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


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

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>List of Symbols</strong></td>
<td>3</td>
</tr>
<tr>
<td><strong>Introduction</strong></td>
<td>6</td>
</tr>
<tr>
<td><strong>1 Asymptotic expansion</strong></td>
<td>8</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>8</td>
</tr>
<tr>
<td>1.2 Asymptotic expansion in 1D</td>
<td>8</td>
</tr>
<tr>
<td>1.3 Asymptotic expansion in two dimensions</td>
<td>14</td>
</tr>
<tr>
<td>1.3.1 Model equation and reduced problem</td>
<td>14</td>
</tr>
<tr>
<td>1.3.2 Local expansion in exponential layers</td>
<td>15</td>
</tr>
<tr>
<td>1.3.3 Corner correction</td>
<td>19</td>
</tr>
<tr>
<td>1.3.4 Parabolic boundary layers</td>
<td>27</td>
</tr>
<tr>
<td>1.4 Numerical experiments</td>
<td>31</td>
</tr>
<tr>
<td><strong>2 Stabilization and Upwind techniques</strong></td>
<td>34</td>
</tr>
<tr>
<td>2.1 Stabilization in 1D</td>
<td>34</td>
</tr>
<tr>
<td>2.1.1 Spurious oscillations</td>
<td>35</td>
</tr>
<tr>
<td>2.1.2 SUPG method</td>
<td>35</td>
</tr>
<tr>
<td>2.1.3 Changing test functions</td>
<td>37</td>
</tr>
<tr>
<td>2.1.4 Adding artificial diffusion</td>
<td>38</td>
</tr>
<tr>
<td>2.1.5 Adding bubble functions</td>
<td>39</td>
</tr>
<tr>
<td>2.1.6 Local Green’s function method</td>
<td>41</td>
</tr>
<tr>
<td>2.1.7 Exponentially fitted schemes</td>
<td>43</td>
</tr>
<tr>
<td>2.2 Uniform convergence of classical Il’in-Allen-Southwell scheme</td>
<td>44</td>
</tr>
<tr>
<td>2.2.1 Consistency</td>
<td>46</td>
</tr>
<tr>
<td>2.2.2 Stability</td>
<td>50</td>
</tr>
<tr>
<td>2.2.3 Convergence</td>
<td>51</td>
</tr>
<tr>
<td><strong>3 Modified SUPG method on convection-oriented meshes</strong></td>
<td>54</td>
</tr>
<tr>
<td>3.1 Introduction and the idea of the method</td>
<td>54</td>
</tr>
<tr>
<td>3.2 Derivation of the method</td>
<td>55</td>
</tr>
<tr>
<td>3.2.1 Monotonicity</td>
<td>57</td>
</tr>
<tr>
<td>3.3 Mesh properties and notation</td>
<td>59</td>
</tr>
<tr>
<td>3.4 Coercivity</td>
<td>63</td>
</tr>
<tr>
<td>3.4.1 Technical lemmas</td>
<td>64</td>
</tr>
<tr>
<td>3.4.2 Coercivity estimates</td>
<td>68</td>
</tr>
<tr>
<td>3.5 Error analysis</td>
<td>71</td>
</tr>
<tr>
<td>3.5.1 SUPG method error analysis</td>
<td>71</td>
</tr>
<tr>
<td>3.5.2 Error analysis of presented method</td>
<td>74</td>
</tr>
<tr>
<td>3.5.3 M-matrix</td>
<td>83</td>
</tr>
<tr>
<td>3.6 ( L^\infty )-convergence improvement</td>
<td>85</td>
</tr>
<tr>
<td>3.6.1 Constant data</td>
<td>85</td>
</tr>
<tr>
<td>3.6.2 Non-constant data</td>
<td>90</td>
</tr>
<tr>
<td>3.7 Higher order finite elements</td>
<td>97</td>
</tr>
<tr>
<td>3.7.1 Definition and properties of the discretized vector field</td>
<td>98</td>
</tr>
</tbody>
</table>
List of Symbols

$(u,v)_\Omega$ \hspace{2cm} $L^2(\Omega)$-inner product, $(u,v)_\Omega = \int_\Omega u(x)v(x)\,dx$, page 124

$\|\cdot\|_{k,p,\Omega}$ \hspace{2cm} Norm on the space $W^{k,p}(\Omega)$,
$\|u\|_{k,p,\Omega} = \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|_{0,p,\Omega}^p\right)^{1/p}$, $p \neq \infty$, page 124

$|\cdot|_{k,p,\Omega}$ \hspace{2cm} Seminorm on the space $W^{k,p}(\Omega)$,
$|u|_{k,p,\Omega} = \left(\sum_{|\alpha| = k} \|D^\alpha u\|_{0,p,\Omega}^p\right)^{1/p}$, $p \neq \infty$, page 125

$\|\cdot\|_{\infty,d}$ \hspace{2cm} Discrete maximum norm,
$\|v_h\|_{\infty,d} = \max_{1 \leq i \leq N} |v_i|$, page 37

$|||\cdot|||_{b}$ \hspace{2cm} Energy norm,
$|||v|||_{b}^2 = \varepsilon |v|_1^2, 0,\Omega + \frac{\eta}{2} \|v\|_{0,\Omega}^2 + \sum_{K \in T_h} \frac{|d_{K,1}|}{|b_{K}|} \|b_{K} \cdot \nabla v\|_{0,K}^2$, page 68

$|||\cdot|||_{b,\ast}$ \hspace{2cm} Energy norm,
$|||v|||_{b,\ast}^2 = \varepsilon |v|_1^2, 0,\Omega + C_b^2 \sum_{K \in T_h} h_K \|v\|_{0,K}^2 + C_b \sum_{K \in T_h} \frac{|d_{K,1}|}{|b_{K}|} \|b_{K} \cdot \nabla v\|_{0,K}^2$, page 70

$a_1$ \hspace{2cm} Bilinear form, $a_1(u, \varphi) = \varepsilon (u', \varphi')_\Omega + (b u', \varphi)_\Omega$, page 34

$a_h, F_h$ \hspace{2cm} Bilinear form and functional used in the definition of the modified SUPG method, page 57

$a_h^{(2)}, F_h^{(2)}$ \hspace{2cm} Bilinear form and functional used in the definition of the modified SUPG method, using second order finite elements, page 98

$a_{\Gamma}, a_{h}^{\infty}$ \hspace{2cm} Bilinear forms used for the $L^\infty$-convergence improvement, page 86

$b_K$ \hspace{2cm} Piecewise constant approximation of $b$,
$b_K = -\frac{1}{|K|} \int_K b \cdot \nabla \lambda_{K,1} \,d\mathbf{x}$, page 56

$b_K'$ \hspace{2cm} Orthogonal $L^2$-projection of the vector $b$ on a given polynomial space, page 75

$\mathbb{B}_s$ \hspace{2cm} Matrix resulting from the discretization of the convective term, page 104

$\beta_j^s$ \hspace{2cm} Weighted average value of $|b_K|$ on $C_j^s$, $\beta_j^s = \frac{1}{|C_j^s|} \sum_{K \subseteq C_j^s} |b_K| |K|$, page 60

$\beta$ \hspace{2cm} Minimum of weighted average values of $|b_K|$ on $C_j^s$, $\beta = \min_{j,s} \{\beta_j^s\}$, page 70

$C_2^s$ \hspace{2cm} Constant appearing in the definition of $|||\cdot|||_{b,\ast}$, $C_2^s = \frac{(4-\delta)\kappa_\beta}{2L^2 R} (n+1)$, page 70

$C_b^s$ \hspace{2cm} Constant appearing in the definition of $|||\cdot|||_{b,\ast}$, $C_b^s = \frac{4-\delta}{4}$, page 70

$C_{\text{inv}}$ \hspace{2cm} Constant from Theorem 4.2.2 (inverse inequality), page 127
\( C_j^s \) Cluster surrounding the edge \( P_j^a P_j^s \), \( C_j^s = \bigcup_{P_{j-1}^s, P_j^s \subset K} P_j^s \), page 59

\( C_K \) Barycentre of the element \( K \), page 55

\( C_{\Pi} \) Constant from Theorem 1.2.3 (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

\( d_{K,j} \) Oriented edge of the \( n \)-simplex \( K \), \( d_{K,j} = P_{K,n+1} - P_{K,j} \), page 99

\( g_j \) Local Green’s function, page 41

\( \Gamma_+, \Gamma_0, \Gamma_- \) Parts of \( \partial\Omega \) determined by the sign of \( b \cdot n \), page 54

\( h^s_j \) Length of the cluster \( C_j^s \) in the streamline direction, \( h^s_j = |d_{K,1}| \), page 60

\( \kappa \) Mesh structure parameter, \( \kappa = \min_{j,s} \left\{ \frac{|C_j^s|}{|\Omega_{s,j}|} \right\} \), page 68

\( \kappa_{ij}^K, \kappa_{pq}^j \) 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, \( L = \max_{j,s} \{N_s h^s_j\} \), page 70

\( L_h \) Method matrix, \( (L_h)_{ki} = a_h(\lambda_i, \lambda_k) \), page 83

\( \tilde{L}_h \) Scaled method matrix, \( (\tilde{L}_h)_{ki} = \frac{(L_h)_{ki}}{|\text{supp}\{\lambda_k\}|} \), 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

\( \mu, \nu, \mu_K, \nu_K \) Coefficients used for the \( L^\infty \)-convergence improvement, page 86

\( N_s \) Number of edges forming the \( s \)-th discrete streamline, page 60

\( \Omega_j^s \) Patch surrounding the node \( P_j^s \), \( \Omega_j^s = \bigcup_{P_{j-1}^s \subset K} P_{j-1}^s \), page 59

\( \Omega_{0,j}^s \) Complementary set, \( \Omega_{0,j}^s = \Omega_j^s \setminus (C_j^s \cup C_{j+1}^s) \), 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, Pe(x)  Péclet number, in 1D there holds Pe = \( \frac{bh}{2\varepsilon} \) and Pe(x) = \( \frac{b(x)h}{2\varepsilon} \), page 35

PeΓ, PeK  Péclet number, in 2D it is PeΓ = \( \frac{h(b \cdot n_Γ)^2}{2\varepsilon|b|} \) and PeK = \( \frac{|d_{K,i}|(b_K \cdot n_K)^2}{2\varepsilon|b_K|} \), in the SUPG method there holds PeK = \( \frac{|b||n_{a,m,k}h_K|}{2\varepsilon} \), page 87

\( \Pi_{b,K}^{(2)} \)  Linear mapping satisfying \( \left( b_K^{(1)} \cdot \nabla \varphi_{K,i}^{(2)} \right) \Pi_{b,K}^{(2)} \left( \varphi_{K,j}^{(2)} \right)_K = 0 \) for all \( i, j = 1, 2, \ldots, 6 \), page 101

\( \varphi_{K,i}^{(2)} \)  Basis functions of \( P_2(K) \) satisfying \( \varphi_{K,i}^{(2)} \left( P_{K,j}^{(2)} \right)_K = \delta_{ij} \) for all \( i, j \in \{1, 2, \ldots, 6\} \), page 98

\( q^s_j \)  Flux through \( C_s^j \), \( q^s_j = -\sum_{K \subset C_s^j} \int_K b \cdot \nabla \lambda_K d\mathbf{x} \), page 60

\( R \)  Mesh structure parameter, \( R = \max_{j,s} \left\{ \max_{K \subset \Omega_s} \frac{h_K}{h_s} \right\} \), page 70

\( \mathcal{R}_{\Gamma}, \mathcal{R}_K \)  Mesh parameters, \( \mathcal{R}_{\Gamma} = \text{Pe}_\Gamma \left( \coth(\text{Pe}_\Gamma) - 1 \right) - 1 \), page 88

\( R_{K}^{(2)} \)  Linear mapping enabling the fulfillment of the discrete maximum principle, page 102

\( S + E \)  S-decomposition of the function \( u = S + E \), page 13

\( \sigma, \nu_K \)  Quantities used for proving coercivity of the bilinear form \( a_h \), page 69

\( \tau, \tau_j, \tau_{\text{upw}}, \tau_s \)  Stabilization parameters used in the SUPG method, page 35

\( \mathcal{T}_h \)  Triangulation of the domain \( \Omega \), page 55

\( \theta_K \)  Mesh parameters,
\[
\theta_K = \frac{1}{|K|} \max \left\{ \max_{2 \leq i \leq n} \left| \int_K b \cdot \nabla \lambda_{K,i} d\mathbf{x} \right|, \left| \sum_{i=2}^{n} \int_K b \cdot \nabla \lambda_{K,i} d\mathbf{x} \right| \right\}, \]
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

\( u_{loc,m}, V_{loc}^{k,m} \)  \( m \)-th order local expansion, page 10

\( W^{k,p}(\Omega) \)  Sobolev space of functions \( v \) for which \( \|v\|_{k,p,\Omega} \) exists and is finite, page 124

\( \mathbf{w}_K \)  Stabilization vector,
\[
\mathbf{w}_K = (P_{K,n+1} - C_K) + \frac{\varepsilon_{\mu_K}}{(b_K \cdot n_K)^2} \mathbf{b}_K + \frac{\varepsilon_{\nu_K}}{b_K \cdot n_K} \mathbf{n}_K \], page 91

\( \xi \)  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

\[ Lu := -\varepsilon u'' + b(x)u' = f \quad \text{in } \Omega = (0,1), \]
\[ u(0) = u(1) = 0, \]

with \( b(x) > \beta > 0 \) 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 \( \varepsilon \) 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 \( \varepsilon \) such that

\[ |u(x) - v(x)| \leq C\varepsilon^{m+1} \quad \text{for all } x \in [0,1] \text{ and } \varepsilon \text{ sufficiently small}. \]
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 \textit{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 \textit{reduced problem}

\[
L_0 u_0 := b(x)u_0'(x) = f(x) \quad \text{in } \Omega, \\
u(0) = 0.
\]

The reduced solution \( u_0 \) is, in fact, the first term of the so-called \textit{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

\[
L_0 u_j = u''_{j-1}, \quad u_j(0) = 0.
\]

This definition immediately implies that

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

\[
\sum_{j=0}^{m} \varepsilon^{j+1} u''_j - \sum_{j=1}^{m} \varepsilon^j u''_{j-1} = \varepsilon^{m+1} u''_m, \quad (1.7)
\]

\[
(u - u_{g,m})(0) = 0 \quad \text{and} \quad (u - u_{g,m})(1) = -\sum_{j=0}^{m} \varepsilon^j u_j(1). \quad (1.8)
\]

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

\[
\xi(x) = \frac{1-x}{\varepsilon} \quad \Rightarrow \quad x(\xi) = 1 - \varepsilon \xi, \quad (1.10)
\]

and consider the following Taylor’s expansion of the function \( b \) at \( x = 1 \)

\[
b(x) = \sum_{j=0}^{\infty} b_j (1-x)^j. \quad (1.11)
\]

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

\[
\frac{du}{dx} = \frac{dU}{d\xi} \quad \frac{d^2u}{dx^2} = \frac{1}{\varepsilon} \frac{dU}{d\xi} \quad \text{and} \quad \frac{d^2u}{dx^2} = \frac{1}{\varepsilon^2} \frac{d^2U}{d\xi^2}. \quad (1.12)
\]

Consequently, we can express the differential operator \( L \) in terms of \( \xi \) as

\[
Lu = \frac{1}{\varepsilon} \left( -\frac{d^2U}{d\xi^2} - b(1-\varepsilon\xi) \frac{dU}{d\xi} \right) = \frac{1}{\varepsilon} \left( -\frac{d^2U}{d\xi^2} - b_0 \frac{dU}{d\xi} \right) - \frac{1}{\varepsilon} \sum_{j=1}^{\infty} b_j \varepsilon^j \xi^j \frac{dU}{d\xi} = L U. \quad (1.13)
\]
We observe that the stretching factor $1/\varepsilon$ in the definition of $\xi$ 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 $\varepsilon$.

In the expression (1.14), let us denote
\[
L^*_0 := -\frac{d^2}{d\xi^2} - b_0 \frac{d}{d\xi} \quad \text{and} \quad L^*_j := -b_j \xi^j \frac{d}{d\xi}, \quad j \geq 1,
\]
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 $\varepsilon$ 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 $t = j - s + m + 1 = j - q - 1$)
\[
\mathcal{L} v_{loc,m+1} = \frac{1}{\varepsilon} \sum_{j=0}^{\infty} \varepsilon^j L^*_j v_{loc,m+1} = \frac{1}{\varepsilon} \sum_{j=0}^{\infty} \varepsilon^j L^*_j \sum_{k=0}^{m+1} \varepsilon^k v_k = \frac{1}{\varepsilon} \sum_{j=0}^{\infty} \sum_{k=0}^{m+1} \varepsilon^{j+k} L^*_j v_k =
\]
\[
= \frac{1}{\varepsilon} \sum_{j=0}^{\infty} \sum_{s=j}^{m+1+j} \varepsilon^s L^*_j v_{s-j} = \frac{1}{\varepsilon} \sum_{j=0}^{\infty} \sum_{s=j}^{m+1} \varepsilon^s L^*_j v_{s-j} + \frac{1}{\varepsilon} \sum_{s=m+2}^{\infty} \sum_{j=s-m-1}^{\infty} \varepsilon^s L^*_j v_{s-j} =
\]
\[
= \frac{1}{\varepsilon} \sum_{j=0}^{\infty} \sum_{s=j}^{m+1} \varepsilon^s L^*_j v_{s-j} + \varepsilon^{m+1} \sum_{q=0}^{\infty} \sum_{l=0}^{m} L^*_{l+q+1} v_{m+1-l}, \tag{1.17}
\]
where the sum equality $\sum_{j=0}^{\infty} \sum_{s=j}^{m+1+j} = \sum_{j=0}^{m+1} \sum_{s=0}^{j} + \sum_{s=m+2}^{\infty} \sum_{j=s-m-1}^{\infty}$ 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
\[
L^*_0 v_s = -\sum_{j=1}^{s} L^*_j v_{s-j} \quad \left(= -\sum_{j=0}^{s-1} L^*_s v_{j} \right), \quad \text{in } (0, \infty), \tag{1.18}
\]
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
\[
v_0(\xi) = -u_0(1) e^{-b_0 \xi} = -u_0(1) e^{-b(1) \xi}, \tag{1.19}
\]
\[
v_1(\xi) = \left(-u_1(1) + \frac{b_1 u_0(1) \xi}{b_0} + \frac{b_1 u_0(1) \xi^2}{2} \right) e^{-b_0 \xi} =
\]
\[
= - \left(u_1(1) + \frac{b' u_0(1) \xi}{b(1)} + \frac{b'(1) u_0(1) \xi^2}{2} \right) e^{-b(1) \xi}. \tag{1.20}
\]

Remark 1.2.1. The above applied equality
\[
\sum_{j=0}^{\infty} \sum_{s=j}^{m+1+j} a_{js} = \sum_{s=0}^{m+1} \sum_{j=0}^{s} a_{js} + \sum_{s=m+2}^{\infty} \sum_{j=s-m-1}^{\infty} a_{js} \tag{1.21}
\]
Let us consider the infinite sums. 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 \( O(\varepsilon^{m+1}) \), the interchange of sums in (1.17) is always correct, up to some \( O(\varepsilon^{m+1}) \)-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 \( b(x) > \beta > 0 \) 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

\[
\begin{align*}
\frac{m}{m}
\sum_{j=0}^{m} \varepsilon^j u_j(x) + \sum_{j=0}^{m} \varepsilon^j v_j \left( \frac{1 - x}{\varepsilon} \right),
\end{align*}
\]

such that for any sufficiently small fixed constant \( \varepsilon_0 \) there holds

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

The constant \( C \) is independent of \( x \) and \( \varepsilon \).

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

\[
\begin{align*}
L \left( u(x) - u_{as,m}^*(x) \right) &= L \left( u(x) - u_{as,m}(x) \right) + L \left( u_{as,m}(x) - u_{as,m}^*(x) \right) = \\
&= \varepsilon^{m+1} u_m \left( \frac{1 - x}{\varepsilon} \right) - L \left( u_{loc,m+1} \left( \frac{1 - x}{\varepsilon} \right) \right) = \\
&= \varepsilon^{m+1} u_m \left( \frac{1 - x}{\varepsilon} \right) - \mathcal{L} u_{loc,m+1}(\xi) = O(\varepsilon^{m+1})
\end{align*}
\]

and

\[
\begin{align*}
\left( u - u_{as,m}^* \right)(0) &= 0 - \sum_{j=0}^{\infty} \varepsilon^j v_j (1/\varepsilon) = O(\varepsilon^\mu) \quad \text{for any } \mu > 0, \\
\left( u - u_{as,m}^* \right)(1) &= - \sum_{j=0}^{\infty} \varepsilon^j (u_j(1) + v_j(0)) - \varepsilon^{m+1} v_{m+1}(0) = \varepsilon^{m+1} u_{m+1}(1).
\end{align*}
\]

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

If we now denote \( w(x) = \max \left\{ 1, \frac{x}{2} \right\} C_\lambda \varepsilon^{m+1}(x + 1) \), we obtain the following
It means that

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

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

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

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

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

It means that

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

Since both \( C_\lambda \) and \( v_{m+1} \) do not depend on \( \varepsilon \) we obtain (1.23). \( \square \)

**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

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

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

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 \( e'_0(\overline{\varepsilon}) = 0 \), i.e. \( u'_0(\overline{\varepsilon}) + \frac{b(1)}{\varepsilon} (u_0(\overline{\varepsilon}) - u_0(1)) = 0 \). In the latter case, using \( b u'_0 = f \), we can estimate

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

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 $\varepsilon$) 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

\[
Lu_{m+1}^* = u_m'' \quad \text{in } (0,1), \tag{1.39}
\]

\[
Lv_{m+1}^* = -\varepsilon^{-(m+1)}L \left( \sum_{j=0}^{m} \varepsilon^j v_j \left( \frac{1-x}{\varepsilon} \right) \right) \quad \text{in } (0,1), \tag{1.40}
\]

equipped with the boundary conditions

\[
u_{m+1}^*(0) = 0 \quad \text{and} \quad v_{m+1}^*(0) = -\varepsilon^{-(m+1)} \sum_{j=0}^{m} \varepsilon^j v_j(0). \tag{1.41}
\]

**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 $E(x) = v_0(\frac{1-x}{\varepsilon}) + \varepsilon v_1^*(x)$, 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 $v_1^*(0) = \frac{f}{b} \exp(-b/\varepsilon)$, $v_1^*(1) = 0$. It means that $u_1^* \equiv 0$ and

\[v_1^*(x) = \frac{f}{b \varepsilon} \frac{\exp \left( -\frac{b}{\varepsilon} x \right) - \exp \left( -\frac{b}{\varepsilon} (2-x) \right)}{1 - \exp \left( -\frac{b}{\varepsilon} \right)}. \tag{1.44}\]

Therefore, the $S$-decomposition takes the form

\[
S(x) = \frac{f}{b} x \tag{1.45}
\]

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

Let us also mention that despite the presence of the factor $1/\varepsilon$ the function $v_1^*$ for all $x \in [0,1]$ satisfies $|v_1^*| \leq |v_1^*(0)| = \frac{|f|}{b \varepsilon} \exp(-b/\varepsilon) \leq \frac{|f|}{b \varepsilon} \varepsilon = \frac{|f|}{b \varepsilon}$, where we have used the inequality $\exp(-x) \leq \frac{1}{e x}$, which is valid for all $x > 0$. 

13
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 \( b(x) > \beta > 0 \) and sufficiently smooth data. Its solution \( u \) can be decomposed as \( u = S + E \), where the smooth part \( S \) satisfies \( LS = f \) and

\[
|S^{(j)}(x)| \leq C_S \quad \text{for } 0 \leq j \leq q, \tag{1.47}
\]

while the layer part \( E \) satisfies \( LE = 0 \) and

\[
|E^{(j)}(x)| \leq C_E \varepsilon^{-j} \exp\left(-\frac{\beta(1-x)}{\varepsilon}\right) \quad \text{for } 0 \leq j \leq q. \tag{1.48}
\]

Here \( C_S \) and \( C_E \) are positive constants independent of \( \varepsilon \).

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

### 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

\[
Lu := -\varepsilon \Delta u(x,y) + b(x,y) \cdot \nabla u(x,y) = f(x,y) \quad \text{in} \quad \Omega \subset \mathbb{R}^2, \tag{1.49}
\]

\[
u(x,y) = 0 \quad \text{on} \quad \partial\Omega, \tag{1.50}
\]

where \( \Omega \) is a convex polygonal domain with boundary \( \partial\Omega \) satisfying

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

with \( \Gamma_+ \), \( \Gamma_0 \) and \( \Gamma_- \) defined as follows:

\[
\Gamma_+ = \{(x,y) \in \partial\Omega, b(x,y) \cdot n(x,y) > 0\},
\]

\[
\Gamma_0 = \{(x,y) \in \partial\Omega, b(x,y) \cdot n(x,y) = 0\},
\]

\[
\Gamma_- = \{(x,y) \in \partial\Omega, b(x,y) \cdot n(x,y) < 0\}. \tag{1.52}
\]

Here \( n(x,y) \) 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 \( b \in C^1(\bar{\Omega})^2 \) and \( f \in L^2(\Omega) \). In what follows
we shall also consider that the vector $b$ possesses the Taylor expansion in $\Omega$, particularly in the neighborhood of $\partial\Omega$.

As $\varepsilon \to 0^+$, the equation (1.49) becomes singularly perturbed and near the boundary $\Gamma_+$ 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 $\Gamma_+$. At first we again formally set $\varepsilon = 0$ in the equation (1.49) and obtain the reduced problem

$$b(x, y) \cdot \nabla u_0(x, y) = f(x, y) \quad \text{in} \quad \Omega \subset \mathbb{R}^2,$$

$$u_0(x, y) = 0 \quad \text{on} \quad \Gamma_-,$$

where we have to consider only the boundary condition on $\Gamma_-$ 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

$$[L_0 u_j](x, y) := b(x, y) \cdot \nabla u_j(x, y) = \Delta u_{j-1}(x, y) \quad \text{in} \quad \Omega \subset \mathbb{R}^2,$$

$$u_j(x, y) = 0 \quad \text{on} \quad \Gamma_-.$$

From this definition it follows that

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

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

$$u - u_{g,m} = 0 \quad \text{on} \quad \Gamma_- \quad \text{and}$$

$$u - u_{g,m} = -u_{g,m} \quad \text{on} \quad \Gamma_+ \cup \Gamma_0.$$
there are only two vertices \( \{P^0, P^H\} = \Gamma_{-} \cap \Gamma_{+} \) and consider \( \Gamma_{+} = \cup_{k=1}^{H} e_k \), where \( e_k \) are the edges of \( \Gamma_{+} \). Then \( P^0 \in e_1, P^H \in e_H \) and the remaining vertices of \( \Gamma_{+} \) satisfy \( P^k = e_k \cap e_{k+1}, k = 1, \ldots, H - 1 \).

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

\[
\begin{align*}
\xi_k(x, y) &= (P_{y}^{k-1} - y) \cos \alpha_k - (P_{x}^{k-1} - x) \sin \alpha_k = (P_{k-1} - X) \cdot n_k, \quad (1.60) \\
\eta_k(x, y) &= (P_{x}^{k-1} - x) \cos \alpha_k + (P_{y}^{k-1} - y) \sin \alpha_k = (P_{k-1} - X) \cdot t_k. \quad (1.61)
\end{align*}
\]

Here \( P_{k-1} = [P_{x}^{k-1}, P_{y}^{k-1}] \), \( t_k = (\cos \alpha_k, \sin \alpha_k)^T \) is the unit (tangent) vector parallel to the edge \( e_k \) and \( n_k = (-\sin \alpha_k, \cos \alpha_k)^T \) 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 \( \Omega \), we may for simplicity assume that the domain \( \Omega \) 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 \( \xi_k \) direction and define the transformation \( \Psi_k^\varepsilon : (x, y) \rightarrow (\xi_k^\varepsilon, \eta_k) \) using the same \( \xi_k, \eta_k \), i.e. \( (\xi_k, \eta_k) = \Psi_k(x, y) \).

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

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

where

\[
\begin{align*}
B_1^k(\xi_k, \eta_k) &= -b \left( \Psi_k^{-1}(\xi_k, \eta_k) \right) \cdot n_k = -b \left( P_{k-1} - \xi_k n_k - \eta_k t_k \right) \cdot n_k, \quad (1.63) \\
B_2^k(\xi_k, \eta_k) &= -b \left( \Psi_k^{-1}(\xi_k, \eta_k) \right) \cdot t_k = -b \left( P_{k-1} - \xi_k n_k - \eta_k t_k \right) \cdot t_k. \quad (1.64)
\end{align*}
\]

Now we assume that both functions \( B_1^k \) and \( B_2^k \) possess a Taylor expansion in the variable \( \xi_k \) and write

\[
\begin{align*}
B_1^k(\xi_k, \eta_k) &= \sum_{j=0}^{\infty} \frac{\varepsilon_j^j}{j!} B_{1,j}^k(0, \eta_k) = \sum_{j=0}^{\infty} \frac{\varepsilon_j^j}{j!} B_{1,j}^k(0, \eta_k) \quad \text{and} \quad (1.65) \\
B_2^k(\xi_k, \eta_k) &= \sum_{j=0}^{\infty} \frac{\varepsilon_j^j}{j!} B_{2,j}^k(0, \eta_k) = \sum_{j=0}^{\infty} \frac{\varepsilon_j^j}{j!} B_{2,j}^k(0, \eta_k), \quad (1.66)
\end{align*}
\]

Figure 1.1: A part of the general convex domain \( \Omega \).
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 $L_k^\psi$ in the local coordinates as

$$L_k^\psi = \frac{1}{\varepsilon} \sum_{j=0}^{\infty} \varepsilon^j L_k^{(j)},$$

(1.67)

where

$$L_k^{(0)} = -\frac{\partial^2}{\partial \xi_k^2} + B^k_1(0, \eta_k) \frac{\partial}{\partial \xi_k},$$

$$L_k^{(1)} = \xi_k B^k_{1,1}(0, \eta_k) \frac{\partial}{\partial \xi_k} + B^k_2(0, \eta_k) \frac{\partial}{\partial \eta_k},$$

$$L_k^{(2)} = -\frac{\partial^2}{\partial \eta_k^2} + \frac{1}{2} \xi_k^2 B^k_{1,2}(0, \eta_k) \frac{\partial}{\partial \eta_k} + \xi_k B^k_{2,1}(0, \eta_k) \frac{\partial}{\partial \eta_k},$$

$$L_k^{(j)} = \frac{1}{j!} \xi_k^j B^k_{1,j}(0, \eta_k) \frac{\partial}{\partial \eta_k} + \frac{1}{(j-1)!} \xi_k^{j-1} B^k_{2,j-1}(0, \eta_k) \frac{\partial}{\partial \eta_k}, \text{ for } j \geq 3.$$  

(1.68)

Expressing the differential operator $L_k^\psi$ in the local coordinates allows us to introduce the local expansion of the $m$-th order

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

(1.69)

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

$$L_k^{(0)} V^k_0 = 0,$$

(1.70)

$$L_k^{(0)} V^k_j = -\sum_{i=1}^{j} L_k^{(i)} V^k_{j-i}, \text{ for } j = 1, 2, \ldots, m,$$

(1.71)

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

$$V^k_j(0, \eta_k) = -u_j \left( \Psi^{-1}_k(0, \eta_k) \right), \text{ and}$$

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

(1.72)

(1.73)

While the first condition ensures the fulfillment 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

$$V^k_0(\xi_k, \eta_k) = -u_0 \left( \Psi^{-1}_k(0, \eta_k) \right) \exp \left( B^k_1(0, \eta_k) \xi_k \right).$$

(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 $\rho \in C^1[0, s_{max}]$ and $g \in C^2[0, s_{max}]$ be arbitrary functions. If $g(s) \leq -g < 0$ for all $s \in [0, s_{max}]$, then there exists a constant $C > 0$ independent of $\varepsilon$ such that

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

(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 \( \varepsilon \) such that \( \left| (\rho(s) - \rho(0)) \exp \left( g(0) \frac{s}{\varepsilon} \right) \right| \leq C_1 \varepsilon \) for all \( s \in [0, s_{\max}] \). Let us therefore denote

\[
e_1(s) = (\rho(s) - \rho(0)) \exp \left( g(0) \frac{s}{\varepsilon} \right).
\] (1.76)

Since \( e_1(0) = 0 \), then either \( |e_1(s)| \leq |e_1(s_{\max})| \leq \frac{2|\rho|_{1, \infty, (0, s_{\max})}}{g_{\max}} \varepsilon \) for all \( s \in [0, s_{\max}] \) or there exists \( s_1 \in (0, s_{\max}) \) such that \( s_1 = \arg \max_{s \in (0, s_{\max})} |e_1(s)| \). Consequently, there holds

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

which results into the inequality

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

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

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

\[
e_2(s) = \exp \left( g(s) \frac{s}{\varepsilon} \right) - \exp \left( g(0) \frac{s}{\varepsilon} \right).
\] (1.79)

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

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

which results into the inequality

\[
|e_2(s)| \leq |e_2(s_2)| = \left| \frac{g(0)}{g'(s_2) s_2 + g(s_2)} - 1 \right| \exp \left( g(0) \frac{s_2}{\varepsilon} \right).
\] (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 \( \varepsilon \) such that \( \left| (r(s_2) - r(0)) \exp \left( g(0) \frac{s_2}{\varepsilon} \right) \right| \leq C_2 \varepsilon \). Combining all estimates we get

\[
\left| \rho(s) \exp \left( g(s) \frac{s}{\varepsilon} \right) - \rho(0) \exp \left( g(0) \frac{s}{\varepsilon} \right) \right| = \\
= \left| \rho(s) \left( \exp \left( g(s) \frac{s}{\varepsilon} \right) - \exp \left( g(0) \frac{s}{\varepsilon} \right) \right) + (\rho(s) - \rho(0)) \exp \left( g(0) \frac{s}{\varepsilon} \right) \right| \leq \\
\leq \|\rho\|_{0, \infty, (0, s_{\max})} C_2 \varepsilon + C_1 \varepsilon \leq C \varepsilon.
\] (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 \(u - u_{ex,0}\) on \(\Gamma_+\).

**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 \(\varepsilon\) such that for each \(\eta_k \in [0,d_k]\) there holds

\[
\begin{align*}
|V_0^{k-1}\left(\frac{1}{\varepsilon}\eta_k \sin \gamma_{k-1}, d_{k-1} - \eta_k \cos \gamma_{k-1}\right) - V_0^{k-1}\left(\frac{1}{\varepsilon}\eta_k \sin \gamma_{k-1}, d_{k-1}\right)| & \leq C_A \varepsilon, \\
|V_0^{k+1}\left(\frac{1}{\varepsilon}(d_k - \eta_k) \sin \gamma_k, (d_k - \eta_k) \cos \gamma_k\right) - V_0^{k+1}\left(\frac{1}{\varepsilon}(d_k - \eta_k) \sin \gamma_k, 0\right)| & \leq C_B \varepsilon.
\end{align*}
\]

**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

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

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

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

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

and the previous lemma provide the desired estimate.

\(\square\)

**Remark 1.3.1.** The previous estimates are valid if the functions \(u_0\) and \(b\) 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 \(\cos \gamma_k > 0\) for all \(k = 0, 1, \ldots, H\). If we want to use these estimates for obtuse angles \(\gamma_k\), \(k = 0, 1, \ldots, H\), the problem data (and consequently the function \(u_0\)) have to be defined also in some neighborhood of \(\Omega\).

### 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

\[
\begin{align*}
u_j(P^k) + V_j^k(\Psi_k(P^k)) + V_j^{k+1}(\Psi_{k+1}^*(P^k)) = u_j(P^k) + V_j^k(0,d_k) + V_j^{k+1}(0,0) &= \\
= u_j(P^k) - u_j(P^k) - u_j(P^k) &= -u_j(P^k),
\end{align*}
\]

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 \(\Phi_k : (x,y) \rightarrow (\xi_k,\xi_{k+1})\) corresponding to the vertex \(P^k\) as

\[
\begin{align*}
\xi_k(x,y) &= (P_y^{k-1} - y) \cos \alpha_k - (P_x^{k-1} - x) \sin \alpha_k, \\
\xi_{k+1}(x,y) &= (P_y^k - y) \cos \alpha_k - (P_x^k - x) \sin \alpha_k,
\end{align*}
\]

(1.86) (1.87) (1.88)
where we used the fact, that the vector $P^k - P^{k-1}$ is perpendicular to the normal vector $n_k = (-\sin \alpha_k, \cos \alpha_k)$.

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

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

$$L_k^{\varepsilon} = \frac{1}{\varepsilon} \left( - \frac{\partial^2}{\partial \xi_k^2} + 2 \cos \gamma_k \frac{\partial^2}{\partial \xi_k \partial \xi_{k+1}} - \frac{\partial^2}{\partial \xi_{k+1}^2} + B_1^k(\varepsilon \xi_k, \varepsilon \xi_{k+1}) \frac{\partial}{\partial \xi_k} + B_2^k(\varepsilon \xi_k, \varepsilon \xi_{k+1}) \frac{\partial}{\partial \xi_{k+1}} \right)$$

(1.89)

with

$$B_1^k(\xi_k, \xi_{k+1}) = -b \left( P_k^{-1}(\xi_k, \xi_{k+1}) \right) \cdot n_k = -b \left( P^k + \frac{1}{\sin \gamma_k} (\xi_{k+1} t_k - \xi_k t_{k+1}) \right) \cdot n_k, \quad (1.90)$$

$$B_2^k(\xi_k, \xi_{k+1}) = -b \left( P_k^{-1}(\xi_k, \xi_{k+1}) \right) \cdot n_{k+1} = -b \left( P^k + \frac{1}{\sin \gamma_k} (\xi_{k+1} t_k - \xi_k t_{k+1}) \right) \cdot n_{k+1}. \quad (1.91)$$

We again assume that both functions $B_1^k$ and $B_2^k$ possess Taylor’s expansion in the form

$$B_1^k(\xi_k, \xi_{k+1}) = \sum_{i,j=0}^{\infty} \frac{\xi_i \xi_j}{i! j!} \partial^{i+j} B_1^k(0,0) = \sum_{i,j=0}^{\infty} \frac{\xi_i \xi_j}{i! j!} B_{1,ij}^k \quad \text{and} \quad (1.92)$$

$$B_2^k(\xi_k, \xi_{k+1}) = \sum_{i,j=0}^{\infty} \frac{\xi_i \xi_j}{i! j!} \partial^{i+j} B_2^k(0,0) = \sum_{i,j=0}^{\infty} \frac{\xi_i \xi_j}{i! j!} B_{2,ij}^k \quad \text{and} \quad (1.93)$$

where it obviously holds $B_{1,00}^k = B_{2,00}^k = (0,0) = -b \left( P^k \right) \cdot n_k = B_{1,0}^k(0,d_k)$ and $B_{2,00}^k = B_{2,00}^k(0,0) = -b \left( P^k \right) \cdot n_{k+1} = B_{1,0}^{k+1}(0,0)$. Since we are going to use the (negative) values $B_{1,00}^k$ and $B_{2,00}^k$ frequently, we also denote $B_1^k = B_{1,00}^k$ and $B_2^k = B_{2,00}^k$.

Using Taylor’s expansions (1.92) and (1.93) of the functions $B_1^k$ and $B_2^k$ we can express the differential operator $L_k^{\varepsilon}$ in the local coordinates as

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

where

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

$$L_k^{(r)} = \left( \sum_{i+j=r} \frac{\xi_i \xi_j}{i! j!} B_{1,ij}^k \right) \frac{\partial}{\partial \xi_k} + \left( \sum_{i+j=r} \frac{\xi_i \xi_j}{i! j!} B_{2,ij}^k \right) \frac{\partial}{\partial \xi_{k+1}}, \quad \text{for } r \geq 1. \quad (1.96)$$
Expressing the operator $L$ in the local coordinates (equality $(1.94)$) allows us to introduce the two-dimensional corner expansion

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

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

$$L_k^\varepsilon Z_{\text{cor}}^{k,m+1}(\xi_k, \xi_{k+1}) = \frac{1}{\varepsilon} \sum_{j=0}^{m+1} \sum_{i=0}^{m+1} \varepsilon^{j+i} L_k^{(j)} Z_i^k(\xi_k, \xi_{k+1}) =$$

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

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

$$L_k^{(0)} Z_0^k = 0,$$

$$L_k^{(0)} Z_j^k = - \sum_{i=1}^{j} L_k^{(i)} Z_{j-i}, \quad \text{for } j = 1, \ldots, m, \quad (1.100)$$

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

$$Z_j^k(\xi_k, 0) = -V_j^k(\xi_k, d_k), \quad (1.101)$$

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

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

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 $\Phi_k, \Phi_{k-1}$ there holds

$$\Psi_{k-1} \Psi_k^{-1}(\xi_k, \eta_k) = [\eta_k \sin \gamma_{k-1} - \xi_k \cos \gamma_{k-1}, d_{k-1} - \eta_k \cos \gamma_{k-1} - \xi_k \sin \gamma_{k-1}],$$

$$\Psi_{k+1} \Psi_k^{-1}(\xi_k, \eta_k) = [(d_k - \eta_k) \sin \gamma_k - \xi_k \cos \gamma_k, (d_k - \eta_k) \cos \gamma_k + \xi_k \sin \gamma_k],$$

$$\Phi_{k-1} \Psi_k^{-1}(\xi_k, \eta_k) = [\eta_k \sin \gamma_{k-1} - \xi_k \cos \gamma_{k-1}, \xi_k],$$

$$\Phi_k \Psi_k^{-1}(\xi_k, \eta_k) = [\xi_k, (d_k - \eta_k) \sin \gamma_k - \xi_k \cos \gamma_k], \quad (1.104)$$

where $\xi_k, \xi_{k+1}$ and $\eta_k$ 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

$$P^{k-2} - \Psi_k^{-1}(\xi_k, \eta_k) = P^{k-2} - X = P^{k-2} - P^{k-1} + \xi_k(X) n_k + \eta_k(X) t_k. \quad (1.105)$$
Hence, using the equalities \( n_k \cdot n_{k-1} = t_k \cdot t_{k-1} = -\cos \gamma_{k-1}, \) \( n_k \cdot t_{k-1} = -\sin \gamma_{k-1}, \) \( t_k \cdot n_{k-1} = \sin \gamma_{k-1} \) and \( P^{k-2} - P^{k-1} = d_{k-1} t_{k-1} \) together with the definition of \( \Psi_{k-1} \) we obtain
\[
(P^{k-2} - \Psi_k^{-1}(\xi_k, \eta_k)) \cdot n_{k-1} = \eta_k \sin \gamma_{k-1} - \xi_k \cos \gamma_{k-1} \quad \text{and} \quad (1.106)
\]
\[
(P^{k-2} - \Psi_k^{-1}(\xi_k, \eta_k)) \cdot t_{k-1} = d_{k-1} - \eta_k \cos \gamma_{k-1} - \xi_k \sin \gamma_{k-1}. \quad (1.107)
\]

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 \( \Phi_k, \Phi_{k-1} \) there holds
\[
\Psi_{k-1}\Psi_k^{-1}(0, \eta_k) = [\eta_k \sin \gamma_{k-1}, d_{k-1} - \eta_k \cos \gamma_{k-1}], \quad (1.108)
\]
\[
\Psi_{k+1}\Psi_k^{-1}(0, \eta_k) = [(d_k - \eta_k) \sin \gamma_k, (d_k - \eta_k) \cos \gamma_k], \quad (1.109)
\]
\[
\Phi_{k-1}\Psi_k^{-1}(0, \eta_k) = [\eta_k \sin \gamma_{k-1}, 0], \quad (1.110)
\]
\[
\Phi_k\Psi_k^{-1}(0, \eta_k) = [0, (d_k - \eta_k) \sin \gamma_k], \quad (1.111)
\]
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 \( b \) are sufficiently smooth functions and let all the characteristics through points of \( \overline{\Omega} \) leave \( \overline{\Omega} \) at points of \( \Gamma_+ \) in finite time. Then the solution of the problem (1.49)–(1.50) has a zeroth-order matched asymptotic expansion of the form
\[
uas_0(x, y) = \]
\[
u_0(x, y) + \sum_{k=1}^{H} V_{loc}^{k,0} \left( \frac{\xi_k(x, y)}{\varepsilon}, \eta_k(x, y) \right) + \sum_{k=1}^{H-1} Z_{cor}^{k,0} \left( \frac{\xi_k(x, y)}{\varepsilon}, \frac{\xi_{k+1}(x, y)}{\varepsilon} \right),
\]
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 \( \varepsilon \) such that
\[
|\nu(x, y) - \nuas_0(x, y)| \leq C\varepsilon \quad \text{for} \quad [x, y] \in \overline{\Omega} \text{ and } \varepsilon \leq \varepsilon_0. \quad (1.113)
\]
Here \( \varepsilon_0 \) is any sufficiently small fixed positive constant.

**Proof.** Instead of \( \nuas_0 \), we firstly prove the theorem considering the function
\[
uas_0(x, y) = \nu_0(x, y) +
\]
\[
+ \sum_{k=1}^{H} V_{loc}^{k,1} \left( \frac{\xi_k(x, y)}{\varepsilon}, \eta_k(x, y) \right) + \sum_{k=1}^{H-1} Z_{cor}^{k,1} \left( \frac{\xi_k(x, y)}{\varepsilon}, \frac{\xi_{k+1}(x, y)}{\varepsilon} \right),
\]
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
\[
L_{\Psi k} V_{loc}^{k,1} = \frac{1}{\varepsilon} \sum_{j=0}^{s-1} \sum_{r=0}^{s} \varepsilon^{j+r} L^{(j)} V_{r}^{k} =
\]
\[
= \frac{1}{\varepsilon} \sum_{j=0}^{s-1} \varepsilon^{j} \sum_{j=0}^{s} L^{(j)} V_{s-j}^{k} + \frac{1}{\varepsilon} \sum_{s=2}^{s} \varepsilon^{s} \sum_{j=s-1}^{s} L^{(j)} V_{s-j}^{k} = \frac{1}{\varepsilon} \sum_{s=2}^{s} \varepsilon^{s} \sum_{j=s-1}^{s} L^{(j)} V_{s-j}^{k}. \quad (1.115)
\]
Consequently, there exists a positive constant $C_V$ (independent of $\varepsilon$) such that for all $(\xi_k, \eta_k) \in \mathbb{R}^+ \times (0, d_k)$ there holds

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

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

$$|L_k^{\varepsilon, \varepsilon} Z_{cor}^{k, 1}| = \left| \frac{1}{\varepsilon} \sum_{s=2}^{\infty} \varepsilon^s \sum_{j=s-1}^{s} L_k^{(j)} Z_{s-j}^{k} \right| \leq C_Z \varepsilon. \quad (1.117)$$

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

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

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 $\Gamma_+$, then for arbitrary $\kappa > 0$ there exists a constant $C_\kappa^- > 0$ such that for every $\varepsilon \leq \varepsilon_0$ there holds

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

Finally, let $e_k \subset \Gamma_+$ be an arbitrary edge and let $X = \Psi_k^{(0)}(0, \eta_k) \in e_k$ 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 $O(\varepsilon)$ due to the presence of exponential functions (exponential decay). Thus, using the boundary conditions (1.101)–(1.102) results in the estimate

$$(u - u_{as,0}^*) (\Psi_k^{(0)}(0, \eta_k)) =$$

$$= O(\varepsilon) - u_0(\Psi_k^{(0)}(0, \eta_k)) - V_{loc}^{k, 1}(0, \eta_k) - V_{loc}^{k-1, 1}(\Psi_k^{(0)}(0, \eta_k)) - Z_{cor}^{k-1, 1}(\Psi_k^{(0)}(0, \eta_k)) - V_{loc}^{k-1, 1}(\Psi_k^{(0)}(0, \eta_k)) - Z_{cor}^{k-1, 1}(\Psi_k^{(0)}(0, \eta_k)) =$$

$$= O(\varepsilon) - u_0(\Psi_k^{(0)}(0, \eta_k)) + V_0^{k, 1}(0, \eta_k) - \varepsilon V_1^{k, 1}(\Psi_k^{(0)}(0, \eta_k)) -$$

$$- \varepsilon \{ V_j^{k-1}(\Psi_k^{(0)}(0, \eta_k)) + Z_j^{k-1}(\Psi_k^{(0)}(0, \eta_k)) \} -$$

$$- \varepsilon \{ V_j^{k+1}(\Psi_k^{(0)}(0, \eta_k)) + Z_j^{k+1}(\Psi_k^{(0)}(0, \eta_k)) \} =$$

$$= O(\varepsilon) - \{ V_0^{k-1} \left( \frac{1}{\varepsilon} \eta_k \sin \gamma_{k-1}, d_{k-1} - \eta_k \cos \gamma_{k-1} \right) + Z_0^{k-1} \left( \frac{1}{\varepsilon} \eta_k \sin \gamma_{k-1}, 0 \right) \} -$$

$$- \{ V_0^{k} \left( \frac{1}{\varepsilon} \eta_k \sin \gamma_k, d_{k-1} - \eta_k \cos \gamma_k \right) + Z_0^{k} \left( 0, \frac{1}{\varepsilon} \eta_k \sin \gamma_k \right) \} =$$

$$= O(\varepsilon) - \{ V_0^{k-1} \left( \frac{1}{\varepsilon} \eta_k \sin \gamma_{k-1}, d_{k-1} - \eta_k \cos \gamma_{k-1} \right) - V_0^{k-1} \left( \frac{1}{\varepsilon} \eta_k \sin \gamma_{k-1}, d_{k-1} \right) \} -$$

$$- \{ V_0^{k} \left( \frac{1}{\varepsilon} \eta_k \sin \gamma_k, d_{k-1} - \eta_k \cos \gamma_k \right) - V_0^{k} \left( \frac{1}{\varepsilon} \eta_k \sin \gamma_k, 0 \right) \}. \quad (1.103)$$
Hence, applying the estimate of Corollary 1.3.1 one can find a constant \( C_0^+ > 0 \) such that \( |(u - u_{as,0})|_{\Gamma_+} \leq C_0^+ \varepsilon \).

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 \( \Gamma_- \) and we consider \( u_0 \) being sufficiently smooth, then it follows that \( |u_0| = O(\varepsilon) \) in \( \mathcal{U}_\varepsilon(P) \) and consequently, we indeed obtain \( |u - u_{as,0}| = O(\varepsilon) \) 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 \( \phi \) such that \( L_0 \phi = b \cdot \nabla \phi \geq \phi_0 > 0 \) in \( \Omega \).

Therefore, if we define the function \( W \) by the relation

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

then employing the inequality (1.118) we may estimate

\[
LW \geq C_0^* \varepsilon \geq L (u - u_{as,0}^*) \geq -C_0^* \varepsilon \geq -LW. \tag{1.122}
\]

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

\[
W \geq \max \left\{ C_0^- , C_0^+ \right\} \varepsilon \geq (u - u_{as,0}^*) |_{\partial \Omega} \geq -\max \left\{ C_0^- , C_0^+ \right\} \varepsilon \geq -W. \tag{1.123}
\]

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

\[
W \geq (u - u_{as,0}^*) \geq -W \quad \text{in} \quad \Omega , \tag{1.124}
\]

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

\[
|(u - u_{as,0})(x, y)| \leq |(u - u_{as,0}^*)(x, y)| + \varepsilon \sum_{k=1}^H V_k^k \left( \frac{\xi_k(x, y)}{\varepsilon} , \eta_k(x, y) \right) + \\
+ \varepsilon \sum_{k=1}^{H-1} Z^k_1 \left( \frac{\xi_k(x, y)}{\varepsilon} , \frac{\xi_{k+1}(x, y)}{\varepsilon} \right) \leq \\
\leq W + \varepsilon \left( \sum_{k=1}^H \|V_k^k\|_{0,\infty, \mathbb{R}^+ \times (0,d_k)} + \sum_{k=1}^{H-1} \|Z^k_1\|_{0,\infty, \mathbb{R}^+ \times \mathbb{R}^+} \right) \leq \\
\leq \varepsilon \left( \max \left\{ \frac{2}{\phi_0} C_0^*, C_0^- , C_0^+ \right\} \left( 1 + \|\phi\|_{0,\infty, \Omega} \right) + \\
+ \sum_{k=1}^H \|V_k^k\|_{0,\infty, \mathbb{R}^+ \times (0,d_k)} + \sum_{k=1}^{H-1} \|Z^k_1\|_{0,\infty, \mathbb{R}^+ \times \mathbb{R}^+} \right). \tag{1.125}
\]
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
\[
Z^k_0(\xi_k, \xi_{k+1}) = u_0(P^k) \left\{ \sum_{j=0}^{\rho_k} \exp\left( p_j^k \xi_k + q_j^k \xi_{k+1} \right) - \sum_{j=0}^{\rho_k-1} \exp\left( p_j^{k+1} \xi_k + q_j^{k+1} \xi_{k+1} \right) \right\},
\]
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 \( Z^k_0(0, \xi_{k+1}) = u_0(P^k) \exp\left( q_0^k \xi_{k+1} \right) \) and \( Z^k_0(\xi_k, 0) = u_0(P^k) \exp\left( p_0^k \xi_k \right) \). 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 \( j \in \{0, 1, \ldots, \rho_k - 1\} \). If, for instance, choose \( p_m^k > 0 \), then the difference \( \exp\left( p_m^k \xi_{k+1} + q_m^k \xi_{k} \right) - \exp\left( p_m^k \xi_{k+1} + q_m^k \xi_{k} \right) \) has to tend to zero as \( \xi_{k+1} \to +\infty \) or \( \xi_k \to +\infty \). However, this is not possible since for fixed \( \xi_k \) one has
\[
\lim_{\xi_{k+1} \to +\infty} \left| \exp\left( p_m^k \xi_{k+1} \right) \left( \exp\left( q_m^k \xi_k \right) - \exp\left( q_m^k \xi_k \right) \right) \right| = +\infty.
\]

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
\[
- \left( p_j^k \right)^2 + 2p_j^k q_j^k \cos \gamma_k - \left( q_j^k \right)^2 + \beta_1^k p_j^k + \beta_2^k q_j^k = 0,
\]
\[
- \left( p_j^{k+1} \right)^2 + 2p_j^{k+1} q_j^{k+1} \cos \gamma_k - \left( q_j^{k+1} \right)^2 + \beta_1^k p_j^{k+1} + \beta_2^k q_j^{k+1} = 0,
\]
\[
- \left( p_j^{k+1} \right)^2 + 2p_j^{k+1} q_j^{k+1} \cos \gamma_k - \left( q_j^{k+1} \right)^2 + \beta_1^k p_j^{k+1} + \beta_2^k q_j^{k+1} = 0.
\]

Subtracting the second equation from the first one and the third one from the second one yields
\[
\left( p_j^{k+1} - p_j^k \right) \left( p_j^{k+1} + p_j^k - 2q_j^k \cos \gamma_k - \beta_1^k \right) = 0,
\]
\[
\left( q_j^{k+1} - q_j^k \right) \left( q_j^{k+1} + q_j^k - 2p_j^{k+1} \cos \gamma_k - \beta_2^k \right) = 0.
\]

If \( p_j^{k+1} = p_j^k \) 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 \( p_j^{k+1} \neq p_j^k \) and \( q_j^{k+1} \neq q_j^k \) for all \( j \in \{0, 1, \ldots, \rho_k - 1\} \). Consequently, from (1.129) it follows that
\[
p_j^{k+1} = -p_j^k + 2q_j^k \cos \gamma_k + \beta_1^k
\]
and using this equality together with (1.130) we get
\[
q_j^{k+1} = -q_j^k + 2p_j^{k+1} \cos \gamma_k + \beta_2^k = -2p_j^k \cos \gamma_k + \left( 4 \cos^2 \gamma_k - 1 \right) q_j^k + 2\beta_1^k \cos \gamma_k + \beta_2^k.
\]

In order to simplify further calculations we denote \( w_j^k = (p_j^k, q_j^k)^T \) for all \( j \in \{0, 1, \ldots, \rho_k\} \). Then for all \( j \in \{1, 2, \ldots, \rho_k\} \) it holds
\[
w_j^k = A w_{j-1}^k + r^k,
\]
where
\[ A = \begin{pmatrix} -1 & 2\cos\gamma_k \\ -2\cos\gamma_k & 4\cos^2\gamma_k - 1 \end{pmatrix} \quad \text{and} \quad r^k = \begin{pmatrix} \beta_1^k \\ \beta_2^k + 2\beta_1^k\cos\gamma_k \end{pmatrix}. \tag{1.135} \]

Since \( p_k^k = \beta_1^k \) and \( q_k^k \neq 0 \) then from the first equation in (1.128) it follows that \( q_k^0 = \beta_2^k + 2\beta_1^k\cos\gamma_k \). Thus, there holds \( w_0^k = r^k \). Let us further define the fixed point \( \bar{w}^k \) of the iterations given by (1.134)

\[ \bar{w}^k = A\bar{w}^k + r^k. \tag{1.136} \]

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

\[
\begin{align*}
\bar{w}_j^k - \bar{w}^k &= A (\bar{w}_j^k - \bar{w}^k) = A (w_j^k - \bar{w}^k) = A (r^k - \bar{w}^k) \\
&= -A^j (-A\bar{w}^k) = -A^j + 1 \bar{w}^k.
\end{align*} \tag{1.137}
\]

Thus \( \bar{w}_j^k = (I - A^{j+1}) \bar{w}^k \) and we would like to find such an index \( p_k \in \mathbb{N} \) for which the second component of \( w_{p_k}^k \) is equal to \( \beta_2^k \). 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 \( \bar{w}^k = (\beta_1^k + 2\beta_2^k\cos\gamma_k, \beta_2^k)^T \) and observe that the vector \( \bar{w}^k \) satisfies \( -A\bar{w}^k = r^k = (I - A)\bar{w}^k \) by (1.136). From this observation it follows that \( \bar{w}_{p_k}^k = \bar{w}^k \) if and only if \( (I - A^{p_k+1}) \bar{w}^k = (I - A^{-1}) \bar{w}^k \), i.e. if

\[ (A^{p_k+1} - A^{-1}) \bar{w}^k = 0. \tag{1.138} \]

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

\[
A^j = \frac{-1}{2\sin\gamma_k} \begin{pmatrix} e^{\frac{j\pi}{2}} & e^{-\frac{j\pi}{2}} \\ e^{-\frac{j\pi}{2}} & e^{\frac{j\pi}{2}} \end{pmatrix} \begin{pmatrix} e^{-2j\pi i} & 0 \\ 0 & e^{2j\pi i} \end{pmatrix} = \frac{1}{\sin\gamma_k} \begin{pmatrix} \sin(1 - 2j)\gamma_k & \sin 2j\gamma_k \\ -\sin 2j\gamma_k & \sin(1 + 2j)\gamma_k \end{pmatrix}. \tag{1.139} \]

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

\[ \frac{-\sin(p_k + 2)\gamma_k}{\sin^2\gamma_k} \begin{pmatrix} \beta_1^k \sin p_k\gamma_k + \beta_2^k \sin(p_k - 1)\gamma_k \\ \beta_1^k \sin p_k\gamma_k + \beta_2^k \sin(p_k + 1)\gamma_k \end{pmatrix} = 0. \tag{1.140} \]

From the definition of \( \beta_1^k \) and \( \beta_2^k \) it follows that both \( \beta_1^k \) and \( \beta_2^k \) 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 \( (p_k + 2)\gamma_k = m\pi \), 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 \( p_k = m\nu_k - 2 \) for any \( m \in \mathbb{N} \) and \( p_j^k, q_j^k \) defined recursively using (1.134) leads to the fulfilment of (1.19) 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

\[
\begin{pmatrix}
p_j^k \\
q_j^k
\end{pmatrix} = (I - A^k)^{j+1} \begin{pmatrix}
w^k \\
\beta^k
\end{pmatrix} = \frac{\sin^2(j+1)\gamma_k}{\sin^2\gamma_k} \begin{pmatrix}
\beta_1^k + \beta_2^k \frac{\sin j\gamma_k}{\sin(j+1)\gamma_k} \\
\beta_2^k + \beta_1^k \frac{\sin(j+1)\gamma_k}{\sin j\gamma_k}
\end{pmatrix} .
\] (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 \( q_j^{k-1} \) we deduce that both terms \( \beta_1^k + \beta_2^k \frac{\sin j\gamma_k}{\sin(j+1)\gamma_k} \) and \( \beta_2^k + \beta_1^k \frac{\sin(j+1)\gamma_k}{\sin j\gamma_k} \) have to be negative and that is why

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

However, this property implies that one can consider only \( \mu_k = 1 \) and \( \rho_k = 1 \), 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 \( j = \left\lfloor \frac{\nu_k}{\rho_k} \right\rfloor \leq \nu_k \leq \frac{\rho_k + 2}{2} \leq \rho_k \) it holds

\[
2\pi > (j+1)\gamma_k = \left( \frac{\nu_k}{\mu_k} + 1 \right) \frac{\mu_k}{\nu_k} \pi > \pi > \frac{\nu_k}{\mu_k} \frac{\mu_k}{\nu_k} \pi = j\gamma_k > 0,
\] (1.143)

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

Similarly, if \( \nu_k = \rho_k - 2 \) and \( \gamma_k = \frac{\mu_k}{\rho_k} \pi \) for some \( \mu_k \geq 2 \), then for \( j = \left\lfloor \frac{\mu_k}{\rho_k} \right\rfloor = \left\lfloor \frac{\rho_k + 2}{\mu_k} \right\rfloor \leq \left\lfloor \frac{\rho_k}{2} + 1 \right\rfloor \leq \rho_k \) 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 \( \pi/m, m \in \mathbb{N}, m \geq 2 \). 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 \( \xi_k \) direction. However, in this case we use the transformation \( \Psi_k^\varepsilon : (x, y) \rightarrow (\varepsilon \xi_k, \eta_k) \), where \( (\xi_k, \eta_k) = \Psi_k(x, y) \).

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

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

A boundary edge \( e_k \subseteq \Gamma_0 \) (where parabolic boundary layer occurs) is characterized by the condition \( B_2^k(0, \eta_k) = -b(\Psi_k^{-1}(0, \eta_k)) \cdot n_k = 0 \), for all \( \eta_k \in [0, d_k] \).
Therefore, the Taylor expansion (1.65) of the function $B^k_1(\xi_k, \eta_k)$ in the variable $\xi_k$ does not contain the first (absolute) term, hence

$$
\frac{B^k_1(\sqrt{\varepsilon}\xi_k, \eta_k)}{\sqrt{\varepsilon}} = \sum_{j=1}^{\infty} \frac{\xi_k^{(j)} (\sqrt{\varepsilon})^{-j}}{j!} \frac{\partial B^k_1(0, \eta_k)}{\partial \xi_k^j} = \sum_{j=1}^{\infty} \frac{\xi_k^{(j)} (\sqrt{\varepsilon})^{-j}}{j!} B^k_{1,j}(0, \eta_k).
$$

(1.145)

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

$$
L^\varepsilon_k = \sum_{j=0}^{\infty} (\sqrt{\varepsilon})^j F_k^{(j)},
$$

(1.146)

where

\[
\begin{align*}
F_k^{(0)} &= -\frac{\partial^2}{\partial \xi_k^2} + \xi_k B^k_{1,1}(0, \eta_k) \frac{\partial}{\partial \xi_k} + B^k_{1}(0, \eta_k) \frac{\partial}{\partial \eta_k}, \\
F_k^{(1)} &= \frac{\xi_k B^k_{1,2}(0, \eta_k) \frac{\partial}{\partial \xi_k} + \eta_k B^k_{1,1}(0, \eta_k) \frac{\partial}{\partial \eta_k}, \\
F_k^{(2)} &= -\frac{\partial^2}{\partial \eta_k^2} + \frac{\xi_k B^k_{1,3}(0, \eta_k) \frac{\partial}{\partial \xi_k} + \eta_k B^k_{1,2}(0, \eta_k) \frac{\partial}{\partial \eta_k}, \\
F_k^{(j)} &= \frac{1}{(j+1)!} \xi_k^{j+1} B^k_{1,j+1}(0, \eta_k) \frac{\partial}{\partial \xi_k} + \frac{1}{j!} \eta_k^{j} B^k_{1,j}(0, \eta_k) \frac{\partial}{\partial \eta_k}, \quad j \geq 3.
\end{align*}
\]

(1.147)

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

\[
\begin{align*}
F_k^{(0)} W^k_0 &= 0, \\
F_k^{(0)} W^k_j &= -\sum_{i=1}^{j} F_k^{(i)} W^k_{j-i}, \quad \text{for } j = 1, 2, \ldots, 2m.
\end{align*}
\]

(1.148, 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)

\[
\begin{align*}
W^k_j(0, \eta_k) &= -u_{j/2}(\Psi^{-1}_k(0, \eta_k)), \quad \forall \eta_k \in (0, d_k), \quad j \text{ even,}
\\
W^k_j(0, \eta_k) &= 0, \quad \forall \eta_k \in (0, d_k), \quad j \text{ odd,}
\\
\lim_{\xi_k \to +\infty} W^k_j(\xi_k, \eta_k) &= 0, \quad \forall \eta_k \in (0, d_k),
\\
W^k_j(\xi_k, 0) &= 0, \quad \forall \xi_k \in \mathbb{R}^+.
\end{align*}
\]

(1.150, 1.151, 1.152, 1.153)

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

\[
\begin{align*}
L \left( u(x, y) - u_{g,m}(x, y) - W^{k,m}_{loc}(\Psi_k(\varepsilon, x, y)) \right) &= \\
&= \varepsilon^{m+1} u_{m}(x, y) - \sum_{i=2m+1}^{\infty} (\sqrt{\varepsilon})^i \sum_{j=0}^{2m} F_k^{(i-j)} W^k_j(\Psi_k(x, y)) = O(\varepsilon^{m+1/2}),
\end{align*}
\]

(1.154)

\[
\begin{align*}
u(u^{-1}_k(0, \eta_k)) - u_{g,m}(u^{-1}_k(0, \eta_k)) - W^{k,m}_{loc}(0, \eta_k) &= \\
&= 0 - \sum_{j=0}^{m} \varepsilon^j u_j(u^{-1}_k(0, \eta_k)) - \sum_{j=0}^{m} \varepsilon^j W^k_{2j}(0, \eta_k) = 0 \quad \text{and}
\end{align*}
\]

(1.155)

\[
\left. \left| u(x, y) - u_{g,m}(x, y) - W^{k,m}_{loc}(\Psi_k(\varepsilon, x, y)) \right| \right|_{\Gamma} = O(\varepsilon^{m+1/2}).
\]

(1.156)
Hence, applying the comparison principle one can show that there exists a constant $C_P > 0$ independent of $\varepsilon$ such that in the vicinity of the edge $e_k$ there holds
\[ |u(x, y) - u_{g,m}(x, y) - W_{\text{loc}}^{k,m}\left(\Psi^k(x, y)\right)| \leq C_P \varepsilon^{m+1/2}. \] (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 $b$ is constant in $\Omega$ and $b \cdot n_k = 0$. Then $B^k_1(\xi_k, \eta_k) = 0$ and $B^k_2(\xi_k, \eta_k) = -b \cdot t_k = |b|$. Thus, function $W^k_0$ is a solution of the parabolic initial-boundary value problem
\[- \frac{\partial^2 W^k_0}{\partial \xi_k^2} + |b| \frac{\partial W^k_0}{\partial \eta_k} = 0 \quad \text{in } \mathbb{R}^+ \times (0, d_k), \] (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 $W^k_0(0, \eta_k) = g(\eta_k) = -u_0(\Psi^k_1(0, \eta_k)) = -u_0(P^{k-1} - \eta_k t_k)$ for all $\eta_k \in (0, d_k)$. This is, in fact, the heat equation whose solution can be expressed in a form
\[ W^k_0(\xi_k, \eta_k) = \int_0^{\eta_k} G(\xi_k, \eta_k - s)g(s) \, ds, \] (1.159)
where
\[ G(\xi, \eta) = \frac{\xi \sqrt{|b|}}{2 \sqrt{\pi} \eta^3} \exp\left(-\frac{|b| \xi^2}{4 \eta}\right). \] (1.160)

If we employ a substitution $s = \eta_k - \frac{|b| \xi^2}{2 \eta^2}$ the expression (1.159) takes a form
\[ W^k_0(\xi_k, \eta_k) = \sqrt{\frac{2}{\pi}} \int_{\xi_k \sqrt{\frac{|b|}{2 \eta_k}}}^{+\infty} g\left(\eta_k - \frac{|b| \xi_k^2}{2 \eta^2}\right) \exp\left(-\frac{t^2}{2}\right) \, dt. \] (1.161)

Further, considering $g(\eta_k) = -\eta_k \frac{f}{|b|}$, we can evaluate the previous integral and obtain
\[ W^k_0(\xi_k, \eta_k) = \frac{f \eta_k}{|b|} \left[ \text{erf}\left(\frac{\xi_k \sqrt{|b|}}{2 \sqrt{\eta_k}}\right) - 1 + \frac{\xi_k^2 |b|}{4 \eta_k \sqrt{\pi}} \Gamma\left(-\frac{1}{2}, \frac{\xi_k^2 |b|}{4 \eta_k}\right)\right]. \] (1.162)

where $\Gamma(s, x) = \int_x^{+\infty} t^{s-1} \text{exp}(-t) \, dt$ is the upper incomplete gamma function and $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \text{exp}(-t^2/2) \, dt = 1 - \frac{1}{\sqrt{\pi}} \Gamma(1/2, x^2)$ 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 $|b| = 1$)
\[ u_0(x, y) + W^k_0\left(\frac{\xi_k(x, y)}{\sqrt{\varepsilon}}, \eta_k(x, y)\right). \] (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 $\lim_{\xi \to 0} \int_A^B G(\xi, \eta - s) \, ds$. Since for any $0 \leq A \leq B \leq \eta$ there holds
\[ \int_A^B G(\xi, \eta - s) \, ds = \text{erf}\left(\frac{\xi \sqrt{|b|}}{2 \sqrt{\eta - B}}\right) - \text{erf}\left(\frac{\xi \sqrt{|b|}}{2 \sqrt{\eta - A}}\right), \] (1.164)
we can split the integral (1.159) into

\[\int_0^\eta G(\xi, \eta - s) g(s) \, ds = \int_0^{\eta - C\xi} G(\xi, \eta - s) g(s) \, ds + \int_{\eta - C\xi}^\eta G(\xi, \eta - s) g(s) \, ds\]

and estimate

\[\left[ \text{erf} \left( \frac{\sqrt{\xi|b|}}{2\sqrt{C}} \right) - \text{erf} \left( \frac{\sqrt{\xi|b|}}{2\sqrt{\eta}} \right) \right] \min_{s \in [0, \eta - C\xi]} g(s) \leq \int_0^{\eta - C\xi} G(\xi, \eta - s) g(s) \, ds \leq \left[ \text{erf} \left( \frac{\sqrt{\xi|b|}}{2\sqrt{C}} \right) - \text{erf} \left( \frac{\sqrt{\xi|b|}}{2\sqrt{\eta}} \right) \right] \max_{s \in [0, \eta - C\xi]} g(s)\] (1.165)

and

\[\left[ 1 - \text{erf} \left( \frac{\sqrt{\xi|b|}}{2\sqrt{C}} \right) \right] \min_{s \in [\eta - C\xi, \eta]} g(s) \leq \int_{\eta - C\xi}^\eta G(\xi, \eta - s) g(s) \, ds \leq \left[ 1 - \text{erf} \left( \frac{\sqrt{\xi|b|}}{2\sqrt{C}} \right) \right] \max_{s \in [\eta - C\xi, \eta]} g(s),\] (1.166)

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

Using the equality \(G(\sqrt{\eta}, \eta) = \frac{|b|}{2\sqrt{\pi}} \exp \left( -\frac{|b|}{4} \right)\) 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 \(\lim_{\xi \to 0} \int_0^\eta G(\xi, \eta - s) \, ds = 1\). Hence, \(G(0, \eta)\) is the Dirac delta function.

\textit{Remark 1.3.3.} If we consider that the function \(g\) possesses the Taylor expansion
of the form \( g \left( \eta_k - \frac{|b| \xi_k^2}{2r^2} \right) = \sum_{j=0}^{-\infty} g^{(j)}(\eta_k) \left( -\frac{|b| \xi_k^2}{4j^2} \right)^j \), then from (1.161) it follows

\[
W_k^0(\xi_k, \eta_k) = \sqrt{\frac{2}{\pi}} \int_{\xi_k}^{+\infty} \frac{1}{\sqrt{\xi_k}} \sum_{j=0}^{\infty} g^{(j)}(\eta_k) \left( -\frac{|b| \xi_k^2}{4j^2} \right)^j \exp \left( -\frac{r^2}{2} \right) \, dr = \end{equation}

\[
= \sqrt{\frac{2}{\pi}} \sum_{j=0}^{\infty} \frac{g^{(j)}(\eta_k)}{j!} \left( -\frac{|b| \xi_k^2}{4} \right)^j \int_{\eta_k}^{+\infty} t^{-j} \exp \left( -\frac{t^2}{2} \right) \, dt = \end{equation}

\[
= \frac{1}{\sqrt{\pi}} \sum_{j=0}^{\infty} \frac{g^{(j)}(\eta_k)}{j!} \left( -\frac{|b| \xi_k^2}{4} \right)^j \Gamma \left( \frac{1}{2} - j, \frac{|b| \xi_k^2}{4\eta_k} \right), \]

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 \( b = (1, 0)^T \) 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.](image)

The mappings \( \Psi_1, \Psi_2 \) corresponding to the edges \( e_1, e_2 \) are then defined by

\[
\Psi_1(x, y) = (\xi_1(x, y), \eta_1(x, y)) \quad \text{and} \quad \Psi_2(x, y) = (\xi_2(x, y), \eta_2(x, y)), \tag{1.168}
\]

where

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

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

31
Consequently, the inverse mappings satisfy

\[
\begin{align*}
\Psi_1^{-1}(0, \eta_1(x, y)) &= \left( \eta_1(x, y) \cos \frac{\gamma}{2}, \eta_1(x, y) \sin \frac{\gamma}{2} - \tan \frac{\gamma}{2} \right), \quad (1.171) \\
\Psi_2^{-1}(0, \eta_2(x, y)) &= (1 - \eta_2(x, y) \cos \frac{\gamma}{2}, \eta_2(x, y) \sin \frac{\gamma}{2}). \quad (1.172)
\end{align*}
\]

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

\[
\begin{align*}
u_{as,0}(x, y) &= u_0(x, y) - u_0 \left( \Psi_1^{-1}(0, \eta_1(x, y)) \right) \exp \left( B_1^1(0, \eta_1(x, y)) \frac{\xi_1(x, y)}{\varepsilon} \right) - \quad - u_0 \left( \Psi_2^{-1}(0, \eta_2(x, y)) \right) \exp \left( B_2^1(0, \eta_2(x, y)) \frac{\xi_2(x, y)}{\varepsilon} \right) + \\
&+ u_0(P^1) \left\{ \sum_{r=0}^r \exp \left( p_j^r \frac{\xi_1(x, y)}{\varepsilon} + q_j^r \frac{\xi_2(x, y)}{\varepsilon} \right) - \sum_{r=0}^r \exp \left( p_{j+1}^r \frac{\xi_1(x, y)}{\varepsilon} + q_{j+1}^r \frac{\xi_2(x, y)}{\varepsilon} \right) \right\}, \quad (1.173)
\end{align*}
\]

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

\[
\begin{align*}
B_1^1 &= B_1^1(0, d_1) = -\mathbf{b}(P^1) \cdot \mathbf{n}_1, \quad p_j^r = \frac{\sin^2((j+1)\gamma)}{\sin^2 \gamma} \left( B_1^1 + B_2^1 \frac{\sin(\gamma)}{\sin((j+1)\gamma)} \right), \quad (1.174) \\
B_2^1 &= B_1^1(0, 0) = -\mathbf{b}(P^1) \cdot \mathbf{n}_2, \quad q_j^r = \frac{\sin^2((j+1)\gamma)}{\sin^2 \gamma} \left( B_1^2 + B_2^1 \frac{\sin((j+2)\gamma)}{\sin((j+1)\gamma)} \right), \quad (1.175)
\end{align*}
\]

with \( \mathbf{n}_1 = \left( \sin \frac{\gamma}{2}, -\cos \frac{\gamma}{2} \right), \mathbf{n}_2 = \left( \sin \frac{\gamma}{2}, \cos \frac{\gamma}{2} \right) \) and \( j \in \{0, 1, \ldots, r\} \).
Numerical experiments were carried out with the use of the discontinuous Galerkin method (see, e.g. [Rivièr (2008)]) with piecewise linear approximations on uniformly refined meshes having approximately 5000 elements for several different values of $\gamma$ and $\varepsilon$. 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 $\varepsilon$. We observe that $EOC \approx 1$ for all considered angles $\gamma$, which is in a good agreement with derived theoretical results of order $O(\varepsilon)$ according to (1.113).

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

Table 1.1: Computational errors in $L^\infty(\Omega)$-norm and experimental orders of convergence for different values of $\gamma$ and $\varepsilon$ (adopted from [Lamac (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

\[- \varepsilon u'' + b(x)u' = f(x) \quad \text{in } \Omega = (0,1), \quad \tag{2.1}\]

\[u|_{\partial \Omega} = 0, \quad \tag{2.2}\]

where \(b(x) > \beta > 0, -b' \geq 0, 1 \gg \varepsilon > 0\) 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 \(\varphi \in H^1_0(\Omega)\), integrating over \(\Omega\) and using Green’s theorem or integration by parts in 1D (Theorem 4.1.1, page 124) the weak formulation reads

Find \(u \in H^1_0(\Omega)\) such that

\[a_1(u, \varphi) = (f, \varphi)_\Omega \quad \forall \varphi \in H^1_0(\Omega), \quad \tag{2.3}\]

where

\[a_1(u, \varphi) = \varepsilon (u', \varphi')_\Omega + (bu', \varphi)_\Omega. \quad \tag{2.4}\]

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

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

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 \(T_h\) of the domain \(\Omega\) 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 \(T_h\) 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 \(H^1_0(\Omega)\) by a finite element subspace \(V_h = X_h \cap H^1_0(\Omega)\), where

\[X_h = \{ \varphi_h \in C(\Omega), \varphi_h|_{I_j} \in P_1(I_j) \ \forall j = 1, 2, \ldots, N \}. \quad \tag{2.6}\]
Then \( u_h \in V_h \) is a discrete solution of (2.1)–(2.2) if
\[
a_1(u_h, \varphi_h) = (f, \varphi_h)_\Omega \quad \forall \varphi_h \in V_h.
\]
(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 \( b = \text{const.}, f = \text{const.} \) and let the partition \( T_h \) of the domain \( \Omega \) 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
\[
\varepsilon \frac{-u_{j-1} + 2u_j - u_{j+1}}{h} + b \frac{u_{j+1} - u_{j-1}}{2} = fh \quad \forall j = 1, 2, \ldots, N - 1.
\]
(2.8)

The solution of this difference equation has a form
\[
u_j = \begin{cases} \frac{f}{b} \left( jh - \frac{r-1}{r} \right), & \text{for } r = \frac{1+P_e}{1-P_e} \text{ and } P_e \neq 1, \\ \frac{f}{b} jh, & \text{for } P_e = 1, \end{cases}
\]
(2.9)

where \( P_e = \frac{bh}{\varepsilon} \) is the so-called Péclet number. When \( b \) is nonconstant we also write \( P_e(x) = \frac{b(x)h}{\varepsilon} \). 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
\[
u(x) = \frac{f}{b} \left( x - \frac{\exp \left( \frac{b}{\varepsilon} x \right) - 1}{\exp \left( \frac{b}{\varepsilon} \right) - 1} \right)
\]
(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 \( u \in H^1_0(\Omega) \cap H^2(\Omega) \). Thus, for any \( \tau_j \in \mathbb{R} \) the SUPG method reads

Find \( u_h \in V_h \) such that
\[
a_1(u_h, \varphi_h) + \sum_{j=1}^{N} \tau_j (R(u_h), b \varphi_h)_I = (f, \varphi_h)_\Omega, \quad \text{for all } \varphi_h \in V_h.
\]
(2.11)
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

$$
\varepsilon (u'_h, \varphi'_h)_{I_j} + \sum_{j=1}^{N} (bu'_h, \varphi_h + \tau_j b \varphi'_h)_{I_j} = \sum_{j=1}^{N} (f, \varphi_h + \tau_j b \varphi'_h)_{I_j}, \quad \text{for all} \quad \varphi_h \in V_h.
$$

(2.12)

The stabilization parameter $\tau_j$ 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 $\tau_j$ are nonzero and equal to $\tau$, the stencil of the method takes the form

$$
(\varepsilon + b^2 \tau) \frac{-u_{j-1} + 2u_j - u_{j+1}}{h} + b \frac{u_{j+1} - u_{j-1}}{2} = fh, \quad \forall j = 1, 2, \ldots, N - 1.
$$

(2.13)

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

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

(2.14)

The solution of the respective difference equation is also easily computable

$$
u_{j}^{upw} = \frac{f}{b} \left( jh - \frac{r^j - 1}{r^N - 1} \right), \quad r = 1 + 2Pe.
$$

(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 $\varepsilon$) since one can prove only that there exist positive constants $\tilde{C}_0, \tilde{C}_1, \tilde{C}_2$ and $\beta^*$ such that (cf. e.g. [Roos et al. 2008, p. 49])

$$
|u(x_j) - u_{j}^{upw}| \leq \begin{cases} 
\tilde{C}_0 h \left[ 1 + \varepsilon^{-1} \exp(-\beta^*(1 - x_j)/\varepsilon) \right] & \text{for } h \leq \varepsilon \\
\tilde{C}_1 h + \tilde{C}_2 \exp(-\beta^*(1 - x_{j+1})/\varepsilon) & \text{for } h \geq \varepsilon 
\end{cases}
$$

(2.16)
Thus, for $1 > h \geq \varepsilon$ and $x_{j+1} \in (1 - \frac{\varepsilon}{b}|\ln h|, 1)$ 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 $\tau_* = \frac{h}{2b} \coth Pe - \frac{r}{2}$. For constant data the solution to the respective difference equation is

$$u^*_j = \frac{f}{b} \left( jh - \frac{r^j - 1}{r^N - 1} \right), \quad r = \exp(2Pe),$$

(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

$$\|v_h\|_{\infty,d} = \max_{1 \leq i \leq N} |v(x_i)|.$$

### 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 $\varphi_h$ 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

$$\sum_{j=1}^{N} \left\{ \varepsilon (u'_h, (\varphi_h + \tau_j b \varphi_h'))_{I_j} + (bu'_h, \varphi_h + \tau_j b \varphi_h')_{I_j} \right\} = \sum_{j=1}^{N} (f, \varphi_h + \tau_j b \varphi_h')_{I_j}. \quad (2.18)$$

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 $H^1_0(\Omega)$, 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)

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

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

$$\varepsilon (u'_h, \tilde{\varphi}_j)'_\Omega + (bu'_h, \tilde{\varphi}_j)_\Omega = \left( \varepsilon + \sigma \frac{bh}{2} \right) -\frac{u_{j-1} + 2u_j - u_{j+1}}{h} + b \frac{u_{j+1} - u_{j-1}}{2}. \quad (2.20)$$

Comparing (2.20) with (2.13) we realize that considering $\sigma = 1$ leads to the simple upwind scheme, whereas taking $\sigma = \coth Pe - 1/Pe$ yields the Il’in-Allen-Southwell scheme.
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 $\varepsilon$. Thus, instead of (2.7) we consider a discrete problem of the form

\[
(\varepsilon + \tilde{\varepsilon})(u_h',\varphi_h')_\Omega + (bu_h',\varphi_h)_\Omega = (f,\varphi_h)_\Omega \quad \forall \varphi_h \in V_h. \tag{2.21}
\]

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

Since there holds $\tilde{\varepsilon}_{upw} > \tilde{\varepsilon}_*$ 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 $\tilde{\varepsilon}$ 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).
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 $I_j \in T_h$ a bubble function $b_j \in L^2(\Omega)$ is any function satisfying $b_j|_{I_j} \in H^1_0(I_j)$ and $\text{supp}\{b_j\} = T_j$. Then the space of bubble functions is defined as $B = \text{span}\{b_j, 1 \leq j \leq N\}$. 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

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

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

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 $\partial I_j$, we have

$$(u'_L, b_j)_{I_j} = [u'_L b_j^x]_{x_j-1} - (u'_L, b_j)_{I_j} = 0 \quad \text{and} \quad (2.24)$$

$$(b u'_B, b_j)_{I_j} = \sum_{i=1}^N c_i (b b'_i, b_j)_{I_j} = c_j (b b'_j, b_j)_{I_j} = \frac{c_j}{2} [b b'_j]_{x_j-1}^2 = 0, \quad (2.25)$$

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

$$\varepsilon(u'_B, b'_j)_{I_j} + (b u'_L, b_j)_{I_j} = (f, b_j)_{I_j} \quad \text{for all } j \in \{1, 2, \ldots, N\}. \quad (2.26)$$

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

$$c_j = \frac{(1, b_j)_{I_j}}{\varepsilon |b_j|^2_{L^2(I_j)}} (f - b u'_L|_{I_j}). \quad (2.27)$$
Figure 2.4: A comparison of the discrete solutions obtained by adding artificial diffusion. Each dashed curve corresponds to a different partition of $\Omega$ 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

\[
(u_B', \varphi_L')_\Omega = \sum_{j=1}^{N} c_j (b_j', \varphi_L')_{I_j} = 0 \quad \text{and} \quad \tag{2.28}
\]

\[
(b u_B', \varphi_L)\Omega = \sum_{j=1}^{N} c_j b (b_j', \varphi_L)_{I_j} = -\sum_{j=1}^{N} c_j b \varphi_L'_{I_j} (b_j, 1)_{I_j} = \]

\[
= \sum_{j=1}^{N} \frac{1}{|I_j|} \frac{(b_j, 1)_{I_j}^2}{|b_j| |I_j|} (bu_L' - f, b\varphi_L')_{I_j} = \sum_{j=1}^{N} \tau_j^B (bu_L' - f, b\varphi_L')_{I_j}. \tag{2.29}
\]

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

\[
\varepsilon(u_L', \varphi_L')_\Omega + (bu_L', \varphi_L)\Omega + \sum_{j=1}^{N} \tau_j^B (bu_L' - f, b\varphi_L')_{I_j} = (f, \varphi_L)\Omega. \tag{2.30}
\]

The formulation (2.30) is equivalent to the SUPG formulation (2.12) and the stabilization parameters $\tau_j^B$ 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

\[
\tau_j^{B_i} = \frac{1}{|I_j|} \frac{(b_j, 1)_{I_j}^2}{|b_j| |I_j|} = \frac{1}{h} \left( \frac{1}{3} h^3 \right)^2 = \frac{h^2}{12 \varepsilon}. \tag{2.31}
\]
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{align*}
-\varepsilon b_j'' + b b_j' &= 1 \quad \text{in } I_j, \quad (2.32) \\
b_j &= 0 \quad \text{on } \partial I_j. \quad (2.33)
\end{align*}

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

\begin{align*}
\tau_{j}^{B_2} = \frac{1}{|I_j|} \frac{(b_j, 1)^2_{I_j}}{\varepsilon \|b_j\|^2_{L_2(I_j)}} = \frac{1}{h} (b_j, 1)_{I_j} = \\
= \frac{1}{h} \left( \frac{h}{2} - \frac{h}{2Pe} \left( 1 - \exp(-2Pe) \right) - h \exp(-2Pe) \right) = \\
= \frac{h}{2b} - \frac{\varepsilon}{b^2} + \frac{h}{b} \frac{1}{\exp(2Pe) - 1} = \frac{h \exp(2Pe) + 1}{2b \exp(2Pe) - 1} - \frac{\varepsilon}{b^2} = \frac{h}{2b} \coth(Pe) - \frac{\varepsilon}{b^2}. \quad (2.34)
\end{align*}

Moreover, in this case, the discrete solution not only is nodally exact, but also coincides with the exact solution everywhere in $\Omega$ (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

\begin{align*}
\tau_{j}^{B_2} = \frac{h}{2b} \left( \frac{\coth(Pe) - 1}{Pe} \right) = \frac{h}{2b} \left( \frac{Pe}{3} + O(\text{Pe}^3) \right) = \frac{h^2}{12\varepsilon} \left( 1 + O(\text{Pe}^2) \right), \quad (2.34)
\end{align*}

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

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 $Lw = -\varepsilon w'' + bw'$

\begin{align*}
L^*w &= -\varepsilon w'' - (bw'). \quad (2.35)
\end{align*}

Thus, if $v, w \in H^1_0(\Omega) \cap H^2(\Omega)$, the following identity holds

\begin{align*}
\int_{\Omega} (Lv)w \, dx &= \int_{\Omega} v(L^*w) \, dx. \quad (2.36)
\end{align*}

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{align*}
L^* g_j &= 0 \quad \text{in } I_j \cup I_{j+1} \quad (2.37) \\
g_j(x_{j-1}) = g_j(x_{j+1}) &= 0 \quad \text{and} \quad (2.38) \\
\varepsilon \left[ g_j'(x_{j-1}) - g_j'(x_{j+1}) \right] &= 1. \quad (2.39)
\end{align*}
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
\[
\int_{x_{j-1}}^{x_{j+1}} (Lu)g_j \, dx = \int_{x_{j-1}}^{x_{j+1}} fg_j \, dx. \tag{2.40}
\]
Using the integration by parts the left-hand side of (2.40) can now be rewritten in the form
\[
\int_{x_{j-1}}^{x_{j+1}} (Lu)g_j \, dx = \int_{x_{j-1}}^{x_{j+1}} (-\varepsilon u'' + bu')g_j \, dx + \int_{x_j}^{x_{j+1}} (-\varepsilon u'' + bu')g_j \, dx =
\]
\[
= [-\varepsilon u'g_j + bug_j]_{x_{j-1}}^{x_j} + \int_{x_{j-1}}^{x_j} -\varepsilon u'g_j' + u(bg_j)' \, dx +
\]
\[
+ [-\varepsilon u'g_j + bug_j]_{x_j}^{x_{j+1}} - \int_{x_j}^{x_{j+1}} -\varepsilon u'g_j' + u(bg_j)' \, dx.
\]

The property (2.38) together with the continuity of \(-\varepsilon u'g_j + bug_j\) at the node \( x_j \) then yields
\[
[-\varepsilon u'g_j + bug_j]_{x_{j-1}}^{x_j} + [-\varepsilon u'g_j + bug_j]_{x_j}^{x_{j+1}} = 0. \tag{2.41}
\]
Thus, we obtain
\[
\int_{x_{j-1}}^{x_{j+1}} (Lu)g_j \, dx = \int_{x_{j-1}}^{x_j} \varepsilon u'g_j' - u(bg_j)' \, dx + \int_{x_j}^{x_{j+1}} \varepsilon u'g_j' - u(bg_j)' \, dx =
\]
\[
= \left[ \varepsilon u g_j \right]_{x_{j-1}}^{x_j} - \int_{x_{j-1}}^{x_j} \varepsilon u g_j'' \, dx - \int_{x_{j}}^{x_{j+1}} u(bg_j)' \, dx +
\]
\[
+ \left[ \varepsilon u g_j' \right]_{x_j}^{x_{j+1}} - \int_{x_j}^{x_{j+1}} \varepsilon u g_j'' \, dx - \int_{x_j}^{x_{j+1}} u(bg_j)' \, dx =
\]
\[
= -\varepsilon g_j'(x_{j-1})u(x_{j-1}) + u(x_j) + \varepsilon g_j'(x_{j+1})u(x_{j+1}), \tag{2.42}
\]
where we used the property (2.39) and the fact that \(-\int_{x_{j-1}}^{x_j+1} u(\varepsilon g_j'' + (bg_j)') \, dx = f_{x_{j-1}}^{x_{j+1}} u(L^*g_j) \, dx = 0\) by (2.37).

Consequently, the equation (2.40) changes to
\[
-\varepsilon g_j'(x_{j-1})u_{j-1} + u_j + \varepsilon g_j'(x_{j+1})u_{j+1} = \int_{x_{j-1}}^{x_{j+1}} fg_j \, dx, \tag{2.43}
\]
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 \( g_j \)

\[
g_j(x) = \begin{cases} 
1 \frac{1 - \exp \left( \frac{-b}{\varepsilon} (x - x_{j-1}) \right) }{1 + \exp \left( -2Pe \right)} & \text{for } x \in [x_{j-1}, x_j], \\
1 \frac{\exp \left( \frac{b}{\varepsilon} (x_{j+1} - x) \right) - 1 }{\exp (2Pe) + 1} & \text{for } x \in [x_j, x_{j+1}], \\
0 & \text{otherwise},
\end{cases} \tag{2.44}
\]
Using this expression, we can evaluate all the terms containing \( g_j \) in the scheme \( (2.43) \) and obtain

\[
- \frac{e^{2Pe}}{e^{2Pe}+1} u_{j-1} + u_j - \frac{1}{e^{2Pe}+1} u_{j+1} = h \frac{e^{2Pe} - 1}{b} e^{2Pe} + f. \tag{2.45}
\]

This can be rearranged to the equivalent form

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

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

\[
\frac{b}{2h} \frac{e^{2Pe}}{e^{2Pe} - 1}(-u_{j-1} + 2u_j - u_{j+1}) + \frac{b}{2h} \frac{e^{2Pe} - 1}{e^{2Pe} - 1} (u_{j+1} - u_{j-1}) = f. \tag{2.47}
\]

Since \( \frac{e^{2Pe}+1}{e^{2Pe}+1} = \coth Pe \), 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 \( \Omega \) 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

\[
\begin{align*}
p_i(x_i - h) + q_i x_i + r_i (x_i + h) &= (L(1))_i = 0, \\
p_i \left( e^{bx_i} e^{-2Pe} \right) + q_i \left( e^{bx_i} \right) + r_i \left( e^{bx_i} e^{2Pe} \right) &= (L(e^{bx_i}))_i = 0,
\end{align*}
\tag{2.49}
\]

and can be rewritten in the form

\[
\begin{align*}
p_i + q_i + r_i &= 0, \\
-h p_i + 0 + hr_i &= b, \\
e^{-2Pe} p_i + q_i + e^{2Pe} r_i &= 0.
\end{align*}
\tag{2.50}
\]
The solution of this system of linear equations takes the form

\[
(p_i, q_i, r_i) = \left( -\frac{b}{h} \frac{e^{2Pe}}{e^{2Pe} - 1}, \frac{b}{h} \frac{e^{2Pe} + 1}{e^{2Pe} - 1}, -\frac{b}{h} \frac{1}{e^{2Pe} - 1} \right),
\]

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 \(\bar{b}_i\). Let us also mention that for \(Pe \to \infty\) we obtain \((p, q, r) \approx \frac{b}{h}(-1, 1, 0)\), which is the backward Euler method, whereas for \(Pe \to 0\) 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

\[
(p, q, r, s, t) = \frac{b}{2h} \left( \frac{e^{12Pe}(e^{12Pe+1})}{(e^{12Pe} - 1)^3}, \frac{-4e^{12Pe}(e^{12Pe+1})}{(e^{12Pe} - 1)^3}, \frac{3(e^{12Pe+1})(e^{12Pe+1})}{(e^{12Pe} - 1)^3}, \frac{-4(e^{12Pe+1})}{(e^{12Pe} - 1)^3}, \frac{e^{12Pe+1}}{(e^{12Pe} - 1)^3} \right) + \frac{\varepsilon}{h^2} \left( \frac{-e^{12Pe}(e^{12Pe+1})}{(e^{12Pe} - 1)^3}, \frac{2e^{12Pe}(e^{12Pe+1})}{(e^{12Pe} - 1)^3}, \frac{-e^{12Pe+4e^{12Pe+1}}}{(e^{12Pe} - 1)^3}, \frac{2(e^{12Pe+1})}{(e^{12Pe} - 1)^3}, \frac{-1}{(e^{12Pe} - 1)^3} \right).
\]

If we again compute the approximation for \(Pe \to \infty\), 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 \(Pe \to 0\) yields \((p, q, r, s, t) \approx \frac{\varepsilon}{h^2} \left( \frac{1}{12}, -\frac{1}{3}, \frac{1}{2}, -\frac{4}{3}, \frac{1}{12} \right) + \frac{b}{h} \left( \frac{1}{12}, \frac{1}{3}, 0, \frac{2}{3}, -\frac{1}{12} \right)\), 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. Rivière [2008] or Dolejsí 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 (\(\varepsilon = 10^{-6}\))

\[
-\varepsilon u'' + u' = 2x \quad \text{in } (0, 1),
\]

\[
u(0) = u(1) = 0.
\]
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

\[
\begin{align*}
u(x) &= x^2 + 2\varepsilon x - (1 + 2\varepsilon) \frac{e^{x/\varepsilon} - 1}{e^{1/\varepsilon} - 1}, \\
u_{upw}^k &= (kh)^2 + (2\varepsilon + h)kh - (1 + 2\varepsilon + h) \frac{(1 + \frac{h}{\varepsilon})^k - 1}{(1 + \frac{h}{\varepsilon})^N - 1}, \\
u_{IAS}^k &= (kh)^2 + kh^2 \frac{e^{h/\varepsilon} + 1}{e^{h/\varepsilon} - 1} - \left(1 + h \frac{e^{h/\varepsilon} + 1}{e^{h/\varepsilon} - 1}\right) \frac{e^{kh/\varepsilon} - 1}{e^{1/\varepsilon} - 1}, \end{align*}
\]

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 \(\varepsilon\)), i.e. for fixed \(\varepsilon\) 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 \(N \approx \varepsilon^{-j+1}\), i.e. for \(h \approx \varepsilon^{-j+1}\).

![Figure 2.6: Comparison of convergence of the simple upwind scheme and the Il'in-Allen-Southwell scheme.](image)

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.

\[
\begin{align*}
-\varepsilon u'' + b(x)u' &= f(x) \quad \text{in } \Omega = (0, 1), \\
u(0) = u(1) &= 0.
\end{align*}
\]
Denoting $u_j = u_i(x_j)$, 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

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

(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

$v(x) = v_0 \left(1 - \frac{x}{\varepsilon}\right) = -u_0(1) \left(-b(1) \frac{1-x}{\varepsilon}\right)$

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

$$z = u - v = S + E - v.$$ 

(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

$$L_h^* v_h = R_h(Lv), \quad (v_h)_0 = v(0), (v_h)_N = v(1),$$

(2.61)

$$L_h^* z_h = R_h(Lz), \quad (z_h)_0 = z(0), (z_h)_N = z(1),$$

(2.62)

where $R_h : C(\Omega) \to \mathbb{R}^{N+1}$ 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

$$\sinh(x) \geq \frac{xe^x}{2(x+1)}, \quad \text{for } x > 0,$$

(2.63)

$$|\sinh(x) - x| \leq \frac{|x|^3 e^x}{6(1 + x^2)}.$$ 

(2.64)

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

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

(2.65)

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

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

(2.66)

where we used the fact that the function $\frac{x^2}{1+x^2}$ is increasing on $(0, +\infty)$. 46
For the case $0 \leq x < 4$ 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

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

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

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

The second inequality in (2.68) comes from the estimate

$$
x^3 \left(1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120}\right) - 6 \left(1 + x^2\right) \left(\frac{x^3}{6} + C_\sigma x^5\right) =
$$

$$
= \left[1 - \left(3C_\sigma + \frac{1}{4}\right)x\right]^2 x^4 + C_T x^6 + \left[30 \left(6C_\sigma - \frac{1}{24}\right) - \frac{x^2}{2}\right] x^6 \geq 0,
$$

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

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 $\beta_1$ and let $h^*_0 \in \left(0, \frac{b(1) - \beta_1}{\beta_1}\right)$ be a positive real number. Then there exists a positive constant $C_z$ independent of $h$ and $\varepsilon$ such that for all $h < h^*_0$ there holds

$$
\left| R_h(Lv) - L^*_h R_h v \right| \leq \frac{C_z h^2}{\varepsilon(\varepsilon + h)} R_h \exp \left(-\frac{\beta_1}{\varepsilon} \left|x\right|\right). \tag{2.69}
$$

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

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

$$
(L^*_h R_h v)(x) = -\frac{2b(x) \sinh(Pe(1))}{h \sinh(Pe(x))} \left(\sinh(Pe(1) - Pe(x))\right) v(x). \tag{2.71}
$$

Further, let us denote $q = Pe(1)$, $p = Pe(x_j)$, $S(x) = \sinh x - x$ 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

\[
\left| (Lv)(x_j) - (L_h^b R_h v)(x_j) \right| = \\
= \frac{4 \varepsilon}{h^2} \left| -q(q-p) + p \frac{\sinh(q)}{\sinh(p)} \sinh(q-p) \right| |v(x_j)| = \\
= \frac{4 \varepsilon}{h^2} \left| p(q-p) S(q) - q(q-p) S(p) + p S(q) S(p - q) \right| \frac{|v(x_j)|}{\sinh(p)} \leq \\
\leq \frac{4 \varepsilon p q |q-p|}{6 b h^2 \sinh(p)} \left[ \psi(q) e^q + \psi(p) e^p + \psi(|q-p|) e^{q-p} (1 + \psi(q) e^q) \right] |v(x_j)| \leq \\
\leq \frac{2 \varepsilon p q |q-p|}{3 h^2} 2 (1+p) \frac{4 \max_{s \in [p,q]} \{|\psi(s)\} e^{q-p} |v(x_j)|.}{3 w^0 (2 \varepsilon + \|b\|_{\infty} h) \exp \left( \frac{h}{\varepsilon} \beta_1 |1-x_j| \right)} |v(x_j)| \leq \\
\leq 4 \beta_1 |1-x_j| \frac{3 h^3 (2 \varepsilon)^2}{4 \varepsilon^2 + \|b\|_{\infty}^2 h^2} \exp \left( \frac{h}{\varepsilon} \beta_1 |1-x_j| \right) |v(x_j)|. 
\]

(2.72)

Since the function \( b \) is Lipschitz-continuous with a constant \( \beta_1 \), we can estimate the difference \(|q-p|\) by

\[
|q-p| \leq \frac{\beta_1 h}{2 \varepsilon} |1-x_j|. 
\]

(2.73)

Consequently, for the consistency error of the function \( v \) there holds

\[
\left| (Lv)(x_j) - (L_h^b R_h v)(x_j) \right| \leq \\
\leq 2 \varepsilon h^2 \|b\|_{\infty} \beta_1 |1-x_j| \frac{4 (2 \varepsilon + \|b\|_{\infty} h)}{4 \varepsilon^2 + \|b\|_{\infty} h} \exp \left( \frac{h}{\varepsilon} \beta_1 |1-x_j| \right) |v(x_j)| \leq \\
\leq 4 \beta_1 |1-x_j| \frac{3 h^3 (2 \varepsilon)^2}{4 \varepsilon^2 + \|b\|_{\infty}^2 h^2} \exp \left( \frac{h}{\varepsilon} \beta_1 |1-x_j| \right) |v(x_j)|. 
\]

(2.74)

Now we rewrite the last two factors in the form

\[
|u_0(1)| \exp \left( \left( h \beta_1 - (b(1) - \beta) \right) \frac{|1-x_j|}{\varepsilon} \right) \exp \left( -\frac{\beta}{\varepsilon} |1-x_j| \right). 
\]

(2.75)

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

\[
\exp \left( \left( h \beta_1 - (b(1) - \beta) \right) \frac{|1-x_j|}{\varepsilon} \right) < \frac{\varepsilon}{|1-x_j| (b(1) - \beta - h_0^* \beta_1)}. 
\]

(2.76)

Consequently for the consistency error there holds

\[
|R_h(Lv) - L_h^b R_h v| \leq \frac{C_z h^2}{\varepsilon (\varepsilon + h)} R_h \exp \left( -\frac{\beta}{\varepsilon} \frac{|1-x|}{\varepsilon} \right), 
\]

(2.77)

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

Let us notice that the quality of the above derived estimate depends on the bound \( \beta \). If it is too small, then the exponential decay is very slow. On the other hand, if \( \beta \) 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 \( L_h^s \) 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
\[
|(Lz)(x_j) - (L_h^s R_h z)(x_j)| \leq C_S (\varepsilon + 2\|b\|_\infty) h + \frac{C_E}{\beta} (1 + 2\|b\|_\infty) \sinh \left( \frac{\beta h}{\varepsilon} \right) \exp \left( -\frac{\beta}{\varepsilon} (1 - t_x) \right). (2.78)
\]

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.
\[
|(Lz)(x_j) - (L_h^s R_h z)(x_j)| = \frac{1}{2} \int_{x-h}^{x+h} |z^{(3)}(t)| + |b(t)||z^{(2)}(t)| + \|b(t)||z^{(2)}(t)| \, dt,
\]
where we used the expressions
\[
D_h^2 z(x) = z''(x) + \frac{1}{2h^2} \left[ \int_{x-h}^{x+h} (x + h - t)^2 z^{(3)}(t) \, dt - \int_{x-h}^{x} (x - h - t)^2 z^{(3)}(t) \, dt \right],
\]
\[
D_h^2 z(x) = \frac{1}{h^2} \int_{x-h}^{x+h} u''(s) \, ds \, dt,
\]
\[
D_h^4 z(x) = z'(x) + \frac{1}{2h} \left[ \int_{x-h}^{x+h} (x + h - t) z^{(2)}(t) \, dt + \int_{x-h}^{x} (x - h - t) z^{(2)}(t) \, dt \right]
\]
resulting in estimates
\[
|D_h^2 z(x) - z''(x)| \leq \frac{1}{2h^2} \left[ \int_{x-h}^{x+h} |z^{(3)}(t)| \, dt + h^2 \int_{x-h}^{x} |z^{(3)}(t)| \, dt \right] = \frac{1}{2} \int_{x-h}^{x+h} |z^{(3)}(t)| \, dt,
\]
\[
|D_h^2 z(x)| \leq \frac{1}{h^2} \int_{x-h}^{x+h} |z''(t)| \, dt \leq \frac{1}{h} \int_{x-h}^{x+h} |z''(t)| \, dt,
\]
\[
|D_h^4 z(x) - z'(x)| \leq \frac{1}{2h} \left[ h \int_{x-h}^{x+h} |z^{(2)}(t)| \, dt + h \int_{x-h}^{x} |z^{(2)}(t)| \, dt \right] = \frac{1}{2} \int_{x-h}^{x+h} |z^{(2)}(t)| \, dt.
\]
Since \( z = S + (E - v) \), we can use Lemma 1.2.1 (page 14) and estimate the derivatives of \( z \) by
\[
|z^{(j)}(t)| \leq C_S + C_E \varepsilon^{1-j} \exp \left( -\frac{\beta}{\varepsilon} (1 - t_x) \right), \tag{2.80}
\]
where the constants \( C_S \) and \( C_E \) are independent from \( \varepsilon \) and \( h \). This estimate contains a factor \( \varepsilon^{1-j} \) instead of \( \varepsilon^j \) which is caused by the fact that the layer part of the S-decomposition of the function \( z \) begins with \( 0.\varepsilon^0 \). Consequently
\[
|(Lz)(x_j) - (L_h^s R_h z)(x_j)| \leq C_S (\varepsilon + 2\|b\|_\infty) h + C_E \left( \frac{1}{2} + \|b\|_\infty \right) \int_{x-h}^{x+h} \exp \left( -\frac{\beta}{\varepsilon} (1 - t_x) \right) \, dt = C_S (\varepsilon + 2\|b\|_\infty) h + \frac{C_E}{\beta} (1 + 2\|b\|_\infty) \sinh \left( \frac{\beta h}{\varepsilon} \right) \exp \left( -\frac{\beta}{\varepsilon} (1 - x_j) \right).
\]

\[\square\]
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 \( b(x) > \beta > 0 \). Then for any function \( \exp \left( \frac{ax}{\varepsilon} \right) \), \( 0 < \alpha < \beta \), there exists positive constant \( C_\alpha \) (independent of \( h \) and \( \varepsilon \)) such that

\[
L^*_h R_h \exp \left( \frac{ax}{\varepsilon} \right) \geq \frac{C_\alpha}{\max \{ h, \varepsilon \}} R_h \exp \left( \frac{ax}{\varepsilon} \right). \tag{2.81}
\]

**Proof.** At first, exact computation gives

\[
L^*_h R_h \exp \left( \frac{ax}{\varepsilon} \right) = 2R_h \left\{ \exp \left( \frac{ax}{\varepsilon} \right) \frac{b(x)}{h} \frac{\alpha h}{2 \varepsilon} \sinh \left( \frac{b(x) - \alpha h}{2 \varepsilon} \right) \right\}. \tag{2.82}
\]

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

\[
L^*_h R_h \exp \left( \frac{ax}{\varepsilon} \right) \geq 2R_h \left\{ \exp \left( \frac{ax}{\varepsilon} \right) \frac{\beta}{h} \frac{\alpha h}{2 \varepsilon} \sinh \left( \frac{b(x) - \alpha h}{2 \varepsilon} \right) \right\} \geq \frac{\beta \alpha (\beta - \alpha)}{2 \varepsilon \sinh \left( \frac{h}{2} \| b \|_\infty \right) R_h \exp \left( \frac{ax}{\varepsilon} \right)}. \tag{2.83}
\]

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

\[
L^*_h R_h \exp \left( \frac{ax}{\varepsilon} \right) \geq 2R_h \left\{ \exp \left( \frac{ax}{\varepsilon} \right) \frac{\sinh \left( \frac{\alpha h}{2 \varepsilon} \right)}{h} \frac{\alpha h}{2 \varepsilon} \sinh \left( \frac{b(x) - \alpha h}{2 \varepsilon} \right) \exp \left( \frac{b(x) - \alpha h}{2 \varepsilon} \right) \right\} \geq \frac{4 \beta \sinh \left( \frac{\alpha h}{2 \varepsilon} \right) \sinh \left( \frac{\beta - \alpha h}{2 \varepsilon} \right)}{h \exp \left( \frac{1}{2} \| b \|_\infty \right) R_h \exp \left( \frac{ax}{\varepsilon} \right)}. \tag{2.84}
\]

Thus the constant \( C_\alpha \) is given by

\[
C_\alpha = \min \left\{ \frac{\beta \alpha (\beta - \alpha)}{2 \sinh \left( \frac{h}{2} \| b \|_\infty \right)}, \frac{4 \beta \sinh \left( \frac{\alpha h}{2 \varepsilon} \right) \sinh \left( \frac{\beta - \alpha h}{2 \varepsilon} \right)}{\exp \left( \frac{1}{2} \| b \|_\infty \right)} \right\}. \tag{2.85}
\]

\( \square \)

**Remark 2.2.1.** The constant \( C_\alpha \) from Lemma 2.2.4 vanishes for \( \alpha = \beta \) and thus the stability of the method is lost. In fact, the difference \( \beta - \alpha \) that occurs in the definition of the constant \( C_\alpha \) is an estimate for the difference \( b(x) - \alpha \), and thus if \( b(x) > \beta \), 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 \( \|(L^*_h)^{-1}\|_{\infty,d} \leq 1/\beta \).

Proof. Since we would like to employ the M-criterion, we firstly examine the sign of the entries of the tridiagonal matrix \( L^*_h \):

\[
(L^*_h)_{j,j} = \frac{b(x_j)}{h} \coth(\text{Pe}(x_j)) > 0, \quad (2.86)
\]

\[
(L^*_h)_{j,j+1} = \frac{b(x_j)}{2h} (1 - \coth(\text{Pe}(x_j))) < 0, \quad (2.87)
\]

\[
(L^*_h)_{j,j-1} = -\frac{b(x_j)}{2h} (1 + \coth(\text{Pe}(x_j))) < 0. \quad (2.88)
\]

Secondly, the vector \( e_h = R_hx \) is positive inside \( \Omega \) and satisfies \( L^*_h e_h = R_h b(x) > 0 \). Thus the matrix \( L^*_h \) is an M-matrix and it holds

\[
\|(L^*_h)^{-1}\|_{\infty,d} \leq \frac{\|e_h\|_{\infty,d}}{\min_k (L^*_h e_h)_k} \leq \frac{1}{\beta}. \quad (2.89)
\]

\[\square\]

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 \( \tilde{C} \) (independent of \( h \) and \( \varepsilon \)) 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

\[
\|R_h u - u^*_h\|_{\infty,d} \leq \tilde{C} h. \quad (2.90)
\]

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

\[
|L^*_h(R_h u - u^*_h)| = |L^*_h R_h u - L^*_h u^*_h| = |L^*_h R_h(z + v) - L^*_h(z_h + v_h)| \leq |L^*_h R_h z - L^*_h z_h| + |L^*_h R_h v - L^*_h v_h|. \quad (2.91)
\]

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

\[
L^*_h(R_h v - v_h) = L^*_h R_h v - R_h(Lv) \leq \frac{C_z h^2}{\varepsilon(\varepsilon + h)} R_h \exp \left(-\frac{\beta}{\varepsilon} \frac{1-x}{\varepsilon} \right) \leq \leq \frac{C_z h^2}{\varepsilon(\varepsilon + h)} R_h \exp \left(-\frac{\beta}{\varepsilon} \frac{1-x}{\varepsilon} \right) \leq \leq \frac{\max\{h,\varepsilon\} C_z h^2}{\varepsilon(\varepsilon + h)} \exp \left(-\frac{\alpha}{\varepsilon} \right) L^*_h R_h \exp \left(\frac{\alpha}{\varepsilon} \right). \quad (2.92)
\]

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

\[
|R_h v - v_h| \leq \frac{\max\{h,\varepsilon\} C_z h^2}{\varepsilon(\varepsilon + h)} R_h \exp \left(-\frac{\alpha}{\varepsilon} \right). \quad (2.93)
\]
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
\[
\|v_h - R_h v\|_{\infty,d} \leq \frac{C_z h^2}{C_\alpha (\varepsilon + h)} \frac{h}{\varepsilon} \exp\left(-\frac{h}{\varepsilon}\right) \leq \frac{C_z h}{C_\alpha} \frac{h}{\varepsilon + h} \frac{1}{\alpha} \leq \frac{C_z h}{\alpha C_\alpha}. \tag{2.95}
\]

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

Again, applying the discrete comparison principle gives
\[
|R_h z - z_h| \leq C_S (\varepsilon + 2\|b\|_{\infty}) h (L_h^*)^{-1} R_h 1 + \frac{\max\{h, \varepsilon\} C_E}{C_\alpha} \frac{1}{\beta} \left(1 + 2\|b\|_{\infty}\right) \sinh \left(\frac{\beta h}{\varepsilon}\right) \exp\left(-\frac{\alpha}{\varepsilon}\right) \frac{1 - x}{\varepsilon}. \tag{2.97}
\]

Since $h \leq \varepsilon$ we can use the inequality $\sinh \left(\frac{\beta h}{\varepsilon}\right) \leq \frac{h}{\varepsilon} \sinh(\beta)$ and find out that
\[
\|R_h z - z_h\|_{\infty,d} \leq \left\{C_S (\varepsilon + 2\|b\|_{\infty}) \frac{1}{\beta} + \frac{C_E}{C_\alpha \beta} \left(1 + 2\|b\|_{\infty}\right) \sinh(\beta)\right\} h. \tag{2.97}
\]

Conversely, when $h \geq \varepsilon$ we apply the inequality (remind that $\alpha \in (0, \beta)$)
\[
\sinh \left(\frac{\beta h}{\varepsilon}\right) \exp\left(-\frac{\beta}{\varepsilon}(1 - x_j)\right) \leq \frac{1}{2} \exp\left(\frac{\beta h}{\varepsilon}\right) \exp\left(-\frac{\beta}{\varepsilon}(1 - x_j)\right) = \frac{1}{2} \exp\left(-\frac{\beta}{\varepsilon}(1 - x_{j+1})\right) \leq \frac{1}{2} \exp\left(-\frac{\alpha}{\varepsilon}(1 - x_{j+1})\right) = \frac{1}{2} \exp\left(-\frac{\alpha}{\varepsilon}(1 - x_{j+1})\right). \tag{2.98}
\]

Lemmas 2.2.3 and 2.2.4 together with this inequality then provide the estimate
\[
L_h^*(R_h z - z_h) = L_h^* R_h z - R_h(Lz) \leq C_S (\varepsilon + 2\|b\|_{\infty}) h L_h^*(L_h^*)^{-1} R_h 1 + \frac{\max\{h, \varepsilon\} C_E}{C_\alpha} \left(1 + 2\|b\|_{\infty}\right) \exp\left(\frac{\alpha h}{\varepsilon}\right) \exp\left(-\frac{\alpha}{\varepsilon}\right) L_h^* R_h \exp\left(\alpha\frac{x}{\varepsilon}\right). \tag{2.99}
\]
Let us emphasize that instead of the factor \(\sinh(\beta h/\varepsilon)\) this estimate contains \(\exp(\alpha h/\varepsilon)\) 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
\[
|R_hz - z_h| \leq C_S(\varepsilon + 2\|b\|_\infty)h(L_h^-)R_h1 + \max\{h, \varepsilon\} C_E\frac{1}{\beta} \left(\frac{1}{2} + \|b\|_\infty\right) \exp\left(\frac{\alpha h}{\varepsilon}\right) R_h \exp\left(-\alpha \frac{1 - x}{\varepsilon}\right),
\]
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
\[
\|R_hz - z_h\|_{\infty,d} \leq \left\{C_S(\varepsilon + 2\|b\|_\infty)\frac{1}{\beta} + \frac{C_E}{C_\alpha \beta} \left(\frac{1}{2} + \|b\|_\infty\right)\right\} h. \quad (2.100)
\]

If we combine all previous estimates we get
\[
\|R_hu - u_h\|_{\infty,d} \leq \|R_hz - z_h\|_{\infty,d} + \|R_hv - v_h\|_{\infty,d} \leq \hat{C}h, \quad (2.101)
\]
where (we can take e.g. \(\alpha = \beta/2\))
\[
\hat{C} = \frac{C_z}{C_\alpha} \max\left\{\frac{1}{2}, \frac{1}{\alpha}\right\} + \frac{3C_S\|b\|_\infty}{\beta} + \frac{C_E(1 + 2\|b\|_\infty)}{C_\alpha \beta} \max\left\{\frac{1}{2}, \sinh(\beta)\right\}. \quad (2.102)
\]
3. Modified SUPG method on convection-oriented meshes

3.1 Introduction and the idea of the method

Let us solve the convection-diffusion equation

\[-\varepsilon \Delta u(x) + b(x) \cdot \nabla u(x) = f(x) \quad \text{in } \Omega \subset \mathbb{R}^n,\]
\[u(x) = 0 \quad \text{on } \partial \Omega,\]

where \( n \in \mathbb{N}, \Omega \) is a polytopic domain with Lipschitz-continuous boundary \( \partial \Omega \), \( b \in W^{1,\infty}(\Omega)^n \) 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

\[\Gamma_+ = \{ x \in \partial \Omega, \ b(x) \cdot n(x) > 0 \}, \]
\[\Gamma_0 = \{ x \in \partial \Omega, \ b(x) \cdot n(x) = 0 \}, \]
\[\Gamma_- = \{ x \in \partial \Omega, \ b(x) \cdot n(x) < 0 \}, \]

satisfying \( \partial \Omega = \Gamma_+ \cup \Gamma_0 \cup \Gamma_- \) and \( \Gamma_+ \cap \Gamma_0 = \Gamma_0 \cap \Gamma_- = \Gamma_- \cap \Gamma_+ = \emptyset \). Here, the vector \( n(x) \) 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 \( \Gamma_+ \) 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 \( \Gamma_0 \), but also along interior layers that propagate from discontinuous boundary conditions at \( \Gamma_- \).

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 \( b \). 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 \( \varepsilon \) 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 \( -b(x) \). 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 \( b \). Then, if \( b_K \) is a constant approximation of \( b \) 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 \( b_K \cdot \nabla \lambda_{K,i}, i \in \{1, 2, \ldots, n+1\} \), 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 \(\varphi \in H^1_0(\Omega)\) and integrate over the whole domain \(\Omega\). Using Green’s theorem (Theorem 4.1.1, page 124) the weak formulation of \(3.1\) reads:

Find \(u \in H^1_0(\Omega)\) such that

\[
\varepsilon(\nabla u, \nabla \varphi) + (b \cdot \nabla u, \varphi) = (f, \varphi) \quad \forall \varphi \in H^1_0(\Omega).
\]

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

To derive the Galerkin finite element discretization of \(3.1\), we define a finite element space \(X_h = X_h^{(1)} = \{v_h \in C(\Omega), v_h|_K \in P_1(K), \forall K \in T_h\}\) and a space of test functions \(V_h = V_h^{(1)} = X_h \cap H^1_0(\Omega)\). The barycentric coordinates \(\{\lambda_{K,j}\}_{j=1}^{n+1}\) of the element \(K \in T_h\) then form a basis of \(P_1(K)\) and we reorder them so that

\[
\int_K b \cdot \nabla \lambda_{K,j} \| \nabla \lambda_{K,j} \| \, dx \leq \int_K \frac{b \cdot \nabla \lambda_{K,j+1}}{\| \nabla \lambda_{K,j+1} \|} \, dx, \quad \text{for } j = 1, 2, \ldots, n.
\]

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

Further, we assume that the barycentric coordinates \(\{\lambda_{K,j}\}_{j=1}^{n+1}\) satisfy for each \(K \in T_h\) the inequality

\[
(\nabla \lambda_{K,j}, \nabla \lambda_{K,i})_K \leq 0 \quad \text{whenever } i \neq j.
\]

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

The SUPG method adds weighted residuals \(R(u) = -\varepsilon \Delta u + b \cdot \nabla u - f\) 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 \(K \in T_h\) by derivative of \(v\) in the direction \(P_{K,n+1} - C_K\) (see Lamác (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 \(1 \leq i, j \leq n + 1\).

Thus, the solution \(u \in H^1_0(\Omega) \cap H^2(\Omega)\) of the problem \(3.6\) satisfies also for all \(\varphi \in H^1_0(\Omega)\)

\[
a(u, \varphi) = F(\varphi),
\]

55
where
\[ F(\varphi) = \sum_{K \in \mathcal{T}_h} (f, \varphi + (P_{K,n+1} - C_K) \nabla \varphi)_K \quad \text{and} \quad (3.10) \]
\[ a(u, \varphi) = \varepsilon(\nabla u, \nabla \varphi)_\Omega + (\mathbf{b} \cdot \nabla u, \varphi)_\Omega + \sum_{K \in \mathcal{T}_h} \left( -\varepsilon \Delta u + \mathbf{b} \cdot \nabla u, (P_{K,n+1} - C_K) \nabla \varphi \right)_K. \quad (3.11) \]

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 \( d_{K,j} = P_{K,n+1} - P_{K,j} \), \( j = 1, 2, \ldots, n \) and consider the element-wise constant approximation \( b_K \) of the vector field \( \mathbf{b} \) by vectors that are parallel with \( d_{K,1} \) on each element \( K \). More precisely, first of all we consider that our mesh is ”well-aligned” with respect to the vector field \( \mathbf{b} \) and then on each element \( K \) we construct a constant approximation \( b_K \) of \( \mathbf{b} \). This ”well-alignment” is provided by the following assumptions.

\( (A1) \) The ordering given by (3.7) on each \( K \in \mathcal{T}_h \) uniquely defines the vector \( d_{K,1} = P_{K,n+1} - P_{K,1} \). We assume that if any edge \( e \) of \( \mathcal{T}_h \) corresponds to \( d_{K,1} \) of some \( K \), then \( e \) corresponds to \( d_{K,1} \) for each \( K \) containing \( e \). We denote by \( \mathcal{E}_h \) the set of such edges.

\( (A2) \) 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
\[ e' \neq e \quad \& \quad e \cap e' \neq \emptyset. \quad (3.12) \]
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 \( (A1) - (A2) \) 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 \( (A2) \). The mesh satisfying \( (A1) - (A2) \) 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 \( \mathbf{b} \). On each element \( K \in \mathcal{T}_h \) it is defined in the following way
\[ b_K = -\frac{1}{|K|} \left( \int_K \mathbf{b} \cdot \nabla \lambda_{K,1} \, dx \right) d_{K,1}. \quad (3.13) \]
Consequently, when \( \mathbf{b} = \alpha d_{K,1} \) in \( K \) for some \( \alpha \in \mathbb{R} \), the previous definition of \( b_K \) implies that \( b_K = -\frac{1}{|K|} \left( \int_K -\alpha \, dx \right) d_{K,1} = \mathbf{b} \) in \( K \).
Finally, we apply the finite element method and the new method reads:

Find \( v_h \in V_h \) such that for all \( \phi_h \in V_h \) there holds

\[
a_h(u_h, \phi_h) = F_h(\phi_h),
\]

where

\[
a_h(u, \phi) = \varepsilon(\nabla u, \nabla \phi)_\Omega + \sum_{K \in T_h} (b_K \cdot \nabla u, \phi)_K + \sum_{K \in T_h} \left( -\varepsilon \Delta u + b_K \cdot \nabla u, (P_{K,n+1} - C_K) \nabla \phi \right)_K,
\]

\[
F_h(\phi) = \sum_{K \in T_h} \left( f, \phi + (P_{K,n+1} - C_K) \nabla \phi \right)_K.
\]

and the vectors \( b_K \) 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 \( A = \{a_{ij}\}_{i=1,j=1}^{p,q} \), \( p \leq q \), is of nonnegative type if the following conditions hold:

\[
a_{ij} \leq 0 \quad \text{whenever} \quad i \neq j \quad \text{and} \quad \sum_{j=1}^{q} a_{ij} \geq 0 \quad \text{for all} \quad i = 1, 2, \ldots, p.
\]

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

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

We can also verify that

\[
A x = ASS^T x + A(\mathbb{I} - SS^T)x = A_r x_r + z - \tilde{z} = z.
\]

**Theorem 3.2.1.** Assume that \( Ax = z \), where \( A \in \mathbb{R}^{p \times q} \), \( p \leq q \), is a matrix of nonnegative type and \( A_r \) is a nonsingular matrix, then the discrete maximum principle holds, i.e.

\[
z \leq 0 \quad \Rightarrow \quad \max_{1 \leq i \leq q} \{x_i\} \leq \max\{0, x_\nu\}, \quad \text{for some} \quad \nu > p.
\]
If in addition $\sum_{j=1}^{q} a_{ij} = 0$ for all $i = 1, 2, \ldots, p$, then there holds
\[ z \leq 0 \quad \Rightarrow \quad \max_{1 \leq i \leq q} \{x_i\} = x_{\nu}, \quad \text{for some } \nu > p. \quad (3.20) \]

Proof. We proceed as in Codina (1993). Let us begin with the first statement (i.e. \((3.19)\)). Since the matrix $A$ is of nonnegative type and does not contain zero rows ($A_{\nu}$ 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
\[
 x_i = \frac{z_i}{a_{ii}} + \frac{1}{a_{ii}} \sum_{j \in S_i} |a_{ij}| x_j \leq \max_{j \in S_i} \{x_j\} \frac{1}{a_{ii}} \sum_{j \in S_i} |a_{ij}| \leq \max \left\{ 0, \max_{j \in S_i} \{x_j\} \right\},
\]
where $S_i = \{j; 1 \leq j \leq q, a_{ij} < 0\}$ is a set of indices of nonzero off-diagonal entries in the $i$-th row of the matrix $A$. In other words, $S_i$ 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
\[ 1 \leq m \leq p \quad \text{and} \quad x_j < x_m \text{ for all } j > p. \quad (3.22) \]
Denoting $x_k = \max_{j \in S_m} \{x_j\}$ 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 $x_k \neq x_m$ we have $k \leq p$ by \((3.22)\) and the maximum is attained at two inner nodes. The system $A \mathbf{x} = \mathbf{z}$ 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 $A$ 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 $\frac{1}{a_{ii}} \sum_{j \in S_i} |a_{ij}| = 1$ for all $i = 1, 2, \ldots, p$. Consequently, the inequality \((3.21)\) changes into $x_i \leq \max_{j \in S_i} \{x_j\}$, for $i = 1, 2, \ldots, p$. □

Remark 3.2.3. Analogously, one can prove that if $A \mathbf{x} = \mathbf{z}$, where $A \in \mathbb{R}^{p \times q}$, $p \leq q$, is a matrix of nonnegative type and $A_{\nu}$ is a nonsingular matrix, then the discrete minimum principle holds, i.e.
\[ z \geq 0 \quad \Rightarrow \quad \min_{1 \leq i \leq q} \{x_i\} \geq \min \{0, x_{\nu}\}, \quad \text{for some } \nu > p. \quad (3.23) \]
If in addition $\sum_{j=1}^{q} a_{ij} = 0$ for all $i = 1, 2, \ldots, p$, then there holds
\[ z \geq 0 \quad \Rightarrow \quad \min_{1 \leq i \leq q} \{x_i\} = x_{\nu}, \quad \text{for some } \nu > p. \quad (3.24) \]

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 $\varphi_h, \hat{\varphi}_h$ be arbitrary basis functions of $V_h$ and let us rewrite the bilinear form $a_h$ in the following form
\[
 a_h(\varphi_h, \hat{\varphi}_h) = \sum_{K \in T_h} \left\{ \varepsilon (\nabla \varphi_h, \nabla \hat{\varphi}_h)_K + (b_K \cdot \nabla \varphi_h, \hat{\varphi}_h + (P_{K,n+1} - C_K) \nabla \hat{\varphi}_h)_K \right\}.
\]
\( (3.25) \)
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

\[
(P_{K,n+1} - C_K) \nabla \lambda_{K,j} = \lambda_{K,j} (P_{K,n+1}) - \lambda_{K,j} (C_K) .
\]

(3.26)

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

\[
\begin{align*}
(b_K \cdot \nabla \lambda_{K,i}, \lambda_{K,j} + (P_{K,n+1} - C_K) \nabla \lambda_{K,j})_K &= \\
= (b_K \cdot \nabla \lambda_{K,i}, \lambda_{K,j} + \lambda_{K,j} (P_{K,n+1}) - \lambda_{K,j} (C_K))_K &= \\
= (b_K \cdot \nabla \lambda_{K,i}, \lambda_{K,j} (P_{K,n+1}))_K &= \\
&= \frac{|b_K|}{|d_{K,1}|} (d_{K,1} \cdot \nabla \lambda_{K,i}, \lambda_{K,j} (P_{K,n+1}))_K = \\
&= \frac{|b_K|}{|d_{K,1}|} |K| (\delta_{i,n+1} - \delta_{i,1}) \delta_{j,n+1},
\end{align*}
\]

(3.27)

where we used the equality

\[
d_{K,1} \cdot \nabla \lambda_{K,i} = (P_{K,n+1} - P_{K,1}) \cdot \nabla \lambda_{K,i} = \lambda_{K,i}(P_{K,n+1}) - \lambda_{K,i}(P_{K,1}) = \delta_{i,n+1} - \delta_{i,1}.
\]

(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). \( \square \)

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

\[
\tilde{\lambda}_{K,j} = \lambda_{K,j} + (P_{K,n+1} - C_K) \cdot \nabla \lambda_{K,j}.
\]

(3.29)

Then for all \( j = 1, 2, \ldots, n \) we obtain \( \tilde{\lambda}_{K,j} = \lambda_{K,j} - \frac{1}{n+1} \) whereas \( \tilde{\lambda}_{K,n+1} = \lambda_{K,n+1} + \frac{n}{n+1} \). 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 \( b \) has a special property: For each mesh node \( P_j^s \) lying on the boundary \( \Gamma_- \) there exists a sequence of nodes \( \{P_j^s\}_{j=1}^{N_j} \) which lay on the same streamline given by the vector field \( b \) (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 \( \Omega_j^s = \cup_{P_j^s \subset \Gamma} K \), a cluster \( C_j^s = \cup_{P_j-1, P_j^s \subset \Gamma} K \) and a complementary set \( \Omega_{0,j}^s = \Omega_j^s \setminus (C_j^s \cup C_{j+1}^s) \) (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 \( K \subset C_j^s \) holds \( P_{K,1} = P_j^s \) and \( P_{K,n+1} = P_j^s \).
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

$$
\sum_{K \in \mathcal{T}_h} = \sum_{s=1}^{P} \sum_{j=1}^{N_s} \sum_{K \subset C_s^j} .
$$

(3.30)

Indeed, for each element $K \in \mathcal{T}_h$ there exists exactly one edge (determining the vector $\mathbf{d}_{K,1}$) which is oriented in the flow direction. This edge certainly lies on some discrete streamline $s$, $1 \leq s \leq P$, and the endpoints of this edge are $P_{j-1}^s, P_j^s$ for suitable $j$, $1 \leq j \leq N_s$. All elements sharing this edge then form the cluster $C_s^j$ and a union of all clusters is the whole domain $\Omega$. 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 $C_s^j$ let us define the quantities

$$
h_j^s = |P_j^s - P_{j-1}^s|, \quad \beta_j^s = \frac{1}{|C_s^j|} \sum_{K \subset C_s^j} |\mathbf{b}_K||K| \quad \text{and}
q_j^s = - \sum_{K \subset C_s^j} \int_K \mathbf{b} \cdot \nabla \lambda_{K,1} \, d\mathbf{x}.
$$

(3.31)

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

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

(3.32)

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

$$
q_j^s = - \sum_{K \subset C_s^j} \int_K \mathbf{b} \cdot \nabla \lambda_{K,1} \, d\mathbf{x} = \sum_{K \subset C_s^j} \left| \int_K \mathbf{b} \cdot \nabla \lambda_{K,1} \, d\mathbf{x} \right| = \\
= \sum_{K \subset C_s^j} |\mathbf{b}_K||K| = \frac{1}{h_j^s} \sum_{K \subset C_s^j} |\mathbf{b}_K||K| = \frac{\beta_j^s |C_s^j|}{h_j^s} > 0,
$$

(3.33)

which results from the Remark 3.2.1.
In the previous definition the quantity $\beta_j$ is the weighted average value of $|b_K|$ on $C_j$ and $q_j$ 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 $\theta_K$ vanish whenever $b$ is parallel to $b_K$ (i.e., to $d_{K,1}$) 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 $q_j$ and $q_{j+1}$. The following lemma provides an inequality resulting from the structure of the mesh.

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

$$\theta_K \leq \frac{\omega}{n + 1}. \quad (3.34)$$

Then $q_j^s \geq q_{j+1}^s + \frac{\omega}{n + 1}|C_{j+1}^s|$ 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 $\lambda_j^s$ satisfying $\text{supp} \lambda_j^s = \Omega_j^s = \Omega_{n,j}^s \cup \Omega_{j+1}^s$. Then for $K \subset C_j$ holds $\nabla \lambda_{j-1}^s = \nabla \lambda_{K,n}^s = -\nabla \lambda_{K,j}^s = -\nabla \lambda_j^s - \sum_{i=2}^n \nabla \lambda_{K,i}^s$ and from the definition of $q_j$ it follows

$$q_j^s = -\sum_{K \subset C_j} \int_K b \cdot \nabla \lambda_{K,1} \, dx = -\sum_{K \subset C_j} \int_K b \cdot \nabla \lambda_{j-1} \, dx =$$

$$= \sum_{K \subset C_j} \sum_{i=2}^n \int_K b \cdot \nabla \lambda_{K,i} \, dx + \sum_{K \subset C_j} \int_K b \cdot \nabla \lambda_j^s \, dx =$$

$$= \sum_{K \subset C_j} \sum_{i=2}^n \int_K b \cdot \nabla \lambda_{K,i} \, dx + \int_{\Omega_j^s} b \cdot \nabla \lambda_j^s \, dx - \int_{\Omega_{n,j}^s} b \cdot \nabla \lambda_j^s \, dx + q_{j+1}^s =$$

$$= q_{j+1}^s + \sum_{K \subset C_j} \sum_{i=2}^n \int_K b \cdot \nabla \lambda_{K,i} \, dx - \int_{\Omega_j^s} \text{div} b \lambda_j^s \, dx - \int_{\Omega_{n,j}^s} b \cdot \nabla \lambda_j^s \, dx \geq$$

$$\geq q_{j+1}^s + \sum_{K \subset C_j} \frac{\omega}{n + 1}|K| + \frac{\omega}{n + 1}|\Omega_j^s| - \sum_{K \subset \Omega_{n,j}^s} \frac{\omega}{n + 1}|K| =$$

$$= q_{j+1}^s + \frac{\omega}{n + 1}|\Omega_j^s \cup (C_j \cup \Omega_{n,j}^s)| = q_{j+1}^s + \frac{\omega}{n + 1}|C_{j+1}^s|. \quad (3.35)$$

\[\square\]

**Corollary 3.3.1.** If $\text{div} \ b < -\omega < 0$ 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 $C_j$, $j = 1, 2, \ldots, N_s$ 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. \[\square\]

**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 $\Omega$ 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 $\Omega$. Further, we divide each small hypercube into $n!$ $n$-simplices. We demonstrate such a partition on the cube $[0,1]^n$:

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

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

Since $K_{\pi}$ 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 $\bigcup_{\pi \in S_n} K_{\pi}$. This is due to the fact that we can always permute (using some permutation $\pi_0$) 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 $K_{\pi_1}$ and $K_{\pi_2}$). 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 $K_{\pi_1}$ and $K_{\pi_2}$). It means that $\{K_{\pi}\}_{\pi \in S_n}$ forms the partition of $[0,1]^n$ (see Figure 3.2 for 3D example).

![Figure 3.2: Partition of the unit cube into simplices $K_{\pi}$, $\pi \in S_3$, in 3D. Images a) – f) correspond to the permutations $(\underline{123})$, $(\underline{132})$, $(\underline{123})$, $(\underline{213})$, $(\underline{241})$, and $(\underline{312})$, respectively.](image)

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

$$ |K_{\pi}| = \int_0^1 \int_0^{x_{\pi(n)}} \cdots \int_0^{x_{\pi(2)}} dx_{\pi(1)} \cdots dx_{\pi(n)} = \int_0^1 \frac{x_{\pi(n)}^{n-1}}{(n-1)!} dx_{\pi(n)} = \frac{1}{n!}. \quad (3.37) $$

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 $S_n$

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

Then the sets $\{K_{\pi} \cap \{x_j = 0\}, \pi \in \mathcal{L}^{(j)}_n\}$ and $\{K_{\pi} \cap \{x_j = 1\}, \pi \in \mathcal{R}^{(j)}_n\}$ form the partitions of $F_0$ and $F_1$, respectively, and the mapping $\sigma^{(j)}_n : \mathcal{L}^{(j)}_n \rightarrow \mathcal{R}^{(j)}_n$ defined
by the relation
\[
\left(\sigma_n^j(\pi)\right)(i) = \pi \left(1 + (i \mod n)\right) \quad \text{for } i = 1, 2, \ldots, n,
\] (3.39)
is a bijection between \(L_n^j\) and \(R_n^j\).

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 \(\pi \in L_n^j\), 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
\[
x\left(\sigma_n^j(\pi)\right)(i) = x\pi(i+1) \leq x\pi(i+2) = x\left(\sigma_n^j(\pi)\right)(i+1),
\] (3.40)
and since it is also \(x\left(\sigma_n^j(\pi)\right)(n-1) \leq 1 = x\left(\sigma_n^j(\pi)\right)(n)\) 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.](image)

### 3.4 Coercivity

Since \(\int_K v_h - v_h(C_K) \, dx = 0\) for all \(v_h \in V_h\), we can write
\[
\left(\mathbf{b}_K \cdot \nabla u_h, v_h + (P_{K,n+1} - C_K) \cdot \nabla v_h\right)_K = \left(\mathbf{b}_K \cdot \nabla u_h, v_h + v_h(P_{K,n+1}) - v_h(C_K)\right)_K = \left(\mathbf{b}_K \cdot \nabla u_h, v_h(P_{K,n+1})\right)_K = |K| \frac{|\mathbf{b}_K|}{d_{K,1}} \left( u_h(P_{K,n+1}) - u_h(P_{K,1}) \right) v_h(P_{K,n+1}).
\]

63
Consequently, for the bilinear form \( a_h \) holds
\[
a_h(v_h, v_h) = \varepsilon |v_h|^2_{\Omega} + \sum_{K \in T_h} |K| \frac{|b_K|}{d_K} \left( v_h(P_{K,n+1}) - v_h(P_{K,1}) \right) v_h(P_{K,n+1}).  
\] (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
\[
\frac{q_{j+1}}{q_j} \leq \rho_j \quad \text{for } j = 1, 2, \ldots, N - 1. 
\] (3.42)

Then for all \( v_j \in \mathbb{R}, j = 1, 2, \ldots, N, \) holds
\[
q_1 v_1^2 + \sum_{j=2}^N q_j (v_j^2 - v_j v_{j-1}) \geq \frac{1}{2} q_N v_N^2 + \frac{1}{2} \sum_{j=1}^{N-1} (1 - \rho_j) q_j v_j^2.  
\] (3.43)

**Proof.** Subtracting the right-hand side of (3.43) from the left-hand side we obtain
\[
q_1 v_1^2 + \sum_{j=2}^N q_j (v_j^2 - v_j v_{j-1}) - \frac{1}{2} q_N v_N^2 - \frac{1}{2} \sum_{j=1}^{N-1} (1 - \rho_j) q_j v_j^2 = 
\]
\[
= \frac{1}{2} \left\{ q_1 v_1^2 + \sum_{j=2}^N q_j (v_j - v_{j-1})^2 + \sum_{j=1}^{N-1} v_j^2 q_j \left( \rho_j - \frac{q_{j+1}}{q_j} \right) \right\}, 
\]
which is nonnegative due to the inequality (3.42). \( \Box \)

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

Then for all \( v_j \in \mathbb{R}, j = 1, 2, \ldots, N, \) holds
\[
q_1 v_1^2 + \sum_{j=2}^N q_j (v_j^2 - v_j v_{j-1}) \geq \frac{4 - \delta}{2N^2} \sum_{j=1}^N q_j v_j^2.  
\] (3.45)

**Proof.** Applying the Young inequality on the left-hand side of (3.45) yields
\[
q_1 v_1^2 + \sum_{j=2}^N q_j (v_j^2 - v_j v_{j-1}) \geq q_1 v_1^2 + \sum_{j=2}^N q_j \left( v_j^2 - \frac{1}{2\sigma_j} v_{j-1}^2 - \frac{\sigma_j}{2} v_j^2 \right) = 
\]
\[
= \left( q_1 - \frac{q_2}{2\sigma_2} \right) v_1^2 + \sum_{j=2}^{N-1} \left( q_j \left( 1 - \frac{\sigma_j}{2} \right) - \frac{q_{j+1}}{2\sigma_{j+1}} \right) v_j^2 + q_N \left( 1 - \frac{\sigma_N}{2} \right) v_N^2, 
\]

64
where \( \sigma_j, j = 2, 3, \ldots, N, \) are positive numbers. Now we must choose the values \( \sigma_j \) in such a way that all terms in the previous expression are positive. Thus, we take

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

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

\[
q_1 - \frac{q_2}{2\sigma_2} = q_1 \left( 1 - \frac{q_2}{2q_1 \sigma_2} \right) > q_1 \left( 1 - \frac{1}{2} \left( 1 + \frac{\delta}{N^2} \right) \frac{6N - 9}{4N - 3} \right) =
\]

\[
= q_1 \left( \frac{2N + 3 - \delta}{8N - 6 - \frac{N^2}{8} 8N - 6} \right) > q_1 \left( \frac{2N + 3 - 4(6N - 9)}{(8N - 6 - \frac{N^2}{8} 8N - 6)} \right) =
\]

\[
= q_1 \left( \frac{1}{4} + \frac{9}{16N} - \frac{165}{64N^2} + \frac{657}{32N^2(8N - 6)} \right) \geq q_1 \frac{4(4 - \delta)}{2N^2}, \tag{3.47}
\]

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

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

for \( 0 \leq \delta < 4 \) and \( N \geq 8. \)

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

\[
q_j \left( 1 - \frac{\sigma_j}{2} \right) - \frac{q_{j+1}}{2\sigma_{j+1}} = q_j \left( 1 - \frac{\sigma_j}{2} - \frac{1}{2q_j \sigma_j} \right) >
\]

\[
> \frac{q_j}{2} \left( 1 - \frac{2}{N} z_{N,j} - \left( 1 + \frac{\delta}{N^2} \right) \left[ 1 - \frac{(z_{N,j} + \frac{2}{N})^2}{\left( 1 - z_{N,j}^2 - \frac{2}{N} z_{N,j} \right) \left( 1 - \frac{z_{N,j}^2}{1 - \frac{2}{N} z_{N,j}} \right)} \right] \right) =
\]

\[
= \frac{q_j}{2N^2} \left( \frac{4}{1 - z_{N,j}^2} \left( 1 - \frac{2}{N} z_{N,j} \right) \left( 1 - \frac{2}{N} z_{N,j} \right) \right) \geq \delta \left[ 1 - \frac{(z_{N,j} + \frac{2}{N})^2}{\left( 1 - z_{N,j}^2 - \frac{2}{N} z_{N,j} \right) \left( 1 - \frac{z_{N,j}^2}{1 - \frac{2}{N} z_{N,j}} \right)} \right] >
\]

\[
= \frac{q_j}{2N^2} \left( 4 - \delta + \delta \left[ z_{N,j}^2 + (1 - z_{N,j}^2)(z_{N,j} + \frac{2}{N})^2 \right] \right) > 4 - \delta q_j. \tag{3.49}
\]

We have estimated \( \left( 1 - \left( z_{N,j} + \frac{2}{N} \right)^2 \right) \left( 1 - z_{N,j}^2 - \frac{2}{N} z_{N,j} \right) \leq 1, \) 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

\[
q \left( v_j^2 + \sum_{j=2}^{N} (v_j^2 - v_j v_{j-1}) \right) = q \left( v_N^T A_N v_N \right) \geq q\lambda_N |v_N|^2, \tag{3.50}
\]

where \( v_N = (v_1, v_2, \ldots, v_N), A_N = \text{tridiag} \{-\frac{1}{2}, 1, -\frac{1}{2}\} \) and \( \lambda_N = 1 - \cos (\pi/N) \) is the minimal eigenvalue of \( A_N. \)
If we now investigate the behavior of the sequence $\Delta_N$ as $N \to +\infty$, we find out that $\lim_{N \to +\infty} \Delta_N N^2 = \frac{\pi^2}{2} \approx 4.935$. Thus, the constant in the estimate (3.49) of Lemma 3.4.2 is not optimal, nevertheless, the order is optimal $\left(\frac{1}{\sqrt{\pi}}\right)$.

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

$$\|v_N\|^2 = \frac{1}{2N} \|w_h\|^2_{I, I} \geq \frac{\pi^2}{2N} \sum_{j=1}^{N} \|w_h\|^2_{0, I} \geq \frac{\pi^2}{2N} \frac{1}{3} \|v_N\|^2 = \frac{\pi^2}{6N^2} \|v_N\|^2,$$  \hspace{1cm} (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 $q_1 v_1^2 + \sum_{j=2}^{N} q_j v_j^2 = 0$ for $(v_1, v_2, v_3, v_4, v_5) = (1, \sqrt{3}, 2, \sqrt{3}, 1)$. Hence, the optimal upper bound for $\delta$ 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 $\delta$ 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

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

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

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

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

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

$$= \frac{1}{2} q_1 v_1^2 + \frac{1}{2} \sum_{j=2}^{N} q_j (v_j^2 - v_{j-1}^2) = \frac{1}{2} q_N v_N^2 + \frac{1}{2} \sum_{j=1}^{N-1} v_j^2(q_j - q_{j+1}) \geq 0.$$  \hspace{1cm} \( \square \)

**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

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

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

$$q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j^2 - v_j v_{j-1}) \geq \frac{4 - \delta}{8} \left\{ q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j - v_{j-1})^2 \right\}. \hspace{1cm} (3.55)$$
Proof. Denoting \( \alpha = \frac{1}{\delta} \in [0, 1) \) and using Young’s inequality we can write

\[
q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j^2 - v_j v_{j-1}) =
\]

\[
= q_1 v_1^2 + \sum_{j=2}^{N} q_j v_j^2 - \sum_{j=2}^{N} q_j (1 - \alpha) v_j v_{j-1} - \sum_{j=2}^{N} q_j \alpha v_j v_{j-1} \geq
\]

\[
\geq q_1 v_1^2 + \sum_{j=2}^{N} q_j v_j^2 - \frac{1}{2} \sum_{j=2}^{N} q_j (1 - \alpha) (v_j - v_{j-1})^2 - \frac{1}{2} \sum_{j=2}^{N} q_j \alpha \left( \frac{v_j^2 - v_{j-1}}{\sigma_j} + v_{j-1}^2 \sigma_j \right) =
\]

\[
= \frac{1 - \alpha}{2} \left( q_1 v_1^2 + \sum_{j=2}^{N} q_j (v_j - v_{j-1})^2 \right) + \left\{ 1 + \alpha - \frac{q_2}{q_1} \left( 1 - \alpha + \frac{\alpha}{\sigma_j} \right) \right\} q_1 v_1^2 + \frac{1}{2} \sum_{j=2}^{N} q_j \left( 1 - \alpha - \frac{q_{j+1}}{q_j} \left( 1 - \alpha + \frac{\alpha}{\sigma_{j+1}} \right) \right) q_j v_j^2 + \left\{ 1 + \alpha - \alpha \sigma_N \right\} q_N v_N^2. \tag{3.56}
\]

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

\[
1 + \alpha - \alpha \sigma_j - \frac{q_{j+1}}{q_j} \left( 1 - \alpha + \frac{\alpha}{\sigma_{j+1}} \right) \geq \tag{3.57}
\]

\[
\geq 1 + \alpha - \alpha \sigma_j - \left( 1 + \frac{\delta}{N^2} \right) \left( 1 - \alpha + \frac{\alpha}{\sigma_{j+1}} \right) =
\]

\[
= 2\alpha \left( 1 - \frac{\sigma_j}{2} - \frac{1}{2} \left( 1 + \frac{\delta}{N^2} \right) \frac{1}{\sigma_j} \right) - \frac{\delta}{N^2} (1 - \alpha) \geq 2\alpha \frac{4 - \delta}{2N^2} - \frac{\delta}{N^2} (1 - \alpha) = 0.
\]

Further, using the estimate (3.47) we obtain

\[
1 + \alpha - \frac{q_2}{q_1} \left( 1 - \alpha + \frac{\alpha}{\sigma_2} \right) \geq 1 + \alpha - \left( 1 + \frac{\delta}{N^2} \right) \left( 1 - \alpha + \frac{\alpha}{\sigma_2} \right) = \tag{3.58}
\]

\[
= 2\alpha \left( 1 - \frac{1}{2} \left( 1 + \frac{\delta}{N^2} \right) \frac{1}{\sigma_2} \right) - \frac{\delta}{N^2} (1 - \alpha) \geq 2\alpha \frac{4 - \delta}{2N^2} - \frac{\delta}{N^2} (1 - \alpha) = 0,
\]

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

\[
1 + \alpha - \alpha \sigma_N = 1 - \alpha + 2\alpha \left( 1 - \frac{\sigma_N}{2} \right) \geq 1 - \alpha + 2\alpha \frac{4 - \delta}{2N^2} =
\]

\[
= \left( 1 + \frac{\delta}{N^2} \right) (1 - \alpha) \geq 0. \tag{3.59}
\]
3.4.2 Coercivity estimates

In the case when \( \text{div } b < 0 \), 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 \( \text{div } b \leq -\omega < 0 \) we estimate the error of the presented method in the energy norm

\[
|||v|||_b^2 = \varepsilon |v|_\Omega^2 + \frac{\omega K}{2} ||v||_{0,\Omega}^2 + \sum_{K \in \mathcal{T}_h} \frac{|d_{K,1}|}{2|b_K|} ||b_K \cdot \nabla v||_{0,K}^2,
\]

where \( \omega = \min_{j,s} \begin{cases} |C_j|/|\Omega_j|, & |C_{j,s}|/|\Omega_j|, \end{cases} \) \( |d_{K,1}| = P_{K,3} - P_{K,1} \) and \( b_K \) is defined in (3.13) (page 56).

**Theorem 3.4.1** (div \( b < 0 \)). Let the assumptions of Lemma 3.3.1 be fulfilled. Further, let there exists a constant \( \kappa \) independent of \( h \) and \( \varepsilon \) such that \( |C_{j,s}|/|\Omega_j| \geq \kappa \) for all \( s = 1, 2, \ldots, \mathcal{P} \) and \( j = 1, 2, \ldots, N_s \). Then the bilinear form defined in (3.13) satisfies

\[
a_h(v_h, v_h) \geq \frac{1}{2} |||v|||_b^2.
\]

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

\[
a_h(v_h, v_h) = \varepsilon |v_h|_{1,\Omega}^2 + \sum_{s=1}^{\mathcal{P}} \sum_{j=1}^{N_s} q^s_j v_h(P^s_j)(v_h(P^s_j) - v_h(P^s_{j-1}))
\]

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

\[
q^s_{j+1}/q^s_j \leq 1 - \frac{\omega |C_{j+1}|}{(n+1)q^s_j} < 1.
\]

The inequality (3.43) then implies

\[
\sum_{s=1}^{\mathcal{P}} \sum_{j=1}^{N_s} q^s_j v_h(P^s_j)(v_h(P^s_j) - v_h(P^s_{j-1})) \geq \sum_{s=1}^{\mathcal{P}} \sum_{j=1}^{N_s} \frac{\omega |C_{j+1}|}{(n+1)} v_h^2(P^s_j) \geq \sum_{s=1}^{\mathcal{P}} \sum_{j=1}^{N_s} \frac{\omega K |\Omega_j|}{2(n+1)} v_h^2(P^s_j) = \sum_{K \in \mathcal{T}_h} \frac{\omega K |\Omega_j|}{2(n+1)} v_h^2(P^s_j) \geq \frac{\omega K}{2} ||v_h||_{0,\Omega}^2.
\]

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

\[
\sum_{s=1}^{N_s} |\Omega_j| v_h^2(P^s_j) = \sum_{s=1}^{N_s} \sum_{K \in \Omega_j} |K| v_h^2(P^s_j) = \sum_{K \in \mathcal{T}_h} |K| v_h^2(P_{K,i}).
\]

Similarly, using the inequality (3.53) we obtain

\[
\sum_{K \in \mathcal{T}_h} (b_K \cdot \nabla v_h, v_h(P_{K,n+1}))_K = \sum_{s=1}^{\mathcal{P}} \sum_{j=1}^{N_s} v_h(P^s_j)(v_h(P^s_j) - v_h(P^s_{j-1}))q^s_j \geq \frac{1}{2} \sum_{s=1}^{\mathcal{P}} \sum_{j=1}^{N_s} q^s_j (v_h(P^s_j) - v_h(P^s_{j-1}))^2 = \frac{1}{2} \sum_{s=1}^{\mathcal{P}} \sum_{j=1}^{N_s} \sum_{K \in \mathcal{T}_h} \frac{h^s_j}{|b_K|} ||b_K \cdot \nabla v_h||_{0,K}^2 = \sum_{K \in \mathcal{T}_h} \frac{|d_{K,1}|}{2|b_K|} ||b_K \cdot \nabla v_h||_{0,K}^2.
\]
Proof.
At first we observe that for each \( q \) the coercivity of the method with respect to the appropriate energy norm. At meshes.

\[
\sum_{K \in C_j^s} \frac{h_j^s}{|b_K|} \| b_K \cdot \nabla v_h \|_{0,K}^2 = \sum_{K \in C_j^s} \frac{h_j^s}{|b_K|} |K| \frac{|b_K|^2}{|d_{K,1}|^2} \left( \sum_{i=1}^{n+1} d_{K,i} \nabla v_h (P_{K,i}) \right)^2 = \left( v_h(P_j^s) - v_h(P_{j-1}^s) \right)^2 \sum_{K \in C_j^s} \frac{h_j^s}{|b_K|} |K| \frac{|b_K|^2}{(h_j^s)^2}. \tag{3.67}
\]

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

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

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

\[
\sigma(P_j^s) = \frac{|C_j^s| + |\Omega_{0,j}^s|}{|C_j^s|} (h_j^s N_s)^2 \frac{\|b\|_{\infty,\Omega_{0,j}^s}}{\beta_j^s} \frac{\max_{K \in \Omega_j^s} h_K}{h_j^s}. \tag{3.68}
\]

Further, for each element \( K \in T_h \) we define the value \( \sigma_K = \max_{1 \leq i \leq n+1} \sigma(P_{K,i}) \).

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

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

For more general data we obtain different values of \( \sigma(P_j^s) \) or \( \sigma_K \), 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 \( \text{div} \ b = 0 \) in \( \Omega \) and let there exists \( \delta \geq 0 \) such that

\[
\theta_K \leq \frac{\delta}{\sigma_K} \|b\|_{\infty,K} h_K \quad \text{for each } K \in T_h. \tag{3.70}
\]

Then for each \( s = 1, 2, \ldots, P \) and \( j = 1, 2, \ldots, N_s \) it holds

\[
\frac{q_{j+1}^s}{q_j^s} \leq 1 + \frac{\delta}{N_s^2}. \tag{3.71}
\]

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

\[
\sigma(P_j^s) \leq \min_{K \in \Omega_j^s} \max_{1 \leq i \leq n+1} \sigma(P_{K,i}) = \min_{K \in \Omega_j^s} \sigma_K. \tag{3.72}
\]

69
We can use this inequality, the equality from (3.35) and estimate

\[
\frac{q_{j+1}^s}{q_j^s} = 1 + \frac{1}{q_j^s} \left\{ \int_{\Omega_{0,j}} b \cdot \nabla \lambda_j^s \, dx - \sum_{K \subset C_j^s} \sum_{i=2}^n \int_K b \cdot \nabla \lambda_{K,i} \, dx \right\} \leq \\
\leq 1 + \frac{h_j^s}{\beta_j^s |C_j^s|} \left\{ \sum_{K \subset \Omega_{0,j}} \delta \frac{\|b\|_{\infty, K} \|h_K \|_{K}}{\nu_K} + \sum_{K \subset C_j^s} \frac{\delta \|b\|_{\infty, K} \|h_K \|_{K}}{\nu_K} \right\} \leq \\
\leq 1 + \frac{h_j^s \delta}{\beta_j^s |C_j^s|} \max_{K \subset \Omega_{0,j}} \frac{h_K}{\nu_K} \frac{\|b\|_{\infty, \Omega_{0,j}}}{|C_j^s|} \|h_K \|_{K} \leq 1 + \frac{\delta}{N^2 s \min_{K \subset C_j^s} \nu_K} \left\{ \frac{|C_j^s|}{|\Omega_{0,j}|} + \frac{|C_j^s|}{|\Omega_{0,j}|} + \frac{|C_j^s|}{|\Omega_{0,j}|} \right\} \leq 1 + \frac{\delta}{N^2 s}, \tag{3.73}
\]

where we used the fact that the minimum (maximum) over larger set does not increase (decrease).

Remark 3.4.2 (div \( b = 0 \)). In the case when div \( b = 0 \) in \( \Omega \) and the assumptions of Lemma 3.4.5 are fulfilled Lemma 3.4.4 provides an inequality analogous to the estimate (3.66)

\[
\sum_{K \in \mathcal{T}_h} (b_K \cdot \nabla v_h, v_h(P_{K,n+1}))_K \geq \frac{4 - \delta}{4} \sum_{K \in \mathcal{T}_h} \left| \frac{d_{K,1}}{2 |b_K|} \right| \|b_K \cdot \nabla v_h\|^2_{0,K}. \tag{3.74}
\]

When \( \text{div} \ b = 0 \) 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 \( \beta, L \) and \( R \) by the relations

\[
\beta = \min_{j,s} \{ \beta_j^s \}, \quad L = \max_{j,s} \{ N_j h_j^s \} \quad \text{and} \quad R = \max_{j,s} \left\{ \frac{\max_{K \subset \Omega_{0,j}} h_K}{h_j^s} \right\}. \tag{3.75}
\]

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 \( b \) is defined as

\[
\|v\|^2_{b,*} = \varepsilon |v|^2_{\Omega, \Omega} + C'_2 \sum_{K \in \mathcal{T}_h} h_K \|v\|^2_{0,K} + C'_b \sum_{K \in \mathcal{T}_h} \frac{|d_{K,1}|}{2 |b_K|} \|b_K \cdot \nabla v\|^2_{0,K}, \tag{3.76}
\]

where \( C'_2 = \frac{(4 - \delta)s \beta}{2L^2 R} (n + 1) \) and \( C'_b = \frac{4 - \delta}{4} \).

**Theorem 3.4.2** (div \( b = 0 \)). Let the assumptions of Lemma 3.4.5 be fulfilled. Then for each \( v_h \in V_h \) there holds

\[
a_h(v_h, v_h) \geq \frac{1}{2} \|v\|^2_{b,*}. \tag{3.77}
\]
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

\[ \sum_{K \in T_h} (b_K \cdot \nabla u_h, v_h(P_{K,n+1}))_K = \sum_{s=1}^{P} \sum_{j=1}^{N_a} v_h(P_J^s)(v_h(P_J^s) - v_h(P_{J-1}^s))q_j^s \geq \]

\[ \geq \frac{4 - \delta}{2} \sum_{s=1}^{P} \sum_{j=1}^{N_a} h_j^s v_h^2(P_J^s) \geq \frac{4 - \delta}{2} \sum_{s=1}^{P} \sum_{j=1}^{N_a} \beta_j^s |C_j^s|^2 h_j^s v_h^2(P_J^s) \geq \]

\[ \geq \frac{(4 - \delta)\kappa\beta}{2L^2} \sum_{s=1}^{P} \sum_{j=1}^{N_a} |\Omega_j^s| h_j^s v_h^2(P_J^s) \geq \frac{(4 - \delta)\kappa\beta}{2L^2 R} \sum_{K \in T_h} h_K |K| \sum_{i=1}^{n+1} v_h^2(P_{K,i}) \geq \]

\[ \geq \frac{(4 - \delta)\kappa\beta}{2L^2 R} (n + 1) \sum_{K \in T_h} h_K \|v_h\|_{0,K}^2. \quad (3.78) \]

Combining the estimates (3.74) and (3.78) completes the proof. \( \square \)

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

\[ -\varepsilon \Delta u + b \cdot \nabla u + cu = f \quad \text{in } \Omega \quad (3.79) \]

equipped with the Dirichlet boundary condition \( u = 0 \) on \( \partial \Omega \). Further, we assume that \( b \in W^{1,\infty}(\Omega)^n \), \( c \in L^{\infty}(\Omega) \), \( f \in L^2(\Omega) \) and \( c - \frac{1}{2} \text{div } b \geq \omega > 0 \) in \( \Omega \).

The finite element space \( V_h^{(k)} \subset V = H_0^1(\Omega) \) is defined as

\[ V_h^{(k)} = \{ \varphi_h \in V, \varphi_h|_K \in P_k(K) \text{ for all } K \in T_h \}, \quad (3.80) \]

where the triangulation \( T_h \) is assumed to be shape-regular (cf. Theorem 4.2.2).

For each \( K \in T_h \) let \( \delta_K \) be any positive numbers (specified later), then the SUPG solution \( u_{SU} \in V_h^{(k)} \) satisfies

\[ a_{SU}(u_{SU}, \varphi_h) = \sum_{K \in T_h} (f, \varphi_h + \delta_K b \cdot \nabla \varphi_h)_K \quad \text{for all } \varphi_h \in V_h^{(k)}, \quad (3.81) \]

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

\[ a_{SU}(v, \varphi) = \varepsilon (\nabla v, \nabla \varphi)_\Omega + (b \cdot \nabla v, \varphi)_\Omega + (cv, \varphi)_\Omega + \]

\[ + \sum_{K \in T_h} \delta_K (-\varepsilon \Delta v + b \cdot \nabla v + cv, b \cdot \nabla \varphi)_K. \quad (3.82) \]
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

\[
|||v|||_{SU} = \left( \varepsilon |v|_{1,\Omega}^2 + \omega \|v\|_{0,\Omega}^2 + \sum_{K \in T_h} \delta_K \|b \cdot \nabla v\|_{0,K}^2 \right)^{1/2}.
\]  

(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 \( \delta_K \leq \min \left\{ \frac{\omega}{\|c\|_{0,\infty,K}^2}, \frac{h_K}{2\varepsilon} \|b\|_{0,\infty,K} \right\} \) for all \( K \in T_h \). Then for all functions \( v_h \in V_h^{(k)} \) holds

\[
a_{SU}(v_h, v_h) \geq \frac{1}{2} |||v_h|||_{SU}^2.
\]  

(3.84)

**Proof.** We use the inverse inequality (Theorem 4.2.2), the Young’s inequality together with the bound for \( \delta_K \) and estimate

\[
a_{SU}(v_h, v_h) = \varepsilon |v_h|_{1,\Omega}^2 + \left( c - \frac{1}{2} \text{div} \ b, v_h^2 \right)_\Omega + \sum_{K \in T_h} \delta_K \|b \cdot \nabla v_h\|_{0,K}^2 + \sum_{K \in T_h} \delta_K (-\varepsilon \Delta v_h + cv_h, b \cdot \nabla v_h)_{K} \geq
\]

\[
\geq \varepsilon |v_h|_{1,\Omega}^2 + \omega \|v_h\|_{0,\Omega}^2 + \sum_{K \in T_h} \delta_K \|b \cdot \nabla v_h\|_{0,K}^2 - \sum_{K \in T_h} \delta_K \left( \varepsilon \frac{\text{min}}{h_K} |v_h|_{1,K} + \|c\|_{0,\infty,K} \|v_h\|_{0,K} \right) \|b \cdot \nabla v_h\|_{0,K} \geq
\]

\[
\geq \varepsilon |v_h|_{1,\Omega}^2 + \omega \|v_h\|_{0,\Omega}^2 + \sum_{K \in T_h} \delta_K \|b \cdot \nabla v_h\|_{0,K}^2 - \sum_{K \in T_h} \delta_K \varepsilon \frac{\text{min}}{h_K} |v_h|_{1,K}^2 \|b\|_{0,\infty,K} - \frac{1}{2} \sum_{K \in T_h} \delta_K \|c\|_{0,\infty,K}^2 \|v_h\|_{0,K}^2 - \frac{1}{2} \sum_{K \in T_h} \delta_K \|b \cdot \nabla v_h\|_{0,K}^2 \geq \frac{1}{2} |||v_h|||_{SU}^2. \quad (3.85)
\]

\[
\square
\]

In the derivation of the error estimate it is necessary to estimate the derivative in the flow direction. Thus, we use the inequality

\[
\|b \cdot \nabla v\|_{0,K} \leq \min \left\{ \frac{1}{\delta_K^{1/2}}, \frac{\|b\|_{0,\infty,K}}{\varepsilon^{1/2}} \right\} \left( \varepsilon |v|_{1,K}^2 + \omega \|v\|_{0,K}^2 + \delta_K \|b \cdot \nabla v\|_{0,K}^2 \right)^{1/2}. \quad (3.86)
\]

Due to the consistency of the method we obtain for all \( u \in H^2(\Omega) \) the Galerkin orthogonality

\[
a_{SU}(u - u_{SU}, v_h) = 0 \quad \text{for all} \ v_h \in V_h^{(k)}. \quad (3.87)
\]

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 \( \eta_h = u - u^f \) and the error of the method \( \xi_h = u^f - u_{SU} \in V_h \). The coercivity (3.84) and the Galerkin orthogonality (3.87) then implies
\[
\frac{1}{2} \| \xi_h \|_{SU}^2 \leq a_{SU}(\xi_h, \xi_h) = a_{SU}(-\eta_h, \xi_h) =
\]
\[
= -\varepsilon(\nabla \eta_h, \nabla \xi_h)_{\Omega} + (\eta_h \nabla b, \xi_h)_{\Omega} + (\eta_h, b \cdot \nabla \xi_h)_{\Omega} - (c \eta_h, \xi_h)_{\Omega} - \sum_{K \in T_h} \delta_K (-\varepsilon \Delta \eta_h + b \cdot \nabla \eta_h + c \eta_h, b \cdot \nabla \xi_h)_K \leq
\]
\[
\leq \sum_{K \in T_h} \varepsilon |\eta_h|_{1,K} |\xi_h|_{1,K} + \sum_{K \in T_h} \left( n \|b\|_{1,\infty,K} + |c|_{0,\infty,K} \right) \|\eta_h\|_{0,K} \|\xi_h\|_{0,K} +
\]
\[
+ \sum_{K \in T_h} \left[ \varepsilon |\eta_h|_{2,K} + \delta_K \left( \varepsilon n |\eta_h|_{2,K} + \|b\|_{0,\infty,K} |\eta_h|_{1,K} + |c|_{0,\infty,K} \|\eta_h\|_{0,K} \right) \right] \|\eta_h\|_{0,K} \|\xi_h\|_{0,K}.
\]
Now we set \( \mu_K = n \|b\|_{1,\infty,K} + |c|_{0,\infty,K} \) for all \( K \in T_h \), use the estimate (3.86), Corollary 4.2.2 with the discrete Cauchy–Schwarz inequality and obtain
\[
\frac{1}{2} \| \xi_h \|_{SU}^2 \leq \| \xi_h \|_{SU} \left( 6 \sum_{K \in T_h} \left( \varepsilon |\eta_h|_{1,K} + \frac{\mu_K}{\omega} \|\eta_h\|_{0,K} + \left[ \|\eta_h\|_{0,K} \right] \right) \right)^{1/2}
\]
\[
+ \frac{\delta_K}{\varepsilon} \left( \varepsilon^2 |\eta_h|_{2,K}^2 + \|b\|_{0,\infty,K}^2 |\eta_h|_{1,K}^2 + |c|_{0,\infty,K}^2 \|\eta_h\|_{0,K}^2 \right) \left( \frac{1}{\delta_K}, \frac{\|b\|_{0,\infty,K}}{\varepsilon} \right)^{1/2}.
\]
(3.88)
The interpolation inequality (Theorem 4.2.1 page 127) then provides for \( u \in H^{r+1}(\Omega), r \leq k \) and \( m \in \{0,1,2\} \) the estimate
\[
\left( \sum_{K \in T_h} \| u - u^f \|_{m,K}^2 \right)^{1/2} \leq C_X \left( \sum_{K \in T_h} h_K^{2(r+1-m)} \| u \|_{r+1,K}^2 \right)^{1/2}.
\]
(3.89)
Dividing by \( \| \xi_h \|_{SU} \) and using (3.89) the inequality (3.88) changes into
\[
\| \xi_h \|_{SU} \leq \left( 24 \sum_{K \in T_h} C_X^2 h_K^4 |u|_{r+1,K}^2 \left( \varepsilon + \frac{\mu_K^2}{\omega} h_K^2 \right) + \frac{\delta_K^2}{\varepsilon} \left[ \frac{\varepsilon^2}{h_K^2} + \|b\|_{0,\infty,K}^2 + |c|_{0,\infty,K}^2 \|h_K^2 \right] \right) \min \left( \frac{1}{\delta_K}, \frac{\|b\|_{0,\infty,K}}{\varepsilon} \right)^{1/2}.
\]
(3.90)
Now we choose the value of \( \delta_K \) in order to obtain the best possible order of convergence. Choosing \( \delta_K = \frac{h_K}{2C_{inv} \|b\|_{0,\infty,K}} \leq \frac{\omega}{\|\eta_h\|_{0,K}} \) leads to estimates
\[
\| \xi_h \|_{SU} \leq C_{SU}^\xi \left( \sum_{K \in T_h} \frac{|b|_{0,\infty,K}}{2} h_K^{2r+1} |u|_{r+1,K}^2 \right)^{1/2},
\]
(3.91)
for \( \text{Pe}_K \geq C_{inv} \) and
\[
\| \xi_h \|_{SU} \leq C_{SU}^\xi \left( \sum_{K \in T_h} C_{inv} \varepsilon h_K^{2r+1} |u|_{r+1,K}^2 \right)^{1/2},
\]
for \( \text{Pe}_K < C_{inv} \). Here \( C_{SU}^\xi = C_X \sqrt{C_{inv}} \left( 2 + \frac{1}{4\varepsilon^{inw}} + 8C_{inv}^3 \right) \left( 1 + n \max_{K \in T_h} \frac{|b|_{0,\infty,K}}{|c|_{0,\infty,K}} \right)^2 \right)^{1/2} \) and the Pécel number is defined as \( \text{Pe}_K = \frac{|b|_{0,\infty,K} h_K}{2\varepsilon} \). From these inequalities it follows that
\[
\| \xi_h \|_{SU} \leq C_{SU}^\xi \left( \sum_{K \in T_h} \max \{ C_{inv}, \text{Pe}_K \} \varepsilon h_K^{2r} |u|_{r+1,K}^2 \right)^{1/2}.
\]
(3.92)
Since for the approximation error holds

\[
\|\eta_h\|_{SU} \leq \left( \sum_{K \in \mathcal{T}_h} C_X^2 \left( \varepsilon + \omega h_K^2 + \frac{\|b\|_{0,\infty,K} h_K}{2C_{inv}} \right) h_K^{2r} \right)^{1/2} \leq C_S^\eta \left( \sum_{K \in \mathcal{T}_h} \max \{C_{inv}, \Pi_K \} \varepsilon h_K^{2r} \right)^{1/2},
\]

with \( C_S^{\eta} = \frac{C_S}{\sqrt{C_{inv}}} \left( 2 + 4\varepsilon^2 C_{inv}^2 \max_{K \in \mathcal{T}_h} \left\{ \frac{\|b\|_{0,\infty,K}^2}{\|\varepsilon\|_{0,\infty,K}} \right\} \right)^{1/2} \), we can for \( r \leq k \) estimate the error \( u - u_{SU} \) in the energy norm

\[
\|u - u_{SU}\|_{SU} \leq \left( C_S^{\xi} + C_S^{\eta} \right) \left( \sum_{K \in \mathcal{T}_h} \max \{C_{inv}, \Pi_K \} \varepsilon h_K^{2r} \right)^{1/2}.
\]

Remark 3.5.1. Since \( C_{inv} > 1 \) (cf. Remark 4.2.1, page 128), the assumption \( \delta_K \leq \min \left\{ \frac{\omega \varepsilon h_K}{2C_{inv} \|b\|_{0,\infty,K}}, \frac{\varepsilon h_K}{C_{inv} \|b\|_{0,\infty,K}} \right\} \) does not allow to choose \( \delta_K = \frac{h_K}{2C_{inv} \|b\|_{0,\infty,K}} \) which is believed to be the optimal choice. Nevertheless, the a priori error estimate of the same order is achieved for the choice \( \delta_K = \frac{h_K}{2C_{inv} \|b\|_{0,\infty,K}} \), 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

\[
a_h(u - u_h, v_h) = \sum_{K \in \mathcal{T}_h} ((b_K - b) \nabla u, v_h + (P_{K,n+1} - C_K) \cdot \nabla v_h)_K
\]

and consequently for any \( w_h \in V_h \) it holds

\[
a_h(w_h - u_h, w_h - u_h) = \varepsilon(\nabla(w_h - u), \nabla(w_h - u_h))_\Omega + \varepsilon \sum_{K \in \mathcal{T}_h} (\Delta u, (P_{K,n+1} - C_K) \cdot \nabla(w_h - u_h))_K +
\]

\[
+ \sum_{K \in \mathcal{T}_h} (b_K - b) \cdot \nabla w_h + b \cdot \nabla(w_h - u), w_h - u_h + (P_{K,n+1} - C_K) \cdot \nabla(w_h - u_h))_K.
\]

*Proof.* The Galerkin quasi-orthogonality property (3.95) follows from the definition of \( a_h, a, F_h \) and \( F \)

\[
a_h(u - u_h, v_h) = a(u, v_h) + \sum_{K \in \mathcal{T}_h} ((b_K - b) \nabla u, v_h)_K +
\]

\[
+ \sum_{K \in \mathcal{T}_h} (b_K - b) \nabla u, (P_{K,n+1} - C_K) \cdot \nabla v_h)_K - a_h(u_h, v_h),
\]

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

\[
a_h(w_h - u_h, w_h - u_h) = a_h(w_h - u, w_h - u_h) + a_h(u - u_h, w_h - u_h)
\]

and the fact that \( \Delta w_h = 0 \) we derive the relation (3.96). \( \square \)
For estimation of the difference \( b_K - b \) that occurs in Lemma \ref{lem:3.5.2} we use the following lemma.

**Lemma 3.5.3.** Let \( b \in W^{1,\infty}(K)^n \) and let us define the vector \( b^I_K \in \mathbb{R}^n \)

\[
 b^I_K = -\frac{1}{|K|} \sum_{j=1}^n \left( \int_K b \cdot \nabla \lambda_{K,j} \, dx \right) d_{K,j}.
\]

(3.99)

Then for every \( v_h \in P_1(K) \) holds

\[
 \|(b^I_K - b) \cdot \nabla v_h\|_{0,K} \leq nC_{\Pi} h_K |b|_{1,\infty,K} |v_h|_{1,K} \quad \text{and} \quad \|(b_K - b^I_K) \cdot \nabla v_h\|_{0,K} \leq \frac{1}{|K|} \sum_{j=2}^n \int_K b \cdot \nabla \lambda_{K,j} \, dx \, h_K |v_h|_{1,K},
\]

(3.100)

(3.101)

where \( b_K \) is a vector defined in \( (3.13) \).

**Proof.** Since \( d_{K,j} = P_{K,n+1} - P_{K,j} \) and \( d_{K,j} \nabla \lambda_{K,i} = -\delta_{ij} \) for \( i \neq n + 1 \), it holds

\[
 \int_K b^I_K \nabla \lambda_{K,i} \, dx = \int_K -\frac{1}{|K|} \sum_{j=1}^n \left( \int_K b \cdot \nabla \lambda_{K,j} \, dy \right) (-\delta_{ij}) \, dx = \int_K b \cdot \nabla \lambda_{K,i} \, dx \tag{3.102}
\]

for all \( i = 1, 2, \ldots, n \) and (using \( \sum_{j=1}^n b \cdot \nabla \lambda_{K,j} = -b \cdot \nabla \lambda_{K,n+1} \))

\[
 \int_K b^I_K \nabla \lambda_{K,n+1} \, dx = \int_K -\frac{1}{|K|} \sum_{j=1}^n \left( \int_K b \cdot \nabla \lambda_{K,j} \, dy \right) \, dx = \int_K b \cdot \nabla \lambda_{K,n+1} \, dx \tag{3.103}
\]

Consequently, \( f_K (b^I_K - b) \cdot \nabla v_h \, dx = 0 \) for all \( v_h \in V_h \) and we can call \( b^I_K \nabla v_h \) the \( P_0 \)-interpolation of the function \( b \cdot \nabla v_h \) on \( K \). Using the approximation property (Theorem 4.2.3 page 128) we therefore obtain

\[
 \|(b^I_K - b) \cdot \nabla v_h\|_{0,K} \leq C_{\Pi} h_K |b \cdot \nabla v_h|_{1,K} \leq nC_{\Pi} h_K |b|_{1,\infty,K} |v_h|_{1,K} \tag{3.104}
\]

The estimate (3.101) results directly from the definition of \( b^I_K \) and \( b_K \) and the fact that \( |d_{K,j}| \leq h_K = \max_{i \neq j} |P_{K,i} - P_{K,j}|. \)

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 \( \kappa \) independent of \( h \) (and \( \varepsilon \)) such that

\[
 \frac{|C_{s,j}|}{|P_j|} \geq \kappa \quad \text{for all} \quad s = 1, 2, \ldots, \mathcal{P} \quad \text{and} \quad j = 1, 2, \ldots, N_s, \quad \text{constant} \quad \omega > 0 \quad \text{such that} \quad \text{div} \, b \leq -\omega < 0 \quad \text{in} \quad \Omega \quad \text{and let for each} \quad K \in \mathcal{T}_h \quad \text{holds}
\]

\[
 \theta_K \leq \frac{\omega}{n+1}.
\]

(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 \( \varepsilon \) such that for the solution \( u_h \in V_h \) obtained by the method \( (3.1) \) (using continuous piecewise linear finite elements) it holds

\[
 |||u - u_h|||_b \leq C_1 \left( \sum_{K \in \mathcal{T}_h} h^2_K \left( |u|_{2,K}^2 + |u|_{1,K}^2 \right) \right)^{1/2}.
\]

(3.106)
If, in addition, the mesh parameter $\theta_K$ satisfies for all $K \in \mathcal{T}_h$

$$\theta_K \leq \min \left\{ \frac{\omega}{n+1}, |b|_{1,\infty,K} \sqrt{\omega} \max \left\{ \frac{h_K}{\varepsilon^{1/2}}, \frac{2}{|b_K|} \varepsilon^{1/2} \right\} \right\}, \quad (3.107)$$

then there exists a constant $C_2 > 0$ independent of $h$ and $\varepsilon$ such that for the solution $u_h \in V_h$ obtained by the method (3.14) there holds

$$\|u - u_h\|_h \leq C_2 \left( \sum_{K \in \mathcal{T}_h} \min \left\{ h_K^2, \max \left\{ \frac{h_K^4}{\varepsilon}, \varepsilon h_K^2 \right\} \right\} \left( |u|_{2,K}^2 + |u|_{1,K}^2 \right) \right)^{1/2}. \quad (3.108)$$

**Proof.** At first, let $u'$ be again the $V_h$-interpolant of the function $u$ and let us denote $\eta_h = u - u'$ and $\xi_h = u' - u_h$. Further, we decompose the error $u - u_h = (u - u') + (u' - u_h) = \eta_h + \xi_h$ and since $\xi_h \in V_h$ we can use the coercivity (3.61) together with (3.96) (setting $w_h = u'$), which yields

$$\frac{1}{2} \|\xi_h\|_h^2 \leq a_h(\xi_h, \xi_h) =$$

$$= -\varepsilon(\nabla \eta_h, \nabla \xi_h)_\Omega + \varepsilon \sum_{K \in \mathcal{T}_h} (\Delta u, (P_{K,n+1} - C_K) \cdot \nabla \xi_h)_K +$$

$$+ \sum_{K \in \mathcal{T}_h} \left( (b_K - b_K') \cdot \nabla u' + (b_K' - b_K) \cdot \nabla u' - b \cdot \nabla \eta_h, \xi_h + (P_{K,n+1} - C_K) \cdot \nabla \xi_h \right)_K. \quad (3.109)$$

Now we estimate each term of (3.109) separately:

1. We use the Cauchy-Schwarz-Bunyakovskiy inequalities (Theorem 4.1.4 page 126), the interpolation inequality (Theorem 4.2.1 page 127) and estimate

$$-\varepsilon(\nabla \eta_h, \nabla \xi_h)_\Omega = -\sum_{K \in \mathcal{T}_h} \varepsilon(\nabla \eta_h, \nabla \xi_h)_K \leq \sum_{K \in \mathcal{T}_h} \varepsilon |\eta_h|_{1,K} |\xi_h|_{1,K} =$$

$$= \sum_{K \in \mathcal{T}_h} \left( \varepsilon^{1/2} |\eta_h|_{1,K} \varepsilon^{1/2} |\xi_h|_{1,K} \right) \leq \left( \sum_{K \in \mathcal{T}_h} \varepsilon |\eta_h|_{1,K}^2 \right)^{1/2} \left( \sum_{K \in \mathcal{T}_h} \varepsilon |\xi_h|_{1,K}^2 \right)^{1/2} \leq$$

$$\leq \left( \sum_{K \in \mathcal{T}_h} \varepsilon C_X h_K^2 |u|_{2,K}^2 \right)^{1/2} \left( \sum_{K \in \mathcal{T}_h} \varepsilon |\xi_h|_{1,K}^2 \right)^{1/2} \leq C_X \left( \sum_{K \in \mathcal{T}_h} \varepsilon h_K^2 |u|_{2,K}^2 \right)^{1/2} \|\xi_h\|_b. \quad (3.110)$$

2. Since $|P_{K,n+1} - C_K| \leq h_K$ then similarly as in the previous case we have

$$\varepsilon \sum_{K \in \mathcal{T}_h} (\Delta u, (P_{K,n+1} - C_K) \cdot \nabla \xi_h)_K \leq \varepsilon \sum_{K \in \mathcal{T}_h} n|u|_{2,K} h_K |\xi_h|_{1,K} =$$

$$= n \sum_{K \in \mathcal{T}_h} \left( \varepsilon^{1/2} h_K |u|_{2,K} \varepsilon^{1/2} |\xi_h|_{1,K} \right) \leq n \left( \sum_{K \in \mathcal{T}_h} \varepsilon h_K^2 |u|_{2,K}^2 \right)^{1/2} \|\xi_h\|_b. \quad (3.111)$$

3. From the inequality (3.101) it follows that there holds $\|(b_K - b_K') \cdot \nabla u'\|_{0,K} \leq \theta_K h_K |u'|_{1,K}$. Consequently, using the inverse inequality (Theorem 4.2.2...
and the estimate \( |u'\|_{1,K} \leq |u|_{1,K} + |\eta_h|_{1,K} \) together with the interpolation inequality (Theorem 4.2.1 page 127) yields

\[
\sum_{K \in T_h} \left( (b_K - b_K') \cdot \nabla u', \xi_h + (P_{K,n+1} - C_K) \cdot \nabla \xi_h \right)_{K} \leq \\
\sum_{K \in T_h} \theta_K h_K |u'|_{1,K} |1 + C_{inv}| \|\xi_h\|_{0,K} \leq \\
\leq \frac{(1 + C_{inv}) \sqrt{2}}{\sqrt{K}} \left( \sum_{K \in T_h} \frac{\theta_K^2 h_K^2 (|u|_{1,K} + C \chi h_K |u|_{2,K}^2)}{\omega} \right)^{1/2} \|\xi_h\|_b. \quad (3.112)
\]

4. Since \( f_K(b_K' - b) \cdot \nabla u' \, dx = 0 \), 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

\[
\sum_{K \in T_h} \left( (b_K' - b) \cdot \nabla u', \xi_h + (P_{K,n+1} - C_K) \cdot \nabla \xi_h \right)_{K} = \\
= \sum_{K \in T_h} \left( (b_K' - b) \cdot \nabla u', \xi_h - \xi_h(C_K) \right)_{K} \leq \sum_{K \in T_h} nC^2_{\Pi} h^2_{K} |b|_{1,\infty,K} |u'|_{1,K} |\xi_h|_{1,K} \leq \\
\leq nC^2_{\Pi} \left( \sum_{K \in T_h} h^4_{K} |b|^2_{1,\infty,K} (|u|_{1,K} + |\eta_h|_{1,K})^2 \min \left\{ \frac{1}{\varepsilon}, \frac{2C_{inv}^2}{\omega K h^2_{K}} \right\} \right)^{1/2} \|\xi_h\|_b \leq \\
\leq nC^2_{\Pi} \left( \sum_{K \in T_h} h^4_{K} |b|^2_{1,\infty,K} (|u|_{1,K} + C \chi h_K |u|_{2,K}^2) \min \left\{ \frac{h^2_{K}}{\varepsilon}, \frac{2C_{inv}^2}{\omega K} \right\} \right)^{1/2} \|\xi_h\|_b, \\
(3.113)
\]

where we employed the inequality \( |\xi_h|^2_{1,K} \leq \min \left\{ \frac{1}{\varepsilon}, \frac{2C^2_{inv}}{\omega K h^2_{K}} \right\} \|\xi_h\|^2_b \) 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

\[
- \sum_{K \in T_h} (b \cdot \nabla \eta_h, \xi_h)_{K} = - (b \cdot \nabla \eta_h, \xi_h)_{\Omega} = (\eta_h \text{div} b, \xi_h)_{\Omega} + (\eta_h, b \cdot \nabla \xi_h)_{\Omega} = \\
= (\eta_h \text{div} b, \xi_h)_{\Omega} + \sum_{K \in T_h} (\eta_h, b_K \cdot \nabla \xi_h)_{K} + \sum_{K \in T_h} (\eta_h, (b - b_K) \cdot \nabla \xi_h)_{K} \leq \\
\leq \sum_{K \in T_h} n|b|_{1,\infty,K} \|\eta_h\|_{0,K} \|\xi_h\|_{0,K} + \sum_{K \in T_h} \|\eta_h\|_{0,K} \|b_K \cdot \nabla \xi_h\|_{0,K} + \\
+ \sum_{K \in T_h} \|\eta_h\|_{0,K} (nC_{\Pi}|b|_{1,\infty,K} + \theta_K) h_K |\xi_h|_{1,K} \leq \\
\leq \left( 3 \sum_{K \in T_h} \gamma_K C^2_{\Pi} h^4_{K} |u|_{2,K}^2 \right)^{1/2} \|\xi_h\|_b, \quad (3.114)
\]

where we denoted \( \gamma_K = \frac{2\varepsilon}{\omega K} |b|^2_{1,\infty,K} + \min \left\{ \frac{2\varepsilon |b|_{1,\infty,K} |b_K|^2}{\omega K}, \frac{|b|^2_{1,\infty,K}}{\varepsilon} \right\} + (nC_{\Pi}|b|_{1,\infty,K} + \theta_K)^2 \min \left\{ \frac{h^2_{K}}{\varepsilon}, \frac{2C^2_{inv}}{\omega K} \right\} \). In the above estimate we decomposed the difference
\[ \mathbf{b} - \mathbf{b}_K = (\mathbf{b} - \mathbf{b}'_K) + (\mathbf{b}'_K - \mathbf{b}_K) \] and employed the inequalities (3.100) and (3.101).

6. Finally, the estimate of the last term takes form

\[ - \sum_{K \in \mathcal{T}_h} \left( \mathbf{b} \cdot \nabla \eta_h, (P_{K,n+1} - C_K) \cdot \nabla \xi_h \right)_K \leq \sum_{K \in \mathcal{T}_h} \| \mathbf{b} \|_{0,\infty,K} |\eta_h|_{1,K} h_K |\xi_h|_{1,K} \leq \left( \sum_{K \in \mathcal{T}_h} \| \mathbf{b} \|^2_{0,\infty,K} C_K^2 h_K^2 |u|^2_{2,K} \min \left\{ \frac{h_K^2}{\varepsilon}, \frac{2C_{\text{inv}}^2}{\omega_K} \right\} \right)^{1/2} \| \xi_h \|_b. \quad (3.115) \]

From the derived estimates it follows that if we consider \( \theta_K \leq \frac{\omega}{n+1} \) for all \( K \in \mathcal{T}_h \), we obtain the estimate

\[ \| \xi_h \|_b \leq C_0 \left( \sum_{K \in \mathcal{T}_h} h_K^2 \left( |u|^2_{2,K} + |u|^2_{1,K} \right) \right)^{1/2}, \quad (3.116) \]

where the constant \( C_0 \) does not depend on \( \varepsilon \) and \( h_K \).

However, if the triangulation \( \mathcal{T}_h \) (and a vector field \( \mathbf{b} \)) 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 \( h_K \geq \varepsilon^{1/2} C_{\text{inv}} \sqrt{\frac{2}{\omega K}} \) for all \( K \in \mathcal{T}_h \), then

\[ |\mathbf{b}|_{1,\infty,K} \sqrt{\omega} \frac{h_K}{\varepsilon^{1/2}} \geq |\mathbf{b}|_{1,\infty,K} C_{\text{inv}} \sqrt{2/K} \geq \frac{\omega}{n} C_{\text{inv}} \sqrt{2(n+1)} \geq \frac{\omega}{n+1}, \quad (3.117) \]

where we used the apparent inequalities \( |\mathbf{b}|_{1,\infty,K} \geq \frac{1}{n} |\text{div} \mathbf{b}| \geq \frac{1}{n} \omega, \: \kappa \leq \frac{1}{n+1} \) and \( C_{\text{inv}} \geq 1 \) (cf. Remark 4.2.1, page 128). Hence, from the assumption (3.107) it follows that the bound \( \theta_K \leq \frac{\omega}{n+1} \) is used in this case, which together with Corollary 4.2.2 (page 129) results in the same estimate as above

\[ \| \xi_h \|_b \leq C_A \left( \sum_{K \in \mathcal{T}_h} h_K^2 \left( |u|^2_{1,K} + |u|^2_{2,K} \right) \right)^{1/2}, \]

where the constant \( C_A \) does not depend on \( \varepsilon \) and \( h_K \). Thus, no improvement is achieved in this case.

B) If the mesh is refined and \( \frac{2\sigma}{|\mathbf{b}_K|} \varepsilon \leq h_K \leq \varepsilon^{1/2} C_{\text{inv}} \sqrt{\frac{2}{\omega K}} \) for all \( K \in \mathcal{T}_h \), then using the inequality

\[ \theta_K \leq |\mathbf{b}|_{1,\infty,K} \sqrt{\omega} \frac{h_K}{\varepsilon^{1/2}} \leq |\mathbf{b}|_{1,\infty,K} C_{\text{inv}} \sqrt{2/K} \quad (3.118) \]

and the estimate \( \varepsilon h_K^2 = \frac{\sigma^2 K^2}{\varepsilon} \leq \frac{|\mathbf{b}_K|^2}{4\sigma^2} \frac{h_K^2}{\varepsilon} \) yields

\[ \| \xi_h \|_b \leq C_B \left( \sum_{K \in \mathcal{T}_h} \frac{h_K^4}{\varepsilon} \left( |u|^2_{1,K} + |u|^2_{2,K} \right) \right)^{1/2}, \]

where the constant \( C_B \) again does not depend on \( \varepsilon \) and \( h_K \). Since in this case there holds \( \frac{2\sigma}{|\mathbf{b}_K|} h_K^3 \leq \frac{h_K^4}{\varepsilon} \leq \frac{2C_{\text{inv}}^2}{\omega K} \), one may expect an acceleration of the convergence in comparison to (3.116).
C) If the mesh step satisfies \( h_K \leq \varepsilon \frac{2\sigma}{|b|} \) for all \( K \in \mathcal{T}_h \), then there holds
\[
\theta_K \leq |b|_{1,\infty,K} \sqrt{\omega} \frac{2\sigma}{|b_K|} \varepsilon^{1/2}.
\] (3.119)

This bound together with the inequality \( \frac{b^2}{\varepsilon} = \frac{b^2}{\varepsilon} \varepsilon h_K^2 \leq \frac{4\sigma^2}{|b_K|} \varepsilon h_K^2 \) then provides the estimate
\[
\| |\xi_h| \|_b \leq C_{\text{C}} \left( \sum_{K \in \mathcal{T}_h} \varepsilon h_K^2 \left( |u|_{1,K}^2 + |u|_{2,K}^2 \right) \right)^{1/2}
\] (3.120)
with the constant \( C_{\text{C}} \) independent of both \( \varepsilon \) and \( h_K \).

Hence, in this case we obtain the same result as for the SUPG method. Since there holds \( \frac{|b|_1}{2\sigma} h_K^2 \leq \varepsilon h_K^2 \leq \varepsilon_0 h_K^2 \) 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
\[
\| |\eta_h| \|_b \leq \left( \sum_{K \in \mathcal{T}_h} \left( \varepsilon + \frac{\omega_K}{2} h_K^2 + \frac{|b_K|}{2} h_K \right) C_X h_K^2 |u|_{2,K}^2 \right)^{1/2}.
\] (3.121)

Consequently, in the first two cases (A) and B)) we obtain the estimate \( \| |\eta_h| \|_b \leq C_{X1} \left( \sum_{K \in \mathcal{T}_h} h_K^3 |u|_{2,K}^2 \right)^{1/2} \leq C_{X1} \left( \sum_{K \in \mathcal{T}_h} \min \{ h_K, h_K^4/\varepsilon \} |u|_{2,K}^2 \right)^{1/2} \), 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 \( \varepsilon \) and \( h_K \).

The statement of the proof therefore results from the triangular inequality
\[
\| |u - u_h| \|_b = \| |\xi_h + \eta_h| \|_b \leq \| |\xi_h| \|_b + \| |\eta_h| \|_b.
\] (3.122)

When \( \text{div } b = 0 \), the situation gets complicated, since we have to use weaker norm \( \| | \cdot \| \|_{h,*} \) (for all \( v \in H^1(\Omega) \) there holds \( \| |v| \|_{h,*} \leq C^*_b \| |v| \|_h^2 \), providing \( h_K \leq \omega v C^*_b h_K^3 \) for all \( K \in \mathcal{T}_h \)). Consequently, we obtain lower order of convergence with respect to the norm \( \| | \cdot \| \|_{h,*} \).

**Theorem 3.5.2.** Let \( \text{div } b = 0 \) and let there exists \( \delta \in (0, 4) \) such that
\[
\theta_K \leq \frac{\delta}{\sigma_K} |b|_{0,\infty,K} h_K \quad \text{for all } K \in \mathcal{T}_h.
\] (3.123)

Further, let there exist positive numbers \( \kappa, L, R \) and \( \beta \) such that for all \( s = 1, 2, \ldots, \mathcal{P} \) and \( j = 1, 2, \ldots, N_s \) it holds
\[
\frac{|C_j^s|}{|\Omega_j^s|} \geq \kappa, \quad N_s h_j^s \leq L, \quad \frac{\max_{K \subset \Omega_j^s} h_K}{h_j^s} \leq R, \quad \text{and} \quad \beta_j^s \geq \beta.
\] (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 $\varepsilon$ such that for the solution obtained by the method (3.14) there holds
\[ \|u - u_h\|_{b,*} \leq C^* \left( \sum_{K \in T_h} \min \left\{ h_K, \max \left\{ \frac{h_K^4}{\varepsilon}, \varepsilon h_K^2 \right\} \right\} \left( |u|_{2,K}^2 + |u|_{1,K}^2 \right) \right)^{1/2}. \] (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
\[ -\varepsilon (\nabla \eta_h, \nabla \xi_h) \Omega \leq \sum_{K \in T_h} \varepsilon |\eta_h|_{1,K} |\xi_h|_{1,K} \leq C_X \left( \sum_{K \in T_h} \varepsilon h_K^2 |u|_{2,K}^2 \right)^{1/2} \|\xi_h\|_{b,*}. \] (3.126)

2*. \[ \varepsilon \sum_{K \in T_h} (\Delta u, (P_{K,n+1} - C_K) \cdot \nabla \xi_h)_K \leq n \left( \sum_{K \in T_h} \varepsilon h_K^2 |u|_{2,K}^2 \right)^{1/2} \|\xi_h\|_{b,*}. \] (3.127)

3*. Since $\text{div } b = 0$, we have to use the norm $||| \cdot |||_{b,*}$ and hence we lost one half of the order in $h_K$.
\[ \sum_{K \in T_h} \left( (b_K - b_K^I) \cdot \nabla u^I, \xi_h + (P_{K,n+1} - C_K) \cdot \nabla \xi_h \right)_K \leq \sum_{K \in T_h} \theta_K h_K |u^I|_{1,K}(1 + C_{\text{inv}}) \|\xi_h\|_{0,K} \leq \frac{1 + C_{\text{inv}}}{\sqrt{C_2^2}} \left( \sum_{K \in T_h} \theta_K^2 h_K \left( |u|_{1,K} + C_X h_K |u|_{2,K} \right)^2 \right)^{1/2} \|\xi_h\|_{b,*}. \] (3.128)

4*. Again one half of the order is lost for large $h_K$.
\[ \sum_{K \in T_h} \left( (b_K - b) \cdot \nabla u^I, \xi_h + (P_{K,n+1} - C_K) \cdot \nabla \xi_h \right)_K = \sum_{K \in T_h} \left( (b_K - b) \cdot \nabla u^I, \xi_h - \xi_h(C_K) \right)_K \leq \sum_{K \in T_h} n C^2 \left( \frac{h_K^2}{\varepsilon} \right) \left( |u|_{1,K} + C_X h_K |u|_{2,K} \right)^2 \left( \frac{h_K^2}{\varepsilon} \right) \min \left\{ \frac{h_K^2}{\varepsilon}, \frac{C_{\text{inv}}^2}{C_2^2 h_K} \right\} \|\xi_h\|_{b,*}. \] (3.129)

5*. When $\text{div } b = 0$ 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 80, ...
the shape-regularity (Assumption 4.2.1 page 127) and the estimates (3.100), (3.101) we get

\(- (b \cdot \nabla \eta_h, \xi_h)_\Omega = \sum_{K \in T_h} (\eta_h, b_K \cdot \nabla \xi_h)_K + \sum_{K \in T_h} (\eta_h, (b - b_K) \cdot \nabla \xi_h)_K \leq \sum_{K \in T_h} \|\eta_h\|_{0,K} \|b_K \cdot \nabla \xi_h\|_{0,K} + \sum_{K \in T_h} \|\eta_h\|_{0,K} (nC_\Pi |b|_{1,\infty,K} + \theta_K) h_K |\xi_h|_{1,K} \leq \sum_{K \in T_h} \gamma_K C_X^2 h_K^4 |u|_{2,K}^2 \|\xi_h\|_{b,*}, \tag{3.130} \)

where \(\gamma_K = \min \left\{ \frac{2\sigma |b_K|}{C^*_h h_K}, \frac{|b_K|^2}{\varepsilon} \right\} + (nC_\Pi |b|_{1,\infty,K} + \theta_K)^2 \min \left\{ \frac{h_K^2}{\varepsilon}, \frac{C^*_h}{C^*_h h_K} \right\} \).

6*. And the estimate of the last term is

\[- \sum_{K \in T_h} \left( b \cdot \nabla \eta_h, (P_{K,n+1} - C_K) \cdot \nabla \xi_h \right)_K \leq \sum_{K \in T_h} \|b\|_{0,\infty,K} |\eta_h|_{1,K} h_K |\xi_h|_{1,K} \leq \left( \sum_{K \in T_h} |b|_{0,\infty,K}^2 C_X h_K^2 |u|_{2,K}^2 \min \left\{ \frac{h_K^2}{\varepsilon}, \frac{C^*_h}{C^*_h h_K} \right\} \right)^{1/2} \|\xi_h\|_{b,*} \tag{3.131} \]

The derived estimates imply that if \(\theta_K\) is bounded for all \(K \in T_h\) (we even consider \(\theta_K \leq \frac{\varepsilon}{|b|_{0,\infty,K} h_K}\)) 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 \(h_K \geq \varepsilon^{1/3} \sqrt{C^*_h/C^*_2}\). Then using the bound \(\theta_K \leq \frac{\varepsilon}{|b|_{0,\infty,K} h_K}\) together with Corollary 4.2.2 (page 129) we obtain the estimate

\[ \|\xi_h\|_{b,*} \leq C_A^* \left( \sum_{K \in T_h} h_K \left( |u|_{1,K} + |u|_{2,K} \right)^2 \right)^{1/2}, \tag{3.132} \]

where the constant \(C_A^*\) does not depend on \(\varepsilon\) and \(h_K\).

B*) If the mesh is refined and \(\frac{2\sigma}{|b|_{0,\infty,K}^2} \varepsilon \leq h_K \leq \varepsilon^{1/3} \sqrt{C^*_h/C^*_2}\) then using the inequalities \(\varepsilon h_K^2 = \left( \frac{2\sigma}{c^*_h} \right) \frac{h_K^4}{\varepsilon} \leq \left( \frac{|b_K| c^*_h}{2\sigma} \right)^2 \frac{h_K^4}{\varepsilon}, h_K^3 = \frac{\varepsilon}{h_K} \leq \left( \frac{C^*_h}{2\sigma} \right) \frac{h_K^4}{\varepsilon}\) and \(\theta_K^2 h_K \leq \left( \frac{\varepsilon}{|b|_{0,\infty,K}} \right)^2 \frac{h_K^3}{h_K} \) we obtain

\[ \|\xi_h\|_{b,*} \leq C_B^* \left( \sum_{K \in T_h} \frac{h_K^4}{\varepsilon} \left( |u|_{1,K} + |u|_{2,K} \right)^2 \right)^{1/2}, \tag{3.133} \]

where the constant \(C_B^*\) again does not depend on \(\varepsilon\) and \(h_K\).

Since in this case there holds \(\frac{2\sigma}{|b|_{0,\infty,K}^2} h_K^3 \leq \frac{h_K^4}{\varepsilon} \leq h_K^3 \frac{C^*_h}{C^*_2}\), one may expect an acceleration of the convergence in comparison to (3.132).
If the mesh step satisfies \( h_K \leq \varepsilon \frac{2\sigma}{|b_K|C_\varepsilon} \) then using the inequalities \( \frac{h_K^4}{\varepsilon^2} = h_K^2 \varepsilon h_K^2 \left( \frac{2\sigma}{|b_K|C_\varepsilon} \right)^2 \) we obtain

\[
|||\xi_h|||_{b,*} \leq C_C^* \left( \sum_{K \in T_h} \varepsilon h_K^2 \left( |u|_{1,K} + |u|_{2,K} \right)^2 \right)^{1/2},
\]

where the constant \( C_C^* \) does not depend on \( \varepsilon \) and \( h_K \).

Since there holds \( \frac{|b_K|C_\varepsilon}{2\sigma} \varepsilon h_K^3 \leq \varepsilon h_K^2 \leq \varepsilon_0 h_K^2 \) 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

\[
|||\eta_h|||_{b,*} \leq \left( \sum_{K \in T_h} \left( \varepsilon + C_2^* h_K^3 + C_\varepsilon^* \frac{|b_K|}{2} h_K \right) C_X^2 h_K^2 |u|_{2,K}^2 \right)^{1/2}.
\]

Thus, in the first two cases (\( A^* \) and \( B^* \)) we obtain the estimate

\[
|||\eta_h|||_{b,*} \leq C_{X,1}^* \left( \sum_{K \in T_h} h_K^3 |u|_{2,K}^2 \right)^{1/2} \leq C_{X,1}^* \left( \sum_{K \in T_h} \min \{ h_K, h_K^2/\varepsilon \} |u|_{2,K}^2 \right)^{1/2},
\]

whereas in the case \( C^* \) there holds

\[
|||\eta_h|||_{b,*} \leq C_{X,2}^* \left( \sum_{K \in T_h} \varepsilon h_K^2 |u|_{2,K}^2 \right)^{1/2},
\]

where \( C_{X,1}^* \) and \( C_{X,2}^* \) are positive constants independent of \( \varepsilon \) and \( h_K \).

The statement of the proof therefore results from the triangular inequality

\[
|||u - u_h|||_{b,*} = |||\xi_h + \eta_h|||_{b,*} \leq |||\xi_h|||_{b,*} + |||\eta_h|||_{b,*}.
\]

Remark 3.5.2. For each numerical method one can define the experimental order of convergence with respect to some norm \( ||| \cdot ||| \) as

\[
\text{EOC} = \frac{\log(e_h)}{\log(h)}.
\]

for suitable constant \( C \) independent of \( h \). From the above derived a priori error estimates it follows which \( EOC \) we should expect. Since \( \varepsilon \) is constant we use powers of \( \varepsilon \) 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 \( \text{div} \ b < 0 \) and \( \text{div} \ b = 0 \). 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 \( \varepsilon \).
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 $L^\infty$-norm of the matrix $L_h^{-1}$. It is the inverse of the matrix $L_h$, which is obtained from $L_h$ by multiplying each row (corresponding to some basis function $\lambda_k$, $1 \leq k \leq N_h$) by the factor $|\text{supp} \{\lambda_k\}|^{-1}$. Hence, the matrix $L_h$ is defined as follows

$$(L_h)_{ki} = \frac{(L_h)_{ki}}{|\text{supp} \{\lambda_k\}|} = \frac{a_k(\lambda_i,\lambda_k)}{|\text{supp} \{\lambda_k\}|}, \quad \text{for all } 1 \leq i, k \leq N_h. \quad (3.137)$$

In our notation, $\lambda_k$ is $\lambda^s_j$ for suitable $s \in \{1,2,\ldots,P\}$, $j \in \{1,2,\ldots,N_s\}$ and therefore $|\text{supp} \{\lambda_k\}| = |\Omega^s_j|$.

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

$$\frac{\max_{K \in \mathcal{T}_h} h_K}{\min_{K \in \mathcal{T}_h} h_K} \leq Q. \quad (3.138)$$

**Theorem 3.5.3.** Let $\Omega_1$ be a domain such that $\overline{\Omega} \subset \Omega_1$ and let $b \in C^1(\Omega_1)^n$ with $\text{div} b \leq -\omega < 0$. Further, let us assume that all the streamlines of the vector field $b$ leave $\overline{\Omega}$ in finite time (i.e. periodic solutions and points with $b(x) = 0$ are not allowed). Then there exists a positive function $\phi \in C^1(\Omega_1)$ so that $b \cdot \nabla \phi \geq \phi_0 > 0$ in $\overline{\Omega}$. Moreover, if $\theta_K \leq \frac{\omega}{n+1}$ for each $K \in \mathcal{T}_h$ and if there holds

$$\varepsilon \leq \min \left\{ \frac{C_1\phi_0}{|\phi|_{2,\infty,\Omega}}, \frac{C_2|\phi|^2}{|\phi|_{2,\infty,\Omega}}, \frac{C_3|\phi|^2}{|\phi|_{2,\infty,\Omega}} \right\}, \quad (3.139)$$

where $C_1 = \frac{\kappa}{\varepsilon} \min \left\{ \frac{1}{1+\kappa C_X \sigma}, \frac{\beta}{2Q^2 C_X L} \right\}$, $C_2 = \frac{\kappa \beta}{128LQ^2\omega^2}$ and $C_3 = \frac{\kappa \beta}{128LQ^2 C_X |b|_{0,\infty,\Omega}}$, then the matrix of the method (3.14)–(3.16) is an M-matrix and

$$\|L_h^{-1}\|_{\infty,d} \leq \frac{2}{\kappa} \max \left\{ \frac{L}{\beta}, \frac{\|\phi\|_{0,\infty,\Omega}}{\phi_0} \right\}. \quad (3.140)$$

**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), $\sigma$ is the shape-regularity constant (cf. Assumption 4.2.1), $\kappa$ is the mesh structure parameter defined in the definition of the energy norm (Definition 3.4.1 page 68) and $\beta$ is a positive constant satisfying $|b| \geq \beta$ in $\Omega$. Since it holds

$$\|b| - |b_K| \leq |b - b'_K| + |b'_K - b_K| \leq (C_{11} + \theta_K)|b|_{1,\infty,K} h_K, \quad (3.141)$$

we consider $\beta$ being sufficiently small such that there also holds $|b_K| \geq \beta$. 

83
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 $\bar{L}_h$) is a matrix of nonnegative type (Definition 3.2.1 page 57). Hence, it remains to find a vector $e > 0$ for which $L_h e > 0$ (and thus $\bar{L}_h e > 0$, as well). We construct such a vector from the function $\phi$.

The existence of the function $\phi$ follows from Lemma 4.1.1 (page 124). Since any function $\phi_e = \phi + c, c \in \mathbb{R}$, satisfies $b \cdot \nabla \phi_e = b \cdot \nabla \phi \geq \phi_0 > 0$, we can choose $\phi$ in such a way that $\phi > 0$ in $\Omega$ 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 $h_K \geq \left(\frac{2\varepsilon L}{\kappa \beta}\right)^{1/2}$, for each $K \in T_h$, and let us define a function $\psi_h \in V_h$ by the relations

$$
\psi_h(P_s^i) = 0 \quad \forall s = 1, 2, \ldots, P \quad \text{and} \quad \psi_h(P_j^s) = \psi_h(P_{j-1}^s) + h_j^s \quad \forall j = 1, 2, \ldots, N_s.
$$

The function $\psi_h$ is positive inside $\Omega$ and satisfies $b_K \cdot \nabla (\psi_h|_K) = |b_K| \geq \beta$ and $\psi_h(P_j^s) = \sum_{i=1}^{j} h_i^s \leq j \max_{1 \leq i \leq j} h_i^s \leq N_s \max_{1 \leq i \leq N_s} h_i^s \leq \max_{1 \leq i \leq N_s} \max_{1 \leq i \leq N_s} h_i^s = L$ for all $s = 1, 2, \ldots, P, j = 1, 2, \ldots, N_s$. Using these inequalities we obtain

$$
a_h(\psi_h, \lambda_j^s) = -\varepsilon(\nabla \psi_h, \nabla \lambda_j^s)_{\Omega_j^s} + \sum_{K \subset C_j^s} (b_K \cdot \nabla \psi_h, 1)_K \geq \sum_{K \subset C_j^s} \left(\frac{L|\Omega_j^s|}{h_K^2} + |b_K||\Omega_j^s|\right) \geq -\frac{1}{2} \frac{\kappa \beta |\Omega_j^s|}{h_K^2} + |b_K||\Omega_j^s| \geq \frac{1}{2} \kappa \beta |\Omega_j^s|.
$$

Thus for large $h_K$ the matrix of the method is an M-matrix and there holds

$$
\|L_h^{-1}\|_{\infty,d} \leq \frac{\|\psi_h\|_{\infty,d}}{\min_{i}(L_h \psi_h)_i} \leq \frac{2L}{\kappa \beta^2}.
$$

If the mesh is refined and $h_K = \left(\frac{2\varepsilon L}{\kappa \beta}\right)^{1/2}$ for some $K \in T_h$ (and the condition $h_K \geq \left(\frac{2\varepsilon L}{\kappa \beta}\right)^{1/2}$ for each $K \in T_h$ is still valid), then for each $K \in T_h$ it also holds

$$
h_K \leq \max_{K \in T_h} h_K \leq Q \min_{K \in T_h} h_K = Q \left(\frac{2\varepsilon L}{\kappa \beta}\right)^{1/2}.
$$

Hence, it remains to prove the theorem for meshes satisfying $h_K \leq Q \left(\frac{2\varepsilon L}{\kappa \beta}\right)^{1/2}$. In these cases using the inequality (3.139) yields

$$
h_K \leq \frac{1}{8} \min \left\{ \frac{\phi_0}{C_X \|b\|_{0,\infty,\Omega} |\phi|_{2,\infty,\Omega}}, \frac{\phi_0}{\omega |\phi|_{1,\infty,\Omega}}, \left( \frac{8\phi_0}{\omega C_X |\phi|_{2,\infty,\Omega}} \right)^{1/2} \right\}
$$

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 $\phi_h = \Pi_h \phi$

$$\sum_{K \subset \Omega_j^s} (b_K \cdot \nabla \phi_h, \lambda_j^s(P_{K,n+1}))_K = \sum_{K \subset \Omega_j^s} (b_K \cdot \nabla \phi_h, 1)_K =$$

$$= \sum_{K \subset \Omega_j^s} \left\{ (b_K - b_K^I, \nabla \phi_h)_K + (b, \nabla \phi_h)_K \right\} =$$

$$= \sum_{K \subset \Omega_j^s} \left\{ (b_K - b_K^I, \nabla(\phi_h - \phi) + \nabla \phi)_K + (b, \nabla(\phi_h - \phi))_K + (b, \nabla \phi)_K \right\} \geq$$

$$\geq \sum_{K \subset \Omega_j^s} |K| \left\{ -(n - 1) \theta K h_K (|\phi_h - \phi|_{1,\infty,K} + |\phi|_{1,\infty,K}) - \|b\|_{0,\infty,K} |\phi_h - \phi|_{1,\infty,K} + \phi_0 \right\} \geq$$

$$\geq \sum_{K \subset \Omega_j^s} |K| \left\{ \phi_0 - C_X \|b\|_{0,\infty,K} h_K |\phi|_{2,\infty,K} - \omega h_K (C_X h_K |\phi|_{2,\infty,K} + |\phi|_{1,\infty,K}) \right\} \geq$$

$$\geq \sum_{K \subset \Omega_j^s} \left( \phi_0 - \frac{1}{8} \phi_0 - \frac{1}{8} \phi_0 - \frac{1}{8} \phi_0 \right) |K| = \frac{5}{8} \phi_0 |C_j^s| \geq \frac{5}{8} \kappa \phi_0 |\Omega_j^s|. \quad (3.147)$$

From the condition (3.139) it also follows that $\varepsilon (1 + n C_X \sigma) |\phi|_{2,\infty,\Omega} \leq \frac{1}{8} \kappa \phi_0$. Consequently, for arbitrary basis function $\lambda_P \in V_h$ there holds

$$\varepsilon (\nabla \phi_h, \nabla \lambda_P)_{\Omega_P} = \varepsilon (\nabla \phi, \nabla \lambda_P)_{\Omega_P} + \varepsilon (\nabla \phi_h - \nabla \phi, \nabla \lambda_P)_{\Omega_P} =$$

$$= -\varepsilon (\Delta \phi, \lambda_P)_{\Omega_P} + \varepsilon \sum_{K \subset \Omega_P} (\nabla (\phi_h - \phi), \nabla \lambda_P)_K \geq$$

$$\geq -\varepsilon \sum_{K \subset \Omega_P} n |\phi|_{2,\infty,K} |K| \frac{1}{n + 1} - \varepsilon \sum_{K \subset \Omega_P} n |\phi_h - \phi|_{1,\infty,K} \frac{\sigma |K|}{h_K} \geq$$

$$\geq -\varepsilon \sum_{K \subset \Omega_P} \left( n C_X \sigma \right) |\phi|_{2,\infty,K} |K| > -\frac{1}{8} \kappa \phi_0 |\Omega_P|.$$  

Summing two previous estimates together gives the inequality

$$a_h(\phi_h, \lambda_j^s) = \varepsilon (\nabla \phi_h, \nabla \lambda_j^s)_{\Omega_j^s} + \sum_{K \subset \Omega_j^s} (b_K \cdot \nabla \phi_h, \lambda_j^s(P_{K,n+1}))_K \geq \frac{1}{2} \kappa \phi_0 |\Omega_j^s|. \quad (3.148)$$

Thus, for the meshes satisfying $h_K \leq Q \left( \frac{2 \ell}{\kappa d} \right)^{1/2}$ together with the assumption 3.139 we obtain the estimate

$$\| \tilde{L}_h^{-1} \|_{\infty,d} \leq \frac{\| \phi_h \|_{\infty,\Omega}}{\min_i (\tilde{L}_h \phi_h)_i} \leq \frac{2 \| \phi \|_{0,\infty,\Omega}}{\kappa \phi_0}. \quad (3.149)$$

$\Box$

3.6 $L^\infty$-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
\[
v_\Gamma(x) = \exp\left(\frac{-b \cdot n_\Gamma}{\varepsilon} \text{dist}_\Gamma(x)\right),
\]
(3.150)
where $n_\Gamma$ is a unit outer normal to $\Gamma \subset \partial \Omega$ and $\text{dist}_\Gamma(x)$ is a distance of $x \in \Omega$ from $\Gamma$.

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 $\Gamma$ for an equidistant partition of $\Omega$, constant vector field $b$ 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 $\Gamma$. Hence, we define a bilinear form $a_\Gamma$ by the relation
\[
a_\Gamma(u_h, v_h) = a_h(u_h, v_h) + \varepsilon \sum_{K \in T_h} \left( b \cdot \nabla u_h, \frac{\mu}{(b \cdot n_\Gamma)^2} b \cdot \nabla v + \frac{\nu}{b \cdot n_\Gamma} n_\Gamma \cdot \nabla v \right)_K,
\]
(3.151)
where $\mu$ and $\nu$ are real numbers which have yet to be defined. Due to the factor $\varepsilon$ 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 $\Omega$ (in the vicinity of $\Gamma$) by congruent triangles (see Figure 3.5). One of each triangle’s edges is always parallel to $b$ and one is parallel to $\Gamma$. This structure of triangulation is not artificial since it naturally appears when we use the convection-oriented mesh with constant $|d_{K,1}| = h$. Further, we denote the inner angles of each triangle by $\alpha$, $\beta$ and $\gamma$ and choose $\gamma$ to be the angle included by $b$ and $\Gamma$. Thus, $\gamma \leq \frac{\pi}{2}$ and $b \cdot n_\Gamma = |b| \cos(\frac{\pi}{2} - \gamma) = |b| \sin \gamma$.

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 $\lambda_\gamma$ are $P_1(K)$-basis functions corresponding to the vertices with vertex angles $\alpha, \beta$ and $\gamma$, respectively. Since $\alpha < \pi - \gamma, \beta < \pi - \gamma$ and $\gamma \leq \frac{\pi}{2}$ we obtain the estimates

\[
A < -\frac{1}{2} \cot (\pi - \gamma) = \frac{1}{2} \cot \gamma = -C, \tag{3.155}
\]

\[
B < -\frac{1}{2} \cot (\pi - \gamma) = \frac{1}{2} \cot \gamma = -C, \tag{3.156}
\]

\[
C \leq -\frac{1}{2} \cot \frac{\pi}{2} = 0. \tag{3.157}
\]

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

\[
2\varepsilon \begin{bmatrix}
0 & B & C \\
A & -2(A + B + C) & A \\
C & B & 0
\end{bmatrix} - 4\varepsilon(B + C)\text{Pe}_\Gamma \begin{bmatrix}
0 & 0 & 0 \\
0 & 1 & 0 \\
0 & -1 & 0
\end{bmatrix} + \varepsilon(2\mu + \nu)(B + C) \begin{bmatrix}
0 & 1 & 0 \\
0 & -2 & 0 \\
0 & 1 & 0
\end{bmatrix} + \nu \varepsilon \begin{bmatrix}
0 & B & C \\
-\text{Pe}_\Gamma C & -2\text{Pe}_\Gamma & -\text{Pe}_\Gamma C \\
\text{Pe}_\Gamma & \text{Pe}_\Gamma & 0
\end{bmatrix}, \tag{3.158}
\]

where in each row one can find the coefficients $a_\Gamma(\lambda_P, \lambda_S)$ with $P$ having identical $\text{dist}_\Gamma(P)$. For the first row one holds $\text{dist}_\Gamma(P) = \text{dist}_\Gamma(S) - h\frac{b_{n_P}}{|b|} = \text{dist}_\Gamma(S) - h\sin \gamma$, for the second row $\text{dist}_\Gamma(P) = \text{dist}_\Gamma(S)$ and for the third row one has $\text{dist}_\Gamma(P) = \text{dist}_\Gamma(S) + h\frac{b_{n_P}}{|b|} = \text{dist}_\Gamma(S) + h\sin \gamma$. Each column of this stencil then corresponds to a different streamline.

The Péclet number $\text{Pe}_\Gamma$ is defined by the relation $\text{Pe}_\Gamma = \frac{h(b_{n_P})^2}{2|b|}$ and the normal $\mathbf{n}_\Gamma$ satisfies either $\mathbf{n}_\Gamma = -\frac{1}{|\nabla \lambda_{K,1}|} \nabla \lambda_{K,1}$ or $\mathbf{n}_\Gamma = \frac{1}{|\nabla \lambda_{K,3}|} \nabla \lambda_{K,3}$ depending on the orientation of $K$. During the derivation of the stencil we used the equalities

\[
2\frac{|b||K|}{h} = 2\frac{|b|}{h} \frac{v_\alpha h_\alpha}{2} = \frac{|b|}{h} \frac{h(b \cdot \mathbf{n}_\Gamma)}{|b|} (h \cos \gamma + h \sin \gamma \cot \beta) = \frac{(b \cdot \mathbf{n}_\Gamma)^2}{|b|} h (\cot \gamma + \cot \beta) = -\varepsilon(B + C)\text{Pe}_\Gamma, \tag{3.159}
\]

\[
\mathbf{b} \cdot \left( -\frac{1}{|\nabla \lambda_{K,1}|} \nabla \lambda_{K,1} \right) = -\frac{1}{|\nabla \lambda_{K,1}|} \frac{|b|}{|d_{K,1}|} d_{K,1} \cdot \nabla \lambda_{K,1} = \frac{|b|}{h|\nabla \lambda_{K,1}|}, \tag{3.160}
\]

\[
\mathbf{b} \cdot \left( \frac{1}{|\nabla \lambda_{K,3}|} \nabla \lambda_{K,3} \right) = \frac{1}{|\nabla \lambda_{K,3}|} \frac{|b|}{|d_{K,1}|} d_{K,1} \cdot \nabla \lambda_{K,3} = \frac{|b|}{h|\nabla \lambda_{K,3}|} \text{ and } \tag{3.161}
\]

\[
-|K||\nabla \lambda_\alpha|^2 = (\nabla \lambda_\alpha, \nabla \lambda_\beta + \nabla \lambda_\gamma)_K = B + C. \tag{3.162}
\]

87
Choosing the coefficients

As we mentioned earlier, we would like to define the coefficients \( \mu \) and \( \nu \) 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 \( \nu \) respectively, and then sum these multiples together. Consequently, we obtain

\[
a_\Gamma(R_h v_\Gamma, \varphi_S) =
\]

\[
= 2\varepsilon(B + C)v_\Gamma(S)\left[e^{2\Per}(\mu + \nu + 1) - 2(\mu + \nu + 1 + \Per) + e^{-2\Per}(\mu + \nu + 1 + 2\Per)\right] =
\]

\[
= 2\varepsilon(B + C)v_\Gamma(S)\left[(\mu + \nu + 1)4\sinh^2(\Per) - 2\Per e^{-\Per}2\sinh(\Per)\right] =
\]

\[
= 8\varepsilon(B + C)\sinh^2(\Per)v_\Gamma(S)\left[\mu + \nu + 1 - \Per\frac{e^{-\Per}}{\sinh(\Per)}\right] =
\]

\[
= 8\varepsilon(B + C)\sinh^2(\Per)v_\Gamma(S)\left[\mu + \nu + 1 - \Per\left(\coth(\Per) - 1\right)\right].
\] (3.163)

Thus, choosing \( \mu + \nu = \Per = \Per\left(\coth(\Per) - 1\right) - 1 \) 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 \( \Gamma \) (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

\[
\begin{align*}
\mu + \nu &= \Per, \\
C(\nu + 2) &\leq 0, \\
2A - \nu C &\leq 0,
\end{align*}
\] (3.164) (3.165) (3.166)

\[
(2\mu + \nu)(B + C) + B(\nu + 2) &\leq 0.
\] (3.167)

If \( C = 0 \), then from (3.166) and (3.167) it follows that it suffices to consider \( 2A \leq 0 \) and \( 2(\mu + \nu + 1)B = 2(\Per + 1)B \leq 0 \), i.e. \( A \leq 0 \) and \( B \leq 0 \). In the case when \( C < 0 \), then from (3.165)–(3.167) we obtain the conditions \( -2 \leq \nu \leq 2\Per \) and \( \nu \leq 2\Per + 2(\Per + 1)\Per \). If we again assume that \( A \leq 0 \) and \( B \leq 0 \), then \( 0 \leq \frac{A}{C} \) and \( 0 \leq \frac{B}{C} \). Consequently, any \( \nu \) satisfying \( -2 \leq \nu \leq 2\Per \) is admissible. Hence, we choose \( \nu = \nu_1 = 2\Per \) and \( \mu = \mu_1 = \Per - \nu_1 = -\Per \).

The assumption \( A, B, C \leq 0 \) is in fact the assumption (3.8) (page 55) meaning that the angles \( \alpha, \beta \) and \( \gamma \) 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 \( A \leq \Per C = |\Per C| \) and \( B \leq \Per C = |\Per C| \) (together with \( C \leq 0 \)), we fulfill the conditions (3.164)–(3.167) if we take \( \mu = \mu_2 = 2\Per(2\Per + 3) \) and \( \nu = \nu_2 = 2\Per(\Per + 2) \). Indeed,
this case there holds
\[ \mu + \nu = -\mathcal{R}_\Gamma(2\mathcal{R}_\Gamma + 3) + 2\mathcal{R}_\Gamma(\mathcal{R}_\Gamma + 2) = \mathcal{R}_\Gamma, \quad (3.168) \]
\[ C(\nu + 2) = 2C(\mathcal{R}_\Gamma(\mathcal{R}_\Gamma + 2) + 1) = 2C(\mathcal{R}_\Gamma + 1)^2 \leq 0, \quad (3.169) \]
\[ 2A - \nu C = 2(A - \mathcal{R}_\Gamma(\mathcal{R}_\Gamma + 2)C) \leq 2C(\mathcal{R}_\Gamma - \mathcal{R}_\Gamma(\mathcal{R}_\Gamma + 2)) = -2C\mathcal{R}_\Gamma(\mathcal{R}_\Gamma + 1) \leq 0 \quad \text{(3.170)} \]
\[ (2\mu + \nu)(B + C) + B(\nu + 2) = -2\mathcal{R}_\Gamma(\mathcal{R}_\Gamma + 1)(B + C) + 2B(\mathcal{R}_\Gamma + 1)^2 \leq \quad (3.171) \]
\[ \leq -2\mathcal{R}_\Gamma(\mathcal{R}_\Gamma + 1)(\mathcal{R}_\Gamma + 1)C + 2C\mathcal{R}_\Gamma(\mathcal{R}_\Gamma + 1)^2 = 0. \]

All considered cases are summarized in the following table. Let us also mention

<table>
<thead>
<tr>
<th>C ≤</th>
<th>A ≤</th>
<th>method’s property</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>upwind scheme, DMP</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>DMP, UNI</td>
</tr>
<tr>
<td>0</td>
<td>R_\Gamma</td>
<td>DMP, UNI, OBT</td>
</tr>
</tbody>
</table>

| R_\Gamma = \text{Pe}_\Gamma(\coth(\text{Pe}_\Gamma) - 1) - 1 |

Table 3.1: Different choices of the coefficients $\mu$ and $\nu$ 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 $\varepsilon$ in the vicinity of the boundary $\Gamma$ (UNI) and admissibility of obtuse inner angles (OBT).

that in the most interesting case $\text{Pe}_\Gamma \gg 0$ the parameter $\mathcal{R}_\Gamma$ is close to $-1$. Thus, the conditions $A \leq \mathcal{R}_\Gamma C$ and $B \leq \mathcal{R}_\Gamma C$ 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 $\text{Pe}_\Gamma$ (approximately smaller than 3) the angles restriction is more significant (see Figure 3.6).

Figure 3.6: Using the Péclet number $\text{Pe}_\Gamma$ one can compute the parameter $\mathcal{R}_\Gamma$ (left) which restricts the maximum angle in triangles. For instance, when $\text{Pe}_\Gamma \approx 0.6282$, then $\mathcal{R}_\Gamma \approx -1/2$ and the angle $\alpha$ (and $\beta$) has to satisfy the inequality $-\frac{1}{2} \cot \alpha = A \leq -\frac{1}{2} C = \frac{1}{2} \cot \gamma$. Consequently, if (for instance) $\gamma = 60^\circ$ then $\alpha, \beta \leq \text{acot} (-1/(2\sqrt{3})) \approx 106.1^\circ$ (right), which is a sharper bound as compared to standard $\alpha, \beta < 180^\circ - 60^\circ = 120^\circ$. The right figure shows the restriction curves for several choices of $\mathcal{R}_\Gamma$ (or $\text{Pe}_\Gamma$).
3.6.2 Non-constant data

In the previous adjustment of the method we have considered constant data and constant \( \mu \) and \( \nu \). Let us now investigate the non-constant case and consider different values \( \mu_K, \nu_K \) 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 \( \mathbf{n}_K \) be any constant unit vector satisfying \( \mathbf{b}_K \cdot \mathbf{n}_K > 0 \) and let us denote \( \text{Pe}_K = \frac{\| \mathbf{d}_{K,1}(\mathbf{b}_K \cdot \mathbf{n}_K) \|^2}{\varepsilon |\mathbf{b}_K|} \). If the coefficients \( \mu_K \) and \( \nu_K \) satisfy

\[
\text{Pe}_K + \mu_K > \frac{1}{4} \nu_K^2, \tag{3.172}
\]

then there exists \( \alpha > 0 \) such that for all \( v_h \in V_h \) there holds

\[
\left( \mathbf{b}_K \cdot \nabla v_h, \frac{\varepsilon \mu_K}{(\mathbf{b}_K \cdot \mathbf{n}_K)^2} \mathbf{b}_K \cdot \nabla v_h + \frac{\varepsilon \nu_K}{\mathbf{b}_K \cdot \mathbf{n}_K} \mathbf{n}_K \cdot \nabla v_h \right)_K \geq \left( \alpha - 1 \right) \varepsilon |v_h|_{1,K}^2 + (2 \alpha - 1) \frac{|d_{K,1}|}{|b_K|} \| \mathbf{b}_K \cdot \nabla v_h \|^2_{0,K}. \tag{3.173}
\]

**Proof.** Since \( \| \mathbf{n}_K \cdot \nabla v_h \|_{0,K} \leq |v_h|_{1,K} \), then using the Cauchy-Schwarz-Bunyakovskiy inequality we can estimate the left-hand side of (3.173) by

\[
\left( \mathbf{b}_K \cdot \nabla v_h, \frac{\varepsilon \mu_K}{(\mathbf{b}_K \cdot \mathbf{n}_K)^2} \mathbf{b}_K \cdot \nabla v_h + \frac{\varepsilon \nu_K}{\mathbf{b}_K \cdot \mathbf{n}_K} \mathbf{n}_K \cdot \nabla v_h \right)_K \geq \frac{\varepsilon \mu_K}{(\mathbf{b}_K \cdot \mathbf{n}_K)^2} \| \mathbf{b}_K \cdot \nabla v_h \|^2_{0,K} - \frac{\varepsilon |\nu_K|}{\mathbf{b}_K \cdot \mathbf{n}_K} \| \mathbf{b}_K \cdot \nabla v_h \|_{0,K} |v_h|_{1,K}. \tag{3.174}
\]

Thus, denoting \( X = \left( \frac{|d_{K,1}|}{|b_K|} \| \mathbf{b}_K \cdot \nabla v_h \|^2_{0,K} \right)^{1/2} \) and \( Y = (\varepsilon |v_h|_{1,K}^2)^{1/2} \) it suffices to prove that there exists \( \alpha > 0 \) such that for all \( X, Y \geq 0 \) holds

\[
\frac{2 \varepsilon |\mathbf{b}_K|}{|d_{K,1}|(\mathbf{b}_K \cdot \mathbf{n}_K)^2} X^2 - \left( \frac{2 \varepsilon |\mathbf{b}_K|^2}{|d_{K,1}|(\mathbf{b}_K \cdot \mathbf{n}_K)^2} \right)^{1/2} XY \geq (\alpha - 1) Y^2 + (2 \alpha - 1) X^2, \tag{3.175}
\]

which can be rewritten in the form

\[
\left( \frac{\nu_K}{2 \text{Pe}_K^{1/2} X - Y} \right)^2 + \left( \frac{\mu_K}{\text{Pe}_K} + 1 - \frac{\nu_K^2}{4 \text{Pe}_K} \right) X^2 \geq \alpha (Y^2 + 2X^2) \tag{3.176}
\]

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 \( \alpha \) such that the right-hand side remains smaller than the left-hand side.

\[\square\]

**Remark 3.6.1.** The previous lemma did not specify the exact value of \( \alpha \). One possible choice is \( \alpha = 1 - \frac{\mu_0}{\nu_0} \) where \( \mu_0 = \frac{\mu_K}{\text{Pe}_K} + 3 > \frac{\nu_K^2}{4 \text{Pe}_K} + 2 = \nu_0 \). Then \( \frac{\mu_K}{\text{Pe}_K} = \mu_0 - 3, \frac{\nu_K^2}{4 \text{Pe}_K} = \nu_0 - 2 \) and from (3.175) it follows that we need to verify the validity of the inequality

\[
(\mu_0 - 3)X^2 - 2\sqrt{\nu_0 - 2XY} + \frac{\nu_0}{\mu_0} Y^2 + \left( 2 \frac{\nu_0}{\mu_0} - 1 \right) X^2 \geq 0. \tag{3.177}
\]
This inequality can be equivalently rewritten in the form

$$\left( \sqrt{\frac{\mu_0(\nu_0 - 2)}{\nu_0}} X - \sqrt{\frac{\nu_0}{\mu_0}} Y \right)^2 + 2 \left( \sqrt{\frac{\nu_0}{\mu_0}} - \sqrt{\frac{\mu_0}{\nu_0}} \right)^2 X^2 \geq 0. \quad (3.178)$$

The maximal possible value of $\alpha$ can be obtained by investigation of the eigenvalues of the matrix corresponding to the bilinear form

$$G(X, Y) = \left( \frac{\mu_K}{\text{Pe}_K} + (1 - 2\alpha) \right) X^2 - \frac{|\nu_K|}{\text{Pe}_K^{1/2}} XY + (1 - \alpha) Y^2. \quad (3.179)$$

This bilinear form is positive-semidefinite if $\alpha \leq \alpha_0 = \min \left\{ 1, \frac{1}{2} \left( 1 + \frac{\mu_K}{\text{Pe}_K} \right) \right\}$ and $\frac{\nu_K^2}{\text{Pe}_K} \leq 4 \left( \frac{\mu_K}{\text{Pe}_K} + (1 - 2\alpha) \right) (1 - \alpha)$, i.e. in the case when

$$g(\alpha) = 2\alpha^2 - \alpha \left( \frac{\mu_K}{\text{Pe}_K} + 3 \right) + \frac{\mu_K}{\text{Pe}_K} + 1 - \frac{\nu_K^2}{4\text{Pe}_K} \geq 0. \quad (3.180)$$

As we already know, since $g(\alpha_0) = -\frac{\nu_K^2}{4\text{Pe}_K} < 0$, the inequality (3.180) has a solution $\alpha \in (0, \alpha_0)$ only if $g(0) > 0$, i.e. when $\frac{\mu_K}{\text{Pe}_K} + 1 > \frac{\nu_K^2}{4\text{Pe}_K}$. The smaller of the two solutions of the equation $g(\alpha) = 0$ is the maximal possible value of $\alpha$ and it has a form

$$\alpha_{\max}^K = \frac{2 \left( \frac{\mu_K}{\text{Pe}_K} + 1 - \frac{\nu_K^2}{\text{Pe}_K} \right)}{\mu_K + \sqrt{\mu_K - 1} + 2 \frac{\nu_K^2}{\text{Pe}_K}} = \frac{2(\mu_0 - \nu_0)}{\mu_0 + \sqrt{\mu_0 - 8(0 \mu_0 - \nu_0)}}. \quad (3.181)$$

**Definition 3.6.1.** For each $K$ we denote by

$$w_K = (P_{K,n+1} - C_K) + \frac{\varepsilon \mu_K}{(b_K \cdot n_K)^2} b_K + \frac{\varepsilon \nu_K}{b_K \cdot n_K} n_K$$

the vector that we use for stabilization. Further, we define a bilinear form

$$a_h(u, v) = \varepsilon (\nabla u, \nabla v)_\Omega + \sum_{K \in T_h} \left\{ (b_K \cdot \nabla u, v)_K + (-\varepsilon \Delta u, b_K \cdot \nabla u, w_K \cdot \nabla v)_K \right\}.$$

The coefficients $\mu_K$ and $\nu_K$ will be defined later.

**Lemma 3.6.2.** For each $K$ let the assumption (3.172) be fulfilled and let us denote $\alpha_\infty = \min_{K \in T_h} \alpha_{\max}^K$. Then

$$a_h^\infty(v_h, v_h) \geq \alpha_\infty \|v_h\|^2_0. \quad (3.183)$$

**Proof.** From the inequalities (3.64) and (3.66) (page 68) it follows that

$$\sum_{K \in T_h} (b_K \cdot \nabla v_h, v_h + (P_{K,n+1} - C_K) \cdot \nabla v_h)_K \geq \alpha_\infty \frac{|\omega_K|}{2} \|v_h\|^2_0 + (1 - \alpha_\infty) \sum_{K \in T_h} \frac{|d^{K,1}|}{2|b_K|} \|b_K \cdot \nabla v_h\|^2_{0,K}. \quad (3.184)$$
Consequently, using the inequality (3.173) we obtain
\[
a_h^\infty(v_h, v_h) \geq \sum_{K \in \mathcal{T}_h} \left\{ \varepsilon |v_h|^2_{1,K} + \alpha_\infty \frac{\omega_K}{2} ||v_h||^2_{0,K} + (1 - \alpha_\infty) \sum_{K \in \mathcal{T}_h} \frac{|d_{K,1}|}{2|b_K|} ||b_K \cdot \nabla v_h||^2_{0,K} + (\alpha_{max}^K - 1)\varepsilon |v_h|^2_{1,K} + (2\alpha_{max}^K - 1)\frac{|d_{K,1}|}{2|b_K|} ||b_K \cdot \nabla v_h||^2_{0,K} \right\} \geq \alpha_\infty \left\{ \varepsilon |v_h|^2_{1,\Omega} + \frac{\omega_K}{2} ||v_h||^2_{0,\Omega} + \sum_{K \in \mathcal{T}_h} \frac{|d_{K,1}|}{2|b_K|} ||b_K \cdot \nabla v_h||^2_{0,K} \right\}.
\]

(3.185)

Remark 3.6.2. If we take \( \mu_K = \nu_K = 0 \) for each \( K \), then the assumption (3.172) is fulfilled and \( \alpha_{max}^K = \frac{2(0+1-0)}{3+\sqrt{(0-1)^2+0}} = \frac{1}{3} \). 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 \( \alpha_0^{(1)} = 1 - \frac{\nu_1}{\mu_1} = 1 - \frac{2+\nu_1^2/(4\mathcal{R}_\Gamma^2)}{3+\mu_1/\mathcal{R}_\Gamma^2} \) and \( \alpha_0^{(2)} \) 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 \( \mathcal{R}_\Gamma \) 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 \( \mu, \nu \) the estimate (3.183) holds with \( \alpha_{max}^{(1)} \) and \( \alpha_{max}^{(2)} \). For arbitrary \( \mathcal{R}_\Gamma \) there holds \( \alpha_0^{(1)} \geq 0.3839 \) 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 \( \mathcal{R}_\Gamma \).

Lemma 3.6.3. Let there exist positive constants \( C_\mu \) and \( C_\nu \) independent of \( \varepsilon \) and \( h_K \) such that for each \( K \in \mathcal{T}_h \) the coefficients \( \mu_K \) and \( \nu_K \) satisfy
\[
|\mu_K| \leq C_\mu \min\{\mathcal{R}_K, 1\} \quad \text{and} \quad |\nu_K| \leq C_\nu \min\{\mathcal{R}_K, 1\}.
\]

Then there exists positive constant \( C_w > 0 \) independent of \( \varepsilon \) and \( h_K \) such that for each \( K \in \mathcal{T}_h \) the vector \( w_K \) satisfies \( |w_K| \leq C_w h_K \). Moreover, for each \( K \in \mathcal{T}_h \) there also holds \( |w_K - (P_{K,n+1} - C_K)| \to 0 \) as \( \varepsilon \to 0 \).

92
Proof. Since there holds
\[ |w_K| \leq h_K + \frac{1}{2} \frac{|\mu_K|}{Pe_K} h_K + \frac{|\nu_K|}{Pe_K} \frac{b_K \cdot n_K}{2|b_K|} h_K, \]  
(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
\[ |w_K - (P_{K,n+1} - C_K)| \leq \varepsilon \left( \frac{C_\mu |b_K|}{(b_K \cdot n_K)^2} + \frac{C_\nu}{b_K \cdot n_K} \right) \]  
(3.188)
and the fact that \( b_K \cdot n_K > 0 \).

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 \( a_\infty \) 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 \( a_\infty \), Lemma 3.6.3 ensures that we can estimate \( \|w_K \nabla \xi_h\|_{0,K} \leq C_w h_K |\xi_h|_{1,K} \). 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 \( \mu \) and \( \nu \) satisfy the conditions (3.172) and (3.186). Indeed, for every (positive) \( \text{Pe}_\Gamma \) there holds (see Figure 3.8 left)
\[ \text{Pe}_\Gamma + \mu_2 \geq \text{Pe}_\Gamma + \mu_1 \geq \text{Pe}_\Gamma \geq \frac{1}{4} \nu_2^2 \geq \frac{1}{4} \nu_1^2, \]  
(3.189)
where \( \mu_1, \nu_1, \mu_2 \) and \( \nu_2 \) equals to \(-R_\Gamma, 2R_\Gamma, -R_\Gamma(2R_\Gamma + 3) \) and \( 2R_\Gamma(R_\Gamma + 2), \) respectively (cf. Remark 3.6.2 and Table 3.1, page 89).

Further, the coefficients \( \mu_1, \nu_1, \mu_2 \) and \( \nu_2 \) also satisfy the inequalities (see Figure 3.8 middle and right)
\[ \mu_1 \leq \max(\text{Pe}_\Gamma, 1), \]  
(3.190)
\[ |\nu_1| \leq 2 \max(\text{Pe}_\Gamma, 1), \]  
(3.191)
\[ \mu_2 \leq 3 \max(\text{Pe}_\Gamma, 3/8) \leq 3 \max(\text{Pe}_\Gamma, 1), \]  
(3.192)
\[ |\nu_2| \leq 4 \max(\text{Pe}_\Gamma, 1/2) \leq 4 \max(\text{Pe}_\Gamma, 1). \]  
(3.193)

Figure 3.8: For constant data and the three-directional mesh both considered choices of the coefficients \( \mu \) and \( \nu \) satisfy the conditions (3.172) and (3.186).
Choosing the coefficients

Let us now describe, how we choose the coefficients $\mu_K$ and $\nu_K$ for a general oriented mesh (i.e. not necessarily three-directional). At first, we have to define the vectors $n_K$ for $K$ laying inside $\Omega$. For each element $K$ with one edge laying on the boundary $\Gamma_+$ we set $n_K = \frac{1}{|V(\lambda_{K,1})|} \nabla \lambda_{K,1}$. In addition, if $C_{N,s} = K \cup T$ is a cluster containing $K$, we set $n_T = \frac{1}{|V(\lambda_{T,3})|}\nabla \lambda_{T,3}$. If $Q$ is any element laying in some cluster on the same discrete streamline $n$ we set $n_Q = \frac{1}{|V(\lambda_{Q,1})|} \nabla \lambda_{Q,1}$ if $Q$ lies on the same side from this streamline as $K$. Otherwise, we set $n_Q = |V(\lambda_{Q,3})| \nabla \lambda_{Q,3}$ (see Figure 3.9 left). Thus, for a fixed discrete streamline $S$ all elements $K$ with $d_{K,1} \in S$ are divided into two groups: elements laying on one side from $S$ satisfy $n_K = \frac{1}{|V(\lambda_{K,1})|} \nabla \lambda_{K,1}$, elements laying on the other side from $S$ satisfy $n_K = \frac{1}{|V(\lambda_{K,3})|} \nabla \lambda_{K,3}$.

![Figure 3.9: Parts of $T_h$ used for the definition of $n_K$, $\nu_j^s$, $\nu_V$ and $\nu_T$.](image)

For each element $K \in T_h$ we define $Pe_K = \frac{|d_K,1| (|b_K| n_K)^2}{2 |b_K|}$ and for each cluster $C_j^s$ we denote $Pe_j^s = \frac{1}{|C_j|} \sum_{K \in C_j^s} |K| Pe_K$. Consequently, we can define $R_j^s = Pe_j^s \left( \coth(Pe_j^s) - 1 \right) - 1 \in (-1, 0)$. For each element $K \in T_h$ we also define values

$$K_i = -\frac{1}{2} \cot \alpha_{K,i}, \quad i = 1, 2, 3,$$

(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 $C_j^s = V \cup T$ with $n_T = \frac{1}{|V(\lambda_{T,1})|} \nabla \lambda_{T,1}$ and $n_V = \frac{1}{|V(\lambda_{V,3})|} \nabla \lambda_{V,3}$. Further, let $N$ be the element neighboring to $V$ satisfying $n_N = n_V$ and let $Y$ be the remaining element neighboring to $V$. Similarly, we denote by $Z$ the element neighboring to $T$ satisfying $n_Z = n_T$ and by $M$ the remaining element neighboring to $T$ (see Figure 3.9 (right)). For each element $K \in T_h$ we also assume that if $n_K = \frac{1}{|V(\lambda_{K,1})|} \nabla \lambda_{K,1}$ then $K_3 < 0$ and if $n_K = \frac{1}{|V(\lambda_{K,3})|} \nabla \lambda_{K,3}$ then $K_1 < 0$ (corresponding acute angles are highlighted in Figure 3.9 (right)).

Then we denote

$$\nu_j^s := \max \left\{ -1, -1 - \frac{M_1}{T_3}, 2R_j^s(R_j^s + 2) \right\},$$

(3.195)

with the values $Y_3, V_1, M_1$ and $T_3$ computed using (3.194).

Consequently, the coefficients $\nu_V$ and $\nu_T$ are defined as follows

$$\nu_V := \nu_j^s \frac{\max \{V_1, T_3\}}{V_1} \quad \text{and} \quad \nu_T := \nu_j^s \frac{\max \{V_1, T_3\}}{T_3},$$

(3.196)
Since \( 0 > \nu_j^* \geq 2 \mathcal{R}_j^3(\mathcal{R}_j^3 + 2) > -2 \), there holds

\[
\nu_V = \nu_j^* \max \left\{ \frac{V_1}{V_1}, \frac{T_3}{T_3} \right\} = \nu_j^* \min \left\{ \frac{1}{V_1}, \frac{T_3}{T_3} \right\} \geq \nu_j^* > -2, \quad (3.197)
\]

\[
\nu_T = \nu_j^* \max \left\{ \frac{V_1}{T_3}, \frac{T_3}{V_1} \right\} = \nu_j^* \min \left\{ \frac{1}{T_3}, \frac{V_1}{V_1} \right\} \geq \nu_j^* > -2. \quad (3.198)
\]

Therefore, we may define the parameters \( \mathcal{R}_V \) and \( \mathcal{R}_T \) by the relations

\[
\mathcal{R}_V := -1 + \sqrt{1 + \nu_V/2} \quad \Rightarrow \quad \nu_V = 2 \mathcal{R}_V (\mathcal{R}_V + 2), \quad (3.199)
\]

\[
\mathcal{R}_T := -1 + \sqrt{1 + \nu_T/2} \quad \Rightarrow \quad \nu_T = 2 \mathcal{R}_T (\mathcal{R}_T + 2). \quad (3.200)
\]

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 \( T_2, V_3 \leq \mathcal{R}_T T_3 \). 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 \( \mathbf{n}_N = \mathbf{n}_V \) and \( \mathbf{n}_Z = \mathbf{n}_T \), hence we assume some restriction on the angles of the triangles with the same \( \mathbf{n}_K \). Conversely, from this observation it immediately follows that we also assume that \( V_3 \leq \mathcal{R}_N N_3 \) and \( T_1 \leq \mathcal{R}_Z Z_1 \).

![Diagram](image)

**Figure 3.10**: A structure of the domain \( \Omega_S \) (left) and a choice of the vectors \( \mathbf{n}_Q, Q \subset \Omega_S \) (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 \( \lambda_S \in \mathcal{X}_h \), the stencil of the method is computed using values \( a_h^\infty(\lambda_X, \lambda_S) \), where \( \lambda_X \) is the basis function corresponding to the node \( X \) neighboring to \( S \). Let the domain \( \Omega_S = \text{supp} \lambda_S \) be a hexagon depicted in Figure 3.10 (left) and let for each element \( Q \subset \Omega_S \) the vector \( \mathbf{n}_Q \) 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

\[
a_h^\infty(\lambda_A, \lambda_S) = \varepsilon \left( L_2 + (1 + \nu_K) K_2 + (2 \mu_L + \mu_L + \nu_L)(L_1 + L_2) + (2 \mu_K + \mu_K)(K_2 + K_3) \right) = (3.201)
\]

\[
= \frac{\varepsilon}{2} \left( (2 \mu_K + \nu_K + 4 \mu_K)(K_2 + K_3) + (2 + \nu_K)K_2 \right) - \frac{\varepsilon}{2} \nu_K K_3 + \frac{\varepsilon}{2} \left( (2 \mu_L + \nu_L + 4 \mu_L)(L_1 + L_2) + (2 + \nu_L)L_2 \right) + \frac{\varepsilon}{2} \nu_L L_1,
\]

\[
a_h^\infty(\lambda_B, \lambda_S) = \varepsilon \left( L_1 + N_3 + \nu_N N_3 \right), \quad (3.202)
\]

95
\[
a_k^\infty(\lambda_C, \lambda_S) = \varepsilon(N_1 + V_3 - \nu_N N_3),
\]
\[
a_k^\infty(\lambda_D, \lambda_S) = \varepsilon\left(T_2 + (1 + \nu_V) V_2 + (\mu_T + \nu_T)(T_2 + T_3) + \mu_V (V_1 + V_2)\right) = (3.204)
\]
\[
= \frac{\varepsilon}{2}\left\{(2\mu_V + \nu_V)(V_1 + V_2) + (2 + \nu_V)V_2\right\} - \frac{\varepsilon}{2} \nu_V V_1 + \\
+ \frac{\varepsilon}{2}\left\{(2\mu_T + \nu_T)(T_2 + T_3) + (2 + \nu_T)T_2\right\} + \frac{\varepsilon}{2} \nu_T T_3,
\]
\[
a_k^\infty(\lambda_E, \lambda_S) = \varepsilon\left(M_1 + T_3 + \nu_M M_1\right),
\]
\[
a_k^\infty(\lambda_F, \lambda_S) = \varepsilon\left(K_1 + M_3 - \nu_M M_1\right).
\]

From the definition of \(\nu_N\) and \(\nu_M\) 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
\[
L_1 + N_3 + \nu_N N_3 \leq L_1 + N_3 + (-N_3 - L_1) = 0, \quad (3.207)
\]
\[
M_1 + T_3 + \nu_M M_1 \leq M_1 + T_3 + (-M_1 - T_3) = 0. \quad (3.208)
\]

Further, if we take into account the inequalities \(N_1, V_3 \leq R_N N_3\) and \(K_1, M_3 \leq R_M M_1\) we can prove
\[
N_1 + V_3 - \nu_N N_3 \leq R_N N_3 + R_N N_3 - 2R_N (R_N + 2) N_3 = \\
= -2R_N (R_N + 1) N_3 \leq 0, \quad (3.209)
\]
\[
K_1 + M_3 - \nu_M M_1 \leq R_M M_1 + R_M M_1 - 2R_M (R_M + 2) M_1 = \\
= -2R_M (R_M + 1) M_1 \leq 0. \quad (3.210)
\]

Since there holds \(\nu_V V_1 = \nu_T T_3\), the inequality \(a_k^\infty(\lambda_D, \lambda_S) \leq 0\) follows immediately from the estimate (3.171) and the inequalities \(V_2 \leq R_V V_1\) and \(T_2 \leq R_T T_3\).

Similarly, the equality \(\nu_K K_3 = \nu_L L_1\) together with the inequalities \(L_2 \leq R_L L_1\) and \(K_2 \leq R_K K_3\) implies \(a_k^\infty(\lambda_A, \lambda_S) \leq 0\). In fact, in this case the inequality \(a_k^\infty(\lambda_A, \lambda_S) \leq 0\) holds mainly due to the presence of the terms \(P_{kK}\) and \(P_{kL}\).

Figure 3.11: A structure of the domain \(\Omega_S\) in the vicinity of a corner of \(\Omega\).

Let us also briefly mention how we choose the coefficients \(\mu\) and \(\nu\) in the vicinity of a domain corner. Since we consider only convex domains, the boundary cluster laying in the corner of the domain \(\Omega\) is formed by two elements \(V\) and \(T\) with \(n_V = \frac{1}{|\nabla \lambda_{V,1}|} \nabla \lambda_{V,1}\) and \(n_T = \frac{1}{|\nabla \lambda_{T,1}|} \nabla \lambda_{T,1}\). Hence, for each element \(Q\) from any cluster laying on the same discrete streamline there holds \(n_Q = \frac{1}{|\nabla \lambda_{Q,1}|} \nabla \lambda_{Q,1}\)
as well (see Figure 3.11). This changes the stencil of the method from the previous paragraph, it now takes the form

\[
a_h^\infty(\lambda_A, \lambda_S) = \varepsilon \left( (1 + \nu_L)L_2 + (1 + \nu_K)K_2 + (2\text{Pe}_L + \mu_L)(L_2 + L_3) + (2\mu_K)(K_2 + K_3) \right),
\]

\[
a_h^\infty(\lambda_B, \lambda_S) = \varepsilon \left( L_1 + N_3 - \nu_N N_1 \right),
\]

\[
a_h^\infty(\lambda_C, \lambda_S) = \varepsilon \left( N_1 + V_3 + \nu_N N_1 \right),
\]

\[
a_h^\infty(\lambda_D, \lambda_S) = \varepsilon \left( T_2 + V_2 + (\mu_T + \nu_T)(T_2 + T_3) + (\mu_V + \nu_V)(V_2 + V_3) \right),
\]

\[
a_h^\infty(\lambda_E, \lambda_S) = \varepsilon \left( M_1 + T_3 + \nu_M M_1 \right),
\]

\[
a_h^\infty(\lambda_F, \lambda_S) = \varepsilon \left( K_1 + M_3 - \nu_M M_1 \right).
\]

Since the boundary layer function near the domain corner is different from the exponential boundary layer function, we no longer require \( \mu + \nu = \mathcal{R}_V \) (moreover, we cannot determine which part of \( \Gamma_+ \) we should use). Hence, for any triangle \( V \) laying on the corner discrete streamline we set \( \mathcal{R}_V = \text{Pe}_V (\cot \text{Pe}_V - 1) - 1 \) and define

\[
\nu_V = \max \left\{ -1 - \frac{N_1}{V_3}, 2\mathcal{R}_V (\mathcal{R}_V + 2) \right\},
\]

where \( N \) is an element laying on the same side from the corner discrete streamline and having different unit vector, i.e. \( n_N \neq n_V \). Finally, we define the coefficient \( \mu_V \) as

\[
\mu_V = \frac{-\mathcal{R}_V}{1 - \rho} - \nu_V,
\]

where \( \rho \in (0, 1) \) is such that \( V_2 \leq \rho \mathcal{R}_V V_3 < \mathcal{R}_V V_3 \). 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

\[
a_h^\infty(\lambda_D, \lambda_S) = \varepsilon \left( T_2 + V_2 - \frac{\mathcal{R}_T}{1 - \rho} (T_2 + T_3) - \frac{\mathcal{R}_V}{1 - \rho} (V_2 + V_3) \right) \leq
\]

\[
\leq \varepsilon \left( \rho \mathcal{R}_T - \frac{\mathcal{R}_T}{1 - \rho} (\rho \mathcal{R}_T + 1) \right) T_3 + \left( \rho \mathcal{R}_V - \frac{\mathcal{R}_V}{1 - \rho} (\rho \mathcal{R}_V + 1) \right) V_3 =
\]

\[
= \varepsilon \left( \rho - \frac{1 - |\rho \mathcal{R}_T|}{1 - \rho} \right) \mathcal{R}_T T_3 + \left( \rho - \frac{1 - |\rho \mathcal{R}_V|}{1 - \rho} \right) \mathcal{R}_V V_3 \leq 0,
\]

since \( \mathcal{R}_T T_3, \mathcal{R}_V V_3 > 0 \) and \( \rho < 1 \). Hence, \( |\rho \mathcal{R}_V| < 1 \).

### 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

\[ X_h(2) = \{ v_h \in C(\Omega), v_h|_K \in P_2(K), \forall K \in T_h \} \]

and we are looking for \( u_h \in V_h(2) = X_h(2) \cap H_0^1(\Omega) \) such that

\[ a_h(2)(u_h, v_h) = F_h(2)(v_h) \quad \text{for all} \quad v_h \in V_h(2), \quad (3.219) \]

where the bilinear form \( a_h(2) \) and the functional \( F_h(2) \) are defined as

\[
a_h(2)(u, v) = \varepsilon(\nabla u, \nabla v)_\Omega + \sum_{K \in T_h} \left( b_K^{(1)} \cdot \nabla u, v \right)_K + \\
+ \sum_{K \in T_h} \left( -\varepsilon \Delta u + b_K^{(1)} \cdot \nabla u, R_K^{(2)} v - \Pi_{b,K}^{(2)} v \right)_K \quad \text{and} \quad (3.220)
\]

\[
F_h(2)(v) = \sum_{K \in T_h} \left( f, v + R_K^{(2)} v - \Pi_{b,K}^{(2)} v \right)_K. \quad (3.221)
\]

Thus, the definition of the method is an analog to the \( P_1 \)-case: \( b_K^{(1)} \) is a piecewise polynomial (linear in each component) vector function parallel with \( d_K \) on each \( K \in T_h \) 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 \( K \in T_h \) we denote \( P_{K,1}^{(2)} = P_{K,1}^{(2)}, P_{K,3}^{(2)} = P_{K,2}^{(2)} \) and \( P_{K,6}^{(2)} = P_{K,3}^{(2)} \) the corner nodes of \( K \). The midpoints of edges \( P_{K,1}^{(2)}P_{K,2}, P_{K,2}P_{K,3}, \) and \( P_{K,3}P_{K,1} \) 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 \( P_2(K) \).](image)

The basis functions \( \{ \varphi_{K,i}^{(2)} \}_{i=1}^6 \) of the space \( P_2(K) \) are standardly defined by the relations

\[
\varphi_{K,i}^{(2)}(P_{K,j}^{(2)}) = \delta_{ij} \quad \text{for all} \quad i, j \in \{1, 2, \ldots, 6\}.
\]

### 3.7.1 Definition and properties of the discretized vector field

Prior to \( b_K^{(1)} \), we firstly define on each \( K \in T_h \) an interpolation \( b_K' \in P_1(K)^2 \) of the vector \( b \). We define \( b_K' \) as an orthogonal \( L^2 \)-projection of the vector \( b \) on the space \( P_1(K)^2 \). Hence, it satisfies the equality

\[
\left( b - b_K', \varphi \right)_{L^2(K)^2} = 0, \quad \text{for all} \quad \varphi \in P_1(K)^2. \quad (3.222)
\]
We can construct this orthogonal $L^2$-projection, for instance, by considering a basis of the space $P_r(K)^2$ in the form $\{\lambda_{K,l} \nabla \lambda_{K,j}, j = 1, 2, l = 1, 2, 3\}$ or $\{\lambda_{K,j} d_{K,j}, j = 1, 2, l = 1, 2, 3\}$, where $d_{K,j} = P_{K,3} - P_{K,j}$ for $j = 1, 2$. Consequently, we can express the vector $b^I_K$ as

$$b^I_K = 2 \left( \sum_{j=1}^2 \sum_{i=1}^3 \alpha^{(j)}_{K,i} \lambda_{K,i} \right) d_{K,j},$$

(3.223)

where $\alpha^{(j)}_{K,i}$, $j = 1, 2$ and $i = 1, 2, 3$, are for each $j = 1, 2$ solutions of the system of linear equations with the same nonsingular matrix

$$\sum_{i=1}^3 \alpha^{(j)}_{K,i} (\lambda_{K,i}, \lambda_{K,l})_K = - (b \cdot \nabla \lambda_{K,j}, \lambda_{K,l})_K, \quad l = 1, 2, 3.$$  

(3.224)

These systems are obtained by considering the vector $b^I_K$ in the form (3.223) and testing the difference in (3.222) by functions $\phi$ 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

$$\alpha^{(j)}_{K,i} = \frac{3}{|K|} (-b \cdot \nabla \lambda_{K,j}, 4\lambda_{K,i} - 1)_K.$$  

(3.225)

Since $b^I_K$ is the orthogonal $L^2$-projection of $b$, 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, $n, r \in \mathbb{N}$, $b \in W^{r+1,\infty}(K)^n$ and $b^I_K \in P_r(K)^n$ satisfies for all $\phi \in P_r(K)^n$

$$\left( b - b^I_K, \phi \right)_{L^2(K)^n} = 0.$$  

(3.226)

If the shape-regularity assumption (4.23) is fulfilled, then there exists a constant $c > 0$ depending only on $n$, $r$ and $\sigma$ such that for all $v_h \in P_{r+1}(K)$ holds

$$\left\| (b - b^I_K) \cdot \nabla v_h \right\|_{0,K} \leq c h^{r+1}_K |b|_{r+1,\infty,K} |v_h|_{1,K}.$$  

(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

$$\left\| (b - b^I_K) \cdot \nabla v_h \right\|_{0,K} = \sum_{i=1}^n \left\| b - b^I_K \right\|_{0,K} \partial v_h \leq \sum_{i=1}^n \left\| b - b^I_K \right\|_{0,K} \partial v_h \leq \sum_{i=1}^n \left\| \nabla v_h \right\|_{0,K} \leq \left( \sum_{i=1}^n \left\| b - b^I_K \right\|_{0,K}^2 \right)^{\frac{1}{2}} \left( \sum_{i=1}^n \left\| \nabla v_h \right\|_{0,K}^2 \right)^{\frac{1}{2}} = \left\| b - b^I_K \right\|_{0,K} |v_h|_{1,K} \leq C h^{r+1}_K |b|_{r+1,\infty,K} |v_h|_{1,K}.$$  

(3.228)

$\square$
Lemma 3.7.2. Let \( n, r \in \mathbb{N} \) and let \( b^l_K \in P_r(K)^n \) be defined by

\[
b^l_K = \sum_{j=1}^{n} \left( \sum_{i=1}^{N_r} \alpha^{(j)}_{K,i} \varphi^{(r)}_{K,i} \right) d_{K,j},
\]

(3.229)

where \( \alpha^{(j)}_{K,i} \in \mathbb{R} \) are suitable coefficients. If we denote \( b^{(r)}_K = \left( \sum_{i=1}^{N_r} \alpha^{(1)}_{K,i} \varphi^{(r)}_{K,i} \right) d_{K,1}, \) then

\[
\left\| \left( b^l_K - b^{(r)}_K \right) \cdot \nabla v_h \right\|_{0,K} \leq \left( \sum_{j=2}^{n} \left( \sum_{i=1}^{N_r} \alpha^{(j)}_{K,i} \varphi^{(r)}_{K,i} \right) \right) h_K |v_h|_{1,K}.
\]

(3.230)

Proof. It results immediately from the fact that \( |d_{K,j}| \leq h_K \) for \( j = 1, 2, \ldots, n \).

\[ \square \]

Remark 3.7.1. In order to fulfil the equality (3.226) the coefficients \( \alpha^{(j)}_{K,i} \) have to be the solutions of \( n \) systems of linear equations (i.e. \( j = 1, 2, \ldots, n \))

\[
\sum_{i=1}^{N_r} \alpha^{(j)}_{K,i} \left( \varphi^{(r)}_{K,i}, \varphi^{(r)}_{K,m} \right)_K = \left( -b \cdot \nabla \lambda_{K,j}, \varphi^{(r)}_{K,m} \right)_K, \quad m = 1, 2, \ldots, N_r.
\]

(3.231)

Let us turn back to the case of piecewise quadratic functions in 2D. As in Lemma 3.7.2 we denote by \( b^{(1)}_K = \left( \sum_{i=1}^{N_r} \alpha^{(1)}_{K,i} \varphi^{(1)}_{K,i} \right) d_{K,1} \) the first term of the sum (3.229), i.e., the first term of \( b^l_K \). Since the bilinear form \( a_h^{(2)} \) (cf. (3.220)) contains terms \( \left( b^{(1)}_K \cdot \nabla u_h, v_h \right)_K \), with \( v_h \in P_1(K) \) and \( u_h \in P_2(K) \), we compute the values \( \left( b^{(1)}_K \cdot \nabla u_h, \lambda_{K,j} \right)_K \) for \( j = 1, 2, 3 \)

\[
\left( b^{(1)}_K \cdot \nabla u_h, \lambda_{K,j} \right)_K = \sum_{i=1}^{6} u_{K,i} \left( b^{(1)}_K \cdot \nabla \varphi^{(2)}_{K,i}, \lambda_{K,j} \right)_K =
\]

\[
= (u_{K,4} - u_{K,1}) \left( -b^{l}_{K} \cdot \nabla \lambda_{K,1}, (4\lambda_{K,1} - 1)\lambda_{K,j} \right)_K + \quad \text{(3.232)}
\]

\[
+ (u_{K,5} - u_{K,2}) \left( -b^{l}_{K} \cdot \nabla \lambda_{K,1}, (4\lambda_{K,2} - 1)\lambda_{K,j} \right)_K + \quad \text{(3.233)}
\]

\[
+ (u_{K,6} - u_{K,3}) \left( -b^{l}_{K} \cdot \nabla \lambda_{K,1}, (4\lambda_{K,3} - 1)\lambda_{K,j} \right)_K = \quad \text{(3.234)}
\]

\[
= (u_{K,4} - u_{K,1}) s_{1j}^K + (u_{K,5} - u_{K,2}) s_{2j}^K + (u_{K,6} - u_{K,3}) s_{3j}^K, \quad \text{(3.235)}
\]

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 \( u_{K,i} = u_h(P^{(2)}_{K,i}) \) 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 \( \nabla \varphi^{(2)}_{K,6} \) in the form

\[
\nabla \varphi^{(2)}_{K,6} = \nabla \left( \lambda_{K,3}(2\lambda_{K,3} - 1) \right) = \nabla \lambda_{K,3}(4\lambda_{K,3} - 1) = -\left( \nabla \lambda_{K,1} + \nabla \lambda_{K,2} \right)(4\lambda_{K,3} - 1),
\]

(3.236)

which results in

\[
\left( b^{(1)}_K \cdot \nabla \varphi^{(2)}_{K,6}, \lambda_{K,j} \right)_K = \left( -b^{(1)}_K \cdot \nabla \lambda_{K,1}, (4\lambda_{K,3} - 1)\lambda_{K,j} \right)_K =
\]

\[
= \left( -b^{(1)}_K \cdot \nabla \lambda_{K,1}, (4\lambda_{K,3} - 1)\lambda_{K,j} \right)_K, \quad \text{(3.237)}
\]

where we applied the equality \( d_{K,j} \cdot \nabla \lambda_{K,j} = -\delta_{ij} \) 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

\[ (b_K^{(1)} \cdot \nabla \varphi_{K,i}, \Pi_{b,K}^{(2)}(\varphi_{K,j}) - \varphi_{K,j})_K = 0 \quad \text{for all } i, j = 1, 2, \ldots, 6. \]  

(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 $b_K^{(1)} \cdot \nabla \varphi = 0$ for $\varphi = 1, \lambda_{K,2}, \lambda_{K,3}$, only three of these six equations are linearly independent. Thus, for each $\varphi_{K,j} \in P_2(K)$ we define $\Pi_{b,K}^{(2)}(\varphi_{K,j}) = \sum_{m=1}^{3} \mu_{K,m}^{(j)} \lambda_{K,m}$, where $\mu_{K,m}^{(j)}$ are solutions of the systems of linear equations

\[ \sum_{m=1}^{3} \mu_{K,m}^{(j)} (-b_K^{(1)} \cdot \nabla \lambda_{K,1}, \lambda_{K,l} \lambda_{K,m})_K = (-b_K^{(1)} \cdot \nabla \lambda_{K,1}, \lambda_{K,l} \lambda_{K,m}^{(2)})_K, \quad l = 1, 2, 3. \]  

(3.239)

If $-b_K^{(1)} \cdot \nabla \lambda_{K,1} > 0$ almost everywhere in $K$, then we obtain for arbitrary $v_1, v_2, v_3 \in \mathbb{R}$ the inequality

\[ \sum_{l,m=1}^{3} v_l v_m (-b_K^{(1)} \cdot \nabla \lambda_{K,1}, \lambda_{K,l}, \lambda_{K,m})_K = (-b_K^{(1)} \cdot \nabla \lambda_{K,1}, \left( \sum_{l=1}^{3} v_l \lambda_{K,l} \right)^2)_K > 0, \]  

(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 $b_K^{(1)}$ is constant vector then $-b_K^{(1)} \cdot \nabla \lambda_{K,1} = \frac{|b_K^{(1)}|}{d_K^{(1)}} > 0$ and $\Pi_{b,K}^{(2)}$ is the orthogonal $L^2$-projection.)

In addition, using the equalities $b_K^{(1)} \cdot \nabla \lambda_{K,2} = 0$ and $b_K^{(1)} \cdot \nabla \lambda_{K,3} = -b_K^{(1)} \cdot \nabla \lambda_{K,5} = -b_K^{(1)} \cdot \nabla \lambda_{K,6}$ we deduce that $b_K^{(1)} \cdot \nabla \varphi_{K,i}^{(2)} = b_K^{(1)} \cdot \nabla \lambda_{K,1} (\nu_1 \lambda_{K,1} + \nu_2 \lambda_{K,2} + \nu_3 \lambda_{K,3})$ for suitable real values $\nu_1, \nu_2, \nu_3$ depending on $i$. This and the definition of the values $\mu_{K,m}^{(j)}$ yields the equality

\[ (-b_K^{(1)} \cdot \nabla \varphi_{K,i}^{(2)}, \sum_{m=1}^{3} \mu_{K,m}^{(j)} \lambda_{K,m} - \varphi_{K,j}^{(2)})_K = 0. \quad \text{for all } i = 1, 2, \ldots, 6, \]  

(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

\[ \left[ \Pi_{b,K}^{(2)}(v_h) \right]_{M_K^{(1)}} = \left( \begin{array}{ccc} \mu_{K,1}^{(1)} & \mu_{K,1}^{(2)} & \cdots & \mu_{K,1}^{(6)} \\ \mu_{K,2}^{(1)} & \mu_{K,2}^{(2)} & \cdots & \mu_{K,2}^{(6)} \\ \mu_{K,3}^{(1)} & \mu_{K,3}^{(2)} & \cdots & \mu_{K,3}^{(6)} \end{array} \right) \left[ v_h \right]_{M_K^{(2)}}, \]

(3.242)

where $\left[ \Pi_{b,K}^{(2)}(v_h) \right]_{M_K^{(1)}}$ and $\left[ v_h \right]_{M_K^{(2)}}$ are coordinates of the functions $\Pi_{b,K}^{(2)}(v_h)$ and $v_h$ with respect to the bases $M_K^{(1)} = \{ \lambda_{K,j} \}_{j=1}^{3}$ and $M_K^{(2)} = \{ \varphi_{K,j} \}_{j=1}^{6}$ of the spaces $P_1(K)$ and $P_2(K)$, respectively.

101
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
\[
\sum_{m=1}^{3} \sum_{j=1}^{6} \mu_{K,m}^{(j)} \left( -b_{K}^{(1)} \cdot \nabla \lambda_{K,1}, \lambda_{K,l} \lambda_{K,m} \right)_{K} = \left( -b_{K}^{(1)} \cdot \nabla \lambda_{K,1}, \lambda_{K,l} \sum_{m=1}^{3} \lambda_{K,m} \right)_{K},
\]
which can be rewritten in the form
\[
\sum_{m=1}^{3} \left( 1 - \sum_{j=1}^{6} \mu_{K,m}^{(j)} \right) \left( -b_{K}^{(1)} \cdot \nabla \lambda_{K,1}, \lambda_{K,l} \lambda_{K,m} \right)_{K} = 0, \quad \forall l = 1, 2, 3. \tag{3.244}
\]

It means that for \(-b_{K}^{(1)} \cdot \nabla \lambda_{K,1} > 0\) 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 \( \Pi_{2}^{(2)} \) preserves polynomials of degree 0, i.e. constants. When the sign of \(-b_{K}^{(1)} \cdot \nabla \lambda_{K,1}\) 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
\[
\begin{pmatrix}
\mu_{K,1}^{(1)}, \mu_{K,1}^{(2)}, \cdots, \mu_{K,1}^{(6)} \\
\mu_{K,2}^{(1)}, \mu_{K,2}^{(2)}, \cdots, \mu_{K,2}^{(6)} \\
\mu_{K,3}^{(1)}, \mu_{K,3}^{(2)}, \cdots, \mu_{K,3}^{(6)}
\end{pmatrix}
= \begin{pmatrix}
2/5, 3/5, -1/5, 3/5, -1/5, -1/5 \\
-1/5, 3/5, 2/5, -1/5, 3/5, -1/5 \\
-1/5, -1/5, -1/5, 3/5, 3/5, 2/5
\end{pmatrix}, \tag{3.245}
\]
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 \( \left( b_K^{(1)} \cdot \nabla u_h, R_K^{(2)}(\varphi_{K,m}^{(2)}) \right)_K = 0 \), 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
\[
[R_K^{(2)}(u_h)]_{M^{(1)}_K} = \begin{pmatrix}
0, 0, 0, r_{14}^K, r_{15}^K, r_{16}^K \\
0, 0, 0, r_{24}^K, r_{25}^K, r_{26}^K \\
0, 0, 0, r_{34}^K, r_{35}^K, r_{36}^K
\end{pmatrix} [u_h]_{M^{(2)}_K}, \tag{3.246}
\]

102
Using the expression \(3.235\) we can further write

\[
(b_K^{(1)} \cdot \nabla u_h, R_K^{(2)} (\tilde{\zeta}_{K,m}))_K = \sum_{j=1}^3 r_{jm}^{K} (b_K^{(1)} \cdot \nabla u_h, \lambda_{K,j})_K =
\]

\[
= (u_{K,1} - u_{K,1}) \sum_{j=1}^3 s_{1j}^K r_j^{K} + (u_{K,5} - u_{K,2}) \sum_{j=1}^3 s_{2j}^K r_j^{K} + (u_{K,6} - u_{K,4}) \sum_{j=1}^3 s_{3j}^K r_j^{K}.
\]

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 \(m = 1, 2, 3\) 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

\[
\begin{pmatrix}
    s_{11}^K & s_{12}^K & s_{13}^K \\
    s_{21}^K & s_{22}^K & s_{23}^K \\
    s_{31}^K & s_{32}^K & s_{33}^K
\end{pmatrix}
\begin{pmatrix}
    r_{14}^K & r_{15}^K & r_{16}^K \\
    r_{24}^K & r_{25}^K & r_{26}^K \\
    r_{34}^K & r_{35}^K & r_{36}^K
\end{pmatrix}
= \begin{pmatrix}
    \kappa_{14}^K & 0 & \kappa_{16}^K \\
    0 & \kappa_{25}^K & 0 \\
    0 & 0 & \kappa_{36}^K
\end{pmatrix},
\]

(3.247)

where \(\kappa_{14}^K, \kappa_{16}^K, \kappa_{25}^K\) and \(\kappa_{36}^K\) 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 \(\Pi_{b,K} v_h = v_h\) 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

\[
\kappa_{14}^K + \kappa_{16}^K = \sum_{j=1}^3 s_{1j}^K = \left( -b \cdot \nabla \lambda_K, 1 \right)_{0,K},
\]

(3.248)

\[
\kappa_{25}^K = \sum_{j=1}^3 s_{2j}^K = \left( -b \cdot \nabla \lambda_K, 2 \right)_{0,K},
\]

(3.249)

\[
\kappa_{36}^K = \sum_{j=1}^3 s_{3j}^K = \left( -b \cdot \nabla \lambda_K, 3 \right)_{0,K}.
\]

(3.250)

Thus, it remains to determine either \(\kappa_{14}^K\) or \(\kappa_{16}^K\), 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 \(b \cdot \nabla \lambda_K,1\) is constant and nonzero on \(K\), then the matrix \((r_{ij}^K)_{i,j=1}^3\) is a solution of the matrix equation

\[
\begin{pmatrix}
    1/3, & 0, & 0 \\
    1/3, & 2/3, & 1/3 \\
    0, & 0, & 1/3
\end{pmatrix}
\begin{pmatrix}
    r_{14}^K & r_{15}^K & r_{16}^K \\
    r_{24}^K & r_{25}^K & r_{26}^K \\
    r_{34}^K & r_{35}^K & r_{36}^K
\end{pmatrix}
= \begin{pmatrix}
    (1 - \gamma_K)/3, & 0, & \gamma_K/3 \\
    0, & 4/3, & 0 \\
    0, & 0, & 1/3
\end{pmatrix},
\]

(3.251)
where $\gamma_K = -\frac{3}{|K|} b_K \lambda_{2K}^{j}$. Solution of this equation has the form

$$
\begin{pmatrix}
v_{14}^{K}, v_{15}^{K}, v_{16}^{K} \\
v_{24}^{K}, v_{25}^{K}, v_{26}^{K} \\
v_{34}^{K}, v_{35}^{K}, v_{36}^{K}
\end{pmatrix}
= \begin{pmatrix}
1 - \gamma_K, 0, -1 + \gamma_K \\
\frac{1}{2}, 2, 0 \\
0, 0, 1
\end{pmatrix}
$$

(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 $\|((\Pi^{(2)}_{K} - R^{(2)}_{K})v_h)_{0,K}\| \leq C(b)v_h|_{1,K}$, 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 $R^{(2)}_{K}$ affects the stability of the method.

### 3.7.4 Stability of the Method

From the construction of the mapping $R^{(2)}_{K}$ it follows that for each $v_h \in V_h$ holds

$$
(b^{(1)}_{K} \cdot \nabla v_h, R^{(2)}_{K}(v_h))_{K} = \sum_{m=4}^{6} v_{K,m} (b^{(1)}_{K} \cdot \nabla v_h, R^{(2)}_{K}(\phi_{K,m}))_{K} =
$$

$$
= (v_{K,4} - v_{K,1})(\kappa_{14}^{K} v_{K,4} + \kappa_{16}^{K} v_{K,6}) + (v_{K,5} - v_{K,2})\kappa_{25}^{K} v_{K,5} + (v_{K,6} - v_{K,4})\kappa_{36}^{K} v_{K,6}.
$$

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 $P_{K,2}$ and $P_{K,5}$ (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

$$
\mathbb{B}_{s} = 
\begin{pmatrix}
\kappa_{14}^{1}, 0, 0, 0, 0, \cdots, 0, 0 \\
0, \kappa_{14}^{1} - \kappa_{36}^{1}, \kappa_{36}^{1}, 0, 0, \cdots, 0, 0 \\
0, 0, -\kappa_{14}^{2}, \kappa_{14}^{2}, 0, 0, \cdots, 0, 0 \\
0, 0, 0, -\kappa_{14}^{2}, \kappa_{14}^{2} - \kappa_{36}^{2}, \kappa_{36}^{2}, 0, 0, \cdots, 0, 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0, 0, 0, 0, 0, \cdots, \kappa_{14}^{N-1}, 0, 0 \\
0, 0, 0, 0, 0, \cdots, \kappa_{14}^{N-1}, \kappa_{14}^{N-1}, 0, 0, \cdots, 0, 0 \\
0, 0, 0, 0, 0, \cdots, 0, 0, \kappa_{14}^{N}, \kappa_{14}^{N}
\end{pmatrix},
$$

(3.253)

where $\kappa_{pq}^{j} = \sum_{K \subset C_{j}^{s}} \kappa_{pq}^{K}$ 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

\[
\begin{pmatrix}
1/\kappa_{14}^1, & 0, & 0, & 0, & 0, & \ldots & 0 \\
\kappa_{36}^1 - \kappa_{16}^1 / \kappa_{14}^1 & 1/\kappa_{36}^1, & 0, & 0, & 0, & \ldots & 0 \\
\kappa_{36}^1 - \kappa_{16}^1 / \kappa_{14}^1 & 1/\kappa_{36}^1, & 1/\kappa_{2}^2, & 0, & 0, & \ldots & 0 \\
\kappa_{36}^1 - \kappa_{16}^1 / \kappa_{14}^1 & 0, & \kappa_{36}^1 - \kappa_{16}^1 / \kappa_{14}^1 & 1/\kappa_{2}^2, & 0, & \ldots & 0 \\
\kappa_{36}^1 - \kappa_{16}^1 / \kappa_{14}^1 & 1/\kappa_{36}^1, & \kappa_{36}^1 - \kappa_{16}^1 / \kappa_{14}^1 & 1/\kappa_{2}^2, & 1/\kappa_{14}^1, & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\kappa_{36}^1 - \kappa_{16}^1 / \kappa_{14}^1 & 1/\kappa_{36}^1, & \kappa_{36}^1 - \kappa_{16}^1 / \kappa_{14}^1 & 0, & 0, & \ldots & 1/\kappa_{36}^{N-1}, 0 \\
\kappa_{36}^1 - \kappa_{16}^1 / \kappa_{14}^1 & 1/\kappa_{36}^1, & \kappa_{36}^1 - \kappa_{16}^1 / \kappa_{14}^1 & 0, & 1/\kappa_{36}^{N-1}, & 1/\kappa_{14}^1 \\
\end{pmatrix} 
\]

(3.254)

Since we expect \( \varepsilon \) 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 \( \kappa_{14}^1 > 0, \kappa_{25}^2 > 0, \kappa_{36}^3 > 0 \) and \( \kappa_{16}^4 \leq \kappa_{36}^3 \). The last assumption is fulfilled if \( \kappa_{16}^4 \leq \kappa_{36}^3 \) 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 \( \varepsilon (\nabla \varphi_{14}^1, \nabla \varphi_{25}^2)_{K} \leq 0 \), 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 \( C \notin \Omega \). Further, let us choose any constant \( \omega > 0 \) and define \( b(x) = \frac{n}{n}(C - x) \), i.e. \( b_i(x) = \frac{n}{n}(C_i - x_i) \), where \( x = [x_1, x_2, \ldots, x_n] \). Then \( \text{div} \, b = -\omega \) and the streamlines of \( b \) are rays ending at the point \( C \).

For each \( K \in \mathcal{T}_h \), let the vector \( d_{K,1} \) (i.e. the corresponding edge \( P_{K,1}P_{K,n+1} \)) lies on some streamline. We are now interested in the evaluating of \( \theta_K \), especially \( \int_{K} b \cdot \nabla \lambda_{K,j} \, dx \), for \( j = 2, 3, \ldots, n \). Let us begin with \( \int_{K} b \cdot \nabla \lambda_{K,n} \, dx \).

Since the vectors \( d_{K,j}, \, j = 1, 2, \ldots, n \), are linearly independent, the matrix \( D = [d_{K,1}, d_{K,2}, \ldots, d_{K,n}] \) is invertible and there exists a uniquely defined QR-decomposition \( D = QR \), where \( R = Q^T \) and \( Q \) is an upper triangular matrix with positive diagonal entries and \( Q \) is an orthonormal matrix (see Golub and Van Loan, 2012 Theorem 5.2.3). Using the matrix \( Q \) one may define the transformation (the translation and the rotation) of the element \( K \) (see Figure 3.13)

\[
\hat{x} = 0 + Q^T(x - P_{K,n+1}).
\]

(3.255)

Then \( R = [d_{K,1}, d_{K,2}, \ldots, d_{K,n}] \) and from the equality \( d_{K,j} \cdot \nabla \lambda_{K,n} = 0 \) for \( j = 1, 2, \ldots, n - 1 \) it follows that \( \nabla \lambda_{K,n} = (0, 0, \ldots, 0, -1/\hat{h}_n) \), where \( \hat{h}_n \) is the height of the \( n \)-simplex \( K \) in the \( \hat{x}_n \) direction.
Since the vector field $b$ 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 $\hat{b}(\hat{x}) = \frac{\omega}{n} (C - \hat{x}) = \frac{\omega}{n} (C - P_{K,n+1} - \hat{x}_1, -\hat{x}_2, \ldots, -\hat{x}_n)$. Consequently

$$\int_K b \cdot \nabla \lambda_{K,n} \, dx = \int_{\hat{K}} \hat{b} \cdot \nabla \lambda_{\hat{K},n} \, d\hat{x} = \int_{\hat{K}} \frac{\omega}{n} \hat{x}_n \, d\hat{x} = \frac{\omega}{n} \int_{\hat{K}} -\lambda_{\hat{K},n} \, d\hat{x} = -\frac{\omega |\hat{K}|}{n(n+1)}. \quad (3.256)$$

Similar approach leads to the same equalities for all $\nabla \lambda_{K,j}$, $j = 2, 3, \ldots, n-1$. Thus, for the mesh parameters $\theta_K$ in this case there holds

$$\theta_K = \frac{1}{|K|} \max \left\{ \max_{2 \leq i \leq n} \left| \int_K b \cdot \nabla \lambda_{K,i} \, dx \right|, \left| \sum_{i=2}^n \int_K b \cdot \nabla \lambda_{K,i} \, dx \right| \right\} = \frac{1}{|K|} \left| \sum_{i=2}^n -\frac{\omega |K|}{n(n+1)} \right| = \frac{n-1}{n} \frac{\omega}{n+1} < \frac{\omega}{n+1}. \quad (3.257)$$

Therefore, the mesh parameters $\theta_K$ satisfy in this case the required inequality $\theta_K \leq \frac{\omega}{n+1}$ (cf. the inequality (3.105), page 75, or the inequality (3.34), page 61). However, since $\theta_K$ 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 $b$. The first type has a form considered above and is defined as

$$b_{1A} = \frac{1}{2} (1-x, 1-y)^T, \quad (\Rightarrow \text{div} \, b_{1A} = -1). \quad (3.258)$$

The second considered type of the vector field (used for a comparison of the matrices of the mappings $R_K^{(2)}$ and $\Pi_{b,K}^{(2)}$)

$$b_{1B} = \frac{1}{\sqrt{(1-x)^2 + (1-y)^2}} (1-x, 1-y)^T \quad (3.259)$$

has the same direction and satisfies $|b_{1B}| = 1$ in $\Omega$. However, div $b_{1B}$ is no longer constant and the condition $\theta_K \leq \frac{\omega}{n+1}$ is unfulfilled (see Figure 3.17).

Further, on $\partial \Omega$ we consider the discontinuous boundary condition $u_{b1}$

$$u_{b1} = 1 \quad \text{in} \quad \{x \in \partial \Omega, |x| \leq 0.3\} \quad \text{and} \quad u_{b1} = 0 \quad \text{otherwise}. \quad (3.260)$$
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)

\[
f_1 = \begin{cases} 
1 & \text{for } -\frac{3}{7} + \frac{10}{7} x \leq y \leq \frac{3}{10} + \frac{7}{10} x \\
0 & \text{otherwise.}
\end{cases}
\] (3.261)

For both data combinations \([b_1, u_1, f_1]\) and \([b_1, u_1, f_1]\) we may compute the reduced solutions \(u_1^A\) and \(u_1^B\) (see Definition 1.3.1, page 15) of the differential equation (3.1). These reduced solutions have the form

\[
u_0^A(x, y) = \left(1 + \min\{-2 \ln(1 - x), -2 \ln(1 - y)\}\right)f_1(x, y), \quad (3.262)
\]

\[
u_0^B(x, y) = \left(1 + \min\left\{\frac{x}{1 - x}, \frac{y}{1 - y}\right\}\sqrt{(1 - x)^2 + (1 - y)^2}\right)f_1(x, y), \quad (3.263)
\]

and they are depicted in Figure 3.15.
reduced solution $u_0$. If $u_0 = xy$, then $b_{1A} \cdot \nabla u_0 = \frac{x}{2} - xy + \frac{y}{2} =: f_2$ and we may construct the (zeroth-order) asymptotic expansion $u^{(E1)}_a$ (Figure 3.16) for the problem with the data $[b_{1A}, u_{b2}, f_2] = [b_{1A}, 0, \frac{x}{2} - xy + \frac{y}{2}]$. It has the form

$$u^{(E1)}_a = xy \left( 1 - \exp \left( \frac{0.05}{\varepsilon} (x - 0.9) \right) \right) \left( 1 - \exp \left( \frac{0.05}{\varepsilon} (y - 0.9) \right) \right) \quad (3.264)$$

and we use it as a continuous test problem ($u^{(E1)}_a$ is the solution of the differential equation (3.1) with the data $[b_{1A}, u_{b2}, f_2] = [b_{1A}, 0, Lu^{(E1)}_a]$).

Figure 3.16: Graphs of the functions $u^{(E1)}_a$, $f_2$ and $Lu^{(E1)}_a$, respectively. The function $u^{(E1)}_a$ is the (zeroth-order) asymptotic expansion of the solution of the boundary value problem (3.1) with the data $[b, u_b, f] = [b_{1A}, 0, f_2]$. It is also the classical solution of the same differential equation with $[b, u_b, f] = [b_{1A}, 0, Lu^{(E1)}_a]$. 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 $\delta_K = h_K / (2\|\text{div} b\|_{\infty, K})$ 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 $\theta_K / \|\text{div} b\|_{\infty, K}$ for $b = b_{1B}$ and all $K \in T_h$.

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 $\varepsilon$ (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 $u^{(E1)}_a$ we may verify experimentally the result of Theorem 3.5.1 (page 75). Hence, we consider $b = b_{1A}$ 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 $b = b_{1A}$ 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 $\varepsilon$ (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 $b = b_{1A}$ and considered $\varepsilon = 10^{-2}$, $10^{-3}$ and $10^{-4}$ ($\varepsilon_k$ stands for $(\sum_K \frac{|d_{K,1}|}{\|b_K\|} \|b_K \cdot \nabla e_h\|_0^2)^{1/2}$).

| $\varepsilon$ | Elms | $|e_h|_{1,\Omega}$ | $\|e_h\|_{0,\Omega}$ | $\|e_b\|_0$ | $\|\varepsilon_h\|_{\infty,d}$ | $\|\varepsilon_h\|_b$ |
|---------------|------|------------------|------------------|----------|------------------|------------------|
| 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  | 1.387E-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 $\varepsilon$ 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 $b_K^{(1)}$ 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 $\mathbf{b} = \mathbf{b}_{1A}$ 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 $\varepsilon$ (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 $b = b_{1A}$ 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 $\varepsilon$ (we consider $\varepsilon = 10^{-3}$, $10^{-4}$ and $10^{-5}$).

Figure 3.21: Solution of Example 1 with $b = b_{1A}$ 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 \( b = b_{1A} \), and considered \( \varepsilon = 10^{-2}, 10^{-3}, \) and \( 10^{-4} \). Here \( \varepsilon_h = u - u_h, \) \( \xi_h = R_h u - u_h, \) \( R_h u \in P_2 \) 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.

<table>
<thead>
<tr>
<th>( \varepsilon )</th>
<th>Elms</th>
<th>|( e_h |_{1,\Omega} )</th>
<th>|( e_h |_{0,2,\Omega} )</th>
<th>|( e_h |_{\infty,d,P_2} )</th>
<th>|( e_h |_{\infty,d,P_1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1E-2</td>
<td>36</td>
<td>7.54E-02</td>
<td>5.003E-03</td>
<td>4.679E-02</td>
<td>1.591E-02</td>
</tr>
<tr>
<td>1E-2</td>
<td>144</td>
<td>3.65E-02</td>
<td>1.895E-03</td>
<td>2.643E-02</td>
<td>7.542E-03</td>
</tr>
<tr>
<td>1E-2</td>
<td>576</td>
<td>1.36E-02</td>
<td>8.839E-04</td>
<td>1.422E-02</td>
<td>2.387E-03</td>
</tr>
<tr>
<td>1E-2</td>
<td>2304</td>
<td>4.99E-03</td>
<td>4.712E-04</td>
<td>7.407E-03</td>
<td>1.098E-03</td>
</tr>
<tr>
<td>1E-2</td>
<td>9216</td>
<td>2.05E-03</td>
<td>2.488E-04</td>
<td>3.785E-03</td>
<td>5.452E-04</td>
</tr>
<tr>
<td>1E-3</td>
<td>36</td>
<td>3.58E-01</td>
<td>1.948E-02</td>
<td>1.07E-01</td>
<td>1.170E-01</td>
</tr>
<tr>
<td>1E-3</td>
<td>144</td>
<td>4.50E-01</td>
<td>1.550E-02</td>
<td>7.54E-02</td>
<td>1.950E-02</td>
</tr>
<tr>
<td>1E-3</td>
<td>576</td>
<td>3.88E-01</td>
<td>6.45E-03</td>
<td>4.68E-02</td>
<td>5.15E-02</td>
</tr>
<tr>
<td>1E-3</td>
<td>2304</td>
<td>1.81E-01</td>
<td>2.24E-03</td>
<td>2.57E-02</td>
<td>1.61E-02</td>
</tr>
<tr>
<td>1E-3</td>
<td>9216</td>
<td>6.03E-02</td>
<td>1.33E-04</td>
<td>4.64E-03</td>
<td>4.38E-03</td>
</tr>
<tr>
<td>1E-4</td>
<td>36</td>
<td>4.45E-01</td>
<td>2.11E-02</td>
<td>1.14E-01</td>
<td>7.83E-02</td>
</tr>
<tr>
<td>1E-4</td>
<td>144</td>
<td>3.43E-01</td>
<td>1.05E-02</td>
<td>7.87E-02</td>
<td>5.29E-02</td>
</tr>
<tr>
<td>1E-4</td>
<td>576</td>
<td>3.85E-01</td>
<td>6.76E-03</td>
<td>5.26E-02</td>
<td>1.06E-01</td>
</tr>
<tr>
<td>1E-4</td>
<td>2304</td>
<td>8.43E-01</td>
<td>7.27E-03</td>
<td>3.64E-02</td>
<td>1.80E-01</td>
</tr>
<tr>
<td>1E-4</td>
<td>9216</td>
<td>1.58E-00</td>
<td>5.74E-03</td>
<td>2.56E-02</td>
<td>1.76E-01</td>
</tr>
</tbody>
</table>

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 \(Y = \frac{7}{30} \sqrt{2}\). The right-hand side \(f = f_3\) satisfies \(f_3(x,y) = 1\) whenever \((\frac{7}{30})^2 \leq x^2 + y^2 \leq (\frac{11}{30})^2\) and \(f_3(x,y) = 0\) otherwise. The boundary condition \(u_{b_2}\) satisfies \(u_{b_2} = 1\) in \(\{x \in \Gamma_-, f(x) = 1\}\) and \(u_{b_2} = 0\) otherwise (see Figure 3.22). Again, we use two definitions of the vector field \(b\).

Example 2A: \(b(x,y) = (-y,x)^T\), Example 2B: \(b(x,y) = \frac{1}{\sqrt{x^2+y^2}} (-y,x)^T\),

where the second one is used for a comparison of the matrices of the mappings \(R_{b,K}^{(2)}\) 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 \(\sqrt{X^2 + Y^2} = \frac{2}{5}\) and thus, the length of the larger part of the diagonal is \(\frac{1}{2} - X \sqrt{2} = \frac{7}{5}\) (hence, the length of the shorter part is \(\frac{3}{5}\)). 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 \(\frac{1}{5}\). Since we would like to obtain an isotropic triangulation of \(\Omega\) and the height of each triangle is approximately given by the distance between two neighboring streamlines (i.e. \(\frac{1}{5}\)), 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 \( |h - h_s| \leq \frac{3\pi}{2L_s} h^2 \), 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.

For both data combinations \([b_{2A}, u_{02}, f_3]\) and \([b_{2B}, u_{02}, f_3]\) we may compute the reduced solutions \( u_{02}^{2A} \) and \( u_{02}^{2B} \) (see Definition 1.3.1, page 15) of the differential equation (3.1). These reduced solutions have the form

\[
\begin{align*}
    u_{02}^{2A}(x, y) &= \left[ 1 + \tan \left( \frac{2}{\sqrt{2 + y^2}} \right) \right] f_3(x, y), \\
    u_{02}^{2B}(x, y) &= \left[ 1 + \left( \tan \left( \frac{2}{\sqrt{2 + y^2}} \right) - \frac{1}{20} \tan \left( \frac{2}{\sqrt{2 + y^2}} \right) \right) \sqrt{x^2 + y^2} \right] f_3(x, y),
\end{align*}
\]

and they are depicted in Figure 3.23.

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 \( b_{2A} \cdot \nabla u_0 = y^2 - yX + xY - x^2 =: f_4 \) and we may construct the (zeroth-order) asymptotic expansion
for the problem with the data \([b, u_b, f] = [b_{2A}, 0, f_4]\). It has the form

\[
u_{as}^{(E2)}(Y-x)(y-X) \left(1 - \exp \left(\frac{y}{\varepsilon}(X-x)\right) \right) \left(1 - \exp \left(\frac{x}{\varepsilon}(y-Y)\right) \right)
\]

(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 \([b_{2A}, 0, Lu_{as}^{(E2)}]\).

Firstly, let us solve Example 2 using the SUPG method with the continuous piecewise linear finite elements, the stabilization parameter \(\delta_K = h_K/(2\|b\|_{\infty,K})\) and as in Example 1 we consider three types of meshes (Figure 3.25).

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 \([b, u_b, f] = [b_{2A}, 0, f_4]\). It is also the classical solution of the same differential equation with \([b, u_b, f] = [b_{2A}, 0, Lu_{as}^{(E2)}]\). In this example we consider \(\varepsilon = 2 \times 10^{-3}\).

Figure 3.25: Meshes considered in Example 2 formed by 284, 1124 and 4498 elements, respectively. The color scale indicates the value \(\theta_K/h_K\) for \(b = b_{2A}\) and all \(K \in T_h\).

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 \(\varepsilon\) (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 \(b = b_{2A}\) 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 $b = b_{2A}$ 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 $\varepsilon$ (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 $b = b_{2A}$ and considered $\varepsilon = 10^{-2}, 10^{-3}$ and $10^{-4}$. We observe, that for small $\varepsilon$ the solution converge only in $L^2$-norm ($e_b$ stands for $\sum_K \frac{|dK|}{2|h_K|b} \|b_K \cdot \nabla e_b\|^2_{h_K})^{1/2}$).

| $\varepsilon$ | Elms | $|e_h|_{1,\Omega}$ | $\|e_h\|_{0,2,\Omega}$ | $e_b$ | $\|e_h\|_{\infty,d}$ | $\|e_h\|_{h,s}$ |
|---------------|------|-------------------|------------------|-------|----------------|------------------|
| 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 | 1.941E-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 $\theta_K = O(h_K)$ (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 $b$ it may be complicated (or even impossible) to construct a mesh satisfying $\theta_K \rightarrow 0$.

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 $b = b_{2A}$, $\varepsilon = 10^{-3}$ 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 $u_{(E2)}^{(u)}$ 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 $b$ affects the structure of the mappings (or corresponding matrices) $R^{(2)}_K$ and $\Pi^{(2)}_{b,K}$ from Section 3.7. As $h \rightarrow 0$, the entries of the matrices of the mappings $R^{(2)}_K$ and $\Pi^{(2)}_{b,K}$ tend to some constant values. The matrix of the mapping $\Pi^{(2)}_{b,K}$ converges (probably under some mesh-related conditions) to the matrix of the orthogonal
Figure 3.27: Solutions of Example 2 with $b = b_{2A}$ 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 $\varepsilon$ (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 $b = b_{2A}$ 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 $\varepsilon$ (we consider $\varepsilon = 10^{-3}$, $10^{-4}$ and $10^{-5}$).

Figure 3.29: Solution of Example 2 with $b = b_{2A}$ 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 $b = b_{2A}$ 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.

| $\varepsilon$ | Elms | $\xi_h|_{1,\Omega}$ | $\|\xi_h\|_{0.2,\Omega}$ | $\|e_h\|_{0.2,\Omega}$ | $\|e_h\|_{\infty,d,P_2}$ | $\|e_h\|_{\infty,d,P_1}$ |
|-------------|-------|----------------|----------------|----------------|----------------|----------------|
| 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-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        | 498   | 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        | 498   | 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 $b = b_{2A}$, $\varepsilon = 10^{-3}$ 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 $R^{(2)}_K$ 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 $b_{1A}$ and $b_{1B}$ are depicted in the upper row, whereas in the bottom one can find the error values of entries given by the vector fields $b_{2A}$ and $b_{2B}$. Similarly, Figures 3.32–3.35 show not only a comparison of the vector fields $b_{2A}$ and $b_{2B}$, 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 $R^{(2)}_K$, $K \in T_h$. The color of each element indicates how close is this entry to its limit state. Up: Example 1, 144 elements, $b_{1B}$ left, $b_{1A}$ right; Down: Example 2, 284 elements, $b_{2B}$ left, $b_{2A}$ 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 \(\Pi_{h,K}^{(2)}, \ K \in \mathcal{T}_h.\) 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 \(\Pi_{h,K}^{(2)}, \ K \in \mathcal{T}_h.\) 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 \( b \) is in \( C^k \) of some neighborhood of \( \Omega \) and \( k \geq 1 \). Every solution to the initial value problem

\[
\zeta'(t) = b(\zeta(t)), \quad \zeta(0) = \zeta_0 \in \Omega, \tag{4.1}
\]

remains in some fixed neighborhood \( \Omega_1 \) of \( \Omega \) for only a finite time in the time interval \( (-\infty, +\infty) \) if and only if there exists a function \( \phi \in C^k(\Omega_1) \) so that \( b \cdot \nabla \phi > 0 \) in \( \Omega \).

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 \( f \in C^1(\Omega)^n \)

\[
\int_{\Omega} \text{div} \, f \, dx = \int_{\partial \Omega} f \cdot n \, ds \tag{4.2}
\]

holds. Here \( n \) is the outward pointing unit normal field of the boundary \( \partial \Omega \).


Definition 4.1.1. For \( p \in [1, \infty) \) we denote by \( L^p(\Omega) \) the Lebesgue space of all functions \( u \) measurable on \( \Omega \) such that \( \int_{\Omega} |u(x)|^p \, dx < +\infty \). The space \( L^p(\Omega) \) is equipped with the norm

\[
\|u\|_{p, \Omega} = \left( \int_{\Omega} |u(x)|^p \, dx \right)^{1/p}. \tag{4.3}
\]

Further, the space \( L^\infty(\Omega) \) consists of such measurable functions on \( \Omega \) for which the norm

\[
\|u\|_{\infty, \Omega} = \text{esssup}_{\Omega} |u| = \inf \left\{ \sup_{z \in \Omega, Z \setminus Z} |u(x)| ; Z \subset \Omega, \text{meas} (Z) = 0 \right\} \tag{4.4}
\]

is finite. The space \( L^2(\Omega) \) is a Hilbert space with the inner product

\[
(u, v)_\Omega = \int_{\Omega} u(x)v(x) \, dx \tag{4.5}
\]

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

\[
\|u\|_{k,p, \Omega} = \left( \sum_{|\alpha| \leq k} \|D^\alpha u\|_{p, \Omega}^p \right)^{1/p} \quad \text{for } p \in [1, \infty), \tag{4.7}
\]

\[
\|u\|_{k,\infty, \Omega} = \max_{|\alpha| \leq k} \{\|D^\alpha u\|_{\infty, \Omega} \} \quad \text{for } p = \infty, \tag{4.8}
\]

124
and the seminorm

\[ |u|_{k,p,\Omega} = \left( \sum_{|\alpha|=k} \|D^\alpha u\|_{0,p,\Omega}^p \right)^{1/p} \quad \text{for } p \in [1,\infty), \quad (4.9) \]

\[ |u|_{k,\infty,\Omega} = \max_{|\alpha|=k} \|D^\alpha u\|_{\infty,\Omega} \quad \text{for } p = \infty. \quad (4.10) \]

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 \( v = (v_1, v_2, \ldots, v_n) \in W^{k,p}(\Omega)^n \) we put

\[ \|v\|_{k,p,\Omega} = \left( \sum_{i=1}^n \|v_i\|_{k,p,\Omega}^2 \right)^{1/2} \quad \text{and} \quad |v|_{k,p,\Omega} = \left( \sum_{i=1}^n |v_i|_{k,p,\Omega}^2 \right)^{1/2}. \quad (4.11) \]

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 \( f : V \rightarrow \mathbb{R} \) be a continuous linear functional \( V \) and let \( a : V \times V \rightarrow \mathbb{R} \) be a bilinear form on \( V \times V \) that is coercive, i.e., there exists a constant \( \alpha > 0 \) such that

\[ a(v,v) \geq \alpha \|v\|^2 \quad \forall \ v \in V, \quad (4.12) \]

and continuous (bounded), i.e. there exists a constant \( M > 0 \) such that

\[ |a(u,v)| \leq M \|u\| \|v\| \quad \forall \ u,v \in V. \quad (4.13) \]

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

\[ a(u_0,v) = f(v) \quad \forall \ v \in V. \quad (4.14) \]

**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 \( \Gamma \), then there exist positive constants \( C_F \) and \( D_F \) depending on \( \Omega \) and \( n \) such that for all \( v \in H^1(\Omega) \) holds

\[ \|v\|_{0,\Omega} \leq C_F |v|_{1,\Omega} + D_F \|v\|_{0,\Gamma}. \quad (4.15) \]

In particular, when \( n = 1 \) then for all \( v \in H^1(I) \), \( I = (a,b) \), holds

\[ \|v\|_{0,I}^2 \leq \left( \frac{2(b-a)}{\pi} \right)^2 |v|_{1,I}^2 + \frac{2(b-a)}{\pi} \left( v^2(a) + v^2(b) \right). \quad (4.16) \]

For \( v \in H^1_0(I) \) one can derive sharper estimate

\[ \|v\|_{0,I}^2 \leq \frac{(b-a)^2}{\pi^2} |v|_{1,I}^2. \quad (4.17) \]

**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
\[
|\langle u, v \rangle| \leq \langle u, u \rangle^{1/2} \langle v, v \rangle^{1/2}. \tag{4.18}
\]
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 \(A\) satisfies \(a_{ij} \leq 0\) for \(i \neq j\). Then \(A\) is an M-matrix if and only if there exists a vector \(e > 0\) such that \(Ae > 0\).

Furthermore, we have
\[
\|A^{-1}\|_{\infty, d} \leq \frac{\|e\|_{\infty, d}}{\min_k (Ae)_k}. \tag{4.19}
\]


Theorem 4.1.6 (Comparison principle). Let \(w \in C^2(\Omega) \cap C(\bar{\Omega})\) and \(L\) is defined by (1.50) (page 14). If there holds
\[
Lw \geq 0 \text{ in } \Omega \text{ and } w \geq 0 \text{ on } \partial \Omega, \tag{4.20}
\]
then \(w \geq 0\) in \(\bar{\Omega}\).

Proof. See, e.g., (Gilbarg and Trudinger 2001, Theorem 3.3).

Definition 4.1.2 (Inverse-monotone matrix). A matrix \(A\) is called inverse-monotone if \(A^{-1}\) exists and \(A^{-1} \geq 0\).

Theorem 4.1.7 (Discrete comparison principle). Let \(A\) be an inverse-monotone matrix. Then \(Aw \leq Aw\) implies \(v \leq w\).

Proof. If \(A(w - v) \geq 0\), then using \(A^{-1} \geq 0\) implies
\[
w - v = A^{-1}[A(w - v)] \geq 0. \tag{4.21}
\]

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 \(\hat{Q}\) of \(\mathbb{R}^n\) are affine-equivalent if there exists an invertible affine mapping \(F : \mathbb{R}^n \to \mathbb{R}^n\), \(F(\hat{x}) = B\hat{x} + r\), such that \(Q = F(\hat{Q})\).

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) \(\Sigma_h = \{\phi_{j,h}, 1 \leq j \leq M\}\) and the basis functions \(w_j\) of \(X_h\) satisfying \(\phi_{j,h}(w_j) = \delta_{ij}\) for all \(1 \leq i, j \leq M\). Then with any function \(v : \hat{Q} \to \mathbb{R}\) sufficiently smooth so that the degrees of freedom \(\phi_{j,h}, 1 \leq j \leq M\), 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 $T_h$). In the finite elements framework we consider the following assumptions

(\(H1\)) We consider a regular family of triangulations $T_h$ in the following sense:

(a) The system of triangulations $\{T_h\}_{h \in (0,h_0)}$ is shape-regular, i.e. there exists a constant $\sigma > 0$ such that for all $K \in T_h$, $h \in (0,h_0)$, it holds

$$\frac{h_K}{\rho_K} \leq \sigma. \quad (4.23)$$

(b) The quantity $h = \max_{K \in T_h} h_K$ approaches zero.

(\(H2\)) The family $(K,P_K,\Sigma_K), K \in T_h, h \in (0,h_0)$, is an affine family of finite elements (used for the construction of $X_h$), i.e. all the finite elements $(K,P_K,\Sigma_K), K \in T_h, h \in (0,h_0)$, are affine-equivalent to a single element $(\tilde{K}, \tilde{P}, \tilde{\Sigma})$.

(\(H3\)) All the finite elements $(K,P_K,\Sigma_K), K \in T_h, h \in (0,h_0)$, are of class $C^0$, i.e. $X_h \subset C^0(\Omega)$ for all $h \in (0,h_0)$.

Theorem 4.2.1 ($X_h$-interpolation). In addition to the assumptions $(H1)$, $(H2)$ and $(H3)$ let there exists integers $k \geq l \geq 0$, such that the following inclusions are for each $\tilde{K} \in T_h$ satisfied

$$P_h(\tilde{K}) \subset P(\tilde{K}) \subset H^{l}(\tilde{K}), \quad (4.24)$$

$$H^{k+1}(\tilde{K}) \hookrightarrow C^{s}(\tilde{K}), \quad (4.25)$$

where $s$ is the maximal order of partial derivatives occurring in the definitions of the set $\Sigma$.

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

$$\|v - \Pi_h v\|_{m,\Omega} \leq C_X h^{k+1-m} |v|_{k+1,\Omega}, \quad \text{for } 0 \leq m \leq \min\{1,l\}, \quad (4.26)$$

$$\left(\sum_{K \in T_h} \|v - \Pi_h v\|_{m,K}^2\right)^{1/2} \leq C_X h^{k+1-m} |v|_{k+1,\Omega}, \quad \text{for } 2 \leq m \leq \min\{k+1,l\}. \quad (4.27)$$

where $\Pi_h v \in V_h$ is the $X_h$-interpolant of the function $v$.

Proof. See [Ciarlet, 1978, Theorem 3.2.1]. \(\square\)

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

$$|v_h|_{m,q,K} \leq C_{inv} \frac{h^m}{h^{m-l+n \max\{0,1/r-1/q\}}} |v_h|_{l,r,K}, \quad \text{for all } v_h \in P(K). \quad (4.28)$$

When $q = \infty$ or $r = \infty$ we set $1/\infty := 0$.

Proof. See [Ciarlet, 1978, Theorem 3.2.6]. \(\square\)
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

$$|v|_{1,K} \leq C_{n,k} \frac{||\partial K||}{|K|} \|v\|_{0,K} \quad \text{for all } v \in P_k(K), k \in \mathbb{N}. \quad (4.29)$$

Using this inequality we can easily compute the constant $C_{\text{inv}}$. For instance, in 2D there holds $\frac{||\partial K||}{|K|} = 2\rho_K$ which results in the estimates

$$4\sqrt{3} h_K \leq ||\partial K|| \leq 2\sigma h_K. \quad (4.30)$$

It means that in 2D it is necessary $\sigma \geq 2\sqrt{3}$ and for $C_{\text{inv}}$ holds $C_{\text{inv}} = 2\sigma C_{2,k}$.

Following table provides several optimal values of the constants $C_{n,k}$ (c.f. Ozisik et al. (2010)).

<table>
<thead>
<tr>
<th>$n$</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$C_{n,k}$</td>
<td>$\sqrt{6}$</td>
<td>$\sqrt{45}/2$</td>
</tr>
</tbody>
</table>

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 $K \in T_h$ 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 $\varphi$ satisfying

$$\langle \pi_{K,r} \varphi - \varphi, v \rangle_K = 0 \quad \forall v \in P_r(K). \quad (4.31)$$

The function $\pi_{K,r} \varphi$ is uniquely defined and we called the mapping $\pi_{K,r} : L^2(K) \rightarrow P_r(K)$ the orthogonal $L^2$-projection onto the space $P_r(K)$.

For each function $\psi \in L^2(K)^n$ we also define a mapping $\pi_{K,r} : L^2(K)^n \rightarrow P_r(K)^n$ by the relation

$$[\pi_{K,r} \psi]_i = \pi_{K,r} \psi_i, \quad \text{for each } i = 1, 2, \ldots, n. \quad (4.32)$$

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 $v \in W^{s,p}(K)$, $K \in T_h$, there holds

$$|\pi_{K,r} v - v|_{m,p,K} \leq C_\Pi h_K^{\mu-m} |v|_{\mu,p,K}, \quad (4.33)$$

where $p \in [1, \infty]$ and $0 \leq m \leq \mu = \min\{r+1, s\}$.


Corollary 4.2.1. If the shape-regularity assumption (4.23) is valid then for all $\psi \in W^{s,p}(K)^n$, $K \in T_h$, there holds

$$|\pi_{K,r} \psi - \psi|_{m,p,K} \leq C_\Pi h_K^{\mu-m} |\psi|_{\mu,p,K}, \quad (4.34)$$

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 $\pi_{K,r}$ it follows
\begin{equation}
|\pi_{K,r}\psi - \psi|_{m,p,K} = \left( \sum_{i=1}^{n} |\pi_{K,r}\psi_i - \psi_i|^2 |_{m,p,K} \right)^{1/2} \leq \pi_{m,p,K} (4.35)
\end{equation}
\begin{equation}
\leq CH_{K,r}^{-m} \left( \sum_{i=1}^{n} |\psi_i|^2 |_{m,p,K} \right)^{1/2} = CH_{K,r}^{-m} |\psi|_{m,p,K}. \tag{4.36}
\end{equation}

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
\begin{equation}
\frac{|K|}{(n + 1)(n + 2)} v_h^2(P_{K,i}) \leq \|v_h\|_{0,K}^2 \leq \frac{|K|}{n + 1} \sum_{i=1}^{n+1} v_h^2(P_{K,i}). \tag{4.37}
\end{equation}

Proof. Let us denote $v = (v_h(P_{K,1}), v_h(P_{K,2}), \ldots, v_h(P_{K,n+1}))^T$ and let $A$ 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{equation}
\|v_h\|_{0,K}^2 = \int_K \left( \sum_{i=1}^{n+1} v_h(P_{K,i}) \lambda_{K,i} \right)^2 \, dx = \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)} v^T A v. \tag{4.38}
\end{equation}

Thus, it remains to determine the eigenvalues of $A$. Since the characteristic polynomial of the matrix $A$ is $\det(A - \lambda I) = (n + 2 - \lambda)(1 - \lambda)^n$, we get $|v|^2 \leq v^T A v \leq (n + 2)|v|^2$ 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
\begin{equation}
\left( \sum_{i=1}^{n} a_i \right)^2 \leq n \sum_{i=1}^{n} a_i^2. \tag{4.39}
\end{equation}

Proof. From the Cauchy-Schwarz-Bunyakovsky inequality (Theorem 4.1.4 page 126) it follows that
\begin{equation}
\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{4.40}
\end{equation}

Corollary 4.2.2. For any $n \in \mathbb{N}$ and $s_{K}^{(i)} \geq 0, i = 1, 2, \ldots, n, K \in T_h$, it holds
\begin{equation}
\sum_{i=1}^{n} \left( \sum_{K \in T_h} s_{K}^{(i)} \right)^{1/2} \leq \left( \sum_{i=1}^{n} \sum_{K \in T_h} s_{K}^{(i)} \right)^{1/2} = \left( \sum_{K \in T_h} \sum_{i=1}^{n} s_{K}^{(i)} \right)^{1/2}. \tag{4.41}
\end{equation}

Proof. It suffices to take $a_i = \left( \sum_{K \in T_h} s_{K}^{(i)} \right)^{1/2}$ 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 \( \pi/m \), \( m \in \mathbb{N}, m \geq 2 \). 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 \( b \) is non-positive, we can discretize \( b \) 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 \( L^\infty \)-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 \( \varepsilon \)) in 2D, or even 3D.
Bibliography


<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>A part of the general convex domain $\Omega$</td>
<td>16</td>
</tr>
<tr>
<td>1.2</td>
<td>The zeroth-order matched asymptotic expansion in the parabolic boundary layer and its contours, $\varepsilon = 0.01$, $f = 1$, $</td>
<td>b</td>
</tr>
<tr>
<td>1.3</td>
<td>A simple triangular domain considered in the numerical experiment.</td>
<td>31</td>
</tr>
<tr>
<td>1.4</td>
<td>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 $\varepsilon = 0.01$.</td>
<td>32</td>
</tr>
<tr>
<td>2.1</td>
<td>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. $Pe = 5$).</td>
<td>36</td>
</tr>
<tr>
<td>2.2</td>
<td>SUPG (left) and continuous Petrov-Galerkin (right) test functions for different choices of $\tau$ and $\sigma$.</td>
<td>38</td>
</tr>
<tr>
<td>2.3</td>
<td>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 $h = 0.02$.</td>
<td>39</td>
</tr>
<tr>
<td>2.4</td>
<td>A comparison of the discrete solutions obtained by adding artificial diffusion. Each dashed curve corresponds to a different partition of $\Omega$ 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.</td>
<td>40</td>
</tr>
<tr>
<td>2.5</td>
<td>Local Green’s functions for $b = 1$ and different choices of $Pe$.</td>
<td>43</td>
</tr>
<tr>
<td>2.6</td>
<td>Comparison of convergence of the simple upwind scheme and the Il’in-Allen-Southwell scheme.</td>
<td>45</td>
</tr>
<tr>
<td>3.1</td>
<td>Definition of the splitting of the domain $\Omega_j$.</td>
<td>60</td>
</tr>
<tr>
<td>3.2</td>
<td>Partition of the unit cube into simplices $K_\pi$, $\pi \in S_3$, in 3D. Images $a) - f$ correspond to the permutations $(123)$, $(123)$, $(123)$, $(123)$, $(123)$, $(123)$, $(123)$, $(123)$, $(123)$, respectively.</td>
<td>62</td>
</tr>
<tr>
<td>3.3</td>
<td>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.</td>
<td>63</td>
</tr>
<tr>
<td>3.4</td>
<td>The expected $EOC$ progression of the original SUPG method and the new method for $\text{div } b &lt; 0$ and $\text{div } b = 0$. Continuous piecewise linear finite elements are used and the error is measured in the norms $|\cdot|<em>{S_D}$, $|\cdot|</em>{S_D}$ and $|\cdot|_{h^x}$, 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 $\varepsilon$.</td>
<td>82</td>
</tr>
<tr>
<td>3.5</td>
<td>Three-directional mesh in the exponential boundary layer.</td>
<td>86</td>
</tr>
</tbody>
</table>
3.6 Using the Péclet number $\text{Pe}_T$ one can compute the parameter $\mathcal{R}_T$ (left) which restricts the maximum angle in triangles. For instance, when $\text{Pe}_T \approx 0.6282$, then $\mathcal{R}_T \approx -1/2$ and the angle $\alpha$ (and $\beta$) has to satisfy the inequality $-\frac{1}{2}\cot \alpha = A \leq -\frac{1}{2}C = \frac{1}{2}\cot \gamma$. Consequently, if (for instance) $\gamma = 60^\circ$ then $\alpha, \beta \leq \text{acot}(-1/(2\sqrt{3}) \approx 106.1^\circ$ (right), which is a sharper bound as compared to standard $\alpha, \beta < 180^\circ - 60^\circ = 120^\circ$. The right figure shows the restriction curves for several choices of $\mathcal{R}_T$ (or $\text{Pe}_T$).

3.7 For constant data, the three-directional mesh and two considered choices of $\mu, \nu$ the estimate (3.183) holds with $\alpha^{(1)}_{\text{max}}$ and $\alpha^{(2)}_{\text{max}}$. For arbitrary $\text{Pe}_T$ there holds $\alpha^{(1)}_{\text{max}} \geq 0.3839$ and $\alpha^{(2)}_{\text{max}} \geq 0.3675$. We also observe that the choices $\alpha = \alpha^{(1)}_0$ and $\alpha = \alpha^{(2)}_0$ are suboptimal (in comparison with $\alpha^{(1)}_{\text{max}}$ and $\alpha^{(2)}_{\text{max}}$). The middle picture shows the detail whereas in the right picture one can see the values of all functions for large $\text{Pe}_T$.

3.8 For constant data and the three-directional mesh both considered choices of the coefficients $\mu$ and $\nu$ satisfy the conditions (3.172) and (3.186).

3.9 Parts of $\mathcal{T}_h$ used for the definition of $\mathbf{n}_K, \mathbf{v}_f, \mathbf{v}_V$ and $\mathbf{v}_T$.

3.10 A structure of the domain $\Omega_S$ (left) and a choice of the vectors $\mathbf{n}_Q$, $Q \subset \Omega_S$ (right).

3.11 A structure of the domain $\Omega_S$ in the vicinity of a corner of $\Omega$.

3.12 Definition of nodes numbering for $P_2(K)$.

3.13 Rotated element $K$ in $2D$.

3.14 Definition of the Example 1 data.

3.15 Reduced solutions $u_0^{LA}$ (up) and $u_0^{LB}$ (down).

3.16 Graphs of the functions $u_0^{LA}, f_2$ and $Lu_0^{LA}$, respectively. The function $u_0^{LA}$ is the (zeroth-order) asymptotic expansion of the solution of the boundary value problem (3.1) with the data $[\mathbf{b}, u_0, f] = [\mathbf{b}_{LA}, 0, f_2]$. It is also the classical solution of the same differential equation with $[\mathbf{b}, u_0, f] = [\mathbf{b}_{LA}, 0, Lu_0^{LA}]$. In this example we considered $\varepsilon = 10^{-3}$. For $\mathbf{b} = \mathbf{b}_{LB}$ and all $K \in \mathcal{T}_h$.

3.17 Meshes considered in Example 1 formed by 144, 576 and 2304 elements, respectively. The color scale indicates the value $\theta_K/\|\text{div } \mathbf{b}\|_{\infty,K}$ for $\mathbf{b} = \mathbf{b}_{LA}$ and all $K \in \mathcal{T}_h$.

3.18 Solutions of Example 1 with $\mathbf{b} = \mathbf{b}_{LA}$ 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 $\varepsilon$ (we consider $\varepsilon = 10^{-3}, 10^{-4}$ and $10^{-5}$). The bottom right solution is displayed enlarged.

3.19 Solutions of Example 1 with $\mathbf{b} = \mathbf{b}_{LA}$ 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 $\varepsilon$ (we consider $\varepsilon = 10^{-3}, 10^{-4}$ and $10^{-5}$). The bottom right solution is displayed enlarged.
3.20 Solutions of Example 1 with \( b = b_{1A} \) 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 \( \varepsilon \) (we consider \( \varepsilon = 10^{-3}, 10^{-4} \) and \( 10^{-5} \)).

3.21 Solution of Example 1 with \( b = b_{1A} \) 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.

3.22 Definition of the Example 2 data.

3.23 Reduced solutions \( u_{0A}^{(\varepsilon)} \) (up) and \( u_{0B}^{(\varepsilon)} \) (down).

3.24 Graphs of the functions \( u_{2A}^{(\varepsilon)} \), \( f_4 \) and \( Lu_{2A}^{(\varepsilon)} \), respectively. The function \( u_{2A}^{(\varepsilon)} \) is the (zeroth-order) asymptotic expansion of the solution of the boundary value problem (3.1) with the data \( [b, u, f] = [b_{2A}, 0, f_4] \). It is also the classical solution of the same differential equation with \( [b, u, f] = [b_{2A}, 0, Lu_{2A}^{(\varepsilon)}] \). In this example we consider \( \varepsilon = 2 \times 10^{-3} \).

3.25 Meshes considered in Example 2 formed by 284, 1124 and 4498 elements, respectively. The color scale indicates the value \( \theta_K/h_K \) for \( b = b_{2A} \) and all \( K \in T_h \).

3.26 Solutions of Example 2 with \( b = b_{2A} \) 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 \( \varepsilon \) (we consider \( \varepsilon = 10^{-3}, 10^{-4} \) and \( 10^{-5} \)). The bottom right solution is displayed enlarged.

3.27 Solutions of Example 2 with \( b = b_{2A} \) 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 \( \varepsilon \) (we consider \( \varepsilon = 10^{-3}, 10^{-4} \) and \( 10^{-5} \)). The bottom right solution is displayed enlarged.

3.28 Solutions of Example 2 with \( b = b_{2A} \) 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 \( \varepsilon \) (we consider \( \varepsilon = 10^{-3}, 10^{-4} \) and \( 10^{-5} \)).

3.29 Solution of Example 2 with \( b = b_{2A} \) 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.

3.30 Solution of Example 2 with \( b = b_{2A}, \varepsilon = 10^{-3} \) 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.
3.31 In any quarter, each square \((i, j), i, j = 1, 2, 3\), corresponds to one entry \(r^K_{ij}\) of the matrix of the mappings \(R^{(2)}_K\), \(K \in T_h\). The color of each element indicates how close is this entry to its limit state. Up: Example 1, 144 elements, \(b_{1B}\) left, \(b_{1A}\) right; Down: Example 2, 284 elements, \(b_{2B}\) left, \(b_{2A}\) right.

3.32 Each square \((m, j), m = 1, 2, 3, j = 1, 2, \ldots, 6\), corresponds to one entry \(\mu^{(j)}_{K,m}\) of the matrix of the mappings \(\Pi^{(2)}_{b,K}\), \(K \in T_h\). The color of each element indicates how close is this entry to its limit state. (Example 2B, 72 elements)

3.33 Each square \((m, j), m = 1, 2, 3, j = 1, 2, \ldots, 6\), corresponds to one entry \(\mu^{(j)}_{K,m}\) of the matrix of the mappings \(\Pi^{(2)}_{b,K}\), \(K \in T_h\). The color of each element indicates how close is this entry to its limit state. (Example 2B, 284 elements)

3.34 Each square \((m, j), m = 1, 2, 3, j = 1, 2, \ldots, 6\), corresponds to one entry \(\mu^{(j)}_{K,m}\) of the matrix of the mappings \(\Pi^{(2)}_{b,K}\), \(K \in T_h\). The color of each element indicates how close is this entry to its limit state. (Example 2A, 72 elements)

3.35 Each square \((m, j), m = 1, 2, 3, j = 1, 2, \ldots, 6\), corresponds to one entry \(\mu^{(j)}_{K,m}\) of the matrix of the mappings \(\Pi^{(2)}_{b,K}\), \(K \in T_h\). The color of each element indicates how close is this entry to its limit state. (Example 2A, 284 elements)
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Computational errors in $L^\infty(\Omega)$-norm and experimental orders of</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>convergence for different values of $\gamma$ and $\varepsilon$ (adopted from</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lamacˇ (2013)).</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Different choices of the coefficients $\mu$ and $\nu$ together with the inner</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>angles restriction lead for the constant data to methods with the following</td>
<td></td>
</tr>
<tr>
<td></td>
<td>properties: the fulfilment of the discrete maximum principle (DMP), the uniform</td>
<td></td>
</tr>
<tr>
<td></td>
<td>convergence in $|\cdot|_{d,\infty}$ norm with respect to $\varepsilon$ in</td>
<td></td>
</tr>
<tr>
<td></td>
<td>the vicinity of the boundary $\Gamma$ (UNI) and admissibility of obtuse inner</td>
<td></td>
</tr>
<tr>
<td></td>
<td>angles (OBT).</td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td>Computational errors in several types of norms. We applied the new method</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>to Example 1 using piecewise linear finite elements with $b = b_{1A}$ and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>considered $\varepsilon = 10^{-2}, 10^{-3}$ and $10^{-4}$ ($e_b$ stands for</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$(\sum_K</td>
<td>d_{K,1}</td>
</tr>
<tr>
<td>3.3</td>
<td>Computational errors in several types of norms. We applied the new method</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td>to Example 1 using piecewise quadratic finite elements with $b = b_{1A}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and considered $\varepsilon = 10^{-2}, 10^{-3}$ and $10^{-4}$. Here $e_h =$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$u - u_h$, $\xi_h = R_h u - u_h$, $R_h u \in P_2$ is the Lagrange interpolation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>of $u$, $| \cdot |_{\infty,d,P_2}$ is the discrete maximum norm over all</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$P_2$-nodes, whereas $| \cdot |_{\infty,d,P_1}$ is the discrete maximum norm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>over all $P_1$-nodes.</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>Computational errors in several types of norms. We applied the new method</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>to Example 2 using piecewise linear finite elements with $b = b_{2A}$ and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>considered $\varepsilon = 10^{-2}, 10^{-3}$ and $10^{-4}$. We observe, that</td>
<td></td>
</tr>
<tr>
<td></td>
<td>for small $\varepsilon$ the solution converge only in $L^2$-norm ($e_b$ stands</td>
<td></td>
</tr>
<tr>
<td></td>
<td>for $(\sum_K</td>
<td>d_{K,1}</td>
</tr>
<tr>
<td>3.5</td>
<td>Computational errors in several types of norms. We applied the new method</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>to Example 2 using piecewise quadratic finite elements with $b = b_{2A}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and considered $\varepsilon = 10^{-2}, 10^{-3}$ and $10^{-4}$. Here $e_h =$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$u - u_h$, $\xi_h = R_h u - u_h$, $R_h u \in P_2$ is the Lagrange interpolation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>of $u$, $| \cdot |_{\infty,d,P_2}$ is the discrete maximum norm over all</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$P_2$-nodes, whereas $| \cdot |_{\infty,d,P_1}$ is the discrete maximum norm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>over all $P_1$-nodes.</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Several optimal values of the constants $C_{n,k}$ for $n = 2$ and $3$.</td>
<td>128</td>
</tr>
</tbody>
</table>