}}{C_\alpha}\frac{C_z\, h^2}{\varepsilon(\varepsilon+h)} \exp\left(-\frac{\alpha}{\varepsilon}\right) L_h^* R_h \exp\left(\alpha\frac{x}{\varepsilon}\right). \end{aligned} \tag{2.92}$$

Since L_h^* is an M-matrix (cf. Lemma 2.2.5), it is inverse-monotone and thus it satisfies the discrete comparison principle (Theorem 4.1.7, page 126). This implies that

$$|R_h v - v_h| \leq \frac{\max\{h,\varepsilon\}}{C_\alpha}\frac{C_z\, h^2}{\varepsilon(\varepsilon+h)} R_h \exp\left(-\alpha\frac{1-x}{\varepsilon}\right). \tag{2.93}$$

Now we distinguish two situations. If $h \leq \varepsilon$ then

$$\|v_h - R_h v\|_{\infty,d} \;\leq\; \frac{C_z\, h^2}{C_\alpha(\varepsilon + h)} \;\leq\; \frac{C_z}{2C_\alpha}\, h. \tag{2.94}$$

In the case $h \geq \varepsilon$ we use the inequality $\exp(-x) \leq x^{-1}$ which holds for all positive x and estimate

$$\|v_h - R_h v\|_{\infty,d} \;\leq\; \frac{C_z\, h^2}{C_\alpha(\varepsilon + h)} \frac{h}{\varepsilon} \exp\left(-\alpha \frac{h}{\varepsilon}\right) \;\leq\; \frac{C_z\, h}{C_\alpha} \frac{h}{\varepsilon + h} \frac{1}{\alpha} \;\leq\; \frac{C_z\, h}{\alpha\, C_\alpha}. \tag{2.95}$$

Similarly, we use Lemmas 2.2.3 and 2.2.4 for proving the consistency and stability of the function z. Let us firstly consider $h \leq \varepsilon$, then

$$\begin{aligned}
L_h^*(R_h z - z_h) \;&=\; L_h^* R_h z - R_h(Lz) \;\leq \\
&\leq\; C_S\,(\varepsilon + 2\|b\|_\infty)\, h + \frac{C_E}{\underline{\beta}}\,(1 + 2\|b\|_\infty) \sinh\left(\frac{\underline{\beta} h}{\varepsilon}\right) R_h \exp\left(-\frac{\underline{\beta}}{\varepsilon}(1 - x)\right) \;\leq \\
&\leq\; C_S\,(\varepsilon + 2\|b\|_\infty)\, h L_h^* (L_h^*)^{-1} R_h 1 \;+ \\
&\quad + \frac{\max\{h, \varepsilon\}}{C_\alpha} \frac{C_E}{\underline{\beta}}\,(1 + 2\|b\|_\infty) \sinh\left(\frac{\underline{\beta} h}{\varepsilon}\right) \exp\left(-\frac{\alpha}{\varepsilon}\right) L_h^* R_h \exp\left(\alpha \frac{x}{\varepsilon}\right).
\end{aligned} \tag{2.96}$$

Again, applying the discrete comparison principle gives

$$\begin{aligned}
|R_h z - z_h| \;&\leq\; C_S\,(\varepsilon + 2\|b\|_\infty)\, h (L_h^*)^{-1} R_h 1 \;+ \\
&\qquad + \frac{\max\{h, \varepsilon\}}{C_\alpha} \frac{C_E}{\underline{\beta}}\,(1 + 2\|b\|_\infty) \sinh\left(\frac{\underline{\beta} h}{\varepsilon}\right) R_h \exp\left(-\alpha \frac{1 - x}{\varepsilon}\right).
\end{aligned}$$

Since $h \leq \varepsilon$ we can use the inequality $\sinh\left(\frac{\underline{\beta} h}{\varepsilon}\right) \leq \frac{h}{\varepsilon} \sinh(\underline{\beta})$ and find out that

$$\|R_h z - z_h\|_{\infty,d} \;\leq\; \left\{ C_S\,(\varepsilon + 2\|b\|_\infty)\, \frac{1}{\underline{\beta}} + \frac{C_E}{C_\alpha \underline{\beta}}\,(1 + 2\|b\|_\infty) \sinh(\underline{\beta}) \right\} h. \tag{2.97}$$

Conversely, when $h \geq \varepsilon$ we apply the inequality (remind that $\alpha \in (0, \underline{\beta})$)

$$\begin{aligned}
&\sinh\left(\frac{\underline{\beta} h}{\varepsilon}\right) \exp\left(-\frac{\underline{\beta}}{\varepsilon}(1 - x_j)\right) \;\leq\; \frac{1}{2} \exp\left(\frac{\underline{\beta} h}{\varepsilon}\right) \exp\left(-\frac{\underline{\beta}}{\varepsilon}(1 - x_j)\right) \;= \\
&=\; \frac{1}{2} \exp\left(-\frac{\underline{\beta}}{\varepsilon}(1 - x_{j+1})\right) \;\leq\; \frac{1}{2} \exp\left(-\frac{\alpha}{\varepsilon}(1 - x_{j+1})\right) \;= \\
&=\; \frac{1}{2} \exp\left(\frac{\alpha h}{\varepsilon}\right) \exp\left(-\frac{\alpha}{\varepsilon}(1 - x_j)\right).
\end{aligned} \tag{2.98}$$

Lemmas 2.2.3 and 2.2.4 together with this inequality then provide the estimate

$$\begin{aligned}
L_h^*(R_h z - z_h) \;&=\; L_h^* R_h z - R_h(Lz) \;\leq \\
&\leq\; C_S\,(\varepsilon + 2\|b\|_\infty)\, h L_h^* (L_h^*)^{-1} R_h 1 \;+ \\
&\quad + \frac{\max\{h, \varepsilon\}}{C_\alpha} \frac{C_E}{\underline{\beta}} \left(\frac{1}{2} + \|b\|_\infty\right) \exp\left(\frac{\alpha h}{\varepsilon}\right) \exp\left(-\frac{\alpha}{\varepsilon}\right) L_h^* R_h \exp\left(\alpha \frac{x}{\varepsilon}\right).
\end{aligned} \tag{2.99}$$

Let us emphasize that instead of the factor $\sinh\left(\underline{\beta}h/\varepsilon\right)$ this estimate contains $\exp\left(\alpha h/\varepsilon\right)$ which is important for estimation in the last layer node. This is the reason, why we cannot simply use the estimate (2.96).

After applying the discrete comparison principle we obtain

$$\begin{aligned} |R_h z - z_h| \;\leq\; & C_S\,(\varepsilon + 2\|b\|_\infty)\,h(L_h^*)^{-1}R_h 1 \;+ \\ & + \frac{\max\{h,\varepsilon\}}{C_\alpha}\,\frac{C_E}{\underline{\beta}}\left(\frac{1}{2} + \|b\|_\infty\right)\exp\left(\frac{\alpha h}{\varepsilon}\right)R_h \exp\left(-\alpha\frac{1-x}{\varepsilon}\right), \end{aligned}$$

which means that for the error corresponding to the smooth part z of the solution u in the case when $h \geq \varepsilon$ there holds

$$\|R_h z - z_h\|_{\infty,d} \;\leq\; \left\{C_S\,(\varepsilon + 2\|b\|_\infty)\,\frac{1}{\underline{\beta}} + \frac{C_E}{C_\alpha \underline{\beta}}\left(\frac{1}{2} + \|b\|_\infty\right)\right\} h. \tag{2.100}$$

If we combine all previous estimates we get

$$\|R_h u - u_h\|_{\infty,d} \;\leq\; \|R_h z - z_h\|_{\infty,d} + \|R_h v - v_h\|_{\infty,d} \;\leq\; \widehat{C}h, \tag{2.101}$$

where (we can take e.g. $\alpha = \underline{\beta}/2$)

$$\widehat{C} \;=\; \frac{C_z}{C_\alpha}\max\left\{\frac{1}{2},\,\frac{1}{\alpha}\right\} \;+\; \frac{3C_S\|b\|_\infty}{\underline{\beta}} \;+\; \frac{C_E\,(1+2\|b\|_\infty)}{C_\alpha\underline{\beta}}\max\left\{\frac{1}{2}, \sinh\left(\underline{\beta}\right)\right\}. \tag{2.102}$$

□

3. Modified SUPG method on convection-oriented meshes

3.1 Introduction and the idea of the method

Let us solve the convection-diffusion equation

$$-\varepsilon\Delta u(x) + \boldsymbol{b}(x)\cdot\nabla u(x) = f(x) \qquad \text{in } \Omega\subset\mathbb{R}^n, \tag{3.1}$$

$$u(x) = 0 \qquad \text{on } \partial\Omega, \tag{3.2}$$

where $n\in\mathbb{N}$, Ω is a polytopic domain with Lipschitz-continuous boundary $\partial\Omega$, $\boldsymbol{b}\in W^{1,\infty}(\Omega)^n$ is a convective vector field, $f\in L^2(\Omega)$ is a given outer source and $\varepsilon>0$ is the constant diffusivity. Further, we divide the boundary $\partial\Omega$ into three subsets

$$\Gamma_+ = \{x\in\partial\Omega,\, \boldsymbol{b}(x)\cdot\boldsymbol{n}(x)>0\}, \tag{3.3}$$

$$\Gamma_0 = \{x\in\partial\Omega,\, \boldsymbol{b}(x)\cdot\boldsymbol{n}(x)=0\}, \tag{3.4}$$

$$\Gamma_- = \{x\in\partial\Omega,\, \boldsymbol{b}(x)\cdot\boldsymbol{n}(x)<0\}, \tag{3.5}$$

satisfying $\overline{\partial\Omega}=\overline{\Gamma}_+\cup\overline{\Gamma}_0\cup\overline{\Gamma}_-$ and $\Gamma_+\cap\Gamma_0=\Gamma_0\cap\Gamma_-=\Gamma_-\cap\Gamma_+=\emptyset$. Here, the vector $\boldsymbol{n}(x)$ denotes a unit outer normal to the boundary $\partial\Omega$.

As $\varepsilon\to 0$, the equation (3.1) becomes singularly perturbed and near the boundary Γ_+ the finite element solution often contains spurious oscillations. We call this region exponential boundary layer. In order to diminish the oscillations at the exponential boundary layers, one may use the SUPG method (cf. Brooks and Hughes (1982)). However, the SUPG method does not diminish all the oscillations, in particular, at the parabolic (characteristic) boundary layers. These regions usually appear near the boundary Γ_0, but also along interior layers that propagate from discontinuous boundary conditions at Γ_- .

Apart from the SUPG method, one can also use the method of Mizukami and Hughes (1985). Unlike the SUPG method, the Mizukami-Hughes method satisfies the discrete maximum principle and therefore it removes all spurious oscillations at the layers. The drawback of the Mizukami-Hughes method is its nonlinearity and the absence of an error analysis. In order to eliminate this drawback we construct a special mesh, which is well-aligned with the vector field $\boldsymbol{b}$. The created linear method then enjoys both positive properties of the Mizukami-Hughes method and the SUPG method - it satisfies the discrete maximum principle and we can apply an error analysis analogous to the SUPG method.

Since ε is considered to be very small, the exact solution at any point $x\in\Omega$ away from layers in fact depends only on the values of u in the direction $-\boldsymbol{b}(x)$. It means that the discretization of the convective term should use only the upwind values. To achieve this, we construct a special mesh $\mathcal{T}_h$. Each element of such a mesh should have one of its edges oriented in the direction of the vector $\boldsymbol{b}$. Then, if $\boldsymbol{b}_K$ is a constant approximation of $\boldsymbol{b}$ on the element $K\in\mathcal{T}_h$ parallel to one of its edges and if we use simplicial finite elements with linear basis functions $\{\lambda_{K,i}\}_{i=1}^{n+1}$, only *two* values of $\boldsymbol{b}_K\cdot\nabla\lambda_{K,i}$, $i\in\{1,2,\ldots,n+1\}$, are nonzero. This property can be used for characterization of a good mesh.

3.2 Derivation of the method

At the beginning of any finite element discretization, we derive the weak formulation of the respective problem. Let us therefore multiply (3.1) by the function $\varphi \in H_0^1(\Omega)$ and integrate over the whole domain Ω. Using Green's theorem (Theorem 4.1.1, page 124) the weak formulation of (3.1) reads:

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

$$\varepsilon(\nabla u, \nabla\varphi)_\Omega + (\boldsymbol{b}\cdot\nabla u, \varphi)_\Omega = (f,\varphi)_\Omega \qquad \forall\varphi \in H_0^1(\Omega). \tag{3.6}$$

Further, let us define the triangulation $\mathcal{T}_h$ of the domain Ω. It consists of a finite number of open simplicial elements K. We assume that $\overline{\Omega} = \cup_{K\in\mathcal{T}_h}\overline{K}$ and that the closures of any two different elements $K, \widetilde{K} \in \mathcal{T}_h$ are either disjoint or possess a common d-dimensional simplex ($d \in \{0,1,\ldots,n-1\}$). We also denote by $\mathcal{M}_h$ the set of nodes of $\mathcal{T}_h$ and by $\mathcal{N}_h \subset \mathcal{M}_h$ the set of all inner nodes of $\mathcal{T}_h$. The number of all nodes of $\mathcal{T}_h$ is then denoted by $M_h = |\mathcal{M}_h|$, whereas $N_h = |\mathcal{N}_h|$ stands for the number off all inner nodes.

To derive the Galerkin finite element discretization of (3.1), we define a finite element space $X_h = X_h^{(1)} = \{v_h \in \mathcal{C}(\Omega), v_h|_K \in P_1(K), \forall K \in \mathcal{T}_h\}$ and a space of test functions $V_h = V_h^{(1)} = X_h \cap H_0^1(\Omega)$. The barycentric coordinates $\{\lambda_{K,j}\}_{j=1}^{n+1}$ of the element $K \in \mathcal{T}_h$ then form a basis of $P_1(K)$ and we reorder them so that

$$\int_K \frac{\boldsymbol{b}\cdot\nabla\lambda_{K,j}}{|\nabla\lambda_{K,j}|}\,\mathrm{d}\boldsymbol{x} \;\le\; \int_K \frac{\boldsymbol{b}\cdot\nabla\lambda_{K,j+1}}{|\nabla\lambda_{K,j+1}|}\,\mathrm{d}\boldsymbol{x}, \quad \text{for } j = 1,2,\ldots,n. \tag{3.7}$$

Remark 3.2.1. Since $\sum_{j=1}^{n+1}\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,j}\,\mathrm{d}\boldsymbol{x} = 0$ for each $K \in \mathcal{T}_h$, then if one of the expressions (3.7) is nonzero we obtain $\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,1}\,\mathrm{d}\boldsymbol{x} < 0$ and $\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,n+1}\,\mathrm{d}\boldsymbol{x} > 0$.

Further, we assume that the barycentric coordinates $\{\lambda_{K,j}\}_{j=1}^{n+1}$ satisfy for each $K \in \mathcal{T}_h$ the inequality

$$(\nabla\lambda_{K,j}, \nabla\lambda_{K,i})_K \;\le\; 0 \quad \text{whenever } i \neq j. \tag{3.8}$$

In 2D this assumption is satisfied for triangulations not containing obtuse triangles.

The SUPG method adds weighted residuals $R(u) = -\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u - f$ to the usual Galerkin finite element method. Since $R(u)$ vanishes for the exact solution, we can add any multiple of $R(u)$ to the weak formulation providing $u \in H^2(\Omega)$. Unlike the original SUPG method, which adds the residual multiplied by the streamline derivative of v, we add the residual multiplied on each $K \in \mathcal{T}_h$ by derivative of v in the direction $P_{K,n+1} - C_K$ (see Lamač (2015)). Here C_K are the barycentres of K and $P_{K,j}$, $j = 1,2,\ldots,n+1$ are the vertices of K satisfying $\lambda_{K,i}(P_{K,j}) = \delta_{ij}$ for $1 \le i,j \le n+1$.

Thus, the solution $u \in H_0^1(\Omega) \cap H^2(\Omega)$ of the problem (3.6) satisfies also for all $\varphi \in H_0^1(\Omega)$

$$a(u,\varphi) \;=\; F(\varphi), \tag{3.9}$$

where

$$F(\varphi) = \sum_{K\in\mathcal{T}_h} (f, \varphi + (P_{K,n+1} - C_K)\nabla\varphi)_K \qquad \text{and} \tag{3.10}$$

$$a(u,\varphi) = \varepsilon(\nabla u, \nabla\varphi)_\Omega + (\boldsymbol{b}\cdot\nabla u, \varphi)_\Omega + \\ + \sum_{K\in\mathcal{T}_h} \Big(-\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u, (P_{K,n+1} - C_K)\nabla\varphi \Big)_K. \tag{3.11}$$

If we now apply the finite element method using the continuous piecewise linear finite elements, the spurious oscillations unfortunately persist (analogous to the original SUPG method). The reason is the presence of the positive off-diagonal entries in the matrix obtained by the discretization of the last two terms in (3.11) resulting in the unfulfilment of the discrete maximum principle.

In order to eliminate these positive entries, we define $\boldsymbol{d}_{K,j} = P_{K,n+1} - P_{K,j}$, $j = 1, 2, \ldots, n$ and consider the element-wise constant approximation $\boldsymbol{b}_K$ of the vector field $\boldsymbol{b}$ by vectors that are parallel with $\boldsymbol{d}_{K,1}$ on each element K. More precisely, first of all we consider that our mesh is "well-aligned" with respect to the vector field $\boldsymbol{b}$ and then on each element K we construct a constant approximation $\boldsymbol{b}_K$ of $\boldsymbol{b}$. This "well-alignment" is provided by the following assumptions.

($\mathcal{A}1$) The ordering given by (3.7) on each $K \in \mathcal{T}_h$ uniquely defines the vector $\boldsymbol{d}_{K,1} = P_{K,n+1} - P_{K,1}$. We assume that if any edge e of $\mathcal{T}_h$ corresponds to $\boldsymbol{d}_{K,1}$ of some K, then e corresponds to $\boldsymbol{d}_{K,1}$ for each K containing e. We denote by $\mathcal{E}_h$ the set of such edges.

($\mathcal{A}2$) Each inner node P of $\mathcal{T}_h$ is the endpoint of exactly two edges of $\mathcal{E}_h$.

Remark 3.2.2. Let us call a *discrete streamline* any set of edges $\mathcal{S} \subset \mathcal{E}_h$ such that for each $e \in \mathcal{S}$ there exists $e' \in \mathcal{S}$ such that

$$e' \neq e \quad \& \quad e \cap e' \neq \emptyset. \tag{3.12}$$

The discrete streamline $\mathcal{S}$ is *closed* if for each $e \in \mathcal{S}$ there exist exactly two different edges e' and e'' satisfying (3.12). Consequently, the assumptions ($\mathcal{A}1$) $-$ $-$($\mathcal{A}2$) do not allow closed discrete streamlines in 2D. Indeed, if there is a closed discrete streamline then there exists a node ("inside" the closed streamline) which does not satisfy ($\mathcal{A}2$). The mesh satisfying ($\mathcal{A}1$) $-$ $-$($\mathcal{A}2$) can be, for instance, constructed by approximation of streamlines by linear spline functions. This will be the subject of future work. Further assumptions on the structure of the mesh will be given by the inequalities (3.34) and (3.70).

It remains to define the piecewise constant approximation of $\boldsymbol{b}$. On each element $K \in \mathcal{T}_h$ it is defined in the following way

$$\boldsymbol{b}_K = -\frac{1}{|K|}\left(\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,1}\,\mathrm{d}\boldsymbol{x}\right)\boldsymbol{d}_{K,1}. \tag{3.13}$$

Consequently, when $\boldsymbol{b} = \alpha\boldsymbol{d}_{K,1}$ in K for some $\alpha \in \mathbb{R}$, the previous definition of $\boldsymbol{b}_K$ implies that $\boldsymbol{b}_K = -\frac{1}{|K|}\left(\int_K -\alpha\,\mathrm{d}\boldsymbol{x}\right)\boldsymbol{d}_{K,1} = \boldsymbol{b}$ in K.

Finally, we apply the finite element method and the new method reads: Find $u_h \in V_h$ such that for all $\varphi_h \in V_h$ there holds

$$a_h(u_h, \varphi_h) = F_h(\varphi_h), \tag{3.14}$$

where

$$\begin{aligned} a_h(u,\varphi) &= \varepsilon(\nabla u, \nabla\varphi)_\Omega + \sum_{K\in\mathcal{T}_h} (\boldsymbol{b}_K \cdot \nabla u, \varphi)_K + \\ &\quad + \sum_{K\in\mathcal{T}_h} \Big(-\varepsilon\Delta u + \boldsymbol{b}_K \cdot \nabla u, (P_{K,n+1} - C_K)\,\nabla\varphi \Big)_K, \qquad (3.15) \\ F_h(\varphi) &= \sum_{K\in\mathcal{T}_h} \Big(f, \varphi + (P_{K,n+1} - C_K)\,\nabla\varphi \Big)_K \qquad (3.16) \end{aligned}$$

and the vectors $\boldsymbol{b}_K$ are defined by (3.13).

3.2.1 Monotonicity

Since we would like to avoid spurious oscillations in the discrete solution, the new method should satisfy the discrete maximum principle. We prove it with a help of matrices of nonnegative type.

Definition 3.2.1. *The matrix $\mathbb{A} = \{a_{ij}\}_{i=1}^{p}{}_{j=1}^{q}$, $p \le q$, is of* nonnegative type *if the following conditions hold:*

$$a_{ij} \le 0 \quad \textit{whenever } i \ne j \qquad \textit{and} \qquad \sum_{j=1}^{q} a_{ij} \ge 0 \quad \textit{for all } i = 1, 2, \ldots, p. \tag{3.17}$$

When solving partial differential equations numerically one usually comes to a system of linear equations $\mathbb{A}\boldsymbol{x} = \boldsymbol{z}$, where $\mathbb{A} = \{a_{ij}\}_{i=1}^{p}{}_{j=1}^{q}$, $p \le q$, is a rectangular matrix, $\boldsymbol{z} = (z_1, z_2, \ldots, z_p)^T$ is a vector obtained by the discretization of the right-hand side of the respective partial differential equation and $\boldsymbol{x} = (x_1, x_2, \ldots, x_q)^T$ is a vector of unknowns. In fact, $q - p$ entries of the vector $\boldsymbol{x}$ are known due to the boundary condition and without loss of generality we use the last $q - p$ entries of $\boldsymbol{x}$ for this purpose. Thus, it remains to compute the first p components of $\boldsymbol{x}$.

In order to obtain a system of equations with a square matrix we denote by $\mathbb{S} = \{s_{ij}\}_{i=1}^{q}{}_{j=1}^{p}$ the $q \times p$ matrix satisfying $s_{ij} = \delta_{ij}$ for all $i = 1, 2, \ldots, q$ and $j = 1, 2, \ldots, p$. Then $\mathbb{A}_r = \mathbb{A}\mathbb{S}$ is a square matrix formed by first p columns of $\mathbb{A}$ and $\boldsymbol{x}_r = \mathbb{S}^T\boldsymbol{x}$ is a restriction of $\boldsymbol{x}$ to the first p rows. Finally, if we define $\widetilde{\boldsymbol{z}} = \boldsymbol{z} - \mathbb{A}(\mathbb{I} - \mathbb{S}\mathbb{S}^T)\boldsymbol{x}$ and if the matrix $\mathbb{A}_r$ is nonsingular, then there exists a unique solution of the equation $\mathbb{A}_r\boldsymbol{x}_r = \widetilde{\boldsymbol{z}}$ ($\widetilde{\boldsymbol{z}}$ is defined using x_i with $i > p$). We can also verify that

$$\mathbb{A}\boldsymbol{x} \;=\; \mathbb{A}\mathbb{S}\mathbb{S}^T\boldsymbol{x} + \mathbb{A}(\mathbb{I} - \mathbb{S}\mathbb{S}^T)\boldsymbol{x} \;=\; \mathbb{A}_r\boldsymbol{x}_r + \boldsymbol{z} - \widetilde{\boldsymbol{z}} \;=\; \boldsymbol{z}. \tag{3.18}$$

Theorem 3.2.1. *Assume that $\mathbb{A}\boldsymbol{x} = \boldsymbol{z}$, where $\mathbb{A} \in \mathbb{R}^{p\times q}$, $p \le q$, is a matrix of nonnegative type and $\mathbb{A}_r$ is a nonsingular matrix, then the discrete maximum principle holds, i.e.*

$$\boldsymbol{z} \le 0 \qquad \Rightarrow \qquad \max_{1\le i\le q}\{x_i\} \;\le\; \max\{0, x_\nu\}, \quad \textit{for some } \nu > p. \tag{3.19}$$

If in addition $\sum_{j=1}^{q} a_{ij} = 0$ *for all* $i = 1, 2, \ldots, p$, *then there holds*

$$\boldsymbol{z} \leq 0 \quad \Rightarrow \quad \max_{1\leq i\leq q}\{x_i\} \;=\; x_\nu, \qquad \text{for some } \nu > p. \tag{3.20}$$

Proof. We proceed as in Codina (1993). Let us begin with the first statement (i.e. (3.19)). Since the matrix $\mathbb{A}$ is of nonnegative type and does not contain zero rows ($\mathbb{A}_r$ is nonsingular) it must have positive entries on the main diagonal. Consequently, from the equality $\sum_{j=1}^{q} a_{ij}x_j = z_i$ it follows

$$x_i \;=\; \frac{z_i}{a_{ii}} + \frac{1}{a_{ii}} \sum_{j\in\mathcal{S}_i} |a_{ij}| x_j \;\leq\; \max_{j\in\mathcal{S}_i}\{x_j\} \frac{1}{a_{ii}} \sum_{j\in\mathcal{S}_i} |a_{ij}| \;\leq\; \max\left\{0, \max_{j\in\mathcal{S}_i}\{x_j\}\right\}, \tag{3.21}$$

where $\mathcal{S}_i = \{j; 1 \leq j \leq q, a_{ij} < 0\}$ is a set of indices of nonzero off-diagonal entries in the i-th row of the matrix $\mathbb{A}$. In other words, $\mathcal{S}_i$ is a set of indices of neighboring nodes of the node corresponding to the value x_i. Let us now denote $x_m = \max_{1\leq i\leq q}\{x_i\}$. If $x_m \leq 0$, then (3.19) holds. Therefore, let us consider the case $x_m > 0$ and for a contradiction let us assume that

$$1 \leq m \leq p \quad \text{and} \quad x_j < x_m \text{ for all } j > p. \tag{3.22}$$

Denoting $x_k = \max_{j\in\mathcal{S}_m}\{x_j\}$ and using (3.21) with $i = m$, we obtain $0 < x_m \leq \max\{0, x_k\}$. Thus, x_k has to be positive and $x_m = x_k$ (x_m is the maximum). Moreover, since $x_k \not< x_m$ we have $k \leq p$ by (3.22) and the maximum is attained at two inner nodes. The system $\mathbb{A}\boldsymbol{x} = \boldsymbol{z}$ of p equations for q variables can be now changed into an equivalent system of $p-1$ equations for $q-1$ variables by eliminating the k-th row and summing the k-th and the m-th column of the matrix $\mathbb{A}$ together (adding the k-th column to the m-th column and then eliminating the k-th column). The resulting matrix is again of nonnegative type and repeating this proof we arrive at $x_1 = x_2 = \cdots = x_p \leq \max_{j>p}\{x_j\}$ by (3.21), which is a contradiction with (3.22).

The second statement of the theorem results from the fact that if $\sum_{j=1}^{q} a_{ij} = 0$ for all $i = 1, 2, \ldots, p$, then $\frac{1}{a_{ii}} \sum_{j\in\mathcal{S}_i} |a_{ij}| = 1$ for all $i = 1, 2, \ldots, p$. Consequently, the inequality (3.21) changes into $x_i \leq \max_{j\in\mathcal{S}_i}\{x_j\}$, for $i = 1, 2, \ldots, p$. □

Remark 3.2.3. Analogously, one can prove that if $\mathbb{A}\boldsymbol{x} = \boldsymbol{z}$, where $\mathbb{A} \in \mathbb{R}^{p\times q}$, $p \leq q$, is a matrix of nonnegative type and $\mathbb{A}_r$ is a nonsingular matrix, then the discrete minimum principle holds, i.e.

$$\boldsymbol{z} \geq 0 \quad \Rightarrow \quad \min_{1\leq i\leq q}\{x_i\} \;\geq\; \min\{0, x_\nu\}, \quad \text{for some } \nu > p. \tag{3.23}$$

If in addition $\sum_{j=1}^{q} a_{ij} = 0$ for all $i = 1, 2, \ldots, p$, then there holds

$$\boldsymbol{z} \geq 0 \quad \Rightarrow \quad \min_{1\leq i\leq q}\{x_i\} \;=\; x_\nu, \qquad \text{for some } \nu > p. \tag{3.24}$$

Theorem 3.2.2. *The method (3.14)–(3.16) satisfies the discrete maximum principle.*

Proof. It suffices to show that the matrix generated by the bilinear form a_h is of nonnegative type. Thus, let $\varphi_h, \tilde{\varphi}_h$ be arbitrary basis functions of V_h and let us rewrite the bilinear form a_h in the following form

$$a_h(\varphi_h, \tilde{\varphi}_h) \;=\; \sum_{K\in\mathcal{T}_h} \Big\{ \varepsilon(\nabla\varphi_h, \nabla\tilde{\varphi}_h)_K + (\boldsymbol{b}_K \cdot \nabla\varphi_h, \tilde{\varphi}_h + (P_{K,n+1} - C_K)\nabla\tilde{\varphi}_h)_K \Big\}. \tag{3.25}$$

Now we investigate the restriction of $a_h(\varphi_h, \tilde{\varphi}_h)$ to a single element K. Without loss of generality we denote $\lambda_{K,i} = \varphi_h|_K$ and $\lambda_{K,j} = \tilde{\varphi}_h|_K$ for some $1 \leq i, j \leq n+1$. Since the first term in the sum (3.25) satisfies the inequality (3.8) (multiplied by $\varepsilon > 0$), it remains to analyze the second term in the sum. From the linearity of the function $\lambda_{K,j}$ it follows that

$$(P_{K,n+1} - C_K)\,\nabla\lambda_{K,j} \;=\; \lambda_{K,j}\,(P_{K,n+1}) - \lambda_{K,j}\,(C_K)\,. \tag{3.26}$$

Using this property, the fact that $\boldsymbol{b}_K \cdot \nabla\lambda_{K,i}$ is a constant function on K, $\boldsymbol{b}_K \,||\, \boldsymbol{d}_{K,1}$ and $\lambda_{K,\mu}(P_{K,\nu}) = \delta_{\mu,\nu}$ for $1 \leq \mu, \nu \leq n+1$, we deduce

$$\begin{aligned}
&(\boldsymbol{b}_K \cdot \nabla\lambda_{K,i}, \lambda_{K,j} + (P_{K,n+1} - C_K)\,\nabla\lambda_{K,j})_K \;= \\
&\quad= \;(\boldsymbol{b}_K \cdot \nabla\lambda_{K,i}, \lambda_{K,j}\,(P_{K,n+1}) - \lambda_{K,j}\,(C_K))_K = \\
&\quad= \;(\boldsymbol{b}_K \cdot \nabla\lambda_{K,i}, \lambda_{K,j}\,(P_{K,n+1}))_K \;=\; \tfrac{|\boldsymbol{b}_K|}{|\boldsymbol{d}_{K,1}|}\,(\boldsymbol{d}_{K,1} \cdot \nabla\lambda_{K,i}, \lambda_{K,j}\,(P_{K,n+1}))_K = \\
&\quad= \;\tfrac{|\boldsymbol{b}_K|}{|\boldsymbol{d}_{K,1}|}\,|K|\,(\delta_{i,n+1} - \delta_{i,1})\delta_{j,n+1},
\end{aligned} \tag{3.27}$$

where we used the equality

$$\boldsymbol{d}_{K,1}\cdot\nabla\lambda_{K,i} \;=\; (P_{K,n+1} - P_{K,1})\cdot\nabla\lambda_{K,i} \;=\; \lambda_{K,i}(P_{K,n+1}) - \lambda_{K,i}(P_{K,1}) \;=\; \delta_{i,n+1} - \delta_{i,1}. \tag{3.28}$$

We observe that when $i = j = n+1$ the term (3.27) is positive, for $j = n+1$ and $i = 1$ it is negative and in all remaining cases it vanishes. Moreover, since $\sum_{i=1}^{n+1}(\nabla\lambda_{K,i}, \nabla\lambda_{K,j})_K = 0$ and $\sum_{i=1}^{n+1}(\delta_{i,n+1} - \delta_{i,1})\delta_{j,n+1} = 0$, the method satisfies the discrete maximum principle (3.20). □

Remark 3.2.4. Instead of adding stabilization term to the weak formulation (3.6) one can change the test functions to

$$\widetilde{\lambda}_{K,j} \;=\; \lambda_{K,j} + (P_{K,n+1} - C_K)\cdot\nabla\lambda_{K,j}. \tag{3.29}$$

Then for all $j = 1, 2, \ldots, n$ we obtain $\widetilde{\lambda}_{K,j} = \lambda_{K,j} - \frac{1}{n+1}$ whereas $\widetilde{\lambda}_{K,n+1} = \lambda_{K,n+1} + \frac{n}{n+1}$. This choice of test functions is the same as in the Mizukami-Hughes method (cf. Mizukami and Hughes (1985) or Knobloch (2006)). It means that the derived method satisfies the discrete maximum principle.

3.3 Mesh properties and notation

In this section we introduce another mesh quantities and labeling. We observe that the mesh whose edges are oriented along $\boldsymbol{b}$ has a special property: For each mesh node P_0^s lying on the boundary Γ_- there exists a sequence of nodes $\{P_j^s\}_{j=1}^{N_s}$ which lay on the same streamline given by the vector field $\boldsymbol{b}$ (of course, that here the verb "lay" in fact means "for a good mesh they almost lay").

Thus, each node P_j^s of the mesh can be characterized by two numbers - the number denoting the streamline (s) and the number determining the order of the node on this streamline (j). For each node P_j^s we can further define the following sets: a patch $\Omega_j^s = \cup_{P_j^s \subset \overline{K}} K$, a cluster $\mathcal{C}_j^s = \cup_{P_{j-1}^s, P_j^s \subset \overline{K}} K$ and a complementary set $\Omega_{0,j}^s = \Omega_j^s \backslash \left(\mathcal{C}_j^s \cup \mathcal{C}_{j+1}^s\right)$ (see Figure 3.1).

From this notation it also follows that each mesh node has double labeling $P_{K,i}$ and P_j^s, in particular, for all $K \subset \mathcal{C}_j^s$ holds $P_{K,1} = P_{j-1}^s$ and $P_{K,n+1} = P_j^s$.

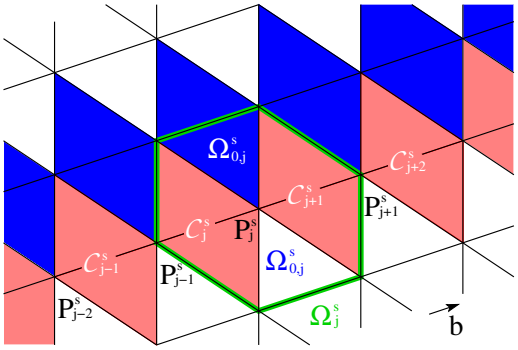

Figure 3.1: Definition of the splitting of the domain Ω_j^s.

Another property resulting from the structure of the mesh is that we can rewrite the sum over all elements $K \in \mathcal{T}_h$ in the form

$$\sum_{K\in\mathcal{T}_h} = \sum_{s=1}^{\mathcal{P}} \sum_{j=1}^{N_s} \sum_{K\subset\mathcal{C}_j^s} . \tag{3.30}$$

Indeed, for each element $K \in \mathcal{T}_h$ there exists exactly one edge (determining the vector $\boldsymbol{d}_{K,1}$) which is oriented in the flow direction. This edge certainly lies on some discrete streamline s, $1 \le s \le \mathcal{P}$, and the endpoints of this edge are P^s_{j-1}, P^s_j for suitable j, $1 \le j \le N_s$. All elements sharing this edge then form the cluster $\mathcal{C}_j^s$ and a union of all clusters is the whole domain Ω. Since each element $K \in \mathcal{T}_h$ lies exactly in one cluster, the expression (3.30) is valid.

Definition 3.3.1. *For each cluster $\mathcal{C}_j^s$ let us define the quantities*

$$\begin{aligned} h_j^s &= |P_j^s - P_{j-1}^s|, \qquad \beta_j^s = \frac{1}{|\mathcal{C}_j^s|} \sum_{K\subset\mathcal{C}_j^s} |\boldsymbol{b}_K||K| \qquad \textit{and} \\ q_j^s &= -\sum_{K\subset\mathcal{C}_j^s} \int_K \boldsymbol{b}\cdot\nabla\lambda_{K,1}\,\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{3.31}$$

For each element $K \in \mathcal{T}_h$ let us also define the mesh parameters θ_K by

$$\theta_K = \frac{1}{|K|} \max\left\{ \max_{2\le i\le n} \left| \int_K \boldsymbol{b}\cdot\nabla\lambda_{K,i}\,\mathrm{d}\boldsymbol{x} \right|, \left| \sum_{i=2}^{n} \int_K \boldsymbol{b}\cdot\nabla\lambda_{K,i}\,\mathrm{d}\boldsymbol{x} \right| \right\}. \tag{3.32}$$

Remark 3.3.1. From the previous definition it follows that h_j^s is the length of the cluster $\mathcal{C}_j^s$ in the streamline direction, i.e. $h_j^s = |P_j^s - P_{j-1}^s| = |P_{K,n+1} - P_{K,1}| = |\boldsymbol{d}_{K,1}|$ for each element $K \subset \mathcal{C}_j^s$. Further, for the quantity q_j^s holds

$$\begin{aligned} q_j^s &= -\sum_{K\subset\mathcal{C}_j^s} \int_K \boldsymbol{b}\cdot\nabla\lambda_{K,1}\,\mathrm{d}\boldsymbol{x}. = \sum_{K\subset\mathcal{C}_j^s} \left| \int_K \boldsymbol{b}\cdot\nabla\lambda_{K,1}\,\mathrm{d}\boldsymbol{x} \right| = \\ &= \sum_{K\subset\mathcal{C}_j^s} \frac{|\boldsymbol{b}_K||K|}{|\boldsymbol{d}_{K,1}|} = \frac{1}{h_j^s} \sum_{K\subset\mathcal{C}_j^s} |\boldsymbol{b}_K||K| = \frac{\beta_j^s |\mathcal{C}_j^s|}{h_j^s} > 0, \end{aligned} \tag{3.33}$$

which results from the Remark 3.2.1.

In the previous definition the quantity β_j^s is the weighted average value of $|\boldsymbol{b}_K|$ on $\mathcal{C}_j^s$ and q_j^s are fluxes for which we derive inequalities in technical Lemmas 3.4.1, 3.4.2, 3.4.3 and 3.4.4 later. The mesh parameters θ_K vanish whenever $\boldsymbol{b}$ is parallel to $\boldsymbol{b}_K$ (i.e., to $\boldsymbol{d}_{K,1}$) in K and therefore we use them for a characterization of a good mesh.

In order to employ the algebraic lemmas from Section 3.4.1, we have to find some relation between the values q_j^s and q_{j+1}^s. The following lemma provides an inequality resulting from the structure of the mesh.

Lemma 3.3.1. *Let there exists $\omega > 0$ such that $\operatorname{div} \boldsymbol{b} \leq -\omega < 0$ in Ω and let for each $K \in \mathcal{T}_h$ holds*

$$\theta_K \ \leq \ \frac{\omega}{n+1}. \tag{3.34}$$

Then $q_j^s \geq q_{j+1}^s + \frac{\omega}{n+1}|\mathcal{C}_{j+1}^s|$ for each $s = 1, 2, \ldots, \mathcal{P}$ and $j = 1, 2, \ldots, N_s$.

Proof. Let us consider any inner node P_j^s and the corresponding basis function λ_j^s satisfying $\operatorname{supp} \lambda_j^s = \overline{\Omega_j^s} = \overline{\mathcal{C}_j^s} \cup \overline{\Omega_{0,j}^s} \cup \overline{\mathcal{C}_{j+1}^s}$. Then for $K \subset \mathcal{C}_j^s$ holds $\nabla \lambda_{j-1}^s = \nabla \lambda_{K,1} = -\nabla \lambda_{K,n+1} - \sum_{i=2}^n \nabla \lambda_{K,i} = -\nabla \lambda_j^s - \sum_{i=2}^n \nabla \lambda_{K,i}$ and from the definition of q_j^s it follows

$$\begin{aligned}
q_j^s \ &= \ -\sum_{K \subset \mathcal{C}_j^s} \int_K \boldsymbol{b} \cdot \nabla \lambda_{K,1} \,\mathrm{d}\boldsymbol{x} \ = \ -\sum_{K \subset \mathcal{C}_j^s} \int_K \boldsymbol{b} \cdot \nabla \lambda_{j-1}^s \,\mathrm{d}\boldsymbol{x} \ = \\
&= \ \sum_{K \subset \mathcal{C}_j^s} \sum_{i=2}^n \int_K \boldsymbol{b} \cdot \nabla \lambda_{K,i} \,\mathrm{d}x + \sum_{K \subset \mathcal{C}_j^s} \int_K \boldsymbol{b} \cdot \nabla \lambda_j^s \,\mathrm{d}\boldsymbol{x} \ = \\
&= \ \sum_{K \subset \mathcal{C}_j^s} \sum_{i=2}^n \int_K \boldsymbol{b} \cdot \nabla \lambda_{K,i} \,\mathrm{d}x + \int_{\Omega_j^s} \boldsymbol{b} \cdot \nabla \lambda_j^s \,\mathrm{d}\boldsymbol{x} - \int_{\Omega_{0,j}^s} \boldsymbol{b} \cdot \nabla \lambda_j^s \,\mathrm{d}x + q_{j+1}^s \ = \\
&= \ q_{j+1}^s + \sum_{K \subset \mathcal{C}_j^s} \sum_{i=2}^n \int_K \boldsymbol{b} \cdot \nabla \lambda_{K,i} \,\mathrm{d}x - \int_{\Omega_j^s} \operatorname{div} \boldsymbol{b} \, \lambda_j^s \,\mathrm{d}x - \int_{\Omega_{0,j}^s} \boldsymbol{b} \cdot \nabla \lambda_j^s \,\mathrm{d}x \ \geq \\
&\geq \ q_{j+1}^s - \sum_{K \subset \mathcal{C}_j^s} \frac{\omega}{n+1}|K| + \frac{\omega}{n+1}|\Omega_j^s| - \sum_{K \subset \Omega_{0,j}^s} \frac{\omega}{n+1}|K| \ = \\
&= \ q_{j+1}^s + \frac{\omega}{n+1}|\Omega_j^s \backslash (\mathcal{C}_j^s \cup \Omega_{0,j}^s)| \ = \ q_{j+1}^s + \frac{\omega}{n+1}|\mathcal{C}_{j+1}^s|.
\end{aligned} \tag{3.35}$$

□

Corollary 3.3.1. If $\operatorname{div} \boldsymbol{b} < -\omega < 0$ and the inequality (3.34) holds for all $K \in \mathcal{T}_h$, then there are not *closed* discrete streamlines in $\mathcal{T}_h$.

Proof. For a contradiction let us assume that the clusters $\mathcal{C}_j^s$, $j = 1, 2, \ldots, N_s$ lay on some closed discrete streamline s. From the inequality (3.34) it then follows that $q_1^s > q_2^s > \cdots > q_{N_s}^s > q_1^s$, which is not possible. □

Remark 3.3.2. Since the method is formulated in arbitrary dimension $\mathbb{R}^n$, $n \in \mathbb{N}$, let us now investigate the number of elements forming one cluster and one patch in $\mathbb{R}^n$. Whereas in 1D the cluster always consists of one element and two neighboring elements form the patch, in higher dimensions these numbers depend on the structure of the mesh. Therefore, let us for simplicity consider a triangulation of Ω by a three-directional mesh (in 2D) or its multidimensional analog. These meshes are constructed in the following way:

Let $\Omega \subset \mathbb{R}^n$ be a hypercube (one can consider any n-dimensional parallelepiped) and we divide it into small hypercubes whose faces are parallel with the faces of Ω. Further, we divide each small hypercube into $n!$ n-simplices. We demonstrate such a partition on the cube $[0,1]^n$:

Let $\mathcal{S}_n$ be a symmetric group of permutations of degree n, i.e. the group of permutations on the set $\{1,2,\ldots,n\}$. Let $\pi \in \mathcal{S}_n$ be any permutation and let us define a set

$$K_\pi = \{(x_1, x_2, \ldots, x_n) \subset \mathbb{R}^n, 0 \le x_{\pi(1)} \le x_{\pi(2)} \le \cdots \le x_{\pi(n)} \le 1\}. \tag{3.36}$$

Since K_π is defined by $n+1$ linearly independent linear inequalities, it is an n-simplex. Further, any point $(x_1, x_2, \ldots, x_n) \in [0,1]^n$ clearly lies in the union of all possible simplices $\bigcup_{\pi\in\mathcal{S}_n} K_\pi$. This is due to the fact that we can always permute (using some permutation π_0) the coordinates $x_1, x_2, \ldots, x_n$ in such a way that they form a non-decreasing sequence and thus $(x_1, x_2, \ldots, x_n) \in K_{\pi_0}$. Finally, if $(x_1, x_2, \ldots, x_n) \in K_{\pi_1} \cap K_{\pi_2}$ for some permutations $\pi_1 \neq \pi_2$, then both sequences $\{x_{\pi_1(j)}\}_{j=1}^n$ and $\{x_{\pi_2(j)}\}_{j=1}^n$ are non-decreasing (by the definition of K_{π_1} and K_{π_2}). It means that they are identical and there must exist at least one couple of coordinates x_p and x_q, $p \neq q$, satisfying $x_p = x_q$. Thus, all points laying in two or more simplices always lay on their boundaries (some inequalities are in fact equalities in the definition of K_{π_1} and K_{π_2}). It means that $\{K_\pi\}_{\pi\in\mathcal{S}_n}$ forms the partition of $[0,1]^n$ (see Figure 3.2 for 3D example).

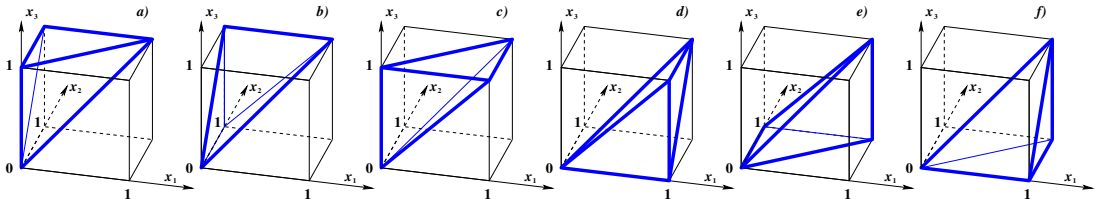

Figure 3.2: Partition of the unit cube into simplices K_π, $\pi \in \mathcal{S}_3$, in 3D. Images a) – f) correspond to the permutations $\binom{123}{123}$, $\binom{123}{132}$, $\binom{123}{213}$, $\binom{123}{231}$, $\binom{123}{312}$ and $\binom{123}{321}$, respectively.

Further, using the definition (3.36) one can compute the volume of each simplex K_π, $\pi \in \mathcal{S}_n$. It is equal to

$$|K_\pi| \;=\; \int_0^1 \int_0^{x_{\pi(n)}} \cdots \int_0^{x_{\pi(2)}} 1 \,\mathrm{d}x_{\pi(1)} \cdots \mathrm{d}x_{\pi(n)} \;=\; \int_0^1 \frac{x_{\pi(n)}^{n-1}}{(n-1)!} \,\mathrm{d}x_{\pi(n)} \;=\; \frac{1}{n!}. \tag{3.37}$$

Hence, all the simplices forming the partition of $[0,1]^n$ have the same volume.

It remains to verify that the opposite faces of $[0,1]^n$ are divided into $(n-1)$-simplices in the same way (we want to set the (hyper)cubes together). Therefore, let $j \in \{1,2,\ldots,n\}$ be arbitrary but fixed and let us consider two faces F_0 and F_1 laying in the hyperplanes $x_j = 0$ and $x_j = 1$, respectively. Further, let us define two subsets of $\mathcal{S}_n$

$$\mathcal{L}_n^{(j)} \;=\; \{\pi \in \mathcal{S}_n,\, \pi(1) = j\} \quad \text{and} \quad \mathcal{R}_n^{(j)} \;=\; \{\pi \in \mathcal{S}_n,\, \pi(n) = j\}. \tag{3.38}$$

Then the sets $\left\{K_\pi \cap \{x_j = 0\}, \pi \in \mathcal{L}_n^{(j)}\right\}$ and $\left\{K_\pi \cap \{x_j = 1\}, \pi \in \mathcal{R}_n^{(j)}\right\}$ form the partitions of F_0 and F_1, respectively, and the mapping $\sigma_n^{(j)} : \mathcal{L}_n^{(j)} \to \mathcal{R}_n^{(j)}$ defined

by the relation

$$\Big(\sigma_n^{(j)}(\pi)\Big)(i) \;=\; \pi\Big(1+(i \mod n)\Big) \quad \text{for } i=1,2,\ldots,n, \tag{3.39}$$

is a bijection between $\mathcal{L}_n^{(j)}$ and $\mathcal{R}_n^{(j)}$.

What remains to show is that if $(x_1, x_2, \ldots, x_{j-1}, 0, x_{j+1}, \ldots, x_n)$ is any point laying in the $(n-1)$-simplex $K_\pi \cap \{x_j = 0\}$, for some $\pi \in \mathcal{L}_n^{(j)}$, then the point $(x_1, x_2, \ldots, x_{j-1}, 1, x_{j+1}, \ldots, x_n)$ lies in $K_{\sigma_n^{(j)}(\pi)} \cap \{x_j = 1\}$. However, this is obvious as for $i < n-1$ there holds

$$x_{\left(\sigma_n^{(j)}(\pi)\right)(i)} \;=\; x_{\pi(i+1)} \;\le\; x_{\pi(i+2)} \;=\; x_{\left(\sigma_n^{(j)}(\pi)\right)(i+1)}, \tag{3.40}$$

and since it is also $x_{\left(\sigma_n^{(j)}(\pi)\right)(n-1)} \le 1 = x_{\left(\sigma_n^{(j)}(\pi)\right)(n)}$ the verification is completed.

From the previous observations it follows that using this special type of mesh each inner mesh node belongs to 2^n hypercubes and in each hypercube it lies in a different position. Equivalently, in each hypercube, there is 2^n types of nodes (corners) depending on their position. Thus, when computing the number of elements forming one patch, it suffices to consider one hypercube and compute for each corner the number of simplices containing this corner. The sum of these numbers equals to the number of simplices forming the hypercube multiplied by the number of their corners, i.e. $n!(n+1) = (n+1)!$. Using this result, one can easily compute the number of elements forming one cluster. It is simply the number of elements forming the $(n-1)$-dimensional boundary patch. Indeed, each boundary node is in fact an endpoint of the streamline and the number of cluster's elements is therefore the same as the number of elements forming the boundary node patch, i.e. $n!$ (see Figure 3.3 for 3D example).

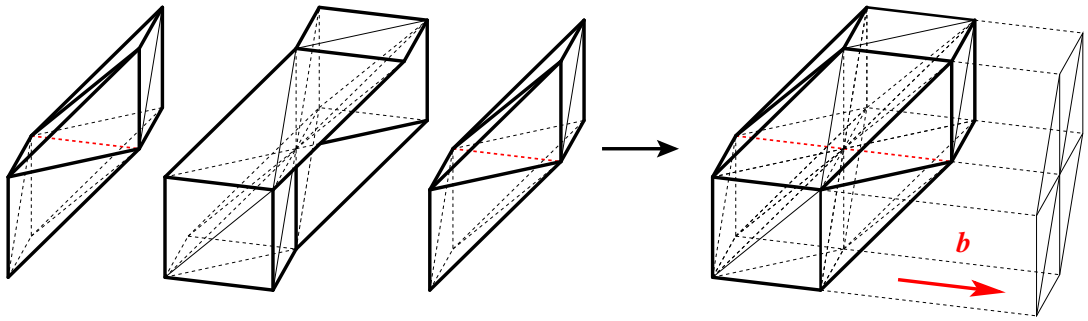

Figure 3.3: Example of clusters, a complementary set and a patch in 3D. The number of elements forming one three-dimensional cluster is the same as the number of elements forming the two-dimensional boundary patch.

3.4 Coercivity

Since $\int_K v_h - v_h(C_K)\,\mathrm{d}\boldsymbol{x} = 0$ for all $v_h \in V_h$, we can write

$$\Big(\boldsymbol{b}_K\cdot\nabla u_h,\, v_h + (P_{K,n+1} - C_K)\cdot\nabla v_h\Big)_K = \Big(\boldsymbol{b}_K\cdot\nabla u_h,\, v_h + v_h(P_{K,n+1}) - v_h(C_K)\Big)_K =$$

$$= \Big(\boldsymbol{b}_K\cdot\nabla u_h, v_h(P_{K,n+1})\Big)_K \;=\; |K|\,\frac{|\boldsymbol{b}_K|}{|\boldsymbol{d}_{K,1}|}\Big(u_h(P_{K,n+1}) - u_h(P_{K,1})\Big)v_h(P_{K,n+1}).$$

Consequently, for the bilinear form a_h holds

$$a_h(v_h, v_h) = \varepsilon |v_h|^2_{1,\Omega} + \sum_{K \in \mathcal{T}_h} |K| \frac{|\boldsymbol{b}_K|}{|\boldsymbol{d}_{K,1}|} \Big(v_h(P_{K,n+1}) - v_h(P_{K,1}) \Big) v_h(P_{K,n+1}). \quad (3.41)$$

Thus, when proving coercivity of the bilinear form a_h, it is necessary to estimate the second term on the right-hand side of (3.41). For this purpose we use the following lemmas.

3.4.1 Technical lemmas

Lemma 3.4.1. *Let $N \in \mathbb{N}$, $0 < \rho_j < 1$, $j = 1, 2, \ldots, N-1$, and q_j, $j = 1, 2, \ldots, N$, are positive numbers satisfying*

$$\frac{q_{j+1}}{q_j} \leq \rho_j \qquad \text{for } j = 1, 2, \ldots, N-1. \quad (3.42)$$

Then for all $v_j \in \mathbb{R}$, $j = 1, 2, \ldots, N$, holds

$$q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j^2 - v_j v_{j-1}) \;\geq\; \frac{1}{2} q_N v_N^2 + \frac{1}{2} \sum_{j=1}^{N-1} (1 - \rho_j) q_j v_j^2. \quad (3.43)$$

Proof. Subtracting the right-hand side of (3.43) from the left-hand side we obtain

$$\begin{aligned} & q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j^2 - v_j v_{j-1}) - \frac{1}{2} q_N v_N^2 - \frac{1}{2} \sum_{j=1}^{N-1} (1 - \rho_j) q_j v_j^2 = \\ = \;& \frac{1}{2} \left\{ q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j - v_{j-1})^2 + \sum_{j=1}^{N-1} v_j^2 q_j \left(\rho_j - \frac{q_{j+1}}{q_j} \right) \right\}, \end{aligned}$$

which is nonnegative due to the inequality (3.42). □

In the case when the fractions $\frac{q_{j+1}}{q_j}$ are not smaller then 1, we can use the following lemma.

Lemma 3.4.2. *Let $N \in \mathbb{N}$, $N \geq 8$, $0 \leq \delta < 4$ and q_j, $j = 1, 2, \ldots, N$, are positive numbers satisfying*

$$\frac{q_{j+1}}{q_j} \leq 1 + \frac{\delta}{N^2} \qquad \text{for } j = 1, 2, \ldots, N-1. \quad (3.44)$$

Then for all $v_j \in \mathbb{R}$, $j = 1, 2, \ldots, N$, holds

$$q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j^2 - v_j v_{j-1}) \;\geq\; \frac{4 - \delta}{2N^2} \sum_{j=1}^{N} q_j v_j^2. \quad (3.45)$$

Proof. Applying the Young inequality on the left-hand side of (3.45) yields

$$\begin{aligned} & q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j^2 - v_j v_{j-1}) \;\geq\; q_1 v_1^2 + \sum_{j=2}^{N} q_j \left(v_j^2 - \frac{1}{2\sigma_j} v_{j-1}^2 - \frac{\sigma_j}{2} v_j^2 \right) = \\ = \;& \left(q_1 - \frac{q_2}{2\sigma_2} \right) v_1^2 + \sum_{j=2}^{N-1} \left(q_j \left(1 - \frac{\sigma_j}{2} \right) - \frac{q_{j+1}}{2\sigma_{j+1}} \right) v_j^2 + q_N \left(1 - \frac{\sigma_N}{2} \right) v_N^2, \end{aligned}$$

where σ_j, $j = 2, 3, \ldots N$, are positive numbers. Now we must choose the values σ_j in such a way that all terms in the previous expression are positive. Thus, we take

$$\sigma_j \;=\; 1 + \frac{2}{N}\,\frac{z_{N,j}}{1 - z_{N,j}^2}, \qquad \text{with } z_{N,j} = \frac{2}{N}\left(j - \frac{N+1}{2}\right). \tag{3.46}$$

Then $z_{N,2} = -1 + \frac{3}{N}$, $\sigma_2 = \frac{2}{3} + \frac{1}{2N-3}$ and consequently for $0 \le \delta < 4$

$$\begin{aligned} q_1 - \frac{q_2}{2\sigma_2} \;&=\; q_1\left(1 - \frac{1}{2}\frac{q_2}{q_1}\frac{1}{\sigma_2}\right) \;>\; q_1\left(1 - \frac{1}{2}\left(1 + \frac{\delta}{N^2}\right)\frac{6N-9}{4N-3}\right) = \\ &=\; q_1\left(\frac{2N+3}{8N-6} - \frac{\delta}{N^2}\frac{6N-9}{8N-6}\right) \;>\; q_1\left(\frac{2N+3}{8N-6} - \frac{4(6N-9)}{N^2(8N-6)}\right) \;= \\ &=\; q_1\left(\frac{1}{4} + \frac{9}{16N} - \frac{165}{64N^2} + \frac{657}{32N^2(8N-6)}\right) \;\ge\; \frac{q_1(4-\delta)}{2N^2}, \end{aligned} \tag{3.47}$$

whenever $N \ge 5$. For $j = N$ we have $z_{N,N} = 1 - \frac{1}{N}$, $\sigma_N = 2 - \frac{1}{2N-1}$ and

$$q_N\left(1 - \frac{\sigma_N}{2}\right) \;=\; \frac{1}{4N-2}\,q_N \;>\; \frac{q_N(4-\delta)}{2N^2} \tag{3.48}$$

for $0 \le \delta < 4$ and $N \ge 8$.

The most complicated case occurs when $2 \le j \le N-1$. Then $-1 + \frac{3}{N} \le z_{N,j} \le 1 - \frac{3}{N}$ and $z_{N,j+1} = z_{N,j} + \frac{2}{N}$. Consequently, it holds

$$\begin{aligned} q_j\left(1 - \frac{\sigma_j}{2}\right) - \frac{q_{j+1}}{2\sigma_{j+1}} \;&=\; q_j\left(1 - \frac{\sigma_j}{2} - \frac{1}{2}\frac{q_{j+1}}{q_j}\frac{1}{\sigma_{j+1}}\right) \;> \\ >\; \frac{q_j}{2}&\left(1 - \frac{\frac{2}{N}z_{N,j}}{1 - z_{N,j}^2} - \left(1 + \frac{\delta}{N^2}\right)\frac{1 - \left(z_{N,j} + \frac{2}{N}\right)^2}{1 - \left(z_{N,j} + \frac{2}{N}\right)^2 + \frac{2}{N}\left(z_{N,j} + \frac{2}{N}\right)}\right) \;= \\ =\; \frac{q_j}{2N^2}&\left(\frac{4}{\left(1 - z_{N,j}^2\right)\left(1 - z_{N,j}^2 - \frac{2}{N}z_{N,j}\right)} - \frac{\delta\left[1 - \left(z_{N,j} + \frac{2}{N}\right)^2\right]}{1 - z_{N,j}^2 - \frac{2}{N}z_{N,j}}\right) \;= \\ =\; \frac{q_j}{2N^2}&\left(\frac{4 - \delta + \delta\left[z_{N,j}^2 + (1 - z_{N,j}^2)(z_{N,j} + \frac{2}{N})^2\right]}{\left(1 - \left(z_{N,j} + \frac{2}{N}\right)^2\right)\left(1 - z_{N,j}^2 - \frac{2}{N}z_{N,j}\right)}\right) \;>\; \frac{4-\delta}{2N^2}\,q_j. \end{aligned} \tag{3.49}$$

We have estimated $\left(1 - \left(z_{N,j} + \frac{2}{N}\right)^2\right)\left(1 - z_{N,j}^2 - \frac{2}{N}z_{N,j}\right) \le 1$, where the equality occurs for $z_{N,j} = -\frac{2}{N}$. We have also used the inequality $|z_{N,j}| < 1$. □

Remark 3.4.1. If we take $\delta = 0$ in Lemma 3.4.2, we obtain a factor $\frac{2}{N^2}$ on the right-hand side of (3.45). One can ask whether it is possible to improve this estimate. Let us therefore consider the worst case $q_j = q$ for $j = 1, 2, \ldots, N$. Then

$$q\left(v_1^2 + \sum_{j=2}^{N}(v_j^2 - v_j v_{j-1})\right) = q\left(\boldsymbol{v}_N^T \mathbb{A}_N \boldsymbol{v}_N\right) \ge q\underline{\lambda}_N |\boldsymbol{v}_N|^2, \tag{3.50}$$

where $\boldsymbol{v}_N = (v_1, v_2, \ldots, v_N)$, $\mathbb{A}_N = \text{tridiag}\,\{-\frac{1}{2}, 1, -\frac{1}{2}\}$ and $\underline{\lambda}_N = 1 - \cos(\pi/N)$ is the minimal eigenvalue of $\mathbb{A}_N$.

If we now investigate the behavior of the sequence $\underline{\lambda}_N$ as $N \to +\infty$, we find out that $\lim_{N\to+\infty} \underline{\lambda}_N N^2 = \frac{\pi^2}{2} \approx 4.935$. Thus, the constant in the estimate (3.49) of Lemma 3.4.2 is not optimal, nevertheless, the order is optimal $\left(\frac{1}{N^2}\right)$.

A suboptimal estimate can be achieved considering the discretization of the second order derivative in 1D by piecewise linear finite elements on equidistant partition of the interval $I = (0, 1)$. Then using Friedrichs' inequality (Theorem 4.1.3, page 125) and Lemma 4.2.1 (page 129) we can prove

$$\boldsymbol{v}_N^T \mathbb{A}_N \boldsymbol{v}_N = \frac{1}{2N} |w_h|^2_{1,I} \geq \frac{\pi^2}{2N} \sum_{j=1}^{N} \|w_h\|^2_{0,I_j} \geq \frac{\pi^2}{2N} \frac{h}{3} |\boldsymbol{v}_N|^2 = \frac{\pi^2}{6N^2} |\boldsymbol{v}_N|^2, \qquad (3.51)$$

where $w_h \in H^1_0(I)$ is a piecewise linear function satisfying $w_h(ih) = v_i$ for $i = 1, 2, \ldots, N$.

The upper bound $\delta < 4$ is not optimal as well. However, if we consider $N = 5$, $\frac{q_{j+1}}{q_j} = 1 + \frac{25}{3}\frac{1}{N^2} = \frac{4}{3}$ for $j = 1, 2, 3, 4$, then $q_1 v_1^2 + \sum_{j=2}^{5} q_j v_j^2 = 0$ for $(v_1, v_2, v_3, v_4, v_5) = (1, \sqrt{3}, 2, \sqrt{3}, 1)$. Hence, the optimal upper bound for δ is not greater than $\frac{25}{3}$. (If we consider only values $N \geq 8$ as in the previous lemma, then we can construct similar example and deduce that the optimal upper bound for δ has to be smaller than approximately $8.478 > \frac{25}{3}$.)

In the previous two lemmas we estimated the left-hand side by the sum that corresponds to the L^2-norm. We would also like to estimate it by the sum that corresponds to the norm of the derivatives in the flow direction. For this purpose we use the next two lemmas.

Lemma 3.4.3. *Let $N \in \mathbb{N}$ and q_j, $j = 1, 2, \ldots, N$, are positive numbers satisfying*

$$\frac{q_{j+1}}{q_j} \leq 1 \qquad \text{for } j = 1, 2, \ldots, N-1. \qquad (3.52)$$

Then for all $v_j \in \mathbb{R}$, $j = 1, 2, \ldots, N$, holds

$$q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j^2 - v_j v_{j-1}) \;\geq\; \frac{1}{2} \left\{ q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j - v_{j-1})^2 \right\}. \qquad (3.53)$$

Proof. Subtracting the right-hand side of (3.53) from the left-hand side we obtain

$$\begin{aligned} & q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j^2 - v_j v_{j-1}) - \frac{1}{2} \left\{ q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j - v_{j-1})^2 \right\} = \\ &= \frac{1}{2} q_1 v_1^2 + \frac{1}{2} \sum_{j=2}^{N} q_j \left(v_j^2 - v_{j-1}^2 \right) = \frac{1}{2} q_N v_N^2 + \frac{1}{2} \sum_{j=1}^{N-1} v_j^2 (q_j - q_{j+1}) \geq 0. \end{aligned}$$

□

Lemma 3.4.4. *Let $N \in \mathbb{N}$, $N \geq 8$, $0 \leq \delta < 4$ and q_j, $j = 1, 2, \ldots, N$, are positive numbers satisfying*

$$\frac{q_{j+1}}{q_j} \leq 1 + \frac{\delta}{N^2} \qquad \text{for } j = 1, 2, \ldots, N-1. \qquad (3.54)$$

Then for all $v_j \in \mathbb{R}$, $j = 1, 2, \ldots, N$, holds

$$q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j^2 - v_j v_{j-1}) \;\geq\; \frac{4-\delta}{8} \left\{ q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j - v_{j-1})^2 \right\}. \qquad (3.55)$$

Proof. Denoting $\alpha = \frac{1}{4}\delta \in [0,1)$ and using Young's inequality we can write

$$\begin{aligned}
&q_1 v_1^2 + \sum_{j=2}^{N} q_j \left(v_j^2 - v_j v_{j-1}\right) \;= \\
&= \; q_1 v_1^2 + \sum_{j=2}^{N} q_j v_j^2 - \sum_{j=2}^{N} q_j (1-\alpha) v_j v_{j-1} - \sum_{j=2}^{N} q_j \alpha v_j v_{j-1} \;\geq \\
&\geq \; q_1 v_1^2 + \sum_{j=2}^{N} q_j v_j^2 - \sum_{j=2}^{N} q_j (1-\alpha) v_j v_{j-1} - \frac{1}{2}\sum_{j=2}^{N} q_j \alpha \left(\frac{v_{j-1}^2}{\sigma_j} + v_j^2 \sigma_j\right) \;= \\
&= \; q_1 v_1^2 + \sum_{j=2}^{N} q_j v_j^2 + \frac{1}{2}\sum_{j=2}^{N} q_j (1-\alpha)(v_j - v_{j-1})^2 - \\
&\quad - \frac{1}{2}\sum_{j=2}^{N} q_j (1-\alpha)\left(v_j^2 + v_{j-1}^2\right) - \frac{1}{2}\sum_{j=2}^{N} q_j \alpha \left(\frac{v_{j-1}^2}{\sigma_j} + v_j^2 \sigma_j\right) \;= \\
&= \; \frac{1-\alpha}{2}\left\{q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j - v_{j-1})^2\right\} + \left\{1 + \alpha - \frac{q_2}{q_1}\left(1 - \alpha + \frac{\alpha}{\sigma_2}\right)\right\}\frac{q_1 v_1^2}{2} + \\
&\quad + \frac{1}{2}\sum_{j=2}^{N-1}\left\{1 + \alpha - \alpha\sigma_j - \frac{q_{j+1}}{q_j}\left(1 - \alpha + \frac{\alpha}{\sigma_{j+1}}\right)\right\} q_j v_j^2 + \\
&\quad + \left\{1 + \alpha - \alpha\sigma_N\right\}\frac{q_N v_N^2}{2}. \qquad (3.56)
\end{aligned}$$

In order to complete the proof we have to show that the last three terms are nonnegative. We use the same definition of σ_j as in Lemma 3.4.2. Let us begin with the terms in the sum and use the estimate (3.49), then

$$\begin{aligned}
&1 + \alpha - \alpha\sigma_j - \frac{q_{j+1}}{q_j}\left(1 - \alpha + \frac{\alpha}{\sigma_{j+1}}\right) \;\geq \qquad (3.57) \\
&\geq \; 1 + \alpha - \alpha\sigma_j - \left(1 + \frac{\delta}{N^2}\right)\left(1 - \alpha + \frac{\alpha}{\sigma_{j+1}}\right) \;= \\
&= \; 2\alpha\left(1 - \frac{\sigma_j}{2} - \frac{1}{2}\left(1 + \frac{\delta}{N^2}\right)\frac{1}{\sigma_{j+1}}\right) - \frac{\delta\,(1-\alpha)}{N^2} \geq 2\alpha\frac{4-\delta}{2N^2} - \frac{\delta\,(1-\alpha)}{N^2} = 0.
\end{aligned}$$

Further, using the estimate (3.47) we obtain

$$\begin{aligned}
&1 + \alpha - \frac{q_2}{q_1}\left(1 - \alpha + \frac{\alpha}{\sigma_2}\right) \;\geq\; 1 + \alpha - \left(1 + \frac{\delta}{N^2}\right)\left(1 - \alpha + \frac{\alpha}{\sigma_2}\right) \;= \qquad (3.58) \\
&= \; 2\alpha\left(1 - \frac{1}{2}\left(1 + \frac{\delta}{N^2}\right)\frac{1}{\sigma_2}\right) - \frac{\delta}{N^2}(1-\alpha) \;\geq\; 2\alpha\frac{4-\delta}{2N^2} - \frac{\delta}{N^2}(1-\alpha) \;=\; 0,
\end{aligned}$$

whenever $N \geq 5$. Finally, for $N \geq 8$ and using (3.48) we have

$$\begin{aligned}
1 + \alpha - \alpha\sigma_N \;&=\; 1 - \alpha + 2\alpha\left(1 - \frac{\sigma_N}{2}\right) \;\geq\; 1 - \alpha + 2\alpha\frac{4-\delta}{2N^2} \;= \\
&=\; \left(1 + \frac{\delta}{N^2}\right)(1-\alpha) \;\geq\; 0. \qquad (3.59)
\end{aligned}$$

□

3.4.2 Coercivity estimates

In the case when $\operatorname{div} \boldsymbol{b} < 0$, we use Lemma 3.3.1 for proving the coercivity of the bilinear form a_h with respect to the energy norm $|||\cdot|||_b$. (See Definition 4.1.1, page 124, for the definition of other norms.)

Definition 3.4.1. *When* $\operatorname{div} \boldsymbol{b} \leq -\omega < 0$ *we estimate the error of the presented method in the energy norm*

$$|||v|||_b^2 \;=\; \varepsilon |v|_{1,\Omega}^2 + \frac{\omega\kappa}{2}\|v\|_{0,\Omega}^2 + \sum_{K\in\mathcal{T}_h} \frac{|\boldsymbol{d}_{K,1}|}{2|\boldsymbol{b}_K|}\|\boldsymbol{b}_K\cdot\nabla v\|_{0,K}^2, \tag{3.60}$$

where $\kappa = \min\limits_{j,s}\left\{\frac{|\mathcal{C}_j^s|}{|\Omega_j^s|}, \frac{|\mathcal{C}_{j+1}^s|}{|\Omega_j^s|}\right\}$, $|\boldsymbol{d}_{K,1}| = P_{K,3} - P_{K,1}$ *and* $\boldsymbol{b}_K$ *is defined in (3.13) (page 56).*

Theorem 3.4.1 ($\operatorname{div}\boldsymbol{b} < 0$)**.** *Let the assumptions of Lemma 3.3.1 be fulfilled. Further, let there exists a constant* κ *independent of* h *(and* ε*) such that* $\frac{|\mathcal{C}_{j+1}^s|}{|\Omega_j^s|} \geq \kappa$ *for all* $s = 1, 2, \ldots, \mathcal{P}$ *and* $j = 1, 2, \ldots, N_s$*. Then the bilinear form defined in (3.15) satisfies*

$$a_h(v_h, v_h) \;\geq\; \frac{1}{2}\,|||v|||_b^2. \tag{3.61}$$

Proof. Combining (3.41) together with (3.30) and (3.33) we realize that

$$a_h(v_h, v_h) \;=\; \varepsilon |v_h|_{1,\Omega}^2 + \sum_{s=1}^{\mathcal{P}}\sum_{j=1}^{N_s} q_j^s\, v_h(P_j^s)(v_h(P_j^s) - v_h(P_{j-1}^s)) \tag{3.62}$$

and it remains to estimate the latter term. In order to do so, we use the inequality from Lemma 3.3.1

$$\frac{q_{j+1}^s}{q_j^s} \;\leq\; 1 - \frac{\omega|\mathcal{C}_{j+1}^s|}{(n+1)q_j^s} \;<\; 1. \tag{3.63}$$

The inequality (3.43) then implies

$$\begin{aligned}\sum_{s=1}^{\mathcal{P}}\sum_{j=1}^{N_s} q_j^s\, v_h(P_j^s)(v_h(P_j^s) - v_h(P_{j-1}^s)) \;&\geq\; \sum_{s=1}^{\mathcal{P}}\sum_{j=1}^{N_s} \frac{\omega\,|\mathcal{C}_{j+1}^s|}{2(n+1)} v_h^2(P_j^s) \;\geq \\ \geq\; \sum_{s=1}^{\mathcal{P}}\sum_{j=1}^{N_s} \frac{\omega\,\kappa\,|\Omega_j^s|}{2(n+1)} v_h^2(P_j^s) &= \sum_{K\in\mathcal{T}_h} \frac{\omega\kappa\,|K|}{2(n+1)} \sum_{i=1}^{n+1} v_h^2(P_{K,i}) \geq \frac{\omega\kappa}{2}\|v_h\|_{0,\Omega}^2,\end{aligned} \tag{3.64}$$

where we used the inequality $\|v_h\|_{0,K}^2 \leq \frac{|K|}{n+1}\sum_{i=1}^{n+1} v_h^2(P_{K,i})$ (cf. Lemma 4.2.1), and the fact that

$$\sum_{s=1}^{\mathcal{P}}\sum_{j=1}^{N_s} |\Omega_j^s| v_h^2(P_j^s) \;=\; \sum_{s=1}^{\mathcal{P}}\sum_{j=1}^{N_s}\sum_{K\subset\Omega_j^s} |K| v_h^2(P_j^s) \;=\; \sum_{K\in\mathcal{T}_h} |K| \sum_{i=1}^{n+1} v_h^2(P_{K,i}). \tag{3.65}$$

Similarly, using the inequality (3.53) we obtain

$$\begin{aligned}\sum_{K\in\mathcal{T}_h} (\boldsymbol{b}_K\cdot\nabla v_h, v_h(P_{K,n+1}))_K \;&=\; \sum_{s=1}^{\mathcal{P}}\sum_{j=1}^{N_s} v_h(P_j^s)(v_h(P_j^s) - v_h(P_{j-1}^s))q_j^s \;\geq \\ \geq\; \frac{1}{2}\sum_{s=1}^{\mathcal{P}}\sum_{j=1}^{N_s} q_j^s\Big(v_h(P_j^s) - v_h(P_{j-1}^s)\Big)^2 \;&=\; \frac{1}{2}\sum_{s=1}^{\mathcal{P}}\sum_{j=1}^{N_s}\sum_{K\subset\mathcal{C}_j^s} \frac{h_j^s}{|\boldsymbol{b}_K|}\,\|\boldsymbol{b}_K\cdot\nabla v_h\|_{0,K}^2 \;= \\ = \sum_{K\in\mathcal{T}_h} \frac{|\boldsymbol{d}_{K,1}|}{2|\boldsymbol{b}_K|}\,\|\boldsymbol{b}_K\cdot\nabla v_h\|_{0,K}^2,\end{aligned} \tag{3.66}$$

which results from the equality

$$\sum_{K\subset\mathcal{C}_j^s}\frac{h_j^s}{|\boldsymbol{b}_K|}\,\|\boldsymbol{b}_K\cdot\nabla v_h\|_{0,K}^2 \;=\; \sum_{K\subset\mathcal{C}_j^s}\frac{h_j^s}{|\boldsymbol{b}_K|}\,|K|\,\frac{|\boldsymbol{b}_K|^2}{|\boldsymbol{d}_{K,1}|^2}\,\left(\sum_{i=1}^{n+1}\boldsymbol{d}_{K,1}\nabla v_h\,(P_{K,i})\right)^2 =$$
$$= \left(v_h(P_j^s)-v_h(P_{j-1}^s)\right)^2\sum_{K\subset\mathcal{C}_j^s}\frac{h_j^s}{|\boldsymbol{b}_K|}\,|K|\,\frac{|\boldsymbol{b}_K|^2}{(h_j^s)^2}. \quad (3.67)$$

Summing halves of the inequalities (3.64) and (3.66) completes the proof. □

In order to prove the coercivity of the bilinear form a_h in the case when $\operatorname{div}\boldsymbol{b}=0$ we use another auxiliary quantities.

Definition 3.4.2. *For each node P_j^s, $s=1,2,\ldots,\mathcal{P}$ and $j=1,2,\ldots,N_s$, let us define the function $\overline{\sigma}(P_j^s)$ by the relation*

$$\overline{\sigma}(P_j^s) \;=\; \frac{|\mathcal{C}_j^s|+|\Omega_{0,j}^s|}{|\mathcal{C}_j^s|}\,(h_jN_s)^2\,\frac{\|\boldsymbol{b}\|_{\infty,\Omega_j^s}}{\beta_j^s}\,\frac{\max\limits_{K\subset\Omega_j^s}h_K}{h_j^s}. \quad (3.68)$$

Further, for each element $K\in\mathcal{T}_h$ we define the value $\overline{\nu}_K=\max\limits_{1\le i\le n+1}\overline{\sigma}(P_{K,i})$.

In the case when $\boldsymbol{b}$ is a constant vector, $h_jN_s=L$ and for the mesh considered in the Remark 3.3.2 it holds $|\mathcal{C}_j^s|=|K|n!$, $|\Omega_{0,j}^s|=|K|(n-1)n!$ and consequently

$$\overline{\sigma}(P_j^s)\;\approx\;\frac{|K|n!+|K|(n-1)n!}{|K|n!}L^2\;=\;nL^2 \qquad \text{for all possible } j,s. \quad (3.69)$$

For more general data we obtain different values of $\overline{\sigma}(P_j^s)$ or $\overline{\nu}_K$, however, the value (3.69) is still a good approximation, in particular for quasi-equidistant meshes.

We use these quantities together with the Lemmas 3.4.2 and 3.4.4 and prove the coercivity of the method with respect to the appropriate energy norm. At first, we again find a relation between q_j^s and q_{j+1}^s.

Lemma 3.4.5. *Let $\operatorname{div}\boldsymbol{b}=0$ in Ω and let there exists $\delta\ge 0$ such that*

$$\theta_K\;\le\;\frac{\delta}{\overline{\nu}_K}\,\|\boldsymbol{b}\|_{\infty,K}\,h_K \qquad \textit{for each } K\in\mathcal{T}_h. \quad (3.70)$$

Then for each $s=1,2,\ldots,\mathcal{P}$ and $j=1,2,\ldots,N_s$ it holds

$$\frac{q_{j+1}^s}{q_j^s}\;\le\;1+\frac{\delta}{N_s^2}. \quad (3.71)$$

Proof. At first we observe that for each $K\subset\Omega_j^s$ it holds $\overline{\sigma}(P_j^s)\le\max\limits_{1\le i\le n+1}\overline{\sigma}(P_{K,i})$. This is due to the fact that P_j^s belongs to each $K\subset\Omega_j^s$. Consequently

$$\overline{\sigma}(P_j^s)\le\min_{K\in\Omega_j^s}\max_{1\le i\le n+1}\overline{\sigma}(P_{K,i})=\min_{K\in\Omega_j^s}\overline{\nu}_K. \quad (3.72)$$

We can use this inequality, the equality from (3.35) and estimate

$$\begin{aligned}
\frac{q_{j+1}^s}{q_j^s} &= 1+\frac{1}{q_j^s}\left\{\int_{\Omega_{0,j}^s}\boldsymbol{b}\cdot\nabla\lambda_j^s\,\mathrm{d}x-\sum_{K\subset\mathcal{C}_j^s}\sum_{i=2}^{n}\int_K\boldsymbol{b}\cdot\nabla\lambda_{K,i}\,\mathrm{d}x\right\}\ \leq\\
&\leq 1+\frac{h_j^s}{\beta_j^s|\mathcal{C}_j^s|}\left\{\sum_{K\subset\Omega_{0,j}^s}\left|\int_K\boldsymbol{b}\cdot\nabla\lambda_j^s\,\mathrm{d}x\right|+\sum_{K\subset\mathcal{C}_j^s}\left|\sum_{i=2}^{n}\int_K\boldsymbol{b}\cdot\nabla\lambda_{K,i}\,\mathrm{d}x\right|\right\}\ \leq\\
&\leq 1+\frac{h_j^s}{\beta_j^s|\mathcal{C}_j^s|}\left\{\sum_{K\subset\Omega_{0,j}^s}\frac{\delta}{\overline{\nu}_K}\,\|\boldsymbol{b}\|_{\infty,K}\,h_K\,|K|+\sum_{K\subset\mathcal{C}_j^s}\frac{\delta}{\overline{\nu}_K}\,\|\boldsymbol{b}\|_{\infty,K}\,h_K\,|K|\right\}\ \leq\\
&\leq 1+\frac{h_j^s\,\delta}{\beta_j^s|\mathcal{C}_j^s|}\left\{\frac{\max\limits_{K\subset\Omega_{0,j}^s}h_K}{\min\limits_{K\in\Omega_{0,j}^s}\overline{\nu}_K}\,\|\boldsymbol{b}\|_{\infty,\Omega_{0,j}^s}\,|\Omega_{0,j}^s|+\frac{\max\limits_{K\subset\mathcal{C}_j^s}h_K}{\min\limits_{K\subset\mathcal{C}_j^s}\overline{\nu}_K}\,\|\boldsymbol{b}\|_{\infty,\mathcal{C}_j^s}\,|\mathcal{C}_j^s|\right\}\ \leq\\
&\leq 1+\frac{\delta}{N_s^2}\,\frac{\overline{\sigma}(P_j^s)}{\min\limits_{K\in\Omega_j^s}\overline{\nu}_K}\left\{\frac{|\Omega_{0,j}^s|}{|\Omega_{0,j}^s|+|\mathcal{C}_j^s|}+\frac{|\mathcal{C}_j^s|}{|\Omega_{0,j}^s|+|\mathcal{C}_j^s|}\right\}\ \leq\ 1+\frac{\delta}{N_s^2},
\end{aligned}\tag{3.73}$$

where we used the fact that the minimum (maximum) over larger set does not increase (decrease). □

Remark 3.4.2 ($\operatorname{div}\boldsymbol{b}=0$). In the case when $\operatorname{div}\boldsymbol{b}=0$ in Ω and the assumptions of Lemma 3.4.5 are fulfilled Lemma 3.4.4 provides an inequality analogous to the estimate (3.66)

$$\sum_{K\in\mathcal{T}_h}(\boldsymbol{b}_K\cdot\nabla v_h,v_h(P_{K,n+1}))_K\ \geq\ \frac{4-\delta}{4}\sum_{K\in\mathcal{T}_h}\frac{|\boldsymbol{d}_{K,1}|}{2|\boldsymbol{b}_K|}\,\|\boldsymbol{b}_K\cdot\nabla v_h\|_{0,K}^2.\tag{3.74}$$

When $\operatorname{div}\boldsymbol{b}=0$ we have to estimate the error of the method with respect to the different type of energy norm than $|||\cdot|||_b$. Hence, we define the energy norm $|||\cdot|||_{b,*}$.

Definition 3.4.3. *Let us define the constants β, L and R by the relations*

$$\beta=\min_{j,s}\{\beta_j^s\},\quad L=\max_{j,s}\{N_s h_j^s\}\quad\textit{and}\quad R=\max_{j,s}\left\{\frac{\max\limits_{K\subset\Omega_j^s}h_K}{h_j^s}\right\}.\tag{3.75}$$

Further, let $\delta\in[0,4)$ be any constant satisfying (3.70) for all $K\in\mathcal{T}_h$. Then the energy norm used in the case of divergence-free vector field $\boldsymbol{b}$ is defined as

$$|||v|||_{b,*}^2\ =\ \varepsilon|v|_{1,\Omega}^2+C_2^*\sum_{K\in\mathcal{T}_h}h_K\|v\|_{0,K}^2+C_b^*\sum_{K\in\mathcal{T}_h}\frac{|\boldsymbol{d}_{K,1}|}{2|\boldsymbol{b}_K|}\|\boldsymbol{b}_K\cdot\nabla v\|_{0,K}^2,\tag{3.76}$$

where $C_2^*=\frac{(4-\delta)\kappa\beta}{2L^2R}(n+1)$ and $C_b^*=\frac{4-\delta}{4}$.

Theorem 3.4.2 ($\operatorname{div}\boldsymbol{b}=0$)**.** *Let the assumptions of Lemma 3.4.5 be fulfilled. Then for each $v_h\in V_h$ there holds*

$$a_h(v_h,v_h)\ \geq\ \frac{1}{2}\,|||v|||_{b,*}^2.\tag{3.77}$$

Proof. Again, we rewrite the convective bilinear form as the sum of streamlines and clusters and use the inequalities (3.45), (3.71) from Lemmas 3.4.2 and 3.4.5

$$\begin{aligned}
\sum_{K\in\mathcal{T}_h}(\boldsymbol{b}_K\cdot\nabla v_h, v_h(P_{K,n+1}))_K &= \sum_{s=1}^{\mathcal{P}}\sum_{j=1}^{N_s} v_h(P_j^s)(v_h(P_j^s)-v_h(P_{j-1}^s))q_j^s \geq \\
&\geq \frac{4-\delta}{2}\sum_{s=1}^{\mathcal{P}}\frac{1}{N_s^2}\sum_{j=1}^{N_s} q_j^s v_h^2(P_j^s) = \frac{4-\delta}{2}\sum_{s=1}^{\mathcal{P}}\sum_{j=1}^{N_s}\frac{\beta_j^s|\mathcal{C}_j^s|}{(N_s h_j^s)^2}h_j^s v_h^2(P_j^s) \geq \\
&\geq \frac{(4-\delta)\kappa\beta}{2L^2}\sum_{s=1}^{\mathcal{P}}\sum_{j=1}^{N_s}|\Omega_j^s|h_j^s v_h^2(P_j^s) \geq \frac{(4-\delta)\kappa\beta}{2L^2R}\sum_{K\in\mathcal{T}_h}h_K|K|\sum_{i=1}^{n+1}v_h^2(P_{K,i}) \geq \\
&\geq \frac{(4-\delta)\kappa\beta}{2L^2R}(n+1)\sum_{K\in\mathcal{T}_h}h_K\|v_h\|_{0,K}^2. \qquad (3.78)
\end{aligned}$$

Combining the estimates (3.74) and (3.78) completes the proof. □

3.5 Error analysis

In this section we recall the error analysis of the standard SUPG method and then use a similar approach for analysis of the error of the presented method.

3.5.1 SUPG method error analysis

We follow the error analysis of the SUPG method presented in Roos et al. (2008). However, in contrast to Roos et al. (2008) we use the same stabilization parameter in both convection-dominated and diffusion-dominated case. Consequently, the error estimate of the same order as in Roos et al. (2008) is derived.

In order to be more general we (just in this case) consider the convection-diffusion-reaction equation

$$-\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u + cu = f \qquad \text{in } \Omega \qquad (3.79)$$

equipped with the Dirichlet boundary condition $u=0$ on $\partial\Omega$. Further, we assume that $\boldsymbol{b}\in W^{1,\infty}(\Omega)^n$, $c\in L^\infty(\Omega)$, $f\in L^2(\Omega)$ and $c-\frac{1}{2}\mathrm{div}\,\boldsymbol{b}\geq\omega>0$ in Ω.

The finite element space $V_h^{(k)}\subset V=H_0^1(\Omega)$ is defined as

$$V_h^{(k)} = \{\varphi_h\in V,\ \varphi_h|_K\in P_k(K)\text{ for all } K\in\mathcal{T}_h\}, \qquad (3.80)$$

where the triangulation $\mathcal{T}_h$ is assumed to be shape-regular (cf. Theorem 4.2.2).

For each $K\in\mathcal{T}_h$ let δ_K be any positive numbers (specified later), then the SUPG solution $u_{SU}\in V_h^{(k)}$ satisfies

$$a_{SU}(u_{SU},\varphi_h) = \sum_{K\in\mathcal{T}_h}(f,\varphi_h+\delta_K\boldsymbol{b}\cdot\nabla\varphi_h)_K \qquad \text{for all } \varphi_h\in V_h^{(k)}, \qquad (3.81)$$

where the bilinear form a_{SU} is for all $v\in H_0^1(\Omega)\cap H^2(\Omega)$ and $\varphi\in H_0^1(\Omega)$ defined by the relation

$$\begin{aligned}
a_{SU}(v,\varphi) &= \varepsilon(\nabla v,\nabla\varphi)_\Omega + (\boldsymbol{b}\cdot\nabla v,\varphi)_\Omega + (cv,\varphi)_\Omega + \\
&\quad + \sum_{K\in\mathcal{T}_h}\delta_K(-\varepsilon\Delta v+\boldsymbol{b}\cdot\nabla v+cv,\boldsymbol{b}\cdot\nabla\varphi)_K. \qquad (3.82)
\end{aligned}$$

We will measure the stability and the error of the SUPG method in the norm $|||\cdot|||_{SU}$ which is for each $v \in H^1(\Omega)$ defined as

$$|||v|||_{SU} \;=\; \left(\varepsilon|v|^2_{1,\Omega} + \omega\|v\|^2_{0,\Omega} + \sum_{K\in\mathcal{T}_h}\delta_K\|\boldsymbol{b}\cdot\nabla v\|^2_{0,K}\right)^{1/2}. \tag{3.83}$$

The stability of the SUPG method with respect to the $|||\cdot|||_{SU}$ provides the following lemma.

Lemma 3.5.1. *Let us assume that* $\delta_K \le \min\left\{\frac{\omega}{\|c\|^2_{0,\infty,K}}, \frac{h_K}{2C_{inv}\|\boldsymbol{b}\|_{0,\infty,K}}\right\}$ *for all* $K \in \mathcal{T}_h$. *Then for all functions* $v_h \in V_h^{(k)}$ *holds*

$$a_{SU}(v_h, v_h) \;\ge\; \frac{1}{2}|||v_h|||^2_{SU}. \tag{3.84}$$

Proof. We use the inverse inequality (Theorem 4.2.2), the Young's inequality together with the bound for δ_K and estimate

$$\begin{aligned}
a_{SU}(v_h, v_h) \;&=\; \varepsilon|v_h|^2_{1,\Omega} + \left(c - \tfrac{1}{2}\mathrm{div}\,\boldsymbol{b}, v_h^2\right)_\Omega + \sum_{K\in\mathcal{T}_h}\delta_K\|\boldsymbol{b}\cdot\nabla v_h\|^2_{0,K} + \\
&\qquad + \sum_{K\in\mathcal{T}_h}\delta_K(-\varepsilon\Delta v_h + cv_h, \boldsymbol{b}\cdot\nabla v_h)_K \;\ge \\
&\ge\; \varepsilon|v_h|^2_{1,\Omega} + \omega\|v_h\|^2_{0,\Omega} + \sum_{K\in\mathcal{T}_h}\delta_K\|\boldsymbol{b}\cdot\nabla v_h\|^2_{0,K} - \\
&\qquad - \sum_{K\in\mathcal{T}_h}\delta_K(\varepsilon\tfrac{C_{inv}}{h_K}|v_h|_{1,K} + \|c\|_{0,\infty,K}\|v_h\|_{0,K})\|\boldsymbol{b}\cdot\nabla v_h\|_{0,K} \;\ge \\
&\ge\; \varepsilon|v_h|^2_{1,\Omega} + \omega\|v_h\|^2_{0,\Omega} + \sum_{K\in\mathcal{T}_h}\delta_K\|\boldsymbol{b}\cdot\nabla v_h\|^2_{0,K} - \sum_{K\in\mathcal{T}_h}\delta_K\varepsilon\tfrac{C_{inv}}{h_K}|v_h|^2_{1,K}\|\boldsymbol{b}\|_{0,\infty,K} - \\
&\qquad - \frac{1}{2}\sum_{K\in\mathcal{T}_h}\delta_K\|c\|^2_{0,\infty,K}\|v_h\|^2_{0,K} - \frac{1}{2}\sum_{K\in\mathcal{T}_h}\delta_K\|\boldsymbol{b}\cdot\nabla v_h\|^2_{0,K} \;\ge\; \frac{1}{2}|||v_h|||^2_{SU}.
\end{aligned} \tag{3.85}$$

□

In the derivation of the error estimate it is necessary to estimate the derivative in the flow direction. Thus, we use the inequality

$$\|\boldsymbol{b}\cdot\nabla v\|_{0,K} \;\le\; \min\left\{\frac{1}{\delta_K^{1/2}}, \frac{\|\boldsymbol{b}\|_{0,\infty,K}}{\varepsilon^{1/2}}\right\}\left(\varepsilon|v|^2_{1,K} + \omega\|v\|^2_{0,K} + \delta_K\|\boldsymbol{b}\cdot\nabla v\|^2_{0,K}\right)^{1/2}. \tag{3.86}$$

Due to the consistency of the method we obtain for all $u \in H^2(\Omega)$ the Galerkin orthogonality

$$a_{SU}(u - u_{SU}, v_h) \;=\; 0 \qquad \text{for all } v_h \in V_h^{(k)}. \tag{3.87}$$

If we want to derive the error estimate of the SUPG method, we use a $V_h^{(k)}$-interpolant u^I of the function u (see Definition 4.2.2, page 126) and decompose the error $e_h = u - u_{SU} = (u - u^I) + (u^I - u_{SU}) = \eta_h + \xi_h$ into the approximation

error $\eta_h = u - u^I$ and the error of the method $\xi_h = u^I - u_{SU} \in V_h^{(k)}$. The coercivity (3.84) and the Galerkin orthogonality (3.87) then implies

$$\begin{aligned}
\frac{1}{2}|||\xi_h|||_{SU}^2 \;\le\; a_{SU}(\xi_h,\xi_h) \;=\; a_{SU}(-\eta_h,\xi_h) \;=\\
=\; -\varepsilon(\nabla\eta_h,\nabla\xi_h)_\Omega + (\eta_h \mathrm{div}\,\boldsymbol{b},\xi_h)_\Omega + (\eta_h,\boldsymbol{b}\cdot\nabla\xi_h)_\Omega - (c\eta_h,\xi_h)_\Omega -\\
- \sum_{K\in\mathcal{T}_h}\delta_K(-\varepsilon\Delta\eta_h + \boldsymbol{b}\cdot\nabla\eta_h + c\eta_h, \boldsymbol{b}\cdot\nabla\xi_h)_K \;\le\\
\le \sum_{K\in\mathcal{T}_h}\varepsilon|\eta_h|_{1,K}|\xi_h|_{1,K} + \sum_{K\in\mathcal{T}_h}\Big(n\|\boldsymbol{b}\|_{1,\infty,K} + \|c\|_{0,\infty,K}\Big)\|\eta_h\|_{0,K}\|\xi_h\|_{0,K} +\\
+\sum_{K\in\mathcal{T}_h}\Big[\|\eta_h\|_{0,K} + \delta_K\Big(\varepsilon n|\eta_h|_{2,K} + \|\boldsymbol{b}\|_{0,\infty,K}|\eta_h|_{1,K} + \|c\|_{0,\infty,K}\|\eta_h\|_{0,K}\Big)\Big]\|\boldsymbol{b}\cdot\nabla\xi_h\|_{0,K}.
\end{aligned}$$

Now we set $\mu_K = n\|\boldsymbol{b}\|_{1,\infty,K} + \|c\|_{0,\infty,K}$ for all $K \in \mathcal{T}_h$, use the estimate (3.86), Corollary 4.2.2 with the discrete Cauchy–Schwarz inequality and obtain

$$\begin{aligned}
\frac{1}{2}|||\xi_h|||_{SU}^2 \;\le\; |||\xi_h|||_{SU}\;\Bigg(6\sum_{K\in\mathcal{T}_h}\Bigg\{\varepsilon|\eta_h|_{1,K}^2 + \frac{\mu_K^2}{\omega}\|\eta_h\|_{0,K}^2 + \Big[\|\eta_h\|_{0,K}^2 +\\
+\delta_K^2\Big(\varepsilon^2|\eta_h|_{2,K}^2 + \|\boldsymbol{b}\|_{0,\infty,K}^2|\eta_h|_{1,K}^2 + \|c\|_{0,\infty,K}^2\|\eta_h\|_{0,K}^2\Big)\Big]\min\Big\{\frac{1}{\delta_K},\frac{\|\boldsymbol{b}\|_{0,\infty,K}^2}{\varepsilon}\Big\}\Bigg\}\Bigg)^{1/2}.
\end{aligned} \tag{3.88}$$

The interpolation inequality (Theorem 4.2.1, page 127) then provides for $u \in H^{r+1}(\Omega)$, $r \le k$ and $m \in \{0,1,2\}$ the estimate

$$\left(\sum_{K\in\mathcal{T}_h}\|u-u^I\|_{m,K}^2\right)^{1/2} \;\le\; C_X\left(\sum_{K\in\mathcal{T}_h}h_K^{2(r+1-m)}|u|_{r+1,K}^2\right)^{1/2}. \tag{3.89}$$

Dividing by $|||\xi_h|||_{SU}$ and using (3.89) the inequality (3.88) changes into

$$\begin{aligned}
|||\xi_h|||_{SU} \;\le\; \Bigg(24\sum_{K\in\mathcal{T}_h}C_X^2h_K^{2r}|u|_{r+1,K}\Bigg(\varepsilon + \frac{\mu_K^2}{\omega}h_K^2 + \qquad (3.90)\\
+\Big[h_K^2 + \delta_K^2\Big(\frac{\varepsilon^2}{h_K^2} + \|\boldsymbol{b}\|_{0,\infty,K}^2 + \|c\|_{0,\infty,K}^2h_K^2\Big)\Big]\min\Big\{\frac{1}{\delta_K},\frac{\|\boldsymbol{b}\|_{0,\infty,K}^2}{\varepsilon}\Big\}\Bigg)\Bigg)^{1/2}.
\end{aligned}$$

Now we choose the value of δ_K in order to obtain the best possible order of convergence. Choosing $\delta_K = \frac{h_K}{2C_{inv}\|\boldsymbol{b}\|_{0,\infty,K}} \le \frac{\omega}{\|c\|_{0,infty,K}^2}$ leads to estimates

$$\begin{aligned}
&\text{for } \mathrm{Pe}_K \ge C_{inv}: \qquad |||\xi_h|||_{SU} \;\le\; C_{SU}^\xi\left(\sum_{K\in\mathcal{T}_h}\frac{\|\boldsymbol{b}\|_{0,\infty,K}}{2}h_K^{2r+1}|u|_{r+1,K}^2\right)^{1/2},\\
&\text{for } \mathrm{Pe}_K < C_{inv}: \qquad |||\xi_h|||_{SU} \;\le\; C_{SU}^\xi\left(\sum_{K\in\mathcal{T}_h}C_{inv}\,\varepsilon h_K^{2r}|u|_{r+1,K}^2\right)^{1/2},
\end{aligned} \tag{3.91}$$

where $C_{SU}^\xi = \frac{C_X}{\sqrt{C_{inv}}}\left(2 + \frac{1}{4C_{inv}^2} + 8C_{inv}^2\left(1 + n\max_{K\in\mathcal{T}_h}\left\{\frac{|\boldsymbol{b}|_{1,\infty,K}}{\|c\|_{0,\infty,K}}\right\}\right)^2\right)^{1/2}$ and the Péclet number is defined as $\mathrm{Pe}_K = \frac{\|\boldsymbol{b}\|_{0,\infty,K}h_K}{2\varepsilon}$. From these inequalities it follows that

$$|||\xi_h|||_{SU} \;\le\; C_{SU}^\xi\left(\sum_{K\in\mathcal{T}_h}\max\{C_{inv},\mathrm{Pe}_K\}\varepsilon h_K^{2r}|u|_{r+1,K}^2\right)^{1/2}. \tag{3.92}$$

Since for the approximation error holds

$$|||\eta_h|||_{SU} \leq \left(\sum_{K\in\mathcal{T}_h} C_X^2 \left(\varepsilon + \omega h_K^2 + \frac{\|\boldsymbol{b}\|_{0,\infty,K} h_K}{2C_{inv}} \right) h_K^{2r} |u|_{r+1,K}^2 \right)^{1/2} \leq$$

$$\leq C_{SU}^{\eta} \left(\sum_{K\in\mathcal{T}_h} \max\{C_{inv}, \mathrm{Pe}_K\} \varepsilon h_K^{2r} |u|_{r+1,K}^2 \right)^{1/2}, \tag{3.93}$$

with $C_{SU}^{\eta} = \frac{C_X}{\sqrt{C_{inv}}} \left(2 + 4n^2 C_{inv}^2 \max\limits_{K\in\mathcal{T}_h} \left\{ \frac{|\boldsymbol{b}|_{1,\infty,K}^2}{\|c\|_{0,\infty,K}^2} \right\} \right)^{1/2}$, we can for $r \leq k$ estimate the error $u - u_{SU}$ in the energy norm

$$|||u - u_{SU}|||_{SU} \leq \left(C_{SU}^{\xi} + C_{SU}^{\eta} \right) \left(\sum_{K\in\mathcal{T}_h} \max\{C_{inv}, \mathrm{Pe}_K\} \varepsilon h_K^{2r} |u|_{r+1,K}^2 \right)^{1/2}. \tag{3.94}$$

Remark 3.5.1. Since $C_{inv} > 1$ (cf. Remark 4.2.1, page 128), the assumption $\delta_K \leq \min\left\{ \frac{\omega}{\|c\|_{0,\infty,K}^2}, \frac{h_K}{2C_{inv}\|\boldsymbol{b}\|_{0,\infty,K}} \right\}$ does not allow to choose $\delta_K = \frac{h_K}{2\|\boldsymbol{b}\|_{0,\infty,K}}$ which is believed to be the optimal choice. Nevertheless, the a priori error estimate of the same order is achieved for the choice $\delta_K = \frac{h_K}{2C_{inv}\|\boldsymbol{b}\|_{0,\infty,K}}$, as well.

3.5.2 Error analysis of presented method

Let us turn back from the finite element space of general order $k \in \mathbb{N}$ to the linear finite elements, i.e. $k = 1$. In order to derive a priori error estimates we have to investigate the consistency error of the presented method.

Lemma 3.5.2. *Let $u \in H_0^1(\Omega) \cap H^2(\Omega)$ be the solution of (3.6) and let $u_h \in V_h$ satisfy (3.14). Then*

$$a_h(u - u_h, v_h) = \sum_{K\in\mathcal{T}_h} \left((\boldsymbol{b}_K - \boldsymbol{b})\nabla u, v_h + (P_{K,n+1} - C_K) \cdot \nabla v_h \right)_K \tag{3.95}$$

and consequently for any $w_h \in V_h$ it holds

$$\begin{aligned} &a_h(w_h - u_h, w_h - u_h) = \\ &= \varepsilon(\nabla(w_h - u), \nabla(w_h - u_h))_\Omega + \varepsilon \sum_{K\in\mathcal{T}_h} (\Delta u, (P_{K,n+1} - C_K) \cdot \nabla(w_h - u_h))_K + \\ &+ \sum_{K\in\mathcal{T}_h} \Big((\boldsymbol{b}_K - \boldsymbol{b}) \cdot \nabla w_h + \boldsymbol{b} \cdot \nabla(w_h - u), w_h - u_h + (P_{K,n+1} - C_K) \cdot \nabla(w_h - u_h) \Big)_K. \end{aligned} \tag{3.96}$$

Proof. The Galerkin quasi-orthogonality property (3.95) follows from the definition of a_h, a, F_h and F

$$\begin{aligned} a_h(u - u_h, v_h) &= a(u, v_h) + \sum_{K\in\mathcal{T}_h} ((\boldsymbol{b}_K - \boldsymbol{b})\nabla u, v_h)_K + \\ &+ \sum_{K\in\mathcal{T}_h} ((\boldsymbol{b}_K - \boldsymbol{b})\nabla u, (P_{K,n+1} - C_K) \cdot \nabla v_h)_K - a_h(u_h, v_h), \end{aligned} \tag{3.97}$$

which gives (3.95) since $a(u, v_h) - a_h(u_h, v_h) = F(v_h) - F_h(v_h) = 0$.

Further, using the decomposition

$$a_h(w_h - u_h, w_h - u_h) = a_h(w_h - u, w_h - u_h) + a_h(u - u_h, w_h - u_h) \tag{3.98}$$

and the fact that $\Delta w_h = 0$ we derive the relation (3.96). □

For estimation of the difference $\boldsymbol{b}_K - \boldsymbol{b}$ that occurs in Lemma 3.5.2 we use the following lemma.

Lemma 3.5.3. *Let $\boldsymbol{b} \in W^{1,\infty}(K)^n$ and let us define the vector $\boldsymbol{b}_K^I \in \mathbb{R}^n$*

$$\boldsymbol{b}_K^I \;=\; -\frac{1}{|K|}\sum_{j=1}^{n}\left(\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,j}\,\mathrm{d}\boldsymbol{x}\right)\boldsymbol{d}_{K,j}. \tag{3.99}$$

Then for every $v_h \in P_1(K)$ holds

$$\|(\boldsymbol{b}_K^I - \boldsymbol{b})\cdot\nabla v_h\|_{0,K} \;\le\; nC_\Pi h_K|\boldsymbol{b}|_{1,\infty,K}|v_h|_{1,K} \qquad \textit{and} \tag{3.100}$$

$$\|(\boldsymbol{b}_K - \boldsymbol{b}_K^I)\cdot\nabla v_h\|_{0,K} \;\le\; \frac{1}{|K|}\left|\sum_{j=2}^{n}\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,j}\,\mathrm{d}\boldsymbol{x}\right| h_K|v_h|_{1,K}, \tag{3.101}$$

where $\boldsymbol{b}_K$ is a vector defined in (3.13).

Proof. Since $\boldsymbol{d}_{K,j} = P_{K,n+1} - P_{K,j}$ and $\boldsymbol{d}_{K,j}\nabla\lambda_{K,i} = -\delta_{ij}$ for $i \neq n+1$, it holds

$$\int_K \boldsymbol{b}_K^I\nabla\lambda_{K,i}\,\mathrm{d}\boldsymbol{x} = \int_K -\frac{1}{|K|}\sum_{j=1}^{n}\left(\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,j}\,\mathrm{d}\boldsymbol{y}\right)(-\delta_{ij})\,\mathrm{d}\boldsymbol{x} = \int_K \boldsymbol{b}\cdot\nabla\lambda_{K,i}\,\mathrm{d}\boldsymbol{x} \tag{3.102}$$

for all $i = 1, 2, \ldots, n$ and (using $\sum_{j=1}^{n}\boldsymbol{b}\cdot\nabla\lambda_{K,j} = -\boldsymbol{b}\cdot\nabla\lambda_{K,n+1}$)

$$\int_K \boldsymbol{b}_K^I\nabla\lambda_{K,n+1}\,\mathrm{d}\boldsymbol{x} = \int_K -\frac{1}{|K|}\sum_{j=1}^{n}\left(\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,j}\,\mathrm{d}\boldsymbol{y}\right)\mathrm{d}\boldsymbol{x} = \int_K \boldsymbol{b}\cdot\nabla\lambda_{K,n+1}\,\mathrm{d}\boldsymbol{x}. \tag{3.103}$$

Consequently, $\int_K(\boldsymbol{b}_K^I - \boldsymbol{b})\cdot\nabla v_h\,\mathrm{d}\boldsymbol{x} = 0$ for all $v_h \in V_h$ and we can call $\boldsymbol{b}_K^I\nabla v_h$ the P_0-interpolation of the function $\boldsymbol{b}\cdot\nabla v_h$ on K. Using the approximation property (Theorem 4.2.3, page 128) we therefore obtain

$$\|(\boldsymbol{b}_K^I - \boldsymbol{b})\cdot\nabla v_h\|_{0,K} \;\le\; C_\Pi h_K|\boldsymbol{b}\cdot\nabla v_h|_{1,K} \;\le\; nC_\Pi h_K|\boldsymbol{b}|_{1,\infty,K}|v_h|_{1,K}. \tag{3.104}$$

The estimate (3.101) results directly from the definition of $\boldsymbol{b}_K^I$ and $\boldsymbol{b}_K$ and the fact that $|\boldsymbol{d}_{K,j}| \le h_K = \max_{i\neq j}|P_{K,i} - P_{K,j}|$. □

Now we use the stability and the Galerkin quasi-orthogonality and derive the error estimates of the presented method.

Theorem 3.5.1. *Let there exists constant κ independent of h (and ε) such that $\frac{|\mathcal{C}_{j+1}^s|}{|\Omega_j^s|} \ge \kappa$ for all $s = 1, 2, \ldots, \mathcal{P}$ and $j = 1, 2, \ldots, N_s$, constant $\omega > 0$ such that $\operatorname{div}\boldsymbol{b} \le -\omega < 0$ in Ω and let for each $K \in \mathcal{T}_h$ holds*

$$\theta_K \;\le\; \frac{\omega}{n+1}. \tag{3.105}$$

If the solution u of the problem (3.1) satisfies $u \in H^2(\Omega)$, then there exists a constant $C_1 > 0$ independent of h and ε such that for the solution $u_h \in V_h$ obtained by the method (3.14) (using continuous piecewise linear finite elements) it holds

$$|||u - u_h|||_b \;\le\; C_1\left(\sum_{K\in\mathcal{T}_h} h_K^2\Big(|u|_{2,K}^2 + |u|_{1,K}^2\Big)\right)^{1/2}. \tag{3.106}$$

If, in addition, the mesh parameter θ_K satisfies for all $K \in \mathcal{T}_h$

$$\theta_K \;\leq\; \min\left\{\frac{\omega}{n+1}, |\boldsymbol{b}|_{1,\infty,K}\sqrt{\omega}\max\left\{\frac{h_K}{\varepsilon^{1/2}}, \frac{2}{|\boldsymbol{b}_K|}\,\varepsilon^{1/2}\right\}\right\}, \tag{3.107}$$

then there exists a constant $C_2 > 0$ independent of h and ε such that for the solution $u_h \in V_h$ obtained by the method (3.14) there holds

$$|||u - u_h|||_b \;\leq\; C_2\left(\sum_{K\in\mathcal{T}_h}\min\left\{h_K^2, \max\left\{\frac{h_K^4}{\varepsilon}, \varepsilon h_K^2\right\}\right\}\Big(|u|_{2,K}^2 + |u|_{1,K}^2\Big)\right)^{1/2}. \tag{3.108}$$

Proof. At first, let u^I be again the V_h-interpolant of the function u and let us denote $\eta_h = u - u^I$ and $\xi_h = u^I - u_h$. Further, we decompose the error $u - u_h = (u - u^I) + (u^I - u_h) = \eta_h + \xi_h$ and since $\xi_h \in V_h$ we can use the coercivity (3.61) together with (3.96) (setting $w_h = u^I$), which yields

$$\begin{aligned}
&\frac{1}{2}|||\xi_h|||_b^2 \;\leq\; a_h(\xi_h,\xi_h) \;=\\
&\quad= \;-\varepsilon(\nabla\eta_h, \nabla\xi_h)_\Omega + \varepsilon\sum_{K\in\mathcal{T}_h}(\Delta u, (P_{K,n+1} - C_K)\cdot\nabla\xi_h)_K + \\
&\qquad+\sum_{K\in\mathcal{T}_h}\Big((\boldsymbol{b}_K - \boldsymbol{b}_K^I)\cdot\nabla u^I + (\boldsymbol{b}_K^I - \boldsymbol{b})\cdot\nabla u^I - \boldsymbol{b}\cdot\nabla\eta_h, \xi_h + (P_{K,n+1} - C_K)\cdot\nabla\xi_h\Big)_K.
\end{aligned} \tag{3.109}$$

Now we estimate each term of (3.109) separately:

1. We use the Cauchy-Schwarz-Bunyakovsky inequalities (Theorem 4.1.4, page 126), the interpolation inequality (Theorem 4.2.1, page 127) and estimate

$$\begin{aligned}
&-\varepsilon(\nabla\eta_h, \nabla\xi_h)_\Omega \;=\; -\sum_{K\in\mathcal{T}_h}\varepsilon(\nabla\eta_h, \nabla\xi_h)_K \;\leq\; \sum_{K\in\mathcal{T}_h}\varepsilon|\eta_h|_{1,K}|\xi_h|_{1,K} \;=\\
&\quad= \sum_{K\in\mathcal{T}_h}\Big(\varepsilon^{1/2}|\eta_h|_{1,K}\Big)\Big(\varepsilon^{1/2}|\xi_h|_{1,K}\Big) \;\leq\; \left(\sum_{K\in\mathcal{T}_h}\varepsilon|\eta_h|_{1,K}^2\right)^{1/2}\left(\sum_{K\in\mathcal{T}_h}\varepsilon|\xi_h|_{1,K}^2\right)^{1/2} \leq\\
&\quad\leq \left(\sum_{K\in\mathcal{T}_h}\varepsilon C_X^2 h_K^2|u|_{2,K}^2\right)^{1/2}\left(\sum_{K\in\mathcal{T}_h}\varepsilon|\xi_h|_{1,K}^2\right)^{1/2} \leq C_X\left(\sum_{K\in\mathcal{T}_h}\varepsilon h_K^2|u|_{2,K}^2\right)^{1/2}|||\xi_h|||_b.
\end{aligned} \tag{3.110}$$

2. Since $|P_{K,n+1} - C_K| \leq h_K$ then similarly as in the previous case we have

$$\begin{aligned}
&\varepsilon\sum_{K\in\mathcal{T}_h}(\Delta u, (P_{K,n+1} - C_K)\cdot\nabla\xi_h)_K \;\leq\; \varepsilon\sum_{K\in\mathcal{T}_h} n|u|_{2,K}\,h_K|\xi_h|_{1,K} \;=\\
&\quad= \;n\sum_{K\in\mathcal{T}_h}\Big(\varepsilon^{1/2}h_K|u|_{2,K}\Big)\Big(\varepsilon^{1/2}|\xi_h|_{1,K}\Big) \;\leq\; n\left(\sum_{K\in\mathcal{T}_h}\varepsilon h_K^2|u|_{2,K}^2\right)^{1/2}|||\xi_h|||_b.
\end{aligned} \tag{3.111}$$

3. From the inequality (3.101) it follows that there holds $\|(\boldsymbol{b}_K - \boldsymbol{b}_K^I)\cdot\nabla u^I\|_{0,K} \leq \theta_K h_K|u^I|_{1,K}$. Consequently, using the inverse inequality (Theorem 4.2.2,

page 127) and the estimate $|u^I|_{1,K} \leq |u|_{1,K} + |\eta_h|_{1,K}$ together with the interpolation inequality (Theorem 4.2.1, page 127) yields

$$\sum_{K\in\mathcal{T}_h}\Big((\boldsymbol{b}_K - \boldsymbol{b}_K^I)\cdot\nabla u^I, \xi_h + (P_{K,n+1} - C_K)\cdot\nabla\xi_h\Big)_K \leq$$
$$\leq \sum_{K\in\mathcal{T}_h}\theta_K h_K |u^I|_{1,K}(1+C_{inv})\|\xi_h\|_{0,K} \leq$$
$$\leq \frac{(1+C_{inv})\sqrt{2}}{\sqrt{\kappa}}\left(\sum_{K\in\mathcal{T}_h}\frac{\theta_K^2}{\omega}h_K^2\Big(|u|_{1,K} + C_X h_K |u|_{2,K}\Big)^2\right)^{1/2}|||\xi_h|||_b. \quad (3.112)$$

4. Since $\int_K(\boldsymbol{b}_K^I - \boldsymbol{b})\cdot\nabla u^I \,\mathrm{d}\boldsymbol{x} = 0$, then using the Cauchy-Schwarz-Bunyakovsky inequality (Theorem 4.1.4, page 126), the approximation property (Theorem 4.2.3, page 128) and the estimate (3.100) we obtain

$$\sum_{K\in\mathcal{T}_h}\Big((\boldsymbol{b}_K^I - \boldsymbol{b})\cdot\nabla u^I, \xi_h + (P_{K,n+1} - C_K)\cdot\nabla\xi_h\Big)_K =$$
$$= \sum_{K\in\mathcal{T}_h}\Big((\boldsymbol{b}_K^I - \boldsymbol{b})\cdot\nabla u^I, \xi_h - \xi_h(C_K)\Big)_K \leq \sum_{K\in\mathcal{T}_h} nC_\Pi^2 h_K^2 |\boldsymbol{b}|_{1,\infty,K}|u^I|_{1,K}|\xi_h|_{1,K} \leq$$
$$\leq nC_\Pi^2\left(\sum_{K\in\mathcal{T}_h}h_K^4|\boldsymbol{b}|_{1,\infty,K}^2\left(|u|_{1,K}+|\eta_h|_{1,K}\right)^2\min\left\{\frac{1}{\varepsilon},\frac{2C_{inv}^2}{\omega\kappa h_K^2}\right\}\right)^{1/2}|||\xi_h|||_b \leq$$
$$\leq nC_\Pi^2\left(\sum_{K\in\mathcal{T}_h}h_K^2|\boldsymbol{b}|_{1,\infty,K}^2\left(|u|_{1,K}+C_X h_K|u|_{2,K}\right)^2\min\left\{\frac{h_K^2}{\varepsilon},\frac{2C_{inv}^2}{\omega\kappa}\right\}\right)^{1/2}|||\xi_h|||_b, \quad (3.113)$$

where we employed the inequality $|\xi_h|_{1,K}^2 \leq \min\left\{\frac{1}{\varepsilon},\frac{2C_{inv}^2}{\omega\kappa h_K^2}\right\}|||\xi_h|||_b^2$ resulting from the inverse inequality (Theorem 4.2.2, page 127).

5. Using the Green theorem (Theorem 4.1.1, page 124), the approximation property (Theorem 4.2.3, page 128), the interpolation inequality (Theorem 4.2.1, page 127), the inverse inequality (Theorem 4.2.2, page 127), the shape-regularity (Assumption 4.2.1, page 127) and the estimates (3.100), (3.101) we get

$$-\sum_{K\in\mathcal{T}_h}(\boldsymbol{b}\cdot\nabla\eta_h,\xi_h)_K = -(\boldsymbol{b}\cdot\nabla\eta_h,\xi_h)_\Omega = (\eta_h\mathrm{div}\,\boldsymbol{b},\xi_h)_\Omega + (\eta_h,\boldsymbol{b}\cdot\nabla\xi_h)_\Omega =$$
$$= (\eta_h\mathrm{div}\,\boldsymbol{b},\xi_h)_\Omega + \sum_{K\in\mathcal{T}_h}(\eta_h,\boldsymbol{b}_K\cdot\nabla\xi_h)_K + \sum_{K\in\mathcal{T}_h}(\eta_h,(\boldsymbol{b}-\boldsymbol{b}_K)\cdot\nabla\xi_h)_K \leq$$
$$\leq \sum_{K\in\mathcal{T}_h} n|\boldsymbol{b}|_{1,\infty,K}\|\eta_h\|_{0,K}\|\xi_h\|_{0,K} + \sum_{K\in\mathcal{T}_h}\|\eta_h\|_{0,K}\|\boldsymbol{b}_K\cdot\nabla\xi_h\|_{0,K} +$$
$$+\sum_{K\in\mathcal{T}_h}\|\eta_h\|_{0,K}(nC_\Pi|\boldsymbol{b}|_{1,\infty,K}+\theta_K)h_K|\xi_h|_{1,K} \leq$$
$$\leq \left(3\sum_{K\in\mathcal{T}_h}\gamma_K C_X^2 h_K^4|u|_{2,K}^2\right)^{1/2}|||\xi_h|||_b, \quad (3.114)$$

where we denoted $\gamma_K = \frac{2n^2}{\omega\kappa}|\boldsymbol{b}|_{1,\infty,K}^2 + \min\left\{\frac{2\sigma|\boldsymbol{b}_K|}{h_K},\frac{|\boldsymbol{b}_K|^2}{\varepsilon}\right\} + (nC_\Pi|\boldsymbol{b}|_{1,\infty,K} + \theta_K)^2\min\left\{\frac{h_K^2}{\varepsilon},\frac{2C_{inv}^2}{\omega\kappa}\right\}$. In the above estimate we decomposed the difference

$\boldsymbol{b} - \boldsymbol{b}_K = (\boldsymbol{b} - \boldsymbol{b}_K^I) + (\boldsymbol{b}_K^I - \boldsymbol{b}_K)$ and employed the inequalities (3.100) and (3.101).

6. Finally, the estimate of the last term takes form

$$-\sum_{K\in\mathcal{T}_h}\Big(\boldsymbol{b}\cdot\nabla\eta_h, (P_{K,n+1}-C_K)\cdot\nabla\xi_h\Big)_K \;\le\; \sum_{K\in\mathcal{T}_h}\|\boldsymbol{b}\|_{0,\infty,K}|\eta_h|_{1,K}h_K|\xi_h|_{1,K} \;\le$$

$$\le \quad \left(\sum_{K\in\mathcal{T}_h}\|\boldsymbol{b}\|_{0,\infty,K}^2 C_X^2 h_K^2 |u|_{2,K}^2 \min\left\{\frac{h_K^2}{\varepsilon}, \frac{2C_{inv}^2}{\omega\kappa}\right\}\right)^{1/2} |||\xi_h|||_b. \tag{3.115}$$

From the above derived estimates it follows that if we consider $\theta_K \le \frac{\omega}{n+1}$ for all $K \in \mathcal{T}_h$, we obtain the estimate

$$|||\xi_h|||_b \;\le\; C_0\left(\sum_{K\in\mathcal{T}_h} h_K^2\Big(|u|_{2,K}^2 + |u|_{1,K}^2\Big)\right)^{1/2}, \tag{3.116}$$

where the constant C_0 does not depend on ε and h_K.

However, if the triangulation $\mathcal{T}_h$ (and a vector field $\boldsymbol{b}$) is such that a sharper bound (3.107) is valid for all $K \in \mathcal{T}_h$, then we may obtain an acceleration of the convergence. To show this we distinguish three situations:

A) At first, let us consider $h_K \ge \varepsilon^{1/2}\, C_{inv}\sqrt{\frac{2}{\omega\kappa}}$ for all $K \in \mathcal{T}_h$, then

$$|\boldsymbol{b}|_{1,\infty,K}\sqrt{\omega}\,\frac{h_K}{\varepsilon^{1/2}} \;\ge\; |\boldsymbol{b}|_{1,\infty,K}C_{inv}\sqrt{\frac{2}{\kappa}} \;\ge\; \frac{\omega}{n}C_{inv}\sqrt{2(n+1)} \;\ge\; \frac{\omega}{n+1}, \tag{3.117}$$

where we used the apparent inequalities $|\boldsymbol{b}|_{1,\infty,K} \ge \frac{1}{n}\,|\mathrm{div}\,\boldsymbol{b}| \ge \frac{1}{n}\,\omega$, $\kappa \le \frac{1}{n+1}$ and $C_{inv} \ge 1$ (cf. Remark 4.2.1, page 128). Hence, from the assumption (3.107) it follows that the bound $\theta_K \le \frac{\omega}{n+1}$ is used in this case, which together with Corollary 4.2.2 (page 129) results in the same estimate as above

$$|||\xi_h|||_b \;\le\; C_A\left(\sum_{K\in\mathcal{T}_h} h_K^2\Big(|u|_{1,K}^2 + |u|_{2,K}^2\Big)\right)^{1/2},$$

where the constant C_A does not depend on ε and h_K. Thus, no improvement is achieved in this case.

B) If the mesh is refined and $\frac{2\sigma}{|\boldsymbol{b}_K|}\,\varepsilon \le h_K \le \varepsilon^{1/2}\, C_{inv}\sqrt{\frac{2}{\omega\kappa}}$ for all $K \in \mathcal{T}_h$, then using the inequality

$$\theta_K \;\le\; |\boldsymbol{b}|_{1,\infty,K}\sqrt{\omega}\,\frac{h_K}{\varepsilon^{1/2}} \quad \Big(\le\; |\boldsymbol{b}|_{1,\infty,K}C_{inv}\sqrt{2/\kappa}\Big) \tag{3.118}$$

and the estimate $\varepsilon h_K^2 = \frac{\varepsilon^2}{h_K^2}\,\frac{h_K^4}{\varepsilon} \le \frac{|\boldsymbol{b}_K|^2}{4\sigma^2}\,\frac{h_K^4}{\varepsilon}$ yields

$$|||\xi_h|||_b \;\le\; C_B\left(\sum_{K\in\mathcal{T}_h} \frac{h_K^4}{\varepsilon}\Big(|u|_{1,K}^2 + |u|_{2,K}^2\Big)\right)^{1/2},$$

where the constant C_B again does not depend on ε and h_K. Since in this case there holds $\frac{2\sigma}{|\boldsymbol{b}_K|}\,h_K^3 \le \frac{h_K^4}{\varepsilon} \le h_K^2\,\frac{2C_{inv}^2}{\omega\kappa}$, one may expect an acceleration of the convergence in comparison to (3.116).

C) If the mesh step satisfies $h_K \leq \varepsilon \frac{2\sigma}{|\boldsymbol{b}_K|}$ for all $K \in \mathcal{T}_h$, then there holds

$$\theta_K \;\leq\; |\boldsymbol{b}|_{1,\infty,K}\sqrt{\omega}\,\frac{2\sigma}{|\boldsymbol{b}_K|}\,\varepsilon^{1/2}. \tag{3.119}$$

This bound together with the inequality $\frac{h_K^4}{\varepsilon} = \frac{h_K^2}{\varepsilon^2}\,\varepsilon h_K^2 \leq \frac{4\sigma^2}{|\boldsymbol{b}_K|^2}\,\varepsilon h_K^2$ then provides the estimate

$$|||\xi_h|||_b \;\leq\; C_C\left(\sum_{K\in\mathcal{T}_h}\varepsilon h_K^2\Big(|u|_{1,K}^2 + |u|_{2,K}^2\Big)\right)^{1/2}, \tag{3.120}$$

with the constant C_C independent of both ε and h_K.

Hence, in this case we obtain the same result as for the SUPG method. Since there holds $\frac{|\boldsymbol{b}_K|}{2\sigma}\,h_K^3 \leq \varepsilon h_K^2 \leq \varepsilon_0 h_K^2$ we observe that the estimate (3.120) is again an improvement of the estimate (3.116).

Finally, for the approximation error in all cases A), B) and C) the following inequality holds

$$|||\eta_h|||_b \;\leq\; \left(\sum_{K\in\mathcal{T}_h}\left(\varepsilon + \frac{\omega\kappa}{2}h_K^2 + \frac{|\boldsymbol{b}_K|}{2}h_K\right)C_X^2 h_K^2 |u|_{2,K}^2\right)^{1/2}. \tag{3.121}$$

Consequently, in the first two cases (A) and B)) we obtain the estimate $|||\eta_h|||_b \leq C_{X1}\left(\sum_{K\in\mathcal{T}_h} h_K^3|u|_{2,K}^2\right)^{1/2} \leq C_{X1}\left(\sum_{K\in\mathcal{T}_h}\min\{h_K^2, h_K^4/\varepsilon\}|u|_{2,K}^2\right)^{1/2}$, whereas in the case C) there holds $|||\eta_h|||_b \leq C_{X2}\left(\sum_{K\in\mathcal{T}_h}\varepsilon h_K^2|u|_{2,K}^2\right)^{1/2}$, where C_{X1} and C_{X2} are positive constants independent of ε and h_K.

The statement of the proof therefore results from the triangular inequality

$$|||u - u_h|||_b \;=\; |||\xi_h + \eta_h|||_b \;\leq\; |||\xi_h|||_b + |||\eta_h|||_b. \tag{3.122}$$

□

When $\operatorname{div}\boldsymbol{b} = 0$, the situation gets complicated, since we have to use weaker norm $|||\cdot|||_{b,*}$ (for all $v \in H^1(\Omega)$ there holds $|||v|||_{b,*}^2 \leq C_b^*|||v|||_b^2$, providing $h_K \leq \frac{\omega\kappa C_b^*}{2C_2^*}$ for all $K \in \mathcal{T}_h$). Consequently, we obtain lower order of convergence with respect to the norm $|||\cdot|||_{b,*}$.

Theorem 3.5.2. *Let* $\operatorname{div}\boldsymbol{b} = 0$ *and let there exists* $\delta \in (0,4)$ *such that*

$$\theta_K \;\leq\; \frac{\delta}{\overline{\nu}_K}\|\boldsymbol{b}\|_{0,\infty,K}h_K \qquad \textit{for all } K \in \mathcal{T}_h. \tag{3.123}$$

Further, let there exist positive numbers κ, L, R *and* β *such that for all* $s = 1,2,\ldots,\mathcal{P}$ *and* $j = 1,2,\ldots,N_s$ *it holds*

$$\frac{|\mathcal{C}_j^s|}{|\Omega_j^s|} \geq \kappa, \qquad N_s h_j^s \leq L, \qquad \frac{\max\limits_{K\subset\Omega_j^s} h_K}{h_j^s} \leq R, \qquad \textit{and} \qquad \beta_j^s \geq \beta. \tag{3.124}$$

If the solution u of the problem (3.1) satisfies $u \in H^2(\Omega)$, then there exists constant $C^* > 0$ independent of h and ε such that for the solution obtained by the method (3.14) there holds

$$|||u - u_h|||_{b,*} \;\le\; C^* \left(\sum_{K\in\mathcal{T}_h} \min\left\{ h_K, \max\left\{ \frac{h_K^4}{\varepsilon}, \varepsilon h_K^2 \right\} \right\} \Big(|u|_{2,K}^2 + |u|_{1,K}^2 \Big) \right)^{1/2}. \tag{3.125}$$

Proof. A similar approach like in the proof of Theorem 3.5.1 leads to the following estimates.

1*. First two inequalities remain the same

$$-\varepsilon(\nabla\eta_h, \nabla\xi_h)_\Omega \;\le\; \sum_{K\in\mathcal{T}_h} \varepsilon |\eta_h|_{1,K} |\xi_h|_{1,K} \;\le\; C_X \left(\sum_{K\in\mathcal{T}_h} \varepsilon h_K^2 |u|_{2,K}^2 \right)^{1/2} |||\xi_h|||_{b,*}. \tag{3.126}$$

2*.

$$\varepsilon \sum_{K\in\mathcal{T}_h} (\Delta u, (P_{K,n+1} - C_K)\cdot\nabla\xi_h)_K \;\le\; n \left(\sum_{K\in\mathcal{T}_h} \varepsilon h_K^2 |u|_{2,K}^2 \right)^{1/2} |||\xi_h|||_{b,*}. \tag{3.127}$$

3*. Since $\operatorname{div} \boldsymbol{b} = 0$, we have to use the norm $|||\cdot|||_{b,*}$ and hence we lost one half of the order in h_K.

$$\begin{aligned} &\sum_{K\in\mathcal{T}_h} \Big((\boldsymbol{b}_K - \boldsymbol{b}_K^I)\cdot\nabla u^I, \xi_h + (P_{K,n+1} - C_K)\cdot\nabla\xi_h \Big)_K \;\le \\ &\le \sum_{K\in\mathcal{T}_h} \theta_K h_K |u^I|_{1,K} (1 + C_{inv}) \|\xi_h\|_{0,K} \;\le \\ &\le \frac{1 + C_{inv}}{\sqrt{C_2^*}} \left(\sum_{K\in\mathcal{T}_h} \theta_K^2 h_K \Big(|u|_{1,K} + C_X h_K |u|_{2,K} \Big)^2 \right)^{1/2} |||\xi_h|||_{b,*}. \end{aligned} \tag{3.128}$$

4*. Again one half of the order is lost for large h_K.

$$\begin{aligned} &\sum_{K\in\mathcal{T}_h} \Big((\boldsymbol{b}_K^I - \boldsymbol{b})\cdot\nabla u^I, \xi_h + (P_{K,n+1} - C_K)\cdot\nabla\xi_h \Big)_K \;= \\ &= \sum_{K\in\mathcal{T}_h} \Big((\boldsymbol{b}_K^I - \boldsymbol{b})\cdot\nabla u^I, \xi_h - \xi_h(C_K) \Big)_K \le \sum_{K\in\mathcal{T}_h} n C_\Pi^2 h_K^2 |\boldsymbol{b}|_{1,\infty,K} |u^I|_{1,K} |\xi_h|_{1,K} \le \\ &\le n C_\Pi^2 \left(\sum_{K\in\mathcal{T}_h} h_K^2 |\boldsymbol{b}|_{1,\infty,K}^2 \Big(|u|_{1,K} + C_X h_K |u|_{2,K} \Big)^2 \min\left\{ \frac{h_K^2}{\varepsilon}, \frac{C_{inv}^2}{C_2^* h_K} \right\} \right)^{1/2} |||\xi_h|||_{b,*}. \end{aligned} \tag{3.129}$$

5*. When $\operatorname{div} \boldsymbol{b} = 0$ then using the Green theorem (Theorem 4.1.1, page 124), the approximation property (Theorem 4.2.3, page 128), the interpolation inequality (Theorem 4.2.1, page 127), the inverse inequality (Theorem 4.2.2, page

127), the shape-regularity (Assumption 4.2.1, page 127) and the estimates (3.100), (3.101) we get

$$-(\boldsymbol{b}\cdot\nabla\eta_h,\xi_h)_\Omega \;=\; \sum_{K\in\mathcal{T}_h}(\eta_h,\boldsymbol{b}_K\cdot\nabla\xi_h)_K+\sum_{K\in\mathcal{T}_h}(\eta_h,(\boldsymbol{b}-\boldsymbol{b}_K)\cdot\nabla\xi_h)_K \;\le$$
$$\le \sum_{K\in\mathcal{T}_h}\|\eta_h\|_{0,K}\|\boldsymbol{b}_K\cdot\nabla\xi_h\|_{0,K}+\sum_{K\in\mathcal{T}_h}\|\eta_h\|_{0,K}(nC_\Pi|\boldsymbol{b}|_{1,\infty,K}+\theta_K)h_K|\xi_h|_{1,K}\le$$
$$\le \left(2\sum_{K\in\mathcal{T}_h}\gamma_K C_X^2 h_K^4|u|_{2,K}^2\right)^{1/2}|||\xi_h|||_{b,*}, \tag{3.130}$$

where $\gamma_K=\min\left\{\frac{2\sigma|\boldsymbol{b}_K|}{C_b^* h_K},\frac{|\boldsymbol{b}_K|^2}{\varepsilon}\right\}+(nC_\Pi|\boldsymbol{b}|_{1,\infty,K}+\theta_K)^2\min\left\{\frac{h_K^2}{\varepsilon},\frac{C_{inv}^2}{C_2^* h_K}\right\}$.

6*. And the estimate of the last term is

$$-\sum_{K\in\mathcal{T}_h}\Big(\boldsymbol{b}\cdot\nabla\eta_h,(P_{K,n+1}-C_K)\cdot\nabla\xi_h\Big)_K \;\le\; \sum_{K\in\mathcal{T}_h}\|\boldsymbol{b}\|_{0,\infty,K}|\eta_h|_{1,K}h_K|\xi_h|_{1,K}\;\le$$
$$\le \left(\sum_{K\in\mathcal{T}_h}\|\boldsymbol{b}\|_{0,\infty,K}^2 C_X^2 h_K^2|u|_{2,K}^2\min\left\{\frac{h_K^2}{\varepsilon},\frac{C_{inv}^2}{C_2^* h_K}\right\}\right)^{1/2}|||\xi_h|||_{b,*}. \tag{3.131}$$

The derived estimates imply that if θ_K is bounded for all $K\in\mathcal{T}_h$ (we even consider $\theta_K\le\frac{\delta}{\overline{\nu}_K}\|\boldsymbol{b}\|_{0,\infty,K}h_K$) we obtain a method of the order 1/2 with respect to the $\|\cdot\|_{b,*}$-norm. Nevertheless, as we will show, when refining the mesh the convergence can again accelerate.

A*) Firstly, we consider $h_K\ge\varepsilon^{1/3}\sqrt[3]{C_{inv}^2/C_2^*}$. Then using the bound $\theta_K\le\frac{\delta}{\overline{\nu}_K}\|\boldsymbol{b}\|_{0,\infty,K}h_K$ together with Corollary 4.2.2 (page 129) we obtain the estimate

$$|||\xi_h|||_{b,*}\;\le\; C_A^*\left(\sum_{K\in\mathcal{T}_h}h_K\Big(|u|_{1,K}+|u|_{2,K}\Big)^2\right)^{1/2}, \tag{3.132}$$

where the constant C_A^* does not depend on ε and h_K.

B*) If the mesh is refined and $\frac{2\sigma}{|\boldsymbol{b}_K|C_b^*}\varepsilon\le h_K\le\varepsilon^{1/3}\sqrt[3]{C_{inv}^2/C_2^*}$ then using the inequalities $\varepsilon h_K^2=\frac{\varepsilon^2}{h_K^2}\frac{h_K^4}{\varepsilon}\le\left(\frac{|\boldsymbol{b}_K|C_b^*}{2\sigma}\right)^2\frac{h_K^4}{\varepsilon}$, $h_K^3=\frac{\varepsilon}{h_K}\frac{h_K^4}{\varepsilon}\le\left(\frac{|\boldsymbol{b}_K|C_b^*}{2\sigma}\right)\frac{h_K^4}{\varepsilon}$ and $\theta_K^2 h_K\le\left(\frac{\delta}{\overline{\nu}_K}\|\boldsymbol{b}\|_{0,\infty,K}\right)^2 h_K^3$ we obtain

$$|||\xi_h|||_{b,*}\;\le\; C_B^*\left(\sum_{K\in\mathcal{T}_h}\frac{h_K^4}{\varepsilon}\Big(|u|_{1,K}+|u|_{2,K}\Big)^2\right)^{1/2}, \tag{3.133}$$

where the constant C_B^* again does not depend on both ε and h_K.

Since in this case there holds $\frac{2\sigma}{|\boldsymbol{b}_K|C_b^*}h_K^3\le\frac{h_K^4}{\varepsilon}\le h_K\frac{C_{inv}^2}{C_2^*}$, one may expect an acceleration of the convergence in comparison to (3.132).

C*) If the mesh step satisfies $h_K \leq \varepsilon \frac{2\sigma}{|\boldsymbol{b}_K|C_b^*}$ then using the inequalities $\frac{h_K^4}{\varepsilon} = \frac{h_K^2}{\varepsilon^2}\varepsilon h_K^2 \leq \left(\frac{2\sigma}{|\boldsymbol{b}_K|C_b^*}\right)^2 \varepsilon h_K^2$ and $\theta_K^2 h_K \leq \left(\frac{\delta}{\overline{\nu}_K}\|\boldsymbol{b}\|_{0,\infty,K}\right)^2 \frac{2\sigma}{|\boldsymbol{b}_K|C_b^*}\varepsilon h_K^2$ we obtain

$$|||\xi_h|||_{b,*} \leq C_C^* \left(\sum_{K\in\mathcal{T}_h} \varepsilon h_K^2 \Big(|u|_{1,K} + |u|_{2,K}\Big)^2\right)^{1/2}, \tag{3.134}$$

where the constant C_C^* does not depend on ε and h_K.

Since there holds $\frac{|\boldsymbol{b}_K|C_b^*}{2\sigma} h_K^3 \leq \varepsilon h_K^2 \leq \varepsilon_0 h_K^2$ we observe that the estimate (3.134) is again an improvement of the estimate (3.132).

Finally, for the approximation error in all cases A*), B*) and C*) there holds

$$|||\eta_h|||_{b,*} \leq \left(\sum_{K\in\mathcal{T}_h} \left(\varepsilon + C_2^* h_K^3 + C_b^* \frac{|\boldsymbol{b}_K|}{2} h_K\right) C_X^2 h_K^2 |u|_{2,K}^2\right)^{1/2}. \tag{3.135}$$

Thus, in the first two cases (A*) and B*)) we obtain the estimate $|||\eta_h|||_{b,*} \leq C_{X1}^* \left(\sum_{K\in\mathcal{T}_h} h_K^3 |u|_{2,K}^2\right)^{1/2} \leq C_{X1}^* \left(\sum_{K\in\mathcal{T}_h} \min\{h_K, h_K^4/\varepsilon\}|u|_{2,K}^2\right)^{1/2}$, whereas in the case C*) there holds $|||\eta_h|||_{b,*} \leq C_{X2}^* \left(\sum_{K\in\mathcal{T}_h} \varepsilon h_K^2 |u|_{2,K}^2\right)^{1/2}$, where C_{X1}^* and C_{X2}^* are positive constants independent of ε and h_K.

The statement of the proof therefore results from the triangular inequality

$$|||u - u_h|||_{b,*} = |||\xi_h + \eta_h|||_{b,*} \leq |||\xi_h|||_{b,*} + |||\eta_h|||_{b,*}. \tag{3.136}$$

□

Remark 3.5.2. For each numerical method one can define the experimental order of convergence with respect to some norm $|||\cdot|||$ as $EOC = \frac{\log(e_{h_1}/e_{h_2})}{\log(h_1/h_2)}$, where $e_h = |||u - u_h|||$. Consequently, one obtain $e_h \approx Ch^{EOC}$ for suitable constant C independent of h. From the above derived a priori error estimates it follows which EOC we should expect. Since ε is constant we use powers of ε as a scale and plot the progression of the dependency of the expected EOC on h (see Figure 3.4).

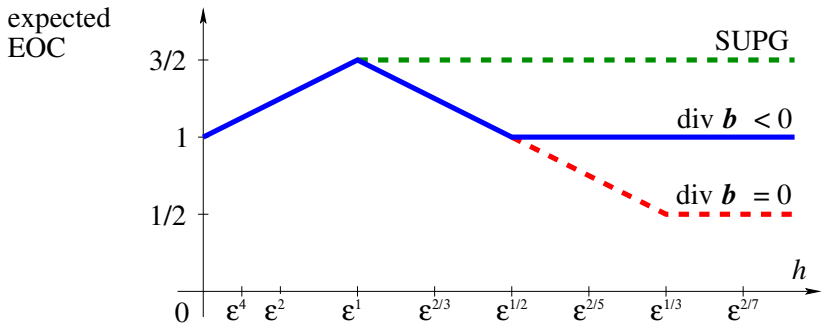

Figure 3.4: The expected EOC progression of the original SUPG method and the new method for $\operatorname{div}\boldsymbol{b} < 0$ and $\operatorname{div}\boldsymbol{b} = 0$. Continuous piecewise linear finite elements are used and the error is measured in the norms $|||\cdot|||_{SD}$, $|||\cdot|||_b$ and $|||\cdot|||_{b,*}$, respectively. We can observe that the theoretical convergence order (the order of the a priori error estimate) depends on the relation between h_K and ε.

3.5.3 M-matrix

While Theorem 3.2.2 (page 58) ensures the fulfilment of the discrete maximum principle (3.19), it does not ensure that the matrix L_h of the presented method is an M-matrix, which can allow us to estimate $\|L_h^{-1}\|_{\infty,d}$.

The following theorem provides some sufficient assumptions under which the matrix of the method is an M-matrix. It also provides an estimate of the discrete L^∞-norm of the matrix $\tilde{L}_h^{-1}$. It is the inverse of the matrix $\tilde{L}_h$, which is obtained from L_h by multiplying each row (corresponding to some basis function λ_k, $1 \leq k \leq N_h$) by the factor $|\mathrm{supp}\,\{\lambda_k\}|^{-1}$. Hence, the matrix $\tilde{L}_h$ is defined as follows

$$(\tilde{L}_h)_{ki} \;=\; \frac{(L_h)_{ki}}{|\mathrm{supp}\,\{\lambda_k\}|} \;=\; \frac{a_h(\lambda_i,\lambda_k)}{|\mathrm{supp}\,\{\lambda_k\}|}, \qquad \text{for all } 1 \leq i,k \leq N_h. \tag{3.137}$$

In our notation, λ_k is λ_j^s for suitable $s \in \{1,2,\ldots,\mathcal{P}\}$, $j \in \{1,2,\ldots,N_s\}$ and therefore $|\mathrm{supp}\,\{\lambda_k\}| = |\Omega_j^s|$.

In order to make our considerations more simple in this section, we assume that our grid is quasi-equidistant.

Definition 3.5.1. *A family $\mathcal{T}_h$ of grids is called* quasi-equidistant *if there exists some constant Q such that for each grid $\mathcal{T}_h$ one has*

$$\frac{\max_{K\in\mathcal{T}_h} h_K}{\min_{K\in\mathcal{T}_h} h_K} \;\leq\; Q. \tag{3.138}$$

Theorem 3.5.3. *Let Ω_1 be a domain such that $\overline{\Omega} \subset \Omega_1$ and let $\boldsymbol{b} \in \mathcal{C}^1(\Omega_1)^n$ with $\mathrm{div}\,\boldsymbol{b} \leq -\omega < 0$. Further, let us assume that all the streamlines of the vector field $\boldsymbol{b}$ leave $\overline{\Omega}$ in finite time (i.e. periodic solutions and points with $\boldsymbol{b}(\boldsymbol{x}) = 0$ are not allowed). Then there exists a positive function $\phi \in \mathcal{C}^1(\Omega_1)$ so that $\boldsymbol{b}\cdot\nabla\phi \geq \phi_0 > 0$ in $\overline{\Omega}$. Moreover, if $\theta_K \leq \frac{\omega}{n+1}$ for each $K \in \mathcal{T}_h$ and if there holds*

$$\varepsilon \;\leq\; \min\left\{\frac{C_1\phi_0}{|\phi|_{2,\infty,\Omega}}, \frac{C_2\phi_0^2}{|\phi|^2_{1,\infty,\Omega}}, \frac{C_3\phi_0^2}{|\phi|^2_{2,\infty,\Omega}}\right\}, \tag{3.139}$$

where $C_1 = \frac{\kappa}{8}\min\left\{\frac{1}{1+nC_X\sigma}, \frac{\underline{\beta}}{2Q^2\omega C_X L}\right\}$, $C_2 = \frac{\kappa\,\underline{\beta}}{128LQ^2\omega^2}$ and $C_3 = \frac{\kappa\,\underline{\beta}}{128LQ^2C_X^2\|\boldsymbol{b}\|^2_{0,\infty,\Omega}}$, then the matrix of the method (3.14)–(3.16) is an M-matrix and

$$\left\|\tilde{L}_h^{-1}\right\|_{\infty,d} \;\leq\; \frac{2}{\kappa}\max\left\{\frac{L}{\underline{\beta}}, \frac{\|\phi\|_{0,\infty,\Omega}}{\phi_0}\right\}. \tag{3.140}$$

Remark 3.5.3. Let us recall that L is the upper bound for the length of any discrete streamline (cf. Definition 3.4.3, page 70), C_X is the constant from the interpolation inequality (Theorem 4.2.1, page 127), σ is the shape-regularity constant (cf. Assumption 4.2.1), κ is the mesh structure parameter defined in the definition of the energy norm (Definition 3.4.1, page 68) and $\underline{\beta}$ is a positive constant satisfying $|\boldsymbol{b}| \geq \underline{\beta}$ in Ω. Since it holds

$$\Big||\boldsymbol{b}| - |\boldsymbol{b}_K|\Big| \;\leq\; |\boldsymbol{b} - \boldsymbol{b}_K^I| + |\boldsymbol{b}_K^I - \boldsymbol{b}_K| \;\leq\; (C_\Pi + \theta_K)|\boldsymbol{b}|_{1,\infty,K}\,h_K, \tag{3.141}$$

we consider $\underline{\beta}$ being sufficiently small such there also holds $|\boldsymbol{b}_K| \geq \underline{\beta}$.

Proof of Theorem 3.5.3. We would like to apply the M-criterion (Theorem 4.1.5, page 126) and show that the matrix of the method is an M-matrix. We already know that the method matrix L_h (and $\tilde{L}_h$) is a matrix of nonnegative type (Definition 3.2.1, page 57). Hence, it remains to find a vector $\boldsymbol{e} > 0$ for which $L_h\boldsymbol{e} > 0$ (and thus $\tilde{L}_h\boldsymbol{e} > 0$, as well). We construct such a vector from the function ϕ.

The existence of the function ϕ follows from Lemma 4.1.1 (page 124). Since any function $\phi_c = \phi + c$, $c \in \mathbb{R}$, satisfies $\boldsymbol{b} \cdot \nabla\phi_c = \boldsymbol{b} \cdot \nabla\phi \geq \phi_0 > 0$, we can choose ϕ in such a way that $\phi > 0$ in Ω and $\|\phi\|_{0,\infty,\Omega}$ is the smallest possible. We use it in the second part of the proof.

At first, let us assume that $h_K \geq \left(2\frac{\varepsilon L}{\kappa\underline{\beta}}\right)^{1/2}$, for each $K \in \mathcal{T}_h$, and let us define a function $\psi_h \in V_h$ by the relations

$$\psi_h(P_0^s) = 0 \quad \forall s = 1, 2, \ldots, \mathcal{P} \quad \text{and} \quad \psi_h(P_j^s) = \psi_h(P_{j-1}^s) + h_j^s \quad \forall j = 1, 2, \ldots, N_s. \tag{3.142}$$

The function ψ_h is positive inside Ω and satisfies $\boldsymbol{b}_K \cdot \nabla(\psi_h|_K) = |\boldsymbol{b}_K| \geq \underline{\beta}$ and $\psi_h(P_j^s) = \sum_{i=1}^{j} h_i^s \leq j \max_{1\leq i\leq j} h_i^s \leq N_s \max_{1\leq i\leq N_s} h_i^s \leq \max_{i,r}\{N_r h_i^r\} = L$ for all $s = 1, 2, \ldots, \mathcal{P}$, $j = 1, 2, \ldots, N_s$. Using these inequalities we obtain

$$\begin{aligned} a_h(\psi_h, \lambda_j^s) &= -\varepsilon(\nabla\psi_h, \nabla\lambda_j^s)_{\Omega_j^s} + \sum_{K\subset\mathcal{C}_j^s} (\boldsymbol{b}_K \cdot \nabla\psi_h, 1)_K \geq \\ &\geq -\varepsilon\frac{L|\Omega_j^s|}{h_K^2} + \underline{\beta}|\mathcal{C}_j^s| \geq -\frac{1}{2}\kappa\underline{\beta}|\Omega_j^s| + \underline{\beta}|\mathcal{C}_j^s| \geq \frac{1}{2}\kappa\underline{\beta}|\Omega_j^s|. \end{aligned} \tag{3.143}$$

Thus for large h_K the matrix of the method is an M-matrix and there holds

$$\|\tilde{L}_h^{-1}\|_{\infty,d} \leq \frac{\|\psi_h\|_{\infty,d}}{\min_i(\tilde{L}_h\psi_h)_i} \leq \frac{2L}{\kappa\underline{\beta}}. \tag{3.144}$$

If the mesh is refined and $h_K = \left(2\frac{\varepsilon L}{\kappa\underline{\beta}}\right)^{1/2}$ for some $K \in \mathcal{T}_h$ (and the condition $h_K \geq \left(2\frac{\varepsilon L}{\kappa\underline{\beta}}\right)^{1/2}$ for each $K \in \mathcal{T}_h$ is still valid), then for each $K \in \mathcal{T}_h$ it also holds

$$h_K \leq \max_{K\in\mathcal{T}_h} h_K \leq Q \min_{K\in\mathcal{T}_h} h_K = Q\left(2\frac{\varepsilon L}{\kappa\underline{\beta}}\right)^{1/2}. \tag{3.145}$$

Hence, it remains to prove the theorem for meshes satisfying $h_K \leq Q\left(2\frac{\varepsilon L}{\kappa\underline{\beta}}\right)^{1/2}$. In these cases using the inequality (3.139) yields

$$h_K \leq \frac{1}{8}\min\left\{\frac{\phi_0}{C_X\|\boldsymbol{b}\|_{0,\infty,\Omega}|\phi|_{2,\infty,\Omega}}, \frac{\phi_0}{\omega|\phi|_{1,\infty,\Omega}}, \left(\frac{8\phi_0}{\omega C_X|\phi|_{2,\infty,\Omega}}\right)^{1/2}\right\} \tag{3.146}$$

We can use this estimate, the estimate (3.101), the interpolation inequality

(Theorem 4.2.1, page 127) and obtain for any basis function $\lambda_j^s \in V_h$ and $\phi_h = \Pi_h \phi$

$$\sum_{K\subset\Omega_j^s} (\boldsymbol{b}_K \cdot \nabla\phi_h, \lambda_j^s(P_{K,n+1}))_K \;=\; \sum_{K\subset\mathcal{C}_j^s} (\boldsymbol{b}_K \cdot \nabla\phi_h, 1)_K \;=$$

$$= \sum_{K\subset\mathcal{C}_j^s} \Big\{ (\boldsymbol{b}_K - \boldsymbol{b}_K^I, \nabla\phi_h)_K + (\boldsymbol{b}, \nabla\phi_h)_K \Big\} \;=$$

$$= \sum_{K\subset\mathcal{C}_j^s} \Big\{ (\boldsymbol{b}_K - \boldsymbol{b}_K^I, \nabla(\phi_h - \phi) + \nabla\phi)_K + (\boldsymbol{b}, \nabla(\phi_h - \phi))_K + (\boldsymbol{b}, \nabla\phi)_K \Big\} \geq$$

$$\geq \sum_{K\subset\mathcal{C}_j^s} |K| \Big\{ -(n-1)\theta_K h_K \Big(|\phi_h - \phi|_{1,\infty,K} + |\phi|_{1,\infty,K} \Big) -$$

$$- \|\boldsymbol{b}\|_{0,\infty,K} |\phi_h - \phi|_{1,\infty,K} + \phi_0 \Big\} \;\geq$$

$$\geq \sum_{K\subset\mathcal{C}_j^s} |K| \Big\{ \phi_0 - C_X \|\boldsymbol{b}\|_{0,\infty,K} h_K |\phi|_{2,\infty,K} - \omega h_K \left(C_X h_K |\phi|_{2,\infty,K} + |\phi|_{1,\infty,K} \right) \Big\} \geq$$

$$\geq \sum_{K\subset\mathcal{C}_j^s} \Big(\phi_0 - \frac{1}{8}\phi_0 - \frac{1}{8}\phi_0 - \frac{1}{8}\phi_0 \Big) |K| \;=\; \frac{5}{8}\,\phi_0\,|\mathcal{C}_j^s| \;\geq\; \frac{5}{8}\,\kappa\,\phi_0\,|\Omega_j^s|. \tag{3.147}$$

From the condition (3.139) it also follows that $\varepsilon(1 + nC_X\sigma)|\phi|_{2,\infty,\Omega} \leq \frac{1}{8}\,\kappa\,\phi_0$. Consequently, for arbitrary basis function $\lambda_P \in V_h$ there holds

$$\begin{aligned}
\varepsilon(\nabla\phi_h, \nabla\lambda_P)_{\Omega_P} &= \varepsilon(\nabla\phi, \nabla\lambda_P)_{\Omega_P} + \varepsilon(\nabla\phi_h - \nabla\phi, \nabla\lambda_P)_{\Omega_P} \;= \\
&= -\varepsilon(\Delta\phi, \lambda_P)_{\Omega_P} + \varepsilon \sum_{K\subset\Omega_P} (\nabla(\phi_h - \phi), \nabla\lambda_P)_K \;\geq \\
&\geq -\varepsilon \sum_{K\subset\Omega_P} n|\phi|_{2,\infty,K} \frac{|K|}{n+1} - \varepsilon \sum_{K\subset\Omega_P} n|\phi_h - \phi|_{1,\infty,K} \frac{\sigma|K|}{h_K} \;\geq \\
&\geq -\varepsilon \sum_{K\subset\Omega_P} \Big(1 + nC_X\sigma \Big) |\phi|_{2,\infty,K} |K| \;>\; -\frac{1}{8}\,\kappa\,\phi_0\,|\Omega_P|.
\end{aligned}$$

Summing two previous estimates together gives the inequality

$$a_h(\phi_h, \lambda_j^s) \;=\; \varepsilon(\nabla\phi_h, \nabla\lambda_j^s)_{\Omega_j^s} + \sum_{K\subset\Omega_j^s} (\boldsymbol{b}_K \cdot \nabla\phi_h, \lambda_j^s(P_{K,n+1}))_K \;\geq\; \frac{1}{2}\,\kappa\,\phi_0\,|\Omega_j^s|. \tag{3.148}$$

Thus, for the meshes satisfying $h_K \leq Q \left(2\frac{\varepsilon L}{\kappa\beta} \right)^{1/2}$ together with the assumption (3.139) we obtain the estimate

$$\|\tilde{L}_h^{-1}\|_{\infty,d} \;\leq\; \frac{\|\phi_h\|_{\infty,d}}{\min_i(\tilde{L}_h\phi_h)_i} \;\leq\; \frac{2\|\phi\|_{0,\infty,\Omega}}{\kappa\phi_0}. \tag{3.149}$$

□

3.6 L^∞-convergence improvement

3.6.1 Constant data

The above derived method is a multi-dimensional analog to the one-dimensional simple upwind scheme. And, as well as the simple upwind scheme, it possesses

an unpleasant property as to stop converge in $\|\cdot\|_{d,\infty}$-norm when $h \approx \varepsilon$. A one-dimensional remedy is the Il'in-Allen-Southwell scheme which provides a nodally exact solution for the equidistant partition and constant data, especially for the zero right-hand side and constant b. In this case the solution is the boundary-layer function. In Chapter 1 we derived the (zeroth-order) asymptotic expansion of the solution of the convection-diffusion equation in some two-dimensional domains. The multi-dimensional boundary-layer function has in the case of constant data form

$$v_\Gamma(\boldsymbol{x}) \;=\; \exp\left(-\frac{\boldsymbol{b}\cdot\boldsymbol{n}_\Gamma}{\varepsilon}\,\mathrm{dist}_\Gamma(\boldsymbol{x})\right), \tag{3.150}$$

where $\boldsymbol{n}_\Gamma$ is a unit outer normal to $\Gamma \subset \partial\Omega$ and $\mathrm{dist}_\Gamma(\boldsymbol{x})$ is a distance of $\boldsymbol{x} \in \Omega$ from Γ.

Inspired by the one-dimensional case we would like to adjust the derived method in such a way that $R_h v_\Gamma$ forms a nodally exact solution in the vicinity of Γ for an equidistant partition of Ω, constant vector field $\boldsymbol{b}$ and a zero right-hand side f, i.e. in such a way that $L_h R_h v_\Gamma = 0$ in the vicinity of Γ. Hence, we define a bilinear form a_Γ by the relation

$$a_\Gamma(u_h, v_h) \;=\; a_h(u_h, v_h) \;+\; \varepsilon \sum_{K\in\mathcal{T}_h} \left(\boldsymbol{b}\cdot\nabla u_h,\; \frac{\mu}{(\boldsymbol{b}\cdot\boldsymbol{n}_\Gamma)^2}\,\boldsymbol{b}\cdot\nabla v \;+\; \frac{\nu}{\boldsymbol{b}\cdot\boldsymbol{n}_\Gamma}\,\boldsymbol{n}_\Gamma\cdot\nabla v\right)_K, \tag{3.151}$$

where μ and ν are real numbers which have yet to be defined. Due to the factor ε before the sum, the added term does not play an important role in the convection-dominant case.

Since we are interested in the two-dimensional case, we consider an equidistant partition of the domain Ω (in the vicinity of Γ) by congruent triangles (see Figure 3.5). One of each triangle's edges is always parallel to $\boldsymbol{b}$ and one is parallel to Γ. This structure of triangulation is not artificial since it naturally appears when we use the convection-oriented mesh with constant $|\boldsymbol{d}_{K,1}| = h$. Further, we denote the inner angles of each triangle by α, β and γ and choose γ to be the angle included by $\boldsymbol{b}$ and Γ. Thus, $\gamma \le \frac{\pi}{2}$ and $\boldsymbol{b}\cdot\boldsymbol{n}_\Gamma = |\boldsymbol{b}|\cos(\frac{\pi}{2}-\gamma) = |\boldsymbol{b}|\sin\gamma$.

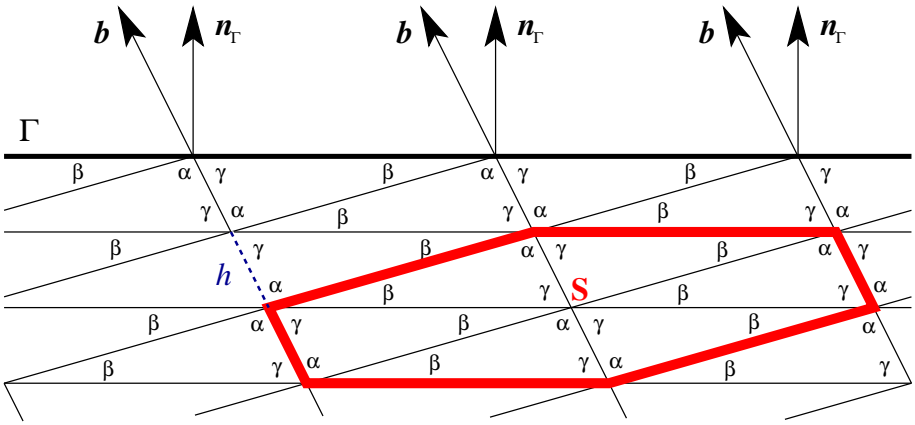

Figure 3.5: Three-directional mesh in the exponential boundary layer.

In each triangle K, we also denote by A, B and C the $L^2(K)$-inner products

$$A = (\nabla\lambda_\beta, \nabla\lambda_\gamma)_K = -\frac{1}{2}\cot\alpha, \tag{3.152}$$

$$B = (\nabla\lambda_\alpha, \nabla\lambda_\gamma)_K = -\frac{1}{2}\cot\beta, \tag{3.153}$$

$$C = (\nabla\lambda_\alpha, \nabla\lambda_\beta)_K = -\frac{1}{2}\cot\gamma, \tag{3.154}$$

where $\lambda_\alpha, \lambda_\beta$ and λ_γ are $P_1(K)$-basis functions corresponding to the vertices with vertex angles α, β and γ, respectively. Since $\alpha < \pi - \gamma$, $\beta < \pi - \gamma$ and $\gamma \le \frac{\pi}{2}$ we obtain the estimates

$$A < -\frac{1}{2}\cot(\pi-\gamma) = \frac{1}{2}\cot\gamma = -C, \tag{3.155}$$

$$B < -\frac{1}{2}\cot(\pi-\gamma) = \frac{1}{2}\cot\gamma = -C, \tag{3.156}$$

$$C \le -\frac{1}{2}\cot\frac{\pi}{2} = 0. \tag{3.157}$$

We can now derive the stencil generated by the method. For an arbitrary inner node S and a corresponding basis function $\lambda_S \in X_h$, the value $a_\Gamma(u_h, \lambda_S)$ can be computed using the scheme

$$2\varepsilon \left[\begin{array}{c|c|c} 0 & B & C \\ \hline A & -2(A+B+C) & A \\ \hline C & B & 0 \end{array}\right] - 4\varepsilon(B+C)\mathrm{Pe}_\Gamma \left[\begin{array}{c|c|c} 0 & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -1 & 0 \end{array}\right] +$$
$$+ \varepsilon(2\mu+\nu)(B+C) \left[\begin{array}{c|c|c} 0 & 1 & 0 \\ \hline 0 & -2 & 0 \\ \hline 0 & 1 & 0 \end{array}\right] + \nu\varepsilon \left[\begin{array}{c|c|c} 0 & B & C \\ \hline -C & -2B & -C \\ \hline C & B & 0 \end{array}\right], \tag{3.158}$$

where in each row one can find the coefficients $a_\Gamma(\lambda_P, \lambda_S)$ with P having identical $\mathrm{dist}_\Gamma(P)$. For the first row there holds $\mathrm{dist}_\Gamma(P) = \mathrm{dist}_\Gamma(S) - h\frac{\boldsymbol{b}\cdot\boldsymbol{n}_\Gamma}{|\boldsymbol{b}|} = \mathrm{dist}_\Gamma(S) - h\sin\gamma$, for the second row $\mathrm{dist}_\Gamma(P) = \mathrm{dist}_\Gamma(S)$ and for the third row one has $\mathrm{dist}_\Gamma(P) = \mathrm{dist}_\Gamma(S) + h\frac{\boldsymbol{b}\cdot\boldsymbol{n}_\Gamma}{|\boldsymbol{b}|} = \mathrm{dist}_\Gamma(S) + h\sin\gamma$. Each column of this stencil then corresponds to a different streamline.

The Péclet number Pe_Γ is defined by the relation $\mathrm{Pe}_\Gamma = \frac{h(\boldsymbol{b}\cdot\boldsymbol{n}_\Gamma)^2}{2\varepsilon|\boldsymbol{b}|}$ and the normal $\boldsymbol{n}_\Gamma$ satisfies either $\boldsymbol{n}_\Gamma = \frac{-1}{|\nabla\lambda_{K,1}|}\nabla\lambda_{K,1}$ or $\boldsymbol{n}_\Gamma = \frac{1}{|\nabla\lambda_{K,3}|}\nabla\lambda_{K,3}$ depending on the orientation of K. During the derivation of the stencil we used the equalities

$$2\frac{|\boldsymbol{b}||K|}{h} = 2\frac{|\boldsymbol{b}|}{h}\frac{v_\alpha h_\alpha}{2} = \frac{|\boldsymbol{b}|}{h}\frac{h(\boldsymbol{b}\cdot\boldsymbol{n}_\Gamma)}{|\boldsymbol{b}|}(h\cos\gamma + h\sin\gamma\cot\beta) =$$
$$= \frac{(\boldsymbol{b}\cdot\boldsymbol{n}_\Gamma)^2}{|\boldsymbol{b}|}h\Big(\cot\gamma + \cot\beta\Big) = -4\varepsilon(B+C)\mathrm{Pe}_\Gamma, \tag{3.159}$$

$$\boldsymbol{b}\cdot\left(\frac{-1}{|\nabla\lambda_{K,1}|}\nabla\lambda_{K,1}\right) = \frac{-1}{|\nabla\lambda_{K,1}|}\frac{|\boldsymbol{b}|}{|\boldsymbol{d}_{K,1}|}\boldsymbol{d}_{K,1}\cdot\nabla\lambda_{K,1} = \frac{|\boldsymbol{b}|}{h|\nabla\lambda_{K,1}|}, \tag{3.160}$$

$$\boldsymbol{b}\cdot\left(\frac{1}{|\nabla\lambda_{K,3}|}\nabla\lambda_{K,3}\right) = \frac{1}{|\nabla\lambda_{K,3}|}\frac{|\boldsymbol{b}|}{|\boldsymbol{d}_{K,1}|}\boldsymbol{d}_{K,1}\cdot\nabla\lambda_{K,3} = \frac{|\boldsymbol{b}|}{h|\nabla\lambda_{K,3}|} \quad \text{and} \tag{3.161}$$

$$-|K||\nabla\lambda_\alpha|^2 = (\nabla\lambda_\alpha, \nabla\lambda_\beta + \nabla\lambda_\gamma)_K = B + C. \tag{3.162}$$

Choosing the coefficients

As we mentioned earlier, we would like to define the coefficients μ and ν in such a way that $a_\Gamma(R_h v_\Gamma, \varphi_S)$ vanishes for each S. Using the above derived stencil we can compute $a_\Gamma(R_h v_\Gamma, \varphi_S)$ simply by multiplying the sum of the first, the second and the third row of the stencil by $v_\Gamma(S)\mathrm{e}^{2\mathrm{Pe}_\Gamma}$, $v_\Gamma(S)$ and $v_\Gamma(S)\mathrm{e}^{-2\mathrm{Pe}_\Gamma}$, respectively, and then sum these multiples together. Consequently, we obtain

$$
\begin{aligned}
&a_\Gamma(R_h v_\Gamma, \varphi_S) \;= \\
&= 2\varepsilon(B+C)v_\Gamma(S)\Big[\mathrm{e}^{2\mathrm{Pe}_\Gamma}(\mu+\nu+1)-2(\mu+\nu+1+\mathrm{Pe}_\Gamma)+\mathrm{e}^{-2\mathrm{Pe}_\Gamma}(\mu+\nu+1+2\mathrm{Pe}_\Gamma)\Big] = \\
&= 2\varepsilon(B+C)v_\Gamma(S)\Big[(\mu+\nu+1)4\sinh^2(\mathrm{Pe}_\Gamma) - 2\mathrm{Pe}_\Gamma\,\mathrm{e}^{-\mathrm{Pe}_\Gamma}2\sinh(\mathrm{Pe}_\Gamma)\Big] = \\
&= 8\varepsilon(B+C)\sinh^2(\mathrm{Pe}_\Gamma)v_\Gamma(S)\left[\mu+\nu+1-\mathrm{Pe}_\Gamma\frac{\mathrm{e}^{-\mathrm{Pe}_\Gamma}}{\sinh(\mathrm{Pe}_\Gamma)}\right] = \\
&= 8\varepsilon(B+C)\sinh^2(\mathrm{Pe}_\Gamma)v_\Gamma(S)\Big[\mu+\nu+1-\mathrm{Pe}_\Gamma\Big(\coth(\mathrm{Pe}_\Gamma)-1\Big)\Big]. \qquad (3.163)
\end{aligned}
$$

Thus, choosing $\mu+\nu = \mathcal{R}_\Gamma = \mathrm{Pe}_\Gamma\Big(\coth(\mathrm{Pe}_\Gamma)-1\Big)-1$ leads to the equality $a_\Gamma(R_h v_\Gamma, \varphi_S) = 0$. Therefore, if the discrete maximum principle is satisfied then for constant data and the three-directional mesh the method provides a solution u_h satisfying $\|u-u_h\|_{d,\infty} = \mathcal{O}(\varepsilon)$ in the exponential boundary layer near the boundary Γ (cf. proof of the uniform convergence of the Il'in-Allen-Southwell scheme, section 2.2).

In order to fulfil the discrete maximum principle the method matrix has to be of nonnegative type (see Definition 3.2.1, page 57), it means that the off-diagonal entries have to be nonpositive. Hence, we obtain the following set of constraints

$$
\begin{aligned}
\mu+\nu &= \mathcal{R}_\Gamma, && (3.164)\\
C(\nu+2) &\le 0, && (3.165)\\
2A-\nu C &\le 0, && (3.166)\\
(2\mu+\nu)(B+C)+B(\nu+2) &\le 0. && (3.167)
\end{aligned}
$$

If $C=0$, then from (3.166) and (3.167) it follows that it suffices to consider $2A\le 0$ and $2(\mu+\nu+1)B = 2(\mathcal{R}_\Gamma+1)B \le 0$, i.e. $A\le 0$ and $B\le 0$. In the case when $C<0$, then from (3.165)–(3.167) we obtain the conditions $-2\le\nu\le 2\frac{A}{C}$ and $\nu\le 2\mathcal{R}_\Gamma+2(\mathcal{R}_\Gamma+1)\frac{B}{C}$. If we again assume that $A\le 0$ and $B\le 0$, then $0\le\frac{A}{C}$ and $0\le\frac{B}{C}$. Consequently, any ν satisfying $-2\le\nu\le 2\mathcal{R}_\Gamma<0$ is admissible. Hence, we choose $\nu=\nu_1=2\mathcal{R}_\Gamma$ and $\mu=\mu_1=\mathcal{R}_\Gamma-\nu_1=-\mathcal{R}_\Gamma$.

The assumption $A,B,C\le 0$ is in fact the assumption (3.8) (page 55) meaning that the angles α,β and γ are not obtuse. However, in thin boundary or interior layers it would be convenient to use elements with obtuse angles. Thus, if we allow obtuse angles and consider only the restrictions $A\le\mathcal{R}_\Gamma C = |\mathcal{R}_\Gamma C|$ and $B\le\mathcal{R}_\Gamma C = |\mathcal{R}_\Gamma C|$ (together with $C\le 0$), we fulfil the conditions (3.164)–(3.167) if we take $\mu=\mu_2-\mathcal{R}_\Gamma(2\mathcal{R}_\Gamma+3)$ and $\nu=\nu_2=2\mathcal{R}_\Gamma(\mathcal{R}_\Gamma+2)$. Indeed, in

this case there holds

$$\begin{aligned}
\mu+\nu &= -\mathcal{R}_\Gamma(2\mathcal{R}_\Gamma+3)+2\mathcal{R}_\Gamma(\mathcal{R}_\Gamma+2) \;=\; \mathcal{R}_\Gamma, && (3.168)\\
C(\nu+2) &= 2C(\mathcal{R}_\Gamma(\mathcal{R}_\Gamma+2)+1) \;=\; 2C(\mathcal{R}_\Gamma+1)^2 \;\le\; 0, && (3.169)\\
2A-\nu C &= 2(A-\mathcal{R}_\Gamma(\mathcal{R}_\Gamma+2)C) \;\le\; 2C(\mathcal{R}_\Gamma-\mathcal{R}_\Gamma(\mathcal{R}_\Gamma+2)) =\\
&= -2C\mathcal{R}_\Gamma(\mathcal{R}_\Gamma+1) \;\le\; 0 && (3.170)\\
(2\mu+\nu)(B+C)+B(\nu+2) &= -2\mathcal{R}_\Gamma(\mathcal{R}_\Gamma+1)(B+C)+2B(\mathcal{R}_\Gamma+1)^2 \;\le && (3.171)\\
&\le -2\mathcal{R}_\Gamma(\mathcal{R}_\Gamma+1)(\mathcal{R}_\Gamma+1)C+2C\mathcal{R}_\Gamma(\mathcal{R}_\Gamma+1)^2 \;=\; 0.
\end{aligned}$$

All considered cases are summarized in the following table Let us also mention

$C\le$	$A\le$	$B\le$	μ	ν	method's property
0	0	0	0	0	upwind scheme, DMP
0	0	0	$-\mathcal{R}_\Gamma$	$2\mathcal{R}_\Gamma$	DMP, UNI
0	$\mathcal{R}_\Gamma C$	$\mathcal{R}_\Gamma C$	$-\mathcal{R}_\Gamma(2\mathcal{R}_\Gamma+3)$	$2\mathcal{R}_\Gamma(\mathcal{R}_\Gamma+2)$	DMP, UNI, OBT

Table 3.1: Different choices of the coefficients μ and ν together with the inner angles restriction lead for the constant data to methods with the following properties: the fulfilment of the discrete maximum principle (DMP), the uniform convergence in $\|\cdot\|_{d,\infty}$ norm with respect to ε in the vicinity of the boundary Γ (UNI) and admissibility of obtuse inner angles (OBT).

that in the most interesting case $\mathrm{Pe}_\Gamma \gg 0$ the parameter $\mathcal{R}_\Gamma$ is close to -1. Thus, the conditions $A \le \mathcal{R}_\Gamma C$ and $B \le \mathcal{R}_\Gamma C$ do not significantly restrict the angles since the inequalities $A < -C$ and $B < -C$ hold for each triangle (cf. inequalities (3.155) and (3.156)). For small Pe_Γ (approximately smaller than 3) the angles restriction is more significant (see Figure 3.6).

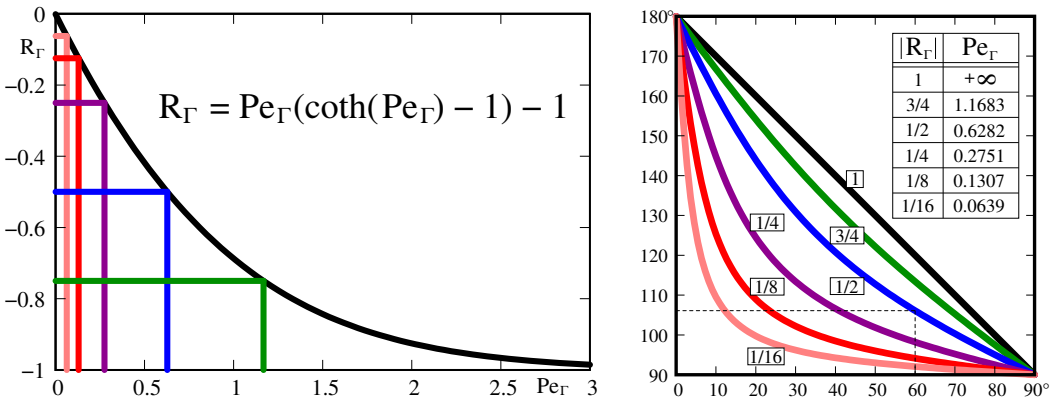

Figure 3.6: Using the Péclet number Pe_Γ one can compute the parameter $\mathcal{R}_\Gamma$ (left) which restricts the maximum angle in triangles. For instance, when $\mathrm{Pe}_\Gamma \approx 0.6282$, then $\mathcal{R}_\Gamma \approx -1/2$ and the angle α (and β) has to satisfy the inequality $-\frac{1}{2}\cot\alpha = A \le -\frac{1}{2}C = \frac{1}{4}\cot\gamma$. Consequently, if (for instance) $\gamma = 60°$ then $\alpha, \beta \le \operatorname{acot}(-1/(2\sqrt{3})) \approx 106.1°$ (right), which is a sharper bound as compared to standard $\alpha, \beta < 180° - 60° = 120°$. The right figure shows the restriction curves for several choices of $\mathcal{R}_\Gamma$ (or Pe_Γ).

3.6.2 Non-constant data

In the previous adjustment of the method we have considered constant data and constant μ and ν. Let us now investigate the non-constant case and consider different values μ_K, ν_K for each $K \in \mathcal{T}_h$.

At first, we determine necessary coercivity conditions. We start with the following lemma.

Lemma 3.6.1. *Let $\boldsymbol{n}_K$ be any constant unit vector satisfying $\boldsymbol{b}_K \cdot \boldsymbol{n}_K > 0$ and let us denote $\mathrm{Pe}_K = \frac{|\boldsymbol{d}_{K,1}|(\boldsymbol{b}_K \cdot \boldsymbol{n}_K)^2}{2\varepsilon|\boldsymbol{b}_K|}$. If the coefficients μ_K and ν_K satisfy*

$$\mathrm{Pe}_K + \mu_K \ > \ \frac{1}{4}\nu_K^2, \tag{3.172}$$

then there exists $\alpha > 0$ such that for all $v_h \in V_h$ there holds

$$\left(\boldsymbol{b}_K \cdot \nabla v_h, \frac{\varepsilon\mu_K}{(\boldsymbol{b}_K \cdot \boldsymbol{n}_K)^2}\,\boldsymbol{b}_K \cdot \nabla v_h + \frac{\varepsilon\nu_K}{\boldsymbol{b}_K \cdot \boldsymbol{n}_K}\,\boldsymbol{n}_K \cdot \nabla v_h\right)_K \ \geq$$
$$\geq \ (\alpha - 1)\varepsilon|v_h|_{1,K}^2 + (2\alpha - 1)\frac{|\boldsymbol{d}_{K,1}|}{2|\boldsymbol{b}_K|}\|\boldsymbol{b}_K \cdot \nabla v_h\|_{0,K}^2. \tag{3.173}$$

Proof. Since $\|\boldsymbol{n}_K \cdot \nabla v_h\|_{0,K} \leq |v_h|_{1,K}$, then using the Cauchy-Schwarz-Bunyakovsky inequality we can estimate the left-hand side of (3.173) by

$$\left(\boldsymbol{b}_K \cdot \nabla v_h, \frac{\varepsilon\mu_K}{(\boldsymbol{b}_K \cdot \boldsymbol{n}_K)^2}\,\boldsymbol{b}_K \cdot \nabla v_h + \frac{\varepsilon\nu_K}{\boldsymbol{b}_K \cdot \boldsymbol{n}_K}\,\boldsymbol{n}_K \cdot \nabla v_h\right)_K \ \geq$$
$$\geq \ \frac{\varepsilon\mu_K}{(\boldsymbol{b}_K \cdot \boldsymbol{n}_K)^2}\,\|\boldsymbol{b}_K \cdot \nabla v_h\|_{0,K}^2 - \frac{\varepsilon|\nu_K|}{\boldsymbol{b}_K \cdot \boldsymbol{n}_K}\|\boldsymbol{b}_K \cdot \nabla v_h\|_{0,K}|v_h|_{1,K}. \tag{3.174}$$

Thus, denoting $X = \left(\frac{|\boldsymbol{d}_{K,1}|}{2|\boldsymbol{b}_K|}\|\boldsymbol{b}_K \cdot \nabla v_h\|_{0,K}^2\right)^{1/2}$ and $Y = \left(\varepsilon|v_h|_{1,K}^2\right)^{1/2}$ it suffices to prove that there exists $\alpha > 0$ such that for all $X, Y \geq 0$ it holds

$$\frac{2\varepsilon|\boldsymbol{b}_K|\mu_K}{|\boldsymbol{d}_{K,1}|(\boldsymbol{b}_K \cdot \boldsymbol{n}_K)^2}\,X^2 - \left(\frac{2\varepsilon|\boldsymbol{b}_K|\nu_K^2}{|\boldsymbol{d}_{K,1}|(\boldsymbol{b}_K \cdot \boldsymbol{n}_K)^2}\right)^{1/2} XY \ \geq \ (\alpha - 1)Y^2 + (2\alpha - 1)X^2, \tag{3.175}$$

which can be rewritten in the form

$$\left(\frac{|\nu_K|}{2\mathrm{Pe}_K^{1/2}}X - Y\right)^2 + \left(\frac{\mu_K}{\mathrm{Pe}_K} + 1 - \frac{\nu_K^2}{4\mathrm{Pe}_K}\right)X^2 \ \geq \ \alpha(Y^2 + 2X^2) \tag{3.176}$$

If $X = 0$ then one can take any $\alpha \in (0, 1]$. If $X > 0$ then (due to the assumption (3.172)) the left-hand side of (3.176) is positive. Consequently, there surely exists sufficiently small positive α such that the right-hand side remains smaller than the left-hand side.

□

Remark 3.6.1. The previous lemma did not specify the exact value of α. One possible choice is $\alpha = 1 - \frac{\nu_0}{\mu_0}$, where $\mu_0 = \frac{\mu_K}{\mathrm{Pe}_K} + 3 > \frac{\nu_K^2}{4\mathrm{Pe}_K} + 2 = \nu_0$. Then $\frac{\mu_K}{\mathrm{Pe}_K} = \mu_0 - 3$, $\frac{\nu_K^2}{4\mathrm{Pe}_K} = \nu_0 - 2$ and from (3.175) it follows that we need to verify the validity of the inequality

$$(\mu_0 - 3)X^2 - 2\sqrt{\nu_0 - 2}XY + \frac{\nu_0}{\mu_0}\,Y^2 + \left(2\,\frac{\nu_0}{\mu_0} - 1\right)X^2 \geq 0. \tag{3.177}$$

This inequality can be equivalently rewritten in the form

$$\left(\sqrt{\frac{\mu_0(\nu_0-2)}{\nu_0}}X-\sqrt{\frac{\nu_0}{\mu_0}}Y\right)^2+2\left(\sqrt{\frac{\nu_0}{\mu_0}}-\sqrt{\frac{\mu_0}{\nu_0}}\right)^2X^2\geq 0. \tag{3.178}$$

The maximal possible value of α can be obtained by investigation of the eigenvalues of the matrix corresponding to the bilinear form

$$G(X,Y) \;=\; \left(\frac{\mu_K}{\mathrm{Pe}_K}+(1-2\alpha)\right)X^2-\frac{|\nu_K|}{\mathrm{Pe}_K^{1/2}}XY+(1-\alpha)Y^2. \tag{3.179}$$

This bilinear form is positive-semidefinite if $\alpha\leq\alpha_0=\min\left\{1,\frac{1}{2}\left(1+\frac{\mu_K}{\mathrm{Pe}_K}\right)\right\}$ and $\frac{\nu_K^2}{\mathrm{Pe}_K}\leq 4\left(\frac{\mu_K}{\mathrm{Pe}_K}+(1-2\alpha)\right)(1-\alpha)$, i.e. in the case when

$$g(\alpha) \;=\; 2\alpha^2-\alpha\left(\frac{\mu_K}{\mathrm{Pe}_K}+3\right)+\frac{\mu_K}{\mathrm{Pe}_K}+1-\frac{\nu_K^2}{4\mathrm{Pe}_K} \;\geq\; 0. \tag{3.180}$$

As we already know, since $g(\alpha_0)=-\frac{\nu_K^2}{4\mathrm{Pe}_K}<0$, the inequality (3.180) has a solution $\alpha\in(0,\alpha_0)$ only if $g(0)>0$, i.e. when $\frac{\mu_K}{\mathrm{Pe}_K}+1>\frac{\nu_K^2}{4\mathrm{Pe}_K}$. The smaller of the two solutions of the equation $g(\alpha)=0$ is the maximal possible value of α and it has a form

$$\alpha_{max}^K \;=\; \frac{2\left(\frac{\mu_K}{\mathrm{Pe}_K}+1-\frac{\nu_K^2}{4\mathrm{Pe}_K}\right)}{\frac{\mu_K}{\mathrm{Pe}_K}+3+\sqrt{\left(\frac{\mu_K}{\mathrm{Pe}_K}-1\right)^2+2\frac{\nu_K^2}{\mathrm{Pe}_K}}} \;=\; \frac{2(\mu_0-\nu_0)}{\mu_0+\sqrt{\mu_0^2-8(\mu_0-\nu_0)}}. \tag{3.181}$$

Definition 3.6.1. *For each K we denote by*

$$\boldsymbol{w}_K \;=\; (P_{K,n+1}-C_K)+\frac{\varepsilon\mu_K}{(\boldsymbol{b}_K\cdot\boldsymbol{n}_K)^2}\,\boldsymbol{b}_K+\frac{\varepsilon\nu_K}{\boldsymbol{b}_K\cdot\boldsymbol{n}_K}\boldsymbol{n}_K \tag{3.182}$$

the vector that we use for stabilization. Further, we define a bilinear form

$$a_h^\infty(u,v) \;=\; \varepsilon(\nabla u,\nabla v)_\Omega+\sum_{K\in\mathcal{T}_h}\Big\{(\boldsymbol{b}_K\cdot\nabla u,v)_K+(-\varepsilon\Delta u+\boldsymbol{b}_K\cdot\nabla u,\boldsymbol{w}_K\cdot\nabla v)_K\Big\}.$$

The coefficients μ_K and ν_K will be defined later.

Lemma 3.6.2. *For each K let the assumption (3.172) be fulfilled and let us denote $\alpha_\infty=\min\limits_{K\in\mathcal{T}_h}\alpha_{max}^K$. Then*

$$a_h^\infty(v_h,v_h) \;\geq\; \alpha_\infty|||v_h|||_b^2. \tag{3.183}$$

Proof. From the inequalities (3.64) and (3.66) (page 68) it follows that

$$\sum_{K\in\mathcal{T}_h}(\boldsymbol{b}_K\cdot\nabla v_h,v_h+(P_{K,n+1}-C_K)\cdot\nabla v_h)_K \;\geq$$

$$\geq\; \alpha_\infty\frac{\omega\kappa}{2}\|v_h\|_{0,\Omega}^2+(1-\alpha_\infty)\sum_{K\in\mathcal{T}_h}\frac{|\boldsymbol{d}_{K,1}|}{2|\boldsymbol{b}_K|}\|\boldsymbol{b}_K\cdot\nabla v_h\|_{0,K}^2. \tag{3.184}$$

Consequently, using the inequality (3.173) we obtain

$$
\begin{aligned}
&a_h^\infty(v_h,v_h) \;\geq \\
&\geq \sum_{K\in\mathcal{T}_h}\left\{\varepsilon|v_h|_{1,K}^2+\alpha_\infty\frac{\omega\kappa}{2}\|v_h\|_{0,K}^2+(1-\alpha_\infty)\sum_{K\in\mathcal{T}_h}\frac{|\boldsymbol{d}_{K,1}|}{2|\boldsymbol{b}_K|}\|\boldsymbol{b}_K\cdot\nabla v_h\|_{0,K}^2+\right.\\
&\qquad\left.+(\alpha_{max}^K-1)\varepsilon|v_h|_{1,K}^2+(2\alpha_{max}^K-1)\frac{|\boldsymbol{d}_{K,1}|}{2|\boldsymbol{b}_K|}\|\boldsymbol{b}_K\cdot\nabla v_h\|_{0,K}^2\right\}\geq\\
&\geq \alpha_\infty\left\{\varepsilon|v_h|_{1,\Omega}^2+\frac{\omega\kappa}{2}\|v_h\|_{0,\Omega}^2+\sum_{K\in\mathcal{T}_h}\frac{|\boldsymbol{d}_{K,1}|}{2|\boldsymbol{b}_K|}\|\boldsymbol{b}_K\cdot\nabla v_h\|_{0,K}^2\right\}. \qquad (3.185)
\end{aligned}
$$

□

Remark 3.6.2. If we take $\mu_K=\nu_K=0$ for each K, then the assumption (3.172) is fulfilled and $\alpha_{max}^K=\frac{2(0+1-0)}{3+\sqrt{(0-1)^2+0}}=\frac{1}{2}$. Hence, $\alpha_\infty=\frac{1}{2}$ and we obtain the original estimate (3.61) (page 68).

In the case of constant data, the three-directional mesh and $\mu_K=\mu_1=-\mathcal{R}_\Gamma$, $\nu_K=\nu_1=2\mathcal{R}_\Gamma$ we can compute $\alpha_0^{(1)}=1-\frac{\nu_0}{\mu_0}=1-\frac{2+\nu_1^2/(4\mathrm{Pe}_\Gamma)}{3+\mu_1/\mathrm{Pe}_\Gamma}$ and $\alpha_{max}^{(1)}$ using the formula (3.181). Similarly, we obtain functions $\alpha_0^{(2)}$ and $\alpha_{max}^{(2)}$ if we consider $\mu_K=\mu_2=-\mathcal{R}_\Gamma(2\mathcal{R}_\Gamma+3)$ and $\nu_K=\nu_2=2\mathcal{R}_\Gamma(\mathcal{R}_\Gamma+2)$ (see Figure 3.7). We observe that for any positive Pe_Γ there holds $\alpha_0^{(1)}\geq 0.2984$, $\alpha_0^{(2)}\geq 0.2933$, $\alpha_{max}^{(1)}\geq 0.3839$ and $\alpha_{max}^{(2)}\geq 0.3675$.

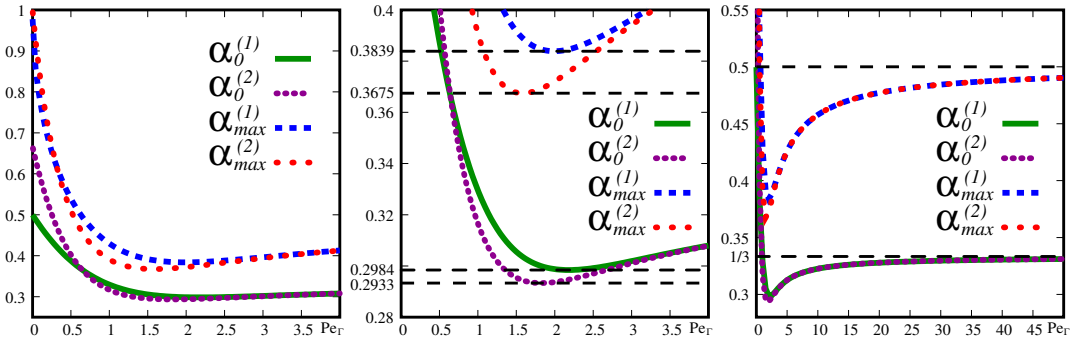

Figure 3.7: For constant data, the three-directional mesh and two considered choices of μ,ν the estimate (3.183) holds with $\alpha_{max}^{(1)}$ and $\alpha_{max}^{(2)}$. For arbitrary Pe_Γ there holds $\alpha_{max}^{(1)}\geq 0.3839$ and $\alpha_{max}^{(2)}\geq 0.3675$. We also observe that the choices $\alpha=\alpha_0^{(1)}$ and $\alpha=\alpha_0^{(2)}$ are suboptimal (in comparison with $\alpha_{max}^{(1)}$ and $\alpha_{max}^{(2)}$). The middle picture shows the detail whereas in the right picture one can see the values of all functions for large Pe_Γ.

Lemma 3.6.3. *Let there exist positive constants C_μ and C_ν independent of ε and h_K such that for each $K\in\mathcal{T}_h$ the coefficients μ_K and ν_K satisfy*

$$
|\mu_K|\;\leq\; C_\mu\min\{\mathrm{Pe}_K,1\}\qquad\text{and}\qquad|\nu_K|\;\leq\; C_\nu\min\{\mathrm{Pe}_K,1\}. \qquad (3.186)
$$

Then there exists positive constant $C_w>0$ independent of ε and h_K such that for each $K\in\mathcal{T}_h$ the vector $\boldsymbol{w}_K$ satisfies $|\boldsymbol{w}_K|\leq C_w h_K$. Moreover, for each $K\in\mathcal{T}_h$ there also holds $|\boldsymbol{w}_K-(P_{K,n+1}-C_K)|\to 0$ as $\varepsilon\to 0$.

Proof. Since there holds

$$|\boldsymbol{w}_K| \;\leq\; h_K + \frac{1}{2}\,\frac{|\mu_K|}{\mathrm{Pe}_K}\,h_K + \frac{|\nu_K|}{\mathrm{Pe}_K}\,\frac{\boldsymbol{b}_K\cdot\boldsymbol{n}_K}{2|\boldsymbol{b}_K|}\,h_K, \tag{3.187}$$

we can take $C_w = 1 + \frac{1}{2}(C_\mu + C_\nu)$. The second statement of the lemma follows directly from the inequality

$$|\boldsymbol{w}_K - (P_{K,n+1} - C_K)| \;\leq\; \varepsilon\left(\frac{C_\mu|\boldsymbol{b}_K|}{(\boldsymbol{b}_K\cdot\boldsymbol{n}_K)^2} + \frac{C_\nu}{\boldsymbol{b}_K\cdot\boldsymbol{n}_K}\right) \tag{3.188}$$

and the fact that $\boldsymbol{b}_K\cdot\boldsymbol{n}_K > 0$. □

Corollary 3.6.1. Let the assumptions of Lemma 3.6.1, Lemma 3.6.3 and Theorem 3.5.1 (page 75) be fulfilled. If we replace the bilinear form a_h by the bilinear form a_h^∞ then there holds the same estimate as in Theorem 3.5.1.

Proof. While Lemma 3.6.1 provides the coercivity of the bilinear form a_h^∞, Lemma 3.6.3 ensures that we can estimate $\|\boldsymbol{w}_K\nabla\xi_h\|_{0,K} \leq C_w h_K|\xi_h|_{1,K}$. Thus, the proof is an analog to the proof of Theorem 3.5.1. □

Remark 3.6.3. In the case of constant data and the three-directional mesh both considered choices of the coefficients μ and ν satisfy the conditions (3.172) and (3.186). Indeed, for every (positive) Pe_Γ there holds (see Figure 3.8 left)

$$\mathrm{Pe}_\Gamma + \mu_2 \;\geq\; \mathrm{Pe}_\Gamma + \mu_1 \;\geq\; \mathrm{Pe}_\Gamma \;\geq\; \tfrac{1}{4}\nu_2^2 \;\geq\; \tfrac{1}{4}\nu_1^2, \tag{3.189}$$

where μ_1, ν_1, μ_2 and ν_2 equals to $-\mathcal{R}_\Gamma, 2\mathcal{R}_\Gamma, -\mathcal{R}_\Gamma(2\mathcal{R}_\Gamma + 3)$ and $2\mathcal{R}_\Gamma(\mathcal{R}_\Gamma + 2)$, respectively (cf. Remark 3.6.2 and Table 3.1, page 89).

Further, the coefficients μ_1, ν_1, μ_2 and ν_2 also satisfy the inequalities (see Figure 3.8 middle and right)

$$\begin{aligned}
\mu_1 &\leq \max(\mathrm{Pe}_\Gamma, 1), && (3.190)\\
|\nu_1| &\leq 2\max(\mathrm{Pe}_\Gamma, 1), && (3.191)\\
\mu_2 &\leq 3\max(\mathrm{Pe}_\Gamma, 3/8) \leq 3\max(\mathrm{Pe}_\Gamma, 1), && (3.192)\\
|\nu_2| &\leq 4\max(\mathrm{Pe}_\Gamma, 1/2) \leq 4\max(\mathrm{Pe}_\Gamma, 1). && (3.193)
\end{aligned}$$

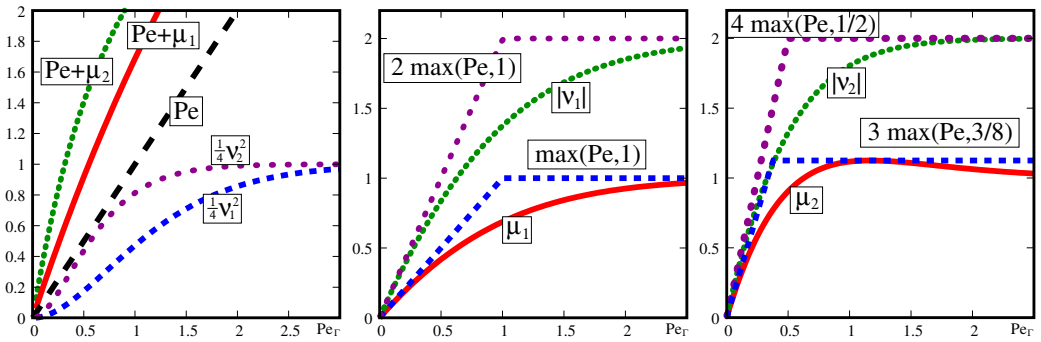

Figure 3.8: For constant data and the three-directional mesh both considered choices of the coefficients μ and ν satisfy the conditions (3.172) and (3.186).

Choosing the coefficients

Let us now describe, how we choose the coefficients μ_K and ν_K for a general oriented mesh (i.e. not necessarily three-directional). At first, we have to define the vectors $\boldsymbol{n}_K$ for K laying inside Ω. For each element K with one edge laying on the boundary Γ_+ we set $\boldsymbol{n}_K = \frac{-1}{|\nabla\lambda_{K,1}|}\nabla\lambda_{K,1}$. In addition, if $\mathcal{C}^s_{N_s} = K \cup T$ is a cluster containing K, we set $\boldsymbol{n}_T = \frac{1}{|\nabla\lambda_{T,3}|}\nabla\lambda_{T,3}$. If Q is any element laying in some cluster on *the same* discrete streamline we set $\boldsymbol{n}_Q = \frac{-1}{|\nabla\lambda_{Q,1}|}\nabla\lambda_{Q,1}$ if Q lies on the same side from this streamline as K. Otherwise, we set $\boldsymbol{n}_Q = \frac{1}{|\nabla\lambda_{Q,3}|}\nabla\lambda_{Q,3}$ (see Figure 3.9 left). Thus, for a fixed discrete streamline $\mathcal{S}$ all elements K with $\boldsymbol{d}_{K,1} \in \mathcal{S}$ are divided into two groups: elements laying on one side from $\mathcal{S}$ satisfy $\boldsymbol{n}_K = \frac{-1}{|\nabla\lambda_{K,1}|}\nabla\lambda_{K,1}$, elements laying on the other side from $\mathcal{S}$ satisfy $\boldsymbol{n}_K = \frac{1}{|\nabla\lambda_{K,3}|}\nabla\lambda_{K,3}$.

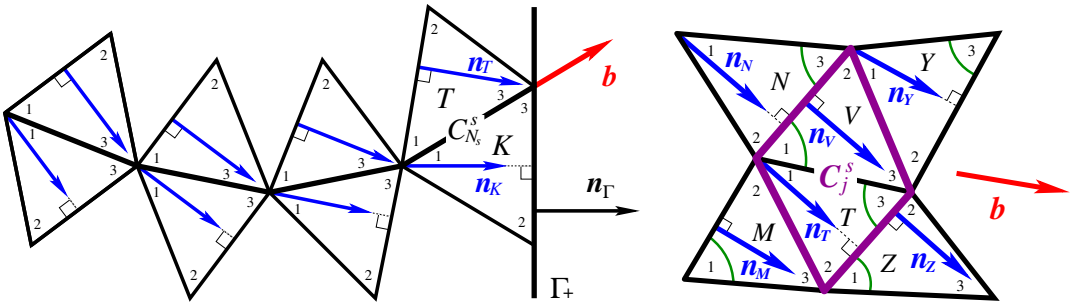

Figure 3.9: Parts of $\mathcal{T}_h$ used for the definition of $\boldsymbol{n}_K$, ν^s_j, ν_V and ν_T.

For each element $K \in \mathcal{T}_h$ we define $\mathrm{Pe}_K = \frac{|\boldsymbol{d}_{K,1}|(\boldsymbol{b}_K\cdot\boldsymbol{n}_K)^2}{2\varepsilon|\boldsymbol{b}_K|}$ and for each cluster $\mathcal{C}^s_j$ we denote $\mathrm{Pe}^s_j = \frac{1}{|\mathcal{C}^s_j|}\sum_{K\subset\mathcal{C}^s_j}|K|\mathrm{Pe}_K$. Consequently, we can define $\mathcal{R}^s_j = \mathrm{Pe}^s_j\Big(\coth(\mathrm{Pe}^s_j) - 1\Big) - 1 \in (-1, 0)$. For each element $K \in \mathcal{T}_h$ we also define values

$$K_i = -\frac{1}{2}\cot\alpha_{K,i}, \qquad i = 1, 2, 3, \tag{3.194}$$

where $\alpha_{K,i}$ are the inner angles of the element (triangle) K corresponding to the vertex $P_{K,i}$.

Let us now consider a cluster $\mathcal{C}^s_j = V \cup T$ with $\boldsymbol{n}_T = \frac{-1}{|\nabla\lambda_{T,1}|}\nabla\lambda_{T,1}$ and $\boldsymbol{n}_V = \frac{1}{|\nabla\lambda_{V,3}|}\nabla\lambda_{V,3}$. Further, let N be the element neighboring to V satisfying $\boldsymbol{n}_N = \boldsymbol{n}_V$ and let Y be the remaining element neighboring to V. Similarly, we denote by Z the element neighboring to T satisfying $\boldsymbol{n}_Z = \boldsymbol{n}_T$ and by M the remaining element neighboring to T (see Figure 3.9 (right)). For each element $K \in \mathcal{T}_h$ we also assume that if $\boldsymbol{n}_K = \frac{-1}{|\nabla\lambda_{K,1}|}\nabla\lambda_{K,1}$ then $K_3 < 0$ and if $\boldsymbol{n}_K = \frac{1}{|\nabla\lambda_{K,3}|}\nabla\lambda_{K,3}$ then $K_1 < 0$ (corresponding acute angles are highlighted in Figure 3.9 (right)). Then we denote

$$\nu^s_j \;:=\; \max\left\{-1 - \frac{Y_3}{V_1}, -1 - \frac{M_1}{T_3}, 2\mathcal{R}^s_j(\mathcal{R}^s_j + 2)\right\}, \tag{3.195}$$

with the values Y_3, V_1, M_1 and T_3 computed using (3.194).

Consequently, the coefficients ν_V and ν_T are defined as follows

$$\nu_V \;:=\; \nu^s_j\frac{\max\{V_1, T_3\}}{V_1} \qquad \text{and} \qquad \nu_T \;:=\; \nu^s_j\frac{\max\{V_1, T_3\}}{T_3}. \tag{3.196}$$

Since $0 > \nu_j^s \geq 2\mathcal{R}_j^s(\mathcal{R}_j^s + 2) > -2$, there holds

$$\nu_V = \nu_j^s \frac{\max\{V_1, T_3\}}{V_1} = \nu_j^s \min\left\{1, \frac{T_3}{V_1}\right\} \geq \nu_j^s > -2, \tag{3.197}$$

$$\nu_T = \nu_j^s \frac{\max\{V_1, T_3\}}{T_3} = \nu_j^s \min\left\{1, \frac{V_1}{T_3}\right\} \geq \nu_j^s > -2. \tag{3.198}$$

Therefore, we may define the parameters $\mathcal{R}_V$ and $\mathcal{R}_T$ by the relations

$$\mathcal{R}_V := -1 + \sqrt{1 + \nu_V/2} \quad \Rightarrow \quad \nu_V = 2\mathcal{R}_V(\mathcal{R}_V + 2), \tag{3.199}$$

$$\mathcal{R}_T := -1 + \sqrt{1 + \nu_T/2} \quad \Rightarrow \quad \nu_T = 2\mathcal{R}_T(\mathcal{R}_T + 2). \tag{3.200}$$

Using the parameters $\mathcal{R}_V$ and $\mathcal{R}_T$ we define $\mu_V = \mathcal{R}_V - \nu_V = -\mathcal{R}_V(2\mathcal{R}_V + 3)$ and $\mu_T = \mathcal{R}_T - \nu_T = -\mathcal{R}_T(2\mathcal{R}_T + 3)$. Moreover, we assume that $N_1, V_2, V_3 \leq \mathcal{R}_V V_1$ and $T_1, T_2, Z_3 \leq \mathcal{R}_T T_3$. While the assumptions $V_2, V_3 \leq \mathcal{R}_V V_1$ and $T_1, T_2 \leq \mathcal{R}_T T_3$ are the same as for the three-directional mesh, the assumptions $N_1 \leq \mathcal{R}_V V_1$ and $Z_3 \leq \mathcal{R}_T T_3$ are additional. Let us recall that $\boldsymbol{n}_N = \boldsymbol{n}_V$ and $\boldsymbol{n}_Z = \boldsymbol{n}_T$, hence we assume some restriction on the angles of the triangles with the same $\boldsymbol{n}_K$. Conversely, from this observation it immediately follows that we also assume that $V_3 \leq \mathcal{R}_N N_3$ and $T_1 \leq \mathcal{R}_Z Z_1$.

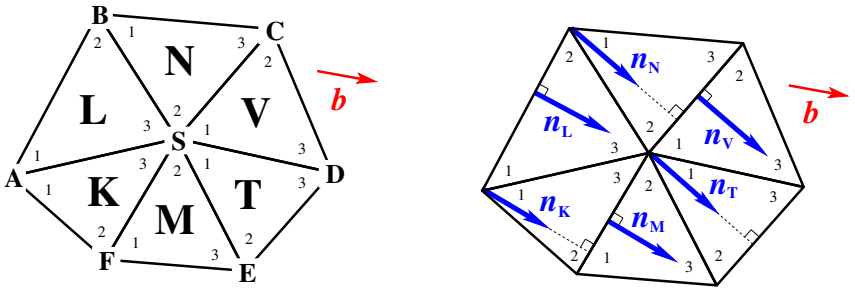

Figure 3.10: A structure of the domain Ω_S (left) and a choice of the vectors $\boldsymbol{n}_Q$, $Q \subset \Omega_S$ (right).

Let us verify the fulfilment of the discrete maximum principle, i.e. let us show that the matrix of the method is of nonnegative type (cf. Definition 3.2.1, page 57). For an arbitrary inner node S and the corresponding basis function $\lambda_S \in X_h$, the stencil of the method is computed using values $a_h^\infty(\lambda_X, \lambda_S)$, where λ_X is the basis function corresponding to the node X neighboring to S. Let the domain $\Omega_S = \operatorname{supp} \lambda_S$ be a hexagon depicted in Figure 3.10 (left) and let for each element $Q \subset \Omega_S$ the vector $\boldsymbol{n}_Q$ be either $\frac{-1}{|\nabla\lambda_{Q,1}|}\nabla\lambda_{Q,1}$ or $\frac{1}{|\nabla\lambda_{Q,3}|}\nabla\lambda_{Q,3}$ depending on the rules stated above. Then there holds

$$\begin{aligned} a_h^\infty(\lambda_A, \lambda_S) &= \varepsilon\Big(L_2 + (1+\nu_K)K_2 + (2\mathrm{Pe}_L + \mu_L + \nu_L)(L_1 + L_2) + \\ &\qquad + (2\mathrm{Pe}_K + \mu_K)(K_2 + K_3)\Big) = \qquad (3.201) \\ &= \tfrac{\varepsilon}{2}\Big\{(2\mu_K + \nu_K + 4\mathrm{Pe}_K)(K_2 + K_3) + (2 + \nu_K)K_2\Big\} - \tfrac{\varepsilon}{2}\nu_K K_3 + \\ &\quad + \tfrac{\varepsilon}{2}\Big\{(2\mu_L + \nu_L + 4\mathrm{Pe}_L)(L_1 + L_2) + (2 + \nu_L)L_2\Big\} + \tfrac{\varepsilon}{2}\nu_L L_1, \\ a_h^\infty(\lambda_B, \lambda_S) &= \varepsilon\Big(L_1 + N_3 + \nu_N N_3\Big), \qquad (3.202) \end{aligned}$$

$$
\begin{aligned}
a_h^\infty(\lambda_C,\lambda_S) &= \varepsilon\Big(N_1+V_3-\nu_N N_3\Big), && (3.203)\\
a_h^\infty(\lambda_D,\lambda_S) &= \varepsilon\Big(T_2+(1+\nu_V)V_2+(\mu_T+\nu_T)(T_2+T_3)+\mu_V(V_1+V_2)\Big) = && (3.204)\\
&= \tfrac{\varepsilon}{2}\Big\{(2\mu_V+\nu_V)(V_1+V_2)+(2+\nu_V)V_2\Big\}-\tfrac{\varepsilon}{2}\nu_V V_1+\\
&\quad+\tfrac{\varepsilon}{2}\Big\{(2\mu_T+\nu_T)(T_2+T_3)+(2+\nu_T)T_2\Big\}+\tfrac{\varepsilon}{2}\nu_T T_3,\\
a_h^\infty(\lambda_E,\lambda_S) &= \varepsilon\Big(M_1+T_3+\nu_M M_1\Big), && (3.205)\\
a_h^\infty(\lambda_F,\lambda_S) &= \varepsilon\Big(K_1+M_3-\nu_M M_1\Big). && (3.206)
\end{aligned}
$$

From the definition of ν_N and ν_M it follows that $\nu_N \geq -1-\frac{L_1}{N_3}$ and $\nu_M \geq -1-\frac{T_3}{M_1}$. Consequently, there holds

$$
\begin{aligned}
L_1+N_3+\nu_N N_3 &\leq L_1+N_3+(-N_3-L_1) = 0, && (3.207)\\
M_1+T_3+\nu_M M_1 &\leq M_1+T_3+(-M_1-T_3) = 0. && (3.208)
\end{aligned}
$$

Further, if we take into account the inequalities $N_1, V_3 \leq \mathcal{R}_N N_3$ and $K_1, M_3 \leq \mathcal{R}_M M_1$ we can prove

$$
\begin{aligned}
N_1+V_3-\nu_N N_3 &\leq \mathcal{R}_N N_3+\mathcal{R}_N N_3-2\mathcal{R}_N(\mathcal{R}_N+2)N_3 =\\
&= -2\mathcal{R}_N(\mathcal{R}_N+1)N_3 \leq 0, && (3.209)\\
K_1+M_3-\nu_M M_1 &\leq \mathcal{R}_M M_1+\mathcal{R}_M M_1-2\mathcal{R}_M(\mathcal{R}_M+2)M_1 =\\
&= -2\mathcal{R}_M(\mathcal{R}_M+1)M_1 \leq 0. && (3.210)
\end{aligned}
$$

Since there holds $\nu_V V_1 = \nu_T T_3$, the inequality $a_h^\infty(\lambda_D,\lambda_S)\leq 0$ follows immediately from the estimate (3.171) and the inequalities $V_2 \leq \mathcal{R}_V V_1$ and $T_2 \leq \mathcal{R}_T T_3$.

Similarly, the equality $\nu_K K_3 = \nu_L L_1$ together with the inequalities $L_2 \leq \mathcal{R}_L L_1$ and $K_2 \leq \mathcal{R}_K K_3$ implies $a_h^\infty(\lambda_A,\lambda_S)\leq 0$. In fact, in this case the inequality $a_h^\infty(\lambda_A,\lambda_S)\leq 0$ holds mainly due to the presence of the terms Pe_K and Pe_L.

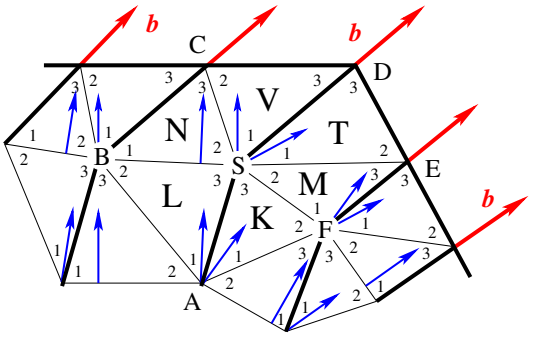

Figure 3.11: A structure of the domain Ω_S in the vicinity of a corner of Ω.

Let us also briefly mention how we choose the coefficients μ and ν in the vicinity of a domain corner. Since we consider only convex domains, the boundary cluster laying in the corner of the domain Ω is formed by two elements V and T with $\boldsymbol{n}_V = \frac{-1}{|\nabla\lambda_{V,1}|}\nabla\lambda_{V,1}$ and $\boldsymbol{n}_T = \frac{-1}{|\nabla\lambda_{T,1}|}\nabla\lambda_{T,1}$. Hence, for each element Q from any cluster laying on the same discrete streamline there holds $\boldsymbol{n}_Q = \frac{-1}{|\nabla\lambda_{Q,1}|}\nabla\lambda_{Q,1}$

as well (see Figure 3.11). This changes the stencil of the method from the previous paragraph, it now takes the form

$$\begin{aligned}
a_h^\infty(\lambda_A,\lambda_S) &= \varepsilon\Big((1+\nu_L)L_2+(1+\nu_K)K_2+(2\mathrm{Pe}_L+\mu_L)(L_2+L_3)+ \\
&\qquad +(2\mathrm{Pe}_K+\mu_K)(K_2+K_3)\Big), && (3.211)\\
a_h^\infty(\lambda_B,\lambda_S) &= \varepsilon\Big(L_1+N_3-\nu_N N_1\Big), && (3.212)\\
a_h^\infty(\lambda_C,\lambda_S) &= \varepsilon\Big(N_1+V_3+\nu_N N_1\Big), && (3.213)\\
a_h^\infty(\lambda_D,\lambda_S) &= \varepsilon\Big(T_2+V_2+(\mu_T+\nu_T)(T_2+T_3)+(\mu_V+\nu_V)(V_2+V_3)\Big), && (3.214)\\
a_h^\infty(\lambda_E,\lambda_S) &= \varepsilon\Big(M_1+T_3+\nu_M M_1\Big), && (3.215)\\
a_h^\infty(\lambda_F,\lambda_S) &= \varepsilon\Big(K_1+M_3-\nu_M M_1\Big). && (3.216)
\end{aligned}$$

Since the boundary layer function near the domain corner is different from the exponential boundary layer function, we no longer require $\mu+\nu=\mathcal{R}_\Gamma$ (moreover, we cannot determine which part of Γ_+ we should use). Hence, for any triangle V laying on the corner discrete streamline we set $\mathcal{R}_V=\mathrm{Pe}_V(\cot\mathrm{Pe}_V-1)-1$ and define

$$\nu_V \;=\; \max\left\{-1-\frac{N_1}{V_3},2\mathcal{R}_V(\mathcal{R}_V+2)\right\}, \tag{3.217}$$

where N is an element laying on the same side from the corner discrete streamline and having different unit vector, i.e. $\boldsymbol{n}_N\neq\boldsymbol{n}_V$. Finally, we define the coefficient μ_V as

$$\mu_V \;=\; \frac{-\mathcal{R}_V}{1-\rho}-\nu_V, \tag{3.218}$$

where $\rho\in(0,1)$ is such that $V_2\le\rho\mathcal{R}_V V_3<\mathcal{R}_V V_3$. This means that using excessively obtuse angles $\alpha_{V,2}$ with $V_2\to\mathcal{R}_V V_3$ results in $\mu_V\to+\infty$, which is not allowed due to the condition (3.186). Thus, the restriction on the inner angles is stronger in this case, however, obtuse angles are still admissible. In the most important downwind node there holds

$$\begin{aligned}
a_h^\infty(\lambda_D,\lambda_S) &= \varepsilon\left(T_2+V_2-\frac{\mathcal{R}_T}{1-\rho}(T_2+T_3)-\frac{\mathcal{R}_V}{1-\rho}(V_2+V_3)\right)\;\le\\
&\le\;\varepsilon\left(\left(\rho\mathcal{R}_T-\frac{\mathcal{R}_T}{1-\rho}(\rho\mathcal{R}_T+1)\right)T_3+\left(\rho\mathcal{R}_V-\frac{\mathcal{R}_V}{1-\rho}(\rho\mathcal{R}_V+1)\right)V_3\right)\;=\\
&=\;\varepsilon\left(\left(\rho-\frac{1-|\rho\mathcal{R}_T|}{1-\rho}\right)\mathcal{R}_T T_3+\left(\rho-\frac{1-|\rho\mathcal{R}_V|}{1-\rho}\right)\mathcal{R}_V V_3\right)\;\le\;0,
\end{aligned}$$

since $\mathcal{R}_T T_3,\mathcal{R}_V V_3>0$ and $\rho<1<\frac{1-|\rho\mathcal{R}_T|}{1-\rho}$, $\rho<1<\frac{1-|\rho\mathcal{R}_V|}{1-\rho}$.

3.7 Higher order finite elements

Although the presented method seems to be designed purely for the piecewise linear finite elements, it is possible to extend it for higher order finite elements in any dimension. We describe the main ideas of such an extension on piecewise quadratic finite elements in 2D (we use an upper index (2) to emphasize the usage

of piecewise quadratic functions). We skip the precise analysis in these cases and just test it on several examples.

The finite element space is defined as $X_h^{(2)} = \{v_h \in \mathcal{C}(\Omega), v_h|_K \in P_2(K),\ \forall K \in \mathcal{T}_h\}$ and we are looking for $u_h \in V_h^{(2)} = X_h^{(2)} \cap H_0^1(\Omega)$ such that

$$a_h^{(2)}(u_h, v_h) \;=\; F_h^{(2)}(v_h) \qquad \text{for all } v_h \in V_h^{(2)}, \tag{3.219}$$

where the bilinear form $a_h^{(2)}$ and the functional $F_h^{(2)}$ are defined as

$$\begin{aligned} a_h^{(2)}(u,v) \;&=\; \varepsilon(\nabla u, \nabla v)_\Omega + \sum_{K\in\mathcal{T}_h} \left(\boldsymbol{b}_K^{(1)} \cdot \nabla u, v\right)_K + \\ &\quad + \sum_{K\in\mathcal{T}_h} \left(-\varepsilon\Delta u + \boldsymbol{b}_K^{(1)} \cdot \nabla u, R_K^{(2)} v - \Pi_{b,K}^{(2)} v\right)_K \quad \text{and} \qquad (3.220) \\ F_h^{(2)}(v) \;&=\; \sum_{K\in\mathcal{T}_h} \left(f, v + R_K^{(2)} v - \Pi_{b,K}^{(2)} v\right)_K . \qquad (3.221) \end{aligned}$$

Thus, the definition of the method is an analog to the P_1-case: $\boldsymbol{b}_K^{(1)}$ is a piecewise polynomial (linear in each component) vector function parallel with $\boldsymbol{d}_{K,1}$ on each $K \in \mathcal{T}_h$ and $R_K^{(2)}, \Pi_{b,K}^{(2)} : P_2(K) \to P_1(K)$ are linear mappings constructed in such a way that the resulting matrix of the method is a positive-definite monotone matrix.

We also use special numbering of nodes and basis functions of $P_2(K)$. On each $K \in \mathcal{T}_h$ we denote $P_{K,1}^{(2)} = P_{K,1}$, $P_{K,3}^{(2)} = P_{K,2}$ and $P_{K,6}^{(2)} = P_{K,3}$ the corner nodes of K. The midpoints of edges $\overline{P_{K,1}P_{K,2}}$, $\overline{P_{K,2}P_{K,3}}$ and $\overline{P_{K,3}P_{K,1}}$ are denoted $P_{K,2}^{(2)}$, $P_{K,5}^{(2)}$ and $P_{K,4}^{(2)}$, respectively. (cf. Figure 3.12)

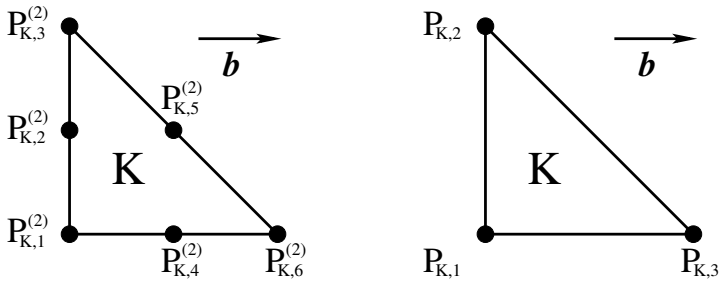

Figure 3.12: Definition of nodes numbering for $P_2(K)$.

The basis functions $\{\varphi_{K,i}^{(2)}\}_{i=1}^6$ of the space $P_2(K)$ are standardly defined by the relations $\varphi_{K,i}^{(2)}\left(P_{K,j}^{(2)}\right) = \delta_{ij}$ for all $i, j \in \{1, 2, \ldots, 6\}$.

3.7.1 Definition and properties of the discretized vector field

Prior to $\boldsymbol{b}_K^{(1)}$, we firstly define on each $K \in \mathcal{T}_h$ an interpolation $\boldsymbol{b}_K^I \in P_1(K)^2$ of the vector $\boldsymbol{b}$. We define $\boldsymbol{b}_K^I$ as an orthogonal L^2-projection of the vector $\boldsymbol{b}$ on the space $P_1(K)^2$. Hence, it satisfies the equality

$$\left(\boldsymbol{b} - \boldsymbol{b}_K^I, \boldsymbol{\varphi}\right)_{L^2(K)^2} = 0, \qquad \text{for all } \boldsymbol{\varphi} \in P_1(K)^2. \tag{3.222}$$

We can construct this orthogonal L^2-projection, for instance, by considering a basis of the space $P_1(K)^2$ in the form $\{\lambda_{K,l}\nabla\lambda_{K,j},\, j=1,2,\, l=1,2,3\}$ or $\{\lambda_{K,l}\boldsymbol{d}_{K,j},\, j=1,2,\, l=1,2,3\}$, where $\boldsymbol{d}_{K,j}=P_{K,3}-P_{K,j}$ for $j=1,2$. Consequently, we can express the vector $\boldsymbol{b}_K^I$ as

$$\boldsymbol{b}_K^I=\sum_{j=1}^{2}\left(\sum_{i=1}^{3}\alpha_{K,i}^{(j)}\lambda_{K,i}\right)\boldsymbol{d}_{K,j}, \tag{3.223}$$

where $\alpha_{K,i}^{(j)}$, $j=1,2$ and $i=1,2,3$, are for each $j=1,2$ solutions of the system of linear equations with the same nonsingular matrix

$$\sum_{i=1}^{3}\alpha_{K,i}^{(j)}\left(\lambda_{K,i},\lambda_{K,l}\right)_K \;=\; -\left(\boldsymbol{b}\cdot\nabla\lambda_{K,j},\lambda_{K,l}\right)_K, \qquad l=1,2,3. \tag{3.224}$$

These systems are obtained by considering the vector $\boldsymbol{b}_K^I$ in the form (3.223) and testing the difference in (3.222) by functions $\boldsymbol{\varphi}$ from the basis $\{\lambda_{K,l}\nabla\lambda_{K,j},\, j=1,2,\, l=1,2,3\}$. For each $j=1,2$ and $i=1,2,3$ the solutions of (3.224) can be expressed in the form

$$\alpha_{K,i}^{(j)} \;=\; \frac{3}{|K|}(-\boldsymbol{b}\cdot\nabla\lambda_{K,j},4\lambda_{K,i}-1)_K. \tag{3.225}$$

Since $\boldsymbol{b}_K^I$ is the orthogonal L^2-projection of $\boldsymbol{b}$, one can prove the following generalization of Lemma 3.5.3 (page 75). We formulate and prove it for general polynomial degree $r\in\mathbb{N}$ and dimension $n\in\mathbb{N}$.

Lemma 3.7.1. *Let $K\subset\mathbb{R}^n$ be a simplex, $n,r\in\mathbb{N}$, $\boldsymbol{b}\in W^{r+1,\infty}(K)^n$ and $\boldsymbol{b}_K^I\in P_r(K)^n$ satisfies for all $\boldsymbol{\varphi}\in P_r(K)^n$*

$$\left(\boldsymbol{b}-\boldsymbol{b}_K^I,\boldsymbol{\varphi}\right)_{L^2(K)^n} \;=\; 0. \tag{3.226}$$

If the shape-regularity assumption (4.23) is fulfilled, then there exists a constant $c>0$ depending only on n, r and σ such that for all $v_h\in P_{r+1}(K)$ holds

$$\left\|\left(\boldsymbol{b}-\boldsymbol{b}_K^I\right)\cdot\nabla v_h\right\|_{0,K} \;\le\; c\,h_K^{r+1}\,|\boldsymbol{b}|_{r+1,\infty,K}\,|v_h|_{1,K}. \tag{3.227}$$

Proof. Using the triangle inequality, the Hölder inequality and the vector version of the approximation property (Corollary 4.2.1, page 128) we obtain

$$\begin{aligned}
&\left\|\left(\boldsymbol{b}-\boldsymbol{b}_K^I\right)\cdot\nabla v_h\right\|_{0,K} \;=\; \left\|\sum_{i=1}^{n}\left[\boldsymbol{b}-\boldsymbol{b}_K^I\right]_i\frac{\partial v_h}{\partial x_i}\right\|_{0,K} \;\le\; \sum_{i=1}^{n}\left\|\left[\boldsymbol{b}-\boldsymbol{b}_K^I\right]_i\frac{\partial v_h}{\partial x_i}\right\|_{0,K} \;\le\\
&\le\sum_{i=1}^{n}\left\|\left[\boldsymbol{b}-\boldsymbol{b}_K^I\right]_i\right\|_{0,\infty,K}\left\|\frac{\partial v_h}{\partial x_i}\right\|_{0,K}\le\left(\sum_{i=1}^{n}\left\|\left[\boldsymbol{b}-\boldsymbol{b}_K^I\right]_i\right\|_{0,\infty,K}^2\right)^{\frac12}\left(\sum_{i=1}^{n}\left\|\frac{\partial v_h}{\partial x_i}\right\|_{0,K}^2\right)^{\frac12}=\\
&=\left|\boldsymbol{b}-\boldsymbol{b}_K^I\right|_{0,\infty,K}|v_h|_{1,K} \;\le\; C_\Pi h_K^{r+1}\,|\boldsymbol{b}|_{r+1,\infty,K}\,|v_h|_{1,K}.
\end{aligned} \tag{3.228}$$

□

Lemma 3.7.2. *Let $n, r \in \mathbb{N}$ and let $\boldsymbol{b}_K^I \in P_r(K)^n$ be defined by*

$$\boldsymbol{b}_K^I \;=\; \sum_{j=1}^{n} \left(\sum_{i=1}^{N_r} \alpha_{K,i}^{(j)} \varphi_{K,i}^{(r)} \right) \boldsymbol{d}_{K,j}, \tag{3.229}$$

where $\alpha_{K,i}^{(j)} \in \mathbb{R}$ are suitable coefficients. If we denote $\boldsymbol{b}_K^{(r)} = \left(\sum_{i=1}^{N_r} \alpha_{K,i}^{(1)} \varphi_{K,i}^{(r)} \right) \boldsymbol{d}_{K,1}$, then

$$\left\| \left(\boldsymbol{b}_K^I - \boldsymbol{b}_K^{(r)} \right) \cdot \nabla v_h \right\|_{0,K} \;\leq\; \left(\sum_{j=2}^{n} \left\| \sum_{i=1}^{N_r} \alpha_{K,i}^{(j)} \varphi_{K,i}^{(r)} \right\|_{\infty,K} \right) h_K |v_h|_{1,K}. \tag{3.230}$$

Proof. It results immediately from the fact that $|\boldsymbol{d}_{K,j}| \leq h_K$ for $j = 1, 2, \ldots, n$. □

Remark 3.7.1. In order to fulfil the equality (3.226) the coefficients $\alpha_{K,i}^{(j)}$ have to be the solutions of n systems of linear equations (i.e. $j = 1, 2, \ldots, n$)

$$\sum_{i=1}^{N_r} \alpha_{K,i}^{(j)} \left(\varphi_{K,i}^{(r)}, \varphi_{K,m}^{(r)} \right)_K \;=\; \left(-\boldsymbol{b} \cdot \nabla \lambda_{K,j}, \varphi_{K,m}^{(r)} \right)_K, \qquad m = 1, 2, \ldots, N_r. \tag{3.231}$$

Let us turn back to the case of piecewise quadratic functions in 2D. As in Lemma 3.7.2 we denote by $\boldsymbol{b}_K^{(1)} = \left(\sum_{i=1}^{3} \alpha_{K,i}^{(1)} \lambda_{K,i} \right) \boldsymbol{d}_{K,1}$ the first term of the sum (3.223), i.e., the first term of $\boldsymbol{b}_K^I$. Since the bilinear form $a_h^{(2)}$ (cf. (3.220)) contains terms $\left(\boldsymbol{b}_K^{(1)} \cdot \nabla u_h, v_h \right)_K$, with $v_h \in P_1(K)$ and $u_h \in P_2(K)$, we compute the values $\left(\boldsymbol{b}_K^{(1)} \cdot \nabla u_h, \lambda_{K,j} \right)_K$ for any $j = 1, 2, 3$

$$\begin{aligned}
\left(\boldsymbol{b}_K^{(1)} \cdot \nabla u_h, \lambda_{K,j} \right)_K \;&=\; \sum_{i=1}^{6} u_{K,i} \left(\boldsymbol{b}_K^{(1)} \cdot \nabla \varphi_{K,i}^{(2)}, \lambda_{K,j} \right)_K \;= \\
&= \quad (u_{K,4} - u_{K,1}) \Big(-\boldsymbol{b}_K^I \cdot \nabla \lambda_{K,1}, (4\lambda_{K,1} - 1) \lambda_{K,j} \Big)_K + && (3.232) \\
&\quad + (u_{K,5} - u_{K,2}) \Big(-\boldsymbol{b}_K^I \cdot \nabla \lambda_{K,1}, 4\lambda_{K,2} \lambda_{K,j} \Big)_K + && (3.233) \\
&\quad + (u_{K,6} - u_{K,4}) \Big(-\boldsymbol{b}_K^I \cdot \nabla \lambda_{K,1}, (4\lambda_{K,3} - 1) \lambda_{K,j} \Big)_K \;= && (3.234) \\
&= \; (u_{K,4} - u_{K,1}) s_{1j}^K \;+\; (u_{K,5} - u_{K,2}) s_{2j}^K \;+\; (u_{K,6} - u_{K,4}) s_{3j}^K, && (3.235)
\end{aligned}$$

where we denoted by s_{1j}^K, s_{2j}^K and s_{3j}^K the integrals in (3.232), (3.233) and (3.234), respectively. We also used the notation $u_{K,i} = u_h(P_{K,i}^{(2)})$ for all $i \in \{1, 2, \ldots, 6\}$. When constructing the integrals in (3.232)–(3.234) we employed the following technique: for instance, when $i = 6$ we express $\nabla \varphi_{K,6}^{(2)}$ in the form

$$\nabla \varphi_{K,6}^{(2)} = \nabla \Big(\lambda_{K,3} (2\lambda_{K,3} - 1) \Big) = \nabla \lambda_{K,3} (4\lambda_{K,3} - 1) = -\Big(\nabla \lambda_{K,1} + \nabla \lambda_{K,2} \Big) (4\lambda_{K,3} - 1), \tag{3.236}$$

which results in

$$\begin{aligned}
\left(\boldsymbol{b}_K^{(1)} \cdot \nabla \varphi_{K,6}^{(2)}, \lambda_{K,j} \right)_K \;&=\; \Big(-\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1}, (4\lambda_{K,3} - 1) \lambda_{K,j} \Big)_K = \\
&=\; \Big(-\boldsymbol{b}_K^I \cdot \nabla \lambda_{K,1}, (4\lambda_{K,3} - 1) \lambda_{K,j} \Big)_K, && (3.237)
\end{aligned}$$

where we applied the equality $\boldsymbol{d}_{K,i} \cdot \nabla \lambda_{K,j} = -\delta_{ij}$ for $i, j \in \{1, 2\}$.

3.7.2 Definition and properties of the mapping $\Pi_{b,K}^{(2)}$

Let us now proceed to the definition of the linear mapping $\Pi_{b,K}^{(2)} : P_2(K) \to P_1(K)$. We would like to define it in such a way that

$$\left(\boldsymbol{b}_K^{(1)} \cdot \nabla \varphi_{K,i}^{(2)}, \Pi_{b,K}^{(2)}\left(\varphi_{K,j}^{(2)}\right) - \varphi_{K,j}^{(2)}\right)_K \;=\; 0 \quad \text{for all } i,j = 1,2,\ldots,6. \tag{3.238}$$

For each $j \in \{1,2,\ldots,6\}$ we need to solve a system of six ($\dim P_2(K) = 6$) equations for three ($\dim P_1(K) = 3$) unknowns. However, due to the equality $\boldsymbol{b}_K^{(1)} \cdot \nabla \varphi = 0$ for $\varphi = 1, \lambda_{K,2}, \lambda_{K,2}^2$, only three of these six equations are linearly independent. Thus, for each $\varphi_{K,j}^{(2)} \in P_2(K)$ we define $\Pi_{b,K}^{(2)}\left(\varphi_{K,j}^{(2)}\right) = \sum_{m=1}^3 \mu_{K,m}^{(j)} \lambda_{K,m}$, where $\mu_{K,m}^{(j)}$ are solutions of the systems of linear equations

$$\sum_{m=1}^{3} \mu_{K,m}^{(j)} \left(-\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1},\, \lambda_{K,l}\lambda_{K,m}\right)_K \;=\; \left(-\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1},\, \lambda_{K,l}\varphi_{K,j}^{(2)}\right)_K,\; l = 1,2,3. \tag{3.239}$$

If $-\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1} > 0$ almost everywhere in K, then we obtain for arbitrary $v_1, v_2, v_3 \in \mathbb{R}$ the inequality

$$\sum_{l,m=1}^{3} v_l v_m \left(-\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1},\, \lambda_{K,l}, \lambda_{K,m}\right)_K = \left(-\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1}, \left(\sum_{l=1}^{3} v_l \lambda_{K,l}\right)^2\right)_K > 0, \tag{3.240}$$

whenever $v_1^2+v_2^2+v_3^2$ is nonzero. It means that the matrix of all 6 systems (3.239) is symmetric, positive definite and therefore nonsingular. Consequently, the values $\mu_{K,m}^{(j)}$ are uniquely defined. (If $\boldsymbol{b}_K^{(1)}$ is constant vector then $-\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1} = \frac{|\boldsymbol{b}_K^{(1)}|}{|\boldsymbol{d}_{K,1}|} > 0$ and $\Pi_{b,K}^{(2)}$ is the orthogonal L^2-projection.)

In addition, using the equalities $\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,2} = 0$ and $\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,3} = -\boldsymbol{b}_K^{(1)} \cdot \nabla(\lambda_{K,1} + \lambda_{K,2}) = -\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1}$ we deduce that $\boldsymbol{b}_K^{(1)} \cdot \nabla \varphi_{K,i}^{(2)} = \boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1}(\nu_1 \lambda_{K,1} + \nu_2 \lambda_{K,2} + \nu_3 \lambda_{K,3})$ for suitable real values ν_1, ν_2, ν_3 depending on i. This and the definition of the values $\mu_{K,m}^{(j)}$ yields the equality

$$\left(-\boldsymbol{b}_K^{(1)} \cdot \nabla \varphi_{K,i}^{(2)},\, \sum_{m=1}^{3} \mu_{K,m}^{(j)} \lambda_{K,m} - \varphi_{K,j}^{(2)}\right)_K \;=\; 0. \qquad \text{for all } i = 1,2,\ldots,6, \tag{3.241}$$

which is the equality (3.238).

Further, since the mapping $\Pi_{b,K}^{(2)}$ is linear, we can construct a matrix of this mapping with respect to the standard FEM basis. Then for each function $v_h \in P_2(K)$ there holds

$$\left[\Pi_{b,K}^{(2)}\left(v_h\right)\right]_{M_K^{(1)}} \;=\; \begin{pmatrix} \mu_{K,1}^{(1)}, & \mu_{K,1}^{(2)}, & \cdots & \mu_{K,1}^{(6)} \\ \mu_{K,2}^{(1)}, & \mu_{K,2}^{(2)}, & \cdots & \mu_{K,2}^{(6)} \\ \mu_{K,3}^{(1)}, & \mu_{K,3}^{(2)}, & \cdots & \mu_{K,3}^{(6)} \end{pmatrix} \left[v_h\right]_{M_K^{(2)}}, \tag{3.242}$$

where $\left[\Pi_{b,K}^{(2)}\left(v_h\right)\right]_{M_K^{(1)}}$ and $\left[v_h\right]_{M_K^{(2)}}$ are coordinates of the functions $\Pi_{b,K}^{(2)}\left(v_h\right)$ and v_h with respect to the bases $M_K^{(1)} = \left\{\lambda_{K,j}\right\}_{j=1}^3$ and $M_K^{(2)} = \left\{\varphi_{K,j}^{(2)}\right\}_{j=1}^6$ of the spaces $P_1(K)$ and $P_2(K)$, respectively.

Finally, if we sum all the equations (3.239) for $j = 1, 2, \ldots, 6$ and use the expressions $\sum_{j=1}^{6} \varphi_{K,j}^{(2)} = 1$ and $\sum_{m=1}^{3} \lambda_{K,m} = 1$ we obtain for all $l = 1, 2, 3$ the equality

$$\sum_{m=1}^{3} \sum_{j=1}^{6} \mu_{K,m}^{(j)} \left(-\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1},\, \lambda_{K,l} \lambda_{K,m} \right)_K \;=\; \left(-\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1},\, \lambda_{K,l} \sum_{m=1}^{3} \lambda_{K,m} \right)_K , \tag{3.243}$$

which can be rewritten in the form

$$\sum_{m=1}^{3} \left(1 - \sum_{j=1}^{6} \mu_{K,m}^{(j)} \right) \left(-\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1}, \lambda_{K,l} \lambda_{K,m} \right)_K \;=\; 0, \qquad \forall\, l = 1, 2, 3. \tag{3.244}$$

It means that for $-\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1} > 0$ almost everywhere in K the vector $\left(1 - \sum_{j=1}^{6} \mu_{K,1}^{(j)}, 1 - \sum_{j=1}^{6} \mu_{K,2}^{(j)}, 1 - \sum_{j=1}^{6} \mu_{K,3}^{(j)}\right)$ is a solution of the system of linear equations with nonsingular matrix and zero right-hand side. Therefore, there holds $\sum_{j=1}^{6} \mu_{K,m}^{(j)} = 1$ for all $m = 1, 2, 3$ (the row sums of the matrix in (3.242) are equal to 1), and thus the mapping $\Pi_{b,K}^{(2)}$ preserves polynomials of degree 0, i.e. constants. When the sign of $-\boldsymbol{b}_K^{(1)} \cdot \nabla \lambda_{K,1}$ changes in K, then the system of equations (3.244) can have more then one solution (it has at least one solution – zero solution – due to the zero right-hand side). If it happens we choose

$$\begin{pmatrix} \mu_{K,1}^{(1)}, \mu_{K,1}^{(2)}, \cdots \mu_{K,1}^{(6)} \\ \mu_{K,2}^{(1)}, \mu_{K,2}^{(2)}, \cdots \mu_{K,2}^{(6)} \\ \mu_{K,3}^{(1)}, \mu_{K,3}^{(2)}, \cdots \mu_{K,3}^{(6)} \end{pmatrix} \;=\; \begin{pmatrix} \frac{2}{5}, & \frac{3}{5}, -\frac{1}{5}, & \frac{3}{5}, -\frac{1}{5}, -\frac{1}{5} \\ -\frac{1}{5}, & \frac{3}{5}, & \frac{2}{5}, -\frac{1}{5}, & \frac{3}{5}, -\frac{1}{5} \\ -\frac{1}{5}, -\frac{1}{5}, -\frac{1}{5}, & \frac{3}{5}, & \frac{3}{5}, & \frac{2}{5} \end{pmatrix}, \tag{3.245}$$

which is a matrix of the orthogonal L^2-projection of $P_2(K)$ onto $P_1(K)$.

3.7.3 Construction of the mapping $R_K^{(2)}$

In order to obtain an upwind scheme we construct the mapping $R_K^{(2)}$ in such a way that the matrix corresponding to the discretization of the convective term is (for a suitable node labeling) triangular. This labeling is carried out successively streamline by streamline. Then the value of u_h in a certain node depends only on the values in the nodes laying on the same discrete streamline in the upwind direction.

The triangular matrix can be achieved by setting $R_K^{(2)}(\varphi_{K,1}^{(2)}) = R_K^{(2)}(\varphi_{K,2}^{(2)}) = R_K^{(2)}(\varphi_{K,3}^{(2)}) = 0$ for each $K \in \mathcal{T}_h$. This configuration results in the equality $\left(\boldsymbol{b}_K^{(1)} \cdot \nabla u_h, R_K^{(2)}\left(\varphi_{K,m}^{(2)}\right)\right)_K = 0$, for each $K \in \mathcal{T}_h$ and $m = 1, 2, 3$. Hence, for a fixed element K none of the values $u_h(P_{K,1})$, $u_h(P_{K,2})$ and $u_h(P_{K,3})$ depends on other $u_h(P_{K,i})$ (some of these $P_{K,i}$ lay in downwind direction, the other on different discrete streamline). The matrix of the mapping $R_K^{(2)}$ with respect to the standard FEM bases then satisfies

$$\left[R_K^{(2)}\left(v_h\right)\right]_{M_K^{(1)}} \;=\; \begin{pmatrix} 0, & 0, & 0, & r_{14}^K, & r_{15}^K, & r_{16}^K \\ 0, & 0, & 0, & r_{24}^K, & r_{25}^K, & r_{26}^K \\ 0, & 0, & 0, & r_{34}^K, & r_{35}^K, & r_{36}^K \end{pmatrix} \left[v_h\right]_{M_K^{(2)}}. \tag{3.246}$$

Using the expression (3.235) we can further write

$$\left(\boldsymbol{b}_K^{(1)} \cdot \nabla u_h, R_K^{(2)}\left(\varphi_{K,m}^{(2)}\right)\right)_K = \sum_{j=1}^{3} r_{jm}^K \left(\boldsymbol{b}_K^{(1)} \cdot \nabla u_h, \lambda_{K,j}\right)_K =$$

$$= (u_{K,4} - u_{K,1})\sum_{j=1}^{3} s_{1j}^K r_{jm}^K + (u_{K,5} - u_{K,2})\sum_{j=1}^{3} s_{2j}^K r_{jm}^K + (u_{K,6} - u_{K,4})\sum_{j=1}^{3} s_{3j}^K r_{jm}^K.$$

Since we would like to obtain an upwind scheme, the first of these three sums has to vanish for $m = 5$ (it already vanishes for $m = 1, 2$ and 3 and since it is multiplied by $(u_{K,4} - u_{K,1})$ it has to vanish for $m = 5$ because the node $P_{K,5}$ does not lay on the same discrete streamline as $P_{K,1}$ and $P_{K,4}$), the second sum has to vanish for $m \in \{4, 6\}$ and the third for $m \in \{4, 5\}$. These requirements can be rewritten in the matrix form

$$\begin{pmatrix} s_{11}^K, & s_{12}^K, & s_{13}^K \\ s_{21}^K, & s_{22}^K, & s_{23}^K \\ s_{31}^K, & s_{32}^K, & s_{33}^K \end{pmatrix} \begin{pmatrix} r_{14}^K, & r_{15}^K, & r_{16}^K \\ r_{24}^K, & r_{25}^K, & r_{26}^K \\ r_{34}^K, & r_{35}^K, & r_{36}^K \end{pmatrix} = \begin{pmatrix} \kappa_{14}^K, & 0, & \kappa_{16}^K \\ 0, & \kappa_{25}^K, & 0 \\ 0, & 0, & \kappa_{36}^K \end{pmatrix}, \tag{3.247}$$

where $\kappa_{14}^K, \kappa_{16}^K, \kappa_{25}^K$ and κ_{36}^K are generally nonzero values.

Recalling the linear finite elements, the derivative of any $v_h \in P_1(K)$ in the direction of the stabilization vector was estimated $\|(P_{K,n+1} - C_K) \cdot \nabla v_h\|_{0,K} \leq h_K |v_h|_{1,K}$ for each $v_h \in P_1(K)$. We would like to extend this property to the quadratic finite elements. This can be achieved by requiring $(\Pi_{b,K}^{(2)} - R_K^{(2)})v_h = 0$, for all $v_h \in P_0(K)$. Since we already know that $\Pi_{b,K}^{(2)} v_h = v_h$ for all $v_h \in P_0(K)$, it suffices to require $R_K^{(2)} v_h = v_h$ for all $v_h \in P_0(K)$, i.e. the row sums of the mapping matrix corresponding to the mapping $R_K^{(2)}$ have to be equal to 1. Using this property and multiplying both sides of the equality (3.247) by the vector $(1, 1, 1)^T$ then gives

$$\kappa_{14}^K + \kappa_{16}^K = \sum_{j=1}^{3} s_{1j}^K = \left(- \boldsymbol{b} \cdot \nabla \lambda_{K,1}, 4\lambda_{K,1} - 1\right)_K, \tag{3.248}$$

$$\kappa_{25}^K = \sum_{j=1}^{3} s_{2j}^K = \left(- \boldsymbol{b} \cdot \nabla \lambda_{K,1}, 4\lambda_{K,2}\right)_K, \tag{3.249}$$

$$\kappa_{36}^K = \sum_{j=1}^{3} s_{3j}^K = \left(- \boldsymbol{b} \cdot \nabla \lambda_{K,1}, 4\lambda_{K,3} - 1\right)_K. \tag{3.250}$$

Thus, it remains to determine either κ_{14}^K or κ_{16}^K, or to give some restriction on them. This will probably follow from the assumptions on the coercivity of the bilinear form $a_h^{(2)}$. Since we omit proof of the coercivity in this case, we use in numerical tests $\kappa_{16}^K = 0$.

If the matrix $(s_{ij}^K)_{i,j=1}^3$ is nonsingular, we can then compute the exact form of the mapping $R_K^{(2)}$. For instance, if $\boldsymbol{b} \cdot \nabla \lambda_{K,1}$ is constant and nonzero on K, then the matrix $(r_{ij}^K)_{i,j=1}^3$ is a solution of the matrix equation

$$\begin{pmatrix} 1/3, & 0, & 0 \\ 1/3, & 2/3, & 1/3 \\ 0, & 0, & 1/3 \end{pmatrix} \begin{pmatrix} r_{14}^K, & r_{15}^K, & r_{16}^K \\ r_{24}^K, & r_{25}^K, & r_{26}^K \\ r_{34}^K, & r_{35}^K, & r_{36}^K \end{pmatrix} = \begin{pmatrix} (1 - \gamma_K)/3, & 0, & \gamma_K/3 \\ 0, & 4/3, & 0 \\ 0, & 0, & 1/3 \end{pmatrix}, \tag{3.251}$$

where $\gamma_K = -\frac{3}{|K|}\frac{\kappa_{14}^K}{\boldsymbol{b}\cdot\nabla\lambda_{K,1}}$. Solution of this equation has the form

$$\begin{pmatrix} r_{14}^K, & r_{15}^K, & r_{16}^K \\ r_{24}^K, & r_{25}^K, & r_{26}^K \\ r_{34}^K, & r_{35}^K, & r_{36}^K \end{pmatrix} = \begin{pmatrix} 1-\gamma_K, & 0, & \gamma_K \\ \frac{-1+\gamma_K}{2}, & 2, & \frac{-1-\gamma_K}{2} \\ 0, & 0, & 1 \end{pmatrix} \tag{3.252}$$

and we use it also when the matrix $(s_{ij}^K)_{i,j=1}^3$ is singular or ill-conditioned.

When the method is coercive with respect to a suitable energy norm, one can use the inequality $\|(\Pi_{b,K}^{(2)} - R_K^{(2)})v_h\|_{0,K} \le C(\boldsymbol{b})h_K|v_h|_{1,K}$, for all $v_h \in P_2(K)$, together with Lemma 3.7.1 (page 99) and derive error estimates analogously to the case of piecewise linear finite elements. This will be the subject of author's future work.

In the final paragraph we show how the construction of the mapping $R_K^{(2)}$ affects the stability of the method.

3.7.4 Stability of the metod

From the construction of the mapping $R_K^{(2)}$ it follows that for each $v_h \in V_h$ holds

$$\begin{aligned} \left(\boldsymbol{b}_K^{(1)}\cdot\nabla v_h, R_K^{(2)}\left(v_h\right)\right)_K &= \sum_{m=4}^{6} v_{K,m}\left(\boldsymbol{b}_K^{(1)}\cdot\nabla v_h, R_K^{(2)}\left(\varphi_{K,m}^{(2)}\right)\right)_K = \\ &= (v_{K,4}-v_{K,1})\left(\kappa_{14}^K v_{K,4} + \kappa_{16}^K v_{K,6}\right) + (v_{K,5}-v_{K,2})\kappa_{25}^K v_{K,5} + (v_{K,6}-v_{K,4})\kappa_{36}^K v_{K,6}. \end{aligned}$$

The matrix generated by this scheme is (due to the structure of the mesh) reducible. Each discrete streamline forms its own submatrix that does not depend on the nodes from the other streamlines (submatrices). Moreover, the nodes numbered $P_{K,2}^{(2)}$ and $P_{K,5}^{(2)}$ (laying between discrete streamlines) also form a chain independent from the other nodes and we can apply the theory from the section devoted to the linear finite elements. Thus, the stability of the method is affected by the remaining nodes (i.e. by the nodes laying on the discrete streamlines).

The matrix corresponding to the discretization of the convective term has for a single discrete streamline (number s) form

$$\mathbb{B}_s = \begin{pmatrix} \kappa_{14}^1, & 0, & 0, & 0, & 0, & \cdots & 0, & 0 \\ \kappa_{16}^1 - \kappa_{36}^1, & \kappa_{36}^1, & 0, & 0, & 0, & \cdots & 0, & 0 \\ 0, & -\kappa_{14}^2, & \kappa_{14}^2, & 0, & 0, & \cdots & 0, & 0 \\ 0, & -\kappa_{16}^2, & \kappa_{16}^2 - \kappa_{36}^2, & \kappa_{36}^2, & 0, & \cdots & 0, & 0 \\ 0, & 0, & 0, & -\kappa_{14}^3 & \kappa_{14}^3, & \cdots & 0, & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0, & 0, & 0, & 0, & \cdots & \kappa_{14}^{N-1}, & 0, & 0 \\ 0, & 0, & 0, & 0, & \cdots & \kappa_{16}^{N-1} - \kappa_{36}^{N-1}, & \kappa_{36}^{N-1}, & 0 \\ 0, & 0, & 0, & 0, & \cdots & 0, & -\kappa_{14}^N, & \kappa_{14}^N \end{pmatrix}, \tag{3.253}$$

where $\kappa_{pq}^j = \sum_{K\subset\mathcal{C}_j^s}\kappa_{pq}^K$ for all suitable indices p, q, j. The inverse of this matrix exists if and only if the diagonal values are nonzero. In this case, we can easily

compute it. It has the following form

$$\mathbb{B}_s^{-1} = \begin{pmatrix} 1/\kappa_{14}^1, & 0, & 0, & 0, & 0, & \cdots & 0 \\ \frac{\kappa_{36}^1-\kappa_{16}^1}{\kappa_{14}^1\kappa_{36}^1}, & 1/\kappa_{36}^1, & 0, & 0, & 0, & \cdots & 0 \\ \frac{\kappa_{36}^1-\kappa_{16}^1}{\kappa_{14}^1\kappa_{36}^1}, & 1/\kappa_{36}^1, & 1/\kappa_{14}^2, & 0, & 0, & \cdots & 0 \\ \frac{\kappa_{36}^1-\kappa_{16}^1}{\kappa_{14}^1\kappa_{36}^1}, & 1/\kappa_{36}^1, & \frac{\kappa_{36}^2-\kappa_{16}^2}{\kappa_{14}^2\kappa_{36}^2}, & 1/\kappa_{36}^2, & 0, & \cdots & 0 \\ \frac{\kappa_{36}^1-\kappa_{16}^1}{\kappa_{14}^1\kappa_{36}^1}, & 1/\kappa_{36}^1, & \frac{\kappa_{36}^2-\kappa_{16}^2}{\kappa_{14}^2\kappa_{36}^2}, & 1/\kappa_{36}^2, & 1/\kappa_{14}^3, & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \frac{\kappa_{36}^1-\kappa_{16}^1}{\kappa_{14}^1\kappa_{36}^1}, & 1/\kappa_{36}^1, & \frac{\kappa_{36}^2-\kappa_{16}^2}{\kappa_{14}^2\kappa_{36}^2}, & 1/\kappa_{36}^2, & \cdots & 1/\kappa_{36}^{N-1}, & 0 \\ \frac{\kappa_{36}^1-\kappa_{16}^1}{\kappa_{14}^1\kappa_{36}^1}, & 1/\kappa_{36}^1, & \frac{\kappa_{36}^2-\kappa_{16}^2}{\kappa_{14}^2\kappa_{36}^2}, & 1/\kappa_{36}^2, & \cdots & 1/\kappa_{36}^{N-1}, & 1/\kappa_{14}^N \end{pmatrix}. \tag{3.254}$$

Since we expect ε to be very small positive number, the stability of the method is mostly affected by the matrices $\mathbb{B}_s$. They are inverse-positive (i.e. $\mathbb{B}_s^{-1} \geq 0$) if for all s and j there holds $\kappa_{14}^j > 0$, $\kappa_{25}^j > 0$, $\kappa_{36}^j > 0$ and $\kappa_{16}^j \leq \kappa_{36}^j$. The last assumption is fulfilled if $\kappa_{16}^K \leq \kappa_{36}^K$ for all $K \in \mathcal{T}_h$.

However, since for any element $K \in \mathcal{T}_h$ and indices $i \neq j$ there does not hold $\varepsilon(\nabla\varphi_{K,i}^{(2)}, \nabla\varphi_{K,j}^{(2)})_K \leq 0$, we cannot prove the discrete maximum principle using Theorem 3.2.1 (page 57). Hence, stabilization terms analogous to the terms from Section 3.6 have to be added. We again postpone it to the forthcoming work.

3.8 Numerical experiments

3.8.1 Example 1, negative divergence

Let us consider $\Omega \subset \mathbb{R}^n$ and let $C = [C_1, C_2, \ldots, C_n] \in \mathbb{R}^n$ be any point such that $C \notin \overline{\Omega}$. Further, let us choose any constant $\omega > 0$ and define $\boldsymbol{b}(\boldsymbol{x}) = \frac{\omega}{n}(C - \boldsymbol{x})$, i.e. $b_i(\boldsymbol{x}) = \frac{\omega}{n}(C_i - x_i)$, where $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$. Then $\operatorname{div} \boldsymbol{b} = -\omega$ and the streamlines of $\boldsymbol{b}$ are rays ending at the point C.

For each $K \in \mathcal{T}_h$, let the vector $\boldsymbol{d}_{K,1}$ (i.e. the corresponding edge $P_{K,1}P_{K,n+1}$) lies on some streamline. We are now interested in the evaluating of θ_K, especially $\int_K \boldsymbol{b} \cdot \nabla\lambda_{K,j}\,\mathrm{d}\boldsymbol{x}$, for $j = 2, 3, \ldots, n$. Let us begin with $\int_K \boldsymbol{b} \cdot \nabla\lambda_{K,n}\,\mathrm{d}\boldsymbol{x}$.

Since the vectors $\boldsymbol{d}_{K,j}$, $j = 1, 2, \ldots, n$, are linearly independent, the matrix $\mathbb{D} = [\boldsymbol{d}_{K,1}, \boldsymbol{d}_{K,2}, \ldots, \boldsymbol{d}_{K,n}]$ is invertible and there exists a uniquely defined QR-decomposition $\mathbb{D} = \mathbb{Q}\mathbb{R}$, where $\mathbb{R} = \mathbb{Q}^T\mathbb{D}$ is an upper triangular matrix with positive diagonal entries and $\mathbb{Q}$ is an orthonormal matrix (Golub and Van Loan, 2012, Theorem 5.2.3). Using the matrix $\mathbb{Q}$ one may define the transformation (the translation and the rotation) of the element K (see Figure 3.13)

$$\widehat{\boldsymbol{x}} \;=\; \boldsymbol{0} + \mathbb{Q}^T(\boldsymbol{x} - P_{K,n+1}). \tag{3.255}$$

Then $\mathbb{R} = [\boldsymbol{d}_{\widehat{K},1}, \boldsymbol{d}_{\widehat{K},2}, \ldots, \boldsymbol{d}_{\widehat{K},n}]$ and from the equality $\boldsymbol{d}_{\widehat{K},j} \cdot \nabla\lambda_{\widehat{K},n} = 0$ for $j = 1, 2, \ldots, n-1$ it follows that $\nabla\lambda_{\widehat{K},n} = (0, 0, \ldots, 0, -1/\widehat{h}_n)$, where $\widehat{h}_n$ is the height of the n-simplex $\widehat{K}$ in the $\widehat{x}_n$ direction.

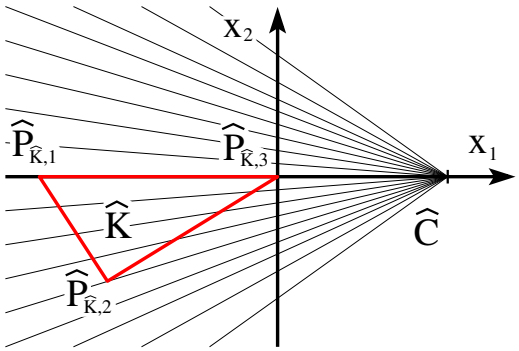

Figure 3.13: Rotated element $\widehat{K}$ in 2D.

Since the vector field $\boldsymbol{b}$ is radial, it is invariant under any rotation around the point C. Hence, it suffices to translate the point C and define the vector field $\widehat{\boldsymbol{b}}(\widehat{\boldsymbol{x}}) = \frac{\omega}{n}(\widehat{C} - \widehat{\boldsymbol{x}}) = \frac{\omega}{n}\Big(|C - P_{K,n+1}| - \widehat{x}_1, -\widehat{x}_2, \ldots, -\widehat{x}_n\Big)$. Consequently

$$\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,n}\,\mathrm{d}\boldsymbol{x} = \int_{\widehat{K}} \widehat{\boldsymbol{b}}\cdot\nabla\lambda_{\widehat{K},n}\,\mathrm{d}\widehat{\boldsymbol{x}} = \int_{\widehat{K}} \frac{\omega}{n}\,\frac{\widehat{x}_n}{\widehat{h}_n}\,\mathrm{d}\widehat{\boldsymbol{x}} = \frac{\omega}{n}\int_{\widehat{K}} -\lambda_{\widehat{K},n}\,\mathrm{d}\widehat{\boldsymbol{x}} = \frac{-\omega|K|}{n(n+1)}. \tag{3.256}$$

Similar approach leads to the same equalities for all $\nabla\lambda_{K,j}$, $j = 2, 3, \ldots, n-1$. Thus, for the mesh parameters θ_K in this case there holds

$$\begin{aligned}\theta_K &= \frac{1}{|K|}\max\left\{\max_{2\le i\le n}\left|\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,i}\,\mathrm{d}\boldsymbol{x}\right|, \left|\sum_{i=2}^{n}\int_K \boldsymbol{b}\cdot\nabla\lambda_{K,i}\,\mathrm{d}\boldsymbol{x}\right|\right\} = \\ &= \frac{1}{|K|}\left|\sum_{i=2}^{n}\frac{-\omega|K|}{n(n+1)}\right| = \frac{n-1}{n}\,\frac{\omega}{n+1} < \frac{\omega}{n+1}.\end{aligned} \tag{3.257}$$

Therefore, the mesh parameters θ_K satisfy in this case the required inequality $\theta_K \le \frac{\omega}{n+1}$ (cf. the inequality (3.105), page 75, or the inequality (3.34), page 61). However, since θ_K are constant for each h (or h_K) one cannot expect its decrement when $h \to 0$.

Let us now be more concrete and specify the data of the example. We consider $n = 2$, $C = [1, 1]$ and $\Omega = (0, 0.9)^2$. We use two types of the vector field $\boldsymbol{b}$. The first type has a form considered above and is defined as

$$\boldsymbol{b}_{1A} = \frac{1}{2}(1 - x, 1 - y)^T, \qquad (\Rightarrow \operatorname{div}\boldsymbol{b}_{1A} = -1). \tag{3.258}$$

The second considered type of the vector field (used for a comparison of the matrices of the mappings $R_K^{(2)}$ and $\Pi_{b,K}^{(2)}$)

$$\boldsymbol{b}_{1B} = \frac{1}{\sqrt{(1-x)^2 + (1-y)^2}}(1 - x, 1 - y)^T \tag{3.259}$$

has the same direction and satisfies $|\boldsymbol{b}_{1B}| = 1$ in Ω. However, $\operatorname{div}\boldsymbol{b}_{1B}$ is no longer constant and the condition $\theta_K \le \frac{\omega}{n+1}$ is unfulfilled (see Figure 3.17).

Further, on $\partial\Omega$ we consider the discontinuous boundary condition u_{b1}

$$u_{b1} = 1 \ \text{ in } \ \{\boldsymbol{x} \in \partial\Omega, |\boldsymbol{x}| \le 0.3\} \qquad \text{and} \qquad u_{b1} = 0 \ \text{ otherwise.} \tag{3.260}$$

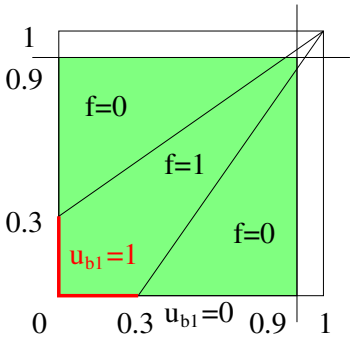

Figure 3.14: Definition of the Example 1 data.

It remains to define the right-hand side $f = f_1$ of the differential equation (3.1). In order to test the behavior of the method in the parabolic layers we define f_1 as a piecewise-constant function satisfying (see Figure 3.14)

$$f_1 = \begin{cases} 1 & \text{for } -\frac{3}{7} + \frac{10}{7}\,x \leq y \leq \frac{3}{10} + \frac{7}{10}\,x \\ 0 & \text{otherwise.} \end{cases} \tag{3.261}$$

For both data combinations $[\boldsymbol{b}_{1A}, u_{b1}, f_1]$ and $[\boldsymbol{b}_{1B}, u_{b1}, f_1]$ we may compute the reduced solutions u_0^{1A} and u_0^{1B} (see Definition 1.3.1, page 15) of the differential equation (3.1). These reduced solutions have the form

$$u_0^{1A}(x,y) = \Big(1 + \min\{-2\ln(1-x), -2\ln(1-y)\}\Big) f_1(x,y), \tag{3.262}$$

$$u_0^{1B}(x,y) = \left(1 + \min\left\{\frac{x}{1-x}, \frac{y}{1-y}\right\}\sqrt{(1-x)^2 + (1-y)^2}\right) f_1(x,y), \tag{3.263}$$

and they are depicted in Figure 3.15.

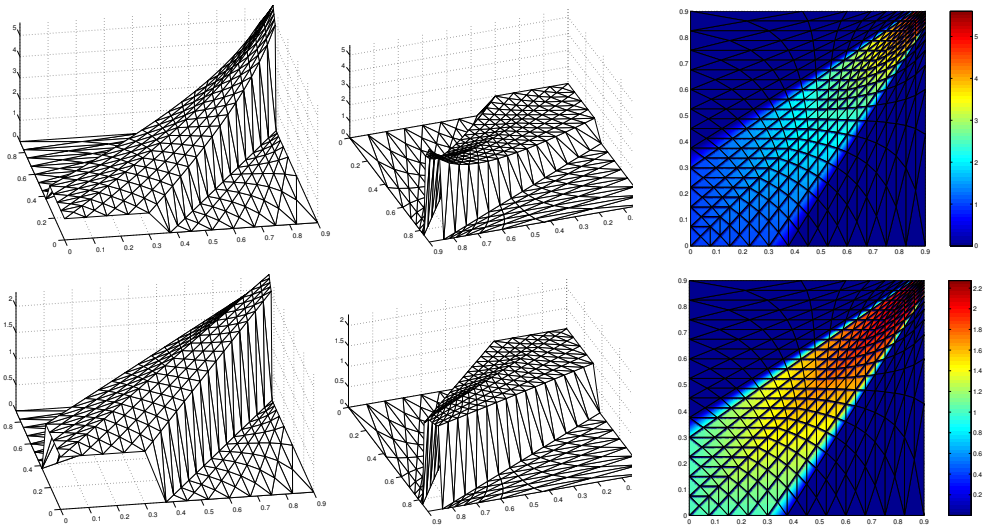

Figure 3.15: Reduced solutions u_0^{1A} (up) and u_0^{1B} (down).

Due to the discontinuous data and the nonzero right-hand side f_1 we are not able to construct the (zeroth-order) asymptotic expansion (see Linß and Stynes (2001) for sufficient assumptions). Hence, we use a small trick and prescribe the

reduced solution u_0. If $u_0 = xy$, then $\boldsymbol{b}_{1A} \cdot \nabla u_0 = \frac{x}{2} - xy + \frac{y}{2} =: f_2$ and we may construct the (zeroth-order) asymptotic expansion $u_{as}^{(E1)}$ (Figure 3.16) for the problem with the data $[\boldsymbol{b}_{1A}, u_{b2}, f_2] = [\boldsymbol{b}_{1A}, 0, \frac{x}{2} - xy + \frac{y}{2}]$. It has the form

$$u_{as}^{(E1)} \;=\; xy\left(1 - \exp\left(\frac{0.05}{\varepsilon}(x - 0.9)\right)\right)\left(1 - \exp\left(\frac{0.05}{\varepsilon}(y - 0.9)\right)\right) \qquad (3.264)$$

and we use it as a continuous test problem ($u_{as}^{(E1)}$ is the solution of the differential equation (3.1) with the data $[\boldsymbol{b}_{1A}, 0, Lu_{as}^{(E1)}]$).

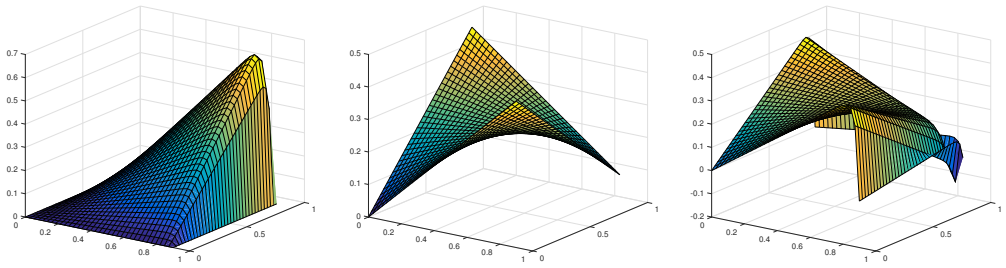

Figure 3.16: Graphs of the functions $u_{as}^{(E1)}$, f_2 and $Lu_{as}^{(E1)}$, respectively. The function $u_{as}^{(E1)}$ is the (zeroth-order) asymptotic expansion of the solution of the boundary value problem (3.1) with the data $[\boldsymbol{b}, u_b, f] = [\boldsymbol{b}_{1A}, 0, f_2]$. It is also the classical solution of the same differential equation with $[\boldsymbol{b}, u_b, f] = [\boldsymbol{b}_{1A}, 0, Lu_{as}^{(E1)}]$. In this example we considered $\varepsilon = 10^{-3}$.

Firstly, let us solve Example 1 using the SUPG method with the continuous piecewise linear finite elements, the stabilization parameter $\delta_K = h_K/(2\|\boldsymbol{b}\|_{\infty,K})$ and consider three types of meshes (see Figure 3.17).

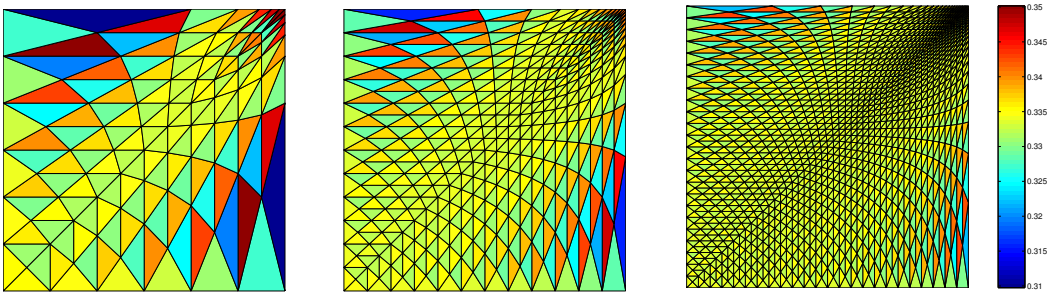

Figure 3.17: Meshes considered in Example 1 formed by 144, 576 and 2304 elements, respectively. The color scale indicates the value $\theta_K/\|\mathrm{div}\,\boldsymbol{b}\|_{\infty,K}$ for $\boldsymbol{b} = \boldsymbol{b}_{1B}$ and all $K \in \mathcal{T}_h$.

Figure 3.18 shows solutions computed using the SUPG method — each column corresponds to a different mesh (with 144, 576 and 2304 elements, respectively) and each row to a different choice of ε (we consider $\varepsilon = 10^{-3}$, 10^{-4} and 10^{-5}, respectively). We observe that the discrete solution contains spurious oscillations at inner characteristic layers, in particular for $\varepsilon = 10^{-5}$.

If we employ the new method we obtain oscillation-free solutions (see Figure 3.19). Further, using our test problem $u_{as}^{(E1)}$ we may verify experimentally the result of Theorem 3.5.1 (page 75). Hence, we consider $\boldsymbol{b} = \boldsymbol{b}_{1A}$ and $\varepsilon = 10^{-2}, 10^{-3}$ and 10^{-4}. Table 3.2 contains the computational errors in several types of norms.

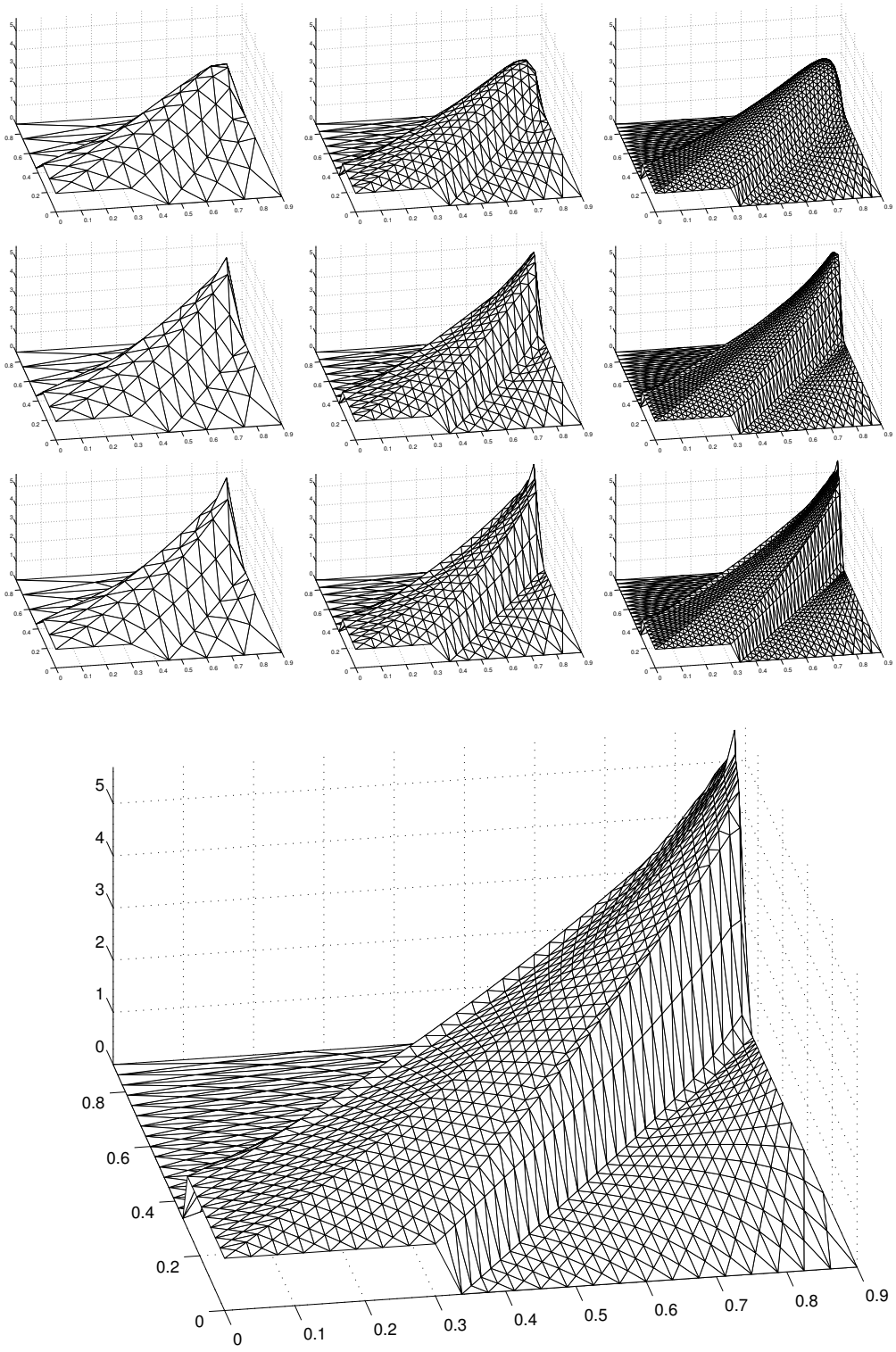

Figure 3.18: Solutions of Example 1 with $\boldsymbol{b} = \boldsymbol{b}_{1A}$ obtained by the SUPG method. Each column corresponds to a different mesh (with 144, 576 and 2304 elements, respectively) and each row to a different choice of ε (we consider $\varepsilon = 10^{-3}$, 10^{-4} and 10^{-5}). The bottom right solution is displayed enlarged.

Table 3.2: Computational errors in several types of norms. We applied the new method to Example 1 using piecewise linear finite elements with $\boldsymbol{b} = \boldsymbol{b}_{1A}$ and considered $\varepsilon = 10^{-2}, 10^{-3}$ and 10^{-4} ($e_{\boldsymbol{b}}$ stands for $(\sum_K \frac{|\boldsymbol{d}_{K,1}|}{2|\boldsymbol{b}_K|}\|\boldsymbol{b}_K \cdot \nabla e_h\|^2_{0,K})^{1/2}$).

ε	Elms	$\lvert e_h\rvert_{1,\Omega}$	$\|e_h\|_{0,2,\Omega}$	$e_{\boldsymbol{b}}$	$\|e_h\|_{\infty,d}$	$\|\|\|e_h\|\|\|_b$
1E-2	36	2.406E-01	3.157E-02	2.515E-02	1.451E-02	3.711E-02
1E-2	144	1.636E-01	1.668E-02	1.160E-02	9.730E-03	2.118E-02
1E-2	576	9.671E-02	8.574E-03	4.644E-03	5.075E-03	1.129E-02
1E-2	2304	5.145E-02	4.353E-03	1.727E-03	2.766E-03	5.711E-03
1E-2	9216	2.629E-02	2.198E-03	6.293E-04	1.499E-03	2.848E-03
1E-3	36	1.390E-00	1.313E-01	1.092E-01	6.233E-02	1.294E-01
1E-3	144	1.867E-00	7.945E-02	9.324E-02	1.344E-01	1.150E-01
1E-3	576	1.533E-00	4.373E-02	5.166E-02	1.350E-01	7.305E-02
1E-3	2304	9.695E-01	2.272E-02	2.259E-02	7.929E-02	3.920E-02
1E-3	9216	5.442E-01	1.150E-02	8.887E-03	4.524E-02	1.993E-02
1E-4	36	1.483E-00	1.588E-01	1.513E-01	1.274E-02	1.653E-01
1E-4	144	2.146E-00	1.182E-01	1.281E-01	2.948E-02	1.386E-01
1E-4	576	2.529E-00	8.376E-02	9.234E-02	6.944E-02	1.017E-01
1E-4	2304	3.664E-00	5.449E-02	8.159E-02	1.337E-01	9.217E-02
1E-4	9216	5.736E-01	3.300E-02	8.906E-02	2.200E-01	1.068E-01

The experimental order of convergence (EOC) with respect to the energy norm $|||\cdot|||_b$ (cf. Remark 3.5.2, page 82) is in the case when $\varepsilon = 10^{-2}$ equal to 0.809, 0.908, 0.983 and 1.004, respectively. Thus, it increases with increasing number of elements (decreasing h), which is in line with our expectations. For $\varepsilon = 10^{-3}$ we obtain $EOC = 0.166$, 0.655, 0.898 and 0.976, respectively. For smaller ε the convergence is achieved when the boundary layer is resolved.

We may also apply the approach derived in Section 3.7 and obtain the solution of Example 1 using continuous piecewise quadratic finite elements — see Figure 3.20. Considering the function values in mesh-nodes only, the solutions are oscillation-free. However, since we are employing the quadratic finite elements, the oscillations occur inside elements (see Figure 3.21).

From the computational errors it follows that in L^2-norm the method provides similar error values as in the case of piecewise linear finite elements (cf. Table 3.3). This could have several causes — either the layers are better resolved by piecewise linear finite elements or the method should be improved. One possible improvement may include the use of curvilinear elements. The vector $\boldsymbol{b}_K^{(1)}$ could then point in a non-constant direction. Nevertheless, comparing the error in mesh-nodes we find out that the use of piecewise quadratic finite elements provides better results. Unfortunately, we were not able to derive the exact form of the energy norm in this case (we did not prove the coercivity).

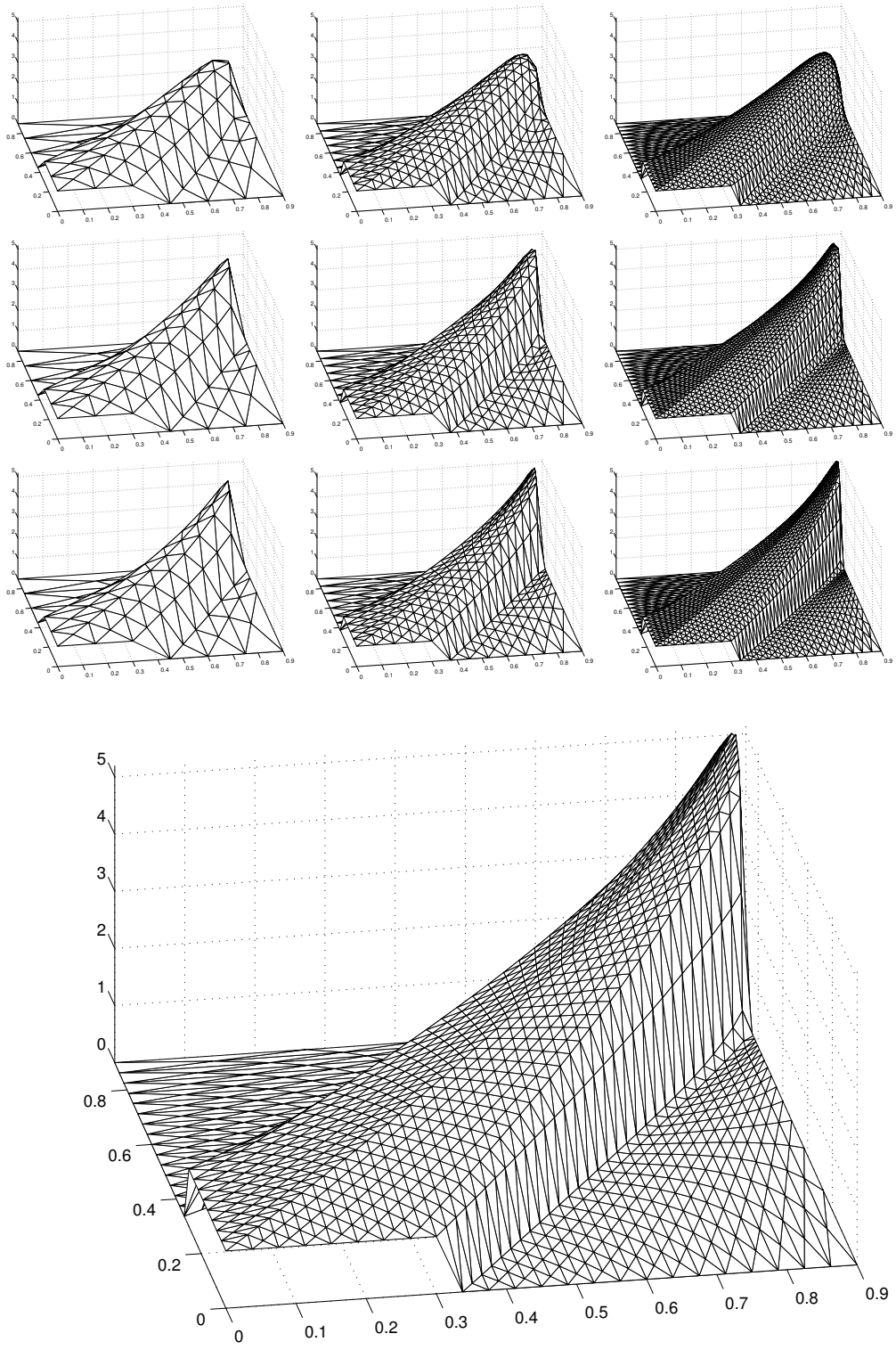

Figure 3.19: Solutions of Example 1 with $\boldsymbol{b} = \boldsymbol{b}_{1A}$ obtained by the new method using continuous piecewise linear finite elements. Each column corresponds to a different mesh (with 144, 576 and 2304 elements, respectively) and each row to a different choice of ε (we consider $\varepsilon = 10^{-3}$, 10^{-4} and 10^{-5}). The bottom right solution is displayed enlarged.

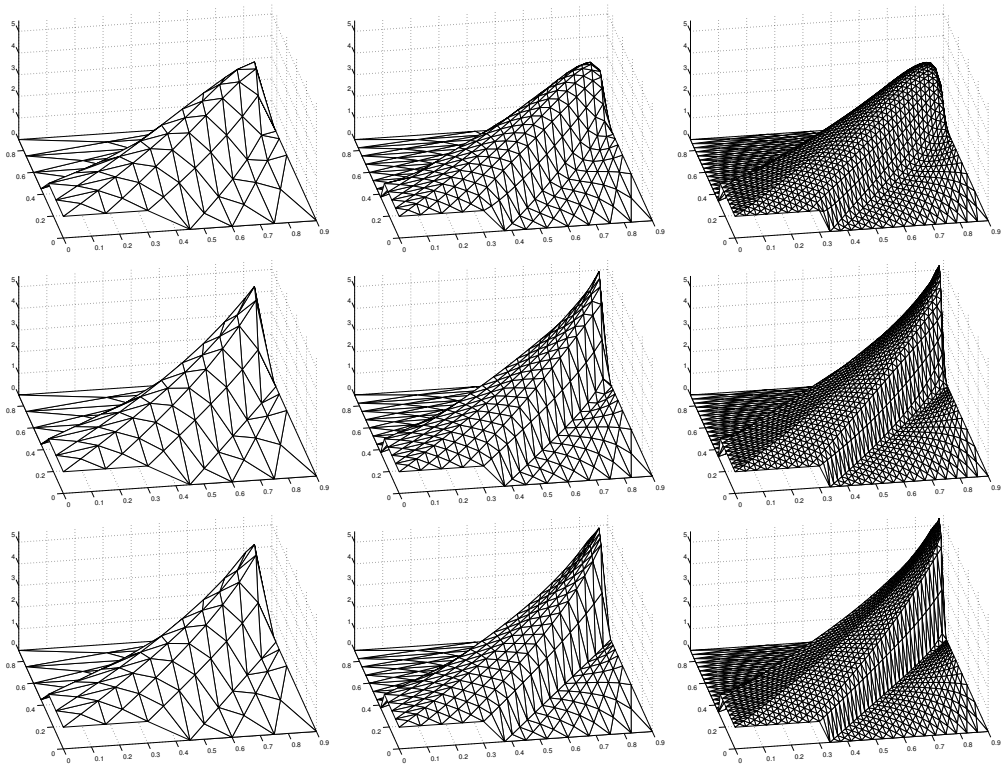

Figure 3.20: Solutions of Example 1 with $\boldsymbol{b} = \boldsymbol{b}_{1A}$ obtained by the new method using continuous piecewise quadratic finite elements. Each column corresponds to a different mesh (with 144, 576 and 2304 elements, respectively) and each row to a different choice of ε (we consider $\varepsilon = 10^{-3}$, 10^{-4} and 10^{-5}).

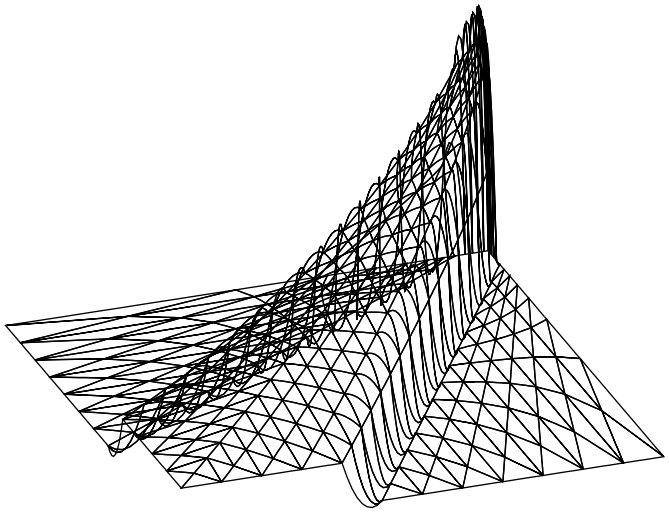

Figure 3.21: Solution of Example 1 with $\boldsymbol{b} = \boldsymbol{b}_{1A}$ obtained by the new method using continuous piecewise quadratic finite elements, mesh with 576 elements and $\varepsilon = 10^{-4}$. Despite the fact that the solution is oscillation-free in mesh-nodes, it contains oscillations inside layer-elements.

Table 3.3: Computational errors in several types of norms. We applied the new method to Example 1 using piecewise quadratic finite elements with $\boldsymbol{b} = \boldsymbol{b}_{1A}$ and considered $\varepsilon = 10^{-2}, 10^{-3}$ and 10^{-4}. Here $e_h = u - u_h$, $\xi_h = R_h u - u_h$, $R_h u \in P_2$ is the Lagrange interpolation of u, $\|\cdot\|_{\infty,d,P_2}$ is the discrete maximum norm over all P_2-nodes, whereas $\|\cdot\|_{\infty,d,P_1}$ is the discrete maximum norm over all P_1-nodes.

ε	Elms	$\lvert\xi_h\rvert_{1,\Omega}$	$\|\xi_h\|_{0,2,\Omega}$	$\|e_h\|_{0,2,\Omega}$	$\|e_h\|_{\infty,d,P_2}$	$\|e_h\|_{\infty,d,P_1}$
1E-2	36	7.524E-02	5.003E-03	4.679E-02	1.591E-02	1.243E-02
1E-2	144	3.651E-02	1.895E-03	2.643E-02	7.542E-03	5.648E-03
1E-2	576	1.366E-02	8.839E-04	1.422E-02	2.387E-03	2.387E-03
1E-2	2304	4.996E-03	4.712E-04	7.407E-03	1.098E-03	1.098E-03
1E-2	9216	2.052E-03	2.488E-04	3.785E-03	5.452E-04	5.452E-04
1E-3	36	3.587E-01	1.948E-02	1.078E-01	1.170E-01	8.867E-02
1E-3	144	4.505E-01	1.550E-02	7.548E-02	1.195E-01	4.283E-02
1E-3	576	3.880E-01	6.459E-03	4.683E-02	5.156E-02	2.841E-02
1E-3	2304	1.816E-01	2.243E-03	2.578E-02	1.610E-02	1.237E-02
1E-3	9216	6.038E-02	9.528E-04	1.334E-02	4.643E-03	4.381E-03
1E-4	36	4.445E-01	2.110E-02	1.144E-01	7.832E-02	4.051E-02
1E-4	144	3.437E-01	1.059E-02	7.870E-02	5.238E-02	5.238E-02
1E-4	576	3.857E-01	6.764E-03	5.256E-02	1.061E-01	5.574E-02
1E-4	2304	8.435E-01	7.274E-03	3.641E-02	1.803E-01	4.613E-02
1E-4	9216	1.586E-00	5.741E-03	2.556E-02	1.769E-01	2.396E-02

3.8.2 Example 2, zero divergence

Let us now consider the equation (3.1) in $(X,Y)^2 \subset (0,1)^2$, where $X = \frac{1}{20}\sqrt{2}$ and $Y = \frac{7}{20}\sqrt{2}$. The right-hand side $f = f_3$ satisfies $f_3(x,y) = 1$ whenever $\left(\frac{7}{30}\right)^2 \le x^2 + y^2 \le \left(\frac{11}{30}\right)^2$ and $f_3(x,y) = 0$ otherwise. The boundary condition u_{b2} satisfies $u_{b2} = 1$ in $\{\boldsymbol{x} \in \Gamma_-,\ f(\boldsymbol{x}) = 1\}$ and $u_{b2} = 0$ otherwise (see Figure 3.22). Again, we use two definitions of the vector field $\boldsymbol{b}$

$$\text{Example } 2A:\ \boldsymbol{b}(x,y) = (-y,x)^T, \quad \text{Example } 2B:\ \boldsymbol{b}(x,y) = \frac{1}{\sqrt{x^2+y^2}}(-y,x)^T, \tag{3.265}$$

where the second one is used for a comparison of the matrices of the mappings $R_K^{(2)}$ and $\Pi_{b,K}^{(2)}$.

The circle (streamline) passing through the vertices $[X,Y]$ and $[Y,X]$ divide the diagonal (with the endpoints $[X,X]$ and $[Y,Y]$) into two parts in the ratio 2:1. Indeed, the length of the square's diagonal is $\frac{3}{5}$, the radius of the considered circle (streamline) is $\sqrt{X^2+Y^2} = \frac{1}{2}$ and thus, the length of the larger part of the diagonal is $\frac{1}{2} - X\sqrt{2} = \frac{2}{5}$ (hence, the length of the shorter part is $\frac{1}{5}$). We can now construct the triangulation of the square $(X,Y)^2$ by constructing the streamlines (circles) in such a way, that the partition of the square's diagonal (with endpoints $[X,X]$ and $[Y,Y]$) is equidistant.

For instance, if we divide the diagonal into $2j + j = 3j$ parts, then the length of each part is $\frac{1}{5j}$. Since we would like to obtain an isotropic triangulation of Ω and the height of each triangle is approximately given by the distance between two neighboring streamlines (i.e. $\frac{1}{5j}$), we have to divide each streamline using

a mesh step $h = \frac{1}{5j}\frac{2\sqrt{3}}{3} = \frac{2\sqrt{3}}{15j}$. In order to obtain an equidistant partition of each streamline s we change h into h_s, which is the closest value to h allowing the equidistant partition of the streamline s (we want to avoid small odd segments near the boundary). Then one can show that there holds $|h - h_s| \leq \frac{3\pi^2}{2L_s}h^2$, where L_s is the length of the streamline s. Hence, away from the corners $[X, X]$ and $[Y, Y]$, the partition of all streamlines is *almost equidistant*.

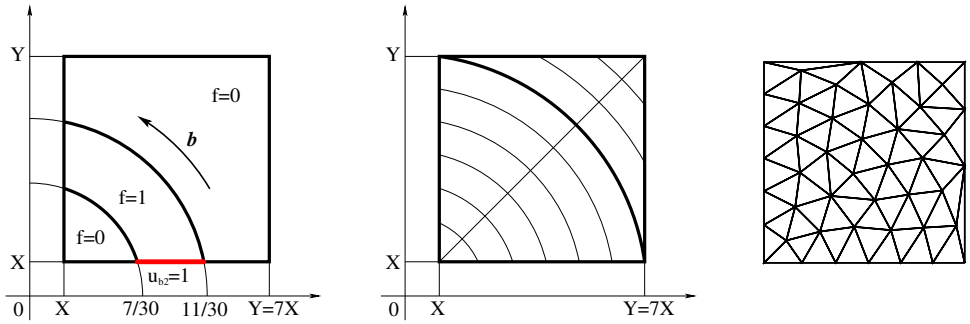

Figure 3.22: Definition of the Example 2 data.

For both data combinations $[\boldsymbol{b}_{2A}, u_{b2}, f_3]$ and $[\boldsymbol{b}_{2B}, u_{b2}, f_3]$ we may compute the reduced solutions u_0^{2A} and u_0^{2B} (see Definition 1.3.1, page 15) of the differential equation (3.1). These reduced solutions have the form

$$\begin{aligned} u_0^{2A}(x,y) &= \left[1 + \operatorname{atan}\left(\frac{y}{x}\right) - \operatorname{asin}\left(\frac{1}{20}\sqrt{\frac{2}{x^2+y^2}}\right)\right] f_3(x,y), \\ u_0^{2B}(x,y) &= \left[1 + \left(\operatorname{atan}\left(\frac{y}{x}\right) - \operatorname{asin}\left(\frac{1}{20}\sqrt{\frac{2}{x^2+y^2}}\right)\right)\sqrt{x^2+y^2}\right] f_3(x,y), \end{aligned}$$

and they are depicted in Figure 3.23.

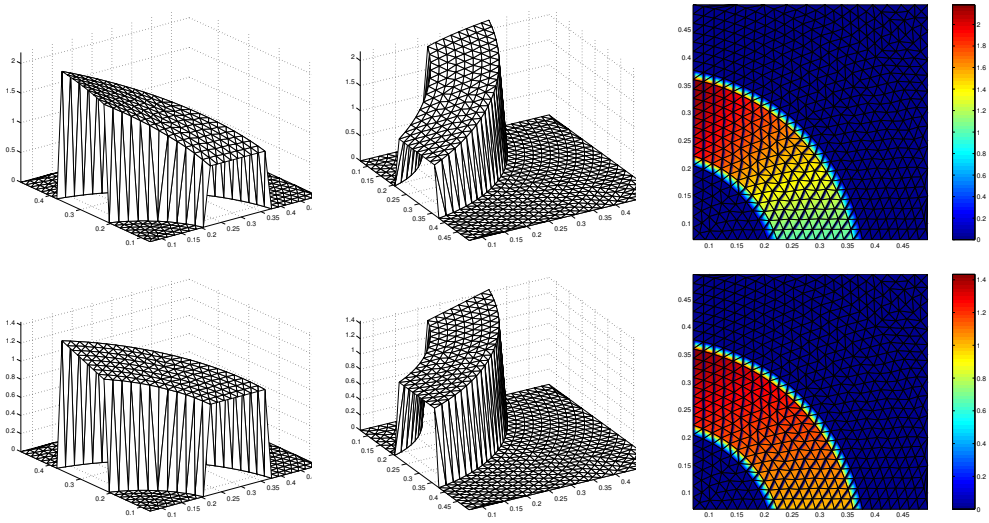

Figure 3.23: Reduced solutions u_0^{2A} (up) and u_0^{2B} (down).

In order to construct the (zeroth-order) asymptotic expansion of the solution of the boundary value problem (3.1) we again prescribe the reduced solution u_0. This time we assume that $u_0 = (Y - x)(y - X)$, then $\boldsymbol{b}_{2A} \cdot \nabla u_0 = y^2 - yX + xY - x^2 =: f_4$ and we may construct the (zeroth-order) asymptotic expansion

$u_{as}^{(E2)}$ (Figure 3.24) for the problem with the data $[\boldsymbol{b}, u_b, f] = [\boldsymbol{b}_{2A}, 0, f_4]$. It has the form

$$u_{as}^{(E2)} \;=\; (Y-x)(y-X)\left(1-\exp\left(\frac{y}{\varepsilon}(X-x)\right)\right)\left(1-\exp\left(\frac{x}{\varepsilon}(y-Y)\right)\right) \quad (3.266)$$

and we use it as a continuous test problem ($u_{as}^{(E2)}$ is a solution of the differential equation (3.1) with the data $[\boldsymbol{b}_{2A}, 0, Lu_{as}^{(E2)}]$).

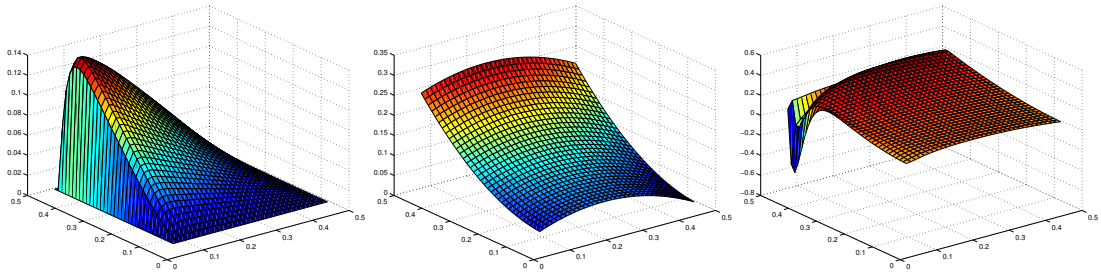

Figure 3.24: Graphs of the functions $u_{as}^{(E2)}$, f_4 and $Lu_{as}^{(E2)}$, respectively. The function $u_{as}^{(E2)}$ is the (zeroth-order) asymptotic expansion of the solution of the boundary value problem (3.1) with the data $[\boldsymbol{b}, u_b, f] = [\boldsymbol{b}_{2A}, 0, f_4]$. It is also the classical solution of the same differential equation with $[\boldsymbol{b}, u_b, f] = [\boldsymbol{b}_{2A}, 0, Lu_{as}^{(E2)}]$. In this example we consider $\varepsilon = 2 \times 10^{-3}$.

Firstly, let us solve Example 2 using the SUPG method with the continuous piecewise linear finite elements, the stabilization parameter $\delta_K = h_K/(2\|\boldsymbol{b}\|_{\infty,K})$ and as in Example 1 we consider three types of meshes (Figure 3.25).

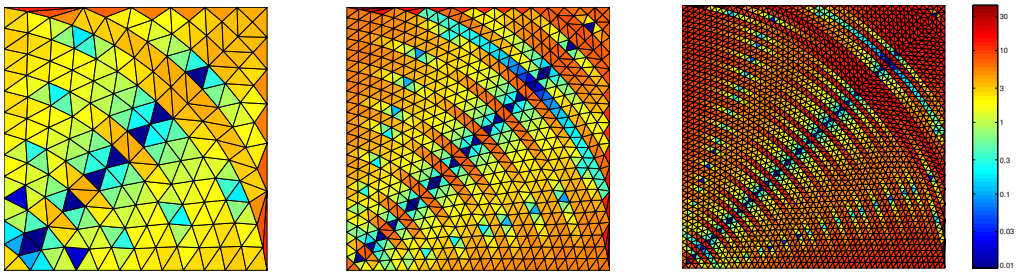

Figure 3.25: Meshes considered in Example 2 formed by 284, 1124 and 4498 elements, respectively. The color scale indicates the value θ_K/h_K for $\boldsymbol{b} = \boldsymbol{b}_{2A}$ and all $K \in \mathcal{T}_h$.

Figure 3.26 shows solutions computed using the SUPG method — each column corresponds to a different mesh (with 284, 1124 and 4498 elements, respectively) and each row to a different choice of ε (we consider $\varepsilon = 10^{-3}$, 10^{-4} and 10^{-5}). We observe that the discrete solution contains spurious oscillations at inner characteristic layers, in particular for $\varepsilon = 10^{-5}$.

If we again employ the new method we obtain oscillation-free solutions (see Figure 3.27). Further, using our test problem $u_{as}^{(E2)}$ we may try to verify experimentally the result of Theorem 3.5.2 (page 79). Hence, we consider $\boldsymbol{b} = \boldsymbol{b}_{2A}$ and $\varepsilon = 10^{-2}, 10^{-3}$ and 10^{-4}. Table 3.4 again shows the computational errors in several types of norms. We observe, that the solution fails to converge in the energy norm $|||\cdot|||_{b,*}$ and it only converges in L^2-norm. This is caused by the fact that,

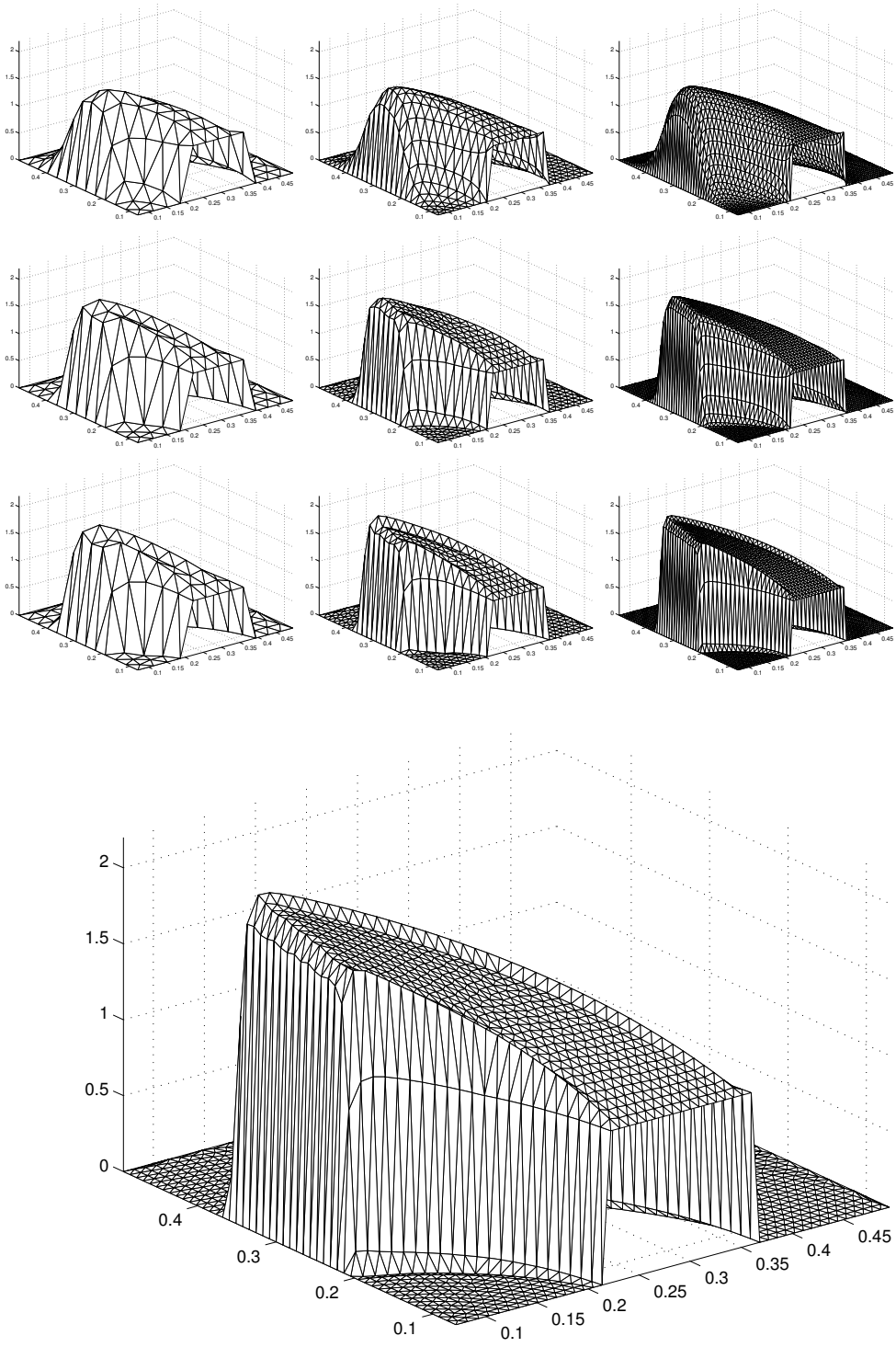

Figure 3.26: Solutions of Example 2 with $\boldsymbol{b} = \boldsymbol{b}_{2A}$ obtained by the SUPG method. Each column corresponds to a different mesh (with 284, 1124 and 4498 elements, respectively) and each row to a different choice of ε (we consider $\varepsilon = 10^{-3}$, 10^{-4} and 10^{-5}). The bottom right solution is displayed enlarged.

Table 3.4: Computational errors in several types of norms. We applied the new method to Example 2 using piecewise linear finite elements with $\boldsymbol{b} = \boldsymbol{b}_{2A}$ and considered $\varepsilon = 10^{-2}, 10^{-3}$ and 10^{-4}. We observe, that for small ε the solution converge only in L^2-norm ($e_{\boldsymbol{b}}$ stands for $(\sum_K \frac{|\boldsymbol{d}_{K,1}|}{2|\boldsymbol{b}_K|}\|\boldsymbol{b}_K \cdot \nabla e_h\|^2_{0,K})^{1/2}$).

ε	Elms	$\lvert e_h\rvert_{1,\Omega}$	$\|e_h\|_{0,2,\Omega}$	$e_{\boldsymbol{b}}$	$\|e_h\|_{\infty,d}$	$\|\|\|e_h\|\|\|_{b,*}$
1E-2	72	2.158E-01	5.008E-03	1.276E-02	2.070E-02	2.510E-02
1E-2	284	2.313E-01	2.778E-03	9.347E-03	1.351E-02	2.496E-02
1E-2	1124	2.396E-01	1.459E-03	6.903E-03	8.015E-03	2.493E-02
1E-2	4498	2.444E-01	7.852E-04	5.056E-03	4.647E-03	2.496E-02
1E-2	17956	2.472E-01	4.188E-04	3.666E-03	2.759E-03	2.499E-02
1E-3	72	3.867E-01	1.099E-02	2.882E-02	1.554E-02	3.146E-02
1E-3	284	4.673E-01	7.533E-03	2.457E-02	1.620E-02	2.871E-02
1E-3	1124	6.512E-01	4.619E-03	2.386E-02	2.025E-02	3.152E-02
1E-3	4498	8.917E-01	2.665E-03	2.094E-02	2.792E-02	3.513E-02
1E-3	17956	1.018E-00	1.490E-03	1.363E-02	3.125E-02	3.495E-02
1E-4	72	4.425E-01	1.170E-02	3.437E-02	1.277E-03	3.480E-02
1E-4	284	6.627E-01	8.333E-03	3.442E-02	7.002E-03	3.509E-02
1E-4	1124	8.864E-01	5.980E-03	3.428E-02	1.384E-02	3.542E-02
1E-4	4498	1.151E-00	4.228E-03	3.326E-02	1.944E-02	3.520E-02
1E-4	17956	1.409E-00	2.912E-03	2.868E-02	2.040E-02	3.196E-02

since the mesh was constructed heuristically, there does not hold $\theta_K = \mathcal{O}(h_K)$ (see Figure 3.25), which is crucial for estimates carried out in Theorem 3.5.2. Moreover, from Example 1 it follows, that for certain types of vector fields $\boldsymbol{b}$ it may be complicated (or even impossible) to construct a mesh satisfying $\theta_K \to 0$.

Again, we may use continuous piecewise quadratic finite elements for solving Example 2 and obtain solutions which are oscillation-free in mesh-nodes (see Figure 3.20). Visualization of the oscillations emerging from the element's interior is depicted in Figure 3.29.

The numerical experiments again provides improved computational errors in the discrete maximum norm in mesh-nodes (as compared to the linear case) and unimproved results in the L^2-norm (cf. Table 3.5).

Let us now verify our result from Section 3.6 considering Example 2 with $\boldsymbol{b} = \boldsymbol{b}_{2A}$, $\varepsilon = 10^{-3}$ and mesh containing 1124 elements. As we already know from Section 2.1.4, upwind scheme in 1D adds too much artificial diffusion to the original finite element method, and thus, the discrete solution is smeared (cf. Figure 2.3, page 39). This happens in 2D as well, therefore we apply the layer correction of Section 3.6 and obtain more accurate solution (see Figure 3.30). We cannot apply it to our test solution $u_{as}^{(E2)}$ since it contains corner expansion (two multiplied exponential functions), and hence, the technique of Section 3.6 fails. The remedy will be a subject of the future work.

Last thing we would like to mention is the way how the vector field $\boldsymbol{b}$ affects the structure of the mappings (or corresponding matrices) $R_K^{(2)}$ and $\Pi_{b,K}^{(2)}$ from Section 3.7. As $h \to 0$, the entries of the matrices of the mappings $R_K^{(2)}$ and $\Pi_{b,K}^{(2)}$ tend to some constant values. The matrix of the mapping $\Pi_{b,K}^{(2)}$ converges (probably under some mesh-related conditions) to the matrix of the orthogonal

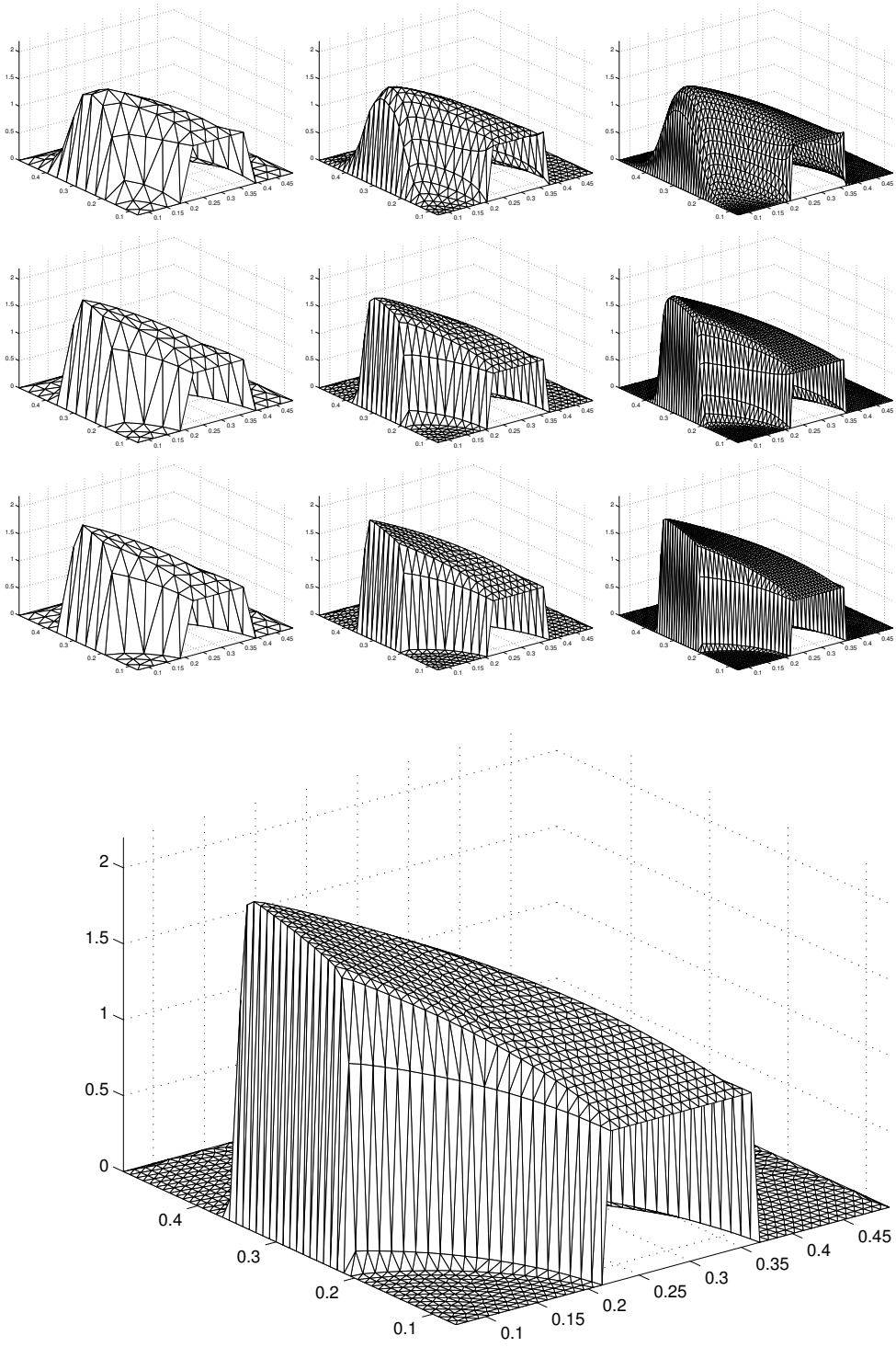

Figure 3.27: Solutions of Example 2 with $\boldsymbol{b} = \boldsymbol{b}_{2A}$ obtained by the new method. Each column corresponds to a different mesh (with 284, 1124 and 4498 elements, respectively) and each row to a different choice of ε (we consider $\varepsilon = 10^{-3}$, 10^{-4} and 10^{-5}). The bottom right solution is displayed enlarged.

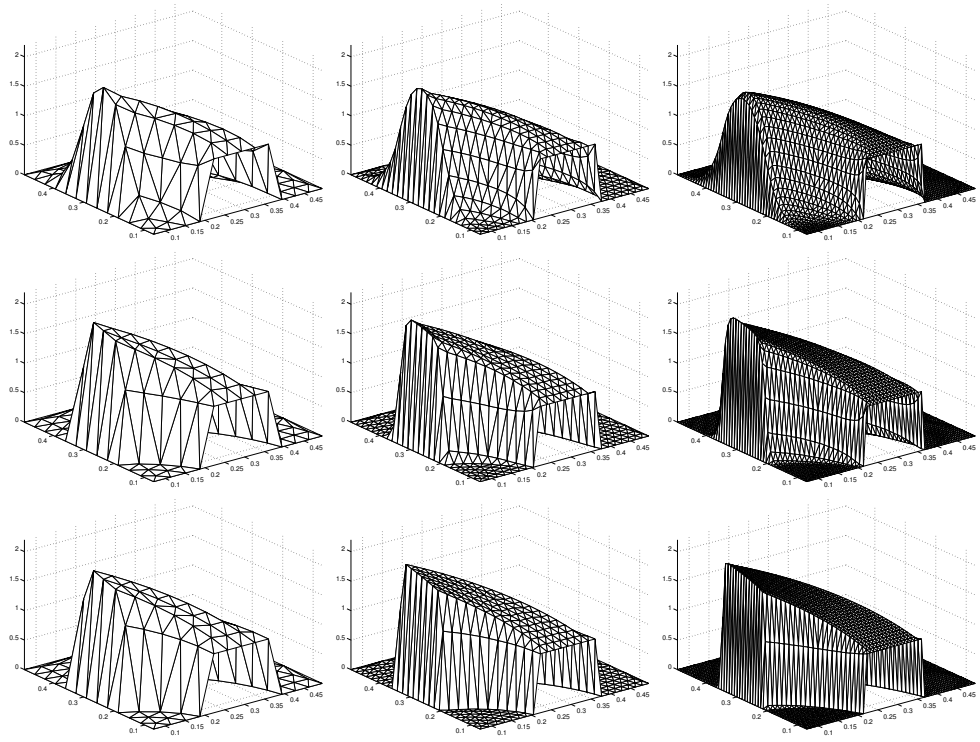

Figure 3.28: Solutions of Example 2 with $\boldsymbol{b} = \boldsymbol{b}_{2A}$ obtained by the new method using continuous piecewise quadratic finite elements. Each column corresponds to a different mesh (with 284, 1124 and 4498 elements, respectively) and each row to a different choice of ε (we consider $\varepsilon = 10^{-3}$, 10^{-4} and 10^{-5}).

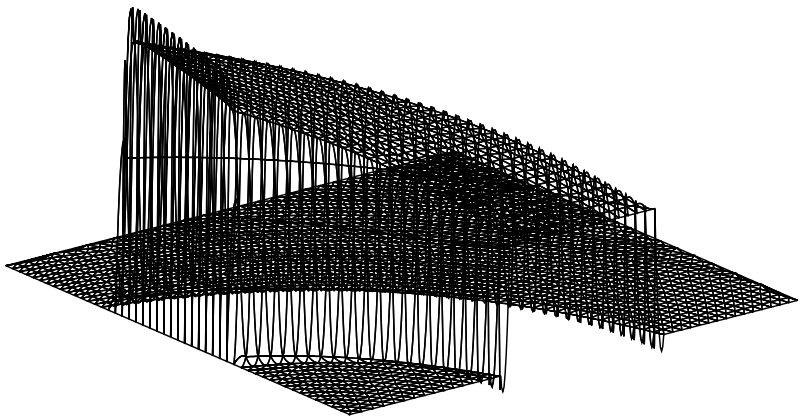

Figure 3.29: Solution of Example 2 with $\boldsymbol{b} = \boldsymbol{b}_{2A}$ obtained by the new method using continuous piecewise quadratic finite elements, mesh with 1124 elements and $\varepsilon = 10^{-4}$. Although the solution is oscillation-free in mesh-nodes, it contains oscillations inside layer-elements.

Table 3.5: Computational errors in several types of norms. We applied the new method to Example 2 using piecewise quadratic finite elements with $\boldsymbol{b} = \boldsymbol{b}_{2A}$ and considered $\varepsilon = 10^{-2}, 10^{-3}$ and 10^{-4}. Here $e_h = u - u_h$, $\xi_h = R_h u - u_h$, $R_h u \in P_2$ is the Lagrange interpolation of u, $\|\cdot\|_{\infty,d,P_2}$ is the discrete maximum norm over all P_2-nodes, whereas $\|\cdot\|_{\infty,d,P_1}$ is the discrete maximum norm over all P_1-nodes.

ε	Elms	$\lvert\xi_h\rvert_{1,\Omega}$	$\|\xi_h\|_{0,2,\Omega}$	$\|e_h\|_{0,2,\Omega}$	$\|e_h\|_{\infty,d,P_2}$	$\|e_h\|_{\infty,d,P_1}$
1E-2	72	3.785E-02	9.884E-04	5.390E-03	1.053E-02	4.797E-03
1E-2	284	1.561E-02	3.087E-04	3.139E-03	3.724E-03	2.025E-03
1E-2	1124	5.964E-03	1.230E-04	1.642E-03	1.428E-03	1.176E-03
1E-2	4498	3.135E-03	7.922E-05	8.665E-04	9.872E-04	9.590E-04
1E-2	17956	2.596E-03	6.965E-05	4.402E-04	9.112E-04	9.056E-04
1E-3	72	8.459E-02	1.067E-03	8.050E-03	1.438E-02	4.418E-03
1E-3	284	1.049E-01	1.017E-03	5.610E-03	1.850E-02	8.684E-03
1E-3	1124	1.316E-01	8.360E-04	4.073E-03	1.732E-02	6.888E-03
1E-3	4498	1.555E-01	4.925E-04	2.809E-03	1.834E-02	3.366E-03
1E-3	17956	1.086E-01	1.803E-04	1.714E-03	1.039E-02	1.100E-03
1E-4	72	3.882E-02	4.718E-04	8.033E-03	9.013E-03	4.944E-03
1E-4	284	1.173E-01	4.217E-04	5.707E-03	1.950E-02	7.196E-03
1E-4	1124	1.454E-01	3.407E-04	4.074E-03	1.773E-02	4.859E-03
1E-4	4498	2.040E-01	3.445E-04	2.917E-03	2.030E-02	4.878E-03
1E-4	17956	2.757E-01	3.229E-04	2.104E-03	1.832E-02	6.354E-03

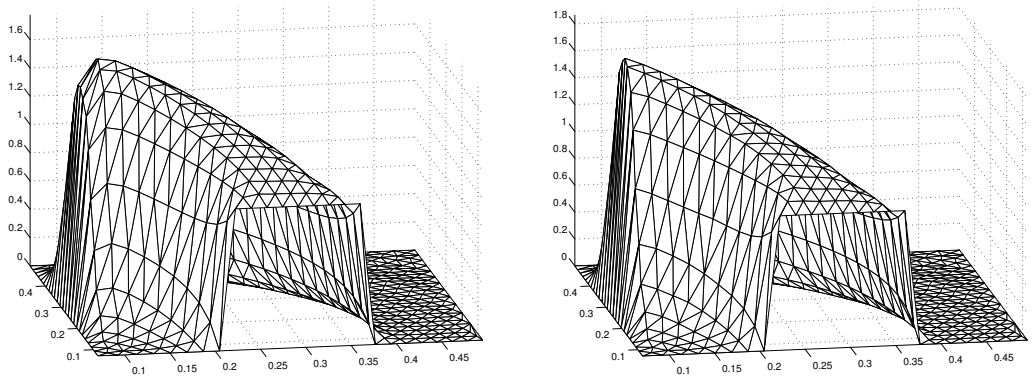

Figure 3.30: Solution of Example 2 with $\boldsymbol{b} = \boldsymbol{b}_{2A}$, $\varepsilon = 10^{-3}$ and the mesh containing 1124 elements, obtained by the new method using continuous piecewise linear finite elements. In the right figure the layer correction was applied and the solution is not smeared.

L^2-projection of $P_2(K)$ onto $P_1(K)$ (cf. equality 3.245), whereas the matrix of the mapping $R_K^{(2)}$ tends to the matrix given by 3.252 (page 104). In Figure 3.31 one can find a comparison of all considered vector fields . The error values of entries given by the vector fields $\boldsymbol{b}_{1A}$ and $\boldsymbol{b}_{1B}$ are depicted in the upper row, whereas in the bottom one can find the error values of entries given by the vector fields $\boldsymbol{b}_{2A}$ and $\boldsymbol{b}_{2B}$. Similarly, Figures 3.32–3.35 show not only a comparison of the vector fields $\boldsymbol{b}_{2A}$ and $\boldsymbol{b}_{2B}$, but also the convergence of the respective matrix entries when the mesh is refined.

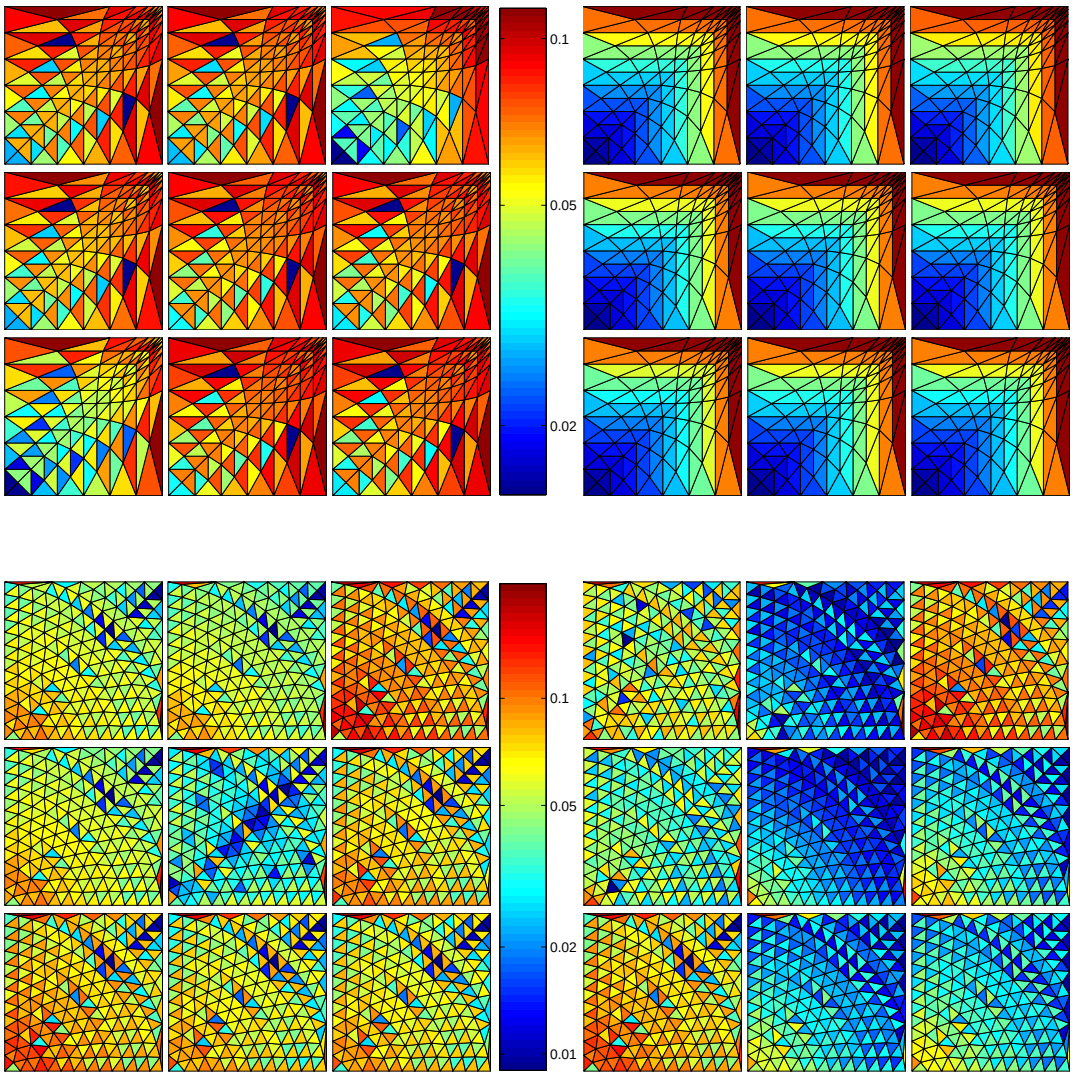

Figure 3.31: In any quarter, each square $(i, j), i, j = 1, 2, 3$, corresponds to one entry r_{ij}^K of the matrix of the mappings $R_K^{(2)}$, $K \in \mathcal{T}_h$. The color of each element indicates how close is this entry to its limit state. Up: Example 1, 144 elements, $\boldsymbol{b}_{1B}$ left, $\boldsymbol{b}_{1A}$ right; Down: Example 2, 284 elements, $\boldsymbol{b}_{2B}$ left, $\boldsymbol{b}_{2A}$ right

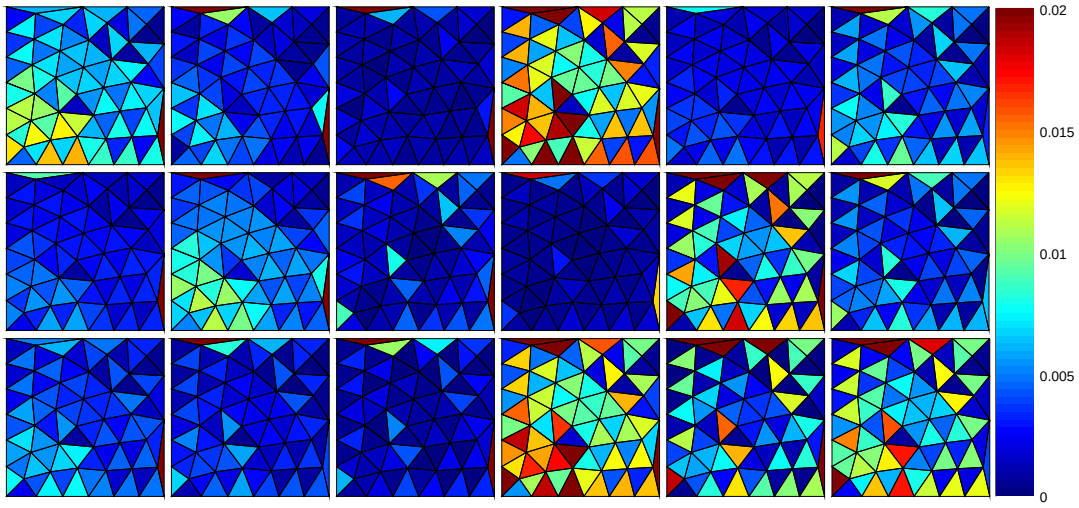

Figure 3.32: Each square $(m, j), m = 1, 2, 3, j = 1, 2, \ldots, 6$, corresponds to one entry $\mu_{K,m}^{(j)}$ of the matrix of the mappings $\Pi_{b,K}^{(2)}$, $K \in \mathcal{T}_h$. The color of each element indicates how close is this entry to its limit state. (Example 2B, 72 elements)

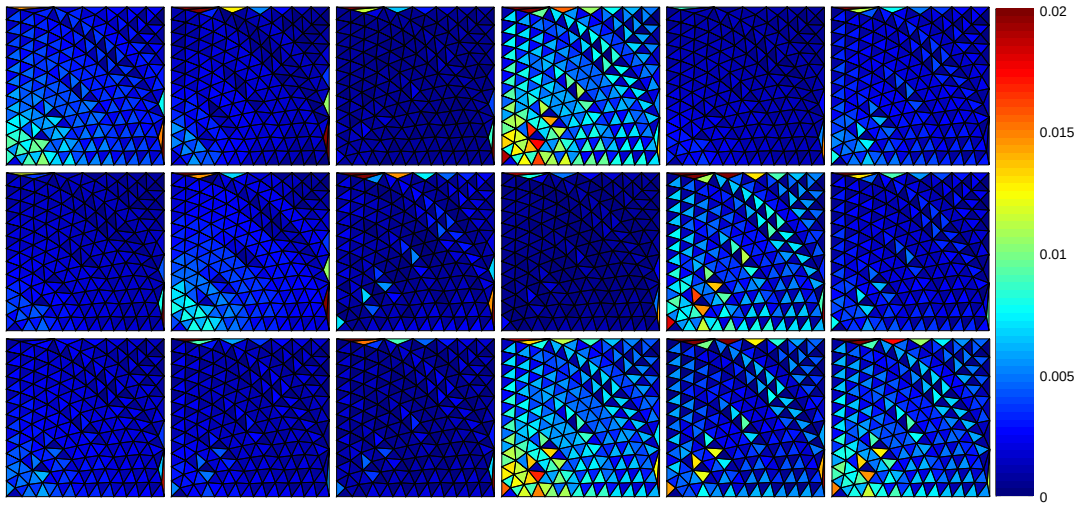

Figure 3.33: Each square $(m, j), m = 1, 2, 3, j = 1, 2, \ldots, 6$, corresponds to one entry $\mu_{K,m}^{(j)}$ of the matrix of the mappings $\Pi_{b,K}^{(2)}$, $K \in \mathcal{T}_h$. The color of each element indicates how close is this entry to its limit state. (Example 2B, 284 elements)

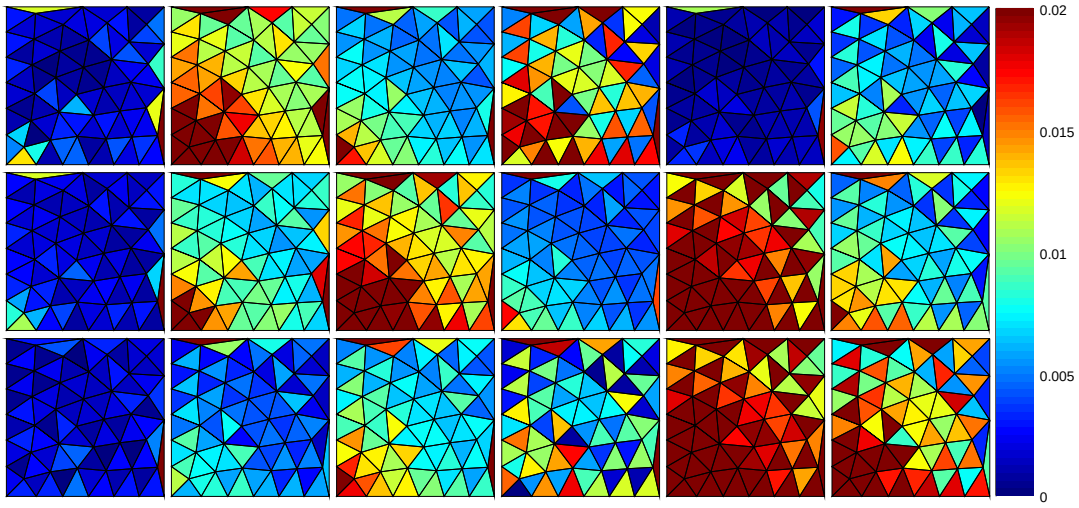

Figure 3.34: Each square $(m, j), m = 1, 2, 3, j = 1, 2, \ldots, 6$, corresponds to one entry $\mu_{K,m}^{(j)}$ of the matrix of the mappings $\Pi_{b,K}^{(2)}$, $K \in \mathcal{T}_h$. The color of each element indicates how close is this entry to its limit state. (Example 2A, 72 elements)

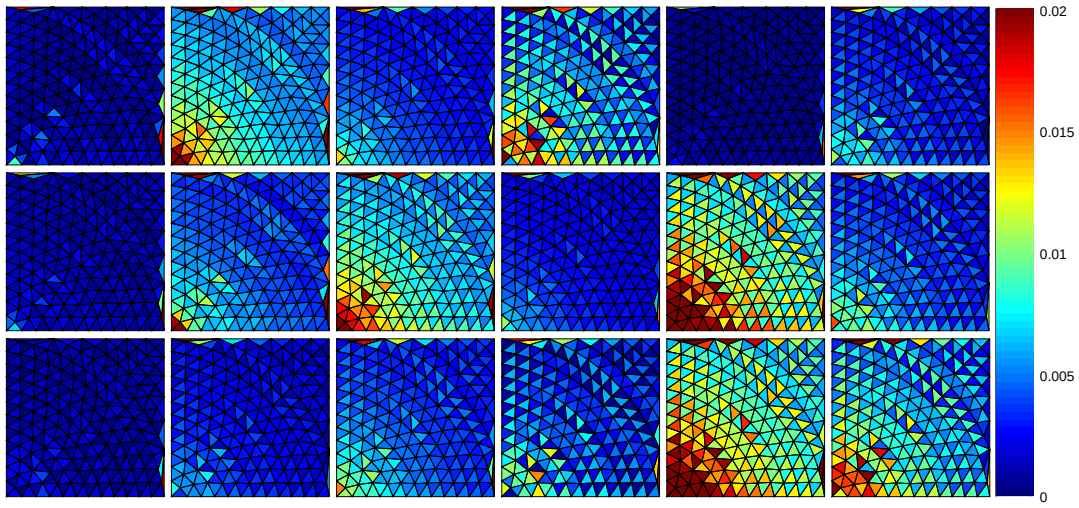

Figure 3.35: Each square $(m, j), m = 1, 2, 3, j = 1, 2, \ldots, 6$, corresponds to one entry $\mu_{K,m}^{(j)}$ of the matrix of the mappings $\Pi_{b,K}^{(2)}$, $K \in \mathcal{T}_h$. The color of each element indicates how close is this entry to its limit state. (Example 2A, 284 elements)

4. Appendix

4.1 Important theorems and lemmas

Lemma 4.1.1. *Suppose that $\boldsymbol{b}$ is in $\mathcal{C}^k$ of some neighborhood of $\overline{\Omega}$ and $k \geq 1$. Every solution to the initial value problem*

$$\zeta'(t) \;=\; \boldsymbol{b}(\zeta(t)), \qquad \zeta(0) \;=\; \zeta_0 \;\in\; \overline{\Omega}, \tag{4.1}$$

remains in some fixed neighborhood Ω_1 of $\overline{\Omega}$ for only a finite time in the time interval $(-\infty, +\infty)$ if and only if there exists a function $\phi \in \mathcal{C}^k(\Omega_1)$ so that $\boldsymbol{b} \cdot \nabla\phi > 0$ in $\overline{\Omega}$.

Proof. See Devinatz et al. (1974). □

Theorem 4.1.1 (Green's theorem)**.** *Let $\Omega \subset \mathbb{R}^n$ be a domain with Lipschitz-continuous boundary $\partial\Omega$. Then for each vector function $\boldsymbol{f} \in \mathcal{C}^1(\Omega)^n$*

$$\int_\Omega \operatorname{div} \boldsymbol{f}\, \mathrm{d}\boldsymbol{x} \;=\; \int_{\partial\Omega} \boldsymbol{f} \cdot \boldsymbol{n}\, \mathrm{d}\boldsymbol{s} \tag{4.2}$$

holds. Here $\boldsymbol{n}$ is the outward pointing unit normal field of the boundary $\partial\Omega$.

Proof. See, e.g., Matthews (1998). □

Definition 4.1.1. *For $p \in [1, \infty)$ we denote by $L^p(\Omega)$ the Lebesgue space of all functions u measurable on Ω such that $\int_\Omega |u(x)|^p \,\mathrm{d}x < +\infty$. The space $L^p(\Omega)$ is equipped with the norm*

$$\|u\|_{0,p,\Omega} \;=\; \left(\int_\Omega |u(x)|^p \,\mathrm{d}x\right)^{1/p}. \tag{4.3}$$

Further, the space $L^\infty(\Omega)$ consists of such measurable functions on Ω for which the norm

$$\|u\|_{\infty,\Omega} \;=\; \operatorname{esssup}_\Omega |u| \;=\; \inf\left\{\sup_{x\in\Omega\setminus Z} |u(x)|;\; Z \subset \Omega,\, \operatorname{meas}(Z) = 0\right\} \tag{4.4}$$

is finite. The space $L^2(\Omega)$ is a Hilbert space with the inner product

$$(u, v)_\Omega \;=\; \int_\Omega u(x)v(x)\,\mathrm{d}x. \tag{4.5}$$

For $k \in \mathbb{N} \cup \{0\}$ and $p \in [1, \infty]$ we define the Sobolev space $W^{k,p}(\Omega)$ as the space of all functions from $L^p(\Omega)$ whose distributional derivatives $D^\alpha u$, up to order k, also belong to $L^p(\Omega)$, i.e.,

$$W^{k,p} \;=\; \{u \in L^p(\Omega);\, D^\alpha u \in L^p(\Omega)\, \forall\alpha :\, |\alpha| \leq k\}. \tag{4.6}$$

The Sobolev space $W^{k,p}(\Omega)$ is equipped with the norm

$$\|u\|_{k,p,\Omega} \;=\; \left(\sum_{|\alpha|\leq k} \|D^\alpha u\|^p_{0,p,\Omega}\right)^{1/p} \qquad \text{for } p \in [1, \infty), \tag{4.7}$$

$$\|u\|_{k,\infty,\Omega} \;=\; \max_{|\alpha|\leq k}\Big\{\|D^\alpha u\|_{\infty,\Omega}\Big\} \qquad \text{for } p = \infty, \tag{4.8}$$

and the seminorm

$$|u|_{k,p,\Omega} = \left(\sum_{|\alpha|=k} \|D^\alpha u\|_{0,p,\Omega}^p \right)^{1/p} \qquad \text{for } p \in [1,\infty), \tag{4.9}$$

$$|u|_{k,\infty,\Omega} = \max_{|\alpha|=k} \left\{ \|D^\alpha u\|_{\infty,\Omega} \right\} \qquad \text{for } p = \infty. \tag{4.10}$$

Further, we denote $H^k(\Omega) = W^{k,2}(\Omega)$ *and if* $k \neq \infty$ *we put* $\|u\|_{k,\Omega} = \|u\|_{k,2,\Omega}$ *and* $|u|_{k,\Omega} = |u|_{k,2,\Omega}$. *For vector-valued functions* $\boldsymbol{v} = (v_1, v_2, \ldots, v_n) \in W^{k,p}(\Omega)^n$ *we put*

$$\|\boldsymbol{v}\|_{k,p,\Omega} = \left(\sum_{i=1}^n \|v_i\|_{k,p,\Omega}^2 \right)^{1/2} \quad \text{and} \quad |\boldsymbol{v}|_{k,p,\Omega} = \left(\sum_{i=1}^n |v_i|_{k,p,\Omega}^2 \right)^{1/2}. \tag{4.11}$$

When there is no misunderstanding we also use $\|b\|_\infty = \|b\|_{\infty,\Omega} = \|b\|_{0,\infty,\Omega}$. *For more details about function spaces, see, e.g., Kufner et al. (1977) or Rudin (1987).*

Theorem 4.1.2 (Lax-Milgram)**.** *Let* V *be a Hilbert space with the norm* $\|\cdot\|$, *let* $f : V \to \mathbb{R}$ *be a continuous linear functional* V *and let* $a : V \times V \to \mathbb{R}$ *be a bilinear form on* $V \times V$ *that is coercive, i.e., there exists a constant* $\alpha > 0$ *such that*

$$a(v,v) \geq \alpha \|v\|^2 \qquad \forall\, v \in V, \tag{4.12}$$

and continuous (bounded), i.e. there exists a constant $M > 0$ *such that*

$$|a(u,v)| \leq M\|u\|\|v\| \qquad \forall\, u, v \in V. \tag{4.13}$$

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

$$a(u_0, v) = f(v) \qquad \forall\, v \in V. \tag{4.14}$$

Proof. See, e.g., (Ciarlet, 1978, Theorem 1.1.3). □

Theorem 4.1.3 (Friedrichs' inequality)**.** *Let* $\Omega \subset \mathbb{R}^n$ *be a bounded domain with Lipschitz continuous boundary* Γ, *then there exist positive constants* C_F *and* D_F *depending on* Ω *and* n *such that for all* $v \in H^1(\Omega)$ *holds*

$$\|v\|_{0,\Omega} \leq C_F |v|_{1,\Omega} + D_F \|v\|_{0,\Gamma}. \tag{4.15}$$

In particular, when $n = 1$ *then for all* $v \in H^1(I)$, $I = (a,b)$, *holds*

$$\|v\|_{0,I}^2 \leq \left(\frac{2(b-a)}{\pi} \right)^2 |v|_{1,I}^2 + \frac{2(b-a)}{\pi} \left(v^2(a) + v^2(b) \right). \tag{4.16}$$

For $v \in H_0^1(I)$ *one can derive sharper estimate*

$$\|v\|_{0,I}^2 \leq \frac{(b-a)^2}{\pi^2} |v|_{1,I}^2. \tag{4.17}$$

Proof. See, for instance, Rektorys (1999). □

Theorem 4.1.4 (Cauchy-Schwarz-Bunyakovsky inequality)**.** *Let $(V, \langle\cdot,\cdot\rangle)$ be an inner product space. Then for each $u, v \in V$ it holds*

$$|\langle u, v\rangle| \;\le\; \langle u, u\rangle \, \langle v, v\rangle. \tag{4.18}$$

The equality occurs if and only if there exists $\alpha \in \mathbb{R}$ such that $u = \alpha v$ or $v = \alpha u$.

Proof. See, e.g., (Garling, 2007, Proposition 2.3.1). □

Theorem 4.1.5 (M-criterion)**.** *Let the matrix $\mathbb{A}$ satisfies $a_{ij} \le 0$ for $i \ne j$. Then $\mathbb{A}$ is an M-matrix if and only if there exists a vector $\boldsymbol{e} > 0$ such that $\mathbb{A}\boldsymbol{e} > 0$. Furthermore, we have*

$$\|\mathbb{A}^{-1}\|_{\infty,d} \;\le\; \frac{\|\boldsymbol{e}\|_{\infty,d}}{\min_k(\mathbb{A}\boldsymbol{e})_k}. \tag{4.19}$$

Proof. See Axelsson and Kolotilina (1990). □

Theorem 4.1.6 (Comparison principle)**.** *Let $w \in \mathcal{C}^2(\Omega) \cap \mathcal{C}(\overline{\Omega})$ and L is defined by (1.50) (page 14). If there holds*

$$Lw \ge 0 \quad \text{in } \Omega \quad \text{and} \quad w \ge 0 \quad \text{on } \partial\Omega, \tag{4.20}$$

then $w \ge 0$ in $\overline{\Omega}$.

Proof. See, e.g., (Gilbarg and Trudinger, 2001, Theorem 3.3). □

Definition 4.1.2 (Inverse-monotone matrix)**.** *A matrix $\mathbb{A}$ is called* inverse-monotone *if $\mathbb{A}^{-1}$ exists and $\mathbb{A}^{-1} \ge 0$.*

Theorem 4.1.7 (Discrete comparison principle)**.** *Let $\mathbb{A}$ be an inverse-monotone matrix. Then $\mathbb{A}\boldsymbol{v} \le \mathbb{A}\boldsymbol{w}$ implies $\boldsymbol{v} \le \boldsymbol{w}$.*

Proof. If $\mathbb{A}(\boldsymbol{w} - \boldsymbol{v}) \ge \boldsymbol{0}$, then using $\mathbb{A}^{-1} \ge 0$ implies

$$\boldsymbol{w} - \boldsymbol{v} \;=\; \mathbb{A}^{-1}\Big[\mathbb{A}(\boldsymbol{w} - \boldsymbol{v})\Big] \;\ge\; \boldsymbol{0}. \tag{4.21}$$

□

4.2 Finite-element theory

Let us recall some basic theorems from the finite-element theory. For details see Ciarlet (1978).

Definition 4.2.1. *We say that two open subsets Q and $\widehat{Q}$ of $\mathbb{R}^n$ are* affine-equivalent *if there exists an invertible affine mapping $F : \mathbb{R}^n \to \mathbb{R}^n$, $F(\widehat{x}) = \mathbb{B}\widehat{x} + \boldsymbol{r}$, such that $Q = F(\widehat{Q})$.*

Definition 4.2.2 (X_h-interpolant)**.** *Let there be given a finite element space X_h with a set of degrees of freedom (functionals) $\Sigma_h = \{\phi_{j,h},\, 1 \le j \le M\}$ and the basis functions w_j of X_h satisfying $\phi_{i,h}(w_j) = \delta_{ij}$ for all $1 \le i, j \le M$. Then with any function $v : \overline{Q} \to \mathbb{R}$ sufficiently smooth so that the degrees of freedom $\phi_{j,h}$, $1 \le j \le M$, are well defined, we associate the function*

$$\Pi_h v \;=\; \sum_{j=1}^{M} \phi_{j,h}(v) w_j. \tag{4.22}$$

*The function $\Pi_h v$ is called the X_h-*interpolant *of the function v.*

Assumption 4.2.1 (Assumptions on $\mathcal{T}_h$)**.** *In the finite elements framework we consider the following assumptions*

($\mathcal{H}1$) We consider a regular family of triangulations $\mathcal{T}_h$ *in the following sense:*

(a) The system of triangulations $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ *is* shape-regular*, i.e. there exists a constant* $\sigma > 0$ *such that for all* $K \in \mathcal{T}_h, h \in (0, h_0)$*, it holds*

$$\frac{h_K}{\rho_K} \leq \sigma. \tag{4.23}$$

(b) The quantity $h = \max_{K\in\mathcal{T}_h} h_K$ *approaches zero.*

($\mathcal{H}2$) The family $(K, P_K, \Sigma_K), K \in \mathcal{T}_h$*,* $h \in (0, h_0)$*, is an* affine *family of finite elements (used for the construction of* X_h*), i.e. all the finite elements* (K, P_K, Σ_K)*,* $K \in \mathcal{T}_h$*,* $h \in (0, h_0)$*, are affine-equivalent to a single element* $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$*.*

($\mathcal{H}3$) All the finite elements (K, P_K, Σ_K)*,* $K \in \mathcal{T}_h$*,* $h \in (0, h_0)$*, are of class* $\mathcal{C}^0$*, i.e.* $X_h \subset \mathcal{C}^0(\overline{\Omega})$ *for all* $h \in (0, h_0)$*.*

Theorem 4.2.1 (X_h-interpolation)**.** *In addition to the assumptions* ($\mathcal{H}1$)*,* ($\mathcal{H}2$) *and* ($\mathcal{H}3$) *let there exists integers* $k \geq l \geq 0$*, such that the following inclusions are for each* $\widehat{K} \in \mathcal{T}_h$ *satisfied*

$$P_k(\widehat{K}) \subset P(\widehat{K}) \subset H^l(\widehat{K}), \tag{4.24}$$

$$H^{k+1}(\widehat{K}) \hookrightarrow \mathcal{C}^s(\widehat{K}), \tag{4.25}$$

where s *is the maximal order of partial derivatives occurring in the definitions of the set* $\widehat{\Sigma}$*.*

Then there exists a constant C_X *independent of* h *such that, for any function* $v \in H^{k+1}(\Omega) \cap V$ *there holds*

$$\|v - \Pi_h v\|_{m,\Omega} \leq C_X h^{k+1-m} |v|_{k+1,\Omega}, \quad \text{for } 0 \leq m \leq \min\{1, l\}, \tag{4.26}$$

$$\left(\sum_{K\in\mathcal{T}_h} \|v - \Pi_h v\|^2_{m,K} \right)^{1/2} \leq C_X h^{k+1-m} |v|_{k+1,\Omega}, \quad \text{for } 2 \leq m \leq \min\{k+1, l\}, \tag{4.27}$$

where $\Pi_h v \in V_h$ *is the* X_h*-interpolant of the function* v*.*

Proof. See (Ciarlet, 1978, Theorem 3.2.1). □

Theorem 4.2.2 (Inverse inequality)**.** *Let the shape-regularity assumption (4.23) be valid and let there be given two pairs* (l, r) *and* (m, q) *with integers* $m \geq l \geq 0$ *and real numbers* $r, q \in [1, \infty]$ *such that* $P(K) \subset W^{l,r}(K) \cap W^{m,q}(K)$*. Then there exists a constant* $C_{inv} = C_{inv}(\sigma, l, r, m, q)$ *such that*

$$|v_h|_{m,q,K} \leq \frac{C_{inv}}{h_K^{m-l+n\max\{0,1/r-1/q\}}} |v_h|_{l,r,K}, \qquad \text{for all } v_h \in P(K). \tag{4.28}$$

When $q = \infty$ *or* $r = \infty$ *we set* $1/\infty := 0$*.*

Proof. See (Ciarlet, 1978, Theorem 3.2.6). □

Remark 4.2.1. The inverse inequality can be also formulated without the shape-regularity assumption (4.23) in the form:

For any dimension $n > 1$ and polynomial degree $k \in \mathbb{N}$ there exists a constant $C_{n,k}$ independent of h_K, v and K such that

$$|v|_{1,K} \;\leq\; C_{n,k}\frac{|\partial K|}{|K|}\,\|v\|_{0,K} \qquad \text{for all } v \in P_k(K), k \in \mathbb{N}. \tag{4.29}$$

Using this inequality we can easily compute the constant C_{inv}. For instance, in 2D there holds $\frac{|\partial K|}{|K|} = \frac{2}{\rho_K}$ which results in the estimates

$$\frac{4\sqrt{3}}{h_K} \;\leq\; \frac{|\partial K|}{|K|} \;\leq\; \frac{2\sigma}{h_K}. \tag{4.30}$$

It means that in 2D it is necessary $\sigma \geq 2\sqrt{3}$ and for C_{inv} holds $C_{inv} = 2\sigma C_{2,k}$. Following table provides several optimal values of the constants $C_{n,k}$ (c.f. Ozısık et al. (2010)).

n	2				3			
k	1	2	3	4	1	2	3	4
$C_{n,k}$	$\sqrt{6}$	$\sqrt{45/2}$	$\approx$7.542	$\approx$10.946	$\sqrt{40}$	$\sqrt{126}$	$\approx$17.175	$\approx$24.365

Table 4.1: Several optimal values of the constants $C_{n,k}$ for $n = 2$ and 3.

Denotation 4.2.1 (Orthogonal L^2-projection)**.** *Let $r \geq 0$ be an integer, then for each $K \in \mathcal{T}_h$ and for each $\varphi \in L^2(K)$ we can construct the polynomial approximation $\pi_{K,r}\varphi \in P_r(K)$ of the function φ satisfying*

$$(\pi_{K,r}\varphi - \varphi, v)_K \;=\; 0 \qquad \forall v \in P_r(K). \tag{4.31}$$

The function $\pi_{K,r}\varphi$ is uniquely defined and we called the mapping $\pi_{K,r} : L^2(K) \to P_r(K)$ the orthogonal L^2-projection *onto the space $P_r(K)$.*

For each function $\boldsymbol{\psi} \in L^2(K)^n$ we also define a mapping $\boldsymbol{\pi}_{K,r} : L^2(K)^n \to P_r(K)^n$ by the relation

$$[\boldsymbol{\pi}_{K,r}\boldsymbol{\psi}]_i \;=\; \pi_{K,r}\psi_i, \qquad \text{for each } i = 1, 2, \ldots, n. \tag{4.32}$$

Theorem 4.2.3 (Approximation property)**.** *When the shape-regularity assumption (4.23) is valid then there exists a constant $C_\Pi > 0$ such that for all $v \in W^{s,p}(K)$, $K \in \mathcal{T}_h$, there holds*

$$|\pi_{K,r}v - v|_{m,p,K} \;\leq\; C_\Pi h_K^{\mu-m}|v|_{\mu,p,K}, \tag{4.33}$$

where $p \in [1, \infty]$ and $0 \leq m \leq \mu = \min\{r+1, s\}$.

Proof. See, for instance, Dolejší and Feistauer (2015). □

Corollary 4.2.1. If the shape-regularity assumption (4.23) is valid then for all $\boldsymbol{\psi} \in W^{s,p}(K)^n$, $K \in \mathcal{T}_h$, there holds

$$|\boldsymbol{\pi}_{K,r}\boldsymbol{\psi} - \boldsymbol{\psi}|_{m,p,K} \;\leq\; C_\Pi h_K^{\mu-m}|\boldsymbol{\psi}|_{\mu,p,K}, \tag{4.34}$$

where $p \in [1, \infty]$ and $0 \leq m \leq \mu = \min\{r+1, s\}$.

Proof. From Theorem 4.2.3 and the definition of $\boldsymbol{\pi}_{K,r}$ it follows

$$\begin{aligned} |\boldsymbol{\pi}_{K,r}\boldsymbol{\psi} - \boldsymbol{\psi}|_{m,p,K} &= \left(\sum_{i=1}^{n} |\pi_{K,r}\psi_i - \psi_i|^2_{m,p,K}\right)^{1/2} \leq \\ &\leq C_\Pi h_K^{\mu-m} \left(\sum_{i=1}^{n} |\psi_i|^2_{\mu,p,K}\right)^{1/2} = C_\Pi h_K^{\mu-m} |\boldsymbol{\psi}|_{\mu,p,K}. \end{aligned} \tag{4.35}$$

□

Lemma 4.2.1. *Let $n \in \mathbb{N}$ and let $K \subset \mathbb{R}^n$ be a simplex with nodes $P_{K,i}$, $i = 1, 2, \ldots, n+1$. Then every $v_h \in P_1(K)$ satisfies*

$$\frac{|K|}{(n+1)(n+2)} \sum_{i=1}^{n+1} v_h^2(P_{K,i}) \leq \|v_h\|^2_{0,K} \leq \frac{|K|}{n+1} \sum_{i=1}^{n+1} v_h^2(P_{K,i}). \tag{4.36}$$

Proof. Let us denote $\boldsymbol{v} = (v_h(P_{K,1}), v_h(P_{K,2}), \ldots, v_h(P_{K,n+1}))^T$ and let $\mathbb{A}$ be a matrix satisfying $a_{ii} = 2$, for $i = 1, 2, \ldots, n+1$, and $a_{ij} = 1$ for $i \neq j$, $i, j = 1, 2, \ldots, n+1$. Then

$$\begin{aligned} \|v_h\|^2_{0,K} &= \int_K \left(\sum_{i=1}^{n+1} v_h(P_{K,i})\lambda_{K,i}\right)^2 \mathrm{d}\boldsymbol{x} = \\ &= \frac{2|K|}{(n+1)(n+2)} \left\{\sum_{i=1}^{n+1} v_h^2(P_{K,i}) + \sum_{1\leq i<j\leq n+1} v_h(P_{K,i})v_h(P_{K,j})\right\} = \\ &= \frac{|K|}{(n+1)(n+2)} \boldsymbol{v}^T \mathbb{A}\boldsymbol{v}. \end{aligned} \tag{4.37}$$

Thus, it remains to determine the eigenvalues of $\mathbb{A}$. Since the characteristic polynomial of the matrix $\mathbb{A}$ is $\det(\mathbb{A} - \lambda\mathbb{I}) = (n+2-\lambda)(1-\lambda)^n$, we get $|\boldsymbol{v}|^2 \leq \boldsymbol{v}^T\mathbb{A}\boldsymbol{v} \leq (n+2)|\boldsymbol{v}|^2$ which completes the proof. □

Lemma 4.2.2. *Let $n \in \mathbb{N}$ and let a_i, $i = 1, 2, \ldots, n$, be arbitrary real numbers. Then*

$$\left(\sum_{i=1}^{n} a_i\right)^2 \leq n \sum_{i=1}^{n} a_i^2. \tag{4.38}$$

Proof. From the Cauchy-Schwarz-Bunyakovsky inequality (Theorem 4.1.4, page 126) it follows that

$$\left(\sum_{i=1}^{n} a_i\right)^2 = \left(\sum_{i=1}^{n} 1 \cdot a_i\right)^2 \leq \left(\sum_{i=1}^{n} 1^2\right)\left(\sum_{i=1}^{n} a_i^2\right) = n\sum_{i=1}^{n} a_i^2. \tag{4.39}$$

□

Corollary 4.2.2. For any $n \in \mathbb{N}$ and $s_K^{(i)} \geq 0$, $i = 1, 2, \ldots, n$, $K \in \mathcal{T}_h$, it holds

$$\sum_{i=1}^{n} \left(\sum_{K\in\mathcal{T}_h} s_K^{(i)}\right)^{1/2} \leq \left(n\sum_{i=1}^{n}\sum_{K\in\mathcal{T}_h} s_K^{(i)}\right)^{1/2} = \left(n\sum_{K\in\mathcal{T}_h}\sum_{i=1}^{n} s_K^{(i)}\right)^{1/2}. \tag{4.40}$$

Proof. It suffices to take $a_i = \left(\sum_{K\in\mathcal{T}_h} s_K^{(i)}\right)^{1/2}$ in Lemma 4.2.2. □

Conclusion

In the first part of this thesis we were concerned with the construction of the asymptotic expansion of singularly perturbed convection-diffusion equations. We adjusted approaches and techniques derived for one-dimensional problems and applied them to the two-dimensional case. Additional corner correction terms had to be added to the sum of the standard layer functions. Consequently, we proved the asymptotic behavior of this structure and derived an exact formula for the zeroth-order matched asymptotic expansion in the two-dimensional domain containing exponential boundary layers and with inner angles of the form π/m, $m \in \mathbb{N}$, $m \geq 2$. Finally, we verified our theoretical results by experiments.

The second part of the thesis was devoted to a brief overview of several stabilizing techniques. We demonstrated their behavior and mutual interconnection on a set of examples. We showed that for constant data almost all of them are equivalent. Several observations were later employed in the rest of the thesis. We concluded this part of the thesis with the proof of the uniform convergence of the Il'in-Allen-Southwell scheme in 1D. We also showed how the constants appearing in this proof depend on problem parameters.

In the third and most important part of this thesis we presented a modification of the classical SUPG finite element method for solving singularly perturbed problems. This modification is based on the observation that when convection dominates the value of the solution at any single point depends only on the values at nodes laying on the same streamline in the upwind direction. Therefore, one should construct the triangulation of the computational domain in such a way that this property holds for the discrete solution as well.

Further, we showed that once we have the mesh oriented along streamlines and the divergence of the given vector field $\boldsymbol{b}$ is non-positive, we can discretize $\boldsymbol{b}$ and add stabilizing terms so that the problem bilinear form is coercive and the method satisfies the discrete maximum principle. Moreover, we were able to derive the a priori error estimates in the SUPG-like energy norms. We also presented the a priori error analysis of the SUPG method itself.

In the remaining sections of the thesis we introduced several modifications of the new method. We used knowledges acquired in the second part of the thesis and proposed several stabilizing terms that can improve the L^∞-convergence at layers. Finally, we demonstrated how to extend the new method to higher order finite elements and carried out several numerical experiments on heuristically constructed meshes. Both — linear and quadratic — finite elements provided satisfactory results and computational errors confirmed our expectations.

To be able to use this method in the future one should firstly design a suitable mesh-generator cooperating with the method. The modifications of the method could be improved as well. The extension to higher dimensions and higher order finite elements is not yet fully resolved and a derivation of further corner expansions may lead to the construction of the uniformly convergent scheme (with respect to ε) in 2D, or even 3D.

Bibliography

M. Ainsworth and W. Dörfler. Fundamental systems of numerical schemes for linear convection-diffusion equations and their relationship to accuracy. *Computing*, 66(2):199–229, 2001. ISSN 0010-485X.

O. Axelsson and L. Kolotilina. Monotonicity and discretization error estimates. *SIAM Journal on Numerical Analysis*, 27(6):1591–1611, 1990. ISSN 0036-1429.

O. Axelsson, E. Glushkov, and N. Glushkova. The local Green's function method in singularly perturbed convection-diffusion problems. *Mathematics of Computation*, 78(265):153–170, 2009. ISSN 0025-5718.

F. Brezzi and A. Russo. Choosing bubbles for advection-diffusion problems. *Mathematical Models and Methods in Applied Sciences*, 04(04):571–587, 1994.

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. *Computer Methods in Applied Mechanics and Engineering*, 32(1):199 – 259, 1982. ISSN 0045-7825.

I Christie, D.F. Griffiths, A.R. Mitchell, and O.C. Zienkiewicz. Finite-element methods for 2nd order differential equations with significant 1st derivatives. *International Journal for Numerical Methods in Engineering*, 10(6):1389–1396, 1976. ISSN 0029-5981.

P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Company, Amsterdam, New York, Oxford, 1978. ISBN 0-444-85028-7, 978-0-444-85028-7.

R. Codina. A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation. *Computer Methods in Applied Mechanics and Engineering*, 110(3):325 – 342, 1993. ISSN 0045-7825.

A. Devinatz, R. Ellis, and A. Friedman. The asymptotic behavior of the first real eigenvalue of second order elliptic operators with a small parameter in the highest derivatives, ii. *Indiana University Mathematics Journal*, 23(11): 991–1011, 1974. ISSN 0022-2518, 1943-5258.

V. Dolejší and M. Feistauer. *Discontinuous Galerkin Method: Analysis and Applications to Compressible Flow*, volume 48 of *Springer Series in Computational Mathematics*. Springer Publishing Company, Incorporated, 1st edition, 2015. ISBN 3-319-19266-3, 978-3-319-19266-6.

E. G. Dutra do Carmo and A. C. Galeão. Feedback Petrov-Galerkin methods for convection-dominated problems. *Computer Methods in Applied Mechanics and Engineering*, 88(1):1–16, 1991. ISSN 0045-7825.

W. Eckhaus. *Asymptotic Analysis of Singular Perturbations*. Studies in Logic and the Foundations of Mathematics. North-Holland Publishing Company, 1979. ISBN 978-0-444-85306-6.

L. P. Franca, S. L. Frey, and T. J. R. Hughes. Stabilized finite element methods. I. Application to the advective-diffusive model. *Computer Methods in Applied Mechanics and Engineering*, 95(2):253–276, 1992. ISSN 0045-7825.

D. J. H. Garling. *Inequalities: A journey into linear analysis*. Cambridge University Press, 1st edition, 2007. ISBN 0-521-69973-8, 978-0-521-69973-0, 0-521-87624-9, 978-0-521-87624-7.

E. C. Gartland. Uniform high-order difference schemes for a singularly perturbed two-point boundary value problem. *Mathematics of Computation*, 48(178): 551+, Apr 1987. ISSN 0025-5718.

D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*. Classics in mathematics. Springer, 2nd ed., rev. 3rd printing edition, 2001. ISBN 3-540-41160-7, 978-3-540-41160-4.

H. Goering, A. Felgenhauer, and G. Lube. *Singularly Perturbed Differential Equations*. Mathematical Research. Akademie-Verlag, Berlin, 1983.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 4th edition, 2012. ISBN 1-4214-0794-9, 978-1-4214-0794-4.

J. C. Heinrich, P. S. Huyakorn, O. C. Zienkiewicz, and A. R. Mitchell. An 'upwind' finite element scheme for two-dimensional convective transport equation. *International Journal for Numerical Methods in Engineering*, 11(1):131–143, 1977. ISSN 0029-5981.

W. Hermanns and A. Einstein. *Einstein and the Poet: In Search of the Cosmic Man*. Branden Press, 1983. ISBN 978-0-82831873-0.

T. J. R. Hughes. A simple scheme for developing 'upwind' finite elements. *International Journal for Numerical Methods in Engineering*, 12(9):1359–1365, 1978. ISSN 1097-0207.

T. J. R. Hughes, L. P. Franca, and G. M. Hulbert. A new finite element formulation for computational fluid dynamics. VIII. The Galerkin/least-squares method for advective-diffusive equations. *Computer Methods in Applied Mechanics and Engineering*, 73(2):173–189, 1989. ISSN 0045-7825.

V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection-diffusion equations. I. A review. *Computer Methods in Applied Mechanics and Engineering*, 196(17-20):2197–2215, 2007. ISSN 0045-7825.

V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection-diffusion equations. II. Analysis for P_1 and Q_1 finite elements. *Computer Methods in Applied Mechanics and Engineering*, 197(21-24):1997–2014, 2008. ISSN 0045-7825.

R. B. Kellogg and A. Tsan. Analysis of some difference approximations for a singular perturbation problem without turning points. *Mathematics of Computation*, 32(144):1025–1039, 1978. ISSN 0025-5718, 1088-6842.

P. Knobloch. Improvements of the Mizukami-Hughes method for convection-diffusion equations. *Computer Methods in Applied Mechanics and Engineering*, 196(1-3):579–594, 2006. ISSN 0045-7825.

A. Kufner, O. John, and S. Fučík. *Function Spaces*. Mechanics: Analysis. Springer, 1 edition, 1977. ISBN 90-286-0015-9, 978-90-286-0015-7.

J. Lamač. Asymptotic expansion of the solution of the singularly perturbed convection-diffusion equation in the 2d convex polygonal domain. *AIP Conference Proceedings*, 1558(1):383–386, 2013.

J. Lamač. Modified SUPG method on oriented meshes. In Petr Knobloch, editor, *Boundary and Interior Layers, Computational and Asymptotic Methods - BAIL 2014*, pages 121–133, Cham, 2015. Springer International Publishing. ISBN 978-3-319-25727-3.

T. Linß and M. Stynes. Asymptotic analysis and Shishkin-type decomposition for an elliptic convection-diffusion problem. *Journal of Mathematical Analysis and Applications*, 261(2):604 – 632, 2001. ISSN 0022-247X.

J. L. López, E. Pérez Sinusía, and N. M. Temme. Asymptotic behaviour of three-dimensional singularly perturbed convection-diffusion problems with discontinuous data. *Journal of Mathematical Analysis and Applications*, 328(2): 931–945, Apr 15 2007. ISSN 0022-247X.

G. I. Marchuk. *Methods of Numerical Mathematics*, volume 2. Springer-Verlag New York, 2nd edition, 1982. ISBN 978-1-4613-8152-5.

P. C. Matthews. *Vector Calculus*. Springer Undergraduate Mathematics Series. Springer, 1 edition, 1998. ISBN 3-540-76180-2, 978-3-540-76180-8, 978-1-4471-0597-8.

G. Matthies, P. Skrzypacz, and L. Tobiska. A unified convergence analysis for local projection stabilisations applied to the Oseen problem. *ESAIM: M2AN Mathematical Modelling and Numerical Analysis*, 41(4):713–742, 2007. ISSN 0764-583X.

A. Mizukami and T. J. R. Hughes. A Petrov-Galerkin finite element method for convection-dominated flows: An accurate upwinding technique for satisfying the maximum principle. *Computer Methods in Applied Mechanics and Engineering*, 50(2):181 – 193, 1985. ISSN 0045-7825.

S. Ozısık, B. Riviére, and T. Warburton. On the Constants in Inverse Inequalities in L2. Technical Report TR10-19, Rice University, Houston, TX, USA, 2010.

G. Payre, M. de Broissia, and J. Bazinet. An 'upwind' finite element method via numerical intergration. *International Journal for Numerical Methods in Engineering*, 18(3):381–396, March 1982. ISSN 1097-0207.

K. Rektorys. *Variační metody v inženýrských problémech a v problémech matematické fyziky*. Česká matice technická. Academia, Prague, 1999. ISBN 978-80-200-0714-8.

B. Riviére. *Discontinuous Galerkin Methods For Solving Elliptic And Parabolic Equations: Theory and Implementation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008. ISBN 0-898716-56-X, 978-0-898716-56-6.

H.-G. Roos. Ten ways to generate the Il'in and related schemes. *Journal of Computational and Applied Mathematics*, 53(1):43–59, 1994. ISSN 0377-0427.

H.-G. Roos, M. Stynes, and L. Tobiska. *Robust numerical methods for singularly perturbed differential equations: convection-diffusion-reaction and flow problems*, volume 24. Springer-Verlag Berlin Heidelberg, 2008. ISBN 978-3-540-34466-7.

W. Rudin. *Real and complex analysis*. Mathematics series. McGraw-Hill Book Company, Singapore, 3rd edition, 1987. ISBN 978-0-07-054234-1.

A. Russo. Streamline-upwind Petrov/Galerkin method (SUPG) vs residual-free bubbles (RFB). *Computer Methods in Applied Mechanics and Engineering*, 195 (13-16):1608–1620, 2006. ISSN 0045-7825.

G. I. Shishkin. On finite difference fitted schemes for singularly perturbed boundary value problems with a parabolic boundary layer. *Journal of Mathematical Analysis and Applications*, 208(1):181–204, 1997. ISSN 0022-247X.

K. Surla and M. Stojanović. Solving singularly perturbed boundary-value problems by spline in tension. *Journal of Computational and Applied Mathematics*, 24(3):355–363, 1988. ISSN 0377-0427.

R. L. Wheeden and A. Zygmund. *Measure and integral: an introduction to real analysis*, volume 308 of *Monographs and textbooks in pure and applied mathematics*. CRC Press, Boca Raton, FL, second edition edition, 2015. ISBN 978-1-4987-0290-4, 1-4987-0290-2.

List of Figures

List of Tables