University Sorbonne Paris Nord
Institut Galilée

in partnership with **Université Paris-Saclay, CEA, Service de Thermo-hydraulique et de Mécanique des Fluides**

Doctoral Thesis

Submitted for the degree of **Doctor of University Sorbonne Paris Nord**
Specialty: **Numerical Analysis** (Field: **Applied Mathematics**)

by

Mayssa MROUEH

STUDY OF THE DISCONTINUOUS GALERKIN FINITE ELEMENT METHOD IN FLUID MECHANICS

Research carried out at the **Université de Paris-Saclay, CEA, Service de Thermo-hydraulique et de Mécanique des Fluides** and at the **Institut Galilée-LAGA**.

Defended on 30 janvier 2026 before the jury composed of:

Reviewer:	Alexandre Ern (Professor, École des Ponts ParisTech)
Reviewer:	Stella Krell (Associate Professor, Université de Nice – Sophia Antipolis)
Examiner:	Fayssal Benkhaldoun (Professor, Sorbonne Paris Nord University)
Examiner:	Patrick Ciarlet (Professor, ENSTA)
Examiner:	Alexei Lozinski (Professor, University of Franche-Comté)
Co-supervisor:	Erell Jamelot (Research Engineer, CEA)
Thesis Supervisor:	Pascal Omnes (Research Director, CEA)

Acknowledgments

Une thèse n'est pas seulement un ensemble de résultats scientifiques ; c'est une expérience qui permet de découvrir un autre point de vue sur la science et sur la vie en général. Pour accomplir cette mission, de nombreuses personnes m'ont aidé, et il est important de les remercier.

Tout d'abord, je tiens à remercier mon directeur de thèse Pascal Omnes et ma co-directrice de these Erell Jamelot. Pascal, j'ai eu beaucoup de chance de travailler avec toi. Ton calme remarquable, qui m'a mise très à l'aise lors de nos discussions sur les problèmes que j'ai rencontrés. J'ai beaucoup appris de ta grande maîtrise des mathématiques et de ta façon d'examiner un même problème sous plusieurs points de vue. L'un des enseignements les plus précieux que j'ai reçus de toi est que, pour chaque problème, il faut d'abord comprendre tout ce qui le précède afin d'en saisir l'origine réelle et de pouvoir le résoudre, tout en gardant toujours son calme.

Erell, j'ai moi aussi eu la grande chance de travailler avec toi et d'apprendre de tes connaissances, de tes points de vue et de tes stratégies. Tu es une femme merveilleuse et intelligente, et j'ai compris grâce à toi qu'il faut toujours persévérer jusqu'au bout, ne jamais abandonner, et qu'il existe toujours de nouvelles idées ou de nouvelles valorisations que l'on peut ajouter en mathématiques. Je te remercie pour ton soutien tout au long de cette thèse : tu as toujours été à mes côtés pour m'aider à surmonter les problèmes.

Je tiens également à remercier mon jury de thèse. Tout d'abord, mes rapporteurs, Alexandre Ern et Stella Krell, pour le temps qu'ils ont consacré à la lecture de mon manuscrit, leurs précieux commentaires et leurs conseils. Je remercie chaleureusement aussi les membres du jury, Fayssal Benkhaldoun, Patrick Ciarlet, et Alexei Lozinski qui ont accepté de participer à mon jury.

Je suis reconnaissante au CEA, en particulier au projet SITHY, pour le financement de cette thèse. Je remercie Nicolas Dorville, notre chef de service.

Je n'oublie pas de remercier celles et ceux qui m'ont donné l'opportunité de venir en France et de financer mon année d'études en Master 2. Je remercie en particulier la

Fondation Mathématique Jacques Hadamard pour m'avoir accordé cette bourse.
Je commence par mes collègues Andrew, Paulin, Blessing, Valentin, Andjib, Grégoire et Antonin, ainsi que tous les stagiaires, les autres doctorants et les permanents. Je souhaite également remercier mes amies que le CEA m'a permis de rencontrer : Nathalie, Jana et Diana. Nathalie m'a soutenue au CEA depuis le début de mon stage jusqu'à la soutenance. Merci d'avoir toujours été à mes côtés et de m'avoir fait rire, même dans les moments les plus difficiles. Merci pour toutes les soirées que tu as organisées, qui ont créé un espace de plaisir pour tous.
Jana, ma colocataire de bureau, avec qui j'ai partagé le quotidien de la thèse, ses hauts comme ses bas. Merci de toujours comprendre mon état d'esprit, que j'aille bien ou non, et d'être restée près de moi en toutes circonstances. Merci aussi pour ton sourire, avec lequel tu m'accueillais chaque matin. Diana, je peux dire que tu as été ma partenaire de tous les déjeuners que nous avons pris ensemble. C'était un grand plaisir de partager avec toi aussi bien mon quotidien que les moments difficiles de ma thèse, mais aussi les moments heureux de la thèse et de ma vie.
Sarah, une amie que j'ai rencontrée en Master et qui fera toujours partie de ma vie. Une amie avec qui j'ai été côte à côte en Master 1, dans les démarches pour venir en France et tout ce qui a suivi. Merci Sarah d'être une personne qui ne m'a jamais refusé une demande. Merci pour les appels vidéo qui m'ont toujours fait plaisir et merci d'être toujours à mes côtés.
Je commence par les amis que j'ai rencontrés au tout début de mon arrivée en France, et avec qui j'ai partagé mes premiers moments ici: Maria, Hadi et Nagham. Il est certain que le début aurait été beaucoup plus difficile sans votre présence et votre soutien.
Maria, je n'oublierai jamais les jours et les nuits que nous avons passés ensemble depuis le début, ni les petits-déjeuners que tu me préparais avant les examens, en particulier les manakich au labné, ni ta phrase célèbre : « Je suis fière de toi ». Je n'oublierai pas non plus ton accueil à mon retour du travail, ni tes histoires et tes aventures sans fin. Merci pour chaque moment où tu m'as soutenue et écoutée. Merci de m'avoir traitée comme une sœur et non comme une simple amie : tu as fait une immense différence dans mon parcours.
Hadi, merci pour tout ton soutien, surtout au début du Master 2 qui n'a pas du tout été facile. Merci pour tout ce que tu as fait pour moi durant cette année, et en particulier pour les délicieux plats que tu nous rapportais du Liban.
Nagham, je pense qu'il est difficile de te décrire ou même de te remercier. Nous avons partagé ensemble des moments de rires et de larmes. Tu as été, et tu restes, une amie qui me connaît profondément, comme une sœur en exil depuis le début du chemin. Un soutien et un appui à toutes les étapes. Tu as été une part essentielle de mon parcours

et tu le resteras. Je n'oublie pas les délicieux repas que tu nous préparais après le travail ou l'université, les soirées, les promenades, les sorties, et surtout ton immense soutien dans les moments difficiles, ou même simplement lorsque je ne me sentais pas bien. Je n'oublie pas non plus les voyages que nous avons faits ensemble, et tant, tant d'autres souvenirs précieux.

Katia, mon amie d'enfance et bien plus que ça. Une personne qui m'a accompagnée depuis l'enfance jusqu'à aujourd'hui, et je sais que tu seras à mes côtés pour toute la vie. Merci de m'avoir rappelé, lorsque j'étais perdue, que les mathématiques étaient le bon choix pour moi. Merci pour tes conseils et les discussions que nous avons eues, qui ont changé mon point de vue sur la vie. Merci pour tous les moments que nous avons passés ensemble, qui me ramènent à mon enfance. Tu es bien plus que ça.

Ali, mon fiancé, je sais que tu sais ce que tu représentes pour moi. J'ai eu la chance immense de t'avoir à mes côtés. Tu savais si j'étais heureuse, triste ou stressée sans que je dise un mot. Merci de comprendre toutes mes humeurs durant ce parcours, merci pour ton soutien inconditionnel, pour tes positions fermes qui m'ont guidée, pour ton aide afin d'améliorer chaque chose vers le meilleur. Mon ami, mon cher, et bien plus encore : merci pour ta capacité à me rendre heureuse dans les moments les plus difficiles, toi qui es mon repère en toute chose. Merci pour tous tes efforts pour me faire rire et me rendre heureuse. Un immense merci également pour les discussions que nous avons eues en mathématiques, pour tes remarques sur le manuscrit et la présentation, ainsi que pour les répétitions afin de bien me préparer pour le jour de la soutenance. Jamais je n'oublierai tes danses et tous tes efforts pour que je garde toujours mon énergie. Tu es un soutien pour toute ma vie.

Manal, ma grande sœur, j'ai toujours ressenti ton amour pour moi comme celui d'une mère plus que d'une sœur. Tes prières m'accompagnent, ton amour me soutient, et ta tendresse restera un appui pour moi toute ma vie. Tu as fait entrer la joie dans mon cœur, et tes enfants, Lana, Ali et Joud, sont ce que tu as de plus précieux.

Zainab, le refuge de mes secrets et mon carnet où je confie tout ce que j'ai, à chaque difficulté ou moment de perdition. Mon amie proche et ma petite sœur — même si tu es plus âgée que moi — je te vois, toi et Fatima, comme mes belles filles. Je résumerai tout par une seule phrase : merci d'exister dans ma vie.

Fatima, si seulement tout le monde avait un cœur comme le tien. Tu m'as soutenue depuis le début de mon parcours, pas seulement en France mais depuis le début de l'université. Tu m'as soutenue de toutes tes forces, comme une montagne sur laquelle je m'appuie. Tu me donnais une énergie incroyable pour continuer ce chemin. Je t'aime énormément, l'une des plus belles bénédictions. La lumière de ma vie, mon enfant chérie et gâtée, ma deuxième mère, la source de mon bonheur et de ma force. Je t'aime énormément.

Ma mère et mon père, il est certain que je ne pourrai jamais vous rendre justice, même si je consacrais toute ma vie pour vous. Mon père merci pour tout, mais je résumerai par deux phrases que vous me disiez toujours. La première : « Si tu veux ou si tu te fatigues, tu peux revenir vers nous ; nous t'aimons, nous sommes fiers de toi et nous n'avons pas besoin d'un doctorat. » Et la seconde : « Étudie bien, mais n'oublie pas toi-même, et je suis très fier de toi. » Merci papa.
Maman, je ne te remercie pas seulement pour un parcours commencé il y a quelques années, mais pour toute une vie que tu as consacrée à moi et à mes frères et sœurs, avec détermination, persévérance et réussite que tu nous as offertes. Sans toi, ni moi ni mes frères et sœurs ne serions ce que nous sommes aujourd'hui. Te parler chaque jour est une source de joie ; ton sourire est ce que j'ai de plus précieux au monde. Sans ton sacrifice, ta fatigue, ton soutien et tes prières, je n'aurais même pas pu commencer ce parcours. Merci depuis toujours. Tu m'as soutenue dès mes premiers pas, tu as veillé sur moi dans les moments difficiles et, grâce à tes sacrifices, je suis arrivée là où je suis aujourd'hui. Un simple "merci" ne suffit pas. Je t'aime énormément.

Contents

List of Figures

List of Tables

Publications, communications and summer school

Publications

During these three-year thesis project, theoretical and numerical results were produced. These findings led to present my work during conferences, to write a proceedings, and an article, which is under progress. I also attended in a summer school.

Article soon to be finalised

2025 *A Priori Error Estimates for Discontinous Galerkin Finite Element Discretization of Stokes Problems with Low Regularity*, E. Jamelot, M. Mroueh, P. Omnes.

2025 Two other articles are currently in preparation: one on the $\mathbf{P}_{dg}^{k} - P_{dg}^{k+1}$ scheme, and another on the non-polynomial basis.

Peer-reviewed conference proceedings

2024 *Discontinuous Galerkin schemes for the Stokes problem*, M. Mroueh, E. Jamelot, P. Omnes, SNA + MC 2024, Oct. 20-24, Paris.

Oral communications without proceedings

2025 **38th Chemnitz Finite Element Symposium**, On Tour in Mülheim an der Ruhr, Germany.

2025 **12^{e} Biennale Française des Mathématiques Appliquées et Industrielles – SMAI**, Carcans Maubuisson, France.

2024 **Présentation de ma thèse en 180 secondes**, Université Sorbonne Paris Nord, Paris, France.

2024 Poster at the **36^{e} Séminaire sur La Mécanique des Fluides Numérique**, CEA GAMNI, Paris, France, 2024.

2024 **Journée scientifique de l'ED**, Université Sorbonne Paris Nord, Paris, France.

2023 **Journées scientifiques du GdR MaNu**, Le Croisic, France.

Summer school

2023 **CEA-EDF-INRIA Numerical Analysis Summer School**, "Discrétisations polyédriques robustes pour la mécanique numérique", EDF Lab Paris-Saclay, Palaiseau, France.

Abstract

Fluid mechanics plays a crucial role in nuclear reactors, as it enables the understanding and control of the coolant flow around the reactor core, which is essential for safe and efficient operation. This flow is modeled by the Navier–Stokes equations coupled with temperature equations. Physically, it occurs in complex geometries, which leads, from a mathematical perspective, to irregular domains and challenging conditions that result in a low-regularity right-hand side in the corresponding model.

In this context, the present work focuses on the study of the Stokes problem in a low-regularity setting, either due to a non-smooth right-hand side or a non-smooth domain. The analysis is carried out both theoretically and numerically using the discontinuous Galerkin finite element method, together with an *a priori* error estimate in the case where the solution $(\mathbf{u}, p) \in \mathbf{H}^{1+s}(\Omega) \times H^s(\Omega)$ with $s > \frac{1}{2}$. We then propose a new scheme in which the polynomial order of the pressure is higher than that of the velocity, which proves particularly effective when the viscosity is very small.

Finally, inspired by the advection–diffusion problem—which can be viewed as a fundamental building block of the Navier–Stokes equations— we introduce a non-polynomial basis involving exponential terms, derived from the exact solution when the right-hand side is zero or constant and when the advection field is piecewise constant. We define this basis in one, two, and three dimensions, and then adapt it to the Oseen and Navier–Stokes problems. A more in-depth study will allow further improvements for these two models.

Key words: Fluid mechanics,Navier–Stokes, Stokes, Oseen, Advection–Diffusion, Discontinuous Galerkin Finite Element Method, Low regularity, Non-polynomial basis, A priori error estimate.

Résumé

La mécanique des fluides est essentielle dans un réacteur nucléaire, car elle permet de comprendre et de contrôler l'écoulement du fluide qui refroidit le cœur, ce qui est indispensable pour un fonctionnement sûr et efficace. Cet écoulement est modélisé par les équations de Navier–Stokes couplées à des équations de température. Physiquement, il se produit dans des géométries complexes, ce qui se traduit mathématiquement par des domaines non réguliers, ainsi que par des conditions difficiles qui donnent lieu à un second membre peu régulier dans le problème modélisé.
Dans ce contexte, ce travail se concentre sur l'étude du problème de Stokes dans un cadre faiblement régulier, soit à cause d'un second membre non régulier, soit à cause d'un domaine non régulier. L'analyse est menée théoriquement et numériquement en utilisant la méthode des éléments finis de Galerkin discontinus, avec une estimation d'erreur a priori dans le cas où la solution $(\mathbf{u}, p) \in \mathbf{H}^{1+s}(\Omega) \times H^s(\Omega)$ avec $s > \frac{1}{2}$. Ensuite, nous proposons un nouveau schéma dans lequel l'ordre polynomial de la pression est plus élevé que celui de la vitesse, ce qui s'avère particulièrement efficace lorsque la viscosité est très faible.
Enfin, en nous inspirant du problème d'advection–diffusion, que l'on peut considérer comme une brique élémentaire du problème de Navier–Stokes, une base non polynomiale comportant des termes exponentiels, dérivée de la solution exacte lorsque le second membre est nul ou constant et que le champ d'advection est constant par morceaux. Nous définissons cette base en dimensions 1, 2 et 3, puis nous l'adaptons aux problèmes d'Oseen et de Navier–Stokes. Une étude plus approfondie permettra d'améliorer encore les résultats pour ces deux derniers modèles.
Mots-clés : Mécanique des fluides, Navier–Stokes, Stokes, Oseen, Advection–Diffusion, Méthode des éléments finis de Galerkin discontinus, Faible régularité, Base non polynomiale, Estimation d'erreur a priori.

Chapter 1

Introduction

Contents

1.1 Context

The unsteady incompressible Navier-Stokes equations [57, 15] allow to model flows in a nuclear core reactor. In this thesis, we study the space discretization of these partial differential equations with Discontinuous Galerkin Finite Elements. We focus on the Symmetric Interior Penalty method [58, 40]. These equations are unsteady, constrained and non linear, which causes theoretical and practical difficulties. Thus, a first step to discretize the unsteady incompressible Navier-Stokes is to study the Stokes problem. Another important step is to consider the Oseen problem.
In the context of fluid mechanics at CEA, several challenges arise, notably the irregularity of the domain due to the presence of obstacles, reentrant corners, and similar features. It is therefore important to study the issues caused by domain irregularities or by the source term. In particular, the solution to Navier-Stokes equations (or Stokes and Oseen problems) may be weakly regular.
Studying the numerical discretization of the Stokes problem helps to build an appropriate approximation of the Navier-Stokes equations. The first part of this work consists in carrying out the numerical analysis of the discretization of the Stokes equations with

Discontinuous Galerkin Finite Elements, using polynomial basis functions and focusing on weakly regular solutions. We consider local polynomial basis.
Another particularly interesting and physically significant issue in numerical simulations is the boundary layer. For this reason, in the second part, we propose a basis adapted to the rapid variations of the numerical solution. We study the use of exponential basis functions, which allow to discretize advection-diffusion problem with an accurate precision [101].
More precisely, in Part I, we present a study of the Stokes problem in a low-regularity setting. We explicitly specify the dependence of the constants in the inequalities on the mesh regularity and also optimize some of these constants. Then, we propose a new numerical scheme designed to reduce computational time and improve the numerical pressure approximation when the viscosity is very small.
In Part II, we introduce a new discontinuous Galerkin finite element method with a non-polynomial basis containing exponential terms, aimed at minimizing numerical oscillations and accurately capturing rapid variations in the solution without requiring strong mesh refinement.

1.2 Overview of the CEA, connection to fluid mechanics

The French Alternative Energies and Atomic Energy Commission (CEA) [88] is a major research institution with ten centers across France. Its historical site in Saclay brings together over 7,000 researchers and focuses on strategic areas such as low-carbon energy, defense, information technologies, and health.
One of the main objectives of the CEA is to study nuclear energy nuclear reactors. A nuclear reactor allows to produce electricity thanks to the energy released by the chain reaction of nuclear fission.
Regardless of the reactor type, several essential components can be identified:

— **The reactor core** contains the nuclear fuel, composed of fissile materials (such as Uranium-235) and fertile materials (such as Uranium-238), which can partially convert into fissile isotopes through neutron absorption. The fuel may come in the form of pellets, particles, or plates, grouped into rods or assemblies.

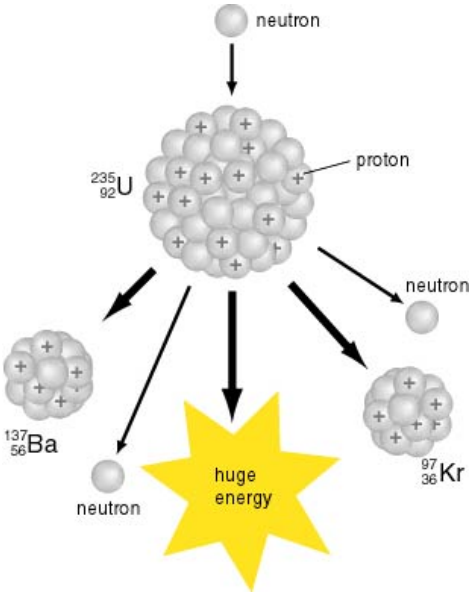


Figure 1.1 – Schematic Drawing of the Fission Process of Uranium-235 [59].

— **The moderator** slows down the fast neutrons produced during fission, increasing the likelihood of further fission events. In thermal neutron reactors (such as those in the current EDF fleet), neutrons must be slowed from 20 000 km/s to around 2 km/s. This is achieved using materials like water or graphite, which effectively decelerate neutrons without absorbing them.

— **The coolant (or heat transfer fluid)** transports the thermal energy produced in the reactor core and transports it to the electricity generation system (turbine and generator). The coolant may be water (also acting as a moderator), a liquid metal (such as sodium or lead), or a gas (such as helium). Cooling the core is essential for the reactor's safety: excessive temperatures could damage structural materials and jeopardize reactor control.

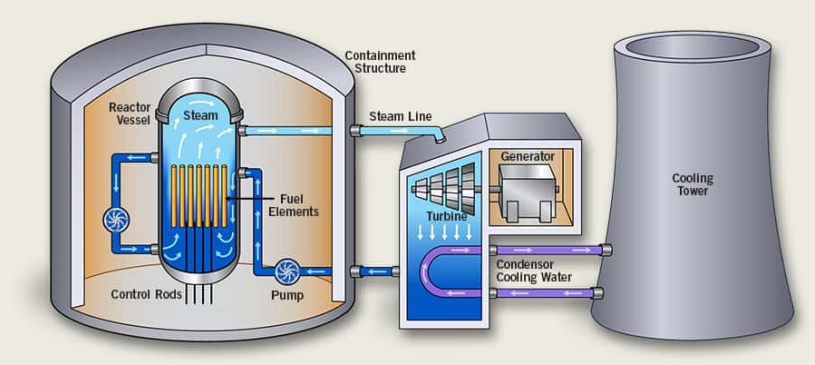


Figure 1.2 – Simplified scheme of plant with boiling water reactor [87].

The flow of this coolant within the reactor, especially under complex thermal and geometric conditions, is a key issue in nuclear engineering. This flow is generally modeled by the *Navier-Stokes equations*, which describe the motion of incompressible viscous fluids. A thorough understanding of these equations, along with efficient numerical methods for solving them, is essential for optimizing and ensuring the safety of nuclear reactor cooling systems.

Within CEA Saclay, the *Service de Thermohydraulique et de Mécanique des Fluides*[1] (STMF) conducts advanced studies and modeling activities to better understand the complex physics of single-phase and multiphase flows involved in various low-carbon energy systems. This service contributes directly to improving the design, safety, and efficiency of nuclear and thermal systems through high-fidelity simulations.

The *Laboratoire de Développement à l'Échelle Locale*[2] (LDEL), part of the STMF, specializes in the physical modeling and numerical simulation of complex local-scale flow phenomena. It develops scientific computing tools, including the TrioCFD code [35]. The TrioCFD software, developed within the STMF, offers various numerical methods for solving local problems inside the reactor, such as water flow around a fuel assembly or inside a primary circuit pipe.

The incompressible Navier-Stokes equations are among the most widely used models for describing the flow of a Newtonian fluid (i.e., a fluid whose viscosity is independent of the external forces applied to it). These equations model the velocity and pressure fields of the fluid. Let $\Omega \subset \mathbb{R}^2$ or $\mathbb{R}^3$ be the physical domain, and T_F the final time of

1. Thermal-hydraulics and Fluid Mechanics Department
2. Local-Scale Code Development Laboratory

the study. The problem consists in finding $(\vec{u}, p)$ such that

$$\begin{cases} \dfrac{\partial \vec{u}}{\partial t} - \nu \Delta \vec{u} + (\vec{u} \cdot \nabla)\vec{u} + \nabla p = \vec{f} & \text{in } (0, T_F) \times \Omega, \\ \nabla \cdot \vec{u} = 0 & \text{in } (0, T_F) \times \Omega, \end{cases}$$

where ν denotes the kinematic viscosity of the fluid, and $\vec{f}$ represents the external body forces acting on it. This system is complemented by suitable initial and boundary conditions. The first equation corresponds to Newton's second law applied to fluid motion, while the second expresses mass conservation for an incompressible fluid.
I conducted my PhD within this laboratory. My research focuses on the development and analysis of numerical methods for simulating flows governed by the Navier-Stokes equations. This work contributes to the STMF's objectives and supports the CEA's mission of advancing safe and efficient energy technologies through robust computational tools.

1.3 The Discontinuous Galerkin Finite Element Method

The Discontinuous Galerkin Finite Element Method (DG-FEM) applied to the study of incompressible fluids, and particularly to the Navier-Stokes equations, has been developped for more than 40 years and is one of the most important numerical methods in fluid mechanics.
Discontinuous Galerkin methods are a combination of ideas from continuous Galerkin methods and finite volume methods. They were proposed to improve and adapt numerical methods capable of handling complex flows with discontinuities, shocks, and regions of high variability.
The first DG method to approximate first-order PDEs was introduced by Reed and Hill [92] in the context of stationary neutron transport, while the first analysis of stationary first-order PDEs was done by Lesaint and Raviart [74] in 1974.
The error estimate was improved by Johnson and Pitkäranta [67] in 1986, who established a convergence order of $(k + 1/2)$ in the L^2 norm when polynomial degree k is used, under the condition that the exact solution is sufficiently regular.
A few years later, the method was developed by Caussignac and Touzani [18] to approximate the three-dimensional boundary layer equations of stationary incompressible flows. Around the same time, DG-FEMs were extended to unsteady hyperbolic PDEs by Chavent and Cockburn [21], using the explicit Euler scheme for time discretization with limiters.

The accuracy order was improved by Cockburn and Shu [33, 32, 34] by using explicit Runge-Kutta schemes for time discretization, while a proof of convergence to the entropic solution was obtained by Jaffré, Johnson, and Szepessy [65].
The first work on DG-FEM for PDEs modeling diffusion was that of Nitsche [86] on penalty methods for boundary conditions in the 1970s. The use of interior penalty techniques to weakly impose continuity conditions on the solution or its derivatives was explored by Babuška [9], Babuška and Zlámal [10], Douglas and Dupont [44], Baker [11], Wheeler [100], and Arnold [7]. The latter work represents a major advancement in the development of DG-FEM with interior penalty methods.
At the end of the 1990s, inspired by hyperbolic formulations, DG-FEMs were formulated using numerical fluxes by considering the mixed formulation of the diffusion term, for example, the work of Bassi and Rebay [12] on compressible Navier-Stokes equations and that of Cockburn and Shu [34] on convection-diffusion systems. These works marked a turning point in the development of DG-FEM.
Arnold, Brezzi, Cockburn, and Marini [43] introduced a unified analysis of DG-FEM for the Poisson problem, while a generalization analysis for diffuse and first-order PDEs within the framework of Friedrichs systems was developed by Ern and Guermond [46, 47].
At the level of the Stokes and Navier-Stokes equations, the application of DG-FEM became more complex due to the specific characteristics of incompressible fluid equations, the velocity-pressure coupling, the zero divergence constraint, and, in addition, for the Navier-Stokes problem, the non-linearity, unsteadiness, and their three-dimensional nature.
The first works focused on the discretization of the Stokes equations with DG-FEM adapted to ensure stability and accuracy. Formulations based on hybrid and mixed finite elements were used to correctly impose the incompressibility constraint ($\nabla \cdot u = 0$). With advancements in Stokes, DG-FEMs were gradually applied to Navier-Stokes. Stabilizing the convective terms was a key issue, addressed with different approaches (DGFDG – Interior Penalty, numerical flux methods [32]). The idea of adapting the polynomial order locally emerged to optimize the trade-off between accuracy and computational cost. For example, the book *Mathematical Aspects of Discontinuous Galerkin Methods* by Di Pietro and Ern in 2012 [40] provided an in-depth theoretical analysis of DG-FEM for incompressible fluids. It became an essential reference for understanding stability, convergence, and the practical application of DG schemes.
Since the 2010s, DG-FEMs have been extended to unsteady equations and more complex models, such as turbulent flows. Significant work has also been done on the local adaptation of polynomial order based on the solution's regularity, a way to optimize the trade-off between accuracy and computational cost [84].

The following years saw major developments in the application of DG-FEMs on high-performance computing architectures. Researchers developed adaptive schemes in time and space to handle singularities and strong gradients in the solutions. These works significantly improved the efficiency of simulations while maintaining high accuracy. Additionally, Cockburn [30], explored integrating DG-FEMs with distributed computing techniques, leading to faster and more accurate results for incompressible fluid problems in complex three-dimensional domains.
Recent approaches based on static condensation have significantly improved the efficiency of high-order methods on general meshes. In particular, Lozinski [76] proposed a condensed variant of the interior penalty discontinuous Galerkin method for elliptic problems, preserving optimal error estimates, while Di Pietro and Krell [91] developed a stable and efficient Hybrid High-Order method for the steady incompressible Navier–Stokes equations on polyhedral meshes.
Today, DG-FEMs are used for turbulent flows, and they have been developed to use higher polynomial orders, such as hybrid DG (HDG) formulations and advanced stabilization methods [85]. Implementation in industrial software like TrioCFD (CEA), FEniCS, or deal.II has allowed testing their efficiency on real-world cases.
Discontinuous Galerkin methods have several advantages for the Navier-Stokes equations:

— **Handling discontinuities and complex flows**

— **Flexibility with unstructured meshes**

— **Ease of adaptation to high orders (hp-adaptivity)**

— **Stability for turbulent flows at high Reynolds numbers**

— **Ease of parallel implementation for intensive computations**

The main drawback of DGFEM is, for a given mesh size, a higher number of degrees of freedom compared to continuous or non conforming Galerkin methods.

1.4 Outline

The objective in Part I is to study the Stokes problem discretized using the Discontinuous Galerkin (DG) method in the weakly regular case, where the velocity–pressure pair $(\mathbf{u}, p)$ satisfies $(\mathbf{u}, p) \in \mathbf{H}^{1+s}(\Omega) \times H^s(\Omega)$ with $s > \frac{1}{2}$, and Ω denotes the computational domain.
We begin by recalling the continuous formulation of the Stokes problem and showing that it is well-posed in the sense of Hadamard, i.e., it admits a unique and stable

solution. Then, we discretize the problem using a Discontinuous Galerkin scheme of the symmetric interior penalty (SIP) type.
In our approach, we extend the classical SIP formulation for the Poisson problem by allowing the stability parameter η to vary across mesh faces, introducing a face-dependent coefficient η_f. We also prove that the discretized problem remains well-posed in the sense of Hadamard. Furthermore, we show that the duality pairing $\langle \cdot, \cdot \rangle_{H^{s-1}(\Omega), H_0^{1-s}(\Omega)}$ can be decomposed over the mesh, which enables us to establish the consistency of the method. Still for the Poisson problem, we demonstrate that the coercivity constant of the bilinear form can be independent of the mesh regularity parameter σ. Next, for the Stokes problem, we then make explicit how the coercivity and continuity constants of the bilinear form depend on σ. We derive a priori error estimates for the weakly regular Stokes problem. Moreover, we propose a new numerical scheme in which the polynomial order of the pressure is one order higher than that of the velocity, as suggested by the coercivity proof of the Stokes bilinear form. Although this is counterintuitive, we have demonstrated the efficiency of this scheme when the viscosity ν is very small, compared to other Discontinuous Galerkin type schemes.
Finally, in the last chapter of Part I, we describe the resolution algorithm and present the corresponding numerical results.
In Part II, we addressed the advection–diffusion problem and established the well-posedness of both the continuous formulation and its discretization using the Discontinuous Galerkin Finite Element Method. Building upon this theoretical framework, we introduced a novel non-polynomial basis and adapted several inequalities, originally developed for polynomial functions, to functions belonging to the space spanned by this new basis. Furthermore, we explicitly computed the constants associated with these inequalities.
A detailed numerical investigation was then carried out through a series of one- and two-dimensional physical test cases. The results demonstrated the strong potential of the proposed basis to enhance both accuracy and stability, particularly in advection-dominated problems characterized by steep gradients or sharp layers.
Finally, we extended the proposed approach to the Oseen problem and to the full Navier–Stokes equations. The corresponding numerical results confirmed the effectiveness and robustness of the method in capturing complex flow features, highlighting its relevance for challenging computational fluid dynamics applications.

Chapter 2

Mathematical Tools

Contents

In this chapter, we present the mathematical tools that will be used throughout this manuscript. We introduce the new elements as well as the technical details needed to justify the integration by parts formulas in fractional Sobolev spaces under the assumptions of Theorems 2.11 and 2.12, and we provide a definition of the duality pairing.

2.1 Sobolev spaces and differential operators

In this section, we define Sobolev spaces along with the operators acting on these spaces. In general, when a space is written in bold, it represents a vector space, while when it is in double-struck, it denotes a tensor space. Let $d \geq 1$ be the spatial dimension, and let $(e_i)_{i=1,\dots,d}$ be an orthonormal basis of $\mathbb{R}^d$. In the study of the proposed models, we generally consider $d \leq 3$.
Let V is Banach space, we denote by $\mathbf{V} = [V]^d$, and $\mathbb{V} = [V]^{d\times d}$. Let Ω be a Lipschitz domain with polytopal boundary in $\mathbb{R}^d$. Let $h_\Omega := \mathrm{diam}(\Omega)$.
We recall that:

— We assume that the functions v is sufficiently regular so that we can define the partial derivative of order α is a scalar-valued operator applied to a scalar function:

$$\partial^\alpha v = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}},$$

where $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ is a multi-index and $|\alpha| = \alpha_1 + \cdots + \alpha_d$.

— Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be two Banach spaces. A linear mapping $A : X \to Y$ is continuous if:

$$\sup_{\substack{v \in X \\ v \neq 0}} \frac{\|Av\|_Y}{\|v\|_X} < \infty.$$

Its norm is then given by:

$$|||A||| := \sup_{\substack{v \in X \\ v \neq 0}} \frac{\|Av\|_Y}{\|v\|_X}. \tag{2.1}$$

The set of all continuous linear mappings from X to Y is denoted by $\mathcal{L}(X, Y)$, When $X = Y$, the notation $\mathcal{L}(X, Y)$ is simplified to $\mathcal{L}(X)$.

2.1.1 Sobolev spaces of integer order

We consider functions $v : \Omega \to \mathbb{R}$ that are Lebesgue measurable, and we denote by $\int_\Omega v$ the Lebesgue integral of v over Ω. Let $1 \leq p \leq \infty$ be a real number. We define the *L^p-norm* of v on Ω by

$$\|v\|_{L^p(\Omega)} := \left(\int_\Omega |v(x)|^p \, dx \right)^{1/p}, \quad \text{for } 1 \leq p < \infty,$$

and

$$\|v\|_{L^\infty(\Omega)} := \operatorname*{ess\,sup}_{x \in \Omega} |v(x)| = \inf \{ M > 0 \mid |v(x)| \leq M \text{ for almost every } x \in \Omega \}.$$

The *Lebesgue space* $L^p(\Omega)$ is then defined as:

$$L^p(\Omega) := \left\{ v \text{ Lebesgue measurable} \mid \|v\|_{L^p(\Omega)} < \infty \right\}.$$

Equipped with the norm $\|\cdot\|_{L^p(\Omega)}$, the space $L^p(\Omega)$ is a Banach space for all $1 \leq p \leq \infty$. Moreover, for $1 \leq p < \infty$, the space $C_c^\infty(\Omega)$ of infinitely differentiable functions with

compact support in Ω is dense in $L^p(\Omega)$. In the particular case $p = 2$, the space $L^2(\Omega)$ becomes a Hilbert space when endowed with the inner product

$$(v, w)_{L^2(\Omega)} := \int_\Omega v(x)w(x)\,dx.$$

Lemma 2.1 (Hölder's inequality) *For $1 \le p, q \le \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$, if $v \in L^p(\Omega)$ and $w \in L^q(\Omega)$, then $vw \in L^1(\Omega)$ and*

$$\int_\Omega |v(x)w(x)|\,dx \le \|v\|_{L^p(\Omega)}\,\|w\|_{L^q(\Omega)}.$$

In the special case $p = q = 2$, this yields the Cauchy–Schwarz inequality*: for all $v, w \in L^2(\Omega)$, we have $vw \in L^1(\Omega)$ and*

$$|(v, w)_{L^2(\Omega)}| \le \|v\|_{L^2(\Omega)}\,\|w\|_{L^2(\Omega)}.$$

Definition 2.1 ([49, Def. 1.29]) *Let Ω be an open set in $\mathbb{R}^d$. The elements of the following space are called locally integrable functions:*

$$L^1_{loc}(\Omega) := \left\{ v \text{ measurable} \;\middle|\; \forall \text{ compact } D \subset \Omega,\ v|_D \in L^1(D) \right\}.$$

Now, we define the Sobolev space with integer-order spaces that will be used in the following sections. Given an integer $m \ge 0$ and $p \in [1, \infty]$, the Sobolev space $W^{m,p}(\Omega)$ is defined as

$$W^{m,p}(\Omega) = \left\{ v \in L^1_{loc}(\Omega) \mid \partial^\alpha v \in L^p(\Omega),\ \forall \alpha \in \mathbb{N}^d,\ |\alpha| \le m \right\}.$$

On this space, if $p < \infty$, we introduce the norm and seminorm:

$$\|v\|_{W^{m,p}(\Omega)} = \left(\sum_{|\alpha| \le m} (h_\Omega)^{|\alpha| p} \|\partial^\alpha v\|^p_{L^p(\Omega)} \right)^{\frac{1}{p}}, \quad |v|_{W^{m,p}(\Omega)} = \left(\sum_{|\alpha| = m} \|\partial^\alpha v\|^p_{L^p(\Omega)} \right)^{\frac{1}{p}}.$$

When $m = 0$, we recover the standard $L^p(\Omega)$ space. When $p = 2$ we denote $W^{m,p}(\Omega)$ by $H^m(\Omega)$.
We let $L^2_{zmv}(\Omega) = \{v \in L^2(\Omega) \mid \int_\Omega v(x)\,dx = 0\}$.

Proposition 2.2 ([2, Thm. 3.22]) *Let $m \in \mathbb{N}$ and $p \in [1, \infty[$. In any open set Ω*

with Lipschitz boundary, the space $\mathcal{C}_c^\infty(\overline{\Omega})$ *is dense in* $W^{m,p}(\Omega)$*, that is,*

$$\forall v \in W^{m,p}(\Omega),\ \exists (v_k)_{k\in\mathbb{N}} \subset \mathcal{C}_c^\infty(\Omega), \quad \lim_{k\to\infty} \|v - v_k\|_{W^{m,p}(\Omega)} = 0.$$

We also introduce the subspace $H_0^m(\Omega)$, defined as the closure of $C_c^\infty(\Omega)$ in $H^m(\Omega)$, which consists of functions vanishing on $\partial\Omega$ in a suitable sense:

$$H_0^m(\Omega) := \overline{\mathcal{C}_c^\infty(\Omega)}^{\|\cdot\|_{H^m(\Omega)}}.$$

From now on, we choose the norm of the space $H_0^1(\Omega)$ as: $\|v\|_{H_0^1(\Omega)} := |v|_{H^1(\Omega)}$. where $|v|_{H^1(\Omega)}$ denotes the seminorm associated with $H^1(\Omega)$.
We denote by $H^{-m}(\Omega) := (H_0^m(\Omega))'$ the dual space of $H_0^m(\Omega)$.
We can define the main differential operators as follows:

— **The gradient**, a vector-valued operator applied to a scalar function:

$$\mathbf{grad}\, v := \sum_{i=1}^{d} \frac{\partial v}{\partial x_i} e_i, \quad \mathbf{grad} \in \mathcal{L}(L^2(\Omega), (\mathbf{H}_0^1(\Omega))'), \quad |||\mathbf{grad}||| \leq \sqrt{d}.$$

— **The tensor gradient**, a tensor-valued operator applied to a vector function:

$$\mathbf{Grad}\, \mathbf{v} := \left(\frac{\partial v_i}{\partial x_j}\right)_{i,j}, \quad \mathbf{Grad} \in \mathcal{L}(\mathbf{L}^2(\Omega), (\mathbb{H}_0^1(\Omega))'), \quad |||\mathbf{Grad}||| \leq \sqrt{d}.$$

— **The divergence**, a scalar-valued operator applied to a vector function:

$$\mathrm{div}\, \mathbf{v} = \sum_{i=1}^{d} \frac{\partial v_i}{\partial x_i}, \quad \mathrm{div} \in \mathcal{L}(\mathbf{H}^1(\Omega), (L^2(\Omega)), \quad |||\mathrm{div}||| \leq \sqrt{d}.$$

— **The vector divergence**, a vector-valued operator applied to a tensor function:

$$\mathrm{Div}\, \underline{\underline{G}} = \sum_{i=1}^{d} \left(\sum_{j=1}^{d} \frac{\partial \underline{\underline{G}}_{(i,j)}}{\partial x_i} \right) e_i, \quad \mathrm{Div} \in \mathcal{L}(\mathbb{H}^1(\Omega), (\mathbf{L}^2(\Omega)), \quad |||\mathrm{Div}||| \leq d.$$

— **The Laplacian**, a scalar-valued operator applied to a scalar function:

$$\Delta v = \sum_{i=1}^{d} \frac{\partial^2 v}{\partial x_i^2} \quad \Delta \in \mathcal{L}(H^1(\Omega), (H_0^1(\Omega))'), \quad |||\Delta||| \leq 1.$$

— **The vector Laplacian**, a vector-valued operator applied to a vector function:

$$\boldsymbol{\Delta}\,\mathbf{v} = \sum_{i=1}^{d} \left(\frac{\partial^2 v_j}{\partial x_j^2} \right) e_i \quad \boldsymbol{\Delta} \in \mathcal{L}(\mathbf{H}^1(\Omega), (\mathbf{H}_0^1(\Omega))'), \quad |||\boldsymbol{\Delta}||| \leq 1.$$

These operators satisfy the fundamental relation:

$$\Delta v = \operatorname{div}(\mathbf{grad}\, v), \quad \boldsymbol{\Delta} = \operatorname{Div}(\mathbf{Grad}\, v).$$

We define the following spaces:

$$\begin{aligned} \mathbf{H}(\operatorname{div};\Omega) &= \left\{ \mathbf{v} \in \mathbf{L}^2(\Omega) \mid \operatorname{div} \mathbf{v} \in L^2(\Omega) \right\}, \\ \mathbb{H}(\operatorname{Div};\Omega) &= \left\{ \underline{\underline{G}} \in \mathbb{L}^2(\Omega) \mid \operatorname{Div} \underline{\underline{G}} \in \mathbf{L}^2(\Omega) \right\}. \end{aligned}$$

This spaces are equipped with the following norms:

$$\begin{aligned} \|\mathbf{v}\|_{\mathbf{H}(\operatorname{div};\Omega)} &= \left(\|\mathbf{v}\|^2_{\mathbf{L}^2(\Omega))} + (h_\Omega)^2 \| \operatorname{div} \mathbf{v}\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}}, \\ \|\underline{\underline{G}}\|_{\mathbb{H}(\operatorname{Div};\Omega)} &= \left(\|\underline{\underline{G}}\|^2_{\mathbb{L}^2(\Omega)} + (h_\Omega)^2 \|\operatorname{Div} \underline{\underline{G}}\|^2_{\mathbf{L}^2(\Omega)} \right)^{\frac{1}{2}}. \end{aligned}$$

Let $\mathbf{V} = \{\mathbf{v} \in \mathbf{H}_0^1(\Omega) \mid \operatorname{div} \mathbf{v} = 0\}$ be a closed subspace of $\mathbf{H}_0^1(\Omega)$. We denote by $\mathbf{V}^\perp$ the orthogonal complement of $\mathbf{V}$ in $\mathbf{H}_0^1(\Omega)$. We recall that:

Property 2.3 ([57, Cor. I.2.4]) *The operator* $\operatorname{div} : \mathbf{H}_0^1(\Omega) \to L^2(\Omega)$ *is an isomorphism from* $\mathbf{V}^\perp$ *to* $L^2_{zmv}(\Omega)$. *It exists a constant* $C_{\operatorname{div}} > 0$ *such that:*

$$\forall q \in L^2_{zmv}(\Omega),\ \exists! \mathbf{v}_q \in \mathbf{V}^\perp \mid \operatorname{div} \mathbf{v}_q = q \text{ and } \|\mathbf{v}_q\|_{\mathbf{H}_0^1(\Omega)} \leq C_{\operatorname{div}} \|q\|_{L^2(\Omega)}. \tag{2.2}$$

2.1.2 Normal trace operator and Green's formula

We recall the trace theorems in the Sobolev spaces $H^1(\Omega)$ and $H(\operatorname{div};\Omega)$, as well as the integration by parts formulas in these spaces. For more details, see [29, Chapter 1].

Theorem 2.2 ([29, Thm. 1.26]) *Let* Ω *be an open subset of* $\mathbb{R}^d$ *with a bounded Lipschitz boundary. The trace operator*

$$\gamma_0 : \mathcal{C}_c^\infty(\overline{\Omega}) \to L^2(\partial\Omega), \quad v \mapsto \gamma_0 v = v|_{\partial\Omega}$$

extends continuously to a linear and continuous mapping, still denoted by γ_0*, from* $H^1(\Omega)$ *into* $L^2(\partial\Omega)$*.*

We introduce the following space:

$$H^{\frac{1}{2}}(\partial\Omega) = \left\{\lambda \in L^2(\partial\Omega) \mid \exists v \in H^1(\Omega),\ \lambda = \gamma_0 v\right\},$$

which is endowed with the norm:

$$\|\lambda\|_{H^{\frac{1}{2}}(\partial\Omega)} = \inf_{\substack{v\in H^1(\Omega)\\ \gamma_0 v=\lambda}} \|v\|_{H^1(\Omega)}.$$

We denote by $(H^{\frac{1}{2}}(\partial\Omega))'$ the dual space of $H^{\frac{1}{2}}(\partial\Omega)$.

Proposition 2.4 (`Integration by parts for functions in` $H^1(\Omega)$) *Let* Ω *be an open subset of* $\mathbb{R}^d$ *with a bounded Lipschitz boundary. For all functions* $v, w \in H^1(\Omega)$*, the following holds for every* $i = 1, \ldots, d$*:*

$$\int_\Omega \left(\frac{\partial v}{\partial x_i} w + v \frac{\partial w}{\partial x_i}\right) d\Omega = \int_{\partial\Omega} \gamma_0 v\, \gamma_0 w\, n_i \, d\Gamma. \tag{2.3}$$

where n_i *is the* i*-th component of the outward normal vector* $\mathbf{n}$ *to the boundary* $\partial\Omega$*.*

Theorem 2.3 (`[29, Thm. 1.31]`) *Let* Ω *be an open subset of* $\mathbb{R}^d$ *with a bounded Lipschitz boundary. The normal trace operator*

$$\gamma_1 : [\mathcal{C}_c^\infty(\overline{\Omega})]^d \to (H^{\frac{1}{2}}(\partial\Omega))', \quad \mathbf{v} \mapsto \gamma_1 \mathbf{v} = (\mathbf{v}\cdot\mathbf{n})|_{\partial\Omega}$$

extends continuously to a linear and surjective mapping, still denoted by γ_1*, from* $\mathbf{H}(\mathrm{div};\Omega)$ *into* $(H^{\frac{1}{2}}(\partial\Omega))'$*. Furthermore, for all* $\mathbf{v} \in \mathbf{H}(\mathrm{div};\Omega)$ *and* $w \in H^1(\Omega)$*, the following integration by parts formula holds:*

$$\int_\Omega (\mathrm{div}\,\mathbf{v} w + \mathbf{v}\cdot\mathbf{grad}\, w)\, d\Omega = \langle \gamma_1 \mathbf{v}, \gamma_0 w\rangle_{H^{\frac{1}{2}}(\partial\Omega)}. \tag{2.4}$$

2.1.3 Fractional Sobolev space

Let $D \subseteq \Omega$. We recall the definition and properties of fractional order Sobolev spaces (see e.g. [39] and references therein). For all $0 < s < 1$, we set:

$$H^s(D) := \left\{ v \in L^2(D) \,|\, \int_D \int_D \frac{|v(\mathbf{x}) - v(\mathbf{y})|^2}{|\mathbf{x} - \mathbf{y}|^{d+2s}} \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} < \infty \right\}. \tag{2.5}$$

Let v, $w \in H^s(D)$. We set:

$$(v, w)_{s,D} = \int_D \int_D \frac{(v(\mathbf{x}) - v(\mathbf{y}))\,(w(\mathbf{x}) - w(\mathbf{y}))}{|\mathbf{x} - \mathbf{y}|^{d+2s}} \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}, \quad |v|_{H^s(D)} = (v, v)_{s,D}^{\frac{1}{2}}.$$

We can define the scalar product and the norm in $H^s(D)$ so that [1]:

$$(u, v)_{H^s(D)} = (u, v)_{L^2(D)} + h_D^{2s}(v, w)_{s,D}, \quad \|v\|_{H^s(D)} = (v, v)_{H^s(D)}^{\frac{1}{2}}. \tag{2.6}$$

We define $\mathbf{H}^s(D) := [H^s(D)]^d$, and the norm in $H^1(D)$ and $H^{1+s}(D)$ as follows:

$$\begin{aligned} \|v\|_{H^1(D)} &:= \left(\|v\|_{L^2(D)}^2 + h_D^2 \, \|\mathbf{grad}\, v\|_{\mathbf{L}^2(D)}^2 \right)^{\frac{1}{2}}. \\ \|v\|_{H^{1+s}(D)} &:= \left(\|v\|_{H^1(D)}^2 + h_D^{2s} |\mathbf{grad}\, v|_{\mathbf{H}^s(D)}^2 \right)^{\frac{1}{2}}. \end{aligned}$$

We can prove a result similar to the Poincaré-Steklov inequality (2.9) to Sobolev spaces with fractional index:

Lemma 2.4 ([48, Lem. 7.1]) *For any $v \in H^s(D) \cap L^2_{zmv}(D)$ with $s \in (0, 1)$, the following inequality holds:*

$$\|v\|_{L^2(D)} \leq h_D^s \left(\frac{h_D^d}{|D|} \right)^{\frac{1}{2}} |v|_{H^s(D)}. \tag{2.7}$$

Let $H_0^s(D)$ be the closure of $\mathcal{C}_c^\infty(D)$ in $H^s(D)$, and $H^{-s}(D) = (H_0^s(D))'$. We will need the following theorem:

Theorem 2.5 ([49, Thm 3.19]) *The following properties hold*

1. *For all $t \in (0, \frac{1}{2})$, one has: $H_0^t(D) = H^t(D)$.*

1. In this article, s is fixed. However, if we consider the limit as $s \to 0$ or $s \to 1$, the semi-norm $|\cdot|_{H^s(\Omega)}$ must be multiplied by $s(1-s)$.

2. *For all* $t \in (0, \frac{1}{2})$, *the mapping* $u \mapsto \tilde{u}$, *where* $\tilde{u}$ *is the extension of* u *by* 0 *outside* D *is a continuous mapping of* $H^t(D) \to H^t(\mathbb{R}^d)$.

Theorem 2.6 ([75, Thms 2.21], Meyers-Serrin) *Let* $s \geq 0$ *and* $p \in [1, \infty)$. *Then*

$$C^\infty(\Omega) \cap W^{s,p}(\Omega) \text{ is dense in } W^{s,p}(\Omega).$$

Let $f \in H^{-r}(\Omega)$, with $r \geq 0$ we define

$$\|f\|_{H^{-r}(\Omega)} := \sup_{\substack{v \in H_0^r(\Omega) \\ v \neq 0}} \frac{\langle f, v \rangle_{H^r(\Omega)}}{\|v\|_{H^r(\Omega)}} \tag{2.8}$$

Note that if $r \in (0, \frac{1}{2})$, the norm $\|\cdot\|_{H^r(\Omega)}$ can be replaced by the semi-norm $|\cdot|_{H_0^r(\Omega)}$, since the norm and the semi-norm are equivalent. However, as r tends to 0, by the Poincaré inequality (2.12), one has $\|f\|_{H^{-r}(\Omega)} \lesssim h_\Omega \|f\|_{L^2(\Omega)}$.

2.1.4 Trace and Poincaré inequalities

Definition 2.5 *A function* $f : \Omega \to \mathbb{R}$ *is said to be of class* $\mathcal{C}^{k,\alpha}(\Omega)$, *with* $k \in \mathbb{N}$ *and* $\alpha \in (0, 1]$, *if* f *is* k*-times continuously differentiable on* Ω, *i.e.,* $f \in \mathcal{C}^k(\Omega)$, *and all partial derivatives of order* k *are* α*-Hölder continuous; that is, there exists a constant* $C > 0$ *such that for all* $x, y \in \Omega$,

$$|\partial^\beta f(x) - \partial^\beta f(y)| \leq C|x - y|^\alpha \quad \text{for all multi-indices } \beta \text{ with } |\beta| = k.$$

Remark 2.6 *A* $C^{0,1}$ *function corresponds to a Lipschitz continuous function.*

Definition 2.7 *A bounded open set* $D \subset \mathbb{R}^d$ *is said to be of class* $\mathcal{C}^{k,\alpha}$ *(*$k \in \mathbb{N}$, $\alpha \in (0, 1]$*) if, for every point* $x_0 \in \partial D$, *there exists a neighborhood* U *of* x_0 *and a function* $\varphi \in \mathcal{C}^{k,\alpha}(\mathbb{R}^{d-1})$ *such that, in a suitable coordinate system,*

$$D \cap U = \{(x', x_d) \in U : x_d > \varphi(x')\}.$$

In other words, the boundary ∂D *is locally the graph of a* $C^{k,\alpha}$ *function.*

We will need the following trace theorem: Let D be a connected subdomain of Ω, with Lipschitz boundary. We recall the following theorem (see §2.1.3 for the definition of $H^s(D)$, $s \in (\frac{1}{2}, 1)$).

Theorem 2.7 (**[79, Thm. 3.37]**, $k=1$) *Define the trace operator $\gamma_0 : \mathcal{C}_c^\infty(\overline{D}) \to \mathcal{C}_c^\infty(\partial D)$ by $\gamma_0 v = v|_{\partial D}$. If D is a Lipschitz domain, and $s \in]\frac{1}{2}, 1]$, then γ_0 has a unique extension to a bounded linear operator*

$$\gamma_0 : H^s(D) \to H^{s-\frac{1}{2}}(\partial D),$$

and this extension has a unique continuous right inverse.

Hence, the norm in $H^{s-\frac{1}{2}}(\partial D)$ is such that:

$$\forall g \in H^{s-\frac{1}{2}}(\partial D),\ \|g\|_{H^{s-\frac{1}{2}}(\partial D)} \equiv \inf_{v \in \{H^s(D)\,|\,v|_{\partial D}=g\}} \|v\|_{H^s(D)}.$$

We deduce that there exists a dimensionless constant $C_{\gamma_0,s,D} > 0$ such that:

$$\forall v \in H^s(D), \quad \|v_{|D}\|_{H^{s-\frac{1}{2}}(\partial D)} \leq C_{\gamma_0,s,D}\,(h_D)^{-\frac{1}{2}} \|v\|_{H^s(D)}. \tag{2.9}$$

Let $H^{-\frac{1}{2}}(\partial\Omega)$ be the dual space of $H^{\frac{1}{2}}(\partial\Omega)$ and $\mathbf{H}^{\frac{1}{2}}(\partial\Omega) := [H^{\frac{1}{2}}(\partial\Omega)]^d$. The norm in $H^{-\frac{1}{2}}(\partial\Omega)$ is defined by:

$$\forall g \in H^{-\frac{1}{2}}(\partial\Omega),\ \|g\|_{H^{-\frac{1}{2}}(\partial\Omega)} = \sup_{v \in H^{\frac{1}{2}}(\partial\Omega)\setminus\{0\}} \frac{\langle g, v\rangle_{H^{-\frac{1}{2}}(\partial\Omega), H^{\frac{1}{2}}(\partial\Omega)}}{\|v\|_{H^{\frac{1}{2}}(\partial\Omega)}}. \tag{2.10}$$

Theorem 2.8 (**[8, Thm. 2.2.22]**, $s=0$) *Define the trace operator $\gamma_1 : (\mathcal{C}_c^\infty(\overline{\Omega}))^d \to \mathcal{C}_c^\infty(\partial\Omega)$ by $\gamma_1 \mathbf{v} = \mathbf{v}\cdot\mathbf{n}|_{\partial\Omega}$. If Ω is a Lipschitz domain, then γ_1 has a unique continuous extension, still denoted by γ_1, from $\mathbf{H}(\mathrm{div};\Omega)$ into $H^{-\frac{1}{2}}(\partial\Omega)$, which is surjective.*

Hence, the norm in $H^{-\frac{1}{2}}(\partial\Omega)$ is such that:

$$\forall g \in H^{-\frac{1}{2}}(\partial\Omega),\ \|g\|_{H^{-\frac{1}{2}}(\partial\Omega)} \equiv \inf_{\mathbf{v} \in \{\mathbf{H}(\mathrm{div};\Omega)\,|\,\mathbf{v}\cdot\mathbf{n}|_{\partial\Omega}=g\}} \|\mathbf{v}\|_{\mathbf{H}(\mathrm{div};\Omega)}.$$

We deduce that there exists a dimensionless constant $C_{\gamma_1,\Omega} > 0$ such that:

$$\forall \mathbf{v} \in \mathbf{H}(\mathrm{div};\Omega), \quad \|\mathbf{v}\cdot\mathbf{n}|_{\partial\Omega}\|_{H^{-\frac{1}{2}}(\partial\Omega)} \leq C_{\gamma_1,\Omega}\,(h_\Omega)^{-\frac{1}{2}} \|\mathbf{v}\|_{\mathbf{H}(\mathrm{div};\Omega)}. \tag{2.11}$$

For all $v \in L^2(\Omega)$, we denote by $\underline{v} := \dfrac{1}{\Omega}\displaystyle\int_\Omega v$. We will recall the Poincaré-Steklov theorem see:

Lemma 2.9 ([49, Lemma 3.24]) *Let Ω be a Lipschitz domain in $\mathbb{R}^d$. Let $h_\Omega := diam(\Omega)$. There exists a constant C_{Pzmv} such that*

$$C_{Pzmv}\|v - \underline{v}\|_{L^2(\Omega)} \leq h_\Omega |v|_{H^1(\Omega)}, \quad \forall v \in H^1(\Omega),$$

If Ω is convex, we have $C_{Pzmv} = \pi$.

Lemma 2.10 ([49, Lem. 3.27]) *Let Ω be a Lipschitz domain, and $v \in W_0^{1,p}(\Omega)$, There exists a constant $C_{PS,p} > 0$ such that*

$$\|v\|_{L^p(\Omega)} \leq C_{PS,p} h_\Omega \|\mathbf{grad}\, v\|_{\mathbf{L}^p(\Omega)}. \tag{2.12}$$

If $p = 2$, we simplify $C_{PS,p}$ to C_{PS}.

In the following two theorems, we generalize the integration by parts formulas (2.3) and (2.4) in a specific subspace of the fractional Sobolev space. Let $D \subset \mathbb{R}^d$ be a connected and bounded domain of $\mathbb{R}^d$, with Lipschitz boundary ∂D for $d = 2$ and polyhedral boundary ∂D for $d = 3$.

Theorem 2.11 *Let $s \in (\frac{1}{2}, 1)$. Consider the operator Δ as follows:*

$$\begin{aligned} \Delta : C^\infty(\overline{D}) &\to (H^{1+s}(D))' \\ u &\mapsto f_u \end{aligned}$$

where the mapping f_u is defined on $H^{1+s}(D)$ as follows:

$$\begin{aligned} f_u : H^{1+s}(D) &\to \mathbb{R} \\ v &\mapsto (\Delta u, v)_{L^2(D)}. \end{aligned}$$

The operator Δ is a linear and continuous operator on $\left(\mathcal{C}_c^\infty(\overline{D}), \|\cdot\|_{H^{1+s}(D)}\right)$. It can be uniquely and continuously extended as an operator acting from $H^{1+s}(D)$ to $(H^{1+s}(D))'$ and the following integration by parts formula holds:

$$\begin{aligned} &\forall (u, v) \in H^{1+s}(D) \times H^{1+s}(D), \\ &\langle f_u, v\rangle_{H^{1+s}(D)} = -(\mathbf{grad}\, u, \mathbf{grad}\, v)_{\mathbf{L}^2(D)} + (\mathbf{grad}\, u \cdot \mathbf{n}, v)_{L^2(\partial D)}. \end{aligned} \tag{2.13}$$

PROOF. Let $(u,v) \in \mathcal{C}^\infty(\overline{D}) \times H^{1+s}(D)$. Using Cauchy-Schwarz and the trace theorem in $H^s(D)$ (cf. Theorem 2.7), we obtain:

$$\begin{aligned}
&|\langle f_u, v\rangle_{H^{1+s}(D)}| = |(\Delta u, v)_{L^2(D)}|,\\
&\quad = \left|-(\mathbf{grad}\, u, \mathbf{grad}\, v)_{\mathbf{L}^2(D)} + (\mathbf{grad}\, u \cdot \mathbf{n}, v)_{L^2(\partial D)}\right|,\\
&\quad \leq \|\mathbf{grad}\, u\|_{\mathbf{L}^2(D)}\, \|\mathbf{grad}\, v\|_{\mathbf{L}^2(D)} + \|\mathbf{grad}\, u \cdot \mathbf{n}\|_{L^2(\partial D)}\, \|v\|_{L^2(\partial D)},\\
&\quad \leq \|\mathbf{grad}\, u\|_{\mathbf{H}^s(D)}\, \|v\|_{H^{1+s}(D)} + \|\mathbf{grad}\, u \cdot \mathbf{n}\|_{H^{s-\frac{1}{2}}(\partial D)}\, \|v\|_{H^{\frac{1}{2}+s}(\partial D)},\\
&\quad \lesssim \|\mathbf{grad}\, u\|_{\mathbf{H}^s(D)} \|v\|_{H^{1+s}(D)} \lesssim \|u\|_{H^{1+s}(D)} \|v\|_{H^{1+s}(D)}.
\end{aligned}$$

We deduce that $\|f_u\|_{(H^{1+s}(D))'} \lesssim \|u\|_{H^{1+s}(D)}$, and since $\left(\mathcal{C}_c^\infty(\overline{D}),\ \|\cdot\|_{H^{1+s}(D)}\right)$ is dense in $H^{1+s}(D)$, according to the Hahn-Banach theorem, we can uniquely extend the operator Δ to an operator acting from $H^{1+s}(D)$ to $(H^{1+s}(D))'$. Consider now $u \in H^{1+s}(D)$. There exists a sequence $(u_n)_n \in \mathcal{C}_c^\infty(\overline{D})$ such that $\lim\limits_{n\to\infty} \|u - u_n\|_{H^{1+s}(D)} = 0$. By using the continuity of the operator Δ, we obtain that $\lim\limits_{n\to\infty} \|f_u - f_{u_n}\|_{(H^{1+s}(D))'} = 0$. Hence, $\lim\limits_{n\to\infty} |\langle f_{u_n}, v\rangle_{H^{1+s}(D)} - \langle f_u, v\rangle_{H^{1+s}(D)}| = 0$ for all $v \in H^{1+s}(\Omega)$ and we have $\lim\limits_{n\to\infty} u_n \to u$ in $H^1(D)$ and in $H^{\frac{1}{2}}(\partial D)$.
We deduce that for all $v \in H^{1+s}(D)$:

$$\begin{aligned}
\langle f_u, v\rangle_{H^{1+s}(D)} &= \lim_{n\to\infty} \langle f_{u_n}, v\rangle_{H^{1+s}(D)} = \lim_{n\to\infty} (\Delta u_n, v)_{L^2(D)}\\
&= \lim_{n\to\infty} \left(-(\mathbf{grad}\, u_n, \mathbf{grad}\, v)_{\mathbf{L}^2(D)} + (\mathbf{grad}\, u_n \cdot \mathbf{n}, v)_{L^2(\partial D)}\right)\\
&= -(\mathbf{grad}\, u, \mathbf{grad}\, v)_{\mathbf{L}^2(D)} + (\mathbf{grad}\, u \cdot \mathbf{n}, v)_{L^2(\partial D)}.
\end{aligned}$$

□

Theorem 2.12 *Let $s \in (\frac{1}{2}, 1)$. Consider the operator* **grad** *as follows:*

$$\begin{aligned}
\mathbf{grad}\ : \mathcal{C}_c^\infty(\overline{D}) &\to (\mathbf{H}^{1+s}(D))'\\
q &\mapsto g_q
\end{aligned},$$

where the mapping g_q is defined on $\mathbf{H}^{1+s}(D)$ as follow:

$$\begin{aligned}
g_q : \mathbf{H}^{1+s}(D) &\to \mathbb{R}\\
\mathbf{v} &\mapsto (\mathbf{grad}\, q, \mathbf{v})_{\mathbf{L}^2(D)}
\end{aligned}$$

The operator **grad** *is a linear and continuous operator on $\left(C_c^\infty(\overline{D}),\ \|\cdot\|_{H^s(D)}\right)$. It can*

be uniquely and continuously extended as an operator acting from $H^s(D)$ *to* $(\mathbf{H}^{1+s}(D))'$ *and we have the following integration by parts formula:*

$$\begin{aligned} &\forall (q,\mathbf{v}) \in H^s(D)\times \mathbf{H}^{1+s}(D),\\ &\langle g_q,\mathbf{v}\rangle_{\mathbf{H}^{1+s}(D)} = -(\operatorname{div}\mathbf{v},q)_{L^2(D)} + (\mathbf{v}\cdot\mathbf{n},q)_{L^2(\partial D)}. \end{aligned} \tag{2.14}$$

Proof. Let $(q,\mathbf{v}) \in C_c^\infty(\overline{D})\times \mathbf{H}^{1+s}(D)$. Using Cauchy-Schwarz and the trace theorem in $H^s(D)$ (cf. Theorem 2.7), we obtain:

$$\begin{aligned} \langle g_q,\mathbf{v}\rangle_{\mathbf{H}^{1+s}(D)} &= (\mathbf{grad}\, q,\mathbf{v})_{\mathbf{L}^2(D)} = -(\operatorname{div}\mathbf{v},q)_{L^2(D)} + (\mathbf{v}\cdot\mathbf{n},q)_{L^2(\partial D)}\\ &\leq \|\operatorname{div}\mathbf{v}\|_{L^2(D)}\|q\|_{L^2(D)} + \|\mathbf{v}\|_{\mathbf{L}^2(\partial D)}\|q\|_{L^2(\partial D)},\\ &\lesssim \|\mathbf{v}\|_{\mathbf{H}^{1+s}(D)}\|q\|_{H^s(D)} + \|\mathbf{v}\|_{\mathbf{H}^{s-\frac{1}{2}}(\partial D)}\|q\|_{H^{s-\frac{1}{2}}(\partial D)},\\ &\lesssim \|\mathbf{v}\|_{\mathbf{H}^{1+s}(D)}\|q\|_{H^s(D)} \end{aligned}$$

We deduce that $\|g_q\|_{(\mathbf{H}^{1+s}(D))'} \lesssim \|q\|_{H^s(D)}$, and since $\left(C_c^\infty(\overline{D}),\ \|\cdot\|_{H^s(D)}\right)$ is dense in $H^s(D)$, according to the Hahn-Banach theorem, we can uniquely extend the operator **grad** as an operator acting from $H^s(D)$ to $(\mathbf{H}^{1+s}(D))'$. Consider now $q \in H^s(D)$. There exists a sequence $(q_n)_n \in C_c^\infty(\overline{D})$ such that $\lim\limits_{n\to\infty}\|q-q_n\|_{H^s(D)} = 0$. By using the continuity of the operator **grad**, we obtain that $\lim\limits_{n\to\infty}\|g_q - g_{q_n}\|_{(H^{1+s}(D))'} = 0$. Hence, $\lim\limits_{n\to\infty}|\langle g_{q_n},\mathbf{v}\rangle_{\mathbf{H}^{1+s}(D)} - \langle g_q,\mathbf{v}\rangle_{\mathbf{H}^{1+s}(D)}| = 0$ for all $\mathbf{v}\in\mathbf{H}^{1+s}(D)$ and we have $\lim\limits_{n\to\infty} q_n \to q$ in $L^2(D)$ and in $L^2(\partial D)$. We deduce that $\forall \mathbf{v}\in\mathbf{H}^{1+s}(D)$:

$$\begin{aligned} \langle g_q,\mathbf{v}\rangle_{\mathbf{H}^{1+s}(D)} &= \lim_{n\to\infty}\langle g_{q_n},\mathbf{v}\rangle_{\mathbf{H}^{1+s}(D)} = \lim_{n\to\infty}(\mathbf{grad}\, q_n,\mathbf{v})_{\mathbf{L}^2(D)}\\ &= \lim_{n\to\infty}\left(-(\operatorname{div}\mathbf{v},q_n)_{L^2(D)} + (\mathbf{v}\cdot\mathbf{n},q_n)_{L^2(\partial D)}\right)\\ &= -(\operatorname{div}\mathbf{v},q)_{L^2(D)} + (\mathbf{v}\cdot\mathbf{n},q)_{L^2(\partial D)} \end{aligned}$$

□

Remark 2.8 *Note that the assumptions in Theorems 2.11 and 2.12 are not optimal. For more details, see [60, Paragraph 5.1.2] and [61, Paragraph 1.2]. We can optimize the choice of the space* $\mathbf{v}$ *in* $\mathbf{H}^1(\Omega)$.

2.2 T-coercivity

Let V and W be two Hilbert spaces equipped with norms $\|\cdot\|_V$ and $\|\cdot\|_W$, respectively. Let V' and W' be the dual spaces of V and W; $(\cdot,\cdot)_V$ and $(\cdot,\cdot)_W$ the inner products of V and W; and $\langle\cdot,\cdot\rangle_V$ and $\langle\cdot,\cdot\rangle_W$ the dual pairings of V and W, respectively. Let Ω be a bounded, open, connected domain in $\mathbb{R}^d$ ($d = 2$ or 3) with a bounded Lipschitz boundary.

Definition 2.9 *Let $a(\cdot,\cdot)$ be a continuous bilinear form defined on $V \times V$. We say that $a(\cdot,\cdot)$ is coercive on V if and only if there exists a constant $c > 0$ such that for all $x \in V$,*

$$|a(x,x)| \geq c\|x\|_V^2.$$

Definition 2.10 (`Basic T-coercivity, [38, Def. 7.1], [26]`) *Let $a(\cdot,\cdot)$ be a continuous bilinear form defined on $V \times W$. We say that $a(\cdot,\cdot)$ is T-coercive on V if and only if:*

$$\exists T \in \mathcal{L}(V,W),\ \textit{bijective},\ \exists \alpha > 0,\ \forall v \in V,\ |a(v,Tv)| \geq \alpha\|v\|_V^2. \tag{2.15}$$

Obviously, if (2.15) holds for some $T \in \mathcal{L}(V,W)$, then T is injective.

Now consider that $a(\cdot,\cdot)$ is a continuous bilinear form defined on $V \times W$ and $f \in W'$. We define the following problem:

$$\begin{cases} \text{Find } v \in V \text{ such that:} \\ \forall w \in W, \quad a(v,w) = \langle f, w\rangle_{W',W}. \end{cases} \tag{2.16}$$

Definition 2.11 (`Well-posedness`) *We say that problem (2.16) is well-posed in the sense of Hadamard if and only if, for all $f \in W'$, problem (2.16) has a unique solution $v \in V$, which depends continuously on the data:*

$$\exists c > 0 \textit{ such that } \|v\|_V \leq c\|f\|_{W'}.$$

Definition 2.12 (`Inf-sup Condition`) *The bilinear form $a(\cdot,\cdot)$ satisfies the inf-sup condition if and only if there exists a constant $\alpha \geq 0$ such that for all $v \in V$,*

$$\sup_{w \in W\setminus\{0\}} \frac{|a(v,w)|}{\|w\|_W} \geq \alpha\|v\|_V. \tag{2.17}$$

Definition 2.13 (`Solvability Condition`) *The bilinear form $a(\cdot,\cdot)$ satisfies the solvability condition if and only if*

$$\{w \in W \mid a(v,w) = 0,\ \forall v \in V\} = \{0\}. \tag{2.18}$$

Remark 2.14 *If $W = V$ and $a(\cdot,\cdot)$ is an Hermitian bilinear form, then the condition (2.17) implies the condition (2.18). Indeed, let $w \in V$ such that for all $v \in V$, $a(v,w) = 0$, then*

$$0 = \sup_{v \in V\setminus\{0\}} \frac{|a(v,w)|}{\|v\|_V} = \sup_{v \in V\setminus\{0\}} \frac{|a(w,v)|}{\|v\|_V} \geq \alpha \|w\|_V.$$

Hence, we obtain $w = 0$.

Theorem 2.13 (`Lax-Milgram [40, Lemma 1.4]`) *If the bilinear form $a(\cdot,\cdot)$ is continuous and coercive defined on $V \times V$, then the problem (2.16) is well-posed.*

Theorem 2.14 (`Banach-Nečas-Babuška (BNB) [40, Thm. 1.1]`) *The problem (2.16) is well-posed if and only if the conditions (2.17) and (2.18) are satisfied.*

Theorem 2.15 (`Well-posedness [27, 22]`) *Let $a(\cdot,\cdot)$ be a continuous bilinear form on $V \times W$. The problem (2.16) is well-posed if, and only if, the form $a(\cdot,\cdot)$ is T-coercive in the sense of definition 2.10.*

Proof.
For all $w \in W \setminus \{0\}$, since T is a continuous and bijective linear operator, there exists $u \in V \setminus \{0\}$ such that $Tu = w$. Moreover, we have

$$0 < \|Tu\|_W \leq |||T||| \, \|u\|_V.$$

Since the bilinear form $a(\cdot,\cdot)$ is T-coercive, we obtain

$$\sup_{w \in W\setminus\{0\}} \frac{|a(u,w)|}{\|w\|_W} \geq \alpha \, |||T|||^{-1} \, \|u\|_V.$$

Assume that there exists $w \in W$ such that $a(v,w) = 0$ for all $v \in V$. Taking $v = T^{-1}(w)$, we get

$$\alpha \|T^{-1}(w)\|_V^2 \leq |a(T^{-1}(w), w)| = 0,$$

which implies $T^{-1}(w) = 0$, and therefore $w = 0$.
Thus, the variational problem (2.16) is well posed according to the Banach–Nečas–Babuška (BNB) theorem (see Theorem 2.14).

□

Chapter 3

Modeling by the Navier–Stokes Problem

Contents

In this chapter, we aim to model incompressible fluid flow on a fixed spatial domain Ω. To achieve this, we first introduce the notion of an incompressible fluid, then apply the conservation of mass and Newton's second law to such a fluid in order to derive the Navier–Stokes problem. Although this chapter does not contain new material, it provides a clear and pedagogical presentation of how the Navier–Stokes equations are obtained.

Many references have addressed the modeling of the Navier-Stokes equations, such as [20, 54, 70, 14]. In this chapter, we mainly rely on the textbook by Boyer and Fabrie [15], as well as the PhD thesis of Peitavy [90].

3.1 Incompressible fluid

A fluid is a material that continuously deforms under the action of a force, in contrast to a solid. This category includes liquids (such as oil or milk), gases (such as air), and plasmas. However, not all fluids respond in the same way when subjected to external forces: for instance, a viscous paste like cement does not flow like air or water.

The Navier-Stokes equations studied in this manuscript describe a specific class of fluids, modeled under several idealized yet reasonable assumptions.
First, the fluid is assumed to be continuous: it is treated as a homogeneous medium without microscopic structure, in which physical quantities such as velocity, pressure, temperature, and density vary smoothly across space.
Second, the fluid is assumed to be incompressible: its volume remains constant over time, although it can still deform. In other words, fluid particles may move and rearrange, but without changing volume.
We also consider homogeneous fluids (with constant density throughout the domain) and isotropic fluids (with no preferred direction in space).
Finally, the analysis focuses on Newtonian fluids, for which viscous stresses are proportional to the rate of deformation. The behavior of such fluids does not depend on the flow history or on the intensity of the applied forces. For example, air and water remain fluid regardless of the flow speed, unlike certain dense suspensions (such as a water–starch mixture), whose viscosity can vary significantly depending on the flow regime.

3.2 Navier-Stokes problem

We begin our introduction with the assumption that the domain $\Omega \subset \mathbb{R}^3$ is a continuous medium, meaning it can be described using macroscopic quantities.

Definition 3.1 *The description as a continuous medium is said to be valid if there exists a function*

$$\rho : [0, T] \times \Omega \to \mathbb{R}^+$$

called the density field, such that the mass contained in any fluid volume $\omega \subset \Omega$ at time t is given by:

$$\textit{Mass in } \omega \textit{ at time } t = \int_\omega \rho(t, X)\, dX.$$

This function ρ allows mass to be represented as a continuous quantity, thus justifying the use of a continuous medium description.

This definition ensures that the material can be modeled as a continuum, which is a necessary condition for applying the tools of fluid mechanics.
The modeling will focus on an incompressible fluid (see 3.1), which means its density ρ is constant. Its unit is kilograms per cubic meter: $[\,\rho\,] = \mathrm{kg} \cdot \mathrm{m}^{-3}$.

We denote by $\mathbf{u} = (u_i)_{1 \leqslant i \leqslant d}$ the velocity of the fluid, with unit meters per second: $[\,\mathbf{u}\,] = \mathrm{m} \cdot \mathrm{s}^{-1}$.
The fluid pressure is denoted by P, with unit Pascal, that is: $[\,P\,] = \mathrm{kg} \cdot \mathrm{m}^{-1} \cdot \mathrm{s}^{-2}$.
The dynamic viscosity of the fluid is denoted by μ, with unit: $[\,\mu\,] = \mathrm{kg} \cdot \mathrm{m}^{-1} \cdot \mathrm{s}^{-1}$.
The kinematic viscosity is defined as the ratio $\nu = \mu/\rho$, and has unit: $[\,\nu\,] = \mathrm{m}^2 \cdot \mathrm{s}^{-1}$.
To describe the motion of particles in a fluid, we use Lagrangian coordinates. This means we fix a particle or a volume of fluid and follow its trajectory over time. In two dimensions, for example, this leads to a curve that describes the motion of the particle. To illustrate this approach, let us consider an initial fluid volume $\omega_0 \subset \Omega$. We introduce a family of bijective mappings $\boldsymbol{\varphi}(t, \cdot)$ such that, for each time t, the volume:

$$\omega_t := \boldsymbol{\varphi}(t, \omega_0) := \boldsymbol{\varphi}_t(\omega_0)$$

contains exactly the same fluid particles at time t as those that were in ω_0 at the initial time t_0.
In order to model the Navier–Stokes equations, we are particularly interested in describing the motion of particles.

Definition 3.2 *The family $(\omega_t)_t$ indexed by time t is called a fluid element. In the case where ω_0 is reduced to a single point $\omega_0 = \{\mathbf{x}_0\}$, the family is referred to as a fluid particle.*

As a consequence, if we take $\omega_0 = \{\mathbf{x}_0\}$, then φ_t describes the position of the particle relative to a fixed reference frame at time t. We introduce the following function:

$$\mathbf{X}(t, t_0, \mathbf{x}_0) := \boldsymbol{\varphi}_t(\mathbf{x}_0),$$

which represents the position, at time t, of the fluid particle that was located at $\mathbf{x}_0$ at the initial time t_0. This function, called the trajectory, describes the evolution of a fluid particle over time.
Knowing $\mathbf{X}$ for all t and all initial points $\mathbf{x}_0 \in \Omega$ would allow us to fully track the motion of particles throughout the domain. However, it is often more practical to adopt a macroscopic description, which leads us to define the velocity field of the fluid:

$$\mathbf{u}(t, \mathbf{x}) = \left(\frac{\partial \mathbf{X}(s, t, \mathbf{x})}{\partial s} \right) \tag{3.1}$$

Since $\mathbf{x}$ is the position of $\mathbf{x}_0$ at time t, we have $\mathbf{x} = \boldsymbol{\varphi}_t(\mathbf{x}_0)$, which can equivalently be written as:

$$\mathbf{u}\left(t, \boldsymbol{\varphi}_t\left(\mathbf{x}_0\right)\right) = \left(\frac{\partial \boldsymbol{\varphi}_s\left(\mathbf{x}_0\right)}{\partial s}\right)_{|s=t}$$

The pair $(t, \mathbf{X})$ defines the Eulerian coordinates of the system. Working in this framework amounts to fixing a point $\mathbf{X}$ in space and formulating the conservation equations at that fixed location.

Let $f(t, \mathbf{X})$ be a characteristic quantity of the fluid (for instance, the density or temperature). To analyze its evolution along the trajectory of a fluid particle, we compute the total time derivative of f along the trajectory:

$$\begin{aligned}\frac{d}{dt}\left(f(t, \mathbf{X}(t, t_0, \mathbf{x}_0))\right) &= \frac{\partial f}{\partial t}(t, \mathbf{X}(t, t_0, \mathbf{x}_0)) + \frac{\partial \mathbf{X}(t, t_0, \mathbf{x}_0)}{\partial t} \cdot \mathbf{grad}\, f(t, \mathbf{X}(t, t_0, \mathbf{x}_0)) \\ &= \frac{\partial f}{\partial t}(t, \mathbf{X}(t, t_0, \mathbf{x}_0)) + \mathbf{u}(t, \mathbf{X}(t, t_0, \mathbf{x}_0)) \cdot \mathbf{grad}\, f(t, \mathbf{X}(t, t_0, \mathbf{x}_0)).\end{aligned}$$

This quantity is called the Eulerian derivative (or material derivative), and is written as:

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \mathbf{u} \cdot \mathbf{grad}\, f.$$

To ensure the validity of this formulation, we assume that the trajectory field $\boldsymbol{\varphi}$ is a $\mathcal{C}^1$-diffeomorphism, that is, a bijective and continuous mapping whose partial derivatives are also continuous.

We recall the following theorem [15, Theorem I.2.1]:

Theorem 3.1 *(Transport theorem) Let f be a function of class $\mathcal{C}^1$ with respect to the variables $(t, \mathbf{X}) \in \mathbb{R} \times \mathbb{R}^3$. Then, for any t, we have:*

$$\frac{d}{dt}\int_{\omega_t} f(t, \mathbf{X})\, d\mathbf{X} = \int_{\omega_t}\left(\frac{\partial f}{\partial t} + \operatorname{div}(f\mathbf{u})\right) d\mathbf{X},$$

where $\mathbf{u}$ is a vector field in $\mathbb{R}^3$ defined by equation (3.1).

Definition 3.3 *Let $\mathbf{a} = (a_1, a_2, \ldots, a_d) \in \mathbb{R}^d$ and $\mathbf{b} = (b_1, b_2, \ldots, b_d) \in \mathbb{R}^d$ be two vectors. The tensor product $\mathbf{a} \otimes \mathbf{b}$ is the $d \times d$ matrix defined by:*

$$(\mathbf{a} \otimes \mathbf{b})_{ij} = a_i b_j, \quad \text{for } i, j = 1, \ldots, d. \tag{3.2}$$

In other words, each element of the matrix is the product of the i-th component of $\mathbf{a}$ *with the j-th component of* $\mathbf{b}$.

3.2.1 Conservation equations

Mass conservation

We can now express the mass conservation equation on the subdomains ω_t, representing the evolution over time of a fluid volume. Using the definition of the density field (see equation (3.1)), this conservation law takes the integral form:

$$\frac{d}{dt}\int_{\omega_t} \rho \, d\mathbf{X} = 0.$$

By applying the transport theorem 3.1, this expression is transformed into a local equation:

$$\int_{\omega_t} \left(\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{u}) \right) d\mathbf{X} = 0, \quad \forall t > t_0.$$

Even though the integral is written over a fluid element $(\omega_t)_t$, this is equivalent to considering all possible subdomains of Ω, since the initial domain ω_0 can be chosen arbitrarily and then evolved through the application φ_t. Since φ_t is assumed to be regular and bijective, the domains $\omega_t = \varphi_t(\omega_0)$ cover the entire accessible region of the fluid. Consequently, the vanishing of the integral for all ω_t implies that the integrand must be zero everywhere in Ω. This justifies the transition to the local form.
Thus, assuming that ρ and $\mathbf{u}$ are regular functions, we obtain the local form of the mass conservation equation:

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{u}) = 0$$

Moreover, since the fluid is incompressible, the density ρ is constant in space, and the equation simplifies to:

$$\operatorname{div}(\mathbf{u}) = 0. \tag{3.3}$$

Newton's second law

The modeling of momentum is based on Newton's second law, which states that the rate of change of momentum of a fluid element is equal to the sum of the external forces applied to it. These forces can be grouped into two main categories:

— Volumetric forces: These act within the interior of the fluid. They are represented by a force field $\mathbf{f}$, generally associated with gravity. The contribution of these

forces over a domain ω_t is given by:

$$\int_{\omega_t} \rho \, \mathbf{f} \, d\mathbf{X},$$

where ρ is the mass density.

— Contact forces: These result from interactions between the fluid and its environment on the boundary $\partial\omega_t$. They are modeled using a stress tensor $\underline{\underline{\Sigma}}$, which represents internal forces. This tensor, expressed in units of $\mathrm{kg} \cdot \mathrm{m}^{-1} \cdot \mathrm{s}^{-2}$, is decomposed as follows:

$$\underline{\underline{\Sigma}} = -P\mathbb{I}_d + \underline{\underline{T}},$$

where P is the pressure of the fluid,$\mathbb{I}_d$ is the identity tensor of dimension $d \times d$, and $\underline{\underline{T}}$ is an additional term that depends on the viscous properties of the fluid (see [15, Section 3.2]).

The contribution of the contact forces is then expressed as a surface integral:

$$\int_{\partial\omega_t} \underline{\underline{\Sigma}} : \mathbf{n} \, d\sigma,$$

where $\mathbf{n}$ is the outward normal vector to the boundary.

In the particular case of a Newtonian fluid, the viscous component $\underline{\underline{T}}$ is given by:

$$\underline{\underline{T}} = 2\mu\underline{\underline{D}} + \lambda \, \mathbf{grad} \cdot \mathbf{u} \, \mathbb{I}_d,$$

where μ is the dynamic viscosity, λ the bulk viscosity, and $\underline{\underline{D}}$ the rate-of-strain tensor, defined by:

$$\underline{\underline{D}} = \frac{1}{2}\left(\mathbf{Grad}\,\mathbf{u} + (\mathbf{Grad}\,\mathbf{u})^T\right). \tag{3.4}$$

We assume that the fields ρ and $\mathbf{u}$ belong respectively to the function spaces C^1 and C^0 with respect to the variables $(t, \mathbf{X}) \in \mathbb{R} \times \mathbb{R}^3$. Under this assumption, it can be shown that the function $(t, \mathbf{x}) \mapsto \underline{\underline{\Sigma}}(t, \mathbf{x}) \cdot \boldsymbol{a}$ is continuous for any vector $\boldsymbol{a} \in \mathbb{R}^3$ [15, Theorem I.3.1].

The stress tensor $\underline{\underline{\Sigma}}$, representing internal forces in the fluid, is given, in the case of an incompressible fluid, by:

$$\underline{\underline{\Sigma}} = -P\mathbb{I}_d + \mu\left(\mathbf{Grad}\,\mathbf{u} + (\mathbf{Grad}\,\mathbf{u})^T\right). \tag{3.5}$$

Applying Newton's second law to a fluid domain ω_t, we obtain:

$$\frac{d}{dt}\int_{\omega_t} \rho \mathbf{u}\, d\mathbf{X} = \int_{\omega_t} \rho \mathbf{f}\, d\mathbf{X} + \int_{\partial\omega_t} \underline{\underline{\Sigma}} : \mathbf{n}_{|\partial\omega_t}\, d\sigma.$$

The contact force term can be transformed into a volume integral using integration by parts:

$$\int_{\partial\omega_t} \underline{\underline{\Sigma}} : \mathbf{n}_{|\partial\omega_t}\, d\sigma = \int_{\omega_t} \mathrm{Div}\,(\underline{\underline{\Sigma}})\, d\mathbf{X}.$$

To treat the momentum term, we apply the transport theorem (3.1) to the function $f = \rho u_i$, for each $i \in \{1, \ldots, 3\}$, which yields:

$$\frac{d}{dt}\int_{\omega_t} \rho u_i\, d\mathbf{X} = \int_{\omega_t} \left(\frac{\partial(\rho u_i)}{\partial t} + \mathrm{div}(\rho u_i \mathbf{u}) \right) d\mathbf{X}.$$

Expanding the divergence of the convective term, we obtain:

$$\mathrm{div}(\rho u_i \mathbf{u}) = \sum_{j=1}^{d} \rho \frac{\partial(u_i u_j)}{\partial x_j} = \sum_{j=1}^{d} \rho \frac{\partial(\mathbf{u}\otimes\mathbf{u})_{i,j}}{\partial x_j},$$

where $\otimes$ denotes the tensor product defined by (3.2). Summing over all components u_i, we get:

$$\frac{d}{dt}\int_{\omega_t} \rho \mathbf{u}\, d\mathbf{X} = \int_{\omega_t} \left(\frac{\partial(\rho\mathbf{u})}{\partial t} + \mathrm{Div}\,(\rho \mathbf{u}\otimes\mathbf{u}) \right) d\mathbf{X}.$$

The conservation of momentum is then written as:

$$\int_{\omega_t} \left(\frac{\partial(\rho\mathbf{u})}{\partial t} + \mathrm{Div}\,(\rho \mathbf{u}\otimes\mathbf{u}) \right) d\mathbf{X} = \int_{\omega_t} \left(\rho \mathbf{f} + \mathrm{Div}\,(\underline{\underline{\Sigma}}) \right) d\mathbf{X}.$$

Using the expression of the stress tensor $\underline{\underline{\Sigma}}$ given in (3.5) for a Newtonian and incompressible fluid, we finally obtain:

$$\begin{aligned} &\int_{\omega_t} \left(\frac{\partial(\rho\mathbf{u})}{\partial t} + \mathrm{Div}\,(\rho \mathbf{u}\otimes\mathbf{u}) \right) d\mathbf{X} \\ &= \int_{\omega_t} \left(\rho \mathbf{f} + \mathrm{Div}\, \left(-P\mathbb{I}_d + \mu \left(\mathbf{Grad}\,\mathbf{u} + (\mathbf{Grad}\,\mathbf{u})^T \right) \right) \right) d\mathbf{X}. \end{aligned} \tag{3.6}$$

We can simplify the formulation (3.6) using the assumption that the fluid is incompressible.

Recall that $\underline{\underline{D}}$ is the deformation rate tensor defined in (3.4). For an incompressible fluid, we have, for all $i \in \{1, \ldots, d\}$,

$$\sum_{j=1}^{d} \frac{\partial \underline{\underline{D}}_{i,j}}{\partial x_j} = \frac{1}{2}\sum_{j=1}^{d}\left(\frac{\partial^2 u_i}{\partial x_j^2} + \frac{\partial^2 u_j}{\partial x_i \partial x_j}\right) = \frac{1}{2}\sum_{j=1}^{d}\frac{\partial^2 u_i}{\partial x_j^2} = \frac{1}{2}\Delta u_i.$$

The term $\sum_{j=1}^{d} \frac{\partial^2 u_j}{\partial x_i \partial x_j}$ vanishes due to the incompressibility condition $\mathrm{div}(\mathbf{u}) = 0$.
To summarize, we have shown that for an incompressible fluid:

$$\mathrm{Div}\left(\mathbf{Grad}\,\mathbf{u} + (\mathbf{Grad}\,\mathbf{u})^T\right) = \Delta\mathbf{u} \tag{3.7}$$

Moreover,

$$\begin{aligned}\mathrm{Div}\,(\rho\mathbf{u}\otimes\mathbf{u}) &= \rho\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial\,(u_i u_j)}{\partial x_j}\boldsymbol{e}_i \\ &= \rho\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial\,(u_i)\,u_j + u_i\partial\,(u_j)}{\partial x_j}\boldsymbol{e}_i\end{aligned}$$

Now since the fluid has zero divergence, we have the following equality:

$$\sum_{i=1}^{d}\sum_{j=1}^{d} u_i\frac{\partial\,(u_j)}{\partial x_j}\boldsymbol{e}_i = \mathbf{0}$$

and therefore:

$$\mathrm{Div}\,(\rho\mathbf{u}\otimes\mathbf{u}) = \rho\sum_{i=1}^{d}\sum_{j=1}^{d} u_j\frac{\partial\,(u_i)}{\partial x_j}\boldsymbol{e}_i = \rho\sum_{i=1}^{d}\sum_{j=1}^{d} u_j\frac{\partial(\mathbf{u})}{\partial x_j}$$

Which ultimately gives us:

$$\mathrm{Div}\,(\rho\mathbf{u}\otimes\mathbf{u}) = \rho(\mathbf{u}\cdot\mathbf{grad}\,)(\mathbf{u}) \tag{3.8}$$

Moreover, with an explicit calculation:

$$\mathrm{Div}\,(P\mathbb{I}_d) = \mathbf{grad}\,P. \tag{3.9}$$

Finally, using (3.7), (3.8), (3.9), and the expression (3.6), we obtain the equation resulting from the fundamental principle of dynamics for a Newtonian and incompressible fluid:

$$\rho\int_{\omega_t}\left(\frac{\partial\mathbf{u}}{\partial t} + (\mathbf{u}\cdot\mathbf{Grad}\,)\mathbf{u}\right)d\mathbf{X} = \int_{\omega_t}(\rho\,\mathbf{f} - \mathbf{grad}\,P + \mu\Delta\mathbf{u})d\mathbf{X}$$

Which gives in local form the equation of conservation of the quantity of movement for

an incompressible fluid:

$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\mathbf{Grad}\,)\mathbf{u}\right) - \mathbf{grad}\,P + \mu\Delta\mathbf{u} = \rho\,\mathbf{f} \quad \forall \mathbf{x} \in \Omega \tag{3.10}$$

3.2.2 Dimensionless variables: a physical reading

The incompressible Navier-Stokes equations express two fundamental physical principles: the conservation of mass and the conservation of momentum. They are written as:

$$\begin{cases} \rho\left(\dfrac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\mathbf{grad}\,)\mathbf{u}\right) - \mu\Delta\mathbf{u} + \mathbf{grad}\,P = \rho\,\mathbf{f} \quad \text{dans } \Omega\times\mathbb{R}^+ \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \operatorname{div}\mathbf{u} = 0 \quad \text{dans } \Omega\times\mathbb{R}^+ \end{cases} \tag{3.11}$$

Each term in the first equation has a physical meaning:

— $\rho\frac{\partial \mathbf{u}}{\partial t}$: the local acceleration of the fluid.

— $(\mathbf{u}\cdot\mathbf{grad}\,)\mathbf{u}$: the convective acceleration due to transport.

— $-\mu\Delta\mathbf{u}$: the internal viscous forces due to friction.

— $\mathbf{grad}\,P$: the pressure effects.

— $\rho\,\mathbf{f}$: external forces (e.g., gravity).

We aim to rewrite this system independently of physical units (meters, seconds, etc.) by scaling it to relative quantities. This allows us to identify dimensionless numbers that characterize the flow regime. To this end, we introduce the following characteristic scales:

— L: a typical length scale (e.g., the diameter of an obstacle).

— U: a characteristic velocity of the fluid.

— $T = \frac{L}{U}$: a typical time scale.

We then define the dimensionless variables:

$$\widetilde{\mathbf{x}} = \frac{\mathbf{x}}{L}, \quad \widetilde{t} = \frac{t}{T}, \quad \widetilde{\mathbf{u}} = \frac{\mathbf{u}}{U}, \quad \widetilde{p} = \frac{p}{\rho U^2}, \quad \widetilde{\mathbf{f}} = \frac{L}{U^2}\mathbf{f}$$

The differential operators are also redefined in the new variables:

$$\widetilde{\mathbf{grad}} = \mathbf{grad}_{\widetilde{x}}, \quad \widetilde{\operatorname{div}} = \operatorname{div}_{\widetilde{x}}, \quad \widetilde{\Delta} = \Delta_{\widetilde{x}}$$

By rewriting the system (3.11) in this framework and multiplying the momentum equation by $\frac{L}{\rho U^2}$, we obtain:

$$\begin{cases} \dfrac{\partial \widetilde{\mathbf{u}}}{\partial \widetilde{t}} + (\widetilde{\mathbf{u}} \cdot \widetilde{\mathbf{grad}})\widetilde{\mathbf{u}} - \dfrac{1}{Re}\widetilde{\Delta}\widetilde{\mathbf{u}} + \widetilde{\mathbf{grad}}\,\widetilde{p} = \widetilde{\mathbf{f}} & \text{dans } \Omega \times \mathbb{R}^+ \\ \widetilde{\mathrm{div}}\widetilde{\mathbf{u}} = 0 & \text{dans } \Omega \times \mathbb{R}^+ \end{cases} \tag{3.12}$$

where the Reynolds number is defined by:

$$Re := \frac{LU}{\nu}$$

with $\nu := \frac{\mu}{\rho}$ the kinematic viscosity of the fluid. The Reynolds number $Re = \dfrac{LU}{\nu}$ is a fundamental dimensionless parameter that characterizes the flow regime of a fluid. It represents the ratio between inertial forces (related to the acceleration of fluid particles) and viscous forces (due to internal friction in the fluid). Its physical interpretation is therefore essential to anticipate the nature of the motion:

— When $Re \ll 1$, viscous forces dominate: the fluid moves slowly, friction effects outweigh acceleration, and the flow is stable, smooth, and predictable. This is known as the Stokes regime or laminar flow.

— On the other hand, when $Re \gg 1$, inertia prevails: the fluid moves rapidly, friction is no longer sufficient to stabilize the motion, and instabilities may occur, leading to a turbulent regime.

The advantage of dimensionless variables is that it allows us to compare very different physical situations (such as pipe flow, atmospheric currents, or blood circulation) within a unified theoretical framework, as long as these systems share the same Reynolds number.
By expressing all physical quantities with dimensionless variables, we obtain the standard form of the Navier–Stokes equations for an incompressible fluid:

$$\begin{cases} \partial_t \mathbf{u} - \nu\Delta\mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f} \\ \operatorname{div} \mathbf{u} = 0 \end{cases} \tag{3.13}$$

where we denote $p = \frac{P}{\rho}$ as the reduced pressure. We then observe that in this system, the Reynolds number is inversely proportional to the viscosity: the lower the viscosity, the higher the Re, which promotes unstable flow regimes.
For the system of equations to be well-posed, it must be supplemented by initial

conditions (velocity values at time $t = 0$) and boundary conditions on the spatial domain boundary.
Two classical types of boundary conditions are commonly used depending on the physical context:

a) Dirichlet condition (imposed velocity on the boundary) This is the most frequent in practice, especially for solid walls or confined flows. It consists in prescribing the velocity on the domain boundary, often set to zero (no-slip condition), together with a normalization condition on the pressure:

$$\begin{cases} \mathbf{u} = \mathbf{0} & \text{on } \partial\Omega \times [0, +\infty) \\ \displaystyle\int_\Omega p \, d\mathbf{x} = 0 & \text{for all } t \geq 0 \end{cases}$$

The latter pressure condition removes the indeterminacy since pressure is defined up to an additive constant.

b) Neumann condition (imposed stress) In some cases, notably at the interface between two fluids or at a free outlet, one may choose to impose the fluid's stress on the boundary instead of the velocity. This amounts to specifying the stress flux on the boundary via the stress tensor $\underline{\underline{\Sigma}}$ defined previously. We then impose:

$$-\frac{1}{2}\,\nu\,(\mathbf{Grad} + \mathbf{Grad}^{\,T})\mathbf{u} : \mathbf{n} + P\mathbb{I}_d = \boldsymbol{g_N} \quad \text{on } \partial\Omega,$$

where $\boldsymbol{g_N}$ represents the external force exerted on the fluid by its environment (external pressure, surface tension, etc.), and $\mathbf{n}$ is the outward normal vector to the boundary. This type of condition is used, for example, to model a free outflow (without obstacles) or to couple several fluid domains.

Remark 3.4 *When boundary conditions other than pure Dirichlet conditions are considered, the use of the symmetric gradient is physically more appropriate. Nevertheless, for the sake of simplicity, in the numerical tests 14, we restrict ourselves to using the full gradient.*

Chapter 4

Spatial discretization notations

Contents

In this chapter, we define notations related to the discretization and we provide several inequalities that apply to polynomial functions, as well as estimates involving the L^2 orthogonal projection. Our aim is to explicit, as far as possible, constants in inequalities, such as in Propositions 4.2 and 4.4. In particular, we exhibit the dependence of these inequalities on the mesh regularity parameter σ, as shown in Propositions 4.5, 4.7, and 4.9.

4.1 Notations related to the mesh

We denote $(O, (x_{d'})_{d'=1}^d)$ the Cartesian coordinate system with an orthonormal basis $(\mathbf{e}_{d'})_{d'=1}^d$. We consider a sequence of meshes of simplices $(\mathcal{T}_h)_h$ in the domain Ω. For a fixed mesh $\mathcal{T}_h$, we use the following index sets and subsets:

— $\mathcal{I}_K$ is the index set of elements such that $\mathcal{T}_h := \bigcup_{\ell \in \mathcal{I}_K} K_\ell$.

— $\mathcal{I}_F$ is the index set of facets[1], such that $\mathcal{F}_h := \bigcup_{f \in \mathcal{I}_F} F_f$.

1. The term "facet" refers to a face (or edge) for $d = 3$ (or $d = 2$).

We let $\mathcal{I}_F = \mathcal{I}_F^i \cup \mathcal{I}_F^b$, where $\forall f \in \mathcal{I}_F^i$, $F_f \subset \Omega$ and $\forall f \in \mathcal{I}_F^b$, $F_f \subset \partial\Omega$.

— $\mathcal{I}_{F,\ell} = \{f \in \mathcal{I}_F \,|\, F_f \subset \partial K_\ell\}$ and $\mathcal{I}_{K,f} = \{\ell \in \mathcal{I}_K \,|\, F_f \subset \partial K_\ell\}$.

— $\mathcal{I}_K^b$ the set of indices of simplices having at least one boundary facet:

$$\mathcal{I}_K^b := \{\ell \in \mathcal{I}_K \,|\, \exists f \in \mathcal{I}_{F,\ell},\, f \in \mathcal{I}_F^b\};$$

— $\mathcal{I}_K^i$ the set of indices of simplices whose all facets lie in the interior of Ω:

$$\mathcal{I}_K^i := \{\ell \in \mathcal{I}_K \,|\, \forall f \in \mathcal{I}_{F,\ell},\, f \in \mathcal{I}_F^i\}.$$

For all $\ell \in \mathcal{I}_K$, we denote by h_ℓ the diameter of K_ℓ and by ρ_ℓ the diameter of its inscribed ball, and we define: $\sigma_\ell = \frac{h_\ell}{\rho_\ell}$. We suppose that the sequence $(\mathcal{T}_h)_h$ is of regular shape, cf. [49, def. 11.2]. Then, there exists a constant $\sigma > 1$, called the shape-regularity parameter, such that for all $\ell \in \mathcal{I}_K$, $\sigma_\ell \leq \sigma$. We denote by $\mathbf{n}_\ell$ the outward unit normal vector of K_ℓ. For $f \in \mathcal{I}_{F,\ell}$, we denote by $\mathbf{n}_{f,\ell}$ the outward unit normal vector of F_f from K_ℓ.

4.2 Discrete functional spaces

We introduce the spaces of piecewise regular functions as follows:

— The broken space $\mathcal{P}_h H^1 = \left\{v \in L^2(\Omega)\,;\quad \forall \ell \in \mathcal{I}_K,\, v_{|K_\ell} \in H^1(K_\ell)\right\}$, equipped with the inner product:

$$(v,w)_h := \sum_{\ell\in\mathcal{I}_K} (\mathbf{grad}\, v, \mathbf{grad}\, w)_{\mathbf{L}^2(K_\ell)},\ \|v\|_h^2 = \sum_{\ell\in\mathcal{I}_K} \|\mathbf{grad}\, v\|^2_{\mathbf{L}^2(K_\ell)}.$$

— The broken space $\mathcal{P}_h\mathbf{H}^1 = [\mathcal{P}_h H^1]^d$, equipped with the inner product: $\forall \mathbf{v}, \mathbf{w} \in \mathcal{P}_h\mathbf{H}^1$

$$(\mathbf{v},\mathbf{w})_h := \sum_{\ell\in\mathcal{I}_K} (\mathbf{Grad}\, \mathbf{v}, \mathbf{Grad}\, \mathbf{w})_{\mathbb{L}^2(K_\ell)},\ \|\mathbf{v}\|_h^2 = \sum_{\ell\in\mathcal{I}_K} \|\mathbf{Grad}\, \mathbf{v}\|^2_{\mathbb{L}^2(K_\ell)}.$$

We define the related piecewise operators as follows:

— The operator $\mathbf{grad}_h : \mathcal{P}_h H^1 \to \mathbf{L}^2(\Omega)$ such that:

$$\forall v \in \mathcal{P}_h H^1,\, \forall \mathbf{w} \in \mathbf{L}^2(\Omega),\ (\mathbf{grad}_h v, \mathbf{w})_{\mathbf{L}^2(\Omega)} = \sum_{\ell\in\mathcal{I}_K} (\mathbf{grad}\, v, \mathbf{w})_{\mathbf{L}^2(K_\ell)}.$$

— The operator $\mathbf{Grad}_h : \mathcal{P}_h\mathbf{H}^1 \to \mathbb{L}^2(\Omega)$ such that:

$$\forall \mathbf{v} \in \mathcal{P}_h\mathbf{H}^1,\ \forall \mathbf{w} \in \mathbf{L}^2(\Omega), \quad (\mathbf{Grad}_h \mathbf{v}, \mathbf{w})_{\mathbb{L}^2(\Omega)} = \sum_{\ell \in \mathcal{I}_K} (\mathbf{Grad}\, \mathbf{v}, \mathbf{w})_{\mathbb{L}^2(K_\ell)}.$$

For any $f \in \mathcal{I}_F$, $\mathbf{n}_f$ denotes the unit normal vector to F_f chosen as follows: if $F_f \subset \partial\Omega$, then $\mathbf{n}_f$ is oriented outward with respect to Ω; if $f \in \mathcal{I}_F^i$ with $F_f = \partial K_L \cap \partial K_R$ and $L < R$, then $\mathbf{n}_f$ is oriented outward from K_L. The jump and the average of a function $v \in \mathcal{P}_h H^1$ across a facet F_f are defined as follows:

$$[\![v]\!]_{F_f} := v_{|K_L} - v_{|K_R} \text{ and } \{v\}_{F_f} := \frac{1}{2}(v_{|K_L} + v_{|K_R}).$$

For $f \in \mathcal{I}_F^b$, we set: $[\![v]\!]_{F_f} := v_{|F_f}$ and $\{v\}_{F_f} := v_{|F_f}$. Let $f \in \mathcal{I}_F^i$ and let F_f be an interior face shared by the elements K_L and K_R. Let $a, b \in \mathcal{P}_h H^1$. We denote by $a_L := a_{|K_L}$, $a_R := a_{|K_R}$, $b_L := b_{|K_L}$, and $b_R := b_{|K_R}$ the traces of a and b on F_f. Then

$$\begin{aligned} [\![a\, b]\!] &= a_L b_L - a_R b_R \\ &= \tfrac{1}{2}(a_L + a_R)(b_L - b_R) + \tfrac{1}{2}(b_L + b_R)(a_L - a_R) \\ &= \{a\}\, [\![b]\!] + \{b\}\, [\![a]\!]. \end{aligned} \tag{4.1}$$

We will need the broken spaces related to the divergence operator:

$$\mathcal{P}_h\mathbf{H}(\mathrm{div}) = \left\{ \mathbf{v} \in \mathbf{L}^2(\Omega)\,; \quad \forall \ell \in \mathcal{I}_K,\ \mathbf{v}_{|K_\ell} \in \mathbf{H}(\mathrm{div}; K_\ell) \right\},$$

$$\mathcal{P}_h\mathbb{H}(\mathrm{Div}) = \left\{ \underline{\underline{G}} \in \mathbb{L}^2(\Omega)\,; \quad \forall \ell \in \mathcal{I}_K,\ \underline{\underline{G}}_{|K_\ell} \in \mathbb{H}(\mathrm{Div}\,; K_\ell) \right\}.$$

We define the related broken operators div_h and Div_h such that:

$$\forall \mathbf{v} \in \mathcal{P}_h\mathbf{H}(\mathrm{div}),\ \forall q \in L^2(\Omega), \quad (\mathrm{div}_h\, \mathbf{v}, q)_{L^2(\Omega)} = \sum_{\ell \in \mathcal{I}_K} (\mathrm{div}\, \mathbf{v}, q)_{L^2(K_\ell)},$$

$$\forall \underline{\underline{G}} \in \mathcal{P}_h\mathbb{H}(\mathrm{Div}),\ \forall \mathbf{q} \in \mathbf{L}^2(\Omega), \quad (\mathrm{Div}_h\, \underline{\underline{G}}, \mathbf{q})_{\mathbf{L}^2(\Omega)} = \sum_{\ell \in \mathcal{I}_K} (\mathrm{Div}\, \underline{\underline{G}}, \mathbf{q})_{\mathbf{L}^2(K_\ell)}.$$

4.3 Geometric Mappings

We now recall classical finite element estimates [49, Chap. 9, 11]. Let $\hat{K}$ be the reference simplex, of vertices $(\hat{S}_i)_{i=0}^d$, which coordinates are such that $\hat{S}_0 = O$ and $\overrightarrow{O\hat{S}_i} = \mathbf{e}_i$.

We have: $|\hat{K}| = (d!)^{-1}$. For $\ell \in \mathcal{I}_K$, we denote by $T_\ell : \hat{K} \to K_\ell$ the geometric affine transformation such that $\forall \hat{\mathbf{x}} \in \hat{K}$, $\mathbf{x}_{|K_\ell} = T_\ell(\hat{\mathbf{x}}) = \mathbb{B}_\ell \hat{\mathbf{x}} + \mathbf{b}_\ell$, and we denote by J_ℓ the Jacobian of T_ℓ. Recall that $|J_\ell| = \det(\mathbb{B}_\ell)$. It holds [49, Eq. (11.3)]:

$$|J_\ell| = \frac{|K_\ell|}{|\hat{K}|}, \quad \|\mathbb{B}_\ell\| \leq \frac{h_\ell}{\rho_{\hat{K}}}, \quad \|\mathbb{B}_\ell{}^{-1}\| \leq \frac{h_{\hat{K}}}{\rho_\ell}. \tag{4.2}$$

In $2D$, classically, the reference triangle $\hat{K}$ is built with vertices $(0,0)$, $(1,0)$, and $(0,1)$. Let K_ℓ, $\ell \in \mathcal{I}_K$ be a triangle with vertices A_1^ℓ, A_2^ℓ, and A_3^ℓ, where $A_i^\ell = (x_i, y_i)$ for $i = 1, 2, 3$. Consider the affine transformation $T_\ell : \hat{K} \to K_\ell$, such that $T_\ell(\hat{K}) = K_\ell$. It reads:

$$T_\ell(x,y) = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + x \begin{bmatrix} a \\ c \end{bmatrix} + y \begin{bmatrix} b \\ d \end{bmatrix}.$$

Assuming the following:

— $T_\ell(0,0) = A_1^\ell \quad \Rightarrow \quad \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = A_1^\ell,$

— $T_\ell(1,0) = A_2^\ell \quad \Rightarrow \quad A_1^\ell + \begin{bmatrix} a \\ c \end{bmatrix} = A_2^\ell,$

— $T_\ell(0,1) = A_3^\ell \quad \Rightarrow \quad A_1^\ell + \begin{bmatrix} b \\ d \end{bmatrix} = A_3^\ell,$

we obtain:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix}.$$

Thus, the affine transformation T_ℓ can be rewritten as:

$$T_\ell(x,y) = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

Since T_ℓ is affine, the Jacobian $J_\ell = 2|K_\ell|$ is constant over $\hat{K}$, and $T_\ell(\hat{K}) = K_\ell$.
Let $f \in \mathcal{I}_{F,\ell}$. Let $T_{f,\ell} = T_{\ell|F_f}$, and $\hat{n}_{f,\ell}$ be the outward normal of $\hat{F}_{f,\ell}$ such that $T_{f,\ell}(\hat{F}_{f,\ell}) = F_{f,\ell}$. For $v \in L^2(K_\ell)$, we set:

$$\hat{v}_\ell = v \circ T_\ell, \quad \forall f \in \mathcal{I}_{F,\ell},\ \hat{v}_{f,\ell} = v \circ T_{f,\ell}. \tag{4.3}$$

We also need to define the local averages:

$$\forall v \in L^1(K_\ell),\ \underline{v}_\ell = \int_{K_\ell} v/|K_\ell|. \tag{4.4}$$

Proposition 4.1 (`Change of variables` [72]) *Changing the variables:*

$$\|v\|^2_{L^2(K_\ell)} = |K_\ell|\,|\hat{K}|^{-1}\,\|\hat{v}_\ell\|^2_{L^2(\hat{K})}, \tag{4.5}$$
$$\|v\|^2_{L^2(F_f)} = |F_f|\,|\hat{F}_{f,\ell}|^{-1}\,\|\hat{v}_{f,\ell}\|^2_{L^2(\hat{F}_{f,\ell})}. \tag{4.6}$$

Let $v \in \mathcal{P}_h H^1$. Changing the variables, $\mathbf{grad}\, v_{|K_\ell} = (\mathbb{B}_\ell{}^{-1})^T \mathbf{grad}_{\hat{\mathbf{x}}} \hat{v}_\ell$, and we have:

$$\begin{array}{lll} \text{(i)} & \|\mathbf{grad}_{\hat{\mathbf{x}}} \hat{v}_\ell\|^2_{\mathbf{L}^2(\hat{K})} \leq & \|\mathbb{B}_\ell\|^2\,|J_\ell|^{-1}\,\|\mathbf{grad}\, v\|^2_{\mathbf{L}^2(K_\ell)} \\ \text{(ii)} & \|\mathbf{grad}\, v\|^2_{\mathbf{L}^2(K_\ell)} \leq & \left\|\mathbb{B}_\ell{}^{-1}\right\|^2\,|J_\ell|\,\|\mathbf{grad}_{\hat{\mathbf{x}}} \hat{v}_\ell\|^2_{\mathbf{L}^2(\hat{K})} \end{array}. \tag{4.7}$$

Let $v \in H^s(K_\ell)$. Changing the variables, it holds [24, §3]:

$$\begin{array}{lll} \text{(i)} & |\hat{v}|^2_{H^s(\hat{K})} \leq & \|\mathbb{B}_\ell\|^{d+2s}\,|J_\ell|^{-2}\,|v|^2_{H^s(K_\ell)} \\ \text{(ii)} & |v|^2_{H^s(K_\ell)} \leq & \|\mathbb{B}_\ell{}^{-1}\|^{d+2s}\,|J_\ell|^2\,|\hat{v}|^2_{H^s(\hat{K})} \end{array}. \tag{4.8}$$

For all $D \subset \mathbb{R}^d$, we denote by $P^k(D)$ the set of polynomials of degree k on D, $\mathbf{P}^k(D) = (P^k(D))^d$, and we consider the space of broken polynomials:

$$P^k_{disc}(\mathcal{T}_h) = \left\{ q \in L^2(\Omega); \quad \forall \ell \in \mathcal{I}_K,\, q_{|K_\ell} \in P^k(K_\ell) \right\}.$$

Table 4.1 – Dimension of the polynomial space $P^k(D)$ for $d \in \{1,2,3\}$ and $k \in \{0,1,2,3\}$

k	$d=1$	$d=2$	$d=3$
0	1	1	1
1	2	3	4
2	3	6	10
3	4	10	20

.

We need a result appearing in [77, p. 565].

Theorem 4.1 *For all $\ell \in \mathcal{I}_K$, we have:*

$$|\partial K_\ell|\,|K_\ell|^{-1} = 2\,d\,(\rho_\ell)^{-1}. \tag{4.9}$$

Proof. Let C_ℓ be the center of the inscribed ball in K_ℓ, of diameter ρ_ℓ. We set $\mathbf{c}_\ell = \overrightarrow{OC}_\ell$. By definition, the closed ball $\overline{B}(\mathbf{c}_\ell, \rho_\ell/2)$ is tangent to *all* faces, so the

distance from the center C_ℓ to F_f is exactly $\rho_\ell/2$:

$$\forall \mathbf{x}_f \in F_f, \qquad (\mathbf{x}_f - \mathbf{c}_\ell) \cdot \mathbf{n}_{f,\ell} = \rho_\ell/2. \tag{4.10}$$

Consider the vector field $X_\ell : \mathbb{R}^d \to \mathbb{R}^d$, defined by $X_\ell(\mathbf{x}) = \mathbf{x} - \mathbf{c}_\ell$. Then $\operatorname{div} X_\ell = d$. Integrating by parts, we get:

$$\int_{K_\ell} \operatorname{div} X_\ell \, d\mathbf{x} = \int_{\partial K_\ell} X_\ell \cdot \mathbf{n}_{|\partial K_\ell} \, ds \Rightarrow d\,|K_\ell| = \int_{\partial K_\ell} (\mathbf{x} - \mathbf{c}_\ell) \cdot \mathbf{n}_{|\partial K_\ell} \, ds.$$

Summing over all faces and using (4.10),

$$\begin{aligned} \int_{\partial K_\ell} (\mathbf{x} - \mathbf{c}_\ell) \cdot \mathbf{n}_{|\partial K_\ell} \, ds &= \sum_{f \in \mathcal{I}_{F,\ell}} \int_{F_f} (\mathbf{x} - \mathbf{c}_\ell) \cdot \mathbf{n}_{f,\ell} \, ds, \\ &= \sum_{f \in \mathcal{I}_{F,\ell}} \frac{\rho_\ell}{2} \, |F_f| = \frac{\rho_\ell}{2} \, |\partial K_\ell|. \end{aligned}$$

We conclude that $d\,|K_\ell| = \dfrac{\rho_\ell}{2}\,|\partial K_\ell|$, hence (4.9).

□

We recall the following theorem.

Theorem 4.2 ([99, Thm. 5]) *Let $v \in P^k(K_\ell)$. The following inverse inequality holds:*

$$\|v\|^2_{L^2(\partial K_\ell)} \leq C_{d,k}\,|\partial K_\ell|\,|K_\ell|^{-1}\,\|v\|^2_{L^2(K_\ell)}, \text{ with } C_{d,k} = \frac{(k+1)(k+d)}{d}. \tag{4.11}$$

We deduce from inequality (4.11) the following result

Proposition 4.2 *For all $\ell \in \mathcal{I}_K$, for all $v \in P^k(K_\ell)$, we have:*

$$\|v\|^2_{L^2(\partial K_\ell)} \leq \tilde{C}_{d,k}\,\sigma_\ell\,(h_\ell)^{-1}\,\|v\|^2_{L^2(K_\ell)}, \tag{4.12}$$

with $\tilde{C}_{d,k} = 2\,(k+1)\,(k+d)$.

Proof. Let $\ell \in \mathcal{I}_K$ and $v \in P^k(K_\ell)$. From inequality (4.11), we have

$$\|v\|^2_{L^2(\partial K_\ell)} \leq C_{d,k}\,|\partial K_\ell|\,|K_\ell|^{-1}\,\|v\|^2_{L^2(K_\ell)}. \tag{4.13}$$

Using Theorem 4.1, we obtain: $|\partial K_\ell|\,|K_\ell|^{-1} = 2\,d\,(\rho_\ell)^{-1} = 2\,d\,\sigma_\ell\,(h_\ell)^{-1}$.

Finally, we have $\|v\|^2_{L^2(\partial K_\ell)} \leq (2\, C_{d,k}\, d)\, \sigma_\ell\, (h_\ell)^{-1}\, \|v\|^2_{L^2(K_\ell)}$.

□

Using [49, Lemma 12.8], we can generalize Proposition 4.2 to the Lebesgue space $L^p(\Omega)$. The proof follows the same methodology as that of Proposition 4.2.

Corollary 4.3 *For any $p \geq 1$, there exists a positive constant $C_{d,k,p}$ such that, for all $\ell \in \mathcal{I}_K$ and for all $v \in P^k(K_\ell)$, we have:*

$$\|v\|_{L^p(\partial K_\ell)} \leq C_{d,k,p}\, \sigma_\ell^{1/p}\, (h_\ell)^{-1/p}\, \|v\|_{L^p(K_\ell)}. \tag{4.14}$$

In order to state a bound similar to (4.12) when K_ℓ is not a simplex, we consider the bound on a single face of the boundary of K_ℓ. Thus, in the case of a general mesh, we will split the elements into simplices and apply Proposition 4.4 below.

Proposition 4.4 *There exists a constant $\hat{C}_{d,k} > 0$ such that for all $f \in \mathcal{I}_F$, for all $\ell \in \mathcal{I}_{K,f}$, for all $v_\ell \in P^k(K_\ell)$, we have:*

$$\|v_\ell\|^2_{L^2(F_f)} \leq \hat{C}_{d,k}\, \sigma_\ell\, (h_\ell)^{-1}\, \|v_\ell\|^2_{L^2(K_\ell)}. \tag{4.15}$$

Proof. Let $N_{d,k} = \dim(P^k(\hat{K}))$. Consider $(\phi_i)_{i=1}^{N_{d,k}}$, a basis of $P^k(\hat{K})$. Let $\mathbb{M}_{\hat{K}} \in \mathbb{R}^{N_{d,k}} \times \mathbb{R}^{N_{d,k}}$ be the matrix such that for all $i, j \in \{1, \ldots, N_{d,k}\}$, $(\mathbb{M}_{\hat{K}})_{i,j} = (\hat{\phi}_i, \hat{\phi}_j)_{L^2(\hat{K})}$. For $\hat{F} \in \partial\hat{K}$, we denote by $(\mathbb{M}_{\hat{F}})_{i,j} \in \mathbb{R}^{N_{d,k}} \times \mathbb{R}^{N_{d,k}}$ the matrix such that for all $i, j \in \{1, \ldots, N_{d,k}\}$, $\mathbb{M}_{\hat{F}} = (\hat{\phi}_i, \hat{\phi}_j)_{L^2(\hat{F})}$. Let $f \in \mathcal{I}_F$, $\ell \in \mathcal{I}_{K,f}$, and $v_\ell \in P^k(K_\ell)$. Let $\hat{V} \in \mathbb{R}^{N_{d,k}}$ be the vector such that $v_\ell(\mathbf{x}) = \sum_{i=1}^{N_{d,k}} \hat{V}_i \hat{\phi}_i(\mathbf{x})$. Hence:

$$\|\hat{v}_{f,\ell}\|^2_{L^2(\hat{F}_{f,\ell})} = (\mathbb{M}_{\hat{F}_{f,\ell}}\, \hat{V} \,|\, \hat{V}) \text{ and } \|\hat{v}_\ell\|^2_{L^2(\hat{K})} = (\mathbb{M}_{\hat{K}}\, \hat{V} \,|\, \hat{V}).$$

We look for $\mu \in \mathbb{R}$ such that $\|\hat{v}_{f,\ell}\|^2_{L^2(\hat{F}_{f,l})} \leq \mu \|\hat{v}_\ell\|^2_{L^2(\hat{K})}$. Let $\hat{F} \in \partial\hat{K}_\ell$ and consider the generalised eigenproblem:

$$\begin{cases} \text{Find } (\hat{V}_{\lambda_{\hat{F}}}, \lambda_{\hat{F}}) \in \mathbb{R}^{N_k} \backslash \{(0,0)\} \text{ such that } (\hat{V}_{\lambda_{\hat{F}}}, \lambda_{\hat{F}}) \text{ solves} \\ \quad \mathbb{M}_{\hat{K}}\, \hat{V}_\lambda = \lambda\, \mathbb{M}_{\hat{F}}\, \hat{V}_\lambda \text{ with } \lambda_{\hat{F}} = \min_\lambda(\lambda) \text{ and } \|\hat{V}_{\lambda_{\hat{F}}}\| = 1. \end{cases}$$

Hence: $\|\hat{v}_{f,\ell}\|^2_{L^2(\hat{F}_{f,l})} \leq \max_{\hat{F} \in \partial\hat{K}} (\lambda_{\hat{F}}^{-1}) \|\hat{v}_\ell\|^2_{L^2(\hat{K})}$. Changing the variables, cf. equations

(4.5)-(4.6) we deduce that:

$$\|v_{f,\ell}\|^2_{L^2(F_{f,\ell})} \leq \max_{\hat{F}\in\partial\hat{K}}(\lambda^{-1}_{\hat{F}})\,(\min_{\hat{F}\in\partial\hat{K}}|\hat{F}|)^{-1}\,|\hat{K}|\,|F_{f,\ell}|\,|K_\ell|^{-1}\,\|v_\ell\|^2_{L^2(K_\ell)}.$$

Using Theorem 4.1, we get: $|F_{f,\ell}|\ |K_\ell|^{-1} \leq |\partial K_\ell|\,|K_\ell|^{-1} = 2\,d\,(\rho_\ell)^{-1} \leq 2\,d\,\sigma_\ell\,(h_\ell)^{-1}$. Finally, we obtain (4.15), with

$$\hat{C}_{d,k} = 2\,d\ \max_{\hat{F}\in\partial\hat{K}}(\lambda^{-1}_{\hat{F}})\,(\min_{\hat{F}\in\partial\hat{K}}|\hat{F}|)^{-1}\,|\hat{K}|.$$

The eigenvalues $\lambda_{\hat{F}}$ depend on the d and the order k and can be computed explicitly. □

Proposition 4.5 *For all $\ell \in \mathcal{I}_K$, for all $v_\ell \in H^s(K_\ell)$ with $s \in (\frac{1}{2}, 1)$, the following estimate holds:*

$$\|v_\ell\|_{L^2(\partial K_\ell)} \lesssim \sigma_\ell^{\frac{1}{2}}\,(h_\ell)^{-\frac{1}{2}}\left(\|v_\ell\|^2_{L^2(K_\ell)} + (\sigma_\ell)^d\,(h_\ell)^{2s}|v_\ell|^2_{H^s(K_\ell)}\right)^{\frac{1}{2}}.$$

Proof. Let $\ell \in \mathcal{I}_K$ and $v_\ell \in H^s(K_\ell)$. We first use (4.6) and then the trace inequality on the reference element (see Equation (2.9)) to obtain

$$\|v_\ell\|^2_{L^2(\partial K_\ell)} \lesssim |\partial K_\ell|\,\|\hat{v}_\ell\|^2_{L^2(\partial\hat{K})} \leq |\partial K_\ell|\,\|\hat{v}_\ell\|^2_{H^{s-\frac{1}{2}}(\partial\hat{K})} \lesssim |\partial K_\ell|\,\|\hat{v}_\ell\|^2_{H^s(\hat{K})}.$$

Then, by definition (2.6) of the H^s norm, changing the variables back to element K_ℓ and using (4.5), (4.8)-(i) and (4.2), we get

$$\|v_\ell\|^2_{L^2(\partial K_\ell)} \lesssim |\partial K_\ell|\,|K_\ell|^{-1}\left(\|v_\ell\|^2_{L^2(K_\ell)} + |K_\ell|^{-1}\,(h_\ell)^{2s+d}|v_\ell|^2_{H^s(K_\ell)}\right),$$

Using then (4.9), the definition of σ_ℓ and the fact that $|K_\ell| \gtrsim \rho_\ell^d$, we obtain

$$\|v_\ell\|^2_{L^2(\partial K_\ell)} \lesssim \sigma_\ell\,(h_\ell)^{-1}\left(\|v_\ell\|^2_{L^2(K_\ell)} + (\sigma_\ell)^d\,(h_\ell)^{2s}|v_\ell|^2_{H^s(K_\ell)}\right),$$

which leads to the result, taking into account the definition (2.6) of the norm in H^s. □

Corollary 4.6 *According to the above proposition 4.5, we obtain*

$$\|v_\ell\|_{L^2(\partial K_\ell)} \lesssim (\sigma_\ell)^{\frac{d+1}{2}} (h_\ell)^{-\frac{1}{2}} \|v\|_{H^s(K_\ell)}. \tag{4.16}$$

4.4 Broken L^2-orthogonal projection

For all $\ell \in \mathcal{I}_K$, we denote by $\pi_\ell^k \in \mathcal{L}(L^2(K_\ell), P^k(K_\ell))$ the L^2-orthogonal projection onto $P^k(K_\ell)$:

$$\forall v \in L^2(K_\ell), \forall v_h \in P^k(K_\ell), \quad (\pi_\ell^k v, v_h)_{L^2(K_\ell)} = (v, v_h)_{L^2(K_\ell)}. \tag{4.17}$$

We now define $\pi_h^k \in \mathcal{L}(L^2(\Omega), P_{disc}^k(\mathcal{T}_h))$ the L^2-orthogonal projection onto $P_{disc}^k(\mathcal{T}_h)$, such that:

$$\forall v \in L^2(\Omega), \forall v_h \in P_{disc}^k(\mathcal{T}_h), \quad (\pi_h^k v, v_h)_{L^2(\Omega)} := \sum_{\ell \in \mathcal{I}_K} (\pi_\ell^k v, v_h)_{L^2(K_\ell)}. \tag{4.18}$$

Since $\underline{v}_\ell \in P^k(K_\ell)$ (cf. (4.4)) and π_ℓ^k minimizes the distance to v in $P^k(K_\ell)$, we have

$$\forall \ell \in \mathcal{I}_K, \forall v \in L^2(K_\ell), \|v - \pi_\ell^k v\|_{L^2(K_\ell)} \leq \|v - \underline{v}_\ell\|_{L^2(K_\ell)}. \tag{4.19}$$

We denote by $\pi_{\hat{K}}^k$ the L^2-orthogonal projection onto $P^k(\hat{K})$. It holds:

$$\forall \ell \in \mathcal{I}_K, \forall v \in L^2(K_\ell), \widehat{\pi_\ell v} = \pi_{\hat{K}}^k \hat{v}_\ell. \tag{4.20}$$

Finally, using a change of variable [72], we notice that $\int_{\hat{K}} \hat{v}_\ell / |\hat{K}| = \underline{v}_\ell$.
The following Proposition can be found in [40, Lemmas 1.58 and 1.59]. Below, we explicit the dependence with respect to the mesh regularity.

Proposition 4.7 *For all $\ell \in \mathcal{I}_K$, for all $v \in H^1(K_\ell)$, and for all $f \in \mathcal{I}_{F,\ell}$, it holds:*

$$\begin{array}{lll} \text{(i)} & \|v - \pi_\ell^k v\|_{L^2(K_\ell)} & \lesssim h_\ell \, |v|_{H^1(K_\ell)} \\ \text{(ii)} & \|v - \pi_\ell^k v\|_{L^2(F_f)} & \lesssim (\sigma_\ell)^{\frac{1}{2}} (h_\ell)^{\frac{1}{2}} \, |v|_{H^1(K_\ell)} \end{array}. \tag{4.21}$$

Proof. Let $\ell \in \mathcal{I}_K$ and $v \in H^1(K_\ell)$. Let us prove inequality (4.21)-(i). Using (4.5), we have

$$\| v_\ell - \pi_\ell^k v_\ell\|_{L^2(K_\ell)}^2 = |\hat{K}|^{-1} \, |K_\ell| \, \|\hat{v}_\ell - \pi_{\hat{K}}^k \hat{v}_\ell\|_{L^2(\hat{K})}^2. \tag{4.22}$$

We obtain, using (4.19) and the Poincaré-Steklov inequality (2.9),

$$\|\hat{v}_\ell - \pi^k_{\hat{K}}\hat{v}_\ell\|^2_{L^2(\hat{K})} \leq \|\hat{v}_\ell - \underline{\hat{v}}_\ell\|^2_{L^2(\hat{K})} \leq \pi^{-1}\, h_{\hat{K}}\, |\hat{v}_\ell|^2_{H^1(\hat{K})}. \tag{4.23}$$

Returning to the element K_ℓ, using (4.7)-(i) and (4.2), we have

$$|\hat{v}_\ell|^2_{H^1(\hat{K})} \leq (h_\ell)^2\, |K_\ell|^{-1}\, |\hat{K}|\, (\rho_{\hat{K}})^{-2}\, |v|^2_{H^1(K_\ell)}. \tag{4.24}$$

Combining (4.22), (4.23), and (4.24) yields (4.21)-(i).

Let us prove inequality (4.21)-(ii). Using (4.6), we obtain

$$\|v - \pi^k_\ell v\|^2_{L^2(F_f)} \leq |F_f|\, |\hat{F}_{f,\ell}|^{-1}\, \|\hat{v}_\ell - \pi^k_{\hat{K}}\hat{v}_\ell\|^2_{L^2(\hat{F}_{f,\ell})}. \tag{4.25}$$

By elementary inequality,

$$\|\hat{v}_\ell - \pi^k_{\hat{K}}\hat{v}_\ell\|^2_{L^2(\hat{F}_{f,\ell})} \leq 2\left(\|\hat{v}_\ell - \underline{\hat{v}}_\ell\|^2_{L^2(\hat{F}_{f,\ell})} + \|\underline{\hat{v}}_\ell - \pi^k_{\hat{K}}\hat{v}_\ell\|^2_{L^2(\hat{F}_{f,\ell})}\right). \tag{4.26}$$

Let us bound the first term in the right-hand side of (4.26). Using Theorem 2.7 and the Poincaré-Steklov inequality (2.9), we obtain

$$\|\hat{v}_\ell - \underline{\hat{v}}_\ell\|^2_{L^2(\hat{F}_{f,\ell})} \lesssim \|\hat{v}_\ell - \underline{\hat{v}}_\ell\|^2_{H^1(\hat{K})} \lesssim |\hat{v}_\ell|^2_{H^1(\hat{K})}. \tag{4.27}$$

Let us bound the second term in the right-hand side of (4.26). Since $\underline{\hat{v}}_\ell - \pi^k_{\hat{K}}\hat{v}_\ell$ is a polynomial, Proposition 4.4 yields

$$\|\underline{\hat{v}}_\ell - \pi^k_{\hat{K}}\hat{v}_\ell\|^2_{L^2(\hat{F}_{f,\ell})} \lesssim \|\underline{\hat{v}}_\ell - \pi^k_{\hat{K}}\hat{v}_\ell\|^2_{L^2(\hat{K})}. \tag{4.28}$$

By elementary inequality, and the Poincaré-Steklov inequality (2.9), we have:

$$\begin{aligned} \|\underline{\hat{v}}_\ell - \pi^k_{\hat{K}}\hat{v}_\ell\|^2_{L^2(\hat{K})} &\leq 2\left(\|\hat{v}_\ell - \underline{\hat{v}}_\ell\|^2_{L^2(\hat{K})} + \|\hat{v}_\ell - \pi^k_{\hat{K}}\hat{v}_\ell\|^2_{L^2(\hat{K})}\right), \\ &\leq 4\,\|\hat{v}_\ell - \underline{\hat{v}}_\ell\|^2_{L^2(\hat{K})}, \\ &\lesssim |\hat{v}_\ell|^2_{H^1(\hat{K})}. \end{aligned} \tag{4.29}$$

Collecting (4.28) and (4.29), we have:

$$\|\underline{\hat{v}}_\ell - \pi^k_{\hat{K}}\hat{v}_\ell\|^2_{L^2(\hat{F}_{f,\ell})} \lesssim |\hat{v}_\ell|^2_{H^1(\hat{K})}. \tag{4.30}$$

Combining (4.27) and (4.30) allows to bound (4.26):

$$\|\hat{v}_\ell - \pi_{\hat{K}}^k \hat{v}_\ell\|_{L^2(\hat{F}_{f,\ell})}^2 \lesssim |\hat{v}_\ell|_{H^1(\hat{K})}^2.$$

Substituting this estimate into (4.25) and using (4.7)-(i), we find

$$\|v - \pi_\ell^k v\|_{L^2(F_f)}^2 \lesssim |F_f|\,|K_\ell|^{-1}\,(h_\ell)^2\,|v|_{H^1(K_\ell)}^2.$$

Finally, using equation (4.9), it holds: $|F_f|\,|K_\ell|^{-1} \lesssim (\rho_\ell)^{-1} \lesssim \sigma_\ell (h_\ell)^{-1}$. Hence we obtain (4.21)-(ii).

□

We can obtain similar estimates for $v \in H^s(\Omega)$ with $s \in (\frac{1}{2}, 1)$. Notice that the constants in the bounds are multiplied by $\sigma^{\frac{d}{2}}$.

Proposition 4.8 *For all $s \in (\frac{1}{2}, 1)$, for all $\ell \in \mathcal{I}_K$, for all $v \in H^s(K_\ell)$, and for all $f \in \mathcal{I}_{F,\ell}$, it holds:*

$$\begin{array}{lrcl} \text{(i)} & \|v - \pi_\ell^k v\|_{L^2(K_\ell)} & \lesssim & (\sigma_\ell)^{\frac{d}{2}}\,(h_\ell)^s |v|_{H^s(K_\ell)} \\ \text{(ii)} & \|v - \pi_\ell^k v\|_{L^2(F_f)} & \lesssim & (\sigma_\ell)^{\frac{d+1}{2}}\,(h_\ell)^{s-\frac{1}{2}}\,|v|_{H^s(K_\ell)} \end{array}, \tag{4.31}$$

where the hidden constants depends only on s and k.

Proof. Let us prove inequality (4.31)-(i). Let $v \in H^s(K_\ell)$. Using (4.19), Poincaré-Steklov inequality (2.7), and $|K_\ell|^{-1} \lesssim \rho_\ell^{-d}$ it follows that

$$\|v - \pi_\ell^k v\|_{L^2(K_\ell)}^2 \;\le\; h_\ell^{2s} \left(\frac{h_\ell^d}{|K_\ell|}\right) |v|_{H^s(K_\ell)}^2 \;\lesssim\; \sigma_\ell^d\,(h_\ell)^{2s}\;|v|_{H^s(K_\ell)}^2. \tag{4.32}$$

Let us prove inequality (4.31)-(ii). Using a change of variables (cf. (4.3) and (4.6)), we obtain

$$\|v - \pi_\ell^k v\|_{L^2(F_f)}^2 \lesssim |F_f|\,\|\hat{v}_\ell - \pi_{\hat{K}}^k \hat{v}_\ell\|_{L^2(\hat{F}_{f,\ell})}^2. \tag{4.33}$$

By elementary inequality,

$$\|\hat{v}_\ell - \pi_{\hat{K}}^k \hat{v}_\ell\|_{L^2(\hat{F}_{f,\ell})}^2 \le 2\left(\|\hat{v}_\ell - \underline{\hat{v}}_\ell\|_{L^2(\hat{F}_{f,\ell})}^2 + \|\underline{\hat{v}}_\ell - \pi_{\hat{K}}^k \hat{v}_\ell\|_{L^2(\hat{F}_{f,\ell})}^2\right). \tag{4.34}$$

Let us bound the first term in the right-hand side of (4.34). Using inequality (4.16) and Poincaré-Steklov inequality (2.7) both on the reference element and the fact that

$|\hat{v}_\ell - \underline{\hat{v}}_\ell|_{H^s(\hat{K})} = |\hat{v}_\ell|_{H^s(\hat{K})}$, we obtain

$$\begin{aligned} \|\hat{v}_\ell - \underline{\hat{v}}_\ell\|^2_{L^2(\hat{F}_{f,\ell})} &\lesssim \|\hat{v}_\ell - \underline{\hat{v}}_\ell\|^2_{H^s(\hat{K})} \\ &\lesssim \|\hat{v}_\ell - \underline{\hat{v}}_\ell\|^2_{L^2(\hat{K})} + |\hat{v}_\ell|^2_{H^s(\hat{K})} \\ &\lesssim |\hat{v}_\ell|^2_{H^s(\hat{K})}. \end{aligned} \tag{4.35}$$

Let us bound the second term in the right-hand side of (4.34). Since $\underline{\hat{v}}_\ell - \pi^k_{\hat{K}} \hat{v}_\ell$ is a polynomial, inequality (4.12) yields

$$\|\underline{\hat{v}}_\ell - \pi^k_{\hat{K}} \hat{v}_\ell\|^2_{L^2(\hat{F}_{f,\ell})} \lesssim \|\underline{\hat{v}}_\ell - \pi^k_{\hat{K}} \hat{v}_\ell\|^2_{L^2(\hat{K})}. \tag{4.36}$$

Then, we use again the elementary inequality

$$\begin{aligned} \|\underline{\hat{v}}_\ell - \pi^k_{\hat{K}} \hat{v}_\ell\|^2_{L^2(\hat{K})} &\leq 2\left(\|\hat{v}_\ell - \underline{\hat{v}}_\ell\|^2_{L^2(\hat{K})} + \|\hat{v}_\ell - \pi^k_{\hat{K}} \hat{v}_\ell\|^2_{L^2(\hat{K})}\right) \\ &\leq 4\,\|\hat{v}_\ell - \underline{\hat{v}}_\ell\|^2_{L^2(\hat{K})}. \end{aligned} \tag{4.37}$$

Applying Poincaré-Steklov inequality (2.7), we get

$$\|\hat{v}_\ell - \underline{\hat{v}}_\ell\|^2_{L^2(\hat{K})} \lesssim |\hat{v}_\ell|^2_{H^s(\hat{K})}. \tag{4.38}$$

Collecting (4.36), (4.37) and (4.38), we have:

$$\|\underline{\hat{v}}_\ell - \pi^k_{\hat{K}} \hat{v}_\ell\|^2_{L^2(\hat{F}_{f,\ell})} \lesssim |\hat{v}_\ell|^2_{H^s(\hat{K})}. \tag{4.39}$$

Combining (4.34), (4.35) and (4.39) gives

$$\|\hat{v}_\ell - \pi^k_{\hat{K}} \hat{v}_\ell\|^2_{L^2(\hat{F}_{f,\ell})} \lesssim |\hat{v}_\ell|^2_{H^s(\hat{K})}.$$

Substituting this estimate into (4.33) and using (4.8)-(i) and (4.2), we find

$$\|v - \pi^k_\ell v\|^2_{L^2(F_f)} \lesssim |F_f|\,|K_\ell|^{-2}\,(h_\ell)^{d+2s}\,|v|^2_{H^s(K_\ell)}.$$

Finally, using (4.9), we get $|F_f|\,|K_\ell|^{-2} \lesssim (\rho_\ell)^{-1-d}$. Using that $(\rho_\ell)^{-1} = \sigma_\ell (h_\ell)^{-1}$, we obtain, for all $f \in \mathcal{I}_{F,\ell}$,

$$\|v - \pi^k_\ell v\|_{L^2(F_f)} \lesssim (\sigma_\ell)^{\frac{d+1}{2}}\,(h_\ell)^{s-\frac{1}{2}}\,|v|_{H^s(K_\ell)}.$$

□

Last, we shall need the following estimate:

Proposition 4.9 *Let $\ell \in \mathcal{I}_K$, and $v \in H^1(K_\ell)$, it holds:*

$$|\pi_\ell^k v|_{H^1(K_\ell)} \lesssim \sigma_\ell \|\mathbf{grad}\, v\|_{\mathbf{L}^2(K_\ell)}. \tag{4.40}$$

Moreover if $v \in H^1(\Omega)$ by summation on ℓ we have:

$$\|\pi_h^k v\|_h \lesssim \sigma \|\mathbf{grad}\, v\|_{\mathbf{L}^2(\Omega)}. \tag{4.41}$$

Proof. Proceeding as in the proof of Proposition 4.4, we can consider the following generalized eigenproblem:

$$\left\{ \begin{array}{l} \text{Find } (\hat{V}_{\lambda_{\hat{K}}}, \lambda_{\hat{K}}) \in \mathbb{R}^{N_k} \backslash \{(0,0)\} \text{ such that } (\hat{V}_{\lambda_{\hat{K}}}, \lambda_{\hat{K}}) \text{ solves} \\ \quad \mathbb{M}_{\hat{K}}\, \hat{V}_\lambda = \lambda\, \mathbb{K}_{\hat{K}}\, \hat{V}_\lambda \text{ with } \lambda_{\hat{K}} = \min_\lambda(\lambda) \text{ and } \|\hat{V}_{\lambda_{\hat{K}}}\| = 1. \end{array} \right.$$

We then have: $\|\mathbf{grad}_{\hat{\mathbf{x}}}(\pi_{\hat{K}}^k \hat{v}_\ell - \underline{\hat{v}}_\ell)\|_{L^2(\hat{K})}^2 \leq \lambda_{\hat{K}}^{-1} \|\pi_{\hat{K}}^k \hat{v}_\ell - \underline{\hat{v}}_\ell\|_{L^2(\hat{K})}^2$, and we let $C_{\hat{K}} = \lambda_{\hat{K}}^{-1}$. Using the Poincaré-Steklov inequality (2.9) in the reference element, it holds:

$$\begin{array}{rcl} \|\mathbf{grad}_{\hat{\mathbf{x}}} \pi_{\hat{K}}^k \hat{v}_\ell\|_{\mathbf{L}^2(\hat{K})} & \leq & (C_{\hat{K}})^{\frac{1}{2}}\, \|\pi_{\hat{K}}^k \hat{v}_\ell - \underline{\hat{v}}_\ell\|_{L^2(\hat{K})}, \\ & \leq & (C_{\hat{K}})^{\frac{1}{2}}\, \|\pi_{\hat{K}}^k (\hat{v}_\ell - \underline{\hat{v}}_\ell)\|_{L^2(\hat{K})}, \\ & \leq & (C_{\hat{K}})^{\frac{1}{2}}\, \|\hat{v}_\ell - \underline{\hat{v}}_\ell\|_{L^2(\hat{K})} \\ & \leq & (C_{\hat{K}})^{\frac{1}{2}}\, \pi^{-1}\, h_{\hat{K}}\, \|\mathbf{grad}_{\hat{\mathbf{x}}} \hat{v}_\ell\|_{\mathbf{L}^2(\hat{K})}, \end{array} \tag{4.42}$$

Using (4.7)-(ii) (resp. (4.7)-(i)) to get a lower (resp. upper) bound of the left-hand (resp. right-hand) side of (4.42), we get:

$$\begin{array}{l} \|\mathbf{grad}\, \pi_\ell^k v\|_{\mathbf{L}^2(K_\ell)}^2 \leq \|\mathbb{B}_\ell^{-1}\|^2 |J_\ell| \|\mathbf{grad}_{\hat{\mathbf{x}}} \pi_{\hat{K}}^k \hat{v}\|_{\mathbf{L}^2(\hat{K})}^2, \\ \quad \leq C_{\hat{K}}\, \pi^{-2}\, (h_{\hat{K}})^2\, \|\mathbb{B}_\ell^{-1}\|^2\, |J_\ell|\, \|\mathbf{grad}_{\hat{\mathbf{x}}} \hat{v}_\ell\|_{\mathbf{L}^2(\hat{K})}^2, \\ \quad \leq C_{\hat{K}}\, \pi^{-2}\, (h_{\hat{K}})^2\, \|\mathbb{B}_\ell^{-1}\|^2\, |J_\ell|\, \|\mathbb{B}_\ell\|^2 |J_\ell|^{-1} \|\mathbf{grad}\, v_\ell\|_{\mathbf{L}^2(K_\ell)}^2. \end{array}$$

Using (4.2), we have: $(\|\mathbb{B}_\ell^{-1}\|\, \|\mathbb{B}_\ell\|)^2 \leq (\sigma_{\hat{K}}\, \sigma_\ell)^2$, so that:

$$\|\mathbf{grad}\, \pi_\ell^k v\|_{\mathbf{L}^2(K_\ell)}^2 \lesssim (\sigma_\ell)^2 \|\mathbf{grad}\, v_\ell\|_{\mathbf{L}^2(K_\ell)}^2.$$

We deduce (4.41) by summation.

□

4.5 Stable and consistent discrete bilinear form

Let U_h be discrete spaces, and let

$$a_h : U_h \times U_h \to \mathbb{R}$$

be a bilinear form, and $l_h(\cdot)$ be a linear form such that:

$$a_h(u_h,\, v_h) = l_h(v_h), \forall u_h,\, v_h \in U_h \tag{4.43}$$

Assume that the finite-dimensional space U_h has dimension N and is spanned by the basis functions $(\phi_i)_{i=1,\dots,N}$. Then, the discrete problem (4.43) can be written in matrix form of dimension N. That is, there exist a matrix $\mathbb{A} \in \mathbb{R}^{N\times N}$ and vectors $\mathbf{x}$, $\mathbf{b} \in \mathbb{R}^N$ such that

$$\mathbb{A}_{i,j} = a_h(\phi_j, \phi_i), \quad \mathbf{b}_i = l_h(\phi_i), \tag{4.44}$$

and $\mathbf{x}$ is the vector of unknowns representing the coefficients of the discrete solution

$$u_h = \sum_{i=1}^{N} \mathbf{x}_i\, \phi_i.$$

Therefore, problem (4.43) is equivalent to solving the following linear system:

$$\mathbb{A}\mathbf{x} = \mathbf{b}. \tag{4.45}$$

Definition 4.10 (`Stability (Uniform Discrete Inf-Sup Condition)`) *The bilinear form a_h is said to be* stable *if there exists a constant $\beta > 0$, independent of the mesh size h, such that*

$$\inf_{u_h \in U_h\setminus\{0\}} \sup_{v_h \in U_h\setminus\{0\}} \frac{a_h(u_h, v_h)}{\|u_h\|_{U_h}\|v_h\|_{U_h}} \geq \beta > 0,$$

where $\|\cdot\|_{U_h}$ is an appropriate norm on U_h.

Remark 4.11 *The uniform* inf-sup *condition ensures the stability of the numerical method and the uniqueness of the discrete solution. The notion of discrete stability is equivalent to the condition that the matrix* $\mathbb{A}$ *in problem* (4.44) *is invertible. This implies that the discrete problem* (4.43) *is well-posed.*

Let $a : U \times U \to \mathbb{R}$ be the continuous bilinear form defined on function spaces U, and let $u \in U$ be the exact solution of the continuous problem, satisfying

$$a(u, v) = l(v), \quad \forall v \in U,$$

where l is the continuous linear form.

To define this consistency, the exact solution must be inserted into the discrete bilinear form a_h. However, since a_h is originally defined only on $U_h \times U_h$, this is not always possible. Therefore, we assume the existence of a space $U_{*,h}$, such that $U \subset U_{*,h}$, and such that the bilinear form a_h can be extended to $U_{*,h} \times U_h$. This extension relies on a regularity assumption on the exact solution u and is necessary to define the notion of consistency.

Definition 4.12 (`Consistency`) *The discrete bilinear form* a_h *is said to be consistent if*

$$a_h(u, v_h) = l_h(v_h), \quad \forall v_h \in U_h, \tag{4.46}$$

where l_h *is the discrete linear form.*

Remark 4.13 *This means that inserting the exact solution* u *into the discrete form reproduces exactly the right-hand side, ensuring that the discretization error arises solely from the approximation of the discrete space* U_h*. The equality* (4.46) *holds true if and only if:*

$$a_h(u - u_h, v_h) = 0, \quad \forall u_h,\, v_h \in U_h.$$

Part I

Stokes problem discretized with the SIP method

Chapter 5

Continuous Stokes problem

Contents

In section 5.1, we analyse the Stokes problem. In section 5.2, we present a concise overview of the theory for the Stokes problem when the solution is weakly regular, i.e. $(\mathbf{u}, p) \in \mathbf{H}^{1+s}(\Omega) \times H^s(\Omega)$, with $s \in (0, 1)$.

The Stokes problem corresponds to the Navier–Stokes system, when we can neglect the term $\frac{\mathrm{d}\mathbf{u}}{dt}$, i.e. the sum of the partial time derivative with the convection term vanishes. It can describe very viscous flows. The Stokes problem is a fundamental building block for the study of the Navier–Stokes problem. It is a simplified model that plays a fundamental role in both the theoretical understanding of fluid flows and the development of robust numerical methods for Navier–Stokes, by focusing on the diffusion term, the divergence free condition, and the coupling between velocity and pressure. In our model, the vector field $\mathbf{u}$ represents the fluid velocity, and the scalar field p represents the pressure divided by the fluid density, which is assumed to be constant.

5.1 Boundary conditions

5.1.1 Dirichlet boundary conditions

We consider the Stokes problem with inhomogeneous Dirichlet boundary conditions:

$$\left\{\begin{array}{rcl} \text{Find } (\mathbf{u}, p) \text{ such that} & & \\ -\nu\,\boldsymbol{\Delta}\,\mathbf{u} + \mathbf{grad}\,p &=& \mathbf{f}, \\ \operatorname{div}\mathbf{u} &=& 0, \\ \mathbf{u}_{|\partial\Omega} &=& \mathbf{g}, \end{array}\right. \tag{5.1}$$

with a normalization condition for p: $\displaystyle\int_\Omega p = 0$, and where $\mathbf{g}$ satisfies $\displaystyle\int_{\partial\Omega} \mathbf{g}\cdot\mathbf{n} = 0$. The first equation of (5.1) is the momentum equation, and the second correponds to the conservation of mass. The constant parameter $\nu > 0$ represents the kinematic viscosity. The vector field $\mathbf{f}$ represents the external force field on the fluid, divided by the fluid density.
We assume that $\mathbf{g} \in \mathbf{H}^{\frac{1}{2}}(\partial\Omega)$. By subtraction, we can transform this problem into a Stokes problem with homogeneous Dirichlet boundary conditions. Indeed, let $\mathbf{u}_g \in \mathbf{H}^1(\Omega)$ satisfy:

$$-\boldsymbol{\Delta}\,\mathbf{u}_g = 0, \quad \mathbf{u}_{g|\partial\Omega} = \mathbf{g}. \tag{5.2}$$

Theorem 5.1 *The harmonic problem (5.2) has a unique solution such that:*

$$\|\mathbf{Grad}\,\mathbf{u}_g\|_{\mathbb{L}^2(\Omega)} \leq C_{\gamma_1,\Omega}\,(h_\Omega)^{-\frac{1}{2}}\,\|\mathbf{g}\|_{\mathbf{H}^{\frac{1}{2}}(\partial\Omega)}. \tag{5.3}$$

Proof. According to Theorem 2.7, there exists $\mathbf{v}_g \in \mathbf{H}^1(\Omega)$ such that $\mathbf{v}_{g|\partial\Omega} = \mathbf{g}$. Let $\mathbf{v}_0 = \mathbf{u}_g - \mathbf{v}_g$. Then, $\mathbf{v}_0 \in \mathbf{H}^1_0(\Omega)$ is such that $-\Delta\mathbf{v}_0 = \Delta\mathbf{v}_g$ and it solves the following variational formulation, which is well-posed:

$$\left\{\begin{array}{l} \text{Find } \mathbf{v}_0 \in \mathbf{H}^1_0(\Omega) \text{ such that, for all } \mathbf{v} \in \mathbf{H}^1_0(\Omega): \\ \quad (\mathbf{Grad}\,\mathbf{v}_0, \mathbf{Grad}\,\mathbf{v})_{\mathbb{L}^2(\Omega)} = \langle \Delta\mathbf{v}_g, \mathbf{v}\rangle_{\mathbf{H}^{-1}(\Omega),\mathbf{H}^1_0(\Omega)}. \end{array}\right. \tag{5.4}$$

Hence, there exists $\mathbf{u}_g \in \mathbf{H}^1(\Omega)$ satisfying (5.2), which is unique since the solution to (5.2) with $\mathbf{g} = 0$ in $\mathbf{H}^{\frac{1}{2}}(\partial\Omega)$ vanishes. Since $\Delta\mathbf{u}_g = 0$, we remark that $\mathbf{u}_g \in \mathbf{H}^1(\Omega;\Delta) := \{\mathbf{v} \in \mathbf{H}^1(\Omega)\,|\,, \Delta\mathbf{v} \in \mathbf{L}^2(\Omega)\}$, and $\mathbf{Grad}\,\mathbf{u}_g \in \mathbb{H}(\mathrm{Div}\,;\Omega)$ with $\mathrm{Div}\,(\mathbf{Grad}\,\mathbf{u}_g) = 0$. Hence:

$$\|\mathbf{Grad}\,\mathbf{u}_g\|_{\mathbb{H}(\mathrm{Div}\,;\Omega)} = \|\mathbf{Grad}\,\mathbf{u}_g\|_{\mathbb{L}^2(\Omega)}. \tag{5.5}$$

Moreover, using Green's formula [8, Thm. 2.2.28] and see (2.4), we have:

$$\begin{aligned} -(\Delta \mathbf{u}_g, \mathbf{u}_g)_{\mathbf{L}^2(\Omega)} &= -(\mathrm{Div}\,(\mathbf{Grad}\,\mathbf{u}_g), \mathbf{u}_g)_{\mathbf{L}^2(\Omega)}, \\ &= \|\mathbf{Grad}\,\mathbf{u}_g\|^2_{\mathbb{L}^2(\Omega)} - \langle \mathbf{Grad}\,\mathbf{u}_g \cdot \mathbf{n}_{|\partial\Omega}, \mathbf{u}_g \rangle_{\mathbf{H}^{-\frac{1}{2}}(\partial\Omega), \mathbf{H}^{\frac{1}{2}}(\partial\Omega)}. \end{aligned} \tag{5.6}$$

Since $-\Delta \mathbf{u}_g = 0$, it holds:

$$\begin{aligned} \|\mathbf{Grad}\,\mathbf{u}_g\|^2_{\mathbb{L}^2(\Omega)} &= \langle \mathbf{Grad}\,\mathbf{u}_g \cdot \mathbf{n}_{|\partial\Omega}, \mathbf{u}_g \rangle_{\mathbf{H}^{-\frac{1}{2}}(\partial\Omega), \mathbf{H}^{\frac{1}{2}}(\partial\Omega)}, & \\ &\leq \|\mathbf{Grad}\,\mathbf{u}_g \cdot \mathbf{n}_{|\partial\Omega}\|_{\mathbf{H}^{-\frac{1}{2}}(\partial\Omega)} \|\mathbf{u}_g\|_{\mathbf{H}^{\frac{1}{2}}(\partial\Omega)} & \text{cf. (2.10)}, \\ &\leq \|\mathbf{Grad}\,\mathbf{u}_g \cdot \mathbf{n}_{|\partial\Omega}\|_{\mathbf{H}^{-\frac{1}{2}}(\partial\Omega)} \|\mathbf{g}\|_{\mathbf{H}^{\frac{1}{2}}(\partial\Omega)} & \text{since } \mathbf{u}_{g|\partial\Omega} = \mathbf{g}, \\ &\leq C_{\gamma_1,\Omega}\,(h_\Omega)^{-\frac{1}{2}} \|\mathbf{Grad}\,\mathbf{u}_g\|_{\mathbb{H}(\mathrm{Div};\partial\Omega)} \|\mathbf{g}\|_{\mathbf{H}^{\frac{1}{2}}(\partial\Omega)} & \text{from (2.11)}, \\ &\leq C_{\gamma_1,\Omega}\,(h_\Omega)^{-\frac{1}{2}} \|\mathbf{Grad}\,\mathbf{u}_g\|_{\mathbb{L}^2(\Omega)} \|\mathbf{g}\|_{\mathbf{H}^{\frac{1}{2}}(\partial\Omega)} & \text{using (5.5)}. \end{aligned}$$

Hence, we obtain (5.3).

□

We define $g_{\mathrm{div}} := -\,\mathrm{div}\,\mathbf{u}_g$. We now consider the generalised Stokes problem with homogeneous Dirichlet boundary conditions on the velocity field:

$$\left\{ \begin{array}{l} \text{Find } (\widetilde{\mathbf{u}}, p) \text{ such that} \\ -\nu\,\boldsymbol{\Delta}\,\widetilde{\mathbf{u}} + \mathbf{grad}\,p = \mathbf{f}, \\ \qquad\qquad \mathrm{div}\,\widetilde{\mathbf{u}} = g_{\mathrm{div}}, \\ \qquad\qquad \widetilde{\mathbf{u}}_{|\partial\Omega} = 0. \end{array} \right. \tag{5.7}$$

It is clear that problem (5.1) and problem (5.7) are equivalent, meaning that problem (5.1) has a unique solution if and only if problem (5.7) has a unique solution, with the relation $\mathbf{u} = \widetilde{\mathbf{u}} + \mathbf{u}_g$. We made this change so that we can use the properties of the Sobolev space $H^1_0(\Omega)$.

Before establishing the variational formulation of Problem (5.7), we provide some definitions and reminders. Let us define $\mathcal{X} = \mathbf{H}^1_0(\Omega) \times L^2_{zmv}(\Omega)$, which is a Hilbert space equipped with the following norm:

$$\|(\mathbf{v}, q)\|^2_{\mathcal{X}} = \|\mathbf{v}\|^2_{\mathbf{H}^1_0(\Omega)} + \nu^{-2}\,\|q\|^2_{L^2(\Omega)}. \tag{5.8}$$

We consider the following continuous and symmetric bilinear form:

$$a_S(\cdot,\cdot) : \left\{ \begin{array}{rcl} \mathcal{X}\times\mathcal{X} & \to & \mathbb{R} \\ ((\mathbf{u}',p'),(\mathbf{v},q)) & \mapsto & \nu(\mathbf{u}',\mathbf{v})_{\mathbf{H}^1_0(\Omega)} \\ & & -(p',\operatorname{div}\mathbf{v})_{L^2(\Omega)} - (q,\operatorname{div}\mathbf{u}')_{L^2(\Omega)} \end{array} \right. . \tag{5.9}$$

We can express problem (5.7) equivalently as follows: Find $(\tilde{\mathbf{u}},p)\in\mathcal{X}$ such that for all $(\mathbf{v},q)\in\mathcal{X}$

$$a_S\left((\tilde{\mathbf{u}},p),(\mathbf{v},q)\right) = \langle \mathbf{f},\mathbf{v}\rangle_{\mathbf{H}^1_0(\Omega)} - (g_{\mathrm{div}},q)_{L^2(\Omega)}. \tag{5.10}$$

Proposition 5.1 ([28, Prop. 2]) *The bilinear form $a_S(\cdot,\cdot)$ is T-coercive.*

Proof. Let us consider $(\mathbf{u}',p')\in\mathcal{X}$ and construct a bijective operator satisfying (2.15). This requires three main steps.

— According to Proposition 2.3, there exists $\tilde{\mathbf{v}}_{p'}\in\mathbf{V}^\perp$ such that:

$$\operatorname{div}\tilde{\mathbf{v}}_{p'} = p' \text{ in } \Omega \quad \text{and} \quad \|\tilde{\mathbf{v}}_{p'}\|_{\mathbf{H}^1_0(\Omega)} \le C_{\mathrm{div}}\|p'\|_{L^2(\Omega)}. \tag{5.11}$$

Let us set $(\mathbf{v}^*,q^*) := (\lambda\mathbf{u}' - \nu^{-1}\tilde{\mathbf{v}}_{p'}, -\lambda p')$, with $\lambda>0$ to be fixed. We obtain:

$$a_S((\mathbf{u}',p'),(\mathbf{v}^*,q^*)) = \nu\lambda\|\mathbf{u}'\|^2_{H^1_0(\Omega)} + \nu^{-1}\|p'\|^2_{L^2(\Omega)} - (\mathbf{u}',\tilde{\mathbf{v}}_{p'})_{H^1_0(\Omega)}. \tag{5.12}$$

— In order to bound the last term of (5.12), we use the Young inequality and then inequality (5.11), and it follows that for all $\eta>0$:

$$(\mathbf{u}',\tilde{\mathbf{v}}_{p'})_{H^1_0(\Omega)} \le \frac{\eta}{2}\|\mathbf{u}'\|^2_{H^1_0(\Omega)} + \frac{\eta^{-1}}{2}(C_{\mathrm{div}})^2\|p'\|^2_{L^2(\Omega)}. \tag{5.13}$$

— Using the bound (5.13) in (5.12), we have:

$$\begin{aligned} &a_S((\mathbf{u}',p'),(\mathbf{v}^*,q^*)) \\ &\quad\ge \left(\nu\lambda - \frac{\eta}{2}\right)\|\mathbf{u}'\|^2_{H^1_0(\Omega)} + \left(\nu^{-1} - \frac{\eta^{-1}}{2}(C_{\mathrm{div}})^2\right)\|p'\|^2_{L^2(\Omega)}. \end{aligned}$$

We look for $\eta>0$ such that $2\nu\lambda>\eta$ and $\eta > \dfrac{\nu(C_{\mathrm{div}})^2}{2}$, which amounts to requiring

$$\lambda > \frac{1}{4}(C_{\mathrm{div}})^2.$$

According to the above, provided that $\lambda > \frac{1}{4}(C_{\rm div})^2$, we have proved that the operator $T_\lambda \in \mathcal{L}(X)$ defined by $T_\lambda((\mathbf{u}', p')) = (\lambda\mathbf{u}' - \nu^{-1}\tilde{\mathbf{v}}_{p'}, -\lambda p')$ is such that:

$$\exists \alpha_\lambda > 0,\ \forall (\mathbf{u}', p') \in X,\ a_S((\mathbf{u}', p'), T_\lambda((\mathbf{u}', p'))) \geq \alpha_\lambda \|(\mathbf{u}', p')\|_{\mathcal{X}}^2.$$

The injectivity of the operator T_λ follows. Given $(\mathbf{v}^*, q^*) \in X$, choosing $(\mathbf{u}', p') = ((\lambda^{-1}\mathbf{v}^* - \nu^{-1}\lambda^{-2}\tilde{\mathbf{v}}_{q^*}, -\lambda^{-1}q^*)$ yields $T_\lambda((\mathbf{u}', p')) = (\mathbf{v}^*, q^*)$. Hence, the operator $T_\lambda \in L(X)$ is bijective. Establishing that the bijective operator $T_\lambda((\mathbf{u}', p')) = (\lambda\mathbf{u}' - \nu^{-1}\tilde{\mathbf{v}}_{p'}, -\lambda p')$ leads to T-coercivity as soon as $\lambda > \frac{1}{4}(C_{\rm div})^2$. Observe that one can have even more flexibility in the choice of T_λ by choosing a different factor in front of $\tilde{\mathbf{v}}_{p'}$, and then choosing λ accordingly.

□

We can split problem (5.7) into two subproblems. Since $\mathbf{H}_0^1(\Omega) = \mathbf{V} \oplus \mathbf{V}^\perp$, there exists a unique decomposition $\widetilde{\mathbf{u}} = \widetilde{\mathbf{u}}_0 + \widetilde{\mathbf{u}}_\perp$ with $\widetilde{\mathbf{u}}_0 \in \mathbf{V}$ and $\widetilde{\mathbf{u}}_\perp \in \mathbf{V}^\perp$. Then, it can be checked that there exist $(\widetilde{p}_0, \widetilde{p}_\perp) \in (L^2_{zmv}(\Omega))^2$ such that $p = \widetilde{p}_0 + \widetilde{p}_\perp$ and such that the pair $(\widetilde{\mathbf{u}}_0, \widetilde{p}_0) \in \mathbf{V} \times L^2_{zmv}(\Omega)$ is solution of:

$$\left\{ \begin{array}{rcl} -\nu\, \boldsymbol{\Delta}\, \widetilde{\mathbf{u}}_0 + \mathbf{grad}\, \widetilde{p}_0 &=& \mathbf{f}, \\ \mathrm{div}\, \widetilde{\mathbf{u}}_0 &=& 0, \\ \widetilde{\mathbf{u}}_{0|\partial\Omega} &=& 0, \end{array} \right. \tag{5.14}$$

and the pair $(\widetilde{\mathbf{u}}_\perp, \widetilde{p}_\perp) \in \mathbf{V}^\perp \times L^2_{zmv}(\Omega)$ is solution of:

$$\left\{ \begin{array}{rcl} -\nu\, \boldsymbol{\Delta}\, \widetilde{\mathbf{u}}_\perp + \mathbf{grad}\, \widetilde{p}_\perp &=& 0, \\ \mathrm{div}\, \widetilde{\mathbf{u}}_\perp &=& g_{\rm div}, \\ \widetilde{\mathbf{u}}_{\perp|\partial\Omega} &=& 0. \end{array} \right. \tag{5.15}$$

We can now prove the following result for the generalized Stokes problem:

Theorem 5.2 ([28, Thm. 2]) *Problem (5.10) is well-posed: it admits one and only one solution for any $(\mathbf{f}, g) \in \mathbf{H}^{-1}(\Omega) \times L^2_{zmv}(\Omega)$. Writing $\widetilde{\mathbf{u}} = \widetilde{\mathbf{u}}_0 + \widetilde{\mathbf{u}}_\perp$ with $\widetilde{\mathbf{u}}_0 \in \mathbf{V}$ and $\widetilde{\mathbf{u}}_\perp \in \mathbf{V}^\perp$, the solution is such that:*

$$\left\{ \begin{array}{l} \textit{For all } \mathbf{f} \in \mathbf{H}^{-1}(\Omega), \textit{ for all } g_{\rm div} \in L^2_{zmv}(\Omega) \\ \|\widetilde{\mathbf{u}}_\perp\|_{\mathbf{H}_0^1(\Omega)} \leq C_{\rm div} \|g_{\rm div}\|_{L^2(\Omega)}, \\ \|\widetilde{\mathbf{u}}_0\|_{\mathbf{H}_0^1(\Omega)} \leq \nu^{-1} \|\mathbf{f}\|_{\mathbf{H}^{-1}(\Omega)} \\ \|p\|_{L^2(\Omega)} \leq C_{\rm div} \|\mathbf{f}\|_{\mathbf{H}^{-1}(\Omega)} + \nu\, (C_{\rm div})^2 \|g_{\rm div}\|_{L^2(\Omega)} \end{array} \right. . \tag{5.16}$$

We note, using (5.3), that

$$\|g_{\text{div}}\|_{L^2(\Omega)} \leq \sqrt{d}\,\|\mathbf{Grad}\,\mathbf{u}_g\|_{\mathbb{L}^2(\Omega)} \leq \sqrt{d}\,C_{\gamma_1,\Omega}\,(h_\Omega)^{-\frac{1}{2}}\,\|\mathbf{g}\|_{\mathbf{H}^{\frac{1}{2}}(\partial\Omega)}. \tag{5.17}$$

Returning to the Stokes problem with inhomogeneous Dirichlet boundary conditions (5.1), and applying inequalities (5.3), (5.16) together with inequality (5.17), we obtain:

$$\left\{\begin{array}{rcl} \|\mathbf{u}\|_{\mathbf{H}^1_0(\Omega)} & \leq & \nu^{-1}\|\mathbf{f}\|_{\mathbf{H}^{-1}(\Omega)} + (1+\sqrt{d}\,C_{\text{div}})\,C_{\gamma_1,\Omega}\,(h_\Omega)^{-\frac{1}{2}}\,\|\mathbf{g}\|_{\mathbf{H}^{\frac{1}{2}}(\partial\Omega)}, \\ \|p\|_{L^2(\Omega)} & \leq & C_{\text{div}}\left(\|\mathbf{f}\|_{\mathbf{H}^{-1}(\Omega)} + \nu\,\sqrt{d}\,C_{\text{div}}\,C_{\gamma_1,\Omega}\,(h_\Omega)^{-\frac{1}{2}}\,\|\mathbf{g}\|_{\mathbf{H}^{\frac{1}{2}}(\partial\Omega)}\right). \end{array}\right. \tag{5.18}$$

5.1.2 Neumann boundary conditions

In part I devoted to the Stokes problem, we focused on the case with Dirichlet boundary conditions. However, in part II, some physical test cases require the use of the Stokes problem with Neumann-type boundary conditions. For this reason, we briefly introduce the Stokes problem with Neumann boundary conditions in the following.

We consider the Stokes problem with Neumann boundary condition:

$$\left\{\begin{array}{l} \text{Find } (\mathbf{u},p) \text{ such that} \\ \qquad\qquad -\nu\,\boldsymbol{\Delta}\,\mathbf{u} + \mathbf{grad}\,p = \mathbf{f} \\ \qquad\qquad \text{div}\,\mathbf{u} = 0 \\ \qquad\qquad \mathbf{u}_{|\Gamma_D} = 0 \\ (\nu\,\mathbf{Grad}\,\mathbf{u} - p\mathbb{I}_d) : \mathbf{n}_{|\Gamma_N} = \mathbf{g} \end{array}\right. \tag{5.19}$$

Where $\mathbb{I}_d$ is the identity tensor on $\mathbb{R}^d$, $\partial\Omega = \Gamma_D \cap \Gamma_N$, $\Gamma_D \cup \Gamma_N = \emptyset$, and $|\Gamma_D| > 0$. Many works have addressed the Stokes problem with Neumann-type boundary conditions, proving its well-posedness in regular cases and considering various types of Neumann conditions, particularly when $|\Gamma_D| = 0$ see, [95, 50, 94, 1, 81, 62, 83]. Additionally, there are studies dealing with this problem under weaker regularity assumptions on the boundary conditions [52, 80, 17, 41, 16]. Finally, reference [15] provides a comprehensive presentation of several boundary condition types for the Stokes problem, along with a broad theoretical analysis.

$$\mathbf{V}_D := \left\{\mathbf{v} \in \mathbf{H}^1(\Omega) \,\middle|\, \gamma_0(\mathbf{v})_{|\Gamma_D} = 0\right\}.$$

In this case, we define the space

$$\mathcal{X} := \mathbf{V}_D \times L^2(\Omega),$$

equipped with the norm

$$\|(\mathbf{u}, q)\|^2 := |\mathbf{u}|^2_{\mathbf{H}^1(\Omega)} + \nu^{-2} \|q\|^2_{L^2(\Omega)}.$$

This defines a valid norm on $\mathcal{X}$, since $|\Gamma_D| > 0$.
The variational formulation of problem (5.19) retains the same structure as that of the Stokes problem with Dirichlet boundary conditions. Find $(\mathbf{u}, p) \in \mathcal{X}$ such that for all $(\mathbf{v}, q) \in \mathcal{X}$

$$a_S\left((\mathbf{u}, p), (\mathbf{v}, q)\right) = \langle \mathbf{f}, \mathbf{v} \rangle_{\mathbf{H}^1_0(\Omega)} + (\mathbf{g}, \mathbf{v})_{L^2(\Gamma_N)}. \tag{5.20}$$

Lemma 5.3 ([50, Lemma 53.9]) *Consider the partition $\partial\Omega = \Gamma_D \cup \Gamma_N$ with $|\Gamma_D| > 0$, $|\Gamma_N| > 0$, and assume there exists a subset $O \subset \Gamma_N$ such that $|O| > 0$ and the normal vector satisfies $\boldsymbol{n}|_O \in \mathbf{H}^{1/2}(O)$. Then the operator*

$$\operatorname{div} : X := \left\{ \mathbf{v} \in \boldsymbol{V}_D \,|\, \gamma_0(\mathbf{v})_{|\Gamma_N} \cdot \mathbf{n} = 0 \right\} \longrightarrow L^2(\Omega)$$

is surjective. Identifying $(L^2(\Omega))'$ with $L^2(\Omega)$, we have the following inf-sup condition:

$$\inf_{q \in L^2(\Omega)} \sup_{\mathbf{v} \in \boldsymbol{V}_D} \frac{|\int_\Omega q \, \operatorname{div} \mathbf{v} \, dx|}{\|q\|_{L^2(\Omega)} \, |\mathbf{v}|_{\mathbf{H}^1(\Omega)}} := \beta_\Omega > 0.$$

That is, for every $q \in L^2(\Omega)$, there exists $\mathbf{v}_q \in \boldsymbol{V}_D$ such that:

$$\operatorname{div} \mathbf{v}_q = q, \quad \text{and} \quad |\mathbf{v}_q|_{\mathbf{H}^1(\Omega)} \leq \beta^{-1} \|q\|_{L^2(\Omega)}.$$

To establish a connection with the Stokes problem with Dirichlet boundary conditions, based on the generalization of property 2.3 by lemma (5.3), one can prove that problem (5.20) is well-posed similarly to the proof of Proposition (5.1).

5.2 Low-regularity solution of the Stokes problem

In our studies, Ω is a connected, bounded domain with a Lipschitz polytopal boundary. We recall the regularity results on Poisson's equation and Stokes problem that will be

useful for our analysis. Let $\mathtt{s} \in (0,1)$. We set for $n = 0, 1$:

$$\dot{H}^{n+\mathtt{s}}(\Omega) = \bigcap_{s\in(0,\mathtt{s})} H^{n+s}(\Omega). \tag{5.21}$$

Consider Poisson's problem with homogeneous Dirichlet boundary conditions:

$$\text{Find } u \text{ such as } -\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega. \tag{5.22}$$

Suppose that $f \in L^2(\Omega)$. If Ω is convex, then $u \in H_0^1(\Omega) \cap H^2(\Omega)$ [61, Cor. 2.6.8]. If Ω is non convex, there exists a constant $\sigma_P \in (\frac{1}{2}, 1)$ depending on the angles of the re-entrant edges such that $u \in H_0^1(\Omega) \cap \dot{H}^{1+\sigma_P}(\Omega)$ (and $\mathbf{grad}\, u \in \dot{\mathbf{H}}^{\sigma_P}(\Omega)$) [36, §1.A], [61, Cor. 2.4.4 and 2.6.7]. Suppose that $f \in H^{s-1}(\Omega)$, with $s \in (0, \sigma_f)$. If Ω is convex, $u \in H_0^1(\Omega) \cap \dot{H}^{1+\sigma_f}(\Omega)$. If Ω is non convex $u \in H_0^1(\Omega) \cap \dot{H}^{1+\sigma_{P,\max}}(\Omega)$ with $\sigma_{P,\max} = \min(\sigma_P, \sigma_f)$. We now use the same notation $\sigma_{P,\max}$ for σ_f (Ω convex and $f \notin L^2(\Omega)$) or σ_P (Ω non convex and $f \in L^2(\Omega)$) or $\min(\sigma_P, \sigma_f)$ (Ω non convex and $f \notin L^2(\Omega)$), see Grivard's books [60, 61]. According to [36, Corollary 23.5] (see also [25, §5.2]), it holds for $s \in (0, \sigma_{P,\max})$:

$$\|u\|_{H^{1+s}(\Omega)} \leq C_P \, \|f\|_{H^{s-1}(\Omega)} \text{ where } C_P > 0. \tag{5.23}$$

We generalize the case to a non-homogeneous Dirichlet condition and use the definition of σ_P in the following lemma:

Lemma 5.4 ([66]) *Consider the following harmonic Poisson problem:*

$$\textit{Find } u \textit{ such as } -\Delta u = 0 \textit{ in } \Omega, \quad u = g \textit{ on } \partial\Omega. \tag{5.24}$$

Let $g \in H^{\frac{1}{2}+s}(\partial\Omega)$, $\forall s \in (\frac{1}{2}, \sigma_P)$. Then, the solution u of (5.24) is such that $u \in \dot{H}^{1+\sigma_P}(\Omega)$, and it holds, for all $s \in (\frac{1}{2}, \sigma_P)$:

$$\|u\|_{H^{1+s}(\Omega)} \leq \tilde{C}_P \, (h_\Omega)^{\frac{1}{2}} \|g\|_{H^{\frac{1}{2}+s}(\partial\Omega)}, \tag{5.25}$$

where $\tilde{C}_P > 0$ is a constant depending on s.

Let Ω be a convex domain. Consider the Stokes problem (5.7) with $\mathbf{f} \in \mathbf{L}^2(\Omega)$ and

$g_{\text{div}} = 0$. It holds [57, Thm. I-1.8]:

$$\begin{aligned}(h_\Omega)^{-1}\,\|\tilde{\mathbf{u}}\|_{\mathbf{H}^2(\Omega)} + \nu^{-1}\,\|p\|_{H^1(\Omega)} \\ \leq C_{S,0}\,h_\Omega\,\nu^{-1}\|\mathbf{f}\|_{\mathbf{L}^2(\Omega)} \text{ where } C_{S,0} > 0.\end{aligned} \tag{5.26}$$

Consider the Stokes problem (5.7). Suppose that there exists $\sigma_{\mathbf{f}} \in (0,1)$, such that for all $s \in (0, \sigma_{\mathbf{f}})$, $\mathbf{f} \in \mathbf{H}^{s-1}(\Omega)$. Suppose moreover that for all $s \in (\frac{1}{2}, \sigma_P)$, $g_{\text{div}} \in L^2_{zmv}(\Omega) \cap H^s(\Omega)$. According to [37], there exists $\sigma_S \in (0,1)$ depending on Ω such that $(\tilde{\mathbf{u}}, p) \in \left(\mathbf{H}^1_0(\Omega) \cap \dot{\mathbf{H}}^{1+\sigma_{\max}}(\Omega)\right) \times (L^2_{zmv}(\Omega) \cap \dot{H}^{\sigma_{\max}}(\Omega))$, where $\sigma_{\max} = \min(\sigma_{\mathbf{f}}, \sigma_P, \sigma_S)$ and it holds, for all $s \in (0, \sigma_{\max})$:

$$\begin{aligned}(h_\Omega)^{-1}\,\|\tilde{\mathbf{u}}\|_{\mathbf{H}^{1+s}(\Omega)} + \nu^{-1}\,\|p\|_{H^s(\Omega)} \\ \leq \tilde{C}_S\,\left(\nu^{-1}\,h_\Omega\,\|\mathbf{f}\|_{\mathbf{H}^{s-1}(\Omega)} + \|g_{\text{div}}\|_{H^s(\Omega)}\right), \text{ where } \tilde{C}_S > 0.\end{aligned} \tag{5.27}$$

Inequality (5.27) holds when $\mathbf{f} \in \mathbf{L}^2(\Omega)$, $g_{\text{div}} = 0$, with $\sigma_{\max} = \sigma_S$.
The difficulty is to compute σ_S. We have, for instance [37, p.75]:

1. If the domain Ω is a $3D$ convex domain, then $\sigma_S \leq 1$.
2. If the domain Ω is a $3D$ convex domain with wedge angles $\leq \frac{2}{3}\pi$ then $\sigma_S \leq \frac{3}{2}$.
3. If the domain Ω is such that $\Omega = Q_1 \backslash Q_2$ and Q_1 and Q_2 are rectangular parallelepipeds with the same axes, then $\sigma_S \lesssim 0.544$. The domain Ω may be a L-shape domain in $2D$ or a prismatic L-shape domain in $3D$, or a Fichera corner in $3D$.

We generalize the analysis by considering a non-homogeneous Dirichlet condition together with an incompressibility condition in the following lemma:

Lemma 5.5 *Consider the Stokes problem (5.1). Suppose that there exists $\sigma_{\mathbf{f}} \in (0,1)$ such that, for all $s \in (0, \sigma_{\mathbf{f}})$, $\mathbf{f} \in \mathbf{H}^{1-s}(\Omega)$. Suppose moreover that for all $s \in (\frac{1}{2}, \sigma_P)$, $\mathbf{g} \in \mathbf{H}^{\frac{1}{2}+s}(\partial\Omega)$. Using the splitting $\mathbf{u} = \tilde{\mathbf{u}} + \mathbf{u}_{\mathbf{g}}$, and by lemma 5.4 (used component by component) we get $\mathbf{u}_{\mathbf{g}} \in \dot{\mathbf{H}}^{1+\sigma_P}(\Omega)$, and thus $g_{\text{div}} \in \dot{H}^{\sigma_P}(\Omega)$. Using (5.27) and (5.25), it holds that $(\mathbf{u}, p) \in \dot{\mathbf{H}}^{1+\sigma_{\max}}(\Omega) \times (L^2_{zmv}(\Omega) \cap \dot{H}^{\sigma_{\max}}(\Omega))$ and for all $s \in (0, \sigma_{\max})$:*

$$\begin{aligned}(h_\Omega)^{-1}\,\|\mathbf{u}\|_{\mathbf{H}^{1+s}(\Omega)} + \nu^{-1}\,\|p\|_{H^s(\Omega)} \\ \leq C_S\,\left(\nu^{-1}\,h_\Omega\|\mathbf{f}\|_{\mathbf{H}^{s-1}(\Omega)} + (h_\Omega)^{-\frac{1}{2}}\|\mathbf{g}\|_{\mathbf{H}^{\frac{1}{2}+s}(\partial\Omega)}\right), \text{ where } C_S > 0.\end{aligned} \tag{5.28}$$

We define:

$$C(\mathbf{f}, \mathbf{g}) = C_S\,\left(\nu^{-1}\,h_\Omega\,\|\mathbf{f}\|_{\mathbf{H}^{s-1}(\Omega)} + (h_\Omega)^{-\frac{1}{2}}\,\|\mathbf{g}\|_{\mathbf{H}^{\frac{1}{2}+s}(\Omega)}\right). \tag{5.29}$$

In our numerical results, cf. Section 9.1, we consider the Stokes problem (5.1) with the following domains and data:

1. The domain Ω is a $2D$ convex domain, $\mathbf{f} \in \mathbf{H}^{s-1}(\Omega)$ with $s \in (0,1)$ and $\mathbf{g} \in \mathbf{H}^{\frac{1}{2}+s}(\partial\Omega)$ with $s \in (\frac{1}{2},1)$.
2. The domain Ω is a $2D$ non convex domain, $\mathbf{f} = 0$ and $\mathbf{g} \in \mathbf{H}^{\frac{1}{2}+s}(\partial\Omega)$ with $s \in (\frac{1}{2},1)$.

In this chapter, we studied the continuous Stokes problem with non homogeneous Dirichlet boundary conditions and explained how to reduce it to homogeneous ones. We derived the variational formulation, showing that Stokes problem is well-posed with the T-coercivity theory. We considered Stokes problem with Neumann boundary conditions. Then, we provide or assume estimates on the dependency of the solution on the data. We can now study the discretization of Stokes problem when the solution is weakly regular.

Chapter 6

Discretization with low regular solution

Contents

In this chapter, in section 6.1, we introduce the SIP discretization of the Poisson problem, considering simplicial triangulation. We allow the stability parameter η to vary across mesh faces, leading to a face-dependent coefficient η_f, which seems to be new. Then, in section 6.2, we generalize the analysis to a polytopal triangulation, and we exhibit the stability constant. We detail the discretization of Stokes problem in section 6.3. Notice that when the solution of Stokes problem is of low-regularity, i.e. $(\mathbf{u}, p) \in \mathbf{H}^{1+s}(\Omega) \times H^s(\Omega)$, we preserve the consistency of the discrete formulation, assuming that $s > \frac{1}{2}$. In section 6.4, we propose new numerical scheme for which the polynomial order of the pressure is higher than that of the velocity. We demonstrate the effectiveness of this approach in the numerical results chapter 9, particularly when ν is very small. Last, in section 6.5, we exhibit the dependence of the coercivity constant for the Stokes problem on the mesh regularity parameter σ.

6.1 Discretization of the Poisson problem

As there is diffusion term, equal to $-\nu\,\mathbf{\Delta}\,\mathbf{u}$, in the first equation of Stokes Problem (5.7), we first study the discretization of Poisson's problem with the Symmetric Interior Penalization method, following [40, Chap. 4], but considering weakly regular solutions. Poisson's problem with inhomogeneous Dirichlet boundary reads:

$$\text{Find } u \text{ such that } \; -\Delta u = f \text{ in } \Omega, \quad u = g \text{ on } \partial\Omega. \tag{6.1}$$

We suppose that the solution is such that $u \in \dot{H}^{1+\sigma_{\max}}(\Omega)$, with $\sigma_{\max} \in (\frac{1}{2}, 1)$. Hence $\mathbf{grad}\, v \in \dot{\mathbf{H}}^{\sigma_{\max}}(\Omega)$ and for all $\ell \in \mathcal{I}_K$, $\mathbf{grad}\, v \cdot \mathbf{n}_\ell \in L^2(\partial K_\ell)$.
Let $k \in \mathbb{N}\backslash\{0\}$. We set $X_h = P^k_{disc}(\mathcal{T}_h)$ and $X_{\star h} := X_h + \dot{H}^{1+\sigma_{\max}}(\Omega)$.
As a consequence of Theorem 2.5, the following property holds:

Property 6.1 *Let $t \in (0, \frac{1}{2})$. It holds: $X_h \subset H^t(\Omega)$.*

Proof. Let $v_h \in X_h$. For all $\ell \in \mathcal{I}_K$, $v_{h|K_\ell} \in H^t(K_\ell)$. Let $\tilde{v}^\ell_h \in L^2(\Omega)$ be the extension of $v_{h|K_\ell}$ by zero outside K_ℓ. According to Theorem 2.5, $\tilde{v}^\ell_h \in H^t(\Omega)$. Moreover, since $v_h = \sum_{\ell\in\mathcal{I}_K} \tilde{v}^\ell_h$, it follows that $v_h \in H^t(\Omega)$.

□

Another similar consequence of Theorem 2.5 is that we can break the duality pairing $\langle\cdot,\cdot\rangle_{H^{-t}(\Omega),H^t(\Omega)}$ into a sum over all the elements of $\mathcal{T}_h$:

Proposition 6.2 *For all $t \in (0, \frac{1}{2})$, for all $\phi \in H^{-t}(\Omega)$, it holds:*

$$\forall v \in H^t(\Omega), \quad \langle \phi, v\rangle_{H^{-t}(\Omega),H^t(\Omega)} = \sum_{\ell\in\mathcal{I}_K} \langle \phi, v\rangle_{H^{-t}(K_\ell),H^t(K_\ell)}, \tag{6.2}$$

where the action of ϕ as an element of $H^{-t}(K_\ell)$ is defined by

$$\langle \phi, v\rangle_{H^{-t}(K_\ell),\,H^t(K_\ell)} = \langle \phi, \tilde{v}^\ell\rangle_{H^{-t}(\Omega),\,H^t(\Omega)},$$

where $\tilde{v}^\ell$ denotes the extension of $v_{|K_\ell}$ by zero over Ω.

Since $u \in \dot{H}^{1+\sigma_{\max}}(\Omega)$, then for all $s \in (\frac{1}{2}, \sigma_{\max})$, $t = 1 - s \in (0, \frac{1}{2})$, and $-\Delta u = f \in H^{-t}(\Omega)$. Using (6.2), we have that for all $w_h \in X_h$:

$$-\langle \Delta u, w_h\rangle_{H^{-t}(\Omega),H^t(\Omega)} = \langle f, w_h\rangle_{H^{-t}(\Omega),H^t(\Omega)} = \sum_{\ell\in\mathcal{I}_K} \langle f, w_h\rangle_{H^{-t}(K_\ell),H^t(K_\ell)}.$$

Using the integration by parts (2.13) with $D = K_\ell$, we obtain that:

$$\begin{aligned}
-\langle \Delta u, w_h\rangle_{H^{-t}(\Omega),H^t(\Omega)} &= \sum_{\ell\in\mathcal{I}_K} -\langle \Delta u, w_h\rangle_{H^{-t}(K_\ell),H^t(K_\ell)},\\
&= \sum_{\ell\in\mathcal{I}_K} (\mathbf{grad}\, u, \mathbf{grad}\, w_h)_{L^2(K_\ell)} - \sum_{\ell\in\mathcal{I}_K}\sum_{f\in\mathcal{I}_{F,\ell}} (\mathbf{grad}\, u\cdot\mathbf{n}_{f,\ell}, w_h)_{L^2(F_f)},\\
&= (u,w_h)_h - \sum_{f\in\mathcal{I}_F} (\mathbf{grad}\, u\cdot\mathbf{n}_f, [\![w_h]\!])_{L^2(F_f)}.
\end{aligned}$$

Noticing that $(\mathbf{grad}\, u\cdot\mathbf{n}_f, [\![w_h]\!])_{L^2(F_f)} = (\{\mathbf{grad}_h\, u\}\cdot\mathbf{n}_f, [\![w_h]\!])_{L^2(F_f)}$, we obtain:

$$(u,w_h)_h - \sum_{f\in\mathcal{I}_F} (\{\mathbf{grad}_h\, u\}\cdot\mathbf{n}_f, [\![w_h]\!])_{L^2(F_f)} = \sum_{\ell\in\mathcal{I}_K} \langle f, w_h\rangle_{H^{-t}(K_\ell),H^t(K_\ell)}.$$

In order to obtain a symmetric numerical scheme, noticing that $[\![u]\!]_{F_f} = 0$, $\forall f\in\mathcal{I}_F^i$, and $[\![u]\!]_{F_f} = g$, $\forall f\in\mathcal{I}_F^b$, we write equivalently:

$$\begin{aligned}
&(u,w_h)_h - \sum_{f\in\mathcal{I}_F} (\{\mathbf{grad}_h\, u\}\cdot\mathbf{n}_f, [\![w_h]\!])_{L^2(F_f)}\\
&\quad - \sum_{f\in\mathcal{I}_F} (\{\mathbf{grad}_h\, w_h\}\cdot\mathbf{n}_f, [\![u]\!])_{L^2(F_f)}\\
&\quad = \sum_{\ell\in\mathcal{I}_K} \langle f, w_h\rangle_{H^{-t}(K_\ell),H^t(K_\ell)} - \sum_{f\in\mathcal{I}_F^b} (\{\mathbf{grad}_h\, w_h\}\cdot\mathbf{n}_f, g)_{L^2(F_f)}.
\end{aligned} \tag{6.3}$$

Let us consider the continuous and symmetric bilinear form $a_h^{CS}(\cdot,\cdot)$ such that for all $(v_{\star h}, w_h)\in X_{\star h}\times X_h$:

$$\begin{aligned}
a_h^{CS}(v_{\star h}, w_h) := (v_{\star h}, w_h)_h &- \sum_{f\in\mathcal{I}_F} (\{\mathbf{grad}_h\, v_{\star h}\}\cdot\mathbf{n}_f, [\![w_h]\!])_{L^2(F_f)}\\
&- \sum_{f\in\mathcal{I}_F} (\{\mathbf{grad}_h\, w_h\}\cdot\mathbf{n}_f, [\![v_{\star h}]\!])_{L^2(F_f)}.
\end{aligned} \tag{6.4}$$

It is not possible to prove that the bilinear form $a_h^{CS}(\cdot,\cdot)$ is coercive on $V_h\times V_h$. Therefore, we introduce a stabilization term that penalizes jumps at the interfaces. For all $(v_{\star h}, w_h)\in X_{\star h}\times X_h$ we set:

$$a_h^{sip}(v_{\star h}, w_h) = a_h^{CS}(v_{\star h}, w_h) + s_h(v_{\star h}, w_h), \tag{6.5}$$

$$s_h(v_{\star h}, w_h) = \sum_{f\in\mathcal{I}_F} \frac{\eta_f}{h_f}([\![v_{\star h}]\!], [\![w_h]\!])_{L^2(F_f)}, \tag{6.6}$$

where the parameters $\eta_f > 0$ are to be determined to achieve coercivity (cf. Theorem

6.3 below), and h_f is a local length scale related to F_f. For $d \geq 2$ we set h_f to be equal to the diameter of the face F_f. For $d = 1$ we set $h_f := \min\{h_{\ell_1}, h_{\ell_2}\}$ if $f \in \mathcal{I}_{F,\ell_1} \cap \mathcal{I}_{F,\ell_2}$, and $h_f = h_\ell$ if $f \in \mathcal{I}_F^b$ and $f \in \mathcal{I}_{F,\ell}$. The exponent *sip* stands for *Symmetric Interior Penalty* method, which is the usual name of this discretization.
Let $s \in (\frac{1}{2}, 1]$ and $t = 1 - s$. Consider the linear form $\ell_{f,g} : X_h \to \mathbb{R}$ such that for all $v_h \in X_h$:

$$\begin{aligned} \ell_{f,g}(v_h) &= \langle f, v_h \rangle_{H^{-t}(\Omega), H^t(\Omega)} - \sum_{f \in \mathcal{I}_F^b} (g, \mathbf{grad}\, v_h \cdot \mathbf{n}_f)_{L^2(F_f)} \\ &\quad + \sum_{f \in \mathcal{I}_F^b} \frac{\eta_f}{h_f} (g, v_h)_{L^2(F_f)}. \end{aligned} \tag{6.7}$$

The discretization of Problem (5.22) with the SIP method reads:

$$\text{Find } u_h \in X_h \text{ such that } \forall v_h \in X_h, \quad a_h^{sip}(u_h, v_h) = \ell_{f,g}(v_h). \tag{6.8}$$

By construction, $\ell_{f,g}(v_h) = a_h^{sip}(u, v_h)$. This follows from (6.3) and from the fact that $s_h(u, v_h) = \sum_{f \in \mathcal{I}_F^b} \frac{\eta_f}{h_f} (g, v_h)_{L^2(F_f)}$.

Definition 6.3 *We define the following norms on $X_{\star h}$: for all $v_{\star h} \in X_{\star h}$,*

$$\begin{aligned} |v_{\star h}|_J^2 &:= \sum_{f \in \mathcal{I}_F} \frac{1}{h_f} \|[\![v_{\star h}]\!]\|_{L^2(F_f)}^2, \\ \|v_{\star h}\|_{sip}^2 &:= \|v_{\star h}\|_h^2 + |v_{\star h}|_J^2, \\ \|v_{\star h}\|_{sip,\star}^2 &:= \|v_{\star h}\|_{sip}^2 + \sum_{\ell \in \mathcal{I}_K} h_\ell \|\mathbf{grad}\, v_{\star h|K_\ell} \cdot \mathbf{n}_\ell\|_{L^2(\partial K_\ell)}^2. \end{aligned}$$

Theorem 6.1 *The mapping $v_h \mapsto \|v_h\|_{sip}$ is a norm on $\mathcal{P}_h H^1$.*

Proof. Let $v_h \in \mathcal{P}_h H^1$ such that $\|v_h\|_{sip} = 0$. Then, for all $\ell \in \mathcal{I}_K$, $\|\mathbf{grad}\, v_h\|_{\mathbf{L}^2(K_\ell)} = 0$, implying that v_h is piecewise constant. Since $|v_h|_J = 0$ as well, jumps at interfaces and boundaries vanish, leading to $v_h = 0$.

□

Remark 6.4 *For all $v \in H_0^1(\Omega)$, $\|v\|_{sip} = \|v\|_{H_0^1(\Omega)}$.*

Remark 6.5 *The coercivity of the bilinear form $a_h^{sip}(\cdot, \cdot)$ depends on the shape-regularity parameter σ and on the jump penalisation parameters η_f. From a numerical viewpoint,*

this penalisation plays a crucial role in ensuring the stability of the method. At the same time, this penalisation should not be too strong in order to preserve the advantages of the discontinuous Galerkin scheme. Therefore, we aim at optimising the choice of the parameters η_f. Lemma 6.2 below focuses on the bound which allows this optimisation.

For all $\ell \in \mathcal{I}_K$, we let $k_\ell \leq k$ be the local polynomial approximation order in K_ℓ, and we set:

$$\tilde{\sigma}_\ell = \tilde{C}_{d,k_\ell-1}\sigma_\ell = 2\,k_\ell\,(k_\ell - 1 + d)\,\sigma_\ell. \tag{6.9}$$

Remark 6.6 *Since $k_\ell \geq 1$, for all $\ell \in \mathcal{I}_K$, $\tilde{C}_{d,k_\ell-1} \geq 2$. Hence, $\tilde{\sigma}_\ell \geq 2$.*

Lemma 6.2 *For all $(v_h, w_h) \in X_h \times X_h$, we have:*

$$\begin{aligned} &\left| \sum_{f\in\mathcal{I}_F} (\,\{\mathbf{grad}_h\, v_h\}\cdot\mathbf{n}_f, [\![w_h]\!])_{L^2(F_f)} \right| \\ &\leq \sum_{\ell\in\mathcal{I}_K} \frac{\tilde{\sigma}_\ell}{2\,\epsilon_\ell}\,\|\mathbf{grad}\, v_h\|^2_{\mathbf{L}^2(K_\ell)} + \sum_{f\in\mathcal{I}_F} \frac{\epsilon_f}{2}(h_f)^{-1}\|[\![w_h]\!]\|^2_{L^2(F_f)}, \end{aligned} \tag{6.10}$$

where for all $f \in \mathcal{I}_F$, $\epsilon_f > 0$ and $\epsilon_\ell = \min\limits_{f\in\mathcal{I}_{F,\ell}} (\epsilon_f\,\mathrm{card}(\mathcal{I}_{K,f}))$.

Proof. Using Young's inequality, we have, for all $f \in \mathcal{I}_F$, $\epsilon_f > 0$:

$$\begin{aligned} &\left|(\,\{\mathbf{grad}_h\, v_h\}\cdot\mathbf{n}_f, [\![w_h]\!])_{L^2(F_f)}\right| \\ &\quad= \left|((h_f)^{\frac{1}{2}}\ \{\mathbf{grad}_h\, v_h\}\cdot\mathbf{n}_f, (h_f)^{-\frac{1}{2}}\,[\![w_h]\!])_{L^2(F_f)}\right| \\ &\quad\leq \frac{1}{2\epsilon_f}\,h_f\,\|\,\{\mathbf{grad}_h\, v_h\}\cdot\mathbf{n}_f\|^2_{L^2(F_f)} + \frac{\epsilon_f}{2}\,(h_f)^{-1}\|[\![w_h]\!]\|^2_{L^2(F_f)}. \end{aligned} \tag{6.11}$$

Let $\delta_f = 2$ if $f \in \mathcal{I}_F^i$, and $\delta_f = 1$ if $f \in \mathcal{I}_F^b$ (i.e. $\delta_f = \mathrm{card}(\mathcal{I}_{K,f})$). It holds:

$$\begin{aligned} \|\,\{\mathbf{grad}_h\, v_h\}\cdot\mathbf{n}_f\|^2_{L^2(F_f)} &\leq \|\,\{\mathbf{grad}_h\, v_h\}\,\|^2_{\mathbf{L}^2(F_f)}, \\ &\leq \frac{1}{\delta_f}\sum_{\ell\in\mathcal{I}_{K,f}} \|\mathbf{grad}\, v_{h|K_\ell}\|^2_{\mathbf{L}^2(F_f)}. \end{aligned}$$

For all $\ell \in \mathcal{I}_K$, we let $\epsilon_\ell := \min\limits_{f\in\mathcal{I}_{F,\ell}} \epsilon_f\,\delta_f$. Notice that for all $f \in \mathcal{I}_F$, $\ell \in \mathcal{I}_{K,f}$, $h_f \leq h_\ell$.

Hence, we have:

$$\begin{aligned}\sum_{f\in\mathcal{I}_F}\frac{h_f}{\epsilon_f}\|\{\mathbf{grad}_h v_h\}\cdot\mathbf{n}_f\|^2_{L^2(F_f)} &\le \sum_{f\in\mathcal{I}_F}\frac{h_f}{\delta_f\,\epsilon_f}\sum_{\ell\in\mathcal{I}_{K,f}}\|\mathbf{grad}\, v_h\|^2_{\mathbf{L}^2(F_f)},\\ &\le \sum_{f\in\mathcal{I}_F}\sum_{\ell\in\mathcal{I}_{K,f}}\frac{h_\ell}{\epsilon_\ell}\|\mathbf{grad}\, v_h\|^2_{\mathbf{L}^2(F_f)},\\ &\le \sum_{\ell\in\mathcal{I}_K}\frac{h_\ell}{\epsilon_\ell}\sum_{f\in\mathcal{I}_{F,\ell}}\|\mathbf{grad}\, v_h\|^2_{\mathbf{L}^2(F_f)},\\ &\le \sum_{\ell\in\mathcal{I}_K}\frac{h_\ell}{\epsilon_\ell}\|\mathbf{grad}\, v_h\|^2_{\mathbf{L}^2(\partial K_\ell)}.\end{aligned}$$

Using Equation (4.12), we deduce that:

$$\sum_{f\in\mathcal{I}_F}\frac{h_f}{\epsilon_f}\|\{\mathbf{grad}_h v_h\}\cdot\mathbf{n}_f\|^2_{L^2(F_f)} \le \tilde{C}_{d,k_\ell-1}\sum_{\ell\in\mathcal{I}_K}\frac{\sigma_\ell}{\epsilon_\ell}\|\mathbf{grad}\, v_h\|^2_{\mathbf{L}^2(K_\ell)}. \tag{6.12}$$

Using (6.9), summing (6.11) over $f\in\mathcal{I}_F$, we deduce (6.10).

□

We can now prove the following stability Theorem.

Theorem 6.3 (`Stability, [40, Lem. 4.12 and 4.16]`) *Let us set the stability parameters $(\eta_f)_{f\in\mathcal{I}_F}$ so that*

$$\forall f\in\mathcal{I}_F,\quad \eta_f = \frac{5/2}{\mathrm{card}(\mathcal{I}_{K,f})}\max_{\ell\in\mathcal{I}_{K,f}}\tilde{\sigma}_\ell, \tag{6.13}$$

where $\tilde{\sigma}_\ell = 2\,k_\ell\,(k_\ell-1+d)\,\sigma_\ell$. Then the bilinear form $a_h^{sip}(\cdot,\cdot)$ is coercive and continuous on $X_h\times X_h$:

$$\forall v_h\in X_h,\quad a_h^{sip}(v_h,v_h) \ge \frac{1}{2}\|v_h\|^2_{sip}, \tag{6.14}$$

$$\forall (v_h,w_h)\in X_h\times X_h,\quad \left|a_h^{sip}(v_h,w_h)\right| \le C_{bnd}\|v_h\|_{sip}\,\|w_h\|_{sip}, \tag{6.15}$$

with $C_{bnd} = \max_{f\in\mathcal{I}_F}\eta_f + \max_{\ell\in\mathcal{I}_K}(2\,\tilde{\sigma}_\ell)^{\frac{1}{2}} \lesssim \sigma$.

PROOF. Recall that

$$\begin{aligned} a_h^{sip}(v_h, w_h) &= (\mathbf{grad}\, v_h, \mathbf{grad}\, w_h)_{\mathbf{L}^2(\Omega)} \\ &\quad - \sum_{f\in\mathcal{I}_F} (\{\mathbf{grad}_h\, v_h\}\cdot\mathbf{n}_f, [\![w_h]\!])_{L^2(F_f)} \\ &\quad - \sum_{f\in\mathcal{I}_F} (\{\mathbf{grad}_h\, w_h\}\cdot\mathbf{n}_f, [\![v_h]\!])_{L^2(F_f)} \\ &\quad + \sum_{f\in\mathcal{I}_F} \frac{\eta_f}{h_f}([\![v_h]\!], [\![w_h]\!])_{L^2(F_f)}. \end{aligned} \tag{6.16}$$

Let us prove (6.14). We have:

$$\begin{aligned} a_h^{sip}(v_h, v_h) &= \|\mathbf{grad}_h v_h\|^2_{\mathbf{L}^2(\Omega)} + \sum_{f\in\mathcal{I}_F} \frac{\eta_f}{h_f}\|[\![v_h]\!]\|^2_{L^2(F_f)} \\ &\quad -2\sum_{f\in\mathcal{I}_F} (\{\mathbf{grad}_h\, v_h\}\cdot\mathbf{n}_f, [\![v_h]\!])_{L^2(F_f)}. \end{aligned}$$

Using Lemma 6.2 with $w_h = v_h$, we have:

$$\begin{aligned} a_h^{sip}(v_h, v_h) &\geq \sum_{\ell\in\mathcal{I}_K} \left(1 - \frac{\tilde{\sigma}_\ell}{\epsilon_\ell}\right) \|\mathbf{grad}\, v_h\|^2_{\mathbf{L}^2(K_\ell)} \\ &\quad + \sum_{f\in\mathcal{I}_F} \frac{\eta_f - \epsilon_f}{h_f}\|[\![v_h]\!]\|^2_{L^2(F_f)}. \end{aligned} \tag{6.17}$$

To ensure coercivity, it is necessary that the parameters $(\eta_f)_{f\in\mathcal{I}_F}$ satisfy:

$$\forall \ell \in \mathcal{I}_K,\ \epsilon_\ell > \tilde{\sigma}_\ell \text{ and } \forall f \in \mathcal{I}_F,\ \eta_f > \epsilon_f. \tag{6.18}$$

Recall that $\epsilon_\ell = \min_{f\in\mathcal{I}_{F,\ell}} \epsilon_f\, \delta_f$. Let us choose

$$\epsilon_f = \frac{\alpha^{-1}}{\delta_f} \max_{\ell'\in\mathcal{I}_{K,f}} \tilde{\sigma}_{\ell'} \text{ with } \alpha < 1.$$

Then, $\epsilon_\ell = \alpha^{-1} \min_{f\in\mathcal{I}_{F,\ell}} \max_{\ell'\in\mathcal{I}_{K,f}} \tilde{\sigma}_{\ell'} \geq \alpha^{-1}\tilde{\sigma}_\ell$ and $1 - \frac{\tilde{\sigma}_\ell}{\epsilon_\ell} \geq 1 - \alpha > 0$. Let

$$\eta_f = \frac{\beta}{\delta_f} \max_{\ell\in\mathcal{I}_{K,f}} \tilde{\sigma}_\ell \geq 1, \text{ where } \beta > \alpha^{-1}.$$

We get (using Remark 6.6 for the last bound):

$$\eta_f - \epsilon_f = \frac{\beta - \alpha^{-1}}{\delta_f} \max_{\ell \in \mathcal{I}_{K,f}} \tilde{\sigma}_\ell \geq 2\, \frac{\beta - \alpha^{-1}}{\delta_f} > 0.$$

Choosing $\alpha = \frac{1}{2}$, $\beta = \frac{5}{2}$, we have:

$$1 - \frac{\tilde{\sigma}_\ell}{\epsilon_\ell} \geq \frac{1}{2} \text{ and } \eta_f - \epsilon_f > \frac{1/2}{\delta_f} \max_{\ell \in \mathcal{I}_{K,f}} \tilde{\sigma}_\ell \geq \frac{1}{\delta_f} \geq \frac{1}{2}.$$

Using these bounds in (6.17), we deduce (6.14).

Let us prove (6.15). Using the Cauchy-Schwarz inequality to bound (6.16), we have:

$$\begin{aligned}
\left|a_h^{sip}(v_h, w_h)\right| \;\leq\; & \|\mathbf{grad}_h v_h\|_{\mathbf{L}^2(\Omega)}\, \|\mathbf{grad}_h w_h\|_{\mathbf{L}^2(\Omega)} \\
& + \sum_{f \in \mathcal{I}_f} \eta_f\, (h_f)^{-1} \|[\![v_h]\!]\|_{L^2(K_\ell)}\, \|[\![w_h]\!]\|_{L^2(K_\ell)} \\
& + \left(\sum_{f \in \mathcal{I}_F} h_f \,\|\, \{\mathbf{grad}_h\, v_h\} \cdot \mathbf{n}_f \|^2_{L^2(F_f)} \right)^{\frac{1}{2}} |w_h|_J \\
& + \left(\sum_{f \in \mathcal{I}_F} h_f \,\|\, \{\mathbf{grad}_h\, w_h\} \cdot \mathbf{n}_f \|^2_{L^2(F_f)} \right)^{\frac{1}{2}} |v_h|_J.
\end{aligned}$$

Using (6.12), with $\epsilon_f = 1$ and $\epsilon_\ell = \min_{f \in \mathcal{I}_{F,\ell}} \delta_f$ here, we deduce that:

$$\begin{aligned}
\left|a_h^{sip}(v_h, w_h)\right| \;\leq\; & \|v_h\|_h\, \|w_h\|_h + \max_{f \in \mathcal{I}_F} \eta_f |v_h|_J\, |w|_J \\
& + \left(\sum_{\ell \in \mathcal{I}_K} \frac{\tilde{\sigma}_\ell}{\epsilon_\ell} \|\mathbf{grad}\, v_h\|^2_{\mathbf{L}^2(K_\ell)} \right)^{\frac{1}{2}} |w_h|_J \\
& + \left(\sum_{\ell \in \mathcal{I}_K} \frac{\tilde{\sigma}_\ell}{\epsilon_\ell} \|\mathbf{grad}\, w_h\|^2_{\mathbf{L}^2(K_\ell)} \right)^{\frac{1}{2}} |v_h|_J.
\end{aligned}$$

Notice that $\max_{f \in \mathcal{I}_F} \eta_f > 1$, and $(\epsilon_\ell)^{-1} < 2$ here. Hence:

$$\begin{aligned}
\left|a_h^{sip}(v_h, w_h)\right| \;\leq\; & \max_{f \in \mathcal{I}_F} \eta_f \left(\|v_h\|_h\, \|w_h\|_h + |v_h|_J\, |w|_J \right) \\
& + \max_{\ell \in \mathcal{I}_K} \left(2\, \tilde{\sigma}_\ell\right)^{\frac{1}{2}} \left(\|v_h\|_h\, |w_h|_J + \|w_h\|_h\, |v_h|_J \right).
\end{aligned}$$

Using that $a\,b + c\,d \leq (a^2 + c^2)^{\frac{1}{2}}\,(b^2 + d^2)^{\frac{1}{2}}$, we have:

$$\left|a_h^{sip}(v_h, w_h)\right| \leq \left(\max_{f\in\mathcal{I}_F} \eta_f + \max_{\ell\in\mathcal{I}_K} (2\,\tilde{\sigma}_\ell)^{\frac{1}{2}}\right) \|v_h\|_{sip}\,\|w_h\|_{sip}.$$

□

Theorem 6.4 *The bilinear form $\ell_{f,g}(\cdot)$ is continuous in the $\|\cdot\|_{sip}$ norm.*

Proof. Let $v_h \in X_h$. As seen before, we have $\ell_{f,g}(v_h) = a_h^{sip}(u, v_h)$. Therefore, it holds, using the Cauchy-Schwarz inequality:

$$\begin{aligned} |\ell_{f,g}(v_h)| &= \left|a_h^{sip}(u, v_h)\right| = \left|(u, v_h)_h - \sum_{f\in\mathcal{I}_F} (\mathbf{grad}\, u\cdot\mathbf{n}_f, [\![v_h]\!])_{L^2(F_f)}\right|, \\ &\leq |u|_{H^1(\Omega)}\,\|v_h\|_h + \left(\sum_{f\in\mathcal{I}_F} h_f\,\|\mathbf{grad}\, u\cdot\mathbf{n}_f\|^2_{\mathbf{L}^2(F_f)}\right)^{1/2} |v_h|_J. \end{aligned}$$

We deduce from (4.16) that:

$$h_f\,\|\mathbf{grad}\, u\cdot\mathbf{n}_f\|^2_{\mathbf{L}^2(F_f)} \leq h_\ell\,\|\mathbf{grad}\, u\|^2_{\mathbf{L}^2(F_f)} \lesssim \sigma^{1+d}\,\|\mathbf{grad}\, u\|^2_{\mathbf{H}^s(K_\ell)}.$$

By summation we get:

$$|\ell_{f,g}(v_h)| \lesssim |u|_{H^1(\Omega)}\,\|v_h\|_h + \sigma^{\frac{d+1}{2}}\,|u|_{H^{1+s}(\Omega)}\,|v_h|_J \lesssim \sigma^{\frac{d+1}{2}}\,|u|_{H^{1+s}(\Omega)}\,\|v_h\|_{sip},$$

which leads to the result.

□

6.2 Generalization to polytopal meshes

We aim to generalize inequality (4.15) to general meshes, provided that suitable regularity assumptions hold. We recall some useful definitions:

Definition 6.7 (`Matching simplicial mesh, [40, Definition 1.36]`) *We say that $(\mathcal{T}_h)_h$ is a* matching simplicial *mesh if it is a simplicial mesh and if for any $T \in \mathcal{T}_h$*

with vertices $(S_i)_{i=0}^d$*, the set* $\partial T \cap \partial T'$*, for any* $T' \in \mathcal{T}_h$*,* $T' \neq T$*, is the convex hull of a (possibly empty) subset of* $(S_i)_{i=0}^d$*.*

Definition 6.8 (`Matching simplicial submesh, [40, Definition 1.37]`) *Let* $\tilde{\mathcal{T}}_{\tilde{h}}$ *be a general mesh, of facets* $\tilde{\mathcal{F}}_{\tilde{h}}$ [1]*. We say that* $\mathcal{T}_{h(\tilde{h})}$ *is a matching simplicial submesh of* $\tilde{\mathcal{T}}_{\tilde{h}}$ *if*

(i) $\mathcal{T}_{h(\tilde{h})}$ *is a matching simplicial mesh,*

(ii) For all $T \in \mathcal{T}_{h(\tilde{h})}$*, there is only one* $K \in \tilde{\mathcal{T}}_{\tilde{h}}$ *such that* $T \subset K$*,*

(iii) For all $F \in \mathcal{F}_{h(\tilde{h})}$*, the set collecting the mesh faces of* $\mathcal{T}_{h(\tilde{h})}$*, there is at most one* $\tilde{F} \in \tilde{\mathcal{F}}_{\tilde{h}}$ *such that* $F \subset \tilde{F}$*.*

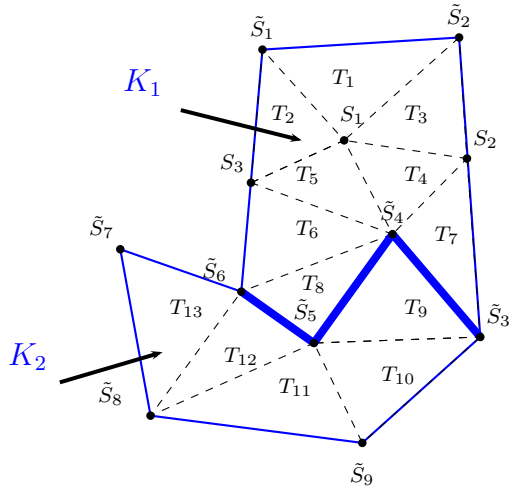


Figure 6.1 – [40, Fig. 1.5]. K_1 and K_2 are two polygonal elements, sharing three faces (in bold). The boundaries of their corresponding subelements, are indicated by dashed lines

Let $\tilde{\mathcal{T}}_{\tilde{h}}$ be a general mesh and let $\mathcal{T}_{h(\tilde{h})}$ be a matching simplicial submesh of $\tilde{\mathcal{T}}_{\tilde{h}}$. We use the following notations:

— $\tilde{\mathcal{I}}_K$ (resp. $\mathcal{I}_T$) is the index set of elements, such that

$$\tilde{\mathcal{T}}_{\tilde{h}} = \bigcup_{\tilde{\ell} \in \tilde{\mathcal{I}}_K} K_{\tilde{\ell}} \quad (\text{resp. } \mathcal{T}_{h(\tilde{h})} = \bigcup_{\ell \in \mathcal{I}_T} T_\ell).$$

1. These facets correspond to edges when $d = 2$, or polygons when $d = 3$.

$\forall \tilde{\ell} \in \tilde{\mathcal{I}}_K$, $\tilde{h}_{\tilde{\ell}}$ is the diameter of $K_{\tilde{\ell}}$.

$\forall \ell \in \mathcal{I}_T$, h_ℓ (resp. ρ_ℓ) is the diameter of T_ℓ (resp. its inscribed ball), and we let $\sigma_\ell := h_\ell \, (\rho_\ell)^{-1}$.

— $\tilde{\mathcal{I}}_{\tilde{F}}$ (resp. $\mathcal{I}_F$) is the index set of facets, such that

$$\tilde{\mathcal{F}}_{\tilde{h}} = \bigcup_{\tilde{f} \in \tilde{\mathcal{I}}_{\tilde{F}}} \tilde{F}_{\tilde{f}} \quad (\text{resp. } \mathcal{F}_{h(\tilde{h})} = \bigcup_{f \in \mathcal{I}_F} F_f).$$

Remark 6.9 *With a polytopal mesh, nonconforming polytopes may occur. The set is defined as*

$$\tilde{\mathcal{F}}_h = \left\{ \bigcup_{\tilde{\ell} \in \tilde{\mathcal{I}}_K} \partial K_{\tilde{\ell}} \cap \partial\Omega \right\} \bigcup \left\{ \bigcup_{\substack{\tilde{\ell},\tilde{\ell}' \in \tilde{\mathcal{I}}_K \\ \tilde{\ell} < \tilde{\ell}'}} \partial K_{\tilde{\ell}} \cap \partial K_{\tilde{\ell}'} \right\}.$$

Let $\tilde{\mathcal{I}}_{\tilde{F}}$ denote the index set associated with $\tilde{\mathcal{F}}_h$. For simplicity, we illustrate an example in 2D in Figure 6.2, where $\tilde{\mathcal{F}}_h = \{\tilde{F}_i\}_{i=1}^{7} \cup \{\tilde{F}_j\}_{j=8}^{10}$.

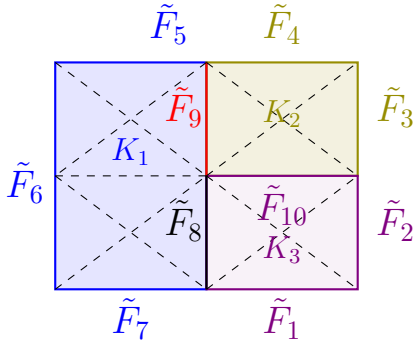


Figure 6.2 – K_1, K_2 and K_3 are three polygonal elements nonconforming, sharing three faces.

For all $\tilde{\ell} \in \tilde{\mathcal{I}}_K$, we set:

$$\tilde{\mathcal{I}}_{\tilde{F},\tilde{\ell}} := \{\tilde{f} \in \tilde{\mathcal{I}}_{\tilde{F}} \,|\, \tilde{F}_{\tilde{f}} \subset \partial K_{\tilde{\ell}}\}, \quad \mathcal{I}_{T,\tilde{\ell}} := \{\ell \in \mathcal{I}_T \,|\, T_\ell \subset K_{\tilde{\ell}}\}$$

For all $\tilde{f} \in \tilde{\mathcal{I}}_{\tilde{F}}$, we set:

$$\mathcal{I}_{F,\tilde{f}} := \{f \in \mathcal{I}_F \,|\, F_f \subset \tilde{F}_{\tilde{f}}\}.$$

For all $\tilde{\ell} \in \tilde{\mathcal{I}}_K$, for all $\tilde{f} \in \tilde{\mathcal{I}}_{\tilde{F},\tilde{\ell}}$, we set:

$$\mathcal{I}_{T,\tilde{\ell},\tilde{f}} := \{\ell \in \mathcal{I}_{T,\tilde{\ell}} \,|\, \mathring{\tilde{F}}_{\tilde{f}} \cap \partial T_\ell \neq \emptyset\}.$$

Definition 6.10 (`Shape and contact regularity, [40, Definition 1.38]`) *We say that the mesh sequence $(\tilde{\mathcal{T}}_{\tilde{h}})_{\tilde{h}}$ is* shape- and contact-regular *if for all $\tilde{h}$, $\tilde{\mathcal{T}}_{\tilde{h}}$ admits a matching simplicial submesh $\mathcal{T}_{h(\tilde{h})}$ such that*

(i) The mesh sequence $(\mathcal{T}_{h(\tilde{h})})_{\tilde{h}}$ is shape-regular in the sense of Ciarlet [23]: there exists a constant $\sigma > 0$, independent of $\tilde{h}$, such that, for all $\tilde{\ell} \in \tilde{\mathcal{I}}_K$, for all $\ell \in \mathcal{I}_{T,\tilde{\ell}}$:

$$\sigma_\ell \leq \sigma_{\tilde{\ell}} \leq \sigma.$$

(ii) For all $\tilde{\ell} \in \tilde{\mathcal{I}}_K$, there exists a constant $\tilde{\sigma}_{\tilde{\ell}}$, such that, for all $\tilde{\ell} \in \tilde{\mathcal{I}}_K$, for all $\ell \in \mathcal{I}_{T,\tilde{\ell}}$,

$$\tilde{h}_{\tilde{\ell}} \leq \tilde{\sigma}_{\tilde{\ell}}\, h_\ell. \tag{6.19}$$

We suppose that there exists a constant $\tilde{\sigma}$, independent of $\tilde{h}$, such that $\forall \tilde{\ell} \in \tilde{\mathcal{I}}_K$,

$$\tilde{\sigma}_{\tilde{\ell}} \leq \tilde{\sigma}.$$

Consider a *shape- and contact-regular* mesh sequence $(\tilde{\mathcal{T}}_{\tilde{h}})_{\tilde{h}}$. Fixing $\tilde{h}$, we can now prove the following proposition:

Proposition 6.11 *For all $\tilde{\ell} \in \tilde{\mathcal{I}}_K$, for all $v \in P^k(K_{\tilde{\ell}})$:*

$$\|v\|_{L^2(\partial K_{\tilde{\ell}})} \leq (\hat{C}_{d,k}\, \tilde{\sigma}_{\tilde{\ell}}\, \sigma_{\tilde{\ell}})^{\frac{1}{2}}\, (\tilde{h}_{\tilde{\ell}})^{-\frac{1}{2}}\, \|v\|_{L^2(K_{\tilde{\ell}})}. \tag{6.20}$$

Proof. Let $\tilde{\ell} \in \tilde{\mathcal{I}}_K$, $v \in P^k(K_{\tilde{\ell}})$. We have:

$$\|v\|^2_{L^2(\partial K_{\tilde{\ell}})} = \sum_{\tilde{f} \in \tilde{\mathcal{I}}_{\tilde{F},\tilde{\ell}}} \|v\|^2_{L^2(\tilde{F}_{\tilde{f}})} = \sum_{\tilde{f} \in \tilde{\mathcal{I}}_{\tilde{F},\tilde{\ell}}} \sum_{f \in \mathcal{I}_{F,\tilde{f}}} \|v\|^2_{L^2(F_f)}.$$

Using inequalities (4.15) and (6.19), we obtain:

$$\begin{aligned}\sum_{\tilde f\in\tilde{\mathcal I}_{\tilde F,\tilde\ell}}\sum_{f\in\mathcal I_{F,\tilde f}}\|v\|^2_{L^2(F_f)} &\le \hat C_{d,k}\sum_{\tilde f\in\tilde{\mathcal I}_{\tilde F,\tilde\ell}}\sum_{\ell\in\mathcal I_{T,\tilde\ell,\tilde f}}\sigma_\ell\,(h_\ell)^{-1}\|v\|^2_{L^2(T_\ell)},\\ &\le \hat C_{d,k}\,\sigma_{\tilde\ell}\,\tilde\sigma_{\tilde\ell}\,(\tilde h_{\tilde\ell})^{-1}\sum_{\tilde f\in\tilde{\mathcal I}_{\tilde F,\tilde\ell}}\sum_{\ell\in\mathcal I_{T,\tilde\ell,\tilde f}}\|v\|^2_{L^2(T_\ell)},\\ &\le \hat C_{d,k}\,\sigma_{\tilde\ell}\,\tilde\sigma_{\tilde\ell}\,(\tilde h_{\tilde\ell})^{-1}\,\|v\|^2_{L^2(K_{\tilde\ell})}.\end{aligned}$$

□

Following the proof of Lemma 6.2 and using Proposition 6.11, one can prove the following Lemma:

Lemma 6.5 *For all $(v_h,w_h)\in X_h\times X_h$, we have:*

$$\begin{aligned}&\left|\sum_{\tilde f\in\tilde{\mathcal I}_{\tilde F}}(\{\mathbf{grad}_h\,v_h\}\cdot\mathbf{n}_{\tilde f},[\![w_h]\!])_{L^2(\tilde F_{\tilde f})}\right|\\ &\quad\le\sum_{\tilde f\in\tilde{\mathcal I}_{\tilde F}}\frac{1}{2\,\epsilon_{\tilde f}}\,h_{\tilde f}\,\|\,\{\mathbf{grad}_h\,v_h\}\,\|^2_{\mathbf{L}^2(\tilde F_{\tilde f})}+\sum_{\tilde f\in\tilde{\mathcal I}_{\tilde F}}\frac{\epsilon_{\tilde f}}{2}(h_{\tilde f})^{-1}\|[\![w_h]\!]\|^2_{L^2(\tilde F_{\tilde f})}\\ &\quad\le\sum_{\tilde\ell\in\tilde{\mathcal I}_{\tilde K}}\frac{\hat C_{d,k-1}\,\sigma_{\tilde\ell}\,\tilde\sigma_{\tilde\ell}}{2\,\epsilon_{\tilde\ell}}\,\|\mathbf{grad}\,v_h\|^2_{\mathbf{L}^2(K_{\tilde\ell})}\\ &\qquad+\sum_{\tilde f\in\tilde{\mathcal I}_{\tilde F}}\frac{\epsilon_{\tilde f}}{2}(h_{\tilde f})^{-1}\|[\![w_h]\!]\|^2_{L^2(\tilde F_{\tilde f})},\end{aligned}\tag{6.21}$$

where for all $\tilde f\in\tilde{\mathcal I}_{\tilde F}$, $\epsilon_{\tilde f}>0$ and $\epsilon_{\tilde\ell}=\min\limits_{\tilde f\in\mathcal I_{\tilde F,\tilde\ell}}\Big(\epsilon_{\tilde f}\,\mathrm{card}(\tilde{\mathcal I}_{K,\tilde f})\Big)$.

We can now generalize Theorem 6.3 in the case where the elements of the mesh are polytopal.

Theorem 6.6 (`Stability, general mesh [40, Lem. 4.12 and 4.16]`) *Let us set the stability parameters $(\eta_{\tilde f})_{\tilde f\in\tilde{\mathcal I}_{\tilde F}}$ so that*

$$\forall\tilde f\in\tilde{\mathcal I}_{\tilde F},\quad \eta_{\tilde f}=\frac{5/2}{\mathrm{card}(\tilde{\mathcal I}_{K,\tilde f})}\max_{\tilde\ell\in\tilde{\mathcal I}_{K,\tilde f}}\tilde{\tilde\sigma}_{\tilde\ell},\tag{6.22}$$

where $\tilde{\tilde\sigma}_{\tilde\ell}=\hat C_{d,k_{\tilde\ell}-1}\,\sigma_{\tilde\ell}\,\tilde\sigma_{\tilde\ell}$. Then the bilinear form $a_h^{sip}(\cdot,\cdot)$ is coercive and continuous on

$X_h \times X_h$:

$$\forall v_h \in X_h, \quad a_h^{sip}(v_h, v_h) \geq \frac{1}{2} \|v_h\|_{sip}^2, \tag{6.23}$$

$$\forall (v_h, w_h) \in X_h \times X_h, \quad \left| a_h^{sip}(v_h, w_h) \right| \leq C_{bnd} \|v_h\|_{sip} \|w_h\|_{sip}, \tag{6.24}$$

with $\tilde{C}_{bnd} = \max_{\tilde{f} \in \tilde{\mathcal{I}}_{\tilde{F}}} \eta_{\tilde{f}} + \max_{\tilde{\ell} \in \tilde{\mathcal{I}}_K} \left(2\, \tilde{\tilde{\sigma}}_{\tilde{\ell}} \right)^{\frac{1}{2}} \lesssim \sigma\, \tilde{\sigma}$.

Proof. We proceed as in the proof of Theorem 6.3, using Lemma 6.5.

□

Let us study the discretization of the Stokes Problem.

6.3 Discretization of the Stokes problem

Let $k_{\mathbf{u}} > 0$ (resp. $k_p \geq 0$) be the discretization order of $\mathbf{u}$ (resp. of p). In what follows, we suppose that there exists $s \in (\frac{1}{2}, 1]$ such that $(\mathbf{u}, p) \in \mathbf{H}^{1+s}(\Omega) \times (L^2_{zmv} \cap H^s(\Omega))$, $\mathbf{g} \in \mathbf{H}^{\frac{1}{2}+s}(\partial\Omega)$ and we let $t = 1 - s \in [0, 1/2)$. Recall that from Theorem 2.5, it holds: $(\mathbf{H}^t(\Omega))' = \mathbf{H}^{-t}(\Omega)$. Recall that $\Delta\mathbf{u}$ and $\mathbf{grad}\, p$ belong to $\mathbf{H}^{-t}(\Omega)$.
We define the following discrete spaces: $\mathbf{X}_h := [P^{k_{\mathbf{u}}}_{disc}(\mathcal{T}_h)]^d$, $\mathbf{X}_{\star h} := \mathbf{X}_h + \mathbf{H}^{1+s}(\Omega)$, and $L_h := L^2_{zmv}(\Omega) \cap P^{k_p}_{disc}(\mathcal{T}_h)$, $L_{\star h} := L_h + L^2_{zmv}(\Omega) \cap H^s(\Omega)$. Recall that from Theorem 2.5, we have $\mathbf{X}_h \subset \mathbf{H}^t(\Omega)$.
The bilinear form related to the discretization of the vector Laplacian operator is expressed as:

$$\forall (\mathbf{w}, \mathbf{v}_h) \in \mathbf{X}_{\star h} \times \mathbf{X}_h; \quad a_h(\mathbf{w}, \mathbf{v}_h) = \sum_{i=1}^{d} a_h^{sip}(w_i, v_{h,i}). \tag{6.25}$$

As a corollary of equalities (2.13)-(2.14), we have the following result

Theorem 6.7 (`Integration by parts`) *Let $\ell \in \mathcal{I}_K$. Let $s \in (\frac{1}{2}, 1)$ and $t = s - 1$. For all $(\mathbf{v}, q) \in \mathbf{H}^{1+s}(K_\ell) \times H^s(K_\ell)$ and $\mathbf{w} \in \mathbf{H}^t(K_\ell)$, the following integrations by parts hold:*

$$\begin{aligned} -\langle \Delta\mathbf{v}, \mathbf{w} \rangle_{\mathbf{H}^{-t}(K_\ell), \mathbf{H}^t(K_\ell)} &= (\mathbf{Grad}\, \mathbf{v}, \mathbf{Grad}\, \mathbf{w})_{\mathbb{L}^2(K_\ell)} - (\mathbf{Grad}\, \mathbf{v} : \mathbf{n}_\ell, \mathbf{w})_{\mathbf{L}^2(\partial K_\ell)}, \\ \langle \mathbf{grad}\, q, \mathbf{v} \rangle_{\mathbf{H}^{-t}(K_\ell), \mathbf{H}^t(K_\ell)} &= -(q, \operatorname{div} \mathbf{v})_{L^2(K_\ell)} + (q, \mathbf{v} \cdot \mathbf{n}_\ell)_{L^2(\partial K_\ell)}. \end{aligned}$$

Hence, for all $\mathbf{v}_h \in \mathbf{X}_h$, it holds:

$$\begin{aligned}\langle \mathbf{grad}\, p, \mathbf{v}_h\rangle_{\mathbf{H}^{-t}(\Omega),\mathbf{H}^t(\Omega)} &= \sum_{\ell\in\mathcal{I}_K} \langle \mathbf{grad}\, p, \mathbf{v}_h\rangle_{\mathbf{H}^{-t}(K_\ell),\mathbf{H}^t(K_\ell)},\\ &= -(\mathrm{div}_h\, \mathbf{v}_h, p)_{L^2(\Omega)} + \sum_{f\in\mathcal{I}_F} (\llbracket \mathbf{v}_h\rrbracket\cdot\mathbf{n}_f,\, \{p\})_{L^2(F_f)}.\end{aligned}$$

Moreover, for all $q_h \in L_h$:

$$\begin{aligned}(\mathrm{div}\,\mathbf{u}, q_h)_{L^2(\Omega)} &= \sum_{\ell\in\mathcal{I}_K} (\mathrm{div}\,\mathbf{u}, q_h)_{L^2(K_\ell)},\\ &= -(\mathbf{grad}_h q_h, \mathbf{u})_{\mathbf{L}^2(\Omega)} + \sum_{f\in\mathcal{I}_F^i} (\,\{\mathbf{u}\}\cdot\mathbf{n}_f, \llbracket q_h\rrbracket)_{L^2(F_f)}.\end{aligned}$$

We can remark that, for all $\mathbf{v} \in \mathbf{X}_{\star h}$, and for all $\ell \in \mathcal{I}_K$, we have $\mathbf{v}_{|K_\ell} \in H^{1+s}(K_\ell)$. We also have that for all $q \in L_{\star h}$ and all $\ell \in \mathcal{I}_K$, then $q_{|K_\ell} \in H^s(K_\ell)$. We then define the bilinear form $b_h(\cdot,\cdot)$ such that for all $(\mathbf{v}, q) \in \mathbf{X}_{\star h} \times L_{\star h}$:

$$\begin{aligned}b_h(\mathbf{v}, q) :=\ & -(\mathrm{div}_h\,\mathbf{v}, q)_{L^2(\Omega)} + \sum_{f\in\mathcal{I}_F} (\llbracket\mathbf{v}\rrbracket\cdot\mathbf{n}_f,\, \{q\})_{L^2(F_f)},\\ =\ & -\sum_{\ell\in\mathcal{I}_K} (\mathrm{div}\,\mathbf{v}, q)_{L^2(K_\ell)} + \sum_{f\in\mathcal{I}_F} (\llbracket\mathbf{v}\rrbracket\cdot\mathbf{n}_f,\, \{q\})_{L^2(F_f)},\\ =\ & \sum_{\ell\in\mathcal{I}_K} \langle \mathbf{grad}\, q, \mathbf{v}\rangle_{\mathbf{H}^{-t}(K_\ell),\mathbf{H}^t(K_\ell)} - \sum_{f\in\mathcal{I}_F} \int_{F_f} \llbracket \mathbf{v}\cdot\mathbf{n}_f\, q\rrbracket\\ & + \sum_{f\in\mathcal{I}_F} (\llbracket\mathbf{v}\rrbracket\cdot\mathbf{n}_f,\, \{q\})_{L^2(F_f)}\\ =\ & \sum_{\ell\in\mathcal{I}_K} \langle \mathbf{grad}\, q, \mathbf{v}\rangle_{\mathbf{H}^{-t}(K_\ell),\mathbf{H}^t(K_\ell)} - \sum_{f\in\mathcal{I}_F^i} (\,\{\mathbf{v}\}\cdot\mathbf{n}_f, \llbracket q\rrbracket)_{L^2(F_f)}.\end{aligned} \tag{6.26}$$

For the transition from the second to the third equality, we used the integration by parts formula (2.14), and in the last line we used relation (4.1).

Recall that for $q \in L_h$, and $\mathbf{v} \in \mathbf{X}_{\star h}$, $\langle \mathbf{grad}\, q, \mathbf{v}\rangle_{\mathbf{H}^{-t}(\Omega),\mathbf{H}^t(\Omega)} = (\mathbf{grad}_h q, \mathbf{v})_{L^2(\Omega)}$.
We deduce that the following consistency properties hold:

$$\forall (\mathbf{v}, q_h) \in (\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^{1+s}(\Omega)) \times L_h, \quad b_h(\mathbf{v}, q_h) = -(\mathrm{div}\,\mathbf{v}, q_h)_{L^2(\Omega)},$$

$$\forall (\mathbf{v}_h, q) \in \mathbf{X}_h \times \left(L^2_{zmv}(\Omega) \cap H^s(\Omega)\right), \quad b_h(\mathbf{v}_h, q) = \sum_{\ell\in\mathcal{I}_K} \langle \mathbf{grad}\, q, \mathbf{v}_h\rangle_{\mathbf{H}^{-t}(K_\ell),\mathbf{H}^t(K_\ell)}.$$

Furthermore, we introduce another stabilization bilinear form $\tilde{s}_h$ that allows to damp

the pressure jumps on the interfaces:

$$\forall (q_h, r_h) \in L_h \times L_h, \quad \tilde{s}_h(q_h, r_h) = \sum_{f \in \mathcal{I}_F^i} h_f \left([\![q_h]\!], [\![r_h]\!] \right)_{L^2(F_f)}. \tag{6.27}$$

The discretization of Problem (5.1) using the Symmetric Interior Penalty method reads:

$$\left\{ \begin{array}{l} \text{Find } (\mathbf{u}_h, p_h) \in \mathbf{X}_h \times L_h \text{ such that:} \\ \begin{array}{rcll} \nu\, a_h(\mathbf{u}_h, \mathbf{v}_h) + b_h(\mathbf{v}_h, p_h) & = & \ell_{\mathbf{f},\mathbf{g}}(\mathbf{v}_h) & \forall \mathbf{v}_h \in \mathbf{X}_h, \\ -b_h(\mathbf{u}_h, q_h) + \lambda\, \tilde{s}_h(p_h, q_h) & = & \displaystyle\sum_{f \in \mathcal{I}_F^b} \left(\{ q_h \}, \mathbf{g} \cdot \mathbf{n}_f \right)_{L^2(F_f)} & \forall q_h \in L_h, \end{array} \end{array} \right. \tag{6.28}$$

where

$$\begin{aligned} \ell_{\mathbf{f},\mathbf{g}}(\mathbf{v}_h) := & \sum_{\ell \in \mathcal{I}_K} \langle \mathbf{f}, \mathbf{v}_h \rangle_{\mathbf{H}^{-t}(K_\ell), \mathbf{H}^t(K_\ell)} \\ & + \sum_{f \in \mathcal{I}_F^b} \nu \left(-(\mathbf{g}, \mathbf{Grad}\, \mathbf{v}_h : \mathbf{n}_f)_{\mathbf{L}^2(F_f)} + \frac{\eta_f}{h_f} (\mathbf{g}, \mathbf{v}_h)_{\mathbf{L}^2(F_f)} \right), \end{aligned} \tag{6.29}$$

and $\lambda = \nu^{-1}$ if $k_{\mathbf{u}} = k_p$ and 0 else. We also define

$$\ell_h((\mathbf{v}_h, q_h)) := \ell_{\mathbf{f},\mathbf{g}}(\mathbf{v}_h) + \sum_{f \in \mathcal{I}_F^b} (q_h, \mathbf{g} \cdot \mathbf{n}_f)_{L^2(F_f)}. \tag{6.30}$$

Definition 6.12 *We introduce the following norms:*

— *For all* $q_h \in L_h$, $|q_h|_{\tilde{s}}^2 := \tilde{s}_h(q_h, q_h)$ *and* $\|q_h\|_{pre}^2 := \|q_h\|_{L^2(\Omega)}^2 + |q_h|_{\tilde{s}}^2$.

— *For all* $\mathbf{v}_h \in \mathbf{X}_h$, $\|\mathbf{v}_h\|_{vel}^2 := \displaystyle\sum_{i=1}^{d} \|v_{h,i}\|_{sip}^2 = (\|\mathbf{Grad}_h \mathbf{v}_h\|_{\mathbb{L}^2(\Omega)}^2 + |\mathbf{v}_h|_J^2)^{\frac{1}{2}}$.

We introduce the space $\mathcal{X}_h := \mathbf{X}_h \times L_h$, which is a Hilbert space when equipped with the norm:

$$\|(\mathbf{v}_h, q_h)\|_{\mathcal{X}_h}^2 := \|\mathbf{v}_h\|_{vel}^2 + \nu^{-2} \|q_h\|_{pre}^2. \tag{6.31}$$

We consider the following symmetric bilinear form:

$$a_{S,h} : \left\{ \begin{array}{rcl} \mathcal{X}_h \times \mathcal{X}_h & \to & \mathbb{R} \\ ((\mathbf{u}'_h, p'_h), (\mathbf{v}_h, q_h)) & \mapsto & \nu\, a_h(\mathbf{u}'_h, \mathbf{v}_h) + b_h(\mathbf{v}_h, p'_h) \\ & & + b_h(\mathbf{u}'_h, q_h) - \lambda \tilde{s}_h(p'_h, q_h) \end{array} \right. . \tag{6.32}$$

Then, problem (6.28) can be rewritten in the form:

$$\left\{\begin{array}{c}\text{Find } (\mathbf{u}_h, p_h) \in \mathcal{X}_h \text{ such that for all } (\mathbf{v}_h, q_h) \in \mathcal{X}_h,\\ a_{S,h}(\,(\mathbf{u}_h, p_h), (\mathbf{v}_h, q_h)\,) = \ell_h((\mathbf{v}_h, q_h)).\end{array}\right. \tag{6.33}$$

Remark 6.13 *In what follows, in the proof of theorem 6.9, we need to replace the symbol "$\lesssim$" by "$\leq$" in inequality (4.21)-(ii), in Lemma 6.8. For this purpose, for all $\ell \in \mathcal{I}_K$, we define C_Π as a positive constant independent of h_ℓ and of the mesh regularity parameter σ, such that*

$$\begin{array}{rcll}\|v - \pi_\ell^k v\|_{L^2(F_f)} & \leq & (\sigma_\ell)^{1/2}\, h_\ell^{1/2}\, C_\Pi\, |v|_{H^1(K_\ell)} & \text{cf. (4.21)-(ii),}\\ \|\pi_h^k v\|_h & \leq & \sigma\, C_\Pi\, \|\mathbf{grad}\, v\|_{\mathbf{L}^2(\Omega)} & \text{cf. (4.41).}\end{array} \tag{6.34}$$

In order to prove the uniform well-posedness of problem (6.33), we need the following lemma:

Lemma 6.8 *Let $v \in H^1(\Omega)$. The following bounds hold:*

$$\forall q_h \in L_h, \quad \left|\sum_{f\in\mathcal{I}_F^i} \left(\left\{\pi_h^k v - v\right\}, [\![q_h]\!]\right)_{L^2(F_f)}\right| \leq \sigma\, C_\Pi\, |v|_{H^1(\Omega)}\, |q_h|_{\tilde{s}}, \tag{6.35}$$

$$\sum_{f\in\mathcal{I}_F^i} \frac{1}{h_f} \left\|[\![\pi_h^k v]\!]\right\|^2_{L^2(F_f)} \leq (\sigma\, C_\Pi)^2\, |v|^2_{H^1(\Omega)}. \tag{6.36}$$

Proof. Continuous and discrete Cauchy-Schwarz inequalities lead to

$$\left|\sum_{f\in\mathcal{I}_F^i} \left(\left\{\pi_h^k v - v\right\}, [\![q_h]\!]\right)_{L^2(F_f)}\right| \leq$$

$$\left(\sum_{f\in\mathcal{I}_F^i} \frac{\|\left\{\pi_h^k v - v\right\}\|^2_{L^2(F_f)}}{h_f}\right)^{\frac{1}{2}} \qquad \left(\sum_{f\in\mathcal{I}_F^i} h_f \|[\![q_h]\!]\|^2_{L^2(F_f)}\right)^{\frac{1}{2}}.$$

Splitting the average into a sum over the two elements that share a given F_f, then using the first equation in (6.34) with the fact that $\frac{h_\ell}{h_f} \leq \sigma_\ell$ when $F_f \subset \partial K_\ell$, and then

rearranging the sum over internal faces into a sum over elements leads to

$$\sum_{f\in\mathcal{I}_F^i}\left|\left(\left\{\pi_h^k v - v\right\},[\![q_h]\!]\right)_{L^2(F_f)}\right| \le \left(\sum_{\ell\in\mathcal{I}_K}(C_\Pi\,\sigma_\ell)^2\,|v|^2_{H^1(K_\ell)}\right)^{\frac12}|q_h|_{\tilde{s}}\,,$$

which leads to (6.35).
Let us prove (6.36). Since $v\in H^1(\Omega)$ we have $[\![v]\!]=0$ on any face F_f and thus

$$\sum_{f\in\mathcal{I}_F^i}\frac{1}{h_f}\left\|[\![\pi_h^k v]\!]\right\|^2_{L^2(F_f)} = \sum_{f\in\mathcal{I}_F^i}\frac{1}{h_f}\left\|[\![\pi_h^k v - v]\!]\right\|^2_{L^2(F_f)}.$$

Bounding the jump by a sum over the two elements that share a given F_f, then using the first equation in (6.34) with the fact that $\frac{h_\ell}{h_f}\le\sigma_\ell$ when $F_f\subset\partial K_\ell$, we get

$$\sum_{f\in\mathcal{I}_F^i}\frac{1}{h_f}\left\|[\![\pi_h^k v - v]\!]\right\|^2_{L^2(F_f)} \le \sum_{\ell\in\mathcal{I}_K}(C_\Pi\,\sigma_\ell)^2\,|v|^2_{H^1(K_\ell)}\,,$$

which leads to (6.36).

□

We need to define the projection operator $\Pi_h^{k_u}\in\mathcal{L}(\mathbf{L}^2(\Omega),\mathbf{X}_h)$ such that, for all $\mathbf{v}\in\mathbf{L}^2(\Omega)$, for all $\ell\in\mathcal{I}_K$, $\Pi_h^{k_u}\mathbf{v} = (\pi_h^{k_u}v)_{i=1}^d$. Following the proofs of [40, Lemmata 6.14 and 6.20], we can prove that the bilinear form $a_{S,h}(\,\cdot,\cdot\,)$ is continuous and satisfies an inf-sup condition:

Theorem 6.9 (`Stability`) *Let $(k_{\mathbf{u}},k_p)\in\mathbb{N}^\star\times\mathbb{N}$, with $k_p\in\{k_{\mathbf{u}},k_{\mathbf{u}}+1\}$. Then the bilinear form $a_{S,h}(\,\cdot,\cdot\,)$ satisfies the following inf-sup condition, $\forall(\mathbf{u}_h',p_h')\in\mathcal{X}_h\backslash(0,0)$:*

$$\sup_{(\mathbf{v}_h,q_h)\in\mathcal{X}_h\backslash(0,0)}\frac{a_{S,h}(\,(\mathbf{u}_h',p_h'),(\mathbf{v}_h,q_h)\,)}{\|(\mathbf{v}_h,q_h)\|_{\mathcal{X}_h}} \;\ge\; \nu\,C_{disc}\,\|(\mathbf{u}_h',p_h')\|_{\mathcal{X}_h}, \tag{6.37}$$

$$|a_{S,h}(\,(\mathbf{u}_h',p_h'),(\mathbf{v}_h,q_h)\,)| \;\le\; C_{ct}\|(\mathbf{u}_h',p_h')\|_{\mathcal{X}_h}\|(\mathbf{v}_h,q_h)\|_{\mathcal{X}_h}, \tag{6.38}$$

where $C_{disc}\approx\sigma^{-3}$ and $C_{ct}\approx\sigma$ are strictly positive constants independent of h.

Proof. Let $(\mathbf{u}_h',p_h')\in\mathcal{X}_h$,

1. According to Proposition 2.3, there exists $\tilde{\mathbf{v}}_{p_h'}\in\mathbf{H}_0^1(\Omega)$ such that $\operatorname{div}\tilde{\mathbf{v}}_{p_h'} = p_h'$ in the domain Ω and $\|\tilde{\mathbf{v}}_{p_h'}\|^2_{\mathbf{H}_0^1(\Omega)}\le C_{\mathrm{div}}\,\|p_h'\|^2_{L^2(\Omega)}$. Let $\mathbf{v}_{p_h'}=\nu^{-1}\,\tilde{\mathbf{v}}_{p_h'}$, so that

$\operatorname{div} \mathbf{v}_{p'_h} = \nu^{-1} p'_h$ in Ω and

$$\|\mathbf{v}_{p'_h}\|_{\mathbf{H}^1_0(\Omega)} \leq \frac{C_{\mathrm{div}}}{\nu} \|p'_h\|_{L^2(\Omega)}. \tag{6.39}$$

In the remainder of this proof, the estimates will be given up to a constant. Our main interest is to understand how the coercivity constant relates to the mesh regularity.
We can now bound the norm $\|\Pi_h^{k_u} \mathbf{v}_{p'_h}\|_{vel}$ as follows. We first recall that:

$$\|\Pi_h^{k_u} \mathbf{v}_{p'_h}\|_{vel}^2 = \|\Pi_h^{k_u} \mathbf{v}_{p'_h}\|_h^2 + \sum_{f \in \mathcal{I}_F} \frac{1}{h_f} \|[\![\Pi_h^{k_u} \mathbf{v}_{p'_h}]\!]\|_{L^2(F_f)}^2.$$

Using the first equation in (6.34), noticing that $[\![\Pi_h^{k_u} \mathbf{v}_{p'_h}]\!] = [\![\Pi_h^{k_u} \mathbf{v}_{p'_h} - \mathbf{v}_{p'_h}]\!]$, together with inequality (6.39), we get:

$$\|\Pi_h^{k_u} \mathbf{v}_{p'_h}\|_{vel} \leq \sigma\, C_\Pi \|\mathbf{v}_{p'_h}\|_{\mathbf{H}^1_0(\Omega)} \leq \sigma \frac{C_{\mathrm{div}}\, C_\Pi}{\nu} \|p'_h\|_{L^2(\Omega)}. \tag{6.40}$$

Choosing $k_p \in \{k_{\mathbf{u}}, k_{\mathbf{u}} + 1\}$, we must choose $\lambda > 0$ in (6.28) in order to obtain a stable scheme. Regarding the physical quantities, we let $\lambda = \nu^{-1}$. Consider $(\mathbf{u}_{\star,h}, p_{\star,h}) := (\tilde{\gamma}\mathbf{u}'_h - \Pi_h^{k_u} \mathbf{v}_{p'_h}, \tilde{\gamma} p'_h)$, with $\tilde{\gamma} > 0$. We obtain:

$$\begin{aligned} a_{S,h}(\,(\mathbf{u}'_h, p'_h), (\mathbf{u}_{\star,h}, p_{\star,h})\,) &= \nu\,\tilde{\gamma}\, a_h(\mathbf{u}'_h, \mathbf{u}'_h) - \nu\, a_h(\mathbf{u}'_h, \Pi_h^{k_u} \mathbf{v}_{p'_h}) \\ &\quad - b_h(\Pi_h^{k_u} \mathbf{v}_{p'_h}, p'_h) + \tilde{\gamma}\, \nu^{-1} |p'_h|_{\tilde{s}}^2, \\ &\geq \nu \frac{\tilde{\gamma}}{2} \|\mathbf{u}'_h\|_{vel}^2 - \nu\, a_h(\mathbf{u}'_h, \Pi_h^{k_u} \mathbf{v}_{p'_h}) \\ &\quad - b_h(\Pi_h^{k_u} \mathbf{v}_{p'_h}, p'_h) + \tilde{\gamma}\, \nu^{-1} |p'_h|_{\tilde{s}}^2. \end{aligned} \tag{6.41}$$

By applying inequality (6.14) component-wise.

2. We will bound the second and third terms on the right-hand side of equation (6.41).

2^{nd} **term** Let us apply inequality (6.15) component-wise, then apply Young's inequality

and inequality (6.40): $\forall \alpha > 0$:

$$\begin{aligned} -\nu\, a_h(\mathbf{u}'_h, \Pi_h^{k_u}\mathbf{v}_{p'_h}) &\geq -\nu\, C_{bnd}\left(\frac{\alpha}{2}\,\|\mathbf{u}'_h\|^2_{vel} + \frac{1}{2\,\alpha}\,\|\Pi_h^{k_u}\mathbf{v}_{p'_h}\|^2_{vel}\right), \\ &\geq -\nu\, C_{bnd}\left(\frac{\alpha}{2}\,\|\mathbf{u}'_h\|^2_{vel} + \nu^{-2}\,\frac{(\sigma\, C_\Pi\, C_{\mathrm{div}})^2}{2\,\alpha}\,\|p'_h\|^2_{L^2(\Omega)}\right). \end{aligned}$$

The sum of the first two terms on the right-hand side of equation (6.41) is bounded as follows:

$$\begin{aligned} &\nu\,\tilde{\gamma}\, a_h(\mathbf{u}'_h, \mathbf{u}'_h) - \nu\, a_h(\mathbf{u}'_h, \Pi_h^{k_u}\mathbf{v}_{p'_h}) \\ &\geq \nu\left(\frac{\tilde{\gamma}}{2} - C_{bnd}\frac{\alpha}{2}\right)\|\mathbf{u}_h\|^2_{vel} - \nu^{-1}\,\frac{(\sigma\, C_\Pi\, C_{\mathrm{div}})^2}{2\,\alpha}\,\|p'_h\|^2_{L^2(\Omega)}. \end{aligned} \tag{6.42}$$

3^{rd} **term** By the definition of $\Pi_h^{k_u}$, and for $k_p \leq k_{\mathbf{u}} + 1$, we have:

$$-(\mathbf{grad}_h p'_h, \Pi_h^{k_u}\mathbf{v}_{p'_h})_{\mathbf{L}^2(\Omega)} = -(\mathbf{grad}_h p'_h, \mathbf{v}_{p'_h})_{\mathbf{L}^2(\Omega)}. \tag{6.43}$$

Integrating by parts the left-hand side in (6.43), we deduce that:

$$\begin{aligned} -\sum_{\ell\in\mathcal{I}_K}(\mathbf{grad}\; p'_h, \Pi_h^{k_u}\mathbf{v}_{p'_h})_{\mathbf{L}^2(K_\ell)} &= -\sum_{\ell\in\mathcal{I}_K}(\mathbf{grad}\, p'_h, \mathbf{v}_{p'_h})_{\mathbf{L}^2(K_\ell)}, \\ &= \sum_{\ell\in\mathcal{I}_K}\left[(p'_h, \operatorname{div}\mathbf{v}_{p'_h})_{L^2(K_\ell)} - \sum_{f\in\mathcal{I}_{F,\ell}}(\mathbf{v}_{p'_h}\cdot\mathbf{n}_{f,\ell}, p'_h)_{L^2(F_f)}\right], \\ &= \sum_{\ell\in\mathcal{I}_K}\nu^{-1}\|p'_h\|^2_{L^2(K_\ell)} - \sum_{f\in\mathcal{I}_F^i}\left(\left\{\mathbf{v}_{p'_h}\right\}\cdot\mathbf{n}_f, [\![p'_h]\!]\right)_{L^2(F_f)}, \end{aligned}$$

where we used that $\operatorname{div}\mathbf{v}_{p'_h} = \nu^{-1}p'_h$; and that for all $f \in \mathcal{I}_F^i$, $\mathbf{v}_{p'_h}|_{F_f} \in \mathbf{H}^{\frac{1}{2}}(F_f)$ and for all $f \in \mathcal{I}_F^b$, $\mathbf{v}_{p'_h}|_{F_f} = 0$. We deduce that:

$$\begin{aligned} &-b_h(\Pi_h^{k_u}\mathbf{v}_{p'_h}, p'_h) = \\ &-\sum_{\ell\in\mathcal{I}_K}(\mathbf{grad}\, p'_h, \Pi_h^{k_u}\mathbf{v}_{p'_h})_{\mathbf{L}^2(K_\ell)} + \sum_{f\in\mathcal{I}_F^i}\left(\left\{\Pi_h^{k_u}\mathbf{v}_{p'_h}\right\}\cdot\mathbf{n}_f, [\![p'_h]\!]\right)_{L^2(F_f)}, \\ &= \nu^{-1}\|p'_h\|^2_{L^2(\Omega)} + \sum_{f\in\mathcal{I}_F^i}\left(\left\{(\Pi_h^{k_u}\mathbf{v}_{p'_h} - \mathbf{v}_{p'_h})\right\}\cdot\mathbf{n}_f, [\![p'_h]\!]\right)_{L^2(F_f)}. \end{aligned} \tag{6.44}$$

Using inequality (6.35), followed by Young's inequality, and inequality (6.39) we have $\forall \zeta > 0$:

$$\begin{aligned}
&\sum_{f \in \mathcal{I}_F^i} \Big(\Big\{ (\Pi_h^{k_u} \mathbf{v}_{p'_h} - \mathbf{v}_{p'_h}) \Big\} \cdot \mathbf{n}_f, [\![p'_h]\!] \Big)_{L^2(F_f)} \\
&\quad \geq -\sigma \, C_\Pi \, |\mathbf{v}_{p'_h}|_{\mathbf{H}_0^1(\Omega)} |p'_h|_{\tilde{s}}, \\
&\quad \geq -\frac{\zeta}{2} \, \nu^{-1} \, |p'_h|_{\tilde{s}}^2 - \frac{(\sigma \, C_\Pi \, C_{\mathrm{div}})^2}{2\,\zeta} \, \nu^{-1} \, \|p'_h\|_{L^2(\Omega)}^2.
\end{aligned} \tag{6.45}$$

We then obtain the following bound for the term (6.44):

$$-b_h(\Pi_h^{k_u} \mathbf{v}_{p'_h}, p'_h) \geq \nu^{-1} \left(\left(1 - \frac{(\sigma \, C_\Pi \, C_{\mathrm{div}})^2}{2\,\zeta} \right) \|p'_h\|_{L^2(\Omega)}^2 - \frac{\zeta}{2} \, |p'_h|_{\tilde{s}}^2 \right). \tag{6.46}$$

3. By substituting the inequalities (6.42) and (6.46) into (6.41), we get:

$$\begin{aligned}
&a_{S,h}(\, (\mathbf{u}'_h, p'_h), (\mathbf{u}_{\star,h}, p_{\star,h}) \,) \\
&\quad \geq \nu \, C_{vel} \, \|\mathbf{u}'_h\|_{vel}^2 + \nu^{-1} \left(C_\Omega \, \|p'_h\|_{L^2(\Omega)}^2 + C_{\tilde{s}} \, |p'_h|_{\tilde{s}}^2 \right),
\end{aligned} \tag{6.47}$$

where we have introduced the following coefficients:

$$C_{vel} = \frac{\tilde{\gamma} - C_{bnd} \, \alpha}{2}, \quad C_\Omega = 1 - \frac{(\sigma \, C_\Pi \, C_{\mathrm{div}})^2}{2} \left(\alpha^{-1} + \zeta^{-1} \right), \quad C_{\tilde{s}} = \tilde{\gamma} - \frac{\zeta}{2}.$$

We need to choose the parameters α, $\tilde{\gamma}$, and ζ such that these coefficients are positive. Choosing $\alpha = \zeta = 2\,(\sigma \, C_\Pi \, C_{\mathrm{div}})^2$, we have:

$$C_{vel} = \frac{\tilde{\gamma}}{2} - C_{bnd} \, (\sigma \, C_\Pi \, C_{\mathrm{div}})^2, \quad C_\Omega = \frac{1}{2}, \quad C_{\tilde{s}} = \tilde{\gamma} - (\sigma \, C_\Pi \, C_{\mathrm{div}})^2.$$

We have $C_{bnd} > 1$. Choosing $\tilde{\gamma} = 4 \, C_{bnd} \, (\sigma \, C_\Pi \, C_{\mathrm{div}})^2$, we obtain

$$C_{vel} = C_{bnd} \, (\sigma \, C_\Pi \, C_{\mathrm{div}})^2 \text{ and } C_{\tilde{s}} = (4 \, C_{bnd} - 1) \, (\sigma \, C_\Pi \, C_{\mathrm{div}})^2.$$

These constants are greater than one. Then we obtain:

$$a_{S,h}(\, (\mathbf{u}'_h, p'_h), (\mathbf{u}_{\star,h}, p_{\star,h}) \,) \geq \frac{\nu}{2} \left(\|\mathbf{u}'_h\|_{vel}^2 + \nu^{-2} \, \|p'_h\|_{pre}^2 \right).$$

Thus, we have:

$$a_{S,h}(\,(\mathbf{u}'_h, p'_h), (\mathbf{u}_{\star,h}, p_{\star,h})\,) \gtrsim \frac{\nu}{2}\,\|(\mathbf{u}'_h, p'_h)\|^2_{\mathcal{X}_h}. \tag{6.48}$$

Now, we write:

$$\begin{aligned}
\|(\mathbf{u}_{\star,h}, p_{\star,h})\|^2_{\mathcal{X}_h} &\leq \|\mathbf{u}'_h\|^2_{vel} + \tilde{\gamma}^2\,\|\Pi^{k_u}_h \mathbf{v}_{p'_h}\|^2_{vel} + \nu^{-2}\,\tilde{\gamma}^2\,\|p'_h\|^2_{pre},\\
&\leq \|\mathbf{u}'_h\|^2_{vel} + \nu^{-2}\,\left((\tilde{\gamma}\,\sigma\,C_\Pi\,C_{\mathrm{div}})^2\|p'_h\|^2_{L^2(\Omega)} + \tilde{\gamma}^2\,\|p'_h\|^2_{pre}\right) \text{ from (6.40)},\\
&\leq (\tilde{C}_{\max})^2\,\|(\mathbf{u}'_h, p'_h)\|^2_{\mathcal{X}_h} \text{ with } \tilde{C}_{\max} = \tilde{\gamma}\,\sigma\,C_\Pi\,C_{\mathrm{div}} = 4\,C_{bnd}\,(\sigma\,C_\Pi\,C_{\mathrm{div}})^3.
\end{aligned}$$

Substituting this result into (6.48), we get:

$$a_{S,h}(\,(\mathbf{u}'_h, p'_h), (\mathbf{u}_{\star,h}, p_{\star,h})\,) \gtrsim \frac{\nu}{2}\,(\tilde{C}_{\max})^{-1}\,\|(\mathbf{u}'_h, p'_h)\|_{\mathcal{X}_h}\,\|(\mathbf{u}_{\star,h}, p_{\star,h})\|_{\mathcal{X}_h}. \tag{6.49}$$

Thus, we deduce (6.37) with an inf-sup condition constant such that $C_{disc} := \frac{1}{2}\,(\tilde{C}_{\max})^{-1}$, independent of the mesh size.

For the continuity of $a_{S,h}(\cdot,\cdot)$, we use [40, Lemma 6.20] where another norm is used, which is equivalent to the norm $\|(\cdot,\,\cdot)\|_{\mathcal{X}_h}$ on $\mathcal{X}_h$, in a way that is independent of h. To prove this, we only need inequality (4.12).

□

Following the proof of Theorem 6.4, we can show the next result:

Theorem 6.10 *The right-hand side in (6.33), $\ell_h(\cdot,\cdot)$, is continuous with respect to the $\|\cdot\|_{\mathcal{X}_h}$ norm.*

Thus, we can state the well-posedness of the discrete Stokes problem (6.33):

Theorem 6.11 *According to Theorems 6.9 and 6.10, the bilinear form $a_{S,h}((\cdot,\cdot))$ is continuous and satisfies the inf-sup condition, and $\ell_h((\cdot,\cdot))$ is continuous. Therefore, problem (6.33) is well-posed.*

We can prove the consistency of the approximation of problem (5.1) with the SIP method, in the case of weakly regular solution.

Theorem 6.12 (`Consistency`) *Let $s \in (\frac{1}{2}, 1]$. The following consistency property holds: Let $(\mathbf{u}, p) \in \mathbf{H}^{1+s}(\Omega) \times \mathbf{H}^s(\Omega)$ be the solution to the Stokes Problem (5.1),*

and $(\mathbf{u}_h, p_h)$ *be the solution to the discrete Stokes Problem* (6.33)*, with the source term* $\mathbf{f} \in \mathbf{H}^{s-1}(\Omega)$ *if* Ω *is convex, and* $\mathbf{f} \in \mathbf{L}^2(\Omega)$ *if* Ω *is non-convex. Then, for all* $(\mathbf{v}_h, q_h) \in \mathcal{X}_h$*, we have:*

$$a_{S,h}((\mathbf{u} - \mathbf{u}_h, p - p_h), (\mathbf{v}_h, q_h)) = 0. \tag{6.50}$$

Proof. Let $(\mathbf{u}_h, p_h) \in \mathcal{X}_h$. We first develop the left-hand side of (6.50) and use Problem (6.33) to obtain

$$\begin{aligned} & a_{S,h}(\,(\mathbf{u} - \mathbf{u}_h, p - p_h), (\mathbf{v}_h, q_h)\,) \\ & \quad = a_{S,h}(\,(\mathbf{u}, p), (\mathbf{v}_h, q_h)\,) - a_{S,h}(\,(\mathbf{u}_h, p_h), (\mathbf{v}_h, q_h)\,) \\ & \quad = a_{S,h}(\,(\mathbf{u}, p), (\mathbf{v}_h, q_h)\,) - \ell_h((\mathbf{v}_h, q_h)). \end{aligned}$$

We recall the definition of $a_{S,h}$ given by (6.32), as well as definitions (6.25) with (6.4)-(6.6), definitions (6.26) and (6.27).

Since $\mathbf{u} \in \mathbf{H}^{1+s}(\Omega)$, it holds for all $f \in \mathcal{I}_F$ $\{\mathbf{Grad}_h \mathbf{u}\} : \mathbf{n}_f = \mathbf{Grad}\,\mathbf{u} : \mathbf{n}_f$. For all $f \in \mathcal{I}_F^i$ we have $[\![\mathbf{u}]\!]_{F_f} = 0$ and for all $f \in \mathcal{I}_F^b$ $[\![\mathbf{u}]\!]_{F_f} = \mathbf{g}$. Similarly, since $p \in H^s(\Omega)$, $\forall f \in \mathcal{I}_F^i$ we have $[\![p]\!]_{F_f} = 0$. Therefore, defining

$$J_{\mathbf{g}}((\mathbf{v}_h, q_h)) := \sum_{f \in \mathcal{I}_F^b} \Big(- (\mathbf{g}, \mathbf{Grad}\,\mathbf{v}_h : \mathbf{n}_f)_{\mathbf{L}^2(F_f)} + \frac{\eta_f}{h_f} (\mathbf{g}, \mathbf{v}_h)_{\mathbf{L}^2(F_f)} \Big),$$

we can write:

$$\begin{aligned} a_h(\mathbf{u}, \mathbf{v}_h) &= (\mathbf{u}, \mathbf{v}_h)_h - \sum_{f \in \mathcal{I}_F} (\mathbf{Grad}\,\mathbf{u} : \mathbf{n}_f, [\![\mathbf{v}_h]\!])_{\mathbf{L}^2(F_f)} + J_{\mathbf{g}}((\mathbf{v}_h, q_h)), \\ &= - \sum_{\ell \in \mathcal{I}_K} \langle \nu \Delta \mathbf{u}, \mathbf{v}_h \rangle_{\mathbf{H}^{-t}(K_\ell), \mathbf{H}^t(K_\ell)} + J_{\mathbf{g}}((\mathbf{v}_h, q_h)), \\ &= \langle \nu \Delta \mathbf{u}, \mathbf{v}_h \rangle_{\mathbf{H}^{-t}(\Omega), \mathbf{H}^t(\Omega)} + J_{\mathbf{g}}((\mathbf{v}_h, q_h)). \end{aligned}$$

Moreover, it holds:

$$b_h(\mathbf{u}, q_h) = \sum_{f \in \mathcal{I}_F^b} (q_h, \mathbf{g} \cdot \mathbf{n}_f)_{L^2(F_f)},$$

$$b_h(\mathbf{v}_h, p) = \sum_{\ell \in \mathcal{I}_K} \langle \mathbf{grad}\, p, \mathbf{v}_h \rangle_{\mathbf{H}^{-t}(K_\ell), \mathbf{H}^t(K_\ell)} = \langle \mathbf{grad}\, p, \mathbf{v}_h \rangle_{\mathbf{H}^{-t}(\Omega), \mathbf{H}^t(\Omega)}.$$

Finally, we have: $\tilde{s}_h(p, q_h) = 0$. Using definitions (6.29) and (6.30), this leads to:

$$\begin{aligned} a_{S,h}((\mathbf{u}, p), (\mathbf{v}_h, q_h)) &= \langle -\nu\Delta\mathbf{u} + \mathbf{grad}\, p, \mathbf{v}_h\rangle_{\mathbf{H}^{-t}(\Omega),\mathbf{H}^t(\Omega)} + \nu J_{\mathbf{g}}((\mathbf{v}_h, q_h)) \\ &\quad + \sum_{f\in\mathcal{I}_F^b} (q_h, \mathbf{g}\cdot\mathbf{n}_f)_{L^2(F_f)} \\ &= \langle \mathbf{f}, \mathbf{v}_h\rangle_{\mathbf{H}^{-t}(\Omega),\mathbf{H}^t(\Omega)} + \nu J_{\mathbf{g}}((\mathbf{v}_h, q_h)) + \sum_{f\in\mathcal{I}_F^b} (q_h, \mathbf{g}\cdot\mathbf{n}_f)_{L^2(F_f)} \\ &= \ell_h((\mathbf{v}_h, q_h)). \end{aligned}$$

□

6.4 New DG numerical scheme of type $\mathbf{P}_{dg}^k - P_{dg}^{k+1}$

According to the variational formulation of the Stokes problem, the coupling term between the velocity and the pressure is represented by the following bilinear form: $\forall(\mathbf{v}_h, q_h) \in \mathcal{X}_h$

$$\begin{aligned} b_h(\mathbf{v}_h, q_h) = &\quad -(\mathrm{div}_h\, \mathbf{v}_h, q_h)_{L^2(\Omega)} + \sum_{f\in\mathcal{I}_F} (\llbracket \mathbf{v}_h\rrbracket\cdot\mathbf{n}_f, \{q_h\})_{L^2(F_f)}, \\ = &\quad \langle \mathbf{grad}\, q_h, \mathbf{v}_h\rangle_{\mathbf{H}^t(\Omega)} - \sum_{f\in\mathcal{I}_F^i} (\{\mathbf{v}_h\}\cdot\mathbf{n}_f, \llbracket q_h\rrbracket)_{L^2(F_f)}. \end{aligned}$$

This term can be expressed in two equivalent forms. The first one, given by the initial equality, is called the divergence form, while the second, given by the latter equality, is called the gradient form. In general, for continuous finite element methods, the Div form is used to prove the stability of the discrete Stokes problem see [57, 23]. However, in the context of discontinuous Galerkin finite element methods, the Grad form is preferred see [40].
In Theorem 6.9, when proving the stability of the discrete Stokes problem, we can observe from equation (6.44) that:

$$\begin{aligned} -b_h(\Pi_h\mathbf{v}_{p'_h}, p'_h) &= -(\mathbf{grad}_h p'_h, \Pi_h\mathbf{v}_{p'_h})_{\mathbf{L}^2(\Omega)} + \sum_{f\in\mathcal{I}_F^i} \left(\left\{\Pi_h\mathbf{v}_{p'_h}\right\}\cdot\mathbf{n}_f, \llbracket p'_h\rrbracket\right)_{L^2(F_f)}, \\ &= \nu^{-1}\|p'_h\|^2_{L^2(\Omega)} + \sum_{f\in\mathcal{I}_F^i} \left(\left\{(\Pi_h\mathbf{v}_{p'_h} - \mathbf{v}_{p'_h})\right\}\cdot\mathbf{n}_f, \llbracket p'_h\rrbracket\right)_{L^2(F_f)}. \end{aligned}$$

The passage from the first to the second line holds under the condition that for all element $\ell \in \mathcal{I}_K$, the polynomial degree of $\mathbf{grad}\, p'_{h|K_\ell}$ can be less than or equal to that of $\mathbf{v}_{p'_h|K_\ell}$. This implies that the polynomial degree of $p'_{h|K_\ell}$ can be lower or equal to one that of $\mathbf{v}_{p'_h|K_\ell}$ plus one. Therefore, the discrete Stokes problem is well-posed when $k_p \leq k_{\mathbf{u}} + 1$.
Classically, k_p is set so that $k_p \leq k_{\mathbf{u}}$. Hence, this observation allows us to propose a new discontinuous Galerkin scheme for which the pressure space is of higher order than the velocity space, i.e. $k_p = k_{\mathbf{u}} + 1$.
We can also handle the case where $(k_{\mathbf{u}}, k_p) = (0, 1)$. The bilinear form (6.5) can then be written as:

$$\begin{aligned} a_h^{sip}(v_{\star h}, w_h) \;=\; & -\sum_{f\in\mathcal{I}_F} (\,\{\mathbf{grad}_h\, v_{\star h}\}\cdot\mathbf{n}_f, [\![w_h]\!])_{L^2(F_f)} \\ & -\sum_{f\in\mathcal{I}_F} (\,\{\mathbf{grad}_h\, w_h\}\cdot\mathbf{n}_f, [\![v_{\star h}]\!])_{L^2(F_f)} \\ & +\sum_{f\in\mathcal{I}_F} \frac{1}{h_f}([\![v_{\star h}]\!]\,,\,[\![w_h]\!])_{L^2(F_f)}. \end{aligned} \tag{6.51}$$

We note that $v_{\star h}$ is piecewise constant on each cell, so $\nabla v_{\star h} = 0$ inside each cell. However, on the interfaces, we must provide a definition for the term $\nabla v_{\star h}\cdot\mathbf{n}$. Inspired by the finite difference method, we define $\nabla v_{\star h}\cdot\mathbf{n}$ as follows.
Let $f \in \mathcal{I}_F^i$. Then there exist $L, R \in \mathcal{I}_K$ such that $\partial K_L \cap \partial K_R = F_f$. We assume $L < R$, and $\mathbf{n}_f$ is oriented outward from K_L. We denoted by $\mathbf{x}_L$ the circumcenter of K_L and $\mathbf{x}_R$ the circumcenter of K_R and we define

$$L_{R,L} = |\mathbf{x}_R - \mathbf{x}_L|.$$

Then, under the assumption that the mesh is regular in the sense that the triangulation is a Delaunay triangulation, we define

$$\nabla v_{\star h}\cdot\mathbf{n}_f = \frac{v_{\star h}(\mathbf{x}_R) - v_{\star h}(\mathbf{x}_L)}{L_{R,L}}. \tag{6.52}$$

The numerical results demonstrating the efficiency of this new scheme will be presented in the chapter on numerical results, section (9.2).

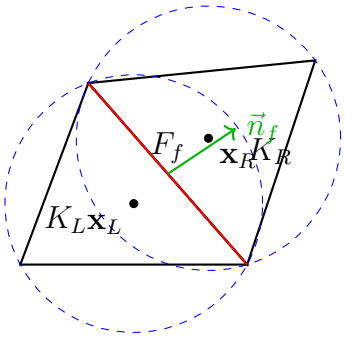
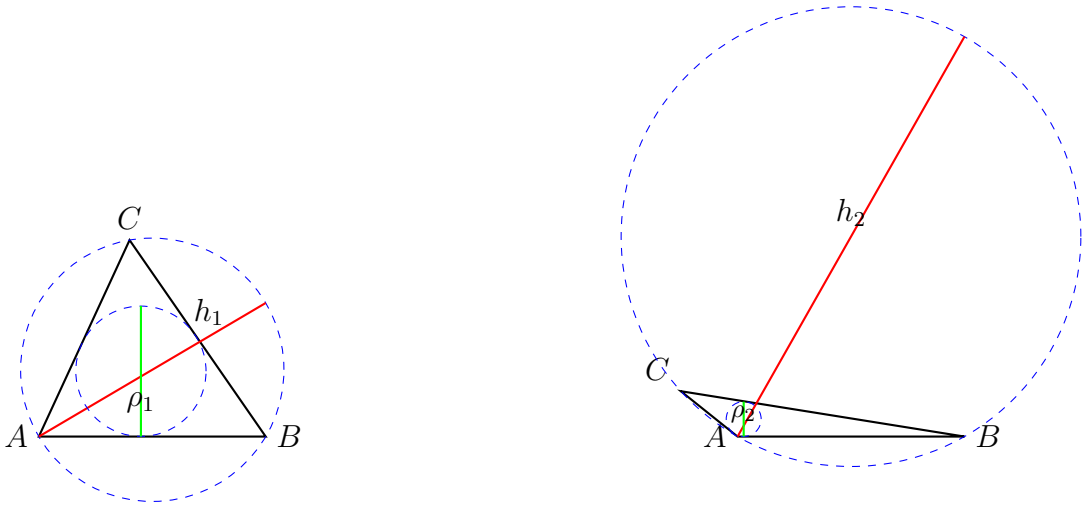


Figure 6.3 – Representation in $d = 2$.

6.5 Dependence of the discrete inf-sup constant on σ

In Theorem 6.9, we exhibited a discrete inf-sup constant C_{disc} such that $C_{disc} = \frac{1}{8}(C_{bnd})^{-1}(\sigma C_{\Pi} C_{\mathrm{div}})^{-3}$, with $C_{bnd} \approx \sigma$. This constant is independent of the mesh size h. It depends on the shape of the domain, the polynomial approximation order $k_{\mathbf{u}}$, and the mesh regularity σ. According to our computations, $C_{disc} \approx \sigma^{-4}$.
To understand the influence of mesh regularity on the coercivity constant C_{disc}, we consider the case of a non-regular mesh. This means that some elements in the mesh are highly distorted: the diameter of their circumscribed circle is much larger than that of the inscribed circle. Such distortion is often measured by a shape parameter σ, which becomes large as the mesh degenerates.
Therefore, when the mesh is not regular, the coercivity constant can become very small, which negatively affects the stability of the numerical method. In practice, this means that maintaining a regular and well-shaped mesh is essential for ensuring the robustness and accuracy of the simulation.
In this figure, we show two types of triangles: a regular triangle K_1, where $\sigma \approx 2$, and a triangle K_2 for which $\sigma > 10$.



Regular triangle $\sigma = 2.02$ **Flattened triangle** $\sigma = 14.5$

Figure 6.4 – Example of triangles to illustrate the regularity of the mesh

In this chapter, we first discretized the Poisson problem with nonhomogeneous Dirichlet boundary conditions using the Symmetric Interior Penalty (SIP) method. Then, we moved on to the discretization of the Stokes problem and showed that its variational formulation is well-posed. We explicitly derived the stability constant and studied its relation with mesh regularity. Following the steps of the proof, we proposed a new discontinuous Galerkin finite element method of type $\mathbf{P}^k - P^{k+1}$. In this method, the number of degrees of freedom for the pressure is higher than that for the velocity, which helps to improve the divergence-free condition.
We have the tools to study the a priori error estimate for the discontinuous Galerkin finite element method applied to the weakly regular Stokes problem.

Chapter 7

Error estimates with low regular solution

What I believe is new in this chapter is the analysis of a priori error estimates for the discontinuous Galerkin finite element method applied to Stokes problems with low-regularity solutions By focusing on solutions in the space $\mathbf{H}^{1+s}(\Omega) \times H^s(\Omega)$ where $s > \frac{1}{2}$.
We introduce the space $\mathcal{X}_{\star h} := \mathbf{X}_{\star h} \times L_{\star h}$, which is a Hilbert space when equipped with the norm:

$$\|(\mathbf{v}_h, q_h)\|^2_{\mathcal{X}_{\star h}} := \|(\mathbf{v}_h, q_h)\|^2_{\mathcal{X}_h} + |\mathbf{v}_h|^2_{J\star} + \nu^{-2}|q_h|^2_{\tilde{s}\star}, \tag{7.1}$$

where:

$$|\mathbf{v}_h|^2_{J\star} := \sum_{\ell\in\mathcal{I}_K} h_\ell \left\| \mathbf{Grad}\, \mathbf{v}_h|_{K_\ell} : \mathbf{n}_\ell \right\|^2_{\mathbf{L}^2(\partial K_\ell)},$$
$$|q_h|^2_{\tilde{s}\star} := \sum_{\ell\in\mathcal{I}_K} h_\ell \|q_h\|^2_{L^2(\partial K_\ell)}.$$

Theorem 7.1 ([40, Lemma 6.20]) *There exists C_{bnd}, independent of h, such that, for all $(\mathbf{u}, p) \in \mathcal{X}_{\star h}$ and all $(\mathbf{v}_h, q_h) \in \mathcal{X}_h$:*

$$|a_{S,h}((\mathbf{u}, p), (\mathbf{v}_h, q_h))| \le C_{bnd}\|(\mathbf{u}, p)\|_{\mathcal{X}_{\star h}}\|(\mathbf{v}_h, q_h)\|_{\mathcal{X}_h}.$$

Using the same method as in the proof of Theorem 7.1, we show that the bilinear form $a_{S,h}(\cdot,\cdot)$ is continuous on $\mathcal{X}_{\star h} \times \mathcal{X}_{\star h}$: For all $((\mathbf{u}, p), (\mathbf{v}, q)) \in \mathcal{X}_{\star h} \times \mathcal{X}_{\star h}$:

$$|a_{S,h}((\mathbf{u}, p), (\mathbf{v}, q))| \le C_{bnd}\|(\mathbf{u}, p)\|_{\mathcal{X}_{\star h}}\|(\mathbf{v}, q)\|_{\mathcal{X}_{\star h}}. \tag{7.2}$$

Remark 7.1 *According to inequality (4.12), for all* $(\mathbf{v}_h, q_h) \in \mathcal{X}_h$*, for all* $\ell \in \mathcal{I}_K$*, we have*

$$\begin{aligned} h_\ell \|\mathbf{Grad}\, \mathbf{v}_h|_{K_\ell} : \mathbf{n}_\ell\|^2_{\mathbf{L}^2(\partial K_\ell)} &\lesssim \sigma_\ell \|\mathbf{Grad}\, \mathbf{v}_h\|^2_{\mathbf{L}^2(K_\ell)}, \\ h_\ell \|q_h\|^2_{L^2(\partial K_\ell)} &\lesssim \sigma_\ell \|q_h\|^2_{L^2(K_\ell)}. \end{aligned}$$

Then, the norms $\|(\cdot,\cdot)\|_{\mathcal{X}_{\star\, h}}$ *and* $\|(\cdot,\cdot)\|_{\mathcal{X}_h}$ *are equivalent over* $\mathcal{X}_h$*. Then the bilinear form* $a_{S,h}(\cdot,\cdot)$ *is coercive on* $\mathcal{X}_h$ *with the norm* $\|(\cdot,\cdot)\|_{\mathcal{X}_{\star\, h}}$*.*

Since we have shown the continuity, the discrete inf-sup condition, and the consistency of the bilinear form $a_{S,h}(\cdot,\cdot)$, we shall use [40, Theorem 1.35] to obtain a Céa-type abstract error estimate. We note that [40, Theorem 1.35] uses as hypothesis the fact that the source term is in $f \in L^2(\Omega)$. However, what really counts is the consistency property (6.50) and, as soon as it is verified, the proof of [40, Theorem 1.35] can be replicated word for word.

Lemma 7.2 *Let* $(\mathbf{u}, p)$ *be the unique solution to the Stokes Problem (5.1) and* $(\mathbf{u}_h, p_h) \in \mathcal{X}_h$ *be the unique solution to the discrete Stokes Problem (6.28). There exists a constant* $C_{eq} > 0$ *independent of* h *such that:*

$$\|(\mathbf{u} - \mathbf{u}_h, p - p_h)\|_{\mathcal{X}_{\star\, h}} \leq C_{eq} \inf_{(\mathbf{v}_h, q_h) \in \mathcal{X}_h} \|(\mathbf{u} - \mathbf{v}_h, p - q_h)\|_{\mathcal{X}_{\star\, h}}.$$

Remark 7.2 *By mapping each cell* K_ℓ *to the reference element and back (change of variables), we obtain the estimate* $C_{eq} \lesssim C$*.*

In order to establish errors estimates, we will use in Lemma 7.2 the quasi-interpolation operator $\mathcal{I}_h^{\text{av}} \in \mathcal{L}(L^1(\Omega), X_h)$ defined in [48, §5]. It is important to note that this inequality (7.3) is not proved for the L^2 projection; therefore, I used this interpolation. In particular, we will use the following approximation result which is a special case of [48, Theorem 5.2] using $p = 2$, $r = 1 + s$ and $m = 1$:

Theorem 7.3 *Let* k *be the order of the polynomial approximation. For any cell* K_ℓ*, let* $\check{\mathcal{T}}_\ell = \{K_{\ell'} \in \mathcal{T}_h \,|K_{\ell'} \cap K_\ell \neq \emptyset\}$ *and* $D_\ell = \text{int}\{\cup_{K_{\ell'} \in \check{\mathcal{T}}_\ell} K_{\ell'}\}$*. There exists a uniform constant* C_{av} *such that for all* $s \in [0, k]$*, for all* $\ell \in \mathcal{I}_K$ *and for all* $v \in H^{1+s}(D_\ell)$ *it holds:*

$$|v - \mathcal{I}_h^{av}(v)|_{H^1(K_\ell)} \leq C_{av} h_\ell^s |v|_{H^{1+s}(D_\ell)}.$$

In the same way, we use [49, Ex. 22.5] to obtain the following approximation properties on the faces of the cells:

Lemma 7.4 *There exists a constant $C_{\tilde{av}} > 0$ such that for all $\ell \in \mathcal{I}_K$, for all $f \in \mathcal{I}_{F,\ell}$, for all $s \in (0, k]$ and for all $v \in H^{1+s}(D_\ell)$, we have:*

$$\|v - \mathcal{I}_h^{av}(v)\|_{L^2(F_f)} \leq C_{\tilde{av}}\, h_\ell^{s+\frac{1}{2}} |v|_{H^{1+s}(D_\ell)}.$$

Furthermore, there exists a constant $C_{\hat{av}} > 0$ such that for all $\ell \in \mathcal{I}_K$, for all $f \in \mathcal{I}_{F,\ell}$, for all $s \in (\frac{1}{2}, k]$ and for all $v \in H^{1+s}(D_\ell)$, we have:

$$\|\mathbf{grad}\,(v - \mathcal{I}_h^{av}(v))\|_{\mathbf{L}^2(F_f)} \leq C_{\hat{av}}\, h_\ell^{s-\frac{1}{2}} |v|_{H^{1+s}(D_\ell)}. \tag{7.3}$$

By summation of the squares of the previous inequalities over all mesh cells, we easily obtain the following corollary, in which the dependence of the constants with respect to σ is due to the fact that the union of all D_ℓ for $\ell \in \mathcal{I}_K$ covers a certain number of times the domain Ω.

Corollary 7.3 *There exists C_σ^{av} (respectively $C_\sigma^{\tilde{av}}$ and $C_\sigma^{\hat{av}}$) depending on C_{av} (respectively $C_{\tilde{av}}$ and $C_{\hat{av}}$) and σ such that:*

$$\|\mathbf{grad}_h(v - \mathcal{I}_h^{av}\, v)\|^2_{\mathbf{L}^2(\Omega)} \leq \quad C_\sigma^{av}\, h^{2s} |v|^2_{H^{1+s}(\Omega)}, \tag{7.4}$$

$$\sum_{\ell \in \mathcal{I}_K} \frac{1}{h_\ell} \|v - \mathcal{I}_h^{av}(v)\|^2_{L^2(\partial K_\ell)} \leq \quad C_\sigma^{\tilde{av}} h^{2s} |v|^2_{H^{1+s}(\Omega)}, \tag{7.5}$$

$$\sum_{\ell \in \mathcal{I}_K} \|\mathbf{grad}\,(v - \mathcal{I}_h^{av}(v))\|^2_{\mathbf{L}^2(\partial K_\ell)} \leq C_\sigma^{\hat{av}}\, h^{2s-1} |v|^2_{H^{1+s}(\Omega)}. \tag{7.6}$$

In order to apply the results above to the velocity field of the solution to the Stokes problem, we define the multi-dimensional version of the interpolant:

$$\boldsymbol{\mathcal{I}}_h^{\mathrm{av}} \in \mathcal{L}(\mathbf{L}^1(\Omega), \mathbf{X}_h), \text{ for all } \mathbf{v} \in \mathbf{L}^1(\Omega),\ \boldsymbol{\mathcal{I}}_h^{\mathrm{av}} \mathbf{v} = (\mathcal{I}_h^{\mathrm{av}} \mathbf{v}_i)_{i=1}^d.$$

Remark 7.4 *The estimates in Theorem 7.3 and in Lemma 7.4 are local, so that we could in practice optimize the choice the polynomial order, choosing $k \geq 2$ where the solution is locally in $\mathbf{H}^r(\Omega) \times H^{r-1}(\Omega)$, with $r > 2$, whereas we would choose $k = 1$ where the solution has lower regularity.*

In order to evaluate the inf in Lemma 7.2, we shall state the following result:

Theorem 7.5 *Let $s \in (\frac{1}{2}, 1]$. There exists a positive constant $C^\star$ independent of h such that for all $(\mathbf{v}, q) \in \mathbf{H}^{1+s}(\Omega) \times H^s(\Omega)$ it holds*

$$\left\| \left(\mathbf{v} - \boldsymbol{\mathcal{I}}_h^{av} \mathbf{v}, q - \pi_h^k q \right) \right\|_{\mathcal{X}_{\star h}} \leq \nu^{-1} C^\star h^s (\nu \, |\mathbf{v}|_{\mathbf{H}^{1+s}(\Omega)} + |q|_{H^s(\Omega)}). \tag{7.7}$$

Proof. Using (7.1), (6.31) and Definitions 6.12 and 6.3, it follows that

$$\left\| \left(\mathbf{v} - \boldsymbol{\mathcal{I}}_h^{\mathrm{av}} \mathbf{v}, q - \pi_h^k q \right) \right\|_{\mathcal{X}_{\star h}}^2 = \mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3 + \nu^{-2}(\mathfrak{T}_4 + \mathfrak{T}_5 + \mathfrak{T}_6),$$

where:

$$\begin{array}{lll} \mathfrak{T}_1 := \|\mathbf{v} - \boldsymbol{\mathcal{I}}_h^{\mathrm{av}} \mathbf{v}\|_h^2, & \mathfrak{T}_2 := |\mathbf{v} - \boldsymbol{\mathcal{I}}_h^{\mathrm{av}} \mathbf{v}|_J^2, & \mathfrak{T}_3 := |\mathbf{v} - \boldsymbol{\mathcal{I}}_h^{\mathrm{av}} \mathbf{v}|_{J_\star}^2, \\ \mathfrak{T}_4 := \|q - \pi_h^k q\|_{L^2(\Omega)}^2, & \mathfrak{T}_5 := |q - \pi_h^k q|_{\tilde{s}}^2, & \mathfrak{T}_6 := |q - \pi_h^k q|_{\tilde{s}\star}^2. \end{array}$$

Using Corollary 7.3, we have:

$$\mathfrak{T}_1 \leq C_\sigma^{av} h^{2s} |\mathbf{v}|_{\mathbf{H}^{1+s}(\Omega)}^2, \quad \mathfrak{T}_2 \lesssim C_\sigma^{\tilde{av}} h^{2s} |\mathbf{v}|_{\mathbf{H}^{1+s}(\Omega)}^2, \quad \mathfrak{T}_3 \lesssim C_\sigma^{\hat{av}} h^{2s} |\mathbf{v}|_{\mathbf{H}^{1+s}(\Omega)}^2$$

Moreover, using Proposition 4.8, we get

$$\mathfrak{T}_4 \lesssim \sigma^d \, h^{2s} |q|_{H^s(\Omega)}^2, \quad \mathfrak{T}_5 \lesssim \sigma^{d+1} \, h^{2s} |q|_{H^s(\Omega)}^2, \quad \mathfrak{T}_6 \lesssim \sigma^{d+1} \, h^{2s} |q|_{H^s(\Omega)}^2.$$

From the estimations above, we deduce (7.7) with

$$C^\star \approx \left(\max \left(C_{d,\sigma}^{av} + C_{d,\sigma}^{\tilde{av}} + C_{d,\sigma}^{\hat{av}}, \sigma^{d+1} \right) \right)^{\frac{1}{2}}.$$

This suggests that the interpolation error behaves at least like $\sigma^{\frac{d+1}{2}}$. It is however difficult to be more specific, since we don't know precisely the constants.

□

With all the results above, we are able to state one the main results of this work. We recall that σ_P, σ_S and $\sigma_{\max}$ are defined in Section 5.2.

Theorem 7.6 (`Convergence weakly solution`) *Suppose that there exists $\sigma_{\mathbf{f}} \in (0, 1)$, such that for all $s \in (0, \sigma_{\mathbf{f}})$, $\mathbf{f} \in \mathbf{H}^{s-1}(\Omega)$. Suppose moreover that for all $s \in (\frac{1}{2}, \sigma_P)$, $\mathbf{g} \in \mathbf{H}^{\frac{1}{2}+s}(\partial\Omega)$. Let $(\mathbf{u}, p) \in \mathbf{H}^{1+s}(\Omega) \times H^s(\Omega)$ for all $s \in (\frac{1}{2}, \sigma_{\max})$ be the unique solution to the Stokes Problem (5.1) with $(\mathbf{u}_h, p_h) \in \mathbf{X}_h \times L_h$ be the unique solution to the discrete Stokes Problem (6.28). Then, there exists a constant $C_{\mathcal{X}_\star} > 0$ independent*

of h such that:

$$\|(\mathbf{u}-\mathbf{u}_h, p-p_h)\|_{\mathcal{X}_{\star h}} \leq C_{\mathcal{X}_\star}\, C(\mathbf{f},\mathbf{g})\, h^s. \tag{7.8}$$

with $C_{\mathcal{X}_\star}$ depending on σ at least like $\sigma^{\frac{d+1}{2}}$.
Suppose that Ω is convex. Then it holds:

$$\|\mathbf{u}-\mathbf{u}_h\|_{\mathbf{L}^2(\Omega)} \leq \tilde{C}_{\mathcal{X}_\star}\, C(\mathbf{f},\mathbf{g})\, h^{1+s}. \tag{7.9}$$

Suppose that Ω is non convex. Then it holds:

$$\|\mathbf{u}-\mathbf{u}_h\|_{\mathbf{L}^2(\Omega)} \leq \tilde{C}_{\mathcal{X}_\star}\, C(\mathbf{f},\mathbf{g})\, h^{2s}. \tag{7.10}$$

Proof. Consider Lemma 7.2. Let $p_h = \pi_\ell^k p$ and $\mathbf{u}_h = \mathcal{I}_h^{\mathrm{av}}(\mathbf{u})$ in order to bound the infimum. Using then inequality (7.7) and (5.27), we obtain (7.8) with $C_{\mathcal{X}_\star} = C_{eq}\, C^\star$. Let us prove (7.9) and (7.10). We use a standard duality argument. Let $(\overline{\mathbf{u}}, \overline{p})$ be the solution to the Stokes (5.1) problem with $\mathbf{g} = 0$, and $\overline{\mathbf{f}} = \mathbf{u}-\mathbf{u}_h \in \mathbf{L}^2(\Omega)$. According to the Stokes regularity result recalled in lemma 5.5, we have $(\overline{\mathbf{u}}, \overline{p}) \in (\mathbf{H}^{1+s'}(\Omega)\cap\mathbf{H}_0^1(\Omega)) \times (H^{s'}(\Omega)\cap L^2_{zmv}(\Omega))$, with $s'=1$ if Ω is convex or $s'=s$ if Ω is not convex. According to inequalities (5.26) and (5.27) so that the following bound holds:

$$\nu\,(h_\Omega)^{-1}\,\|\overline{\mathbf{u}}\|_{\mathbf{H}^{1+s'}(\Omega)} + \|\overline{p}\|_{H^{s'}(\Omega)} \leq C_{\overline{\mathbf{f}}}\, h_\Omega\, \|\mathbf{u}-\mathbf{u}_h\|_{\mathbf{L}^2(\Omega)}\,, \tag{7.11}$$

Since $\overline{\mathbf{u}} \in \mathbf{H}^{1+s'}(\Omega)\cap\mathbf{H}_0^1(\Omega)$, for all $f\in\mathcal{I}_F^i$, $[\![\mathbf{Grad}\,\overline{\mathbf{u}}:\mathbf{n}_f]\!]_{F_f} = 0$, and for all $f\in\mathcal{I}_F$, $[\![\overline{\mathbf{u}}]\!]_{F_f} = 0$. Therefore, we have the following simplifications

$$\begin{aligned}
a_h(\overline{\mathbf{u}}, \mathbf{u}-\mathbf{u}_h) &= (\overline{\mathbf{u}}, \mathbf{u}-\mathbf{u}_h)_h - \sum_{f\in\mathcal{I}_F} (\mathbf{Grad}\,\overline{\mathbf{u}}:\mathbf{n}_f, [\![\mathbf{u}-\mathbf{u}_h]\!])_{L^2(F_f)},\\
b_h(\mathbf{u}-\mathbf{u}_h, \overline{p}) &= -(\mathrm{div}_h\,(\mathbf{u}-\mathbf{u}_h), \overline{p})_{L^2(\Omega)} + \sum_{f\in\mathcal{I}_F} ([\![(\mathbf{u}-\mathbf{u}_h)\cdot\mathbf{n}_f]\!], \overline{p})_{L^2(F_f)},\\
b_h(\overline{\mathbf{u}}, p-p_h) &= 0, \quad \tilde{s}_h\,(p-p_h, \tilde{p}) = 0.
\end{aligned}$$

Hence, setting $\underline{\underline{G}} := -\nu\,\mathbf{Grad}\,\overline{\mathbf{u}} + \overline{p}\,\mathbb{I}_d$ and $\mathcal{G} = \sum_{f\in\mathcal{I}_F} (\underline{\underline{G}}:\mathbf{n}_f, [\![\mathbf{u}-\mathbf{u}_h]\!])_{\mathbf{L}^2(F_f)}$, we get

$$\begin{aligned}
a_{S,h}(\,(\overline{\mathbf{u}},\overline{p}),(\mathbf{u}-\mathbf{u}_h, p-p_h)\,) &= \nu a_h\,(\overline{\mathbf{u}}, \mathbf{u}-\mathbf{u}_h) + b_h(\mathbf{u}-\mathbf{u}_h, \overline{p}),\\
&= \nu(\overline{\mathbf{u}}, \mathbf{u}-\mathbf{u}_h)_h - (\mathrm{div}_h(\mathbf{u}-\mathbf{u}_h), \overline{p})_{L^2(\Omega)} + \mathcal{G},\\
&= \nu(\mathbf{Grad}\;\overline{\mathbf{u}}, \mathbf{Grad}_h(\mathbf{u}-\mathbf{u}_h))_{\mathbb{L}^2(\Omega)} - (\overline{p}\,\mathbb{I}_d, \mathbf{Grad}_h(\mathbf{u}-\mathbf{u}_h))_{\mathbb{L}^2(\Omega)} + \mathcal{G},\\
&= -(\underline{\underline{G}}, \mathbf{Grad}_h(\mathbf{u}-\mathbf{u}_h))_{\mathbb{L}^2(\Omega)} + \mathcal{G} = (\mathrm{Div}\,\underline{\underline{G}}, \mathbf{u}-\mathbf{u}_h)_{\mathbf{L}^2(\Omega)},\\
&= (-\nu\Delta\overline{\mathbf{u}} + \mathbf{grad}\,\overline{p}, \mathbf{u}-\mathbf{u}_h)_{\mathbf{L}^2(\Omega)}.
\end{aligned}$$

Finally, we obtain

$$a_{S,h}(\,(\overline{\mathbf{u}},\overline{p}),(\mathbf{u}-\mathbf{u}_h,p-p_h)\,)=\|\mathbf{u}-\mathbf{u}_h\|^2_{\mathbf{L}^2(\Omega)}\,.$$

Using the consistency of $a_{S,h}(\cdot,\cdot)$, cf. equation (6.50), then its continuity, cf. inequality (7.2), the interpolation estimate (7.7) and the bound (7.11), we obtain:

$$\begin{aligned}
\|\mathbf{u}-\mathbf{u}_h\|^2_{\mathbf{L}^2(\Omega)} &= a_{S,h}\left((\mathbf{u}-\mathbf{u}_h,p-p_h)\,,\left(\overline{\mathbf{u}}-\mathcal{I}_h^{\mathrm{av}}\,\overline{\mathbf{u}},\overline{p}-\pi_h^k\overline{p}\right)\right),\\
&\leq C_{bnd}\,\|(\mathbf{u}-\mathbf{u}_h,p-p_h)\|_{\mathcal{X}_{\star\,h}}\left\|\left(\overline{\mathbf{u}}-\mathcal{I}_h^{\mathrm{av}}\,\overline{\mathbf{u}},\overline{p}-\pi_h^k\overline{p}\right)\right\|_{\mathcal{X}_{\star\,h}},\\
&\leq h^{s'}C_{bnd}\,\|(\mathbf{u}-\mathbf{u}_h,p-p_h)\|_{\mathcal{X}_{\star\,h}}\,C_{\mathcal{X}_\star}\,\left(|\overline{\mathbf{u}}|_{H^{1+s'}(\Omega)}+\nu^{-1}\,|\overline{p}|_{H^{s'}(\Omega)}\right),\\
&\leq h^{s'}\nu^{-1}\,C_{bnd}\,C_{\overline{\mathbf{f}}}\,C_{\mathcal{X}_\star}\,\|(\mathbf{u}-\mathbf{u}_h,p-p_h)\|_{\mathcal{X}_{\star\,h}}\,(h_\Omega)^{-s}\,\|\mathbf{u}-\mathbf{u}_h\|_{\mathbf{L}^2(\Omega)},
\end{aligned}$$

where $s'=1$ if Ω is convex and $s'=s$ if Ω is non-convex. Finally:

$$\|\mathbf{u}-\mathbf{u}_h\|_{\mathbf{L}^2(\Omega)}\leq h^{s'}\;C_{bnd}\,C_{\overline{\mathbf{f}}}\,C_{\mathcal{X}_\star}\,(h_\Omega)^{-s}\;\|(\mathbf{u}-\mathbf{u}_h,p-p_h)\|_{\mathcal{X}_{\star\,h}}\,.$$

By using estimate (7.8), we obtain (7.9) and (7.10) with $\tilde{C}_{\mathcal{X}_\star}=C_{bnd}\,C_{\overline{\mathbf{f}}}\,(C_{\mathcal{X}_\star})^2\,C(\mathbf{f},\mathbf{g})\,(h_\Omega)^{-s}$.

□

Remark 7.5 *If Ω is convex and $\mathbf{f}\in\mathbf{L}^2(\Omega)$, then $(\mathbf{u},p)\in\mathbf{H}^2(\Omega)\times H^1(\Omega)$. The convergence rates for the velocity and the pressure are given by*

$$\|\mathbf{u}-\mathbf{u}_h\|_{\mathbf{L}^2(\Omega)}\lesssim h^2,\quad \|\mathbf{u}-\mathbf{u}_h\|_{\mathbf{H}^1(\Omega)}\lesssim h,\ \textit{and}\quad \|p-p_h\|_{L^2(\Omega)}\lesssim h.$$

In this chapter, we studied the a priori error estimate for the weakly regular Stokes problem discretized with the discontinuous Galerkin scheme $\mathbf{P}^1 - P^1$, which is well suited to the regularity of this problem. We found that the regularity of the domain, whether it is convex or not, has an impact on the convergence rate. In particular, when the domain is convex, the convergence rate for the velocity is $1 + s$ and s for the pressure. On the other hand, if the domain is non-convex, the convergence rate is $2\,s$ for the velocity and s for the pressure.
Now, we can perform numerical tests to numerically verify all the results we have obtained.

Chapter 8

Algorithms

Contents

8.1 Structure of matrices

Let $d \in \{2,3\}$ be the dimension and $\mathcal{I}_d = \{x,y\}$ for $d=2$, $\mathcal{I}_d = \{x,y,z\}$ for $d=3$. Let (O,x,y) for $d=2$ and (O,x,y,z) for $d=3$ be the Cartesian coordinates system, of orthonormal basis $(\mathbf{e}_{\tilde{d}})_{\tilde{d}\in\mathcal{I}_d}$. Denote by $N_{\mathbf{u}} = N_{k_{\mathbf{u}}} \times N_K$ (resp. $N_p = N_{k_p} \times N_K$) the number of degrees of freedom of $P^{k_{\mathbf{u}}}_{disc}(\mathcal{T}_h)$ (resp. $P^{k_p}_{disc}(\mathcal{T}_h)$). For $k \in \mathbb{N}$, $k>0$, we let $\mathcal{I}_k = \{1,\dots,\frac{(k+d)!}{k!\,d!}\}$. Let's call $(\phi_{\ell,i})_{i\in\mathcal{I}_k}$ a polynomial basis of $P^k(K_\ell)$, and $A \in \mathbb{R}^{N_{\mathbf{u}}} \times \mathbb{R}^{N_{\mathbf{u}}}$ the matrix such that

$$\forall (\ell,i),\,(\ell',j) \in (\mathcal{I}_K \times \mathcal{I}_{k_{\mathbf{u}}})^2, \quad A_{(\ell,i),(\ell',j)} := a_h^{sip}(\phi_{\ell,i},\phi_{\ell',j}). \tag{8.1}$$

Let $\mathbb{A} \in \mathbb{R}^{d\times N_{\mathbf{u}}} \times \mathbb{R}^{d\times N_{\mathbf{u}}}$ be the block diagonal matrix such that: $\mathbb{A} = (\delta_{\tilde{d},\tilde{d}'}A)_{\tilde{d},\tilde{d}'\in\mathcal{I}_d}$. For $\tilde{d} \in \mathcal{I}_d$, we call $B_{\tilde{d}} \in \mathbb{R}^{N_p} \times \mathbb{R}^{N_{\mathbf{u}}}$ the matrix such that for all $(\ell,i) \in \mathcal{I}_K \times \mathcal{I}_{k_p}$, $(\ell',j) \in \mathcal{I}_K \times \mathcal{I}_{k_{\mathbf{u}}}$, $(B_{\tilde{d}})_{(\ell,i),(\ell',j)} := b_h(\phi_{\ell',j}\,\mathbf{e}_{\tilde{d}},\phi_{\ell,i})$ and $\mathbb{B} = (B_{\tilde{d}})_{\tilde{d}\in\mathcal{I}_d} \in \mathbb{R}^{N_p} \times \mathbb{R}^{d\times N_{\mathbf{u}}}$. Let $M_p \in \mathbb{R}^{N_p} \times \mathbb{R}^{N_p}$ and $S_p \in \mathbb{R}^{N_p} \times \mathbb{R}^{N_p}$ the matrices such that for all (ℓ,i), $(\ell',j) \in \mathcal{I}_K \times \mathcal{I}_{k_p}$, $(M_p)_{(\ell,i),(\ell',j)} = \delta_{\ell,\ell'}(\phi_{\ell,i},\phi_{\ell,j})_{L^2(K_\ell)}$ and $(S_p)_{(\ell,i),(\ell',j)} = s_h(\phi_{\ell,i},\phi_{\ell',j})$. Notice that A, $B_{\tilde{d}}$ and S_p are sparse: for ℓ, $\ell' \in \mathcal{I}_K$ such that $\partial K_\ell \cap \partial K_{\ell'} = \emptyset$, $A_{(\ell,i),(\ell',j)} = 0$, $(B_{\tilde{d}})_{(\ell,i),(\ell',j)} = 0$ and $(S_p)_{(\ell,i),(\ell',j)} = 0$. The matrix A is symmetric, positive definite.

When $k_{\mathbf{u}} > k_p$, matrix $\mathbb{B}$ is of rank $N_p - 1$. Notice that the matrix S_p is also of rank $N_p - N_{GC}$, where $N_{GC} := N_S + (k_p - 1)\, N_E + \frac{(k_p-1)(k_p-2)}{2}\, N_K$, it is the dimension of the space of continuous functions, where N_E denotes the number of edges and N_S the number of vertices in the mesh. When $k_{\mathbf{u}} \leq k_p$, matrix $\mathbb{B}$ is of rank $< N_p - 1$. Let $F_{\mathbf{u}} = (F_{\mathbf{u},\tilde{d}})_{\tilde{d}\in\mathcal{I}_d} \in \mathbb{R}^{d\times N_{\mathbf{u}}}$ the vector such that for all $(\ell, i) \in \mathcal{I}_K \times \mathcal{I}_{k_{\mathbf{u}}}$, $(F_{\mathbf{u},\tilde{d}})_{\ell,i} = f_h(\phi_{\ell,i}\, \mathbf{e}_{\tilde{d}})$. It stands:

$$\mathbf{u}_h = \sum_{\ell\in\mathcal{I}_K} \sum_{i\in\mathcal{I}_{k_{\mathbf{u}}}} \mathbf{u}_{\ell,i}\, \phi_{\ell,i} \text{ and } p_h = \sum_{\ell\in\mathcal{I}_K} \sum_{i\in\mathcal{I}_{k_p}} p_{\ell,i}\, \phi_{\ell,i}.$$

Let's call $U_h = (U_{h,\tilde{d}})_{\tilde{d}\in\mathcal{I}_d} \in \mathbb{R}^{d\times N_{\mathbf{u}}}$ (resp. $P_h \in \mathbb{R}^{N_p}$) the vector of the discrete velocity (resp. pressure) degrees of freedom: for all $(\ell, i) \in \mathcal{I}_K \times \mathcal{I}_{k_{\mathbf{u}}}$ (resp. $(\ell, i) \in \mathcal{I}_K \times \mathcal{I}_{k_p}$), $(U_{h,\tilde{d}})_{\ell,i} = \mathbf{u}_{\ell,i} \cdot \mathbf{e}_{\tilde{d}}$ (resp. $(P_h)_{\ell,i} = p_{\ell,i}$). The linear system related to Problem (6.28) reads: Find $(U_h, P_h) \in \mathbb{R}^{d\times N_{\mathbf{u}}} \times \mathbb{R}^{N_p}$ such that:

$$\left\{ \begin{array}{rcl} \nu\mathbb{A}U_h + \mathbb{B}^T P_h &=& F_{\mathbf{u}} \\ -\mathbb{B}U_h + \beta\, S_p\, P_h &=& F_p \end{array} \right. \text{ and } \sum_{\ell\in\mathcal{I}_K} \sum_{i\in\mathcal{I}_{k_p}} p_{\ell,i} \int_{K_\ell} \phi_{\ell,i} = 0, \tag{8.2}$$

where $F_p \in \mathbb{R}^{N_p}$ vanishes when considering homogeneous Dirichlet boundary conditions. To compute $p_h \in L_h$, we can either impose the last equation in (8.2) at each iteration of our solver or cancel a row from matrix $\mathbb{B}$ and the same row and corresponding column from matrices M_p and S_p. We make the abuse of keeping the same notation. Let's call $\mathbb{K} = \mathbb{B}\,\mathbb{A}^{-1}\mathbb{B}^T + \beta\nu\, S_p \in \mathbb{R}^{N_p-1} \times \mathbb{R}^{N_p-1}$, which is a symmetric positive definite matrix. To solve the coupled velocity-pressure problem (8.2), one relies usually on the three + one steps below (the fourth step being straightforward):

$$\begin{array}{ll} \text{Prediction:} & \text{Solve in } U_h^\star \text{ such that } \nu\, \mathbb{A}U_h^\star = F_{\mathbf{u}}. \\ \text{Pressure solver:} & \text{Solve in } P_h \text{ such that } \mathbb{K}P_h = \nu(F_p + \mathbb{B}U_h^\star). \\ \text{Correction:} & \text{Solve in } \delta U_h \text{ such that } \nu\, \mathbb{A}\delta U_h = -\mathbb{B}^T\, P_h. \\ \text{Update:} & U_h = \delta U_h + U_h^\star. \end{array} \tag{8.3}$$

One can prove that matrix $\mathbb{K}$ is spectrally equivalent to $M = M_p + \beta S_p$, so that M can be used as a precontionner to solve $\mathbb{K}P_h = \nu(F_p + \mathbb{B}U_h^\star)$ [89, Lemma 5.9]. Using some iterative solver for this system, each matrix-vector product with $\mathbb{K}$ requires the solution of linear systems such as $AX = b$ at each iteration, where A is defined by Eq. (8.1). Let us give details on our resolution algorithm.

8.2 Domain decomposition and Schur complement

Let $N \in \mathbb{N}^\star$, $\mathcal{I}_N = \{1, \cdots, N\}$ and $\tilde{\mathcal{I}}_N = \{1, \cdots, N-1\}$. We split $\mathcal{T}_h$ into N disjoint subsets $(\mathcal{T}_{h,n})_{n\in\mathcal{I}_N}$. Let's denote by $(\mathcal{F}_{h,n})_{n\in\mathcal{I}_N}$ the associated sets of facets. The splitting is such that for $n,\, m \in \tilde{\mathcal{I}}_N$, for $n \neq m$, $\mathcal{F}_{h,n} \cap \mathcal{F}_{h,m} = \emptyset$. For $n \in \mathcal{I}_N$, we consider matrices $A_{n,n} = \Big((A_{\ell,\ell'})_{\ell\in\mathcal{T}_{h,i},\ell'\in\mathcal{T}_{h,n}} \Big)$, $A_{n,N} = \Big((A_{\ell,\ell'})_{\ell\in\mathcal{T}_{h,i},\ell'\in\mathcal{T}_{h,N}} \Big)$ and for $n,\, m \in \tilde{\mathcal{I}}_N$, $A_{n,m} = 0$. The matrix A can be rewritten by blocks as follows: $A = \big((A_{n,m})_{n,m\in\mathcal{I}_N\times\mathcal{I}_N} \big)$. For all $n \in \mathcal{I}_N$, $A_{n,n}$ is a symmetric positive definite matrix, and for all $n \in \tilde{\mathcal{I}}_N$, $A_{n,N}$ is the coupling matrix between $\mathcal{F}_{h,n}$ and $\mathcal{F}_{h,N}$ (we have: $A_{N,n} = (A_{n,N})^T$). Similarly, we set $X = (X_n)_{n\in\mathcal{I}_N}$ and $b = (b_n)_{n\in\mathcal{I}_N}$. The linear system $A\,X = b$ is solved as follows:

$$
\begin{array}{l}
\text{For all } n \in \tilde{\mathcal{I}}_N, \text{ compute} \quad \left\{ \begin{array}{rcl} \tilde{b}_n &=& A_{n,N}^T\, A_{n,n}^{-1}\, b_n \\ \\ \tilde{A}_{n,n} &=& A_{n,N}^T\, A_{n,n}^{-1}\, A_{n,N} \end{array} \right. . \\
\\
\text{Compute} \qquad\qquad\qquad\qquad \left\{ \begin{array}{rcl} \tilde{b}_N &=& b_N - \displaystyle\sum_{n\in\tilde{\mathcal{I}}_N} \tilde{b}_n \\ \tilde{A}_{N,N} &=& A_{N\,N} - \displaystyle\sum_{n\in\tilde{\mathcal{I}}_N} \tilde{A}_{n,n} \end{array} \right. . \\
\\
\text{Solve for } X_N \text{ such that} \qquad \tilde{A}_{N,N}\, X_N = \tilde{b}_N. \\
\text{For all } n \in \tilde{\mathcal{I}}_N, \text{ solve for } X_n \text{ such that } A_{n,n}\, X_n = b_n - A_{n,N} X_N.
\end{array}
\tag{8.4}
$$

For $n \in \tilde{\mathcal{I}}_N$, we store the Cholesky decomposition of matrix $A_{n,n}$ to compute vector $\tilde{b}_i$ and matrix $\tilde{A}_{n,n}$. These computations can be done in parallel. After, we build $\tilde{A}_{N,N}$ and store its LU decomposition. The matrices $(\tilde{A}_{n,n})_{n\in\tilde{\mathcal{I}}_N}$ can be calculated once and for all in parallel. For the resolution, we can also calculate the vectors $(\tilde{b}_n)_{n\in\tilde{\mathcal{I}}_N}$ and $(X_n)_{n\in\tilde{\mathcal{I}}_N}$in parallel.
We present the linear system to be solved:

— The discretized Stokes problem can be written in matrix form as follows:

$$
\left\{ \begin{array}{rcl} \nu\,\mathbb{A}\,U_{xh} + \mathbb{B}_x^T\,P_h &=& F_{\mathbf{u}_x} \\ \nu\,\mathbb{A}\,U_{yh} + \mathbb{B}_y^T\,P_h &=& F_{\mathbf{u}_y} \\ -\left(\mathbb{B}_x\,U_{xh} + \mathbb{B}_y\,U_{yh}\right) + \lambda\,S_p\,P_h &=& F_p \end{array} \right. \quad \Longleftrightarrow \quad \left\{ \begin{array}{rcl} \mathbb{K}\,P_h &=& F_{(\mathbf{u},p)} \\ \nu\,\mathbb{A}\,U_{xh} &=& F_{\mathbf{u}_x} - \mathbb{B}_x^T\,P_h \\ \nu\,\mathbb{A}\,U_{yh} &=& F_{\mathbf{u}_y} - \mathbb{B}_y^T\,P_h \end{array} \right.
$$

— $\mathbb{K} = \mathbb{B}_x\,\mathbb{A}^{-1}\mathbb{B}_x^T + \mathbb{B}_y\,\mathbb{A}^{-1}\mathbb{B}_y^T + \nu\,\lambda\,S_p, \quad F_{(\mathbf{u},p)} = \nu\,F_p + \mathbb{B}_x\,\mathbb{A}^{-1}F_{\mathbf{u}_x} + \mathbb{B}_y\,\mathbb{A}^{-1}F_{\mathbf{u}_y}$.

— The matrices $\mathbb{K}$ and $\mathbb{A}$ are symmetric and positive definite.

—

Prediction (Cholesky): Find $U^{\star}_{xh}$, $U^{\star}_{yh}$ such that $\nu\,\mathbb{A}\,U^{\star}_{xh} = F_{\mathbf{u}_x}$ and $\nu\,\mathbb{A}\,U^{\star}_{yh} = F_{\mathbf{u}_y}$,

Pressure Solver (PCG): Find P_h such that $M\,\mathbb{K}\,P_h = \nu\,M\,(F_p + \mathbb{B}_x\,U^{\star}_{xh} + \mathbb{B}_y\,U^{\star}_{yh})$,

Correction (Cholesky): Find δU_{xh}, δU_{yh} such that $\begin{cases} \nu\,\mathbb{A}\,\delta U_{xh} = -\mathbb{B}_x^T\,P_h, \\ \nu\,\mathbb{A}\,\delta U_{yh} = -\mathbb{B}_y^T\,P_h, \end{cases}$
then update: $\begin{array}{l} U_{xh} = \delta U_{xh} + U^{\star}_{xh}, \\ U_{yh} = \delta U_{yh} + U^{\star}_{yh}. \end{array}$

— Solving $\mathbb{A}\,X = b$, based on domain decomposition and the Schur complement method.

The Domain Decomposition Method and Schur Complement
To simplify, we present the case when $N = 5$.

— $\mathbb{A}$ is a large symmetric positive definite matrix, and we aim to solve the linear system $\mathbb{A}\,X = b$.

— The matrix $\mathbb{A}$ can be written in the following block form:

$$\mathbb{A} = \begin{bmatrix} A_{11} & 0 & 0 & 0 & A_{15} \\ 0 & A_{22} & 0 & 0 & A_{25} \\ 0 & 0 & A_{33} & 0 & A_{35} \\ 0 & 0 & 0 & A_{44} & A_{45} \\ A_{15}^T & A_{25}^T & A_{35}^T & A_{45}^T & A_{55} \end{bmatrix}.$$

— Below is a diagram illustrating the domain decomposition into 5 subdomains Ω_i, $i = 1, \cdots, 5$:

— For $i = 1, \cdots, 4$, we define:

$$\begin{cases} \tilde{b}_i &= A_{i,5}^T\,A_{i,i}^{-1}\,b_i\,(Parallel), \\ \tilde{b}_5 &= b_5 - \sum_{i=1}^4 \tilde{b}_i, \end{cases} \qquad \begin{cases} \tilde{A}_{i,i} &= A_{i,5}^T\,A_{i,i}^{-1}\,A_{i,5}\,(Parallel), \\ \tilde{A}_{5,5} &= A_{5,5} - \sum_{i=1}^4 \tilde{A}_{i,i}, \end{cases}$$

$$\begin{cases} \tilde{A}_{5,5}\,X_5 &= \tilde{b}_5, \\ A_{i,i}\,X_i &= b_i - A_{i,5}\,X_5\,(Parallel). \end{cases}$$

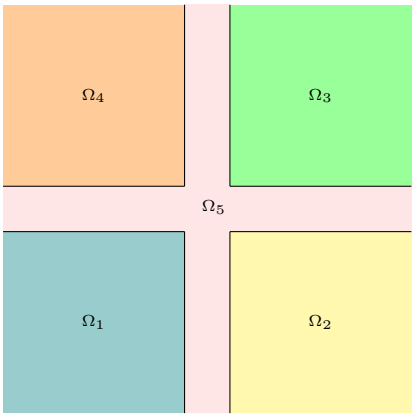


Figure 8.1 – Domain decomposition into 5 subdomains.

In order to validate our theoretical analysis and our numerical developments , we study low-regular test-cases, when Ω is non-convex and $\mathbf{f} = 0$; and when Ω is convex and $\mathbf{f} \notin \mathbf{L}^2(\Omega)$. In the next section, we give details on this last implementation.

8.3 Numerical integration of a low-regular source term

Consider here the $2D$ case where $\Omega = (0,1)^2$ and the meshes $(\mathcal{T}_h)_h$ are made of triangles. Let $S_0 = (x_0, y_0)$, where $x_0 = y_0 = 0.5$ and $\mathbf{x}_0 = \overrightarrow{OS}_0$.
Call (r, θ) the polar coordinates centered in S_0: $r = |\mathbf{x} - \mathbf{x}_0|$, $\theta = \arctan\left(\frac{y-y_0}{x-x_0}\right) + \lambda\pi$, with $\lambda = 1$ if $x \leq x_0$, and elsewhere.
Let's set $\mathcal{I}_{K,\mathbf{x}_0} := \{\ell \in \mathcal{I}_K \ : \ \mathbf{x}_0 \in K_\ell\}$ and define $\Omega_{\mathbf{x}_0} = \{K_\ell,\ \ell \in \mathcal{I}_{K,\mathbf{x}_0}\}$.
The polar basis $(\mathbf{e}_r, \mathbf{e}_\theta)$ is such that: $\mathbf{e}_r = \cos\theta \mathbf{e}_x + \sin\theta \mathbf{e}_y$, $\mathbf{e}_\theta = -\sin\theta \mathbf{e}_x + \cos\theta \mathbf{e}_y$.
Let $\alpha,\ \beta \in \mathbb{R}^\star$. We will compute the following prescribed solution to Problem (5.7):

$$(\mathbf{u}, p) = (r^\alpha \mathbf{e}_\theta, r^\beta - p_0), \qquad \text{where } p_0 = \frac{1}{|\Omega|}\int_\Omega r^\beta \, d\mathbf{x} \tag{8.5}$$

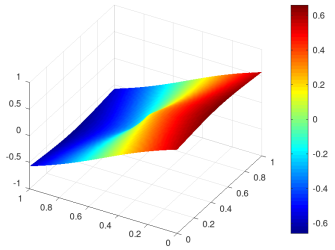


Figure 8.2 – velocity $\mathbf{u}_x$

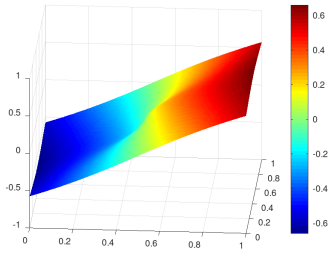


Figure 8.3 – velocity $\mathbf{u}_y$

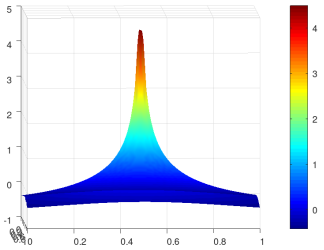


Figure 8.4 – Pressure p

When $\alpha = 1$, $\beta = 2$, it is close to the Rankine vortex model [82, Annexe B]. The source term is such that $\mathbf{f} \in (\mathbf{H}^t(\Omega))'$, with $t = 1 - s$, $s = \min\{\alpha, \beta + 1\}$ and:

$$\mathbf{f} = \nu\,(1 - \alpha^2)\, r^{\alpha-2}\, \mathbf{e}_\theta + \beta\, r^{\beta-1}\, \mathbf{e}_r. \tag{8.6}$$

Notice that $\mathbf{f}_{|\Omega_{\mathbf{x}_0}} \in \mathbf{L}^1(\Omega)$ and $\mathbf{f}_{|\Omega\setminus\Omega_{\mathbf{x}_0}} \in \mathbf{L}^2(\Omega)$. Let $\mathbf{v}_h \in \mathbf{X}_h$. It stands:

$$\sum_{\ell\in\mathcal{I}_K} \langle \mathbf{f}, \mathbf{v}_h\rangle_{\mathbf{H}^t(K_\ell)} = \sum_{\ell\in\mathcal{I}_K\setminus\mathcal{I}_{K,\mathbf{x}_0}} \int_{K_\ell} \mathbf{f}\cdot\mathbf{v}_h d\mathbf{x} + \sum_{\ell\in\mathcal{I}_{K,\mathbf{x}_0}} \langle \mathbf{f}, \mathbf{v}_h\rangle_{\mathbf{H}^t(K_\ell)}. \tag{8.7}$$

Hence, computing the right-hand side of Equation (6.28) is not straightforward.
The seven-point Gauss quadrature [98] is used to approximate the first term in the right-hand side of (8.7).
To approximate the second term in the right-hand side of (8.7), we change the variables to polar coordinates, integrating first with respect to r and then using a trapezoidal rule with respect to θ.
Let $K_\ell = S_0S_1S_2$, $\ell \in \mathcal{I}_{K,\mathbf{x}_0}$ be a triangle such that $S_i = (x_i, y_i)$ (resp. $S_i = (R(\Theta_i), \Theta_i)$ in Cartesian (resp. polar) coordinates for $i = 1, 2$.
Let $(X_i, Y_i) := (x_i - 0.5, y_i - 0.5)$, $(X, Y) := (x - 0.5, y - 0.5)$. We aim to evaluate $\int_{K_\ell} r^\gamma\, f(\theta)\, X^k\, Y^{k'} d\mathbf{x}$, for $\gamma \in \{\alpha - 2, \beta - 1\}$ and $k, k' \in \mathbb{N}$ and where $f(\theta) = \sin\theta$ or $\cos\theta$. We need to compute the intersection of lines $\mathcal{D}_{12} := (S_1S_2)$ and $\mathcal{D}_\theta = (S_0, \mathbf{e}_r)$:

— In the case where $X_1 \neq X_2$, the equations of $\mathcal{D}_{12}$, $\mathcal{D}_\theta$ are such that:

$$\mathcal{D}_{12} \,:\, Y = \frac{Y_1 - Y_2}{X_1 - X_2} X + \frac{X_1\,Y_2 - X_2\,Y_1}{X_1 - X_2}\,; \quad \mathcal{D}_\theta \,:\, Y = \tan\theta\, X.$$

$\mathcal{D}_{12}$ and $\mathcal{D}_\theta$ intersect at (X_θ, Y_θ) such that:

$$X_\theta = \frac{X_1\,Y_2 - X_2\,Y_1}{\mathbf{e}_\theta \cdot \overrightarrow{S_1S_2}} \cos\theta, \quad Y_\theta = \frac{X_1\,Y_2 - X_2\,Y_1}{\mathbf{e}_\theta \cdot \overrightarrow{S_1S_2}} \sin\theta,$$

Let $R_\theta := (X_\theta^2 + Y_\theta^2)^{\frac{1}{2}} = \dfrac{X_1\,Y_2 - X_2\,Y_1}{\mathbf{e}_\theta \cdot \overrightarrow{S_1S_2}}$.

— In the case where $X_1 = X_2$ the equations of $\mathcal{D}_{12}$, $\mathcal{D}_\theta$ are such that:

$$\mathcal{D}_{12} \,:\, X = X_1 \quad \mathcal{D}_\theta \,:\, Y = \tan\theta\, X.$$

$\mathcal{D}_{12}$ and $\mathcal{D}_\theta$ intersect at (X_θ, Y_θ) such that: $X_\theta = X_1$, $Y_\theta = \tan\theta\, X_1$.
Let $R_\theta := (X_\theta^2 + Y_\theta^2)^{\frac{1}{2}} = \dfrac{X_1}{|\cos\theta|}$.

We have to evaluate:

$$\int_{K_\ell} r^\gamma\, f(\theta)\, X^k\, Y^{k'}\, d\,\mathbf{x} = \int_{\theta=\Theta_1}^{\Theta_2} \int_{r=0}^{R(\theta)} r^\gamma f(\theta)\, X(r,\theta)^k\, Y(r,\theta)^{k'}\, r\, d\, r\, f(\theta)\, \mathrm{d}\theta.$$

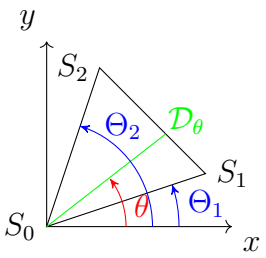


Figure 8.5 – Triangle such that $X_1 \neq X_2$.

Let $\gamma' = \gamma + k + k'$ and $\tilde{f}(\theta) = (\cos\theta)^k\, (\sin\theta)^{k'}\, f(\theta)$.
We get: $\displaystyle\int_{K_\ell} r^\gamma\, f(\theta)\, X^k\, Y^{k'} d\mathbf{x} = \frac{1}{\gamma' + 2} \int_{\Theta_1}^{\Theta_2} R(\theta)^{\gamma'+2}\, \tilde{f}(\theta)\, d\,\theta.$
Let $N_\theta \in \mathbb{N}^\star$, $\theta_0 = \Theta_1$, $\theta_{N_\theta+1} = \Theta_2$, and $\Delta\theta = \frac{\Theta_2 - \Theta_1}{N_\theta+1}$. Using the following trapezoidal rule, it stands:

$$\int_{\theta=\Theta_1}^{\Theta_2} R(\theta)^{\gamma'+2}\, \tilde{f}(\theta)\, d\,\theta \approx \sum_{n=0}^{N_\theta} \frac{\Delta\theta}{2} \left(R(\theta_{n+1})^{\gamma'+2}\, \tilde{f}(\theta_{n+1}) + R(\theta_n)^{\gamma'+2}\, \tilde{f}(\theta_n) \right).$$

We explained how we numerically solved the problem using the Uzawa method combined with a preconditioned conjugate gradient method, where the preconditioner is the matrix representing the pressure norm $\left(\|p\|^2_{L^2(\Omega)} + |p|_J\right)^{1/2}$, which is well suited for this problem. Then, to address the large size of the matrix $\mathbb{A}$, we applied a domain decomposition method with the Schur complement. We also detailed how the right-hand side was computed when $\mathbf{f} \notin \mathbf{L}^2(\Omega)$.

Chapter 9

Numerical illustrations

Contents

In this chapter, in Section 9.1, we confirm the theoretical results of Chapter 7 on both convex and non-convex domains. In Section 9.2, we demonstrate the effectiveness of the new method introduced earlier when ν is very small. Finally, in Section 9.3, we show the influence of the mesh regularity on the numerical results.

9.1 Validation of the theoretical convergence rates

Denote by $\mathbf{P}^{k_\mathbf{u}}_{dg} - P^{k_p}_{dg}$ the DG scheme s.t. $(\mathbf{u}_h, p_h) \in \mathbf{P}^{k_\mathbf{u}}_{dg} \times P^{k_p}_{dg}$. We recall that Π_h, (resp. π_h) is the L^2-orthogonal projection onto $\mathbf{X}_h$ (resp. X_h). The errors are given by:

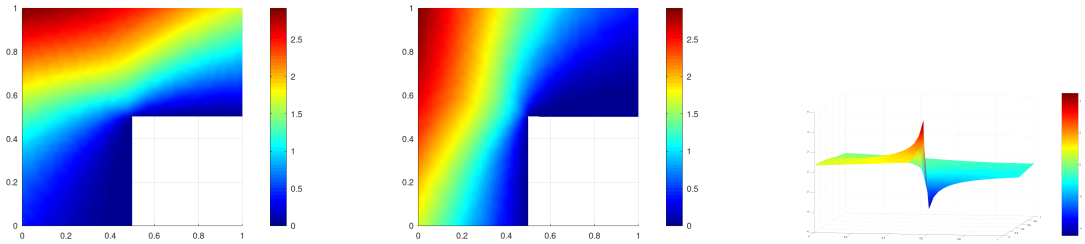
$$\begin{aligned} \varepsilon_0(\mathbf{u}_h) &:= \|\Pi_h\mathbf{u} - \mathbf{u}_h\|_{\mathbf{L}^2(\Omega)} \quad \text{and} \quad \varepsilon_0(p_h) := \|\pi_h p - p_h\|_{L^2(\Omega)} \\ \varepsilon_1(\mathbf{u}_h) &:= \|\mathbf{Grad}_h(\Pi_h\mathbf{u} - \mathbf{u}_h)\|_{\mathbf{L}^2(\Omega)} \quad \text{and} \quad \varepsilon_d(\mathbf{u}_h) := \|\operatorname{div}_h \mathbf{u}_h\|_{\mathbf{L}^2(\Omega)} \end{aligned}$$

The convergence rates of $\varepsilon_0(\mathbf{u}_h)$ and $\varepsilon_0(p_h)$ are denoted by $\tau_\mathbf{u}$ and τ_p.
The results are given here for $\nu = 1$.
For $(\alpha, \beta) = (0.6, -0.4)$. Then $(\mathbf{u}, p) \in \dot{\mathbf{H}}^{1+\sigma_{\max}}(\Omega) \times \dot{H}^{\sigma_{\max}}(\Omega)$, with $\sigma_{\max} = 0.6$. Hence $(\tau_\mathbf{u}, \tau_p) = (1 + \sigma_{\max}\,,\, \sigma_{\max})$.

For the non-convex domain, consider Problem (5.7) with $\Omega = (0,1)^2 \backslash [\frac{1}{2}, 1] \times [0, \frac{1}{2}]$ and $(\mathbf{u}, p) \in \dot{\mathbf{H}}^{1+\sigma_{\max}}(\Omega) \times \dot{H}^{\sigma_{\max}}(\Omega)$ given in [37]. It is studied in [97] and [19], with $\sigma_{\max} = \alpha \approx 0.54$, $\alpha_{\pm} = \alpha \pm 1$ and $\omega = 3\pi/2$:

$$\begin{cases} \mathbf{u}(r\,,\,\theta) &= r^{\alpha} \begin{pmatrix} \cos(\theta)\psi\prime(\theta) + \alpha_+ \sin(\theta)\psi(\theta) \\ \sin(\theta)\psi\prime(\theta) - \alpha_+ \cos(\theta)\psi(\theta) \end{pmatrix} \\ \\ p(r\,,\,\theta) &= -r^{\alpha_-}\,(\alpha_-)^{-1}\,((\alpha_+)^2\psi\prime(\theta) + \psi^{(3)}(\theta)) \end{cases} \tag{9.1}$$

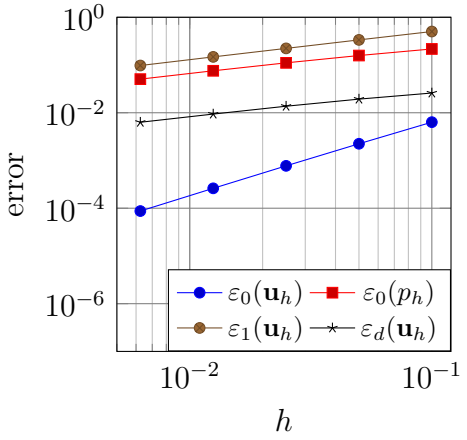
Where $\psi(\theta) = (\alpha_+)^{-1}\,\sin(\alpha_+\,\theta)\cos(\alpha\,\omega) - \cos(\alpha_+\,\theta) - (\alpha_-)^{-1}\,\sin(\alpha_-\,\theta)\cos(\alpha\,\omega) + \cos(\alpha_-\,\theta)$. We expect $\tau_{\mathbf{u}} = 2\,\alpha \approx 1.09$ and $\tau_p = \alpha$.



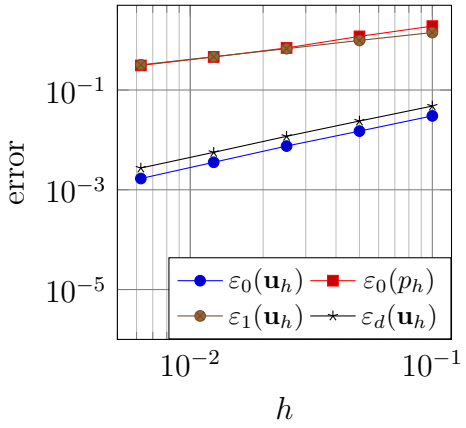
(a) Velocity $\mathbf{u}_x$ (b) Velocity $\mathbf{u}_y$ (c) Pressure p

Figure 9.1 – Velocity components and pressure field.

Let us consider the $\mathbf{P}^1_{dg} - P^1_{dg}$ scheme. Figures 9.2b and 9.2a represent $\varepsilon^{\nu}_0(\mathbf{u}_h)$, $\varepsilon^{\nu}_1(\mathbf{u}_h)$, $\varepsilon^{\nu}_2(\mathbf{u}_h)$ and $\varepsilon^{\nu}_0(p_h)$ against the mesh size.



(a) Test-case (8.5) (Ω convex), $(\alpha, \beta) = (0.6, -0.4), \nu = 1$



(b) Test-Case (9.1) (Ω non-convex).

Figure 9.2 – Plots of $\varepsilon_0(\mathbf{u}_h)$ and $\varepsilon_0(p_h)$ against the mesh size with $\mathbf{P}^1_{dg} - P^1_{dg}$ scheme.

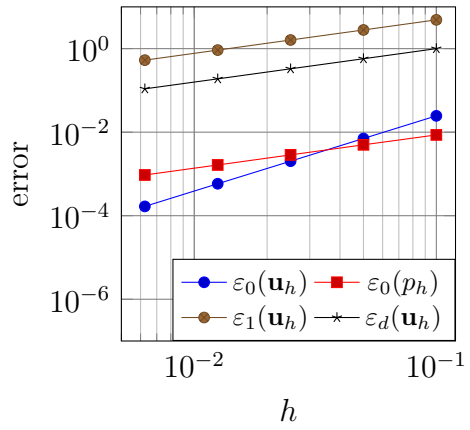
Convergence rates between two consecutive meshes are recorded in Table 9.1. In both test-cases, we observe a good agreement with the theoretical ones (theorem 7.6).

test-case	τ	theory	mesh 1-2	mesh 2-3	mesh 3-4	mesh 4-5
(8.5) (Ω convex)	$\tau_{\mathbf{u}}$	1.6	1.50	1.54	1.57	1.58
	$\tau_{\mathbf{Grad\,u}}$	0.6	0.58	0.59	0.6	0.6
	τ_p	0.6	0.46	0.51	0.55	0.57
(9.1) (Ω non-convex)	$\tau_{\mathbf{u}}$	1.09	1.02	0.99	1.08	1.08
	$\tau_{\mathbf{Grad\,u}}$	0.54	0.61	0.56	0.55	0.54
	τ_p	0.54	0.68	0.77	0.60	0.56

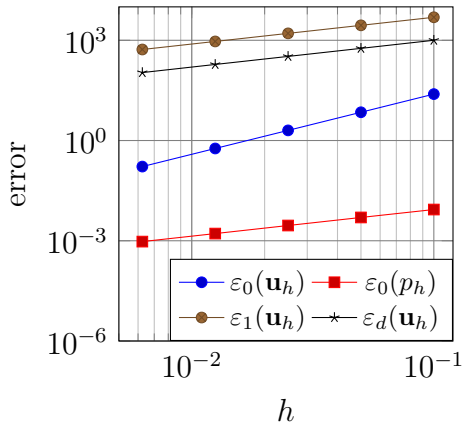
Table 9.1 – Convergence rates computed between two consecutive meshes.

The test case corresponding to the non-convex domain (9.1) is valid only for $\nu = 1$. Otherwise, the right-hand side $\mathbf{f}$ would not be zero and would not necessarily belong to $\mathbf{L}^2(\Omega)$. On the other hand, the test case associated with the convex domain (8.5) is valid for any ν.
Therefore, we choose different values for the parameters α, β, and ν. For example, for $\alpha = 0.8$ and $\beta = -0.2$, we obtain $s = 0.8$. The convergence rate for the velocity is then $\tau_{\mathbf{u}} = 1.8$, and for the velocity gradient and the pressure we have $\tau_{\mathbf{Grad\,u}} = \tau_p = 0.8$. These results hold for any ν; in our tests, we chose $\nu = 10^{-3}$ and $\nu = 10^{-6}$:
Let us consider the $\mathbf{P}^1_{dg} - P^1_{dg}$ scheme. Figures 9.2b and 9.2a represent $\varepsilon^\nu_0(\mathbf{u}_h)$, $\varepsilon^\nu_1(\mathbf{u}_h)$, $\varepsilon^\nu_2(\mathbf{u}_h)$ and $\varepsilon^\nu_0(p_h)$ against the mesh size. Observe that the errors are lower when the domain is

convex.



(a) Test-case (8.5) (Ω convex), $(\alpha, \beta) = (0.8, -0.2), \nu = 10^{-3}$



(b) Test-case (8.5) (Ω convex), $(\alpha, \beta) = (0.8, -0.2), \nu = 10^{-6}$

Figure 9.3 – Plots of $\varepsilon_0(\mathbf{u}_h)$ and $\varepsilon_0(p_h)$ against the mesh size.

test-case	τ	theory	mesh 1-2	mesh 2-3	mesh 3-4	mesh 4-5
$(\alpha, \beta) = (0.8, -0.2), \nu = 10^{-3}$	$\tau_{\mathbf{u}}$	1.8	1.50	1.8	1.8	1.8
	$\tau_{\mathbf{Grad\,u}}$	0.8	0.8	0.8	0.8	0.8
	τ_p	0.8	0.79	0.8	0.8	0.8
$(\alpha, \beta) = (0.8, -0.2), \nu = 10^{-6}$	$\tau_{\mathbf{u}}$	1.8	1.8	1.8	1.8	1.8
	$\tau_{\mathbf{Grad\,u}}$	0.8	0.8	0.8	0.8	0.8
	τ_p	0.8	0.78	0.79	0.8	0.8
$(\alpha, \beta) = (0.55, -0.45), \nu = 10^{-2}$	$\tau_{\mathbf{u}}$	1.55	1.54	1.54	1.55	1.55
	$\tau_{\mathbf{Grad\,u}}$	0.55	0.55	0.55	0.55	0.55
	τ_p	0.55	0.54	0.55	0.55	0.55

Table 9.2 – Convergence rates computed between two consecutive meshes for the Test-case (8.5) (Ω convex), with different values of α, β and ν.

Although this is not covered by our theory, in order to delve deeper into the convex case, we take a value strictly lower than $\frac{1}{2}$, for example, $s = 0.4$. This corresponds to the test case (8.5) where $\alpha = 0.4$ and $\beta = -0.6$.

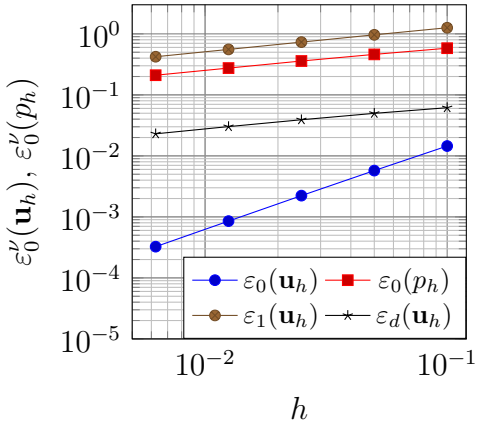


Figure 9.4 – Test-case $s = 0.4$ (Ω convex)

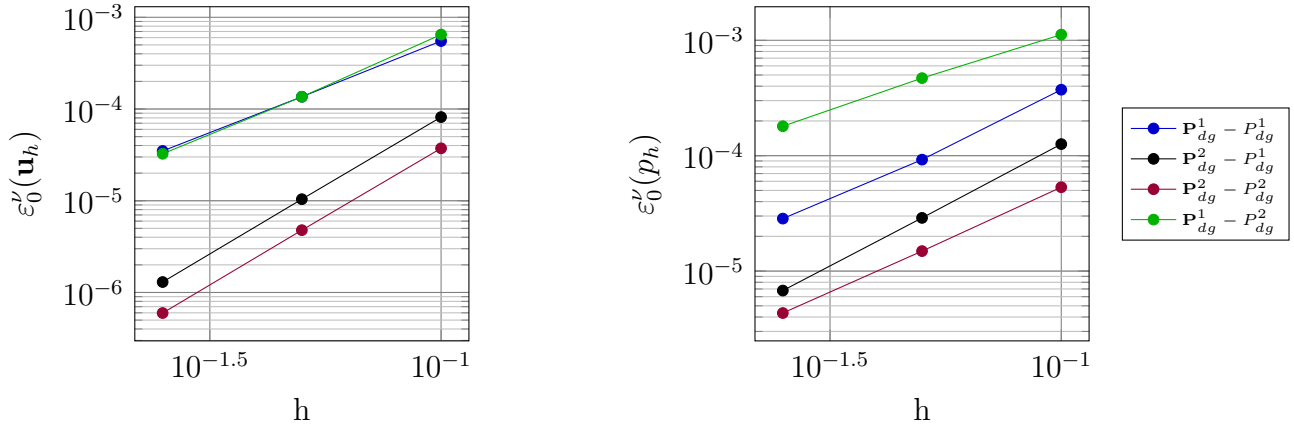
In Figure 9.4, we observe that the convergence rate stabilizes around 1.4 for velocity and 0.4 for pressure; they thus behave like $1 + s$ and s, which is similar to the case where $s > \frac{1}{2}$. We have not developed a theory for the case where $s < \frac{1}{2}$ because it is impossible to break the duality brackets in equation (6.4).

9.2 Comparison of Different Numerical Schemes

In this section, we compare the numerical results obtained using the classical discontinuous Galerkin scheme, where $k_p \leq k_{\mathbf{u}}$, with those obtained using the new numerical scheme, where $k_p = k_{\mathbf{u}} + 1$.

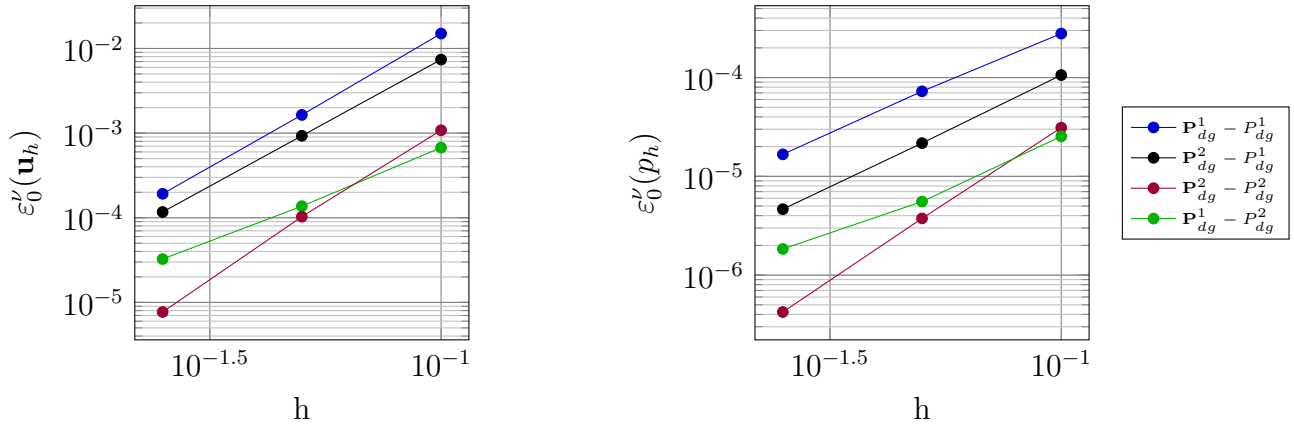
Consider test-case (8.5) with $(\alpha, \beta) = (6, 3)$. We compare DG schemes with different pairs of polynomial orders $(k_{\mathbf{u}}, k_p)$.

On figure 9.5 we plot the errors $\varepsilon_0^\nu(\mathbf{u}_h)$ and $\varepsilon_0^\nu(p_h)$ against the mesh size with $\nu = 10^{-2}$.



(a) $\varepsilon_0^\nu(\mathbf{u}_h)$ (b) $\varepsilon_0^\nu(p_h)$

Figure 9.5 – Plots of $\epsilon_0^\nu(\mathbf{u}_h)$, and $\epsilon_0^\nu(p_h)$ against mesh size, with $\nu = 10^{-2}$.



(a) $\varepsilon_0^\nu(\mathbf{u}_h)$ (b) $\varepsilon_0^\nu(p_h)$

Figure 9.6 – Plots of $\epsilon_0^\nu(\mathbf{u}_h)$, and $\epsilon_0^\nu(p_h)$ against mesh size, with $\nu = 10^{-4}$.

test-case	τ	$\mathbf{P}^1_{dg} - P^1_{dg}$	$\mathbf{P}^2_{dg} - P^1_{dg}$	$\mathbf{P}^1_{dg} - P^2_{dg}$	$\mathbf{P}^2_{dg} - P^2_{dg}$
$\nu = 1$	$\tau_{\mathbf{u}}$	1.96	2.99	2	2.97
	$\tau_{\mathbf{Grad\,u}}$	1	1.95	0.94	1.93
	τ_p	1	1.94	1.3	1.73
$\nu = 10^{-2}$	$\tau_{\mathbf{u}}$	1.97	2.99	2	3
	$\tau_{\mathbf{Grad\,u}}$	1.02	1.98	0.94	1.97
	τ_p	2	2	1.4	1.77
$\nu = 10^{-4}$	$\tau_{\mathbf{u}}$	3.1	2.99	2.07	3.74
	$\tau_{\mathbf{Grad\,u}}$	2	1.98	0.96	2.83
	τ_p	1.94	2.22	1.6	3.14
$\nu = 10^{-6}$	$\tau_{\mathbf{u}}$	3.11	2.99	3.70	3.74
	$\tau_{\mathbf{Grad\,u}}$	2.03	1.99	2.92	2.86
	τ_p	2.11	2.22	2.97	3.15

Table 9.3 – Convergence rates computed between meshes 2 and 3, for $\alpha = 6$, $\beta = 3$, and several values of ν.

In Figures 9.5, 9.6, and 9.7, we observe that when ν is small, the $\mathbf{P}^1-P^2$ scheme is more effective than the other schemes presented.

Regarding the experimental convergence rates reported in table 9.3, we first recall that the expected convergence rates for the $\mathbf{P}^k - P^k$ and the $\mathbf{P}^k - P^{k-1}$ pairs are $k+1$ for $\tau_{\mathbf{u}}$ and k for both $\tau_{\mathbf{Gradu}}$ and τ_p. Moreover, a closer look at the proof of Theorem 7.5 tells us that the expected rates of convergence for the $\mathbf{P}^k - P^{k+1}$ pair should also be k for both $\tau_{\mathbf{Gradu}}$ and τ_p.

So, for $\nu = 1$ and $\nu = 10^{-2}$, the expected theoretical rates are confirmed, except for the pressure convergence in the $\mathbf{P}^1$–P^1 scheme and $\nu = 10^{-2}$, where we observe a superconvergence phenomenon. Observe that the pressure converges slightly faster than expected for the $\mathbf{P}^1$–P^2 scheme, and slightly slower than expected for the $\mathbf{P}^2-P^2$ scheme.

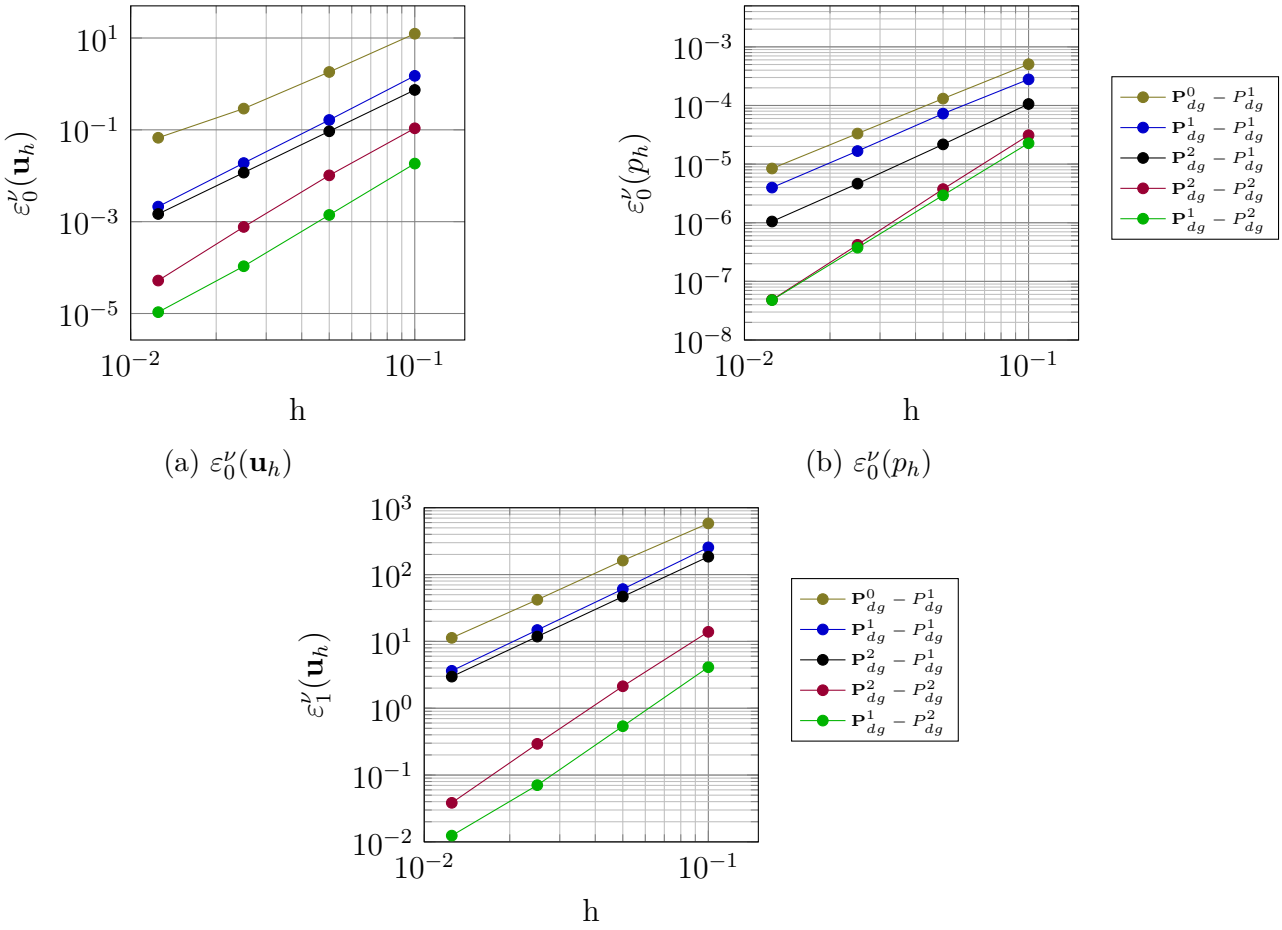
For $\nu = 10^{-4}$ or $\nu = 10^{-6}$, we again observe superconvergence for all schemes except for the $\mathbf{P}^2-P^1$ scheme. Superconvergence can be explained most likely because we are in the pre-asymptotic convergence regime (h not small enough); with sufficient refinement, the convergence rate should tend toward the expected asymptotic behaviour.

Now we fix $\nu = 10^{-6}$ which is close to the kinematic viscosity of the pressurized water in nuclear reactor core and. In order to compare DG schemes with different pairs of polynomial orders $(k_{\mathbf{u}}, k_p)$, we plot the error $\varepsilon_0^{\nu}(\mathbf{u}_h)$, $\varepsilon_1^{\nu}(\mathbf{u}_h)$ (resp. $\varepsilon_0^{\nu}(p_h)$) against the mesh size on Fig. 9.7 or against the CPU time on Fig. 9.8.

— $\mathbf{P}^1_{dg} - P^1_{dg}$ *vs* $\mathbf{P}^2_{dg} - P^1_{dg}$: for a given CPU time, the error $\varepsilon^\nu_0(\mathbf{u}_h)$ and the error $\varepsilon^\nu_0(p_h)$ are similar for both schemes (with a slight advantage for $\mathbf{P}^2_{dg} - P^1_{dg}$ on the pressure error). However, using the $\mathbf{P}^1_{dg} - P^1_{dg}$ scheme requires less memory footprint and seems therefore more efficient.

— $\mathbf{P}^1_{dg} - P^2_{dg}$ *vs* $\mathbf{P}^2_{dg} - P^2_{dg}$: for a given CPU time, the error $\varepsilon^\nu_0(p_h)$ are similar for both schemes, and the error $\varepsilon^\nu_0(\mathbf{u}_h)$ is better using the $\mathbf{P}^1_{dg} - P^2_{dg}$ scheme, which therefore seems more efficient.

— $\mathbf{P}^0 - P^1$ scheme combines a finite-volume discretization for the velocity, where the definition of the gradient is applied on the interfaces as indicated in (6.52), with a discontinuous Galerkin finite-element method of order 1 for the pressure. This scheme is numerically stable. It is less accurate than the other scheme but significantly faster than the other methods. These results should be regarded as preliminary.

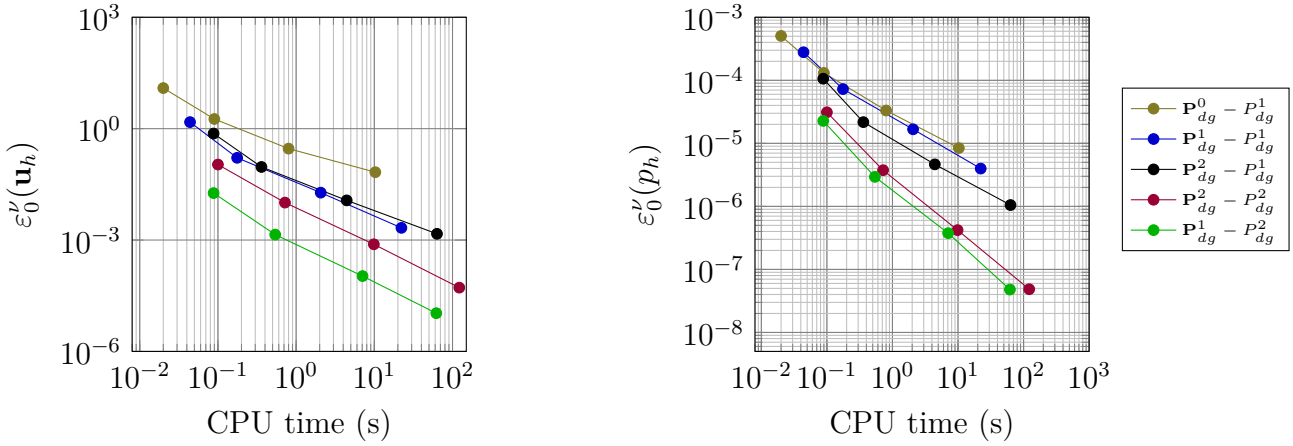


(a) $\varepsilon_0^\nu(\mathbf{u}_h)$

(b) $\varepsilon_0^\nu(p_h)$

(c) $\varepsilon_1^\nu(\mathbf{u}_h)$

Figure 9.7 – Plots of $\varepsilon_0^\nu(\mathbf{u}_h)$, $\varepsilon_1^\nu(\mathbf{u}_h)$ and $\varepsilon_0^\nu(p_h)$ against the mesh size.



(a) $\varepsilon_0^\nu(\mathbf{u}_h)$ (b) $\varepsilon_0^\nu(p_h)$

Figure 9.8 – Plots of $\varepsilon_0^\nu(\mathbf{u}_h)$ and $\varepsilon_0^\nu(p_h)$ for $\nu = 10^{-6}$ against CPU time (s).

Observe that for a given CPU time, the $\mathbf{P}^1_{dg} - P^2_{dg}$ scheme yields better results. Additionally, it exhibits the fastest convergence rate.

9.3 Impact of mesh regularity on the stability and accuracy of the DG scheme $\mathbf{P}^1_{dg}$–P^1_{dg}

In order to analyze the impact of mesh regularity on the accuracy and stability of the numerical scheme, we construct two 2D meshes with the same number of degrees of freedom: one is a regular (isotropic) mesh, and the other is strongly anisotropic.

We compare the numerical results obtained using these two types of meshes for the test case (8.5), with parameters $\nu = 10^{-2}$, $\alpha = 4$, and $\beta = 2$. In this case, the exact solution $(\mathbf{u}, p)$ belongs to the space $\mathbf{H}^2(\Omega) \times H^1(\Omega)$, which implies that both the solution and the domain are regular. This study is not only intended to observe differences in terms of accuracy, but also aims to analyze the numerical stability of the method , which is reflected in the convergence rate.

9.3. Impact of mesh regularity on the stability and accuracy of the DG scheme $\mathbf{P}^1_{dg}$–P^1_{dg}

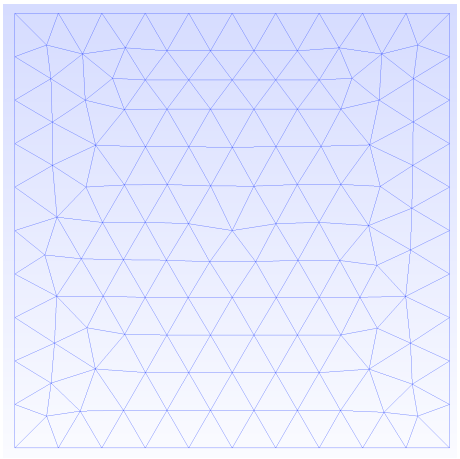


Figure 9.9 – Regular (isotropic) mesh

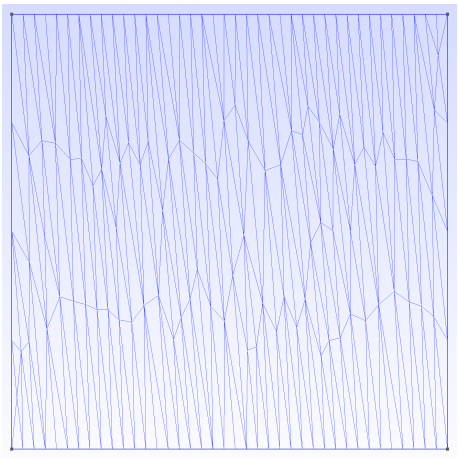


Figure 9.10 – Anisotropic mesh

The figures above show the two meshes used in the numerical simulations. For the anisotropic mesh, the mesh regularity parameter σ_ℓ varies in the interval [4.3, 169.69], while for the regular mesh, we have $\sigma_\ell \in [2, 2.64]$.

Numerical simulations are carried out on 5 mesh levels. The first level corresponds to the meshes shown above, and each triangle is refined by subdividing it into four at each level.

We first present the results obtained with the anisotropic mesh using $\mathbf{P}^1_{dg} - P^1_{dg}$ scheme:

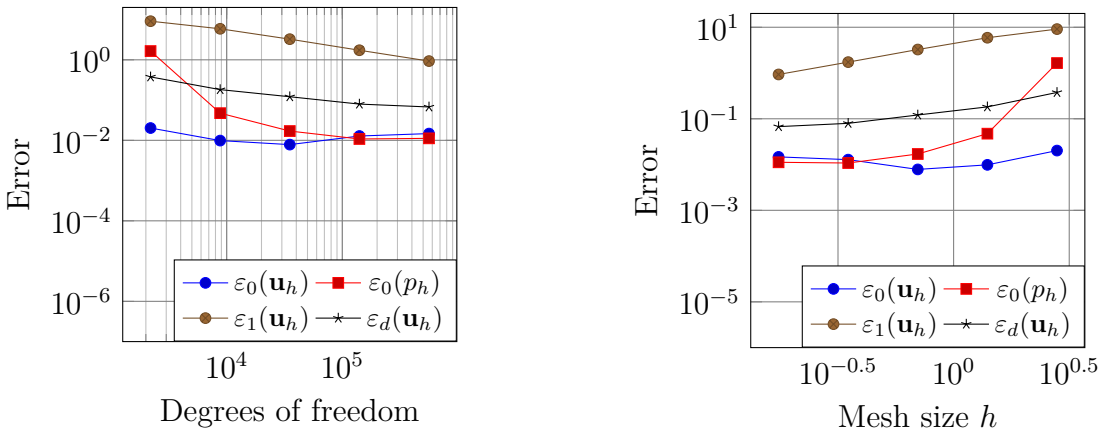


Figure 9.11 – Error plots with respect to the mesh size and the number of degrees of freedom for the anisotropic mesh.

The scheme does not converge as h tends to zero. This is consistent with the dependence of the coercivity constant on the mesh regularity.

Now we present the results obtained with the isotropic mesh:

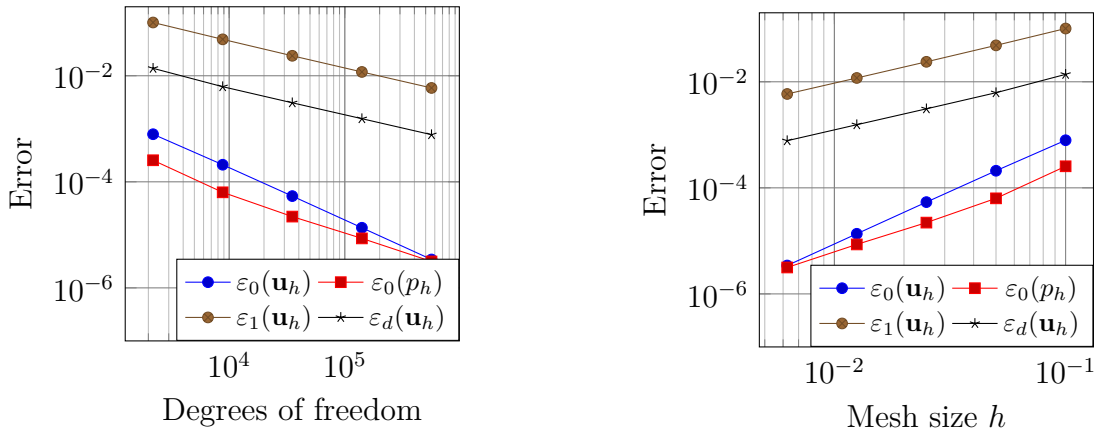


Figure 9.12 – Error plots with respect to the mesh size and the number of degrees of freedom for the isotropic mesh.

It is clear that with the anisotropic mesh, the discontinuous Galerkin scheme P^1_{dg}–P^1_{dg} is less stable and less accurate compared to the results obtained with the isotropic mesh. This observation is in line with theoretical expectations, since the coercivity constant (which ensures stability) depends on the mesh regularity.

9.3. Impact of mesh regularity on the stability and accuracy of the DG scheme $\mathbf{P}^1_{dg}$–P^1_{dg}

In this chapter, we perform a numerical validation of the theoretical convergence rate results. We propose a test case in a convex domain, where the solution regularity depends on the parameters α and β, as well as a test case from the literature in a non-convex domain. The numerical experiments confirm that the obtained convergence rates are optimal.

In the second section, we compare the numerical results of several DG-type schemes and demonstrate that the new $\mathbf{P}^k$–P^{k+1} scheme, with $k \geq 1$, is the most robust when the viscosity ν is very small.

Finally, we investigate the effect of mesh regularity on the numerical schemes and show that mesh irregularity significantly affects both the accuracy and stability of the numerical method.

Part II

DGFEM with exponential Basis

We carry out this work with the goal of integrating it into the numerical simulation of the Navier-Stokes equations, which model many physical phenomena related to fluid flow. Our approach is progressive, ensuring a thorough understanding and rigorous validation at each step. We begin with the advection-diffusion equations, which capture the balance between convective transport and diffusion. These equations form a fundamental starting point, as they allow us to test the robustness of our method in simple yet representative scenarios.
Next, we move on to the Stokes equations, which describe flows at low Reynolds numbers, where viscous effects dominate. Once these equations are mastered, we address the Oseen equations, a generalization of the Stokes equations that incorporates linearized convective terms. This intermediate step is crucial to verify the stability and accuracy of our approach in regimes where convective effects begin to play a significant role.
Finally, this progression naturally leads us to the full Navier-Stokes equations, which model more complex phenomena, including the nonlinear interaction between convective and viscous terms. This gradual approach ensures that our non-polynomial discretization method is well-suited and effective for addressing a wide range of problems, from diffusion-dominated to convection-dominated regimes.

Chapter 10

Advection-Diffusion Problem

Contents

10.1 Brief Bibliography

There are many works on the advection-diffusion problem, both in the continuous and discrete settings. For the continuous problem, classical references include the work of Ladyzhenskaya, Solonnikov, and Ural'tseva [69], where the existence and regularity of solutions in Sobolev spaces are discussed, as well as more recent publications such as those by Renardy and Rogers [93] or Evans [51], which provide a rigorous analytical framework adapted to parabolic equations. On the numerical side, many studies have focused on this problem. I mention a few where the advection-diffusion equation is discretized using the discontinuous Galerkin (DG) method, such as the work of Georgoulis, Hall and Houston [56]. Other works focus on the hp-analysis of the method, like those by Houston, Schwab, and Süli [64]. Some very applied contributions can also be mentioned, such as Mazaheri and Nishikawa [78] for high-order stable schemes, or Lasser and Toselli [71], who designed an overlapping domain decomposition preconditioner for DG discretizations of advection-diffusion problems.

10.2 Continuous Advection-Diffusion problem

Let Ω be a connected and bounded domain in $\mathbb{R}^d$, with Lipschitz boundary $\partial\Omega$ for $d = 2$ and polyhedral boundary $\partial\Omega$ for $d = 3$. We consider the following problem:

$$\text{Find } u \in H^1(\Omega) \text{ such that : } \left\{ \begin{array}{rcl} -\nu\,\Delta u + (\boldsymbol{\beta}\cdot\mathbf{grad}\,)\,u & = & f, \\ u_{|\partial\Omega} & = & g, \end{array} \right. \tag{10.1}$$

with $\boldsymbol{\beta}$ is a modeling parameter that we suppose for the moment to be bounded: $\boldsymbol{\beta} \in \mathbf{L}^\infty(\Omega)$. The unknown u represents the transported quantity, which may be, for example, the concentration of a chemical species, the temperature, or, in fluid mechanics, a (component of a) velocity field,.... The constant parameter $\nu > 0$ represents the diffusion coefficient (or kinematic viscosity). The vector field $f \in L^2(\Omega)$ represents the external force field on u, and we assume that $g \in H^{\frac{3}{2}}(\Omega)$.
By a lifting, we can transform problem (10.1) into a problem with a homogeneous Dirichlet condition. Let $u_\perp \in H^1(\Omega)$ be such that

$$-\Delta u_\perp = 0, \quad \text{and} \quad u_{\perp|\partial\Omega} = g. \tag{10.2}$$

This problem has a unique solution, see Theorem 5.1 and its gradient is in $\mathbf{L}^2(\Omega)$; since $\boldsymbol{\beta}$ is bounded, defining $f_\perp := f - (\boldsymbol{\beta}\cdot\mathbf{grad}\,)\,u_\perp$, we have $f_\perp \in L^2(\Omega)$. Then problem (10.1) is equivalent to the following problem with $\tilde{u} = u - u_\perp$:

$$\text{Find } \tilde{u} \in H^1(\Omega) \text{ such that : } \left\{ \begin{array}{rcl} -\nu\,\Delta \tilde{u} + (\boldsymbol{\beta}\cdot\mathbf{grad}\,)\,\tilde{u} & = & f_\perp, \\ \tilde{u}_{|\partial\Omega} & = & 0. \end{array} \right. \tag{10.3}$$

Thus we can always assume, without loss of generality, that $g = 0$.

We consider the following continuous bilinear form:

$$a(\cdot,\cdot) : \left\{ \begin{array}{rcl} H_0^1(\Omega)\times H_0^1(\Omega) & \to & \mathbb{R} \\ u, v & \mapsto & \nu\,(u\,,\,v)_{H_0^1(\Omega)} + ((\boldsymbol{\beta}\cdot\mathbf{grad}\,)\,u\,,\,v)_{L^2(\Omega)} \end{array} \right. \tag{10.4}$$

We can express problem (10.1) equivalently as follows: Find $u \in H_0^1(\Omega)$ such that for all $v \in H_0^1(\Omega)$:

$$a(u, v) = (f, v)_{L^2(\Omega)}. \tag{10.5}$$

The fact that $\boldsymbol{\beta}$ is bounded is sufficient to verify the hypotheses of [45, Theorem 2.1], which

tells us that problem (10.5) is well-posed. Therefore, there exists a unique solution $u \in H_0^1(\Omega)$ of problem (10.1).

Remark 10.1 *In the case we are most interested in, when* $\operatorname{div} \boldsymbol{\beta} = 0$*, we can show in an easier way that problem* (10.1) *is well-posed. It is sufficient to make a variational formulation for the advection term* $(\boldsymbol{\beta} \cdot \mathbf{grad})u$ *that is consistent and antisymmetric. Let* $(u, v) \in H_0^1(\Omega) \times H_0^1(\Omega)$*, we define :*

$$t(u, v) := ((\boldsymbol{\beta} \cdot \mathbf{grad})u, v)_{L^2(\Omega)}, \quad \forall v \in H_0^1(\Omega)$$

and we have $t(v, v) = 0, \forall v \in H_0^1(\Omega)$*.*

10.3 Discretization of the Advection-Diffusion problem

We will present the discretization using the method of discontinuous Galerkin finite elements for the Advection-Diffusion problem (10.1). For the Laplacian problem, this has already been done in section 6.1. As the discretization of the Laplacian $a_h(\cdot, \cdot)$ is coercive, our goal is to have a discretization for the advection term $(\boldsymbol{\beta} \cdot \mathbf{grad})u$ that is consistent and controllable in terms of stability. Moreover, we specialize now the vector $\boldsymbol{\beta}$ to be constant per cell, with its normal component continuous across the cells. Such vectors can be for example constructed in the following way: take a P^1-Lagrange vector field $\mathbf{\Phi_h}$ defined on the considered mesh and set on each cell K_ℓ: $\boldsymbol{\beta}_{|K_\ell} = \nabla \times (\mathbf{\Phi_h})_{|K_\ell}$. The constructed vector field $\boldsymbol{\beta}$ is evidently constant per cell (and thus in $\mathbf{L}^\infty(\Omega)$) and it can be easily shown that its normal components are continuous across the edges (in 2D) or the faces (in 3D).
We recall that k_u is the discretization order of u, and that $X_h = [P_{disc}^{k_u}(\mathcal{T}_h)]^d$. Let $(u_h, v_h) \in X_h \times X_h$, we define $t_h^0(u_h, v_h) := \sum_{\ell \in \mathcal{I}_K} ((\boldsymbol{\beta} \cdot \mathbf{grad})u_h, v_h))_{L^2(K_\ell)}$. Since the normal component of $\boldsymbol{\beta}$ is supposed continuous across the mesh interfaces, we get that for all $v_h \in X_h$, we have (note that we could simply write $\boldsymbol{\beta} \cdot \mathbf{n}_f$, but we keep the average sign for later purposes)

$$t_h^0(v_h, v_h) = \sum_{f \in \mathcal{I}_F} (\{\boldsymbol{\beta}\} \cdot \mathbf{n}_f, [\![v_h]\!] \cdot \{v_h\})_{L^2(F_f)}.$$

So we define:

$$t_h(u_h, v_h) := t_h^0(u_h, v_h) - \sum_{f \in \mathcal{I}_F} (\{\boldsymbol{\beta}\} \cdot \mathbf{n}_f, [\![u_h]\!] \cdot \{v_h\})_{L^2(F_f)}. \tag{10.6}$$

Lemma 10.1 *By construction, we have for all* $u_h \in X_h$*,* $t_h(u_h, u_h) = 0$*.*

Now we can propose the discrete variational formulation of the advection-diffusion problem (10.1):

$$\begin{cases} \text{Find } u_h \in X_h \text{ such that for all } v_h \in X_h, \\ \quad a_h(u_h, v_h) + t_h(u_h, v_h) = (f, v_h)_{L^2(\Omega)} \end{cases} \tag{10.7}$$

We denote by $c_h(u_h, v_h) := a_h(u_h, v_h) + t_h(u_h, v_h)$.

Remark 10.2 *At the discrete level, the condition $\nabla \cdot \boldsymbol{\beta} = 0$ can be enforced by introducing a vector potential in a Nédélec ($H(rot)$-conforming) space whose curl belongs to a Raviart–Thomas ($H(\mathrm{div})$-conforming) space. This yields a discrete field that is naturally $H(\mathrm{div})$-conforming and (discretely) divergence-free.*

We recall the definition of the norm $\|(\cdot,\cdot)\|_{sip}$ given in Definition 6.3. In what follows, to prove the continuity of the discrete form $c_h(\cdot,\cdot)$, we need the following theorem:

Theorem 10.2 ([40, Thm. 5.3] Discrete Sobolev embeddings) *There exist two positive constants σ_2 and σ_4 such that:*

$$\forall v_h \in \mathbb{P}_d^k(\mathcal{T}_h), \qquad \|v_h\|_{L^2(\Omega)} \leq \sigma_2 \|v_h\|_{sip} \qquad \text{and} \qquad \|v_h\|_{L^4(\Omega)} \leq \sigma_4 \|v_h\|_{sip}. \tag{10.8}$$

The quantities σ_2 and σ_4 additionally depend on Ω, d, k and mesh regularity σ.

Theorem 10.3 *Under the same conditions as Theorem 6.3, the bilinear form $c_h(\cdot,\cdot)$ is coercive and continuous on $X_h \times X_h$:*

$$\begin{aligned} \forall v_h \in X_h,\ c_h(v_h, v_h) &\geq \tfrac{1}{2}\|v_h\|_{sip}^2, \\ \forall (u_h,\, v_h) \in X_h \times X_h,\ c_h(u_h,\, v_h) &\leq C_{cnt}\|u_h\|_{sip}\|v_h\|_{sip}. \end{aligned}$$

Proof. For coercivity, we apply Theorem 6.3 , we have $a_h(u_h, u_h) \geq \frac{1}{2}\|u_h\|_{sip}^2$, and according to Lemma 10.1, we have $t_h(u_h, u_h) = 0$. Therefore, we obtain:

$$c_h(u_h, u_h) \geq \frac{1}{2}\|u_h\|_{sip}^2.$$

To establish continuity, it remains to prove the continuity of $t_h(\cdot,\cdot)$. We have $t_h(u_h, v_h) = \mathcal{I}_1 + \mathcal{I}_2$, with:

$$\begin{aligned} \mathcal{I}_1 &:= (\boldsymbol{\beta} \cdot \mathbf{grad}\, u_h, v_h)_{L^2(\Omega)}, \\ \mathcal{I}_2 &:= -\sum_{f \in \mathcal{I}_F} \left(\{\boldsymbol{\beta}\} \cdot \mathbf{n}_F, [\![u_h]\!] \cdot \{v_h\} \right)_{L^2(F)}. \end{aligned}$$

Using inequality (10.8), for the first term we get

$$|\mathcal{I}_1| \leq \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)} \|\mathbf{Grad}_h u_h\|_{L^2(\Omega)} \|v_h\|_{L^2(\Omega)} \leq \sigma_2 \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)} \|u_h\|_{sip} \|v_h\|_{sip}$$

We recall that $|u_h|_J^2 = \sum_{f \in \mathcal{I}_F} h_f^{-1} \|[\![u_h]\!]\|_{L^2(F_f)}^2$.
For $\mathcal{I}_2$, using the Cauchy–Schwarz inequality, Theorem 10.2, and corollary 4.3, we get:

$$\begin{aligned} |\mathcal{I}_2| &\leq \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)} \sum_{f \in \mathcal{I}_F} |(h_f^{-\frac{1}{2}} [\![u_h]\!], h_f^{\frac{1}{2}} \{v_h\})_{L^2(F_f)}|, \\ &\leq \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)} \left(\sum_{\ell \in \mathcal{I}_K} h_\ell \|v_h\|_{L^2(\partial K_\ell)}^2 \right)^{\frac{1}{2}} |u_h|_J, \\ &\leq C_{d,k,2}\, \sigma^{\frac{1}{2}}\, \sigma_2 \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)} \|u_h\|_{sip} \|v_h\|_{sip}. \end{aligned}$$

Using the continuity result from Theorem 6.3 component by component, we obtain:

$$c_h(u_h, v_h) \leq C_{cont} \, \|u_h\|_{sip} \, \|v_h\|_{sip},$$

where

$$C_{cont} = C_{bnd} + \left(C_{d,k,2}\, \sigma^{\frac{1}{2}} + 1 \right) \sigma_2 \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)}.$$

Therefore, problem (10.7) is well-posed in the sense of Hadamard.

□

These problems are especially interesting in the advection-dominated case ($\nu \ll |\boldsymbol{\beta}|$). It is generally acknowledged that there is a connection between the advection-diffusion evolution problem and the incompressible Navier-Stokes equations. From this perspective, designing stable numerical methods for the simplified advection-diffusion model can serve as a good criterion for selecting appropriate methods for the incompressible Navier-Stokes problem.

In this chapter, we studied the advection–diffusion problem in both the continuous and discrete settings. We showed that the problem is well-posed in the continuous case, and we performed a discontinuous Galerkin finite element discretization while preserving consistency and coercivity.

Chapter 11

Non-Polynomial basis

Contents

In sections 11.1, 11.2 and 11.3, we introduce a new idea in discontinuous Galerkin finite elements, where we propose an exponential basis in one, two and three dimensions and explain the motivation behind it. Moreover in section 11.4 we compute the parameter $C_{\mathbb{NP}}$, which is required to ensure the coercivity of the bilinear form for the Poisson problem. In the numerical results of chapter 14, we demonstrate the effectiveness of this basis, especially for the advection–diffusion and Oseen problems.

11.1 Non-polynomial basis in 1d

We propose a new non-polynomial discretization space designed to exploit the properties of solutions to partial differential equations characterized by exponential and polynomial terms. In one dimension, consider a constant $\beta \neq 0$ and a viscosity parameter $\nu > 0$. The solution of the equation

$$-\nu u'' + \beta\, u' = 0 \tag{11.1}$$

is given by $C_1 + C_2 e^{\frac{\beta x}{\nu}}$, where C_1 and C_2 are unique and determined from the boundary conditions. The exact solution is generated by the basis $\{1,\, e^{\frac{\beta x}{\nu}}\}$.
Moreover, the solution of the associated inhomogeneous equation,

$$-\nu u'' + \beta\, u' = 1, \tag{11.2}$$

is given by $\frac{x}{\beta} + C_1 + C_2 e^{\frac{\beta\, x}{\nu}}$, where C_1 and C_2 are unique and determined from the boundary conditions. The exact solution is generated by the basis $\{1,\, e^{\frac{\beta x}{\nu}},\, x\}$. These analytical forms of the solutions serve as the foundation for designing an adapted local basis for discretization spaces in 1d.
The main idea is to define, on each element, a local basis that includes the exponential terms, while taking into account both their limit behavior (for instance, when certain parameters become very large or very small) and their numerical effects, such as stability and accuracy of the computations as we shall see in the numerical results presented in chapter 14 below.

Remark 11.1 *In what follows, $\boldsymbol{\beta}$ is not necessarily piecewise constant.*

For order 1, the proposed basis on cell K_ℓ, with $\ell \in \mathcal{I}_K$, is defined as:

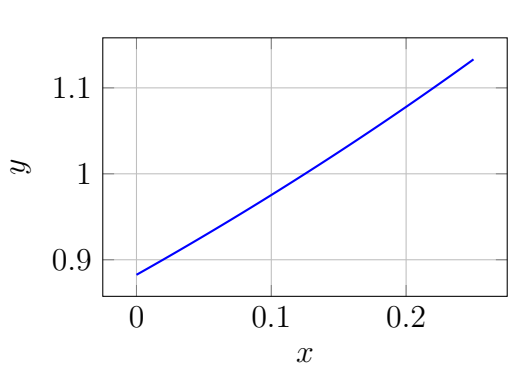
$$\left\{1, e^{\frac{b_\ell (x - x_\ell)}{\nu}}\right\},$$

where x_ℓ represents the barycenter of the cell, and $b_\ell = \beta_{|K_\ell}$ if β is piecewise constant. On the other hand, if β is not a constant parameter, one may choose for b_ℓ a local representative value like $b_\ell = \beta(x_\ell)$ or other choices that will be detailed in the section 12.4. We denote by b the vector of $\mathbb{R}^{N_K}$ that contains all the b_ℓ, for all $\ell \in \mathcal{I}_K$. In figures 11.1 and 11.2, I show the non-polynomial basis function, denoted by ϕ

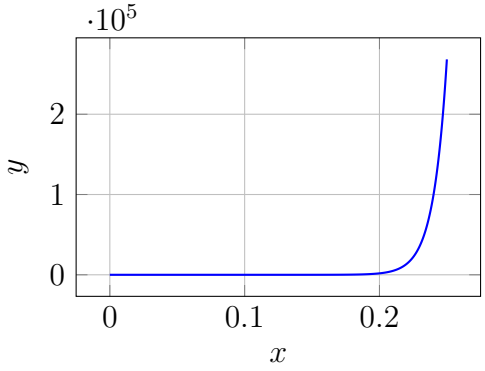
$$\phi := e^{\frac{b_\ell (x - x_\ell)}{\nu}},$$

on a cell $[0,\, 0.25]$, as a function of the ratio $\frac{b_\ell h}{\nu}$. Since we fix $h = 0.25$, we vary the parameters

ν and b_ℓ. For example, if we set $b_\ell = 1$, we get Figure 11.1:



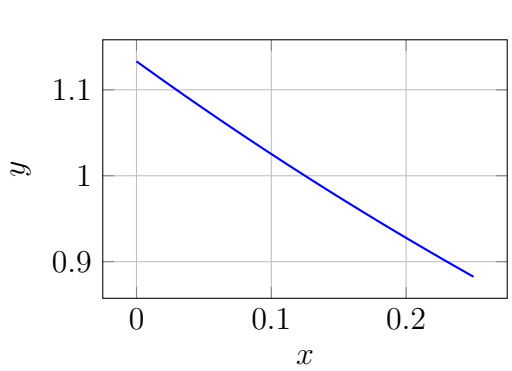
(a) ϕ with $b_\ell = 1$, $\nu = 1$



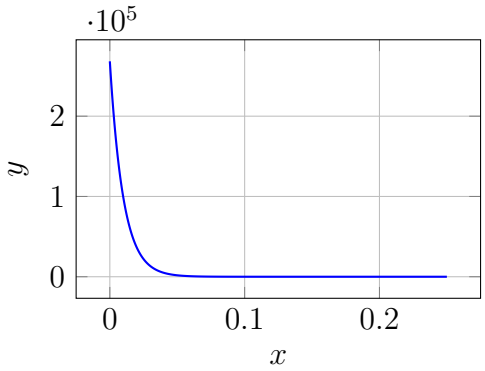
(b) ϕ with $b_\ell = 1$, $\nu = 10^{-2}$

Figure 11.1 – Comparison of the shape of the basis function ϕ with varying ν

If $b_\ell = -1$, for example, we simply change the sign of b_ℓ, and the shape of ϕ becomes:



(a) ϕ with $b_\ell = -1$, $\nu = 1$



(b) ϕ with $b_\ell = -1$, $\nu = 10^{-2}$

Figure 11.2 – Comparison of the shape of the basis function ϕ with changing ν

Remark 11.2 *When the ratio $\frac{b_\ell h}{\nu}$ is very small, ϕ converges to 1, which corresponds to the first component of the basis, so in that case the two basis functions would become almost colinear. In order to avoid this possible numerical issue, we can replace ϕ by $\phi - 1$, and additionally multiply by $\frac{\nu}{b_\ell h_\ell}$ so that the term*

$$(\phi - 1)\frac{\nu}{b_\ell h_\ell}$$

converges to $\frac{x - x_\ell}{h_\ell}$ as the ratio $\frac{b_\ell h}{\nu}$ becomes very small, as a Taylor expansion shows easily.

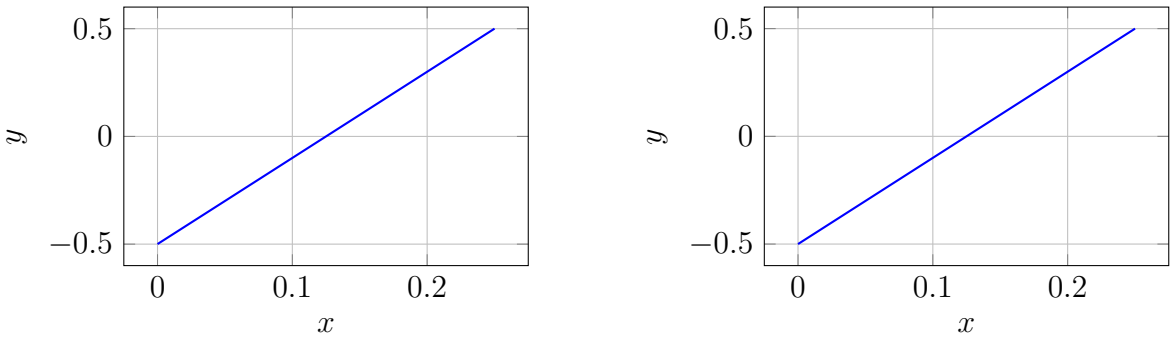
This modification helps stabilize the computations. Therefore the considered basis is:

$$\left\{1, \left(e^{\frac{b_\ell(x-x_\ell)}{\nu}} - 1\right) \frac{\nu}{b_\ell\, h_\ell}\right\}, \tag{11.3}$$

We denote:

$$\mathbb{NP}^1(K_\ell) := \operatorname{span}\left(\left\{1, \left(e^{\frac{b_\ell(x-x_\ell)}{\nu}} - 1\right) \frac{\nu}{b_\ell\, h_\ell}\right\}\right).$$

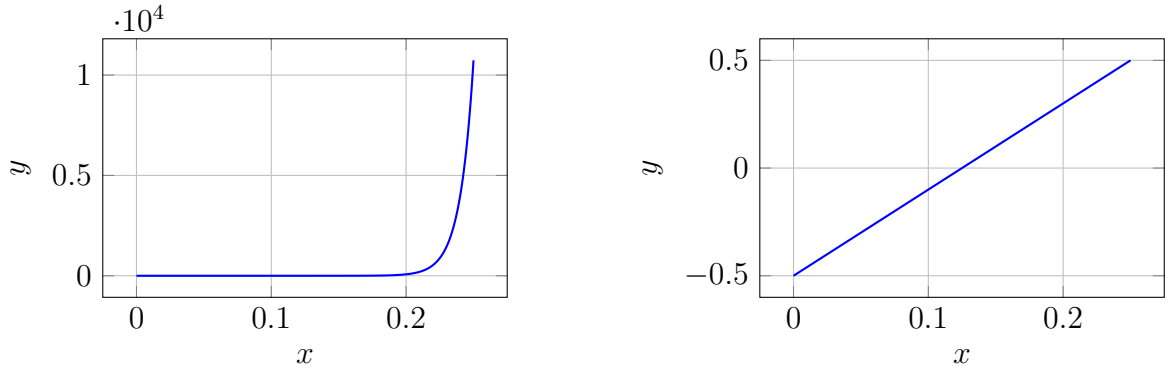
By Taylor expansion, this basis converges to the polynomial basis when the ratio $\frac{b_\ell\, h}{\nu}$ is small. In this figure, we plot the polynomial basis $\psi_P := \frac{x-x_\ell}{h}$ and the non-polynomial basis $\psi_{NP} := \left(e^{\frac{b_\ell(x-x_\ell)}{\nu}} - 1\right) \frac{\nu}{b_\ell\, h_\ell}$, when the ratio $\frac{b_\ell\, h}{\nu}$ is small, for example with $b_\ell = 0.01$, $\nu = 1$, and $h = 0.25$:



(a) ψ_{NP} (b) ψ_P

Figure 11.3 – Comparison of the shapes of the basis functions ψ_P and ψ_{NP} when $\frac{b_\ell\, h}{\nu} = 0.0025$

We can observe the difference when the ratio $\frac{b_\ell\, h}{\nu}$ is far from zero. For example, we take $\nu = 10^{-2}$, $b_\ell = 1$, and $h = 0.25$.



(a) ψ_{NP} (b) ψ_P

Figure 11.4 – Comparison of the shapes of the basis functions ψ_P and ψ_{NP} when $\frac{b_\ell h}{\nu} = 25$

For higher orders, this basis is enriched with additional terms to capture more complex variations; for example, the third term in this family enriches the spanned space by a first order polynomial, but when $\frac{bh}{\nu}$ tends to 0, the spanned space is very much like the second order polynomial space, like again a Taylor expansion shows:

$$\left\{1, \left(e^{\frac{b_\ell(x-x_\ell)}{\nu}} - 1\right) \frac{\nu}{b_\ell\, h_\ell}, \left(e^{\frac{b_\ell(x-x_\ell)}{\nu}} - 1 - \frac{b_\ell(x-x_\ell)}{\nu}\right) \left(\frac{\nu}{b_\ell\, h_\ell}\right)^2\right\}.$$

We denote:

$$\mathbb{NP}^2(K_\ell) := \operatorname{span}\left\{1, \left(e^{\frac{b_\ell(x-x_\ell)}{\nu}} - 1\right) \frac{\nu}{b_\ell\, h_\ell}, \left(e^{\frac{b_\ell(x-x_\ell)}{\nu}} - 1 - \frac{b_\ell(x-x_\ell)}{\nu}\right) \left(\frac{\nu}{b_\ell\, h_\ell}\right)^2\right\}.$$

The dependence of this space on the parameter b_ℓ is a major advantage for adapting the basis to the problem to be solved. This feature is particularly beneficial, provided that one knows how to choose b_ℓ appropriately, since in the Oseen or Navier–Stokes problem the vector field $\boldsymbol{\beta}$ is not constant, especially in dimensions 2 and 3. In what follows, we define the non-polynomial basis in two and three dimensions.

11.2 Non-polynomial basis in two dimensions

In dimension $d = 2$, we have $\mathbf{b} \in \mathbb{R}^{N_K \times 2}$ as a piecewise constant vector field, we denote by $\boldsymbol{\beta}_\ell$ the value of $\boldsymbol{\beta}$ in cell K_ℓ, we choose $\mathbf{b}_\ell = \boldsymbol{\beta}_\ell$, for all $\ell \in \mathcal{I}_K$. The point $\mathbf{x}_\ell$ represents the barycenter of the cell K_ℓ. We denote by $\mathbf{b}_\ell^\perp$ the vector orthogonal to $\mathbf{b}_\ell$, i.e., $\mathbf{b}_\ell^\perp \cdot \mathbf{b}_\ell = 0$,

having the same norm as $\mathbf{b}_\ell$.
The local basis of order 1 is constructed on each triangle K_ℓ based on the form:

$$\left\{1,\, e^{\frac{\mathbf{b}_\ell\cdot(\mathbf{x}-\mathbf{x}_\ell)}{\nu}},\, \frac{\mathbf{b}_\ell^\perp\cdot(\mathbf{x}-\mathbf{x}_\ell)}{h_\ell\,\|\mathbf{b}_\ell\|}\right\}.$$

We denote by x and y the first and second components of $\mathbf{x}$, respectively, and by x_ℓ and y_ℓ the first and second components of $\mathbf{x}_\ell$, respectively. We define the polynomial basis in $2d$ as

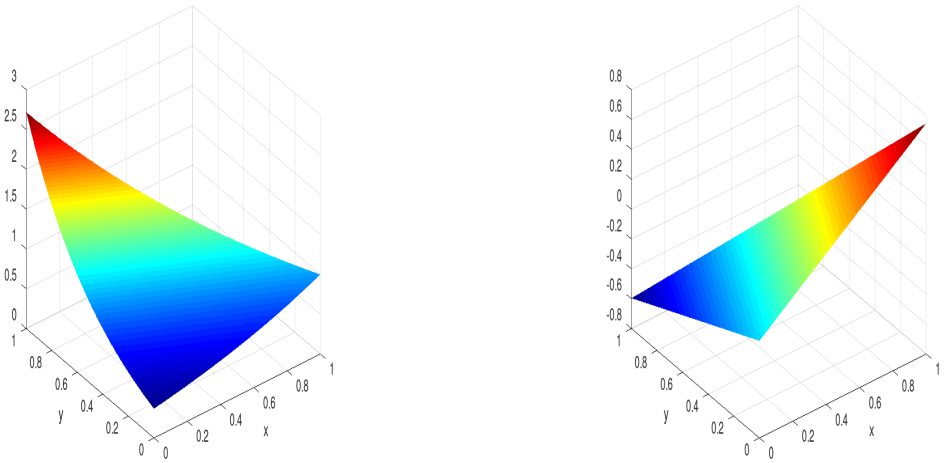
$$\left\{1,\, \frac{x-x_\ell}{h_\ell},\, \frac{y-y_\ell}{h_\ell}\right\}.$$

Remark 11.3 *The principle of Remark 11.2 in* $1d$ *also applies in* $2d$*. To avoid numerical issues when the ratio* $\frac{\|\mathbf{b}_\ell\|\,h_\ell}{\nu}$ *is very small, we replace* $e^{\frac{\mathbf{b}_\ell\cdot(\mathbf{x}-\mathbf{x}_\ell)}{\nu}}$ *by* $(e^{\frac{\mathbf{b}_\ell\cdot(\mathbf{x}-\mathbf{x}_\ell)}{\nu}}-1)\frac{\nu}{h_\ell\,\|\mathbf{b}_\ell\|}$ *which helps stabilize the computations. The resulting basis is then*

$$\left\{1,\, \left(e^{\frac{\mathbf{b}_\ell\cdot(\mathbf{x}-\mathbf{x}_\ell)}{\nu}}-1\right)\frac{\nu}{h_\ell\,\|\mathbf{b}_\ell\|},\, \frac{\mathbf{b}_\ell^\perp\cdot(\mathbf{x}-\mathbf{x}_\ell)}{h_\ell\,\|\mathbf{b}_\ell\|}\right\}.$$

Moreover, this basis converges to the polynomial basis when the ratio $\frac{\|\mathbf{b}_\ell\|\,h_\ell}{\nu}$ *is very small, since the second function in the basis tends to a linear one in the direction orthogonal to the third one, so that both directions are indeed taken into account.*

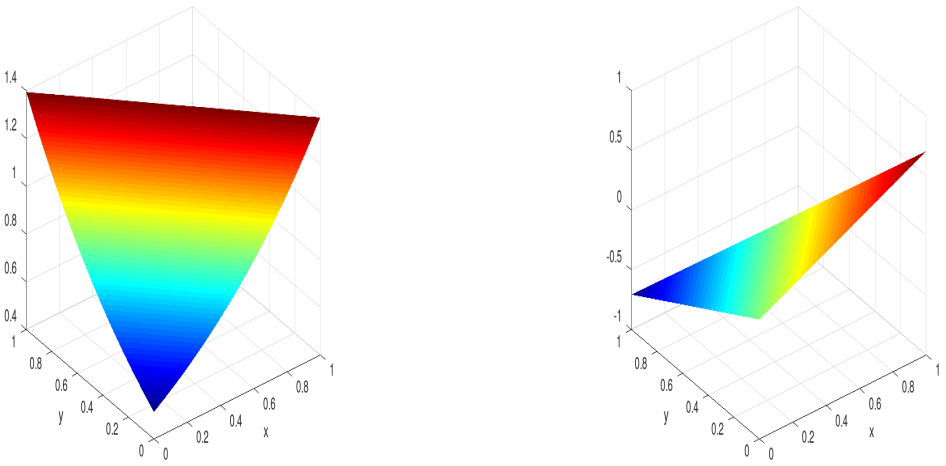
In figures 11.5, 11.6, 11.7 and 11.8, we show the form of this basis with different forms of ν, $\mathbf{b}_\ell$, and also when $\mathbf{b}_\ell$ is orthogonal to one of the edges of the cell K_ℓ, we denote in these figures $\phi_1 := e^{\frac{\mathbf{b}_\ell\cdot(\mathbf{x}-\mathbf{x}_\ell)}{\nu}}$, and $\phi_2 := \frac{\mathbf{b}_\ell^\perp\cdot(\mathbf{x}-\mathbf{x}_\ell)}{h_\ell\,\|\mathbf{b}_\ell\|}$:



(a) ϕ_1 (b) ϕ_2

Figure 11.5 – The shapes of the basis functions ϕ_1 and ϕ_2 when $\nu = 1$, $\mathbf{b}_\ell = [1, 2]$, and $h = \frac{1}{2}$.

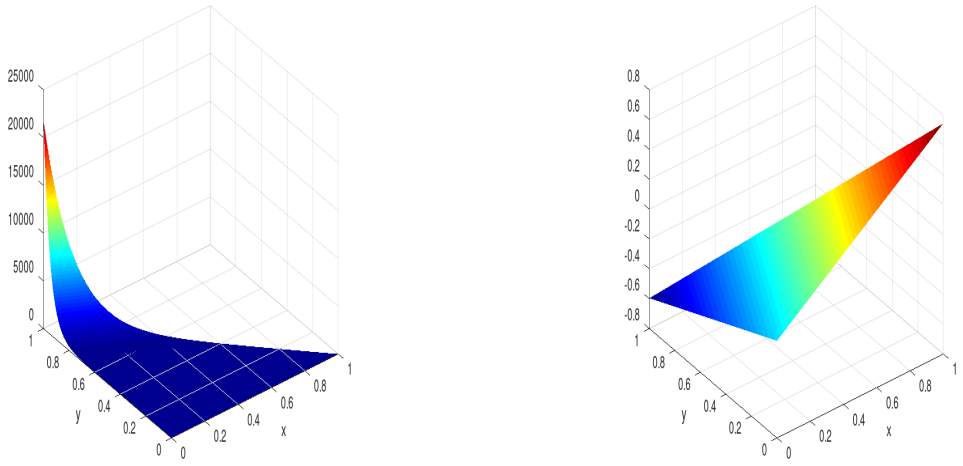
In this figure we take $\mathbf{b}_\ell$ perpendicularly to an edge of the cell K_ℓ, and $\nu = 1$.



(a) ϕ_1 (b) ϕ_2

Figure 11.6 – The shapes of the basis functions ϕ_1 and ϕ_2 when $\nu = 1$, $\mathbf{b}_\ell = [1, 1]$, and $h = \frac{1}{2}$.

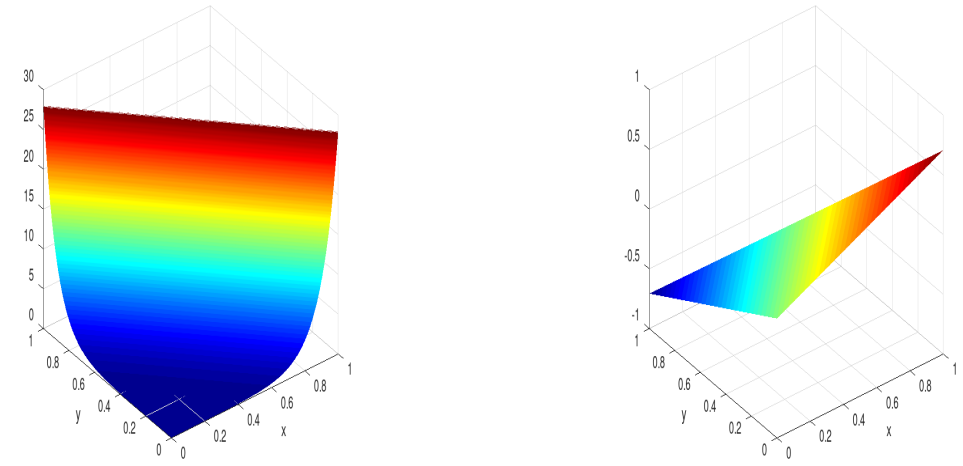
In this figure alose we take $\mathbf{b}_\ell$ perpendicularly to an edge of the cell K_ℓ, but $\nu = 10^{-1}$.



(a) ϕ_1 (b) ϕ_2

Figure 11.7 – The shapes of the basis functions ϕ_1 and ϕ_2 when $\nu = 10^{-1}$, $\mathbf{b}_\ell = [1, 2]$, and $h = \frac{1}{2}$.

To understand the effect of ν on the exponential basis, we redraw the same previous figures under the same conditions on h and $\mathbf{b}_\ell$, but with $\nu = 10^{-1}$.



(a) ϕ_1 (b) ϕ_2

Figure 11.8 – The shapes of the basis functions ϕ_1 and ϕ_2 when $\nu = 10^{-1}$, $\mathbf{b}_\ell = [1, 1]$, and $h = \frac{1}{2}$.

The ratio $\frac{\mathbf{b}_\ell \cdot (\mathbf{x}-\mathbf{x}_\ell)}{\nu}$ changes sign inside the element. If this ratio is positive and very large, there is a risk of obtaining terms that blow up in the implementation, as for ϕ_1 in Figure 11.7. We note that with the above parameters for h and $\mathbf{b}_\ell$, then if $\nu \leq 10^{-2}$, some terms tend numerically to diverge to very large values. To address this, we impose an additional condition: let K_ℓ be an element, we choose $\tilde{\mathbf{x}}_\ell$ such that $\mathbf{b}_\ell \cdot (\mathbf{x} - \tilde{\mathbf{x}}_\ell) \leq 0$ for any $\mathbf{x}$ in the triangle K_ℓ. This construction ensures that the exponential basis function remains bounded in asymptotic regimes and at the same time allows for a more accurate representation of local solutions to the considered partial differential equation. This is particularly relevant for problems where convective terms dominate diffusive ones ($||\mathbf{b}|| \gg \nu$). Consequently, this approach provides increased flexibility for numerically solving partial differential equations across various regimes.

Proposition 11.4 *For a given triangle K_ℓ and a given vector $\mathbf{b}_\ell$, there exists a point $\tilde{\mathbf{x}}_\ell \in \overline{K_\ell}$ such that*

$$\forall \mathbf{x} \in \overline{K_\ell}, \quad \mathbf{b}_\ell \cdot (\mathbf{x} - \tilde{\mathbf{x}}_\ell) \leq 0, \tag{11.4}$$

and $\tilde{\mathbf{x}}_\ell$ is one of the vertices of K_ℓ. If $\mathbf{b}_\ell$ is not orthogonal to any edge of K_ℓ, then $\tilde{\mathbf{x}}_\ell$ is unique.

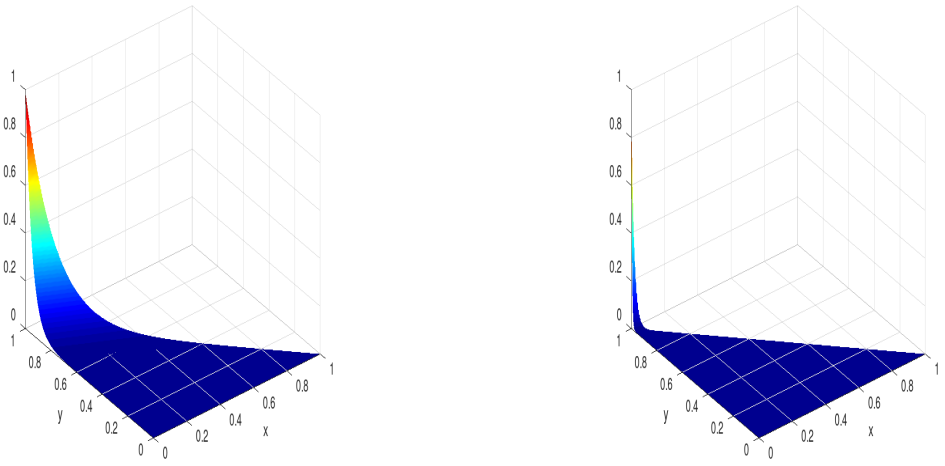
Proof. Let, S_1^ℓ, S_2^ℓ, S_3^ℓ be the vertices of triangle K_ℓ and $M_{\max} = \max\{\mathbf{b}_\ell \cdot S_1^\ell, \mathbf{b}_\ell \cdot S_2^\ell, \mathbf{b}_\ell \cdot S_3^\ell\}$. We know that for every point $\mathbf{x}$ in $\overline{K_\ell}$, there exist α_1, α_2, α_3 three positive real numbers such that $\sum_{i=1}^3 \alpha_i = 1$ and $\mathbf{x} = \sum_{i=1}^3 \alpha_i S_i$. Then we have $\mathbf{b}_\ell \cdot \mathbf{x} \leq (\sum_{i=1}^3 \alpha_i)\, M_{\max} = M_{\max}$. Now we can choose

$$\tilde{\mathbf{x}}_\ell = \arg \max_{S_1^\ell, S_2^\ell, S_3^\ell} \{\mathbf{b}_\ell \cdot S_1^\ell, \mathbf{b}_\ell \cdot S_2^\ell, \mathbf{b}_\ell \cdot S_3^\ell\} \tag{11.5}$$

and we get (11.4). For uniqueness, suppose that $\mathbf{b}_\ell$ is not orthogonal to any edge of K_ℓ, and assume that there exist two points satisfying (11.5). Let these points be S_1^ℓ and S_2^ℓ. Then we obtain $\mathbf{b}_\ell \cdot (S_1^\ell - S_2^\ell) = 0$, which means that $\mathbf{b}_\ell$ is orthogonal to the edge $S_1^\ell S_2^\ell$ of K_ℓ, a contradiction. Hence $\tilde{\mathbf{x}}_\ell$ is unique.

□

So we define $\tilde{\phi}_1 := e^{\frac{\mathbf{b}_\ell \cdot (\mathbf{x}-\tilde{\mathbf{x}}_\ell)}{\nu}}$. In Fig. 11.7, we can see the shape of ϕ_1 when $\mathbf{b}_\ell = [1, 1]$ and $h = \frac{1}{2}$ with two different values of ν:



(a) $\tilde{\phi}_1$ (b) $\tilde{\phi}_2$

Figure 11.9 – The shapes of the basis function $\tilde{\phi}_1$ for $\nu = 10^{-1}$ and $\nu = 10^{-2}$, respectively.

We note that ϕ_1 and $\tilde{\phi}_1$, shown in Figures 11.7 and 11.9, respectively, have the same shape. However, ϕ_1 is unbounded in the sense that if ν tends to 0, it numerically diverges to $+\infty$ (it already has large values when $\nu < 10^{-2}$ whereas $\tilde{\phi}_1$ remains bounded by 1 for any value of ν. So now we can define the space $\mathbb{NP}^1(K_\ell)$, for any $\ell \in \mathcal{I}_K$:

$$\mathbb{NP}^1(K_\ell) := \text{span}\left(\left\{1,\ \left(e^{\frac{\mathbf{b}_\ell \cdot (\mathbf{x}-\tilde{\mathbf{x}}_\ell)}{\nu}} - 1\right)\frac{\nu}{h_\ell \left\|\mathbf{b}_\ell\right\|},\ \frac{\mathbf{b}_\ell^\perp \cdot (\mathbf{x}-\tilde{\mathbf{x}}_\ell)}{h_\ell \left\|\mathbf{b}_\ell\right\|}\right\}\right).$$

For higher-order bases, this construction is enriched by adding additional terms of polynomial degree 2 or higher, or even non-polynomial terms. Thus, when we write $\mathbb{NP}^k(K_\ell)$, we mean a space of dimension $\frac{(k+d)!}{k!,d!}$, which contains $\mathbb{NP}^1(K_\ell)$. Now we can define the non polynomial discrete space:

$$\begin{aligned}
\mathbb{NP}^1_{disc}(\mathcal{T}_h) &:= \left\{v \in L^2(\Omega)\,|\,\forall \ell \in \mathcal{I}_K,\, v_{|K_\ell} \in \mathbb{NP}^1(K_\ell)\right\},\\
\mathbb{NP}^k_{disc}(\mathcal{T}_h) &:= \left\{v \in L^2(\Omega)\,|\,\forall \ell \in \mathcal{I}_K,\, v_{|K_\ell} \in \mathbb{NP}^k(K_\ell)\right\},\\
\mathbf{NP}^k_{disc}(\mathcal{T}_h) &:= [\mathbb{NP}^k(K_\ell)]^d, \quad d = 1,\, 2.
\end{aligned}$$

Remark 11.5 *Why do we chose the dimension of $\mathbb{P}^k(K_\ell)$ to be $\frac{(k+d)!}{k!\,d!}$? This choice ensures that, when comparing the numerical results of the polynomial and non-polynomial bases, both bases have the same number of degrees of freedom.*

11.3 Non-polynomial basis in three dimension

In dimension $d = 3$, we consider $\mathbf{b} \in \mathbb{R}^{N_K \times 3}$ as a piecewise constant vector field. We denote by $\mathbf{b}_\ell$ the value of $\mathbf{b}$ in cell K_ℓ, and set $\mathbf{b}_\ell = \boldsymbol{\beta}_\ell$ for all $\ell \in \mathcal{I}_K$. The point $\mathbf{x}_\ell$ represents the barycenter of the cell K_ℓ, and $\tilde{\mathbf{x}}_\ell$ is one of the vertices of the tetrahedron K_ℓ such that $\mathbf{b}_\ell \cdot \tilde{\mathbf{x}}_\ell \leq 0$ for all $\mathbf{x} \in K_\ell$. We denote by $\mathbf{b}_\ell^{1,\perp}$ and $\mathbf{b}_\ell^{2,\perp}$ a linearly independent pair of vectors orthogonal to $\mathbf{b}_\ell$, i.e., $\mathbf{b}_\ell^{1,\perp} \cdot \mathbf{b}_\ell = 0$ and $\mathbf{b}_\ell^{2,\perp} \cdot \mathbf{b}_\ell = 0$, and they are linearly independent.
The local basis of order 1 is constructed on each triangle K_ℓ based on the form:

$$\left\{1,\ \left(e^{\frac{\mathbf{b}_\ell \cdot (\mathbf{x} - \tilde{\mathbf{x}}_\ell)}{\nu}} - 1\right) \frac{\nu}{\|\mathbf{b}_\ell\|\, h_\ell},\ \frac{\mathbf{b}_\ell^{1,\perp} \cdot (\mathbf{x} - \mathbf{x}_\ell)}{h_\ell \,\|\mathbf{b}_\ell\|},\ \frac{\mathbf{b}_\ell^{2,\perp} \cdot (\mathbf{x} - \mathbf{x}_\ell)}{h_\ell \,\|\mathbf{b}_\ell\|}\right\}.$$

So now we can define the space $\mathbb{NP}^1(K_\ell)$, when $K_\ell \in \mathbb{R}^2$, will be:

$$\mathbb{NP}^1(K_\ell) := \operatorname{span}\left(\left\{1,\ \left(e^{\frac{\mathbf{b}_\ell \cdot (\mathbf{x} - \tilde{\mathbf{x}}_\ell)}{\nu}} - 1\right) \frac{\nu}{\|\mathbf{b}_\ell\|\, h_\ell},\ \frac{\mathbf{b}_\ell^{1,\perp} \cdot (\mathbf{x} - \mathbf{x}_\ell)}{h_\ell \,\|\mathbf{b}_\ell\|},\ \frac{\mathbf{b}_\ell^{2,\perp} \cdot (\mathbf{x} - \mathbf{x}_\ell)}{h_\ell \,\|\mathbf{b}_\ell\|}\right\}\right).$$

Remark 11.6 *A natural question arises: in 1D, the solution of the advection-diffusion problem (11.1) is given by*

$$C_1 + C_2 e^{\frac{\beta x}{\nu}},$$

but in 2D or 3D, the problem becomes

$$-\nu \Delta u + (\boldsymbol{\beta} \cdot \mathbf{grad}\,) u = 0. \tag{11.6}$$

Without boundary conditions, multiple solutions may exist. Let us denote $\boldsymbol{\beta}_{(i)} := \boldsymbol{\beta}_i \,\mathbf{e}_i$ for $i = 1, \ldots, d$, where $\boldsymbol{\beta}_i$ is the i-th component of $\boldsymbol{\beta}$, and $\boldsymbol{\beta}_{(i,j)} := \boldsymbol{\beta}_i \,\mathbf{e}_i + \boldsymbol{\beta}_j \,\mathbf{e}_j$ for $i, j = 1, \ldots, d$, $i < j$.

We define

$$v_0 := e^{\frac{\boldsymbol{\beta} \cdot \mathbf{x}}{\nu}}, \quad v_i := e^{\frac{\boldsymbol{\beta}_{(i)} \cdot \mathbf{x}}{\nu}},\ i = 1, \ldots, d, \quad v_{i,j} := e^{\frac{\boldsymbol{\beta}_{(i,j)} \cdot \mathbf{x}}{\nu}},\ i, j = 1, \ldots, d,\ i < j.$$

Any function $u \in \operatorname{span}\left(\{1, v_t, v_{i,j} \mid t = 0, \ldots, d,\ i, j = 1, \ldots, d,\ i < j\}\right)$ satisfies equation (11.6). This raises the question: in 2D there are 3 forms and in 3D there are 7 different forms for $\boldsymbol{\beta}_\ell$; why did we choose the form $\mathbf{b}_\ell = \boldsymbol{\beta}$ in our basis?

The answer is that it is possible to choose a form of $\mathbf{b}_\ell$ different from $\boldsymbol{\beta}$, but $\mathbf{b}_\ell$ must be nonzero. We have chosen $\mathbf{b}_\ell = \boldsymbol{\beta}$ in dimensions $d = 2, 3$ because this represents the natural

extension of the 1D basis to 2D and 3D. Other forms could be introduced if we choose $k > 1$.

11.4 On constants in inverse trace inequalities of finite elements with exponential basis

In the Polynomial case, the central result to prove that the discrete Poisson problem is well-posed was the coercivity-continuity property stated in Theorem 6.3. Looking at its proof, the only ingredient that depends on the fact that the basis is polynomial is inequality (4.12) (and its generalisation (4.15)) that is used in Lemma 6.2. In other words, one must show that (4.15) still holds in the non polynomial case with a constant that does not diverge as $h \to 0$. We first recall the following results

Theorem 11.1 ([63, Theorem 4.2.2] Rayleigh-Ritz) *Let $\mathbb{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Then*

$$\max_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \mathbf{x}\neq 0}} \frac{\mathbf{x}^T\mathbb{A}\mathbf{x}}{\|\mathbf{x}\|^2} = \lambda_{\max}(\mathbb{A}), \qquad \min_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \mathbf{x}\neq 0}} \frac{\mathbf{x}^T\mathbb{A}\mathbf{x}}{\|\mathbf{x}\|^2} = \lambda_{\min}(\mathbb{A}),$$

where $\lambda_{\max}(\mathbb{A})$ and $\lambda_{\min}(\mathbb{A})$ denote the largest and smallest eigenvalues of $\mathbb{A}$, respectively.

Corollary 11.7 *Let $\mathbb{A} \in \mathbb{R}^{n\times n}$ be symmetric. Then, for all $\mathbf{x} \in \mathbb{R}^n$,*

$$\lambda_{\min}(\mathbb{A})\,\|\mathbf{x}\|^2 \;\le\; \mathbf{x}^T\mathbb{A}\mathbf{x} \;\le\; \lambda_{\max}(\mathbb{A})\,\|\mathbf{x}\|^2.$$

Then, we are ready to generalise (4.15)

Theorem 11.2 *For all $\ell \in \mathcal{I}_K$, for all $v \in \mathbb{NP}^k(K_\ell)$, and for all $f \in \mathcal{I}_{F,\ell}$, we have*

$$\|v\|_{L^2(F_f)} \le \sigma^{\frac{1}{2}} C_{\mathbb{NP}}^{1/2}\, h_\ell^{-1/2}\, \|v\|_{L^2(K_\ell)}, \tag{11.7}$$

where $C_{\mathbb{NP}}$ is a positive constant depending on d, k, and the shape of K_ℓ. Moreover, as $h_\ell \to 0$, $C_{\mathbb{NP}}$ converges to a strictly positive constant and does not diverge.

Proof. Let $\hat{K}$ be the reference element (a segment in 1D, a triangle in 2D, and a tetrahedron in 3D), and let $\hat{F}$ be a face of $\hat{K}$. Let

$$N = \dim \mathbb{NP}^k(\hat{K}) = \binom{k+d}{k},$$

11.4. On constants in inverse trace inequalities of finite elements with exponential basis

and let $(\psi_n)_{n=1,\dots,N}$ be an $L^2(\hat{K})$-orthonormal basis of $\mathbb{NP}^k(\hat{K})$, i.e.,

$$(\psi_i, \psi_j)_{\hat{K}} = \int_{\hat{K}} \psi_i(x)\, \psi_j(x)\, \mathrm{d}x = \delta_{ij}, \quad \forall\, i, j \in \{1, \dots, N\}.$$

For any $v \in \mathbb{NP}^k(\hat{K})$, there exist coefficients $(\hat{v}_n)_{n=1,\dots,N} \in \mathbb{R}$ such that

$$v(x) = \sum_{n=1}^{N} \hat{v}_n\, \psi_n(x).$$

We denote the vector of coefficients by

$$\vec{v} = (\hat{v}_n)_{n=1,\dots,N}.$$

Since the basis (ψ_n) is $L^2(\hat{K})$-orthonormal, it follows that

$$\|v\|^2_{L^2(\hat{K})} = \|\vec{v}\|^2_{\ell^2}.$$

Moreover, the coefficients $\hat{v}_n$ can be expressed as

$$\hat{v}_n = \int_{\hat{K}} v(x)\, \psi_n(x)\, \mathrm{d}x.$$

Now, let $\mathbb{F} \in \mathbb{R}^{N\times N}$ be the matrix defined by

$$\mathbb{F}_{i,j} = \int_{\hat{F}} \psi_i(x)\, \psi_j(x)\, \mathrm{d}x, \quad 1 \le i, j \le N,$$

where $\hat{F}$ denotes a face of $\hat{K}$. We denote by $\rho(\mathbb{F})$ the spectral radius of $\mathbb{F}$. Then, for any $v \in \mathbb{NP}^k(\hat{K})$, one has

$$\|v\|^2_{L^2(\hat{F})} = \vec{v}^T \mathbb{F}\, \vec{v}.$$

Since $\mathbb{F}$ is symmetric using corollary 11.7, we obtain the following inequality:

$$\|v\|^2_{L^2(\hat{F})} \le \rho(\mathbb{F})\, \|v\|^2_{L^2(\hat{K})}.$$

For all $\ell \in \mathcal{I}_K$, K_ℓ be an element, h_ℓ the diameter of K_ℓ, and F_f a face of K i.e $f \in \mathcal{I}_{F,\ell}$. Using inequalities (4.5), (4.6) and (4.9), we obtain

$$\|v\|^2_{L^2(F_f)} \le h_\ell^{-1}\, (2\, d\, \sigma)\, \frac{|\hat{K}|}{|\hat{F}_{f,\ell}|} \rho(\mathbb{F})\, \|v\|^2_{L^2(K_\ell)}.$$

We denote by $C_{\mathbb{NP}} := 2\,d\,\frac{|\hat{K}|}{\min_f |\hat{F}_{f,\ell}|}\rho(\mathbb{F})$. The next step is to realize that, contrarily to polynomial basis, $\rho(\mathbb{F})$ depends on h_ℓ and we must thus look at what happens when $h \to 0$ and show that $\rho(\mathbb{F})$ remains bounded as $h \to 0$. We demonstrate this theoretically in 1D for $k = 1$; the same approach extends to $d \geq 1$ and $k \geq 1$.
Following [99], the proof is carried out in dimension $d = 1$, where the $L^2(F_f)$ norm is replaced by the Euclidean norm for the points representing the interface. We first map the basis to a reference element. Consider an arbitrary element $K_\ell = [a_1^\ell, a_2^\ell]$ and map it to the reference interval $[0, 1]$ via

$$x = a_1^\ell + t(a_2^\ell - a_1^\ell).$$

We denote by $h_\ell = |a_2^\ell - a_1^\ell|$, and let b_ℓ be a nonzero scalar. The basis functions transform as:

$$\begin{cases} 1 \longrightarrow v_1 := 1, \\ e^{b_\ell \frac{x - x_\ell}{\nu}} \longrightarrow v_2(t) = e^{\frac{b_\ell\, h_\ell}{\nu}(t-1/2)}. \end{cases}$$

Next, we construct an orthogonal basis on $[0, 1]$ using the Gram-Schmidt process:

$$\begin{cases} v_1 \longrightarrow e_1 := 1, \\ v_2 \longrightarrow e_2 := \frac{e^{\frac{b_\ell\, h_\ell}{\nu}(t-1/2)} - \alpha}{\beta}, \end{cases}$$

with

$$\begin{aligned} \alpha &:= \int_{[}0,1](v_2, 1)_{L^2([0,1])} &&= \frac{\nu}{b_\ell\, h_\ell}\left(e^{\frac{b_\ell\, h_\ell}{2\nu}} - e^{-\frac{b_\ell\, h_\ell}{2\nu}}\right), \\ \beta^2 &:= \|v_2 - \alpha\|^2_{L^2([0,1])} &&= -\alpha^2 + \frac{\nu}{2b_\ell\, h_\ell}\left(e^{\frac{b_\ell\, h_\ell}{\nu}} - e^{-\frac{b_\ell\, h_\ell}{\nu}}\right). \end{aligned}$$

For any $u \in \mathbb{NP}^k(\hat{K}) = \hat{u}_1 e_1 + \hat{u}_2 e_2$, we then have

$$|u(x_i)| \leq \left[(\hat{u}_1)^2 + (\hat{u}_2)^2\right]^{\frac{1}{2}} \left[(e_1(x_i))^2 + (e_2(x_i))^2\right]^{\frac{1}{2}}$$

Hence,

$$u(x_i)^2 \leq \rho(\mathbb{F})\, \|u\|^2_{L^2(\Omega)}, \quad \text{with} \quad \rho(\mathbb{F}) = 1 + e_2(x_i)^2.$$

A Taylor expansion shows that $e_2(x_i)^2 \to 3$ as $h \to 0$, it follows that $\rho(\mathbb{F})$ converges to 4.
For $k, d > 1$, the convergence of $\rho(\mathbb{F})$ can be verified numerically using the same procedure.
□

We will present the numerical convergence of $\rho(\mathbb{F})$ as h_ℓ tends to zero. In one dimension:

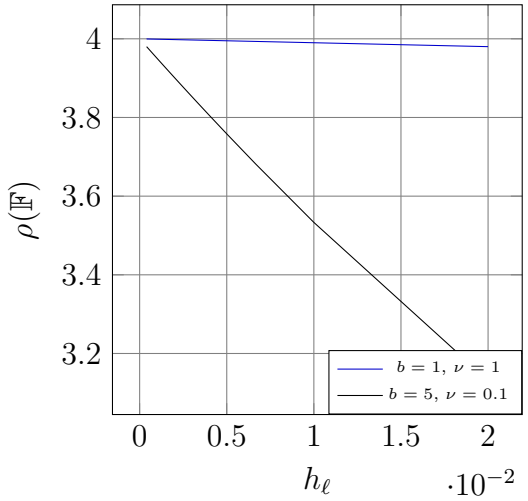


Figure 11.10 – $d = 1$, order 1.

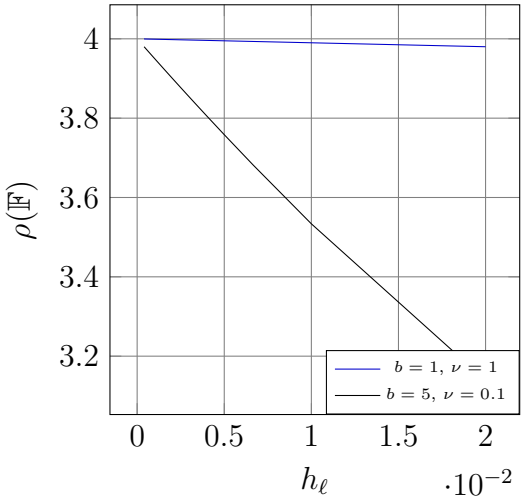


Figure 11.11 – $d = 1$, order 2.

For the two-dimensional case, we fix $\nu = 10^{-1}$ and work on a triangle K_ℓ defined by its vertices $S_1^\ell = [0, 0]$, $S_2^\ell = \epsilon[1, 0]$, and $S_3^\ell = \epsilon[0, 1]$, and $F_\ell = [S_1^\ell, S_2^\ell]$ with ϵ varying from 1 to $\frac{1}{200}$, which is equivalent to h_ℓ varying from $\sqrt{2}$ to $\frac{\sqrt{2}}{200}$.We present the convergence of $\rho(\mathbb{F})$ for several values of $\mathbf{b}_\ell$, for instance when $\mathbf{b}_\ell$ is collinear or perpendicular to F_f.

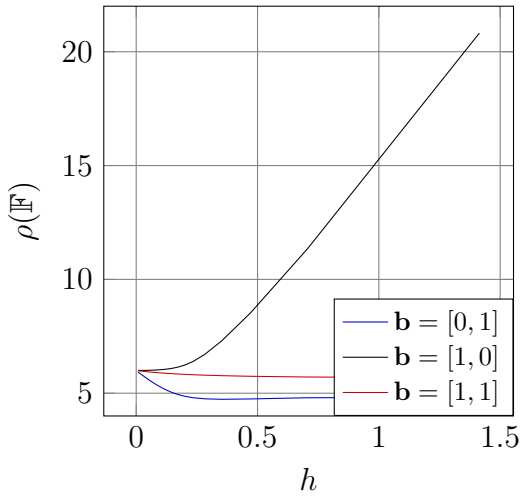


Figure 11.12 – $d = 2$, variation of $\rho(\mathbb{F})$ as a function of h.

These figures validate that the constant C_{NP} in the inequality (11.7) converges to a positive constant as h tends to 0.

11.5 Exact evaluation of the integrals involved in the matrices

Numerically, we implemented the exact integrals for the mass and stiffness matrices, as well as for the advection term, when using the non-polynomial basis, and similarly for the right-hand side. To achieve this, I wrote functions that compute the integrals as accurately as possible. In this section, I present a small part of the computation to give an idea of how these integrals were evaluated. Let $\hat{K}$ be the reference triangle with vertices $(0,0)$, $(1,0)$, and $(0,1)$, and let K_ℓ be a triangle in the mesh with vertices S_1^ℓ, S_2^ℓ, and S_3^ℓ, where $S_i^\ell = (x_i^\ell, y_i^\ell)$ for $i = 1,2,3$. We define an affine transformation $T_\ell : \hat{K} \to K_\ell$. According to Paragraph (4.3) it can be written in the following form:

$$T_\ell(\hat{x},\hat{y}) = \begin{bmatrix} x_1^\ell \\ y_1^\ell \end{bmatrix} + \begin{bmatrix} x_2^\ell - x_1^\ell & x_3^\ell - x_1^\ell \\ y_2^\ell - y_1^\ell & y_3^\ell - y_1^\ell \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix}. \tag{11.8}$$

For simplicity, we denote by $\mathbb{B}_\ell := \begin{bmatrix} x_2^\ell - x_1^\ell & x_3^\ell - x_1^\ell \\ y_2^\ell - y_1^\ell & y_3^\ell - y_1^\ell \end{bmatrix}$, $\hat{\mathbf{x}} := (\hat{x},\hat{y})$. Since T_ℓ is affine, the Jacobian J_ℓ is constant over $\hat{K}$, and $T_\ell(\hat{K}) = K_\ell$. Recall that $|J_\ell| = \det(\mathbb{B}_\ell)$. It holds:

$$|J_\ell| = 2\,|K_\ell|.$$

11.5.1 Exact integration of an exponential over a triangle K_ℓ

Let $\mathbf{b}_\ell \in \mathrm{R}^2\backslash\{0\}$, $\mathbf{x}_1^\ell \in \mathrm{R}^2$. In this section, we aim to compute the exact integration of the term

$$\int_{K_\ell} \exp(\mathbf{b}_\ell \cdot (\mathbf{x} - \mathbf{x}_1^\ell))\, d\mathbf{x}.$$

According to the geometric mapping T_ℓ (11.8), for any $\mathbf{x} \in K_\ell$, there exists $\hat{\mathbf{x}}$ such that

$$\mathbf{x} = \mathbb{B}_\ell \hat{\mathbf{x}} + S_1^\ell = S_1^\ell + (S_2^\ell - S_1^\ell)\hat{x} + (S_3^\ell - S_1^\ell)\hat{y}.$$

Thus, we can write

$$\begin{aligned} \mathbf{x} - \mathbf{x}_1^\ell &= \mathbb{B}_\ell(\hat{\mathbf{x}} - \hat{\mathbf{x}}_1^\ell) \\ \mathbf{b}_\ell \cdot (\mathbb{B}_\ell \hat{\mathbf{x}}) &= \alpha\hat{x} + \beta\hat{y} \\ \mathbf{b}_\ell \cdot (\mathbb{B}_\ell \hat{\mathbf{x}}_1^\ell) &= \mathbf{b}_\ell \cdot (\mathbf{x}_1^\ell - S_1^\ell) \end{aligned}$$

where

$$\alpha = \mathbf{b}_\ell \cdot (S_2^\ell - S_1^\ell), \quad \beta = \mathbf{b}_\ell \cdot (S_3^\ell - S_1^\ell).$$

11.5. Exact evaluation of the integrals involved in the matrices

By transforming to the reference triangle, we obtain

$$\int_{K_\ell} \exp(\mathbf{b}_\ell \cdot (\mathbf{x} - \mathbf{x}_1^\ell))\, d\mathbf{x} = 2|K_\ell| \exp(\mathbf{b}_\ell \cdot (S_1^\ell - \mathbf{x}_1^\ell)) \int_{\hat{K}} \exp(\alpha\hat{x} + \beta\hat{y})\, d\hat{\mathbf{x}} \tag{11.9}$$

Let us begin with the calculation:

$$\begin{aligned}
\int_{\hat{K}} \exp(\alpha\hat{x} + \beta\hat{y})\, d\hat{\mathbf{x}} &= \int_0^1 \exp(\alpha\hat{x}) \int_0^{-\hat{x}+1} \exp(\beta\hat{y})\, d\hat{y}\, d\hat{x} \\
(\text{if } \beta \neq 0) &= \tfrac{1}{\beta} \int_0^1 \exp(\alpha\hat{x}) \left(\exp(\beta(-\hat{x}+1)) - 1\right) d\hat{x} \\
&= \tfrac{1}{\beta} \left(\exp(\beta) \int_0^1 \exp((\alpha-\beta)\hat{x}) - \int_1^0 \exp(\alpha\hat{x}) \right) \\
(\text{if } \alpha \neq 0 \text{ and } \alpha - \beta \neq 0) &= \tfrac{1}{\beta} \left(\tfrac{\exp(\beta)}{\alpha-\beta} \left(\exp(\alpha-\beta) - 1\right) - \tfrac{1}{\alpha}\left(\exp(\alpha) - 1\right) \right) \\
&= \tfrac{1}{\beta} \left(\tfrac{1}{\alpha-\beta} \left(\exp(\alpha) - \exp(\beta)\right) - \tfrac{1}{\alpha}\left(\exp(\alpha) - 1\right) \right)
\end{aligned}$$

By substituting the expressions of α and β, we obtain:

$$\begin{aligned}
\int_{\hat{K}} \exp(\alpha\hat{x} + \beta\hat{y})\, d\hat{\mathbf{x}} &= \tfrac{1}{\mathbf{b}_\ell \cdot (S_3^\ell - S_1^\ell)} \Big(\tfrac{1}{\mathbf{b}_\ell \cdot (S_2^\ell - S_3^\ell)} \left(\exp(\mathbf{b}_\ell \cdot (S_2^\ell - S_1^\ell)) - \exp(\mathbf{b}_\ell \cdot (S_3^\ell - S_1^\ell)) \right) \\
&\quad - \tfrac{1}{\mathbf{b}_\ell \cdot (S_2^\ell - S_1^\ell)} \left(\exp(\mathbf{b} \cdot (S_2^\ell - S_1^\ell)) - 1 \right) \Big)
\end{aligned}$$

Finally, if $\beta \neq 0,\ \alpha \neq 0$ and $\alpha - \beta \neq 0$ we obtain:

$$\begin{aligned}
\int_{K_\ell} \exp(\mathbf{b}_\ell \cdot (\mathbf{x} - \mathbf{x}_1^\ell))\, d\mathbf{x} &= \tfrac{2|K_\ell|}{\mathbf{b} \cdot (S_3^\ell - S_1^\ell)} \Big[\tfrac{1}{\mathbf{b}_\ell \cdot (S_2^\ell - S_3^\ell)} \left(\exp(\mathbf{b}_\ell \cdot (S_2^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_\ell \cdot (S_3^\ell - \mathbf{x}_1^\ell)) \right) \\
&\quad - \tfrac{1}{\mathbf{b}_\ell \cdot (S_2^\ell - S_1^\ell)} \left(\exp(\mathbf{b}_\ell \cdot (S_2^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_\ell \cdot (S_1^\ell - \mathbf{x}_1^\ell)) \right) \Big].
\end{aligned}$$

Now, let us treat the case $\beta = 0$; then $\alpha \neq 0$ since $\mathbf{b}_\ell$ is nonzero and we get

$$\begin{aligned}
\int_{\hat{K}} \exp(\alpha\hat{x})\, \hat{\mathbf{x}}\, h\hat{\mathbf{x}} &= \int_0^1 \exp(\alpha\hat{x}) \int_0^{-\hat{x}+1} d\hat{y}\, d\hat{x} \\
&= \int_0^1 (-\hat{x}+1) \exp(\alpha\hat{x}) \\
&= \tfrac{1}{\alpha} \left(- \left(-\tfrac{1}{\alpha} (\exp(\alpha) - 1) \right) - 1 \right)
\end{aligned}$$

By substituting the expressions of α, we obtain:

$$\int_{\hat{K}} \exp(\alpha\hat{x})\, \hat{\mathbf{x}} = \tfrac{1}{\mathbf{b}_\ell \cdot (S_2^\ell - S_1^\ell)} \left(- \left(-\tfrac{1}{\mathbf{b} \cdot (S_2^\ell - S_1^\ell)} (\exp(\mathbf{b}_\ell \cdot (S_2^\ell - S_1^\ell)) - 1) \right) - 1 \right)$$

Finally, if $\beta = 0$ we obtain:

$$\begin{aligned}\int_{K_\ell} \exp(\mathbf{b}_\ell \cdot (\mathbf{x} - \mathbf{x}_1^\ell))\, d\mathbf{x} &= \frac{2|K_\ell|}{\mathbf{b}_\ell \cdot (S_2^\ell - S_1^\ell)} \Big(\frac{1}{\mathbf{b}_\ell \cdot (S_2^\ell - S_1^\ell)} (\exp(\mathbf{b}_\ell \cdot (S_2^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_\ell \cdot (S_1^\ell - \mathbf{x}_1^\ell))) \\ &\quad - \exp(\mathbf{b}_\ell \cdot (S_1^\ell - \mathbf{x}_1^\ell)) \Big).\end{aligned}$$

Now, let us treat the case $\alpha = 0$; then $\beta \neq 0$ since $\mathbf{b}_\ell$ is nonzero and we get

$$\begin{aligned}\int_{\hat{K}} \exp(\beta \hat{y}) &= \int_0^1 \int_0^{-\hat{x}+1} \exp(\beta \hat{y})\, d\hat{y}\, d\hat{x} \\ &= \tfrac{1}{\beta} \int_0^1 (\exp(\beta(-\hat{x}+1)) - 1) \\ &= \tfrac{1}{\beta} \left(\exp(\beta)(\exp(-\beta) - 1) - 1\right) \\ &= \tfrac{1}{\beta} \left(-\tfrac{1}{\beta}(1 - \exp(\beta)) - 1\right)\end{aligned}$$

By substituting the expressions of β, we obtain:

$$\int_{\hat{K}} \exp(\beta \hat{y})\, d\hat{\mathbf{x}} = \frac{1}{\mathbf{b}_\ell \cdot (S_3^\ell - S_1^\ell)} \left(-\frac{1}{\mathbf{b}_\ell \cdot (S_3^\ell - S_1^\ell)} \left(1 - \exp(\mathbf{b}_\ell \cdot (S_3^\ell - S_1^\ell))\right) - 1 \right)$$

Finally, if $\alpha = 0$ we obtain:

$$\begin{aligned}\int_{K_\ell} \exp(\mathbf{b}_\ell \cdot (\mathbf{x} - \mathbf{x}_1^\ell))\, d\mathbf{x} &= 2|K_\ell| \tfrac{1}{\mathbf{b}_\ell \cdot (S_3^\ell - S_1^\ell)} \Big(-\tfrac{1}{\mathbf{b}_\ell \cdot (S_3^\ell - S_1^\ell)} \left(\exp(\mathbf{b}_\ell \cdot (S_1^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_\ell \cdot (S_3^\ell - \mathbf{x}_1^\ell))\right) \\ &\quad - \exp(\mathbf{b}_\ell \cdot (S_1^\ell - \mathbf{x}_1^\ell)) \Big)\end{aligned}$$

The last special case is when $\alpha \neq 0$ and $\beta \neq 0$, but $\alpha - \beta = 0$, and we get

$$\begin{aligned}\int_{\hat{K}} \exp(\alpha \hat{x} + \beta \hat{y}) d\hat{\mathbf{x}} &= \int_0^1 \exp(\alpha \hat{x}) \int_0^{-\hat{x}+1} \exp(\beta \hat{y})\, d\hat{y} d\hat{x} \\ &= \tfrac{1}{\beta} \int_0^1 \exp(\alpha \hat{x}) \left(\exp(\beta(-\hat{x}+1)) - 1\right) d\hat{x} \\ &= \tfrac{1}{\beta} \left(\exp(\beta) - \int_0^1 \exp(\alpha \hat{x}) \right) \\ &= \tfrac{1}{\beta} \left(\exp(\beta) - \tfrac{1}{\alpha} \left(\exp(\alpha) - 1\right) \right)\end{aligned}$$

By substituting the expressions of α and β, we obtain:

$$\int_{\hat{K}} \exp(\alpha \hat{x} + \beta \hat{y})\, d\hat{\mathbf{x}} = \frac{1}{\mathbf{b}_\ell \cdot (S_3^\ell - S_1^\ell)} \left(\exp(\mathbf{b}_\ell \cdot (S_3^\ell - S_1^\ell)) - \frac{1}{\mathbf{b}_\ell \cdot (S_2^\ell - S_1^\ell)} \left(\exp(\mathbf{b}_\ell \cdot (S_2^\ell - S_1^\ell)) - 1\right) \right)$$

Finally, if $\alpha - \beta = 0$ we obtain:

$$\begin{aligned}\int_{K_\ell} \exp(\mathbf{b}_\ell \cdot (\mathbf{x} - \mathbf{x}_1^\ell))\, d\mathbf{x} &= \frac{2|K_\ell|}{\mathbf{b}_\ell \cdot (S_3^\ell - S_1^\ell)} \Big[- \frac{1}{\mathbf{b}_\ell \cdot (S_2^\ell - S_1^\ell)} \Big(\exp(\mathbf{b}_\ell \cdot (S_2^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_\ell \cdot (S_1^\ell - \mathbf{x}_1^\ell)) \Big) \\ &\quad + \exp(\mathbf{b}_\ell \cdot (S_3^\ell - \mathbf{x}_1^\ell)) \Big]\end{aligned}$$

Remark 11.8 *Numerically, we have* $\mathbf{b}_\ell = \frac{\tilde{\mathbf{b}}_\ell}{\nu}$. *If the ratio* $\frac{\tilde{\mathbf{b}}_\ell\, h_\ell}{\nu}$ *is very small, where* h_ℓ *denotes the diameter of* K_ℓ, *it is preferable to approximate this integral using methods such as Gauss quadrature, or by performing a Taylor expansion up to a high order, because the exponential term rapidly approaches machine zero, and we then multiply by the inverse of this quantity, i.e.,* $\frac{\nu}{\tilde{\mathbf{b}}_\ell\, h_\ell}$, *which is very large. This combination leads to numerical difficulties.*

11.5.2 Exact integration of an exponential times a linear function over a triangle K_ℓ

Let $\mathbf{b}_1^\ell$, $\mathbf{b}_2^\ell \in \mathrm{R}^2 \backslash \{0\}$, and $\mathbf{x}_1^\ell$, $\mathbf{x}_2^\ell \in \mathrm{R}^2$. For $i = 1, 2$, we can write:

$$\begin{aligned}\mathbf{b}_i^\ell \cdot (\mathbb{A}\hat{\mathbf{x}}) &= \alpha_i \hat{x} + \beta_i \hat{y}, \\ \mathbf{b}_i^\ell \cdot (\mathbb{A}\hat{\mathbf{x}_i^\ell}) &= \mathbf{b}_i^\ell \cdot (\mathbf{x}_i^\ell - S_1^\ell).\end{aligned}$$

where

$$\alpha_i = \mathbf{b}_i^\ell \cdot (S_2^\ell - S_1^\ell),\ \beta_i = \mathbf{b}_i^\ell \cdot (S_3^\ell - S_1^\ell).$$

In this section, we aim to compute the exact integration of the term

$$\int_{K_\ell} \exp(\mathbf{b}_1^\ell \cdot (\mathbf{x} - \mathbf{x}_1^\ell))\, \mathbf{b}_2^\ell \cdot (\mathbf{x} - \mathbf{x}_2^\ell)\, d\mathbf{x}.$$

By transforming to the reference triangle, we obtain

$$\begin{aligned}&\int_{K_\ell} \exp(\mathbf{b}_1^\ell \cdot (\mathbf{x} - \mathbf{x}_1^\ell))\, \mathbf{b}_2^\ell \cdot (\mathbf{x} - \mathbf{x}_2^\ell)\, d\mathbf{x} = \\ &2|K_\ell| \exp(\mathbf{b}_1^\ell \cdot (S_1^\ell - \mathbf{x}_1^\ell)) \int_{\hat{K}} \exp(\alpha_1\, \hat{x} + \beta_1\, \hat{y})(\alpha_2\, \hat{x} + \beta_2\, \hat{y})\, d\hat{\mathbf{x}} \\ &+ \mathbf{b}_2^\ell \cdot (S_1^\ell - \mathbf{x}_2^\ell) \int_{K_\ell} \exp(\mathbf{b}_1^\ell \cdot (\mathbf{x} - \mathbf{x}_1^\ell))\, d\mathbf{x}.\end{aligned} \tag{11.10}$$

The second term of equality (11.10) has already been treated in Section 11.5.1. For the first term:

$$\int_{\hat{K}} \exp(\alpha_1\,\hat{x}+\beta_1\,\hat{y})(\alpha_2\hat{x}+\beta_2\hat{y})\,d\hat{\mathbf{x}} = \int_0^1 \exp(\alpha_1\,\hat{x})\Big[\alpha_2\,\hat{x}\int_0^{-\hat{x}+1}\exp(\beta_1\,\hat{y})\,d\hat{y}+\beta_2\int_0^{-\hat{x}+1}\hat{y}\exp(\beta_1\,\hat{y})\,d\hat{y}\Big]\,d\hat{x}.$$

We denote:

$$\begin{aligned}\mathcal{I}_1 &:= \alpha_2\int_0^1 \hat{x}\exp(\alpha_1\,\hat{x})\int_0^{-\hat{x}+1}\exp(\beta_1\,\hat{y})\,d\hat{y}\,d\hat{x},\\ \mathcal{I}_2 &:= \beta_2\int_0^1 \exp(\alpha_1\,\hat{x})\int_0^{-\hat{x}+1}\hat{y}\exp(\beta_1\,\hat{y})\,d\hat{y}\,d\hat{x}\end{aligned}$$

We start with the computation:
if $\beta_1 \neq 0$, $\alpha_1-\beta_1\neq 0$, and $\alpha_1\neq 0$:

$$\begin{aligned}\mathcal{I}_1 &= \tfrac{\alpha_2}{\beta_1}\int_0^1 \hat{x}\exp(\alpha_1\,\hat{x})\Big[\exp(\beta_1\,(-\hat{x}+1))-1\Big]\,d\hat{x}\\ &= \tfrac{\alpha_2}{\beta_1}\Big[\Big(\tfrac{1}{\alpha_1-\beta_1}\Big(\exp(\alpha_1)-\tfrac{1}{\alpha_1-\beta_1}\left(\exp(\alpha_1)-\exp(\beta_1)\right)\Big)\Big)-\tfrac{1}{\alpha_1}\Big(\exp(\alpha_1)-\tfrac{1}{\alpha_1}\left(\exp(\alpha_1)-1\right)\Big)\Big]\end{aligned}$$

For $\mathcal{I}_2$:

$$\begin{aligned}\mathcal{I}_2 &= \beta_2\int_0^1\exp(\alpha_1\,\hat{x})\int_0^{-\hat{x}+1}\hat{y}\exp(\beta_1\,\hat{y})\,d\hat{y}\,d\hat{x}\\ &= \tfrac{\beta_2}{\beta_1}\Big[\tfrac{1}{\alpha_1-\beta_1}\Big(-\Big(\exp(\alpha_1)-\tfrac{1}{\alpha_1-\beta_1}\left(\exp(\alpha_1)-\exp(\beta_1)\right)\Big)+\left(\exp(\alpha_1)-\exp(\beta_1)\right)\Big)\\ &\quad -\tfrac{1}{\beta_1}\Big(\tfrac{1}{\alpha_1-\beta_1}\left(\exp(\alpha_1)-\exp(\beta_1)\right)-\Big(\tfrac{1}{\alpha_1}\left(\exp(\alpha_1)-1\right)\Big)\Big)\Big]\end{aligned}$$

By substituting the expressions of α and β, we obtain:

$$\begin{aligned}\exp(S_1^\ell-\mathbf{x}_1^\ell)\,\mathcal{I}_1 = \frac{\mathbf{b}_2^\ell\cdot(S_2^\ell-S_1^\ell)}{\mathbf{b}_1^\ell\cdot(S_3^\ell-S_1^\ell)}\Big[&\frac{1}{\mathbf{b}_1^\ell\cdot(S_2^\ell-S_3^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_2^\ell-\mathbf{x}_1^\ell))\\ &-\frac{1}{\mathbf{b}_1^\ell\cdot(S_2^\ell-S_3^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_2^\ell-\mathbf{x}_1^\ell))-\exp(\mathbf{b}_1^\ell\cdot(S_3^\ell-\mathbf{x}_1^\ell))\Big)\Big)\\ &-\frac{1}{\mathbf{b}_1^\ell\cdot(S_2^\ell-S_1^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_2^\ell-\mathbf{x}_1^\ell))\\ &-\frac{1}{\mathbf{b}_1^\ell\cdot(S_2^\ell-S_1^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_2^\ell-\mathbf{x}_1^\ell))-\exp(\mathbf{b}_1^\ell\cdot(S_1^\ell-\mathbf{x}_1^\ell))\Big)\Big)\Big]\end{aligned}$$

and

$$\begin{aligned}
\exp(S_1^\ell - \mathbf{x}_1^\ell)\mathcal{I}_2 &= \tfrac{\mathbf{b}_2^\ell\cdot(S_3^\ell - S_1^\ell)}{\mathbf{b}_1^\ell\cdot(S_3^\ell - S_1^\ell)}\Big[\tfrac{1}{\mathbf{b}_1^\ell\cdot(S_2^\ell - S_3^\ell)}\Big(- \Big(\exp(\mathbf{b}_1^\ell\cdot(S_2^\ell - \mathbf{x}_1^\ell)) \\
&\quad - \tfrac{1}{\mathbf{b}_1^\ell\cdot(S_2^\ell - S_3^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_2^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_1^\ell\cdot(S_3^\ell - \mathbf{x}_1^\ell))\Big)\Big) \\
&\quad + \Big(\exp(\mathbf{b}_1^\ell\cdot(S_2^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_1^\ell\cdot(S_3^\ell - \mathbf{x}_1^\ell))\Big)\Big) \\
&\quad - \tfrac{1}{\mathbf{b}_1^\ell\cdot(S_3^\ell - S_1^\ell)}\Big(\tfrac{1}{\mathbf{b}_1^\ell\cdot(S_2^\ell - S_3^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_2^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_1^\ell\cdot(S_3^\ell - \mathbf{x}_1^\ell))\Big) \\
&\quad - \Big(\tfrac{1}{\mathbf{b}_1^\ell\cdot(S_2^\ell - S_1^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_2^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_1^\ell\cdot(S_1^\ell - \mathbf{x}_1^\ell))\Big)\Big)\Big)\Big]
\end{aligned}$$

Let us now treat the case $\beta_1 = 0$:

$$\begin{aligned}
\mathcal{I}_1 + \mathcal{I}_2 &= \alpha_2 \int_0^1 \hat{x}\exp(\alpha_1\,\hat{x}) \int_0^{-\hat{x}+1} 1\,d\hat{y}\,d\hat{x} + \beta_2 \int_0^1 \exp(\alpha_1\,\hat{x}) \int_0^{-\hat{x}+1} \hat{y}\,d\hat{y}\,d\hat{x} \\
&= (\tfrac{\beta_2}{2} - \alpha_2)\Big(\tfrac{1}{\alpha_1}\Big(\exp(\alpha_1) - \tfrac{2}{\alpha_1}\exp(\alpha_1) + \tfrac{2}{\alpha_1^2}(\exp(\alpha_1) - 1)\Big)\Big) \\
&\quad + (\alpha_2 - \beta_2)\Big(\tfrac{1}{\alpha_1}\Big(\exp(\alpha_1) - \tfrac{1}{\alpha_1}(\exp(\alpha_1) - 1)\Big)\Big) \\
&\quad + \tfrac{\beta_2}{2}\Big(\tfrac{1}{\alpha_1}(\exp(\alpha_1) - 1)\Big)
\end{aligned}$$

By substituting the expressions of α and β, we obtain:

$$\begin{aligned}
\exp(S_1^\ell - \mathbf{x}_1^\ell)(\mathcal{I}_1 + \mathcal{I}_2) &= \frac{(\frac{\mathbf{b}_2^\ell \cdot (S_3^\ell - S_1^\ell)}{2} - \mathbf{b}_2^\ell \cdot (S_2^\ell - S_1^\ell))}{\mathbf{b}_1^\ell \cdot (S_2^\ell - S_1^\ell)} \Big(\exp(\mathbf{b}_1^\ell \cdot (S_2^\ell - \mathbf{x}_1^\ell)) - \frac{2}{\mathbf{b}_1^\ell \cdot (S_2^\ell - S_1^\ell)} \exp(\mathbf{b}_1^\ell \cdot (S_2^\ell - \mathbf{x}_1^\ell)) \\
&\quad + \frac{2}{(\mathbf{b}_1^\ell \cdot (S_2^\ell - S_1^\ell))^2} \Big(\exp(\mathbf{b}_1^\ell \cdot (S_2^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_1^\ell \cdot (S_1^\ell - \mathbf{x}_1^\ell)) \Big) \Big) \\
&\quad + \frac{(\mathbf{b}_2^\ell \cdot (S_2^\ell - S_3^\ell))}{\mathbf{b}_1^\ell \cdot (S_2^\ell - S_1^\ell)} \Big(\exp(\mathbf{b}_1^\ell \cdot (S_2^\ell - S_1^\ell)) \\
&\quad - \frac{1}{\mathbf{b}_1^\ell \cdot (S_2^\ell - S_1^\ell)} \Big(\exp(\mathbf{b}_1^\ell \cdot (S_2^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_1^\ell \cdot (S_1^\ell - \mathbf{x}_1^\ell)) \Big) \Big) \\
&\quad + \frac{\mathbf{b}_2^\ell \cdot (S_3^\ell - S_1^\ell)}{2\,(\mathbf{b}_1^\ell \cdot (S_2^\ell - S_1^\ell))} \Big(\exp(\mathbf{b}_1^\ell \cdot (S_2^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_1^\ell \cdot (S_1^\ell - \mathbf{x}_1^\ell)) \Big)
\end{aligned}$$

Let us now treat the case $\alpha_1 = 0$:

$$\begin{aligned}
\mathcal{I}_1 &= \frac{\alpha_2}{\beta_1} \int_0^1 \hat{x} \Big[\exp(\beta_1 (-\hat{x} + 1)) - 1 \Big] \, d\hat{x} \\
&= \frac{\alpha_2}{\beta_1} \Big[-\frac{1}{\beta_1} \Big(1 - \frac{1}{\beta_1} (\exp(\beta_1) - 1) \Big) - \frac{1}{2} \Big] \\
\mathcal{I}_2 &= \beta_2 \int_0^1 \int_0^{-\hat{x}+1} \hat{y} \exp(\beta_1 \hat{y}) \, d\hat{y} \, d\hat{x} \\
&= \frac{\beta_2}{\beta_1} \Big[\frac{1}{\beta_1} \Big(\Big(1 + \frac{1}{\beta_1} (1 - \exp(\beta_1)) \Big) - (1 - \exp(\beta_1)) \Big) - \frac{1}{\beta_1} \Big(\frac{1}{-\beta_1} (1 - \exp(\beta_1)) - 1 \Big) \Big]
\end{aligned}$$

By substituting the expressions of α and β, we obtain:

$$\begin{aligned}
\exp(S_1^\ell - \mathbf{x}_1^\ell)\mathcal{I}_1 &= \tfrac{\mathbf{b}_2^\ell\cdot(S_2^\ell - S_1^\ell)}{\mathbf{b}_1^\ell\cdot(S_3^\ell - S_1^\ell)}\Big[-\tfrac{1}{2}\exp(\mathbf{b}_1^\ell\cdot(S_1^\ell - \mathbf{x}_1^\ell)) - \tfrac{1}{\mathbf{b}_1^\ell\cdot(S_3^\ell - S_1^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_1^\ell - \mathbf{x}_1^\ell)) \\
&\quad -\tfrac{1}{\mathbf{b}_1^\ell\cdot(S_3^\ell - S_1^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_3^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_1^\ell\cdot(S_1^\ell - \mathbf{x}_1^\ell))\Big)\Big)\Big] \\
\exp(S_1^\ell - \mathbf{x}_1^\ell)\mathcal{I}_2 &= \tfrac{\mathbf{b}_2^\ell\cdot(S_3^\ell - S_1^\ell)}{\mathbf{b}_1^\ell\cdot(S_3^\ell - S_1^\ell)}\Big[\tfrac{1}{\mathbf{b}_1^\ell\cdot(S_3^\ell - S_1^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_1^\ell - \mathbf{x}_1^\ell)) + \tfrac{1}{\mathbf{b}_1^\ell\cdot(S_3^\ell - S_1^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_1^\ell - \mathbf{x}_1^\ell)) \\
&\quad -\exp(\mathbf{b}_1^\ell\cdot(S_3^\ell - \mathbf{x}_1^\ell))\Big) - \Big(\exp(\mathbf{b}_1^\ell\cdot(S_1^\ell - \mathbf{x}_1^\ell)) - \exp(\mathbf{b}_1^\ell\cdot(S_3^\ell - \mathbf{x}_1^\ell))\Big)\Big) \\
&\quad +\tfrac{1}{\mathbf{b}_1^\ell\cdot(S_3^\ell - S_1^\ell)}\Big(\tfrac{1}{\mathbf{b}_1^\ell\cdot(S_3^\ell - S_1^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_1^\ell - \mathbf{x}_1^\ell)) \\
&\quad -\exp(\mathbf{b}_1^\ell\cdot(S_3^\ell - \mathbf{x}_1^\ell))\Big) + \exp(\mathbf{b}_1^\ell\cdot(S_1^\ell - \mathbf{x}_1^\ell))\Big)\Big]
\end{aligned}$$

Let us now treat the case $\alpha_1 - \beta_1 = 0$:

$$\begin{aligned}
\mathcal{I}_1 &= \tfrac{\alpha_2}{\beta_1}\int_0^1 \hat{x}\exp(\alpha_1\,\hat{x})\Big[\exp(\beta_1\,(-\hat{x}+1)) - 1\Big]\,d\hat{x} \\
&= \tfrac{\alpha_2}{\beta_1}\Big[\exp(\beta_1)\int_0^1 \hat{x} - \int_0^1 \hat{x}\exp(\alpha_1\,\hat{x})\Big] \\
&= \tfrac{\alpha_2}{\beta_1}\Big[\tfrac{1}{2}\exp(\beta_1) - \tfrac{1}{\alpha_1}\Big(\exp(\alpha_1) - \tfrac{1}{\alpha_1}\,(\exp(\alpha_1) - 1)\Big)\Big]
\end{aligned}$$

$$
\begin{aligned}
\mathcal{I}_2 &= \beta_2 \int_0^1 \exp(\alpha_1\,\hat{x}) \int_0^{-\hat{x}+1} \hat{y}\exp(\beta_1\,\hat{y})\,d\hat{y}\,d\hat{x} \\
&= \tfrac{\beta_2}{\beta_1} \int_0^1 \exp(\alpha_1\,\hat{x}) \Big[(-\hat{x}+1)\exp(\beta_1\,(-\hat{x}+1)) - \frac{1}{\beta_1}\left(\exp(\beta_1\,(-\hat{x}+1))-1\right)\Big] \\
&= \tfrac{\beta_2}{\beta_1} \Big[\int_0^1 \exp(\beta_1)\Big(-\hat{x}+1\Big) - \frac{1}{\beta_1}\int_0^1 \exp(\beta_1) - \exp(\alpha_1\,\hat{x})\Big] \\
&= \tfrac{\beta_2}{\beta_1} \Big[\tfrac{1}{2}\exp(\beta_1) - \tfrac{1}{\beta_1}\Big(\exp(\beta_1) - \tfrac{1}{\alpha_1}(\exp(\alpha_1)-1)\Big)\Big]
\end{aligned}
$$

By substituting the expressions of α_i and β_i, $i = 1,\, 2$, we obtain:

$$
\begin{aligned}
\exp(S_1^\ell - \mathbf{x}_1^\ell)\mathcal{I}_1 &= \tfrac{\mathbf{b}_2^\ell\cdot(S_2^\ell - S1)}{\mathbf{b}_1^\ell\cdot(S_3^\ell - S_1^\ell)}\Big[\tfrac{1}{2}\exp(\mathbf{b}_1^\ell\cdot(S_3^\ell-\mathbf{x}_1^\ell)) - \tfrac{1}{\mathbf{b}_1^\ell\cdot(S_2^\ell-S_1^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_2^\ell-\mathbf{x}_1^\ell)) \\
&\quad - \tfrac{1}{\mathbf{b}_1^\ell\cdot(S_2^\ell-S_1^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_2^\ell-\mathbf{x}_1^\ell)) - \exp(\mathbf{b}_1^\ell\cdot(S_1^\ell-\mathbf{x}_1^\ell))\Big)\Big)\Big] \\
\exp(S_1^\ell - \mathbf{x}_1^\ell)\mathcal{I}_2 &= \tfrac{\mathbf{b}_2^\ell\cdot(S_3^\ell - S_1^\ell)}{\mathbf{b}_1^\ell\cdot(S_3^\ell - S_1^\ell)}\Big[\tfrac{1}{2}\exp(\mathbf{b}_1^\ell\cdot(S_3^\ell-\mathbf{x}_1^\ell)) - \tfrac{1}{\mathbf{b}_1^\ell\cdot(S_3^\ell-S_1^\ell)}\Big(\exp(\mathbf{b}_1^\ell\cdot(S_3^\ell-\mathbf{x}_1^\ell)) \\
&\quad - \tfrac{1}{\mathbf{b}_1^\ell\cdot(S_2^\ell-S_1^\ell)}(\exp(\mathbf{b}_1^\ell\cdot(S_2^\ell-S_1^\ell)) - \exp(\mathbf{b}_1^\ell\cdot(S_1^\ell-\mathbf{x}_1^\ell)))\Big)\Big]
\end{aligned}
$$

From a pedagogical point of view, one could, for example, compute integrals of the form

$$\int_{K_\ell} \exp(\mathbf{b}_1^\ell\cdot(\mathbf{x}-\mathbf{x}_1^\ell))\,(\mathbf{b}_2^\ell\cdot(\mathbf{x}-\mathbf{x}_2^\ell))^2\,d\mathbf{x}$$

and other similar expressions. However, we do not present these computations here, as they would involve a large amount of complex calculations.

Remark 11.9 *Remark 11.8 applies to this integral, but with* $\mathbf{b}$ *replaced by* $\mathbf{b}_1^\ell$.

11.5.3 Exact integration of an exponential over a edge F_f

Let $\widehat{F}$ be the reference segment with vertices $(0,0)$, $(1,0)$, and let F_f be a segment in the mesh with vertices P_1^f and P_2^f, where $P_i^f = (x_i^f, y_i^f)$ for $i = 1, 2$.

11.5. Exact evaluation of the integrals involved in the matrices

We define an affine transformation $T_f : [0,\,1] \to F^f$. It can be written in the following form:

$$T_f(t) = (P_1^f - P_0^f)\, t + P_0^f \tag{11.11}$$

Since T_f is affine, the Jacobian J_f is constant over $\widehat{F}$, and $T_f([0,\,1]) = F_f$ and $|J_f| = |F_f|$. In what follows, we only present the final result of the computation, since the detailed derivation has already been carried out on a triangle. The case of a segment is simpler and does not require further elaboration.
Let $\mathbf{b}_f, \mathbf{x}_1^f \in \mathrm{R}^2$. We aim to compute the exact integration of the term

$$\int_{F_f} \exp(\mathbf{b}_f \cdot (\mathbf{x} - \mathbf{x}_1^f)).$$

According to the geometric mapping T_f (11.11), for any $\mathbf{x} \in F_f$, there exists t such that

$$\mathbf{x} = (P_1^f - P_0^f)\, t + P_0^f.$$

For simplicity we denote $\delta P = P_1^f - P_0^f$. By transforming to the reference segment, we obtain

$$\int_{F_f} \exp(\mathbf{b}_f \cdot (\mathbf{x} - \mathbf{x}_1^f)) = |F| \exp(\mathbf{b}_f \cdot (P_0^f - \mathbf{x}_1^f)) \int_0^1 \exp(\mathbf{b}_f \cdot \delta P\, t)\, dt \tag{11.12}$$

By a simple calculation if $\mathbf{b}_f \cdot \delta P \neq 0$:

$$\int_{F_f} \exp(\mathbf{b}_f \cdot (\mathbf{x} - \mathbf{x}_1^f)) = \frac{|F_f|}{\mathbf{b}_f \cdot (P_1^f - P_0^f)} \left(\exp(\mathbf{b}_f \cdot (P_1^f - \mathbf{x}_1^f)) - \exp(\mathbf{b}_f \cdot (P_0^f - \mathbf{x}_1^f)) \right).$$

On the other hand, if $\mathbf{b}_f \cdot (P_1^f - P_0^f) = 0$, we get:

$$\int_{F_f} \exp(\mathbf{b}_f \cdot (\mathbf{x} - \mathbf{x}_1^f)) = |F_f| \exp(\mathbf{b}_f \cdot (P_0^f - \mathbf{x}_1^f)).$$

We can use the same principle to continue finding the exact values of the integral of the exponential against polynomials, we will not write it here since the calculations are very complex and long.

In this chapter, we introduced a new non-polynomial basis, consisting of exponential-type functions in one, two, and three dimensions. This basis is inspired by the exact solution of the advection–diffusion problem in $1d$. In $2d$ and $3d$, the basis can be enriched with additional terms so that it matches exactly the exact solution of the advection–diffusion problem in these dimensions.
Then, we proved the inverse inequality (4.4) in the space generated by this non-polynomial basis. Finally, we briefly presented how to compute the exact integrals for this basis using a reference element.

Chapter 12

The Oseen problem

Contents

What is new in this chapter in section 12.3 is the demonstration that the discretized Oseen problem is well-posed for the polynomial basis using T-coercivity and we replace $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_h$ in such a way that $\boldsymbol{\beta}_h$ preserves, as much as possible, the main characteristics of $\boldsymbol{\beta}$, such as $\nabla \cdot \boldsymbol{\beta} = 0$ and its boundary values. And in section 12.4 several methods to choose $\mathbf{b}$ in the non polynomial basis.

12.1 Brief Bibliography

Several studies have investigated both the uniqueness of solutions and the development of robust numerical methods for the Oseen problem. Anaya and al. [6] provide a detailed error analysis for finite element methods applied to the steady problem in vorticity form, demonstrating stability and convergence. Cockburn, Kanschat, and Schötzau [31] focus on stabilized schemes for the Oseen problem, ensuring optimal convergence. Theoretical results on uniqueness and stability are also thoroughly covered in the book by Galdi [55]. Finally, there are many other publications dedicated to the Oseen problem in a more general context

as these publications [5, 3, 4, 53], $\cdots$.
In this work, we study the Oseen problem to show that it is well-posed using the $T-$coercivity method, and to perform numerical simulations using the non-polynomial basis defined in the previous chapter 11.

12.2 Continuous Oseen problem

The Oseen problem can be seen as an intermediate model between the Stokes and the Navier–Stokes equations. It is obtained by linearizing the stationary Navier–Stokes system around a given velocity field, keeping a convection term with a prescribed velocity. This formulation allows us to account for moderate convective effects while preserving the linearity of the system. The Oseen problem is particularly useful in the study of steady flows and is a crucial step in the analysis and numerical approximation of the full Navier–Stokes equations. It retains the key features of the Navier–Stokes problem, such as the coupling between velocity and pressure and the incompressibility condition, while simplifying the mathematical treatment by fixing the convective velocity. In our model, the vector field $\mathbf{u}$ denotes the fluid velocity, p is the pressure and $\boldsymbol{\beta}$ corresponds to a given approximate velocity field used to simplify the convection term.

Let $\Gamma_1, \Gamma_2 \subset \partial\Omega$ be such that $\Gamma_1 \cup \Gamma_2 = \partial\Omega$ and $\Gamma_1 \cap \Gamma_2 = \emptyset$. We consider the Oseen problem with different types of boundary conditions:

$$\left\{\begin{array}{l} \text{Find } (\mathbf{u}, p) \text{ such that:} \\ -\nu\,\boldsymbol{\Delta}\,\mathbf{u} + (\boldsymbol{\beta}\cdot\mathbf{grad}\,)\mathbf{u} + \mathbf{grad}\,p = \mathbf{f}, \\ \qquad\qquad \operatorname{div}\mathbf{u} = g, \\ \qquad\qquad \mathbf{u}_{|\Gamma_1} = \mathbf{g}, \\ \qquad (-\nu\,\mathbf{Grad}\,\mathbf{u} : \mathbf{n} + pI_d)_{|\Gamma_2} = \mathbf{s}. \end{array}\right. \tag{12.1}$$

Here, g, $\mathbf{g}$, and $\mathbf{s}$ are regular functions. However, in the theoretical analysis, we simplify the boundary conditions by considering only Dirichlet conditions and setting $g = 0$ in order to satisfy the incompressibility condition. We thus consider the following Oseen problem with

homogeneous Dirichlet boundary condition:

$$\left\{\begin{array}{l} \text{Find } (\mathbf{u},p) \text{ such that} \\ -\nu\,\boldsymbol{\Delta}\,\mathbf{u} + (\boldsymbol{\beta}\cdot\mathbf{grad}\,)\,\mathbf{u} + \mathbf{grad}\,p \;=\; \mathbf{f} \\ \qquad\qquad\qquad\qquad\qquad\qquad \operatorname{div}\mathbf{u} \;=\; 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad \mathbf{u}_{|\partial\Omega} \;=\; 0 \end{array}\right. \tag{12.2}$$

with normalization condition for p: $\displaystyle\int_\Omega p = 0$, The constant parameter $\nu > 0$ represents the kinematic viscosity. The vector field $\mathbf{f}$ represents the external force field on the fluid, and we assume that there exists $s > \frac{1}{2}$ such that $\boldsymbol{\beta} \in \mathbf{H}^{1+s}(\Omega)$, so has the same properties as $\mathbf{u}$, that is:

$$\boldsymbol{\beta} \in \mathbf{H}^1_0(\Omega) \cap \mathbf{H}^{1+s}(\Omega) \quad \text{and} \quad \operatorname{div}\boldsymbol{\beta} = 0.$$

Remark 12.1 *We assume that $\boldsymbol{\beta} \in \mathbf{H}^{1+s}(\Omega)$ with $s > \frac{1}{2}$. Then, using [49, Thm. 2.31], we obtain that $\boldsymbol{\beta} \in \mathbf{L}^\infty(\Omega)$ in both two and three dimensions.*

We consider the following continuous bilinear form:

$$a_O(\cdot,\cdot) : \left\{\begin{array}{rcl} \mathcal{X}\times\mathcal{X} & \to & \mathbb{R} \\ ((\mathbf{u}',p'),(\mathbf{v},q)) & \mapsto & a_S\left((\mathbf{u}',p'),(\mathbf{v},q)\right) + t(\mathbf{u}',\,\mathbf{v}) \end{array}\right. , \tag{12.3}$$

where $t(\mathbf{u}',\,\mathbf{v}) = \displaystyle\int_\Omega (\boldsymbol{\beta}\cdot\mathbf{Grad}\,)\mathbf{u}'\cdot\mathbf{v}$. This integral is well defined since the three fields are in $\mathbf{H}^1(\Omega)$, as shows a Hölder inequality with $(\mathbf{L}^4,\mathbf{L}^2,\mathbf{L}^4)$ combined with injection of $\mathbf{H}^1$ into $\mathbf{L}^4$. We can express problem (12.2) equivalently as follows: Find $(\mathbf{u},p) \in \mathcal{X}$ such that for all $(\mathbf{v},q) \in \mathcal{X}$:

$$a_O\left((\mathbf{u},p),(\mathbf{v},q)\right) = \langle \mathbf{f},\mathbf{v}\rangle_{\mathbf{H}^1_0(\Omega)}. \tag{12.4}$$

Remark 12.2 *Let $\mathbf{q} \in H(\operatorname{div};\Omega)$ and $\mathbf{w},\,\mathbf{v} \in \mathbf{H}^1(\Omega)$. By integration by parts, we have*

$$\int_\Omega (\mathbf{q}\cdot\mathbf{Grad}\,)\mathbf{w}\cdot\mathbf{v} = -\int_\Omega (\mathbf{q}\cdot\mathbf{Grad}\,)\mathbf{v}\cdot\mathbf{w} - \int_\Omega (\operatorname{div}\mathbf{q})\,\mathbf{v},\,\mathbf{w} + \int_{\partial\Omega} \mathbf{q}\cdot\mathbf{n}\,\mathbf{v}\cdot\,\mathbf{w}. \tag{12.5}$$

Proposition 12.3 *The bilinear form $a_O(\cdot,\cdot)$ is T-coercive.*

Proof. The proof is similar to that of Proposition 5.1. For all

$$((\mathbf{u}',p'),(\mathbf{v},q)) \in \mathcal{X}\times\mathcal{X},$$

we have

$$a_O\left((\mathbf{u}',p'),(\mathbf{v},q)\right) = a_S\big((\mathbf{u}',p'),(\mathbf{v},q)\big) + t\big((\mathbf{u}',p'),(\mathbf{v},q)\big), \tag{12.6}$$

Let us consider $(\mathbf{u}', p') \in \mathcal{X}$ and construct a bijective operator satisfying (2.15).
According to Proposition 2.3 and Lemma 2.10, there exists $\tilde{\mathbf{v}}_{p'} \in \mathbf{V}^{\perp}$ such that

$$\text{div } \tilde{\mathbf{v}}_{p'} = p' \quad \text{in } \Omega, \quad \|\tilde{\mathbf{v}}_{p'}\|_{\mathbf{H}_0^1(\Omega)} \leq C_{\text{div}} \|p'\|_{L^2(\Omega)}, \quad \text{and} \quad \|\tilde{\mathbf{v}}_{p'}\|_{\mathbf{L}^2(\Omega)} \leq C_{\text{div},2} \|p'\|_{L^2(\Omega)}, \tag{12.7}$$

where : $C_{\text{div},2} = C_{\text{div}}\, C_{PS}\, h_\Omega$.
Let us set

$$(\mathbf{v}^*, q^*) := (\lambda \mathbf{u}' - \nu^{-1} \tilde{\mathbf{v}}_{p'},\, -\lambda p'),$$

where $\lambda > 0$. From the proof of Proposition 5.1, we obtain for all $\eta > 0$:

$$a_S((\mathbf{u}', p'), (\mathbf{v}^*, q^*)) \geq \left(\nu\lambda - \frac{\eta}{2}\right) \|\mathbf{u}'\|^2_{\mathbf{H}_0^1(\Omega)} + \left(\nu^{-1} - \frac{\eta^{-1}}{2}(C_{\text{div}})^2\right) \|p'\|^2_{L^2(\Omega)}, \tag{12.8}$$

where η is a positive parameter to be determined, introduced in inequality (5.13).
For the second term in the right-hand side of (12.6), we have:

$$t(\mathbf{u}',\, \mathbf{v}^*) = -\nu^{-1}\, t(\mathbf{u}',\, \tilde{\mathbf{v}}_{p'})$$

Using (12.7) and Lemma 2.10 and remark 12.1, then we obtain:

$$\begin{aligned} \nu^{-1} \left|t(\mathbf{u}', \tilde{\mathbf{v}}_{p'})\right| &\leq \nu^{-1} \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)} \|\mathbf{Grad}\, \mathbf{u}'\|_{\mathbf{L^2(\Omega)}} \|\tilde{\mathbf{v}}_{\mathbf{p}'}\|_{\mathbf{L^2(\Omega)}}, \\ &\leq \nu^{-1} (C_{PS}) h_\Omega \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)} \|\mathbf{u}'\|_{\mathbf{H}_0^1(\Omega)} \|\tilde{\mathbf{v}}_{p'}\|_{\mathbf{H}_0^1(\Omega)} \\ &\leq \nu^{-1} C_{PS}\, C_{\text{div}} h_\Omega \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)} \|\mathbf{u}'\|_{\mathbf{H}_0^1(\Omega)} \|p'\|_{L^2(\Omega)} \end{aligned}$$

Using Young's inequality, for all $\alpha > 0$, we have:

$$\nu^{-1} \left|t(\mathbf{u}', \tilde{\mathbf{v}}_{p'})\right| \leq \frac{\alpha}{2} (C_{PS})^2 \|\boldsymbol{\beta}\|^2_{\mathbf{L}^\infty(\Omega)} \|\mathbf{u}'\|^2_{\mathbf{H}_0^1(\Omega)} + \frac{\alpha^{-1}}{2} (C_{\text{div}}\, \nu^{-1})^2 \|p\|^2_{L^2(\Omega)} \tag{12.9}$$

Using bounds (12.8) and (12.9) in (12.6):

$$a_O\left((\mathbf{u}', p'), (\mathbf{v}^*, q^*)\right) \geq \begin{pmatrix} \left(\nu\lambda - \frac{\eta}{2} - \frac{\alpha}{2} (C_{PS})^2 \|\boldsymbol{\beta}\|^2_{\mathbf{L}^\infty(\Omega)}\right) \|\mathbf{u}'\|^2_{\mathbf{H}_0^1(\Omega)} \\ + \left(\nu^{-1} - \frac{\eta^{-1}}{2}(C_{\text{div}})^2 - \frac{\alpha^{-1}}{2}(C_{\text{div}}\, \nu^{-1})^2\right) \|p'\|^2_{L^2(\Omega)} \end{pmatrix}. \tag{12.10}$$

We first choose $\eta = 2\,\nu\,(C_{\text{div}})^2$ and $\alpha = 2\,(C_{\text{div}})^2\,\nu^{-1}$ so that the coefficient in front of $\|p'\|^2_{L^2(\Omega)}$ is equal to $\frac{\nu^{-1}}{2}$, and we look for $\lambda > 0$ such that $\left(\nu\lambda - \frac{\eta}{2} - \frac{\alpha}{2} (C_{PS})^2 \|\boldsymbol{\beta}\|^2_{\mathbf{L}^\infty(\Omega)}\right) > 0$, which amounts to requiring

$$\lambda > (C^2_{\text{div}} + (\nu^{-1}\, C_{\text{div}}\, C_{PS}\, \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)})^2). \tag{12.11}$$

According to the above, under the sufficient condition that $\lambda > (C_{\text{div}}^2 + (\nu^{-1} C_{\text{div},2} \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)})^2)$, we have proved that the operator $T_\lambda \in \mathcal{L}(X)$ defined by $T_\lambda((\mathbf{u}', p')) = (\lambda \mathbf{u}' - \nu^{-1}\tilde{\mathbf{v}}_{p'}, -\lambda p')$ is such that:

$$\exists \alpha_\lambda > 0,\ \forall (\mathbf{u}', p') \in X,\ a_O((\mathbf{u}', p'), T_\lambda((\mathbf{u}', p'))) \geq \alpha_\lambda \|(\mathbf{u}', p')\|_{\mathcal{X}}^2. \tag{12.12}$$

The continuity and bijectivity of T_λ are established in the proof of Proposition (5.1).

□

12.3 Discretisation of the Oseen problem

The Oseen problem can be seen as a Stokes problem with an additional convection term. Numerically, we need to implement the advection matrix corresponding to $(\boldsymbol{\beta} \cdot \nabla)\mathbf{u}$, which depends on $\boldsymbol{\beta}$. Since $\boldsymbol{\beta}$ may take various forms—possibly very complicated—the exact computation of the integrals involving $\boldsymbol{\beta}$ against the basis functions of the approximation space can become difficult, especially because, with the non-polynomial basis, I compute the integrals exactly for accuracy and stability reasons explained above. For this reason, I chose to replace $\boldsymbol{\beta}$ with $\boldsymbol{\beta}_h$, where $\boldsymbol{\beta}_h$ is the velocity part of the numerical approximation (with usual DG polynomial spaces) of problem (12.13) below (whose solution is obviously $(\boldsymbol{\beta}, 0)$). One may ask why $\boldsymbol{\beta}_h$ is not defined directly as an interpolation or projection of $\boldsymbol{\beta}$. The purpose is to transfer, as much as possible in a weak sense, the property $\nabla \cdot \boldsymbol{\beta} = 0$ to $\boldsymbol{\beta}_h$. The Stokes problem we approach to find an accurate approximation of $\boldsymbol{\beta}$ by an element of the DG space is given by

$$\left\{ \begin{array}{l} \text{Find } (\boldsymbol{\beta}, p_\beta) \text{ such that} \\ -\nu\, \boldsymbol{\Delta}\, \boldsymbol{\beta} + \nabla p_\beta = -\nu\, \boldsymbol{\Delta}\, \boldsymbol{\beta}, \\ \qquad\qquad \operatorname{div} \boldsymbol{\beta} = 0, \\ \qquad\qquad \boldsymbol{\beta}|_{\partial\Omega} = 0. \end{array} \right. \tag{12.13}$$

Remark 12.4 *We choose $\boldsymbol{\beta}_h$ to satisfy a Stokes problem rather than a Poisson problem in order to ensure that*

$$b_h(\boldsymbol{\beta}_h, q_h) = 0, \quad \forall q_h \in L_h.$$

It is clear that $(\boldsymbol{\beta}, 0)$ is the solution of problem (12.13).

Since we replace $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_h$, in the variational formulation of the advection term, then it's modified as follows:

Based on the form $t_h^0(\mathbf{u}_h,\, \mathbf{v}_h) := ((\boldsymbol{\beta}_h \cdot \mathbf{Grad}_h)\mathbf{u}_h,\, \mathbf{v}_h)_{\mathbf{L}^2(\Omega)}$, we have:

$$\begin{aligned} t_h^0(\mathbf{u}_h,\, \mathbf{u}_h) &= \tfrac{1}{2} \sum_{f\in\mathcal{I}_F} (\llbracket \boldsymbol{\beta}_h \rrbracket \cdot \mathbf{n}_f,\ \{\mathbf{u}_h \cdot \mathbf{u}_h\})_{\mathbf{L}^2(F_f)} + \sum_{f\in\mathcal{I}_F^i} (\,\{\boldsymbol{\beta}_h\}\cdot\mathbf{n}_f,\, \llbracket \mathbf{u}_h \rrbracket \cdot \{\mathbf{u}_h\})_{\mathbf{L}^2(F_f)} \\ &\quad -\tfrac{1}{2}(\operatorname{div} \boldsymbol{\beta}_h\, \mathbf{u}_h,\, \mathbf{u}_h)_{L^2(\Omega)}. \end{aligned}$$

We then define:

$$\begin{aligned} \tilde{t}_h(\mathbf{u}_h,\, \mathbf{v}_h) &= t_h^0(\mathbf{u}_h,\, \mathbf{v}_h) + \tfrac{1}{2}(\operatorname{div}_h \boldsymbol{\beta}_h\, \mathbf{u}_h,\, \mathbf{u}_h)_{\mathbf{L}^2(\Omega)} \\ &\quad -\tfrac{1}{2} \sum_{f\in\mathcal{I}_F} (\llbracket \boldsymbol{\beta}_h \rrbracket \cdot \mathbf{n}_f,\ \{\mathbf{u}_h\, \mathbf{v}_h\})_{\mathbf{L}^2(F_f)} - \sum_{f\in\mathcal{I}_F^i} (\,\{\boldsymbol{\beta}_h\}\cdot\mathbf{n}_f,\, \llbracket \mathbf{u}_h \rrbracket \cdot \{\mathbf{v}_h\})_{L^2(F_f)} \end{aligned} \tag{12.14}$$

Remark 12.5 *We compute $\boldsymbol{\beta}_h$ using the discontinuous Galerkin scheme $\mathbf{P}^1_{\text{disc}} - P^1_{\text{disc}}$. Using Theorem 7.6, we obtain:*

$$\|\boldsymbol{\beta}_h\|_{vel} \leq (C_{\boldsymbol{\beta}}\, h^s + \|\boldsymbol{\beta}\|_{vel}) \tag{12.15}$$

Where $C_{\boldsymbol{\beta}} = C_{\mathcal{X}_\star}\, C_S \left(\nu^{-1}\, (h_\Omega)^{1-s}\, \|\nu\, \Delta\boldsymbol{\beta}\|_{\mathbf{H}^{s-1}(\Omega)}\right)$.

Lemma 12.1 *For all $\mathbf{u}_h \in \mathbf{X}_h$, we have $\tilde{t}_h(\mathbf{u}_h,\, \mathbf{u}_h) = 0$.*

Therefore, for the discretization of the Oseen problem, everything is in place; it remains only to verify whether the formulation is well-posed. Recall that the bilinear forms $a_h(\cdot,\cdot)$, $b_h(\cdot,\cdot)$, and $\tilde{s}_h(\cdot,\cdot)$ are already defined in equations (6.25), (6.26), and (6.27), respectively. Now we can propose the discrete variational formulation of the Oseen problem (12.2)

$$\left\{ \begin{array}{l} \text{Find } (\mathbf{u}_h, p_h) \in \mathbf{X}_h \times L_h \text{ such that:} \\ \nu\, a_h(\mathbf{u}_h, \mathbf{v}_h) + \tilde{t}_h(\mathbf{u}_h,\, \mathbf{v}_h) + b_h(\mathbf{v}_h, p_h) = (\mathbf{f}, \mathbf{v}_h)_{\mathbf{L}^2(\Omega)} \qquad \forall \mathbf{v}_h \in \mathbf{X}_h, \\ \qquad\qquad -b_h(\mathbf{u}_h, q_h) + \lambda\, \tilde{s}_h(p_h, q_h) = 0 \qquad\qquad \forall q_h \in L_h. \end{array} \right. \tag{12.16}$$

We consider the following bilinear form:

$$a_{O,h} : \left\{ \begin{array}{rcl} \mathcal{X}_h \times \mathcal{X}_h & \to & \mathbb{R} \\ (\,(\mathbf{u}_h', p_h'), (\mathbf{v}_h, q_h)\,) & \mapsto & a_{S,h}(\,(\mathbf{u}_h', p_h'), (\mathbf{v}_h, q_h)\,) + \tilde{t}_h(\mathbf{u}_h', \mathbf{v}_h) \end{array} \right. . \tag{12.17}$$

Problem (12.16) can be rewritten in the form:

$$\left\{\begin{array}{c}\text{Find } (\mathbf{u}_h, p_h) \in \mathcal{X}_h \text{ such that for all } (\mathbf{v}_h, q_h) \in \mathcal{X}_h,\\ a_{O,h}(\,(\mathbf{u}_h, p_h), (\mathbf{v}, q_h)\,) = (\mathbf{f},\, \mathbf{v}_h)_{\mathbf{L}^2(\Omega)}\end{array}\right. \tag{12.18}$$

We need the following lemma to prove the coercivity of the bilinear form $a_{O,h}(\cdot,\,\cdot)$.

Lemma 12.2 ([40, lemma 6.11]) *there exists a constant $\hat{C}_\Pi$ such that for all $\mathbf{v} \in \mathbf{H}_0^1(\Omega)$, we have:*

$$\|\Pi_h^{k_u}\mathbf{v}\|_{vel} \leq \sigma\,\hat{C}_\Pi \|\mathbf{v}\|_{\mathbf{H}_0^1(\Omega)}.$$

Proof. Let $\mathbf{v} \in \mathbf{H}_0^1(\Omega)$. Owing to the H^1-stability of the L^2-projector onto the broken polynomial space $\mathcal{P}_{disc}^k(\mathcal{T}_h)$, see remark 6.13 and inequality (4.41), there exists a constant C_Π such that we have:

$$\|\mathbf{Grad}_h \Pi_h^{k_u}\mathbf{v}\|^2_{\mathbb{L}^2(\Omega)} \leq (\sigma\, C_\Pi)^2\, \|\mathbf{v}\|^2_{\mathbf{H}_0^1(\Omega)}.$$

Moreover also see remark 6.13 and inequality (6.36), there exists a constant $C_{F\,\Pi}$ such that

$$\sum_{F\in\mathcal{F}_h} h_f^{-1}\|[\![\Pi_h^{k_u}\mathbf{v}]\!]\|^2_{\mathbf{L}^2(F_f)} = \sum_{F\in\mathcal{F}_h} h_f^{-1}\|[\![\Pi_h^{k_u}\mathbf{v} - \mathbf{v}]\!]\|^2_{\mathbf{L}^2(F_f)} \leq (\sigma\, C_{F\,\Pi})^2\, \|\mathbf{v}\|^2_{\mathbf{H}_0^1}.$$

The result follows from the definition of the $\|\cdot\|_{vel}$ norm, see Definition 6.12.

□

We can prove that the bilinear form $a_{O,h}(\,\cdot,\cdot\,)$ is continuous and satisfies an inf-sup condition:

Theorem 12.3 *Let $(k_{\mathbf{u}}, k_p) \in \mathbb{N}^\star \times \mathbb{N}$, with $k_{\mathbf{u}} \in \{k_p - 1,\, k_p,\, k_p + 1\}$. Then the bilinear form $a_{O,h}(\,\cdot,\cdot\,)$ satisfies the following inf-sup condition, $\forall (\mathbf{u}'_h, p'_h) \in \mathcal{X}_h\backslash(0,0)$:*

$$\sup_{(\mathbf{v}_h,q_h)\in\mathcal{X}_h\backslash(0,0)} \frac{a_{O,h}(\,(\mathbf{u}'_h, p'_h), (\mathbf{v}_h, q_h)\,)}{\|(\mathbf{v}_h, q_h)\|_{\mathcal{X}_h}} \geq \nu\, C_{O,disc}\, \|(\mathbf{u}'_h, p'_h)\|_{\mathcal{X}_h}, \tag{12.19}$$

$$\left|a_{O,h}(\,(\mathbf{u}'_h, p'_h), (\mathbf{v}_h, q_h)\,)\right| \leq C_{O,ct}\|(\mathbf{u}'_h, p'_h)\|_{\mathcal{X}_h}\|(\mathbf{v}_h, q_h)\|_{\mathcal{X}_h}. \tag{12.20}$$

where $C_{O,disc}$ and $C_{O,ct}$ are strictly positive constants independent of h.

Proof. We proceed as in the proof of Theorem 6.9.
Let $(\mathbf{u}'_h, p'_h) \in \mathcal{X}_h$, we will construct $(\mathbf{v}_h^\star, q_h^\star) \in \mathcal{X}_h$ satisfying (12.19). According to proposition 2.3, there exists $\tilde{\mathbf{v}}_{p'_h} \in \mathbf{H}_0^1(\Omega)$ such that $\operatorname{div}\tilde{\mathbf{v}}_{p'_h} = p'_h$ in the domain Ω and $\|\tilde{\mathbf{v}}_{p'_h}\|^2_{\mathbf{H}_0^1(\Omega)} \leq$

$C_{\rm div}\,\|p'_h\|^2_{L^2(\Omega)}$. Let $\mathbf{v}_{p'} = \nu^{-1}\,\tilde{\mathbf{v}}_{p'_h}$, so that $\operatorname{div}\mathbf{v}_{p'_h} = \nu^{-1}\,p'_h$ in Ω and

$$\|\mathbf{v}_{p'_h}\|_{\mathbf{H}^1_0(\Omega)} \leq \frac{C_{\rm div}}{\nu}\,\|p'_h\|_{L^2(\Omega)}. \tag{12.21}$$

Using Lemma 12.2, we obtain:

$$\|\Pi^{k_u}_h\mathbf{v}_{p'_h}\|_{vel} \leq \sigma\,C_{\rm div}\,C_\Pi\nu^{-1}\,\|p'_h\|_{L^2(\Omega)}. \tag{12.22}$$

Consider $T_{\tilde\gamma}(\mathbf{u}'_h, p'_h) := (\tilde\gamma\mathbf{u}'_h - \Pi^{k_u}_h\mathbf{v}_{p'_h},\, \tilde\gamma p'_h)$. From inequality (6.47), there exist strictly positive constants α, $\tilde\gamma$, and ζ such that:

$$a_{S,h}(\,(\mathbf{u}'_h, p'_h), T_{\tilde\gamma}(\mathbf{u}'_h, p'_h)\,) \geq \quad \nu\,C_{vel}\,\|\mathbf{u}'_h\|^2_{vel} + \nu^{-1}\,\left(C_\Omega\,\|p'_h\|^2_{L^2(\Omega)} + C_{\tilde s}\,|p'_h|^2_{\tilde s}\right), \tag{12.23}$$

where:

$$C_{vel} = \frac{\tilde\gamma}{2} - C_{bnd}\,\frac{\alpha}{2}, \quad C_\Omega = 1 - \frac{(\sigma\,C_\Pi\,C_{\rm div})^2}{2\,\alpha} - \frac{(\sigma\,C_{F\,\Pi}\,C_{\rm div})^2}{2\,\zeta}, \quad C_{\tilde s} = \tilde\gamma - \frac{\zeta}{2}.$$

For the advection term:

$$\tilde t_h(\mathbf{u}_h,\, \tilde\gamma\mathbf{u}_h - \Pi^{k_u}_h\mathbf{v}_{p'_h}) = -\tilde t_h(\mathbf{u}_h,\, \Pi^{k_u}_h\mathbf{v}_{p'_h})$$

We have $\tilde t_h(\mathbf{u}_h, \mathbf{v}_h) = \mathcal{I}_1 + \mathcal{I}_2 + \frac{1}{2}\mathcal{I}_3 + \mathcal{I}_4$. with:

$$\begin{aligned}
\mathcal{I}_1 &:= (\boldsymbol{\beta}_h \cdot \mathbf{Grad}_h\mathbf{u}_h, \Pi^{k_u}_h\mathbf{v}_{p',h})_{\mathbf{L}^2(\Omega)},\\
\mathcal{I}_2 &:= \sum_{\ell\in\mathcal{I}_K} (\operatorname{div}\boldsymbol{\beta}_h\,\mathbf{u}_h, \Pi^{k_u}_h\mathbf{v}_{p',h})_{\mathbf{L}^2(K_\ell)},\\
\mathcal{I}_3 &:= -\sum_{f\in\mathcal{I}_F} (\llbracket\boldsymbol{\beta}_h\rrbracket\cdot\mathbf{n}_f\,,\, \left\{\mathbf{u}_h\cdot\Pi^{k_u}_h\mathbf{v}_{p',h}\right\})_{L^2(F_f)},\\
\mathcal{I}_4 &:= -\sum_{f\in\mathcal{I}^i_F} (\,\{\boldsymbol{\beta}_h\}\cdot\mathbf{n}_f,\, \llbracket\mathbf{u}_h\rrbracket\cdot\left\{\Pi^{k_u}_h\mathbf{v}_{p',h}\right\})_{L^2(F_f)},
\end{aligned}$$

With the same methodology as in the proof of Theorem 10.3 to show the continuity of $t_h(\cdot,\,\cdot)$, and using inequalities (12.22) and (10.8), we obtain:

For the first term:

$$\begin{aligned} |\mathcal{I}_1| &\leq \|\boldsymbol{\beta}_h\|_{\mathbf{L}^4(\Omega)}\|\mathbf{Grad}\,\mathbf{u}_h\|_{\mathbb{L}^2(\Omega)}\|\Pi_h^{k_u}\mathbf{v}_{p'_h}\|_{\mathbf{L}^4(\Omega)} \\ &\leq \sigma_4\,\|\boldsymbol{\beta}_h\|_{\mathbf{L}^4(\Omega)}\|\mathbf{Grad}\,\mathbf{u}_h\|_{\mathbb{L}^2(\Omega)}\,\|\Pi_h^{k_u}\mathbf{v}_{p'_h}\|_{vel} \\ &\leq \sigma_4^2\,\sigma\,C_{\mathrm{div}}\,C_\Pi\nu^{-1}\,\|\boldsymbol{\beta}_h\|_{vel}\|\mathbf{u}_h\|_{vel}\,\|p'_h\|_{L^2(\Omega)} \end{aligned}$$

For the second term:

$$\begin{aligned} |\mathcal{I}_2| &\leq \|\mathbf{u}_h\|_{\mathbf{L}^4(\Omega)}\|\Pi_h^{k_u}\mathbf{v}_{p'_h}\|_{\mathbf{L}^4(\Omega)}\|\,\mathrm{div}_h\,\boldsymbol{\beta}_h\|_{L^2(\Omega)} \\ &\leq \sigma_4^2\,\sigma\,C_{\mathrm{div}}\,C_\Pi\nu^{-1}\,\|\mathbf{u}_h\|_{vel}\|p'_h\|_{L^2(\Omega)}\|\mathbf{Grad}_h\boldsymbol{\beta}_h\|_{\mathbb{L}^2(\Omega)} \\ &\leq \sigma_4^2\sigma\,C_{\mathrm{div}}\,C_\Pi\nu^{-1}\,\|\mathbf{u}_h\|_{vel}\|p'_h\|_{L^2(\Omega)}\|\boldsymbol{\beta}_h\|_{vel} \end{aligned}$$

For the third term,also using the Cauchy–Schwarz inequality, Theorem 10.2, and inequality (4.14):

$$\begin{aligned} |\mathcal{I}_3| &\leq \left(\sum_{\ell\in\mathcal{I}_K} h_\ell\|\mathbf{u}_h\|^4_{\mathbf{L}^4(\partial K_\ell)}\right)^{\frac14}\left(\sum_{\ell\in\mathcal{I}_K} h_\ell\|\Pi_h^{k_u}\mathbf{v}_{p'_h}\|^4_{\mathbf{L}^4(\partial K_\ell)}\right)^{\frac14}|\boldsymbol{\beta}_h|_J \\ &\leq (C_{d,k,4})^2\,\sigma^{\frac12}\,\sigma_4\|\mathbf{u}_h\|_{\mathbf{L}^4(\Omega)}\|\boldsymbol{\beta}_h\|_{vel}\|\Pi_h^{k_u}\mathbf{v}_{p'_h}\|_{vel} \\ &\leq (C_{d,k,4})^2\,\sigma^{\frac32}\,\sigma_4^2\,C_{\mathrm{div}}\,C_\Pi\nu^{-1}\,\|\boldsymbol{\beta}_h\|_{\mathbf{L}^4(\Omega)}\|\mathbf{u}_h\|_{vel}\|\,p'_h\|_{L^2(\Omega)} \\ &\leq (C_{d,k,4})^2\,\sigma^{\frac32}\,\sigma_4^2\,C_{\mathrm{div}}\,C_\Pi\nu^{-1}\,\|\boldsymbol{\beta}_h\|_{vel}\|\mathbf{u}_h\|_{vel}\|\,p'_h\|_{L^2(\Omega)} \end{aligned}$$

For the fourth term, using the Cauchy–Schwarz inequality, Theorem 10.2, and inequality (4.14):

$$\begin{aligned} |\mathcal{I}_4| &\leq \left(\sum_{\ell\in\mathcal{I}_K} h_\ell\|\boldsymbol{\beta}_h\|^4_{\mathbf{L}^4(\partial K_\ell)}\right)^{\frac14}\left(\sum_{\ell\in\mathcal{I}_K} h_f\|\Pi_h^{k_u}\mathbf{v}_{p'_h}\|^4_{\mathbf{L}^4(\partial K_\ell)}\right)^{\frac14}|\mathbf{u}_h|_J \\ &\leq (C_{d,k,4})^2\,\sigma^{\frac12}\,\sigma_4\|\boldsymbol{\beta}_h\|_{\mathbf{L}^4(\Omega)}\|\mathbf{u}_h\|_{vel}\|\Pi_h^{k_u}\mathbf{v}_{p'_h}\|_{vel} \\ &\leq (C_{d,k,4})^2\,\sigma^{\frac32}\,\sigma_4\,C_{\mathrm{div}}\,C_\Pi\nu^{-1}\,\|\boldsymbol{\beta}_h\|_{\mathbf{L}^4(\Omega)}\|\mathbf{u}_h\|_{vel}\|\,p'_h\|_{L^2(\Omega)} \\ &\leq (C_{d,k,4})^2\,\sigma^{\frac32}\,\sigma_4^2\,C_{\mathrm{div}}\,C_\Pi\nu^{-1}\,\|\boldsymbol{\beta}_h\|_{vel}\|\mathbf{u}_h\|_{vel}\|\,p'_h\|_{L^2(\Omega)} \end{aligned}$$

We sum the estimates of $\mathcal{I}_i$, for $i = 1,\ldots,4$, and obtain:

$$\begin{aligned} \tilde{t}_h(\mathbf{u}_h,\,\Pi_h^{k_u}\mathbf{v}_{p'_h}) &\leq \sigma\,C_{\mathrm{div}}\,C_\Pi\nu^{-1}\,\sigma_4^2\,\|\boldsymbol{\beta}_h\|_{vel}(1+(C_{d,k,4})^2\,\sigma^{\frac12})\,\|\mathbf{u}_h\|_{vel}\|\,p'_h\|_{L^2(\Omega)} \\ &\leq \sigma\,C_{\mathrm{div}}\,C_\Pi\nu^{-1}\,\sigma_4^2\,(1+(C_{d,k,4})^2\,\sigma^{\frac12})\,\left(\tfrac{\iota}{2}\|\boldsymbol{\beta}_h\|^2_{vel}\|\mathbf{u}\|^2_{vel}+\tfrac{1}{2\iota}\,\|p'\|^2_{L^2(\Omega)}\right) \end{aligned}$$

Finally, we obtain:

$$a_{O,h}(\,(\mathbf{u}'_h,p'_h),T(\mathbf{u}'_h,p'_h)\,)\;\geq\;C_{O,vel}\,\|\mathbf{u}'_h\|^2_{vel}+\nu^{-1}\,\left(C_{O,\Omega}\,\|p'_h\|^2_{L^2(\Omega)}+C_{\tilde{s}}\,|p'_h|^2_{\tilde{s}}\right)$$

Where $C_{O,vel} = \nu\, C_{vel} - \sigma\, C_{\rm div}\, C_\Pi \nu^{-1}\, \sigma_4^2\, (1 + (C_{d,k,4})^2\, \sigma^{\frac{1}{2}})\, \|\boldsymbol{\beta}_h\|_{vel}^2\, \frac{\iota}{2}$ and $C_{O,\Omega} = C_\Omega - \frac{1}{2\iota}\, \sigma\, C_{\rm div}\, C_\Pi \nu^{-1}\, \sigma_4^2\, (1 + (C_{d,k,4})^2\, \sigma^{\frac{1}{2}})$.
We need to choose the parameters α, $\tilde{\gamma}$, ζ and ι such that these coefficients are positive.
Let us take $\alpha = 4\, (\sigma\, C_\Pi\, C_{\rm div})^2$ and $\zeta = 4\, (\sigma\, C_{F\,\Pi}\, C_{\rm div})^2$. and $\iota = 4\, \sigma\, C_{\rm div}\, C_\Pi\, \nu^{-1} \sigma_4^2\, (1 + (C_{d,k,4})^2\, \sigma^{\frac{1}{2}})$. Then we have:

$$\begin{aligned} C_{O,vel} &= \nu\, \Big(\tfrac{\tilde{\gamma}}{2} - 2\, C_{bnd}\, (\sigma\, C_\Pi\, C_{\rm div})^2\Big) - 2\, (\sigma\, C_{\rm div}\, C_\Pi\, \nu^{-1} \sigma_4^2\, (1 + (C_{d,k,4})^2\, \sigma^{\frac{1}{2}}))^2\, \|\boldsymbol{\beta}_h\|_{vel}^2, \\ C_{O,\Omega} &= \tfrac{5}{8}, \quad C_{\tilde{s}} = \tilde{\gamma} - \tfrac{1}{2}\, (\sigma\, C_{\rm div})^2. \end{aligned}$$

We can choose $\tilde{\gamma} = 4\, (\sigma\, C_{\rm div})^2\, \max\Big(\, (C_\Pi)^2\, (C_{bnd} + \nu^{-1}\, (\nu^{-1} \sigma_4^2\, (1 + (C_{d,k,4})^2)^2\, \|\boldsymbol{\beta}_h\|_{vel}^2),\, (C_{F\,\Pi})^2\, 1\Big)$.
As $h \to 0$, we have $\|\boldsymbol{\beta}_h\|_{vel} \to \|\boldsymbol{\beta}\|_{vel}$. Therefore, the dependence of the coercivity constant on $\|\boldsymbol{\beta}_h\|_{vel}$ does not affect the stability of the numerical scheme.

□

12.4 Adaptation of the non-polynomial basis: how should we choose b?

Previously, for the advection–diffusion problem, we assumed that $\boldsymbol{\beta}$ was piecewise constant and took $\mathbf{b}_\ell = \boldsymbol{\beta}$. However, for the Oseen problem, $\boldsymbol{\beta}$ is a function that is not necessarily constant. We must therefore address the following question: how should we choose $\mathbf{b}_\ell$?
There are some intuitive choices, such as:

— $\mathbf{b}_\ell = \boldsymbol{\beta}(\mathbf{x}_\ell)$, where $\mathbf{x}_\ell$ is the barycenter of K_ℓ,

— $\mathbf{b}_\ell = \dfrac{1}{|K_\ell|} \displaystyle\int_{K_\ell} \boldsymbol{\beta}$.

Other, more complex, choices include:

— If we know that locally (or globally) β behaves like an exponential term that dominates other terms, we can use the following method: find C_1 and the two components of $\mathbf{b}_\ell$ by an interpolation of $\boldsymbol{\beta}$ at the vertices S_i, $i = 1, \ldots, 3$ of the triangle K_ℓ. In other words, component by component we write $C_1 \exp\Big(\dfrac{\mathbf{b} \cdot S_i}{\nu}\Big) = \boldsymbol{\beta}_j(S_i)$ for $i = 1, \ldots, 3$, where $\boldsymbol{\beta}_j$ is the j-th component of $\boldsymbol{\beta}$ with $j \in \{1, 2\}$. This will provide us with a value for $\mathbf{b}$ that we will use in the approximation space of $\mathbf{u}_j$ (the different components of $\mathbf{u}$ will thus have different exponential functions in their approximation bases). By applying the logarithm to the absolute value of this system and denoting $\tilde{C}_\ell = \log(|C_\ell|)$, we obtain

a system of three unknowns and three equations:

$$\begin{bmatrix} \frac{(S_1)_1}{\nu} & \frac{(S_1)_2}{\nu} & 1 \\ \frac{(S_2)_1}{\nu} & \frac{(S_2)_2}{\nu} & 1 \\ \frac{(S_3)_1}{\nu} & \frac{(S_3)_2}{\nu} & 1 \end{bmatrix} \begin{bmatrix} (\mathbf{b}_\ell)_1 \\ (\mathbf{b}_\ell)_2 \\ \tilde{C}_\ell \end{bmatrix} = \begin{bmatrix} \log(|\boldsymbol{\beta}(S_1)|) \\ \log(|\boldsymbol{\beta}(S_2)|) \\ \log(|\boldsymbol{\beta}(S_3)|) \end{bmatrix}.$$

— The fourth method, that we apply separately to each component of $\boldsymbol{\beta}$, is based on the fact that if $\beta_j(x, y) = \exp(\delta x + \gamma y)$, then a way to compute the coefficients δ and γ is to notice that $\frac{1}{\beta_j}(\nabla \beta_j) = (\delta, \gamma)^T$. Thus, if we denote by $\alpha_i := \beta_j(S_i)$, $i = 1, \ldots, 3$, and let $a\,x + c\,y + d$ be the interpolating function passing through the points (S_i, α_i), then by differentiating this interpolation, we may set

$$\mathbf{b}_\ell = \frac{(a,\, c)^T}{\beta_j(\mathbf{x}_\ell)},$$

which is a practical way to define the coefficient vector in the exponential basis of the velocity component if we know the function $\boldsymbol{\beta}$ or some approximation of it. Note that if $\beta_j(\mathbf{x}_\ell)$ is very small, we replace it by 10^{-3} or -10^{-3}, depending on the sign of $\beta_j(\mathbf{x}_\ell)$.

For the approximation space of pressure we may either choose a polynomial basis with a uniform choice of $\mathbf{b}$ of small amplitude, so that the exponential basis is close to the first-order polynomial basis, or select one of the two spaces adapted to $\boldsymbol{\beta}$. This method is very efficient when ν is not too small, see Chapter 14.

For all the methods we presented, numerically, if $\|\mathbf{b}_\ell\|$ is very small (close to zero), it is replaced by a small number ensuring that the ratio $\dfrac{\|\mathbf{b}_\ell\|\, h_\ell}{\nu} \approx 10^{-3}$ for example $\boldsymbol{\beta}_\ell = 10^{-3}\frac{\nu}{h_\ell}[1,\, 0]$. With this choice, the non-polynomial basis becomes very close to the polynomial basis.

In this chapter, we studied the Oseen problem and showed that it is well-posed using the T-coercivity method. We performed a discontinuous Galerkin discretization and also demonstrated that the discrete problem is well-posed using T-coercivity. Moreover, we proposed methods to adapt the choice of $\mathbf{b}$ in the non-polynomial basis for the Oseen problem.

Chapter 13

Navier-Stokes problem

Contents

In the previous chapters, we studied separately several model problems associated with the Navier–Stokes equations. The Stokes problem, presented in part I, allowed us to analyze the role of viscosity, the velocity–pressure coupling, as well as the incompressibility constraint $\operatorname{div} \mathbf{u} = 0$. The advection–diffusion problem, discussed in Chapter 10, served to isolate and understand the effects of transport and diffusion, while also providing a first insight into the treatment of the nonlinear convective term $(\mathbf{u} \cdot \mathbf{Grad})\mathbf{u}$. Finally, the Oseen problem, introduced in the previous Chapter 12, can be seen as a linearization of the full equations, including both convective and viscous terms, as well as velocity-pressure coupling.

The Navier–Stokes equations thus combine all these physical phenomena into a single, generally unsteady, framework.

13.1 Brief Literature Review

The Navier–Stokes equations describe the motion of an incompressible fluid over a time interval $(0, T_F)$, with $T_F > 0$ and form a nonlinear system coupling the conservation of

momentum with the incompressibility constraint:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \mathbf{Grad}\,)\mathbf{u} - \nu\Delta\mathbf{u} + \mathbf{grad}\, p = \mathbf{f}, \quad \operatorname{div} \mathbf{u} = 0, \quad \text{in } (0, T_F) \times \Omega.$$

From a theoretical standpoint, the Navier–Stokes equations have been extensively studied with respect to the existence, uniqueness, and regularity of solutions within suitable functional spaces. Classical references such as [73, 68, 57, 96] present both the strong and weak formulations of the problem, as well as the analytical framework required for the mathematical study of solutions.

From a numerical point of view, the discontinuous Galerkin finite element method has proven to be particularly well suited for the unsteady Navier–Stokes equations, especially in convection-dominated regimes. This approach allows each mesh element to be treated independently while maintaining stable numerical fluxes at interfaces—an essential feature for handling the nonlinear convective term $(\mathbf{u} \cdot \mathbf{Grad}\,)\mathbf{u}$. The work of [13] provide stable and consistent DG formulations for the Navier–Stokes equations, while references, such as [42], offer a comprehensive overview ranging from the theoretical formulation to numerical discretization, thereby providing a solid foundation for the development of accurate and reliable simulations.

13.2 Types of Boundary Conditions

Let $C^m([0, T_F]; \Omega)$ be the Banach space of m-continuously differentiable functions with values in Ω. Let also

$$\begin{aligned} \mathbf{H} &= \overline{\{\mathbf{v} \in C_c^\infty(\Omega)^d \mid \operatorname{div} \mathbf{v} = 0\}}^{\mathbf{L}^2(\Omega)} &&= \{\mathbf{v} \in \mathbf{L}^2(\Omega) \mid \operatorname{div} \mathbf{v} = 0, \gamma_1(\mathbf{v}) = 0\}, \\ \mathbf{W} &= \overline{\{\mathbf{v} \in C_c^\infty(\Omega)^d \mid \operatorname{div} \mathbf{v} = 0\}}^{\mathbf{H}_0^1(\Omega)} &&= \{\mathbf{v} \in \mathbf{H}_0^1(\Omega), \mid \operatorname{div} \mathbf{v} = 0\}. \end{aligned}$$

In what follows, we assume that $\mathbf{f} \in C^0([0, T_F]; \mathbf{L}^2(\Omega))$, and $\mathbf{u}_0 \in \mathbf{H}$.

Several types of boundary conditions can be considered. The simplest one for numerical

studies is the Dirichlet boundary condition. The Navier–Stokes problem then reads:

$$\left\{\begin{array}{l}\text{Find } (\mathbf{u},p) \text{ such that:}\\ \dfrac{\partial \mathbf{u}}{\partial t} - \nu\,\boldsymbol{\Delta}\,\mathbf{u} + (\mathbf{u}\cdot\mathbf{Grad}\,)\mathbf{u} + \mathbf{grad}\,p \;=\; \mathbf{f}, \quad \text{in } (0,T_F)\times\Omega,\\ \qquad\qquad \operatorname{div}\mathbf{u} \;=\; 0, \quad \text{in } (0,T_F)\times\Omega,\\ \qquad\qquad \mathbf{u}_{|\partial\Omega} \;=\; \mathbf{g}, \quad \text{in } (0,T_F),\\ \qquad\qquad \mathbf{u}(\mathbf{x},0) \;=\; \mathbf{u}_0, \quad \text{in } \Omega,\ t=0.\end{array}\right. \tag{13.1}$$

with a normalization condition for p: $\displaystyle\int_\Omega p = 0$, and where $\mathbf{g}$ satisfies $\displaystyle\int_{\partial\Omega}\mathbf{g}\cdot\mathbf{n} = 0$, $\forall t\in(0,\,T_F)$. In the case where the Dirichlet boundary conditions are homogeneous, according to [15, Thm. V.1.4], the solution of problem (13.1) satisfies

$$(\mathbf{u},p)\in\Big(L^\infty(0,T_F;\mathbf{H})\cap L^2(0,T_F;\mathbf{W})\Big)\times W^{-1,\infty}(0,T_F;L^2_{zmv}(\Omega)).$$

Let $\partial\Omega=\Gamma_1\cup\Gamma_2$, with $\Gamma_1\cap\Gamma_2=\emptyset$. Physically, the Navier–Stokes problem often involves both Dirichlet and Neumann boundary conditions, leading to:

$$\left\{\begin{array}{l}\text{Find } (\mathbf{u},p) \text{ such that:}\\ \dfrac{\partial \mathbf{u}}{\partial t} - \nu\,\boldsymbol{\Delta}\,\mathbf{u} + (\mathbf{u}\cdot\mathbf{Grad}\,)\mathbf{u} + \mathbf{grad}\,p \;=\; \mathbf{f}, \quad \text{in } (0,T_F)\times\Omega,\\ \qquad\qquad \operatorname{div}\mathbf{u} \;=\; 0, \quad \text{in } (0,T_F)\times\Omega,\\ \qquad\qquad \mathbf{u}_{|\Gamma_1} \;=\; \mathbf{g}, \quad \text{in } (0,T_F)\\ \qquad\qquad (\nu\mathbf{Grad}\,\mathbf{u}\cdot\mathbf{n} - p\,\mathbf{n})_{|\Gamma_2} \;=\; \mathbf{s}, \quad \text{in } (0,T_F),,\\ \qquad\qquad \mathbf{u}(\mathbf{x},0) \;=\; \mathbf{u}_0, \quad \text{in } \Omega,\ t=0.\end{array}\right. \tag{13.2}$$

13.3 Spatial Discretization of the Navier–Stokes problem

Consider now the stationary Navier–Stokes problem:

$$\left\{\begin{array}{l}\text{Find } (\mathbf{u},p) \text{ such that:}\\ -\nu\,\boldsymbol{\Delta}\,\mathbf{u} + (\mathbf{u}\cdot\mathbf{Grad}\,)\mathbf{u} + \mathbf{grad}\,p \;=\; \mathbf{f}, \quad \text{in } (0,T_F)\times\Omega,\\ \qquad\qquad \operatorname{div}\mathbf{u} \;=\; 0, \quad \text{in } (0,T_F)\times\Omega,\\ \qquad\qquad \mathbf{u}_{|\partial\Omega} \;=\; 0, \quad \text{in } (0,T_F)\times\Omega.\end{array}\right. \tag{13.3}$$

To solve the Navier–Stokes equations numerically, we proceed to the spatial discretization. Since the Stokes and Oseen problems have already been discretized, most of the necessary components are available. We now focus on the discretization of the advection term.

If we recall the bilinear form $\tilde{t}_h(\cdot,\,\cdot)$ discretized in the Oseen problem (12.14), and if we treat $\boldsymbol{\beta}_h$ as a variable, then $\tilde{t}_h(\cdot,\,\cdot)$ becomes a trilinear form corresponding to the discretization of the advection term. We therefore define the trilinear form $\bar{t}_h(\cdot,\,\cdot,\,\cdot)$ as follows:

$$\begin{aligned}\bar{t}_h(\mathbf{u}_h,\,\mathbf{v}_h,\,\mathbf{w}_h) \;:=\; & ((\mathbf{u}_h\cdot\mathbf{grad}_h)\mathbf{v}_h,\,\mathbf{w}_h)_{\mathbf{L}^2(\Omega)} + \tfrac{1}{2}\,(\mathrm{div}_h\,\mathbf{u}_h\,\mathbf{v}_h,\,\mathbf{w}_h)_{\mathbf{L}^2(\Omega)} \\ & -\tfrac{1}{2}\sum_{f\in\mathcal{I}_F}(\llbracket\mathbf{u}_h\rrbracket\cdot\mathbf{n}_f,\,\{\mathbf{v}_h\cdot\mathbf{w}_h\}\,)_{L^2(F_f)} - \sum_{f\in\mathcal{I}_F^i}(\,\{\mathbf{u}_h\}\cdot\mathbf{n}_f,\,\llbracket\mathbf{v}_h\rrbracket\cdot\{\mathbf{w}_h\}\,)_{L^2(F_f)}.\end{aligned} \tag{13.4}$$

Recall that the bilinear forms $a_h(\cdot,\cdot)$, $b_h(\cdot,\cdot)$, and $\tilde{s}_h(\cdot,\cdot)$ are already defined in equations (6.25), (6.26), and (6.27), respectively. The discrete variational formulation of the stationary Navier–Stokes problem (13.3) then reads:

$$\left\{\begin{array}{l}\text{Find } (\mathbf{u}_h,p_h)\in\mathbf{X}_h\times L_h \text{ such that:}\\ \begin{aligned}\nu\,a_h(\mathbf{u}_h,\mathbf{v}_h)+\bar{t}_h(\mathbf{u}_h,\,\mathbf{u}_h,\,\mathbf{v}_h)+b_h(\mathbf{v}_h,p_h) &= (\mathbf{f},\mathbf{v}_h)_{\mathbf{L}^2(\Omega)} && \forall\mathbf{v}_h\in\mathbf{X}_h,\\ -b_h(\mathbf{u}_h,q_h)+\lambda\,\tilde{s}_h(p_h,q_h) &= 0 && \forall q_h\in L_h.\end{aligned}\end{array}\right. \tag{13.5}$$

We then define the following form:

$$a_{NS,h}:\left\{\begin{array}{rcl}\mathcal{X}_h\times\mathcal{X}_h &\to& \mathbb{R},\\ ((\mathbf{u}_h',p_h'),(\mathbf{v}_h,q_h)) &\mapsto& a_{S,h}((\mathbf{u}_h',p_h'),(\mathbf{v}_h,q_h))+\bar{t}_h(\mathbf{u}_h',\,\mathbf{u}_h',\,\mathbf{v}_h).\end{array}\right. \tag{13.6}$$

Thus, problem (12.16) can be rewritten as:

$$\left\{\begin{array}{l}\text{Find } (\mathbf{u}_h,p_h)\in\mathcal{X}_h \text{ such that for all } (\mathbf{v}_h,q_h)\in\mathcal{X}_h,\\ a_{NS,h}((\mathbf{u}_h,p_h),(\mathbf{v}_h,q_h))=(\mathbf{f},\,\mathbf{v}_h)_{\mathbf{L}^2(\Omega)}.\end{array}\right. \tag{13.7}$$

Often, the stationary Navier–Stokes problem involves both Dirichlet and Neumann boundary

conditions, as in:

$$\left\{\begin{array}{rcll} \text{Find } (\mathbf{u},p) \text{ such that:} & & & \\ -\nu\,\boldsymbol{\Delta}\,\mathbf{u} + (\mathbf{u}\cdot\mathbf{Grad}\,)\mathbf{u} + \mathbf{grad}\,p &=& \mathbf{f}, & \text{in } \Omega, \\ \mathrm{div}\,\mathbf{u} &=& 0, & \text{in } \Omega, \\ \mathbf{u}_{|\Gamma_1} &=& \mathbf{g}, & \\ (\nu\mathbf{Grad}\,\mathbf{u}\cdot\mathbf{n} - p\mathbf{n})_{|\Gamma_2} &=& \mathbf{s}. & \end{array}\right. \tag{13.8}$$

In this situation, it is sufficient to remove the zero-mean pressure constraint and apply the following modifications to the variational formulation. Let $\mathcal{I}_F^D = \mathcal{I}_F^i \cup \mathcal{I}_F^{b,D}$, where for all $f \in \mathcal{I}_F^{b,D}$, $F_f \subset \Gamma_1$, and $\mathcal{I}_F^{b,N}$ where for all $f \in \mathcal{I}_F^{b,N}$, $F_f \subset \Gamma_2$.

$$\begin{aligned} a_h(\mathbf{v}_h,\mathbf{w}_h) &= (\mathbf{v}_h,\mathbf{w}_h)_h - \sum_{f\in\mathcal{I}_F^D} \big(\,\{\mathbf{Grad}_h\mathbf{v}_h\}\cdot\mathbf{n}_f,\,[\![\mathbf{w}_h]\!]\big)_{\mathbf{L}^2(F_f)} \\ &\quad - \sum_{f\in\mathcal{I}_F^D} \big(\,\{\mathbf{Grad}_h\mathbf{w}_h\}\cdot\mathbf{n}_f,\,[\![\mathbf{v}_h]\!]\big)_{\mathbf{L}^2(F_f)} + \sum_{f\in\mathcal{I}_F^D}\frac{\eta}{h_f}([\![\mathbf{v}_h]\!],[\![\mathbf{w}_h]\!])_{\mathbf{L}^2(F_f)}, \\ b_h(\mathbf{v}_h,q_h) &= -(\mathrm{div}_h\,\mathbf{v}_h,q_h)_{L^2(\Omega)} + \sum_{f\in\mathcal{I}_F^D}\big([\![\mathbf{v}_h]\!]\cdot\mathbf{n}_f,\;\{q_h\}\,\big)_{L^2(F_f)}, \\ \bar{t}_h(\mathbf{u}_h,\mathbf{v}_h,\mathbf{w}_h) &= ((\mathbf{u}_h\cdot\mathbf{grad}\,)\mathbf{v}_h,\,\mathbf{w}_h)_{\mathbf{L}^2(\Omega)} + \frac{1}{2}(\,\mathrm{div}\,\mathbf{u}_h\,\mathbf{v}_h,\,\mathbf{w}_h)_{\mathbf{L}^2(\Omega)} \\ &\quad - \frac{1}{2}\sum_{f\in\mathcal{I}_F^D}\big([\![\mathbf{u}_h]\!]\cdot\mathbf{n}_f,\;\{\mathbf{v}_h\cdot\mathbf{w}_h\}\,\big)_{L^2(F_f)} - \sum_{f\in\mathcal{I}_F^i}\big(\,\{\mathbf{u}_h\}\cdot\mathbf{n}_f,\,[\![\mathbf{v}_h]\!]\cdot\{\mathbf{w}_h\}\,\big)_{L^2(F_f)}. \end{aligned}$$

We then define the linear functional:

$$\begin{aligned} \ell_{NS}(\mathbf{v}_h) \;:=\; & \nu\sum_{f\in\mathcal{I}_F^{b,D}} -(\mathbf{g},\mathbf{Grad}\,\mathbf{v}_h\cdot\mathbf{n}_f)_{L^2(F_f)} + \frac{\eta}{h_F}(\mathbf{g},\mathbf{v}_h)_{\mathbf{L}^2(F_f)} \\ & -\tfrac{1}{2}\sum_{f\in\mathcal{I}_F^{b,D}}(\mathbf{g}\cdot\mathbf{n}_f,\,\mathbf{g}\cdot\mathbf{v}_h)_{L^2(F_f)} + \sum_{f\in\mathcal{I}_F^{b,N}}(\mathbf{v}_h,\,\mathbf{s})_{\mathbf{L}^2(F_f)}. \end{aligned}$$

Finally, the discrete problem becomes:

$$\begin{cases} \text{Find } (\mathbf{u}_h, p_h) \in \mathbf{X}_h \times X_h \text{ such that:} \\ \nu\, a_h(\mathbf{u}_h, \mathbf{v}_h) + \bar{t}_h(\mathbf{u}_h,\, \mathbf{u}_h,\, \mathbf{v}_h) + b_h(\mathbf{v}_h, p_h) &= (\mathbf{f}, \mathbf{v}_h)_{\mathbf{L}^2(\Omega)} + \ell_{NS}(\mathbf{v}_h) & \forall \mathbf{v}_h \in \mathbf{X}_h, \\ -b_h(\mathbf{u}_h, q_h) + \lambda\, \tilde{s}_h(p_h, q_h) &= \sum_{f \in \mathcal{I}_F^{b,D}} (\,\{q_h\}\,,\, [\![\mathbf{g}]\!] \cdot \mathbf{n}_f)_{L^2(F_f)} & \forall q_h \in X_h. \end{cases}$$

13.4 Temporal Discretization: Semi-Implicit Euler Scheme

The implicit Euler scheme evaluates the Navier–Stokes terms at the future time step $\mathbf{u}_h^{n+1}$, according to

$$\mathbf{u}_h^{n+1} = \mathbf{u}_h^n + \delta t\, F(\mathbf{u}_h^n,\, \mathbf{u}_h^{n+1}, p_h^{n+1}), \quad \operatorname{div}_h \mathbf{u}_h^{n+1} = 0.$$

This approach is unconditionally stable for the diffusive and linearized terms, allowing larger time steps without compromising stability. However, it requires solving a linear or nonlinear system at each time step (according to whether the convective term is treated semi-implicitly or fully implicitly), which increases the computational cost.
In the numerical results, we used a semi-implicit Euler scheme.
We define the discretization space $\mathcal{X}_h^n = \mathbf{X}_h^n \times L_h^n$. If we use a polynomial basis, then $\mathcal{X}_h^n = \mathcal{X}_h$. If we use a non-polynomial basis, it can be adapted at each time step to compute the solution at time step $n+1$. The solution at time step n can be used to adapt the non-polynomial basis more precisely, in particular for the choice of $\mathbf{b}$, which will be denoted by $\mathbf{b}_n$. Thus, we define $\mathbb{NP}^k_{\mathrm{disc},b_n}(\mathcal{T}_h)$ as the space $\mathbb{NP}^k_{\mathrm{disc}}(\mathcal{T}_h)$ generated by $\mathbf{b}_n$, and we set

$$X_h^n = \mathbb{NP}^k_{\mathrm{disc},b_n}(\mathcal{T}_h), \quad L_h^n = \mathbb{NP}^k_{\mathrm{disc},b_n}(\mathcal{T}_h) \cap L^2_{\mathrm{zmv}}(\Omega).$$

The space-time variational formulation is:

$$\begin{cases} \text{Find } (\mathbf{u}_h^{n+1}, p_h^{n+1}) \in \mathcal{X}_h^{n+1} \text{ such that } \forall (\mathbf{v}_h^{n+1},\, q_h^{n+1}) \in \mathcal{X}_h^{n+1} : \\ \qquad \dfrac{1}{\delta t}\,(\mathbf{u}_h^{n+1} - \mathbf{u}_h^n, \mathbf{v}_h^{n+1})_{\mathbf{L}^2(\Omega)} + \nu\, a_h(\mathbf{u}_h^{n+1}, \mathbf{v}_h^{n+1}) \\ \qquad\qquad + \bar{t}_h(\mathbf{u}_h^n,\, \mathbf{u}_h^{n+1},\, \mathbf{v}_h^{n+1}) + b_h(\mathbf{v}_h^{n+1}, p_h^{n+1}) = (\mathbf{f}^{n+1}, \mathbf{v}_h^{n+1})_{\mathbf{L}^2(\Omega)}, \\ \qquad\qquad - b_h(\mathbf{u}_h^{n+1}, q_h^{n+1}) + \lambda\, \tilde{s}_h(p_h^{n+1}, q_h^{n+1}) = 0. \end{cases} \tag{13.9}$$

We will rewrite the problem in matrix form with Dirichlet boundary conditions, applying

the same methodology for other types of boundary conditions according to the variational formulation.

We recall that $d \in \{2,3\}$ is the dimension and $\mathcal{I}_d = \{x,y\}$ for $d=2$, $\mathcal{I}_d = \{x,y,z\}$ for $d=3$. Let (O,x,y) for $d=2$ and (O,x,y,z) for $d=3$ be the Cartesian coordinates system, of orthonormal basis $(\mathbf{e}_{\tilde{d}})_{\tilde{d}\in\mathcal{I}_d}$. Denote by $N_{\mathbf{u}} = N_{k_{\mathbf{u}}} \times N_K$ (resp. $N_p = N_{k_p} \times N_K$) the number of degrees of freedom of $V_{disc}^{k_{\mathbf{u}},n+1}(\mathcal{T}_h)$ (resp. $V_{disc}^{k_p,n+1}(\mathcal{T}_h)$). Let us call $(\phi_{\tilde{d},i,\ell}^{n+1})_{i\in\mathcal{I}_k}$ a basis of $P_k^{n+1}(K_\ell)$. This basis can be either polynomial or non-polynomial. If it is a classical polynomial basis, then it does not change with time. In contrast, if it is non-polynomial, it depends on $\mathbf{b}$, which may vary from one time step to another and $A_{\tilde{d}}^{n+1} \in \mathbb{R}^{N_{\mathbf{u}}} \times \mathbb{R}^{N_{\mathbf{u}}}$ the matrix s.t.:

$$\forall (i,\ell),\ (j,\ell') \in \mathcal{I}_{k_{\mathbf{u}}} \times \mathcal{I}_K, \quad (A_{\tilde{d}}^{n+1})_{(i,\ell),(j,\ell')} := a_h^{sip}(\phi_{\tilde{d},i,\ell}^{n+1}, \phi_{\tilde{d},j,\ell'}^{n+1}), \tag{13.10}$$

for $\tilde{d} \in \mathcal{I}_d$. Let $\mathbb{A}^{n+1} \in \mathbb{R}^{d\times N_{\mathbf{u}}} \times \mathbb{R}^{d\times N_{\mathbf{u}}}$ be the block diagonal matrix s.t.: $\mathbb{A}^{n+1} = (\delta_{\tilde{d},\tilde{d}'} A_{\tilde{d}}^{n+1})_{\tilde{d},\tilde{d}'\in\mathcal{I}_d}$.

Simply, if $d=2$: $\mathbb{A}^{n+1} = \begin{bmatrix} A_1^{n+1} & 0 \\ 0 & A_2^{n+1} \end{bmatrix}$, if $d=3$, $\mathbb{A}^{n+1} = \begin{bmatrix} A_1^{n+1} & 0 & 0 \\ 0 & A_2^{n+1} & 0 \\ 0 & 0 & A_3^{n+1} \end{bmatrix}$.

For $\tilde{d} \in \mathcal{I}_d$, we call $B_{\tilde{d}}^{n+1} \in \mathbb{R}^{N_p} \times \mathbb{R}^{N_{\mathbf{u}}}$ the matrix s.t. for all $(i,\ell) \in \mathcal{I}_{k_p} \times \mathcal{I}_K$, $(j,\ell') \in \mathcal{I}_{k_{\mathbf{u}}} \times \mathcal{I}_K$,

$$(B_{\tilde{d}}^{n+1})_{(i,\ell),(j,\ell')} := b_h(\phi_{\tilde{d},j,\ell'}^{n+1} \mathbf{e}_{\tilde{d}}, \psi_{i,\ell}^{n+1})$$

and

$$\mathbb{B}^{n+1} = (B_{\tilde{d}}^{n+1})_{\tilde{d}\in\mathcal{I}_d} \in \mathbb{R}^{N_p} \times \mathbb{R}^{d\times N_{\mathbf{u}}}.$$

Let $M_p^{n+1} \in \mathbb{R}^{N_p} \times \mathbb{R}^{N_p}$ and $S_p^{n+1} \in \mathbb{R}^{N_p} \times \mathbb{R}^{N_p}$ the matrices s.t. for all (i,ℓ), $(j,\ell') \in \mathcal{I}_{k_p} \times \mathcal{I}_K$,

$$(M_p^{n+1})_{(i,\ell),(j,\ell')} = \delta_{\ell,\ell'}(\psi_{i,\ell}^{n+1}, \psi_{j,\ell'}^{n+1})_{L^2(K_\ell)}, \quad \text{and} \quad (S_p^{n+1})_{(i,\ell),(j,\ell')} = s_h(\psi_{i,\ell}^{n+1}, \psi_{j,\ell'}^{n+1}).$$

Notice that $A_{\tilde{d}}^{n+1}$, $B_{\tilde{d}}^{n+1}$ and S_p^{n+1} are sparse: for $\ell, \ell' \in \mathcal{I}_K$ s.t. $\partial K_\ell \cap \partial K_{\ell'} = \emptyset$, $(A_{\tilde{d}}^{n+1})_{(i,\ell),(j,\ell')} = 0$, $(B_{\tilde{d}}^{n+1})_{(i,\ell),(j,\ell')} = 0$ and $(S_p^{n+1})_{(i,\ell),(j,\ell')} = 0$. The matrix $A_{\tilde{d}}^{n+1}$ is symmetric, positive definite. When $k_{\mathbf{u}} > k_p$, matrix $\mathbb{B}^{n+1}$ is of rank $N_p - 1$. Notice that matrix S_p^{n+1} is also of rank $N_p - 1$. When $k_{\mathbf{u}} \leq k_p$, matrix $\mathbb{B}^{n+1}$ is of rank $< N_p - 1$. Let $T_{\tilde{d}}^n \in \mathbb{R}^{N_{\mathbf{u}}} \times \mathbb{R}^{N_{\mathbf{u}}}$ the matrix s.t.:

$$\forall (i,\ell),\ (j,\ell') \in \mathcal{I}_{k_{\mathbf{u}}} \times \mathcal{I}_K, \quad (T_{\tilde{d}}^n)_{(i,\ell),(j,\ell')} := \bar{t}_h(\mathbf{u}_h^n, \phi_{\tilde{d},i,\ell}^{n+1}, \phi_{\tilde{d},j,\ell'}^{n+1}).$$

Let $\mathbb{T}^n \in \mathbb{R}^{d\times N_{\mathbf{u}}} \times \mathbb{R}^{d\times N_{\mathbf{u}}}$ be the block diagonal matrix s.t.: $\mathbb{T}^{n+1} = (\delta_{\tilde{d},\tilde{d}'} T_{\tilde{d}}^{n+1})_{\tilde{d},\tilde{d}'\in\mathcal{I}_d}$.
The discrete Navier-Stokes problem reads: at each time step, Find $(U_h^{n+1}, P_h^{n+1}) \in \mathbb{R}^{N_D\times N_{\mathbf{u}}} \times \mathbb{R}^{N_p}$ s.t.:

$$\left\{ \begin{array}{rcl} \nu\mathbb{A}^{n+1}U_h^{n+1} + \mathbb{T}^{n+1}U^{n+1} + (\mathbb{B}^{n+1})^T P_h^{n+1} &=& F_{\mathbf{u}}^{n+1} \\ -\mathbb{B}^{n+1}U_h^{n+1} + \beta\, S_p^{n+1}\, P_h^{n+1} &=& F_p^{n+1} \end{array} \right. \quad \text{and} \quad \sum_{\ell\in\mathcal{I}_K} \sum_{i\in\mathcal{I}_{k_p}} p_{i,\ell} \int_{K_\ell} \phi_{i,\ell} = 0, \tag{13.11}$$

where $F_p \in \mathbb{R}^{N_p}$ vanishes when considering homogeneous Dirichlet boundary conditions. Finally the matrix form of the problem reads

$$\begin{bmatrix} \nu\mathbb{A}^{n+1} + \mathbb{T}^{n+1} & (\mathbb{B}^{n+1})^T \\ -\mathbb{B}^{n+1} & \beta S_p^{n+1} \end{bmatrix} \begin{bmatrix} U_h^{n+1} \\ P_h^{n+1} \end{bmatrix} = \begin{bmatrix} F_{\mathbf{u}}^{n+1} \\ F_p^{n+1} \end{bmatrix} \tag{13.12}$$

In our code, we solve this problem using the classical GMRES method.

In this chapter, we discretized the Navier–Stokes problem using the semi-implicit scheme in the polynomial basis and the exponential basis, noting that the approximation space for the non-polynomial basis can be changed with the time step, since at each time step we can modify the parameter **b**.

Chapter 14

Numerical Illustrations for Exponential Basis

Contents

In this chapter, we present the numerical results for the non-polynomial basis and compare them with those obtained using the polynomial basis in order to evaluate its efficiency. We have shown that the non-polynomial basis is more effective in terms of accuracy and in reducing oscillations. Even on a coarse mesh, the non-polynomial basis does not produce oscillations, whereas the polynomial basis does, as seen for the advection–diffusion problem in Section 14.1 and Subsection 14.2. For the Oseen problem in Subsection 14.2.2, we also observed that the non-polynomial basis outperforms the polynomial one, although the results could be further improved by using a more adaptive choice of $\mathbf{b}$. For the Navier–Stokes problem, we have not yet performed a sufficiently extensive numerical study, but we have included preliminary results. It would be preferable to carry out additional experiments to more thoroughly assess the performance of the non-polynomial basis.

14.1 Comparison of numerical results between polynomial and non-polynomial bases in one dimension

Consider here the one-dimensional case where $\Omega = (0,\, 1)$ and the meshes $(\mathcal{T}_h)_h$ are made of segments. We compare the numerical results obtained using the polynomial basis and the non-polynomial (exponential) basis for the advection–diffusion problem (10.1).
The physical test case is defined as

$$f(x) = \begin{cases} 0 & \text{if } x < \frac{1}{2}, \\ 1 & \text{if } x \geq \frac{1}{2}, \end{cases} \qquad u(0) = 1,\ u(1) = 0.$$

The analytical solution is given by

$$u(x) = \begin{cases} C_1(-1 + e^{\frac{\beta x}{\nu}}) + 1, & x < \frac{1}{2}, \\ C_2(e^{\frac{\beta x}{\nu}} - e^{\frac{\beta}{\nu}}) - \frac{1}{\beta} + \frac{x}{\beta}, & x \geq \frac{1}{2}, \end{cases}$$

with

$$\begin{aligned} C_1 &= \nu \tfrac{1}{\beta} e^{-\frac{\beta x}{2\nu}} \left(C_2 \left(\tfrac{\beta}{\nu} e^{\frac{\beta x}{2\nu}} + \tfrac{1}{\beta} \right) \right), \\ C_2 &= \frac{(e^{\frac{\beta}{2\nu}} - 1)}{-1 + e^{\frac{\beta}{\nu}}} \left(-\nu \tfrac{1}{\beta^2} e^{-\frac{\beta}{2\nu}} - \frac{1 + \frac{1}{2\beta}}{-1 + e^{\frac{\beta}{2\nu}}} \right). \end{aligned}$$

Instead of testing this case for several values of ν and β separately, we consider the non-polynomial basis defined in (11.3) with $b_\ell = \beta$ for all $\ell \in \mathcal{I}_K$, and study different values of the ratio $\dfrac{b_\ell\, h}{\nu}$. This ratio is independent of ℓ, since $b_\ell = \beta$ for all $\ell \in \mathcal{I}_K$. We fix $\nu = 10^{-1}$. and take several values of β. The first example is for $\beta = 1$. Simulations were performed on five meshes, from the coarsest one with $h = \dfrac{1}{20}$ to the finest with $h = \dfrac{1}{320}$.

14.1. Comparison of numerical results between polynomial and non-polynomial bases in one dimension

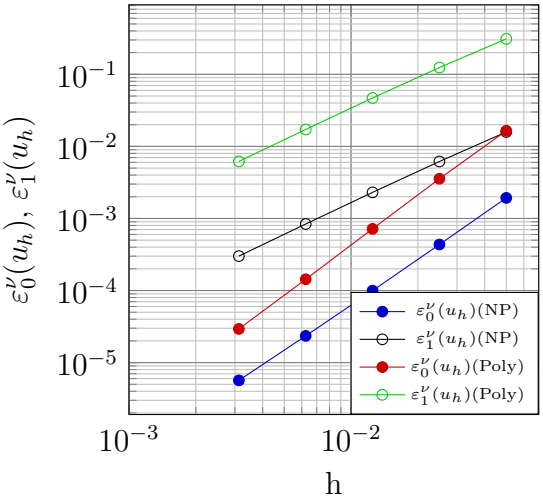


Figure 14.1 – Plots of $\varepsilon_0^\nu(u_h)$ and $\varepsilon_1^\nu(u_h)$ against the mesh size.

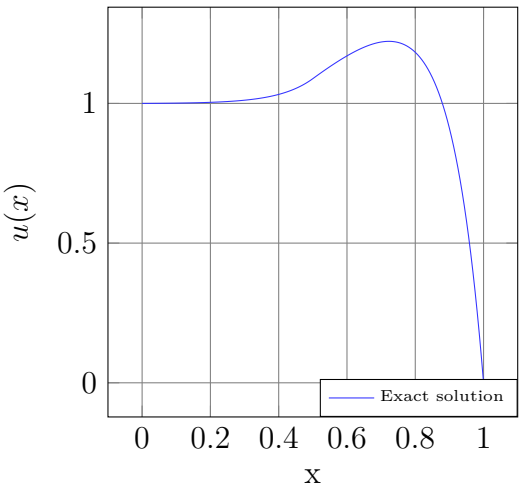


Figure 14.2 – Plot of exact solution $u(x)$, with $\beta = 1$ and $\nu = 10^{-1}$.

Basis	τ	mesh 1-2	mesh 2-3	mesh 3-4	mesh 4-5
Polynomial basis	τ_u	2.21	2.3	2.32	2.29
	$\tau_{\mathbf{grad}\, u}$	1.31	1.4	1.45	1.47
Non-polynomial basis	τ_u	2.15	2.13	2.11	2.1
	$\tau_{\mathbf{grad}\, u}$	1.36	1.41	1.46	1.48

Table 14.1 – Convergence rates computed with polynomial basis and non polynomial basis.

For $\beta = 40$, in this test case, to ensure numerical stability for the polynomial basis, the simulation was performed on 6 meshes, starting from the first mesh with $h = \frac{1}{20}$ up to the finest mesh with $h = \frac{1}{640}$.

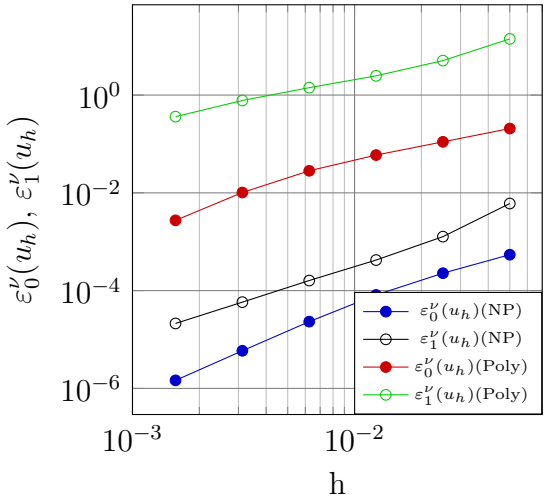


Figure 14.3 – Plots of $\varepsilon_0^\nu(u_h)$ and $\varepsilon_1^\nu(u_h)$ against the mesh size.

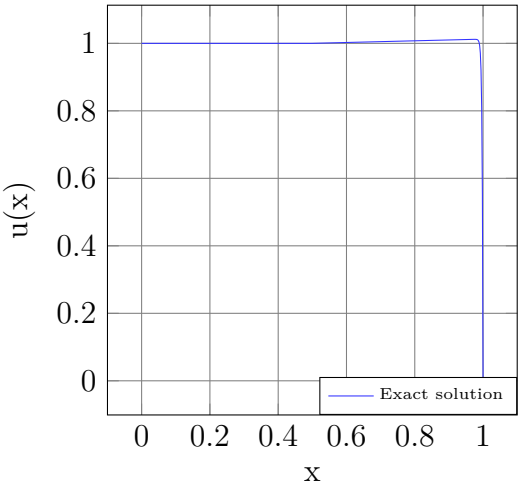


Figure 14.4 – Plot of exact solution $u(x)$, with $\beta = 40$ and $\nu = 10^{-1}$.

Base	τ	mesh 1-2	mesh 2-3	mesh 3-4	mesh 4-5	mesh 5-6
Polynomial base	τ_u	0.9	0.9	1.07	1.48	1.9
	$\tau_{\mathbf{grad}\, u}$	1.47	1.04	0.8	0.87	1.1
Non-polynomial base	τ_u	1.25	1.48	1.81	2	2.01
	$\tau_{\mathbf{grad}\, u}$	2.25	1.6	1.4	1.47	1.45

Table 14.2 – Convergence rates computed between the polynomial and non-polynomial bases.

We now study the advection–diffusion problem on another academic test case. We take the exact solution $u(x) = x(x-1)$, and the source term depends on the choice of β. In this test case, we consider a non-constant β in order to understand how the choice of $(b_\ell)_\ell$ should be adapted with respect to β. Let $\beta = a \sin\left(\frac{x}{N}\right)$ and $\nu = 10^{-1}$. In what follows, we choose $a = 5$ and $N = 16$. We consider several choices for $(b_\ell)_\ell$. For all $\ell \in \mathcal{I}_K$, we test the following natural choices for $(b_\ell)_\ell$:

$$\begin{cases} b_\ell &= \dfrac{1}{h}\displaystyle\int_{K_\ell} \beta \quad (1), \\ b_\ell &= \beta(x_\ell) \quad (2), \\ b_\ell &= 10^{-3} \quad (3). \end{cases} \tag{14.1}$$

We considered the last choice $b_\ell = 10^{-3}$ since we know that the exact solution is a quadratic polynomial, and when the ratio $\dfrac{b_\ell\, h}{\nu}$ is very small, the non-polynomial basis tends to a

polynomial one.

We now propose a fourth method based on the analytical solution. This serves as an introduction to how one can adapt $(b_\ell)_\ell$ for the unsteady Navier–Stokes problem, since in the time-dependent problem, the solution at the previous time step $n-1$ is known.

Let $u(x)$ be the analytical solution of the advection–diffusion problem; we aim to approximate it locally in this form: there exist b_ℓ, C_1, and C_2 such that

$$u(x) \approx C_1 + C_2 \exp\left(\frac{b_\ell(x - x_\ell)}{\nu}\right).$$

We have three unknowns, thus we need three equations. Let us assume that $K_\ell = [x_\ell^1, x_\ell^2 = x_\ell^1 + h]$. Then:

$$\begin{aligned} u(x_\ell^1) &= C_1 + C_2 \exp\left(\frac{b_\ell(-h)}{2\nu}\right), \\ u(x_\ell^2) &= C_1 + C_2 \exp\left(\frac{b_\ell(h)}{2\nu}\right), \\ u'(x_\ell) &= \frac{b_\ell}{\nu} C_2. \end{aligned} \tag{14.2}$$

By combining these three equations, we obtain

$$u(x_\ell^2) - u(x_\ell^1) = \frac{\nu\, u'(x_\ell)}{b_\ell}\left(\exp\left(\frac{b_\ell\, h}{2\nu}\right) - \exp\left(-\frac{b_\ell\, h}{2\nu}\right)\right).$$

Assuming that h is sufficiently small, by using a third-order Taylor expansion we obtain:

$$|b_\ell| = \frac{\nu}{h}\sqrt{24\left|\frac{u(x_\ell^2) - u(x_\ell^1)}{h\, u'(x_\ell)} - 1\right|}. \tag{14.3}$$

To correctly choose the sign of b_ℓ, we observe from the third equality in (14.2), that $u'(x_\ell)$ and $C_2\, b_\ell$ have the same sign. Therefore, we assume that $u'(x_\ell)$ and b_ℓ share the same sign. Thus, we define λ as follows: $\lambda = \begin{cases} 1 & \text{if } u'(x_\ell) \geq 0 \\ -1 & \text{if } u'(x_\ell) < 0 \end{cases}$, and $b_\ell = \lambda |b_\ell|$.

Numerically, if $u'(x_\ell)$ is very small, we replace it by 10^{-3} or -10^{-3} according to its sign. The value of b_ℓ behaves like $\nu\sqrt{\left|\frac{u'''(x_\ell)}{u'(x_\ell)}\right|}$. In what follows, we fix $a = 5$, and in figure 14.5, we compare the numerical results obtained with the non-polynomial basis for several choices of $(b_\ell)_\ell$ see (14.1). The index in parentheses indicates the choice of $(b_\ell)_\ell$; for example, $\varepsilon_0^\nu(u_h)(1)$ denotes the error $\varepsilon_0^\nu(u_h)$ associated with the first choice of $(b_\ell)_\ell$. We can see that the third and fourth choices for b_ℓ yield a more precise result compared to the first and second choices,

which are almost the same.

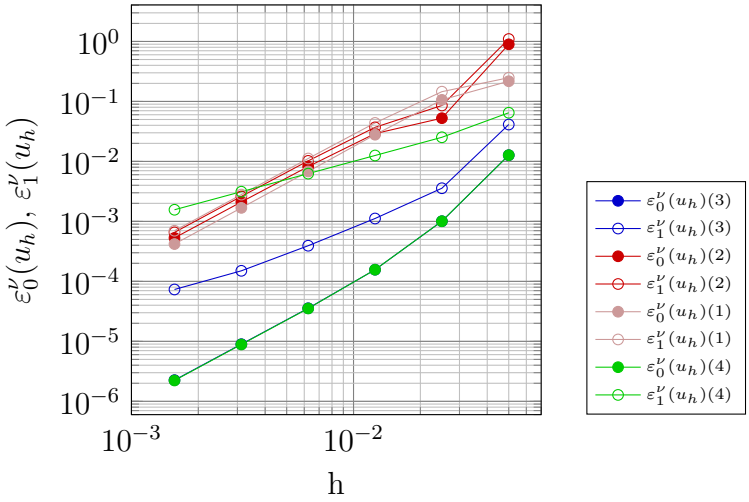


Figure 14.5 – Plots of $\varepsilon_0^\nu(u_h)$ and $\varepsilon_1^\nu(u_h)$ against the mesh size.

In this figure (14.6), we compare the polynomial basis with the non-polynomial basis using the forth choice of $(b_\ell)_\ell$.

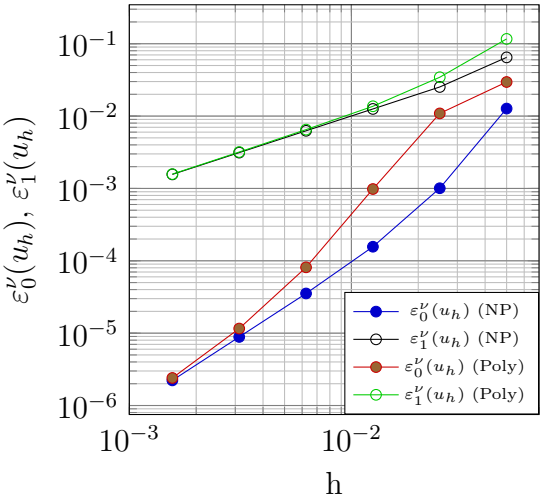


Figure 14.6 – Plots of $\varepsilon_0^\nu(u_h)$ and $\varepsilon_1^\nu(u_h)$ against the mesh size for polynomial and non-polynomial basis.

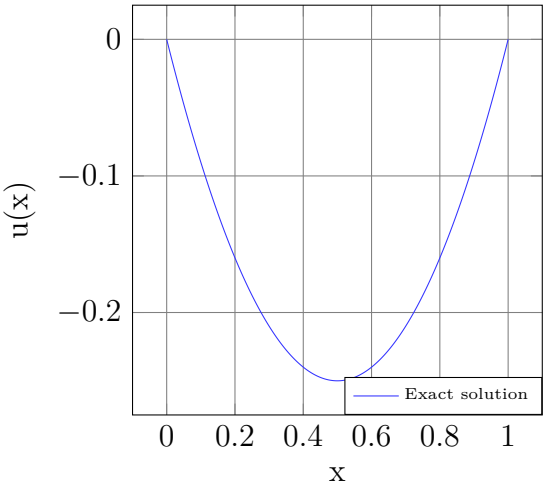


Figure 14.7 – Exact solution $u(x)$.

Basis	τ	mesh 1-2	mesh 2-3	mesh 3-4	mesh 4-5	mesh 5-6
Polynomial basis	τ_u	1.44	3.47	3.6	2.81	2.61
	$\tau_{\mathbf{grad}\, u}$	2.15	2.1	1.6	1.53	1.51
Non polynomial basis	τ_u	3.67	2.67	2.15	2	2
	$\tau_{\mathbf{grad}\, u}$	3.53	1.67	1.53	1.52	1.51

Table 14.3 – Computed convergence rates with polynomial basis and non-polynomial basis.

To evaluate the efficiency of the proposed methods and to observe the impact of the choice of $(b_\ell)_\ell$ on the numerical results, we construct a test case with the manufactured analytical solution $u(x) = \cos(\frac{2\pi x}{N})$. This function exhibits N oscillations over the domain $[0, 1]$. We set $\beta = 1$. For the non-polynomial basis, we test it with four different configurations shown in Figure 14.8 and we compare the polynomial and non-polynomial bases. We set $N = 10$ and $\nu = 10^{-1}$. In this figure, we compare the numerical results computed using the non-polynomial basis for several choices of $(b_\ell)_\ell$. One of the methods is to choose $(b_\ell)_\ell = \boldsymbol{\beta}$, and the other three (denoted by (2), (3) and (4)) correspond to the methods proposed in (14.1) applied to $u(x)$ instead of β for choices (2) and (3), and to (14.3) for choice (4).

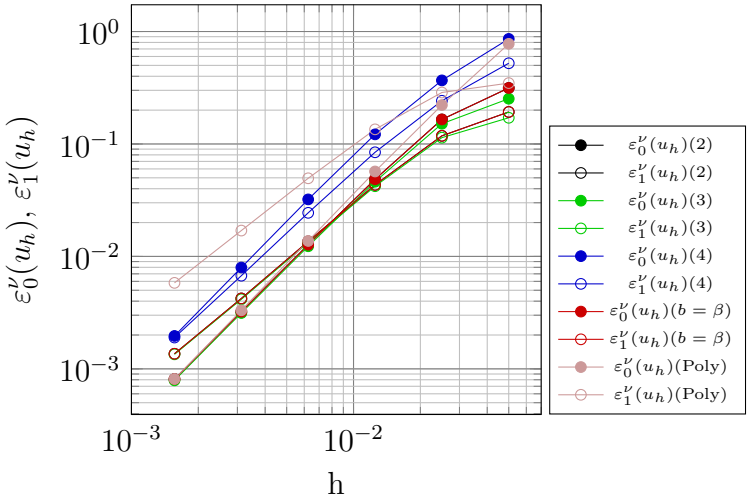


Figure 14.8 – Plots of $\varepsilon_0^\nu(u_h)$ and $\varepsilon_1^\nu(u_h)$ against the mesh size.

We can observe that the fourth choice of $(b_\ell)_\ell$ loses its efficiency compared to the other choices, while the remaining choices of $(b_\ell)_\ell$ yield more efficient results than those obtained with the polynomial basis.

14.2 Comparison of the numerical results between the polynomial and non-polynomial bases in 2D

14.2.1 Advection-diffusion problem

For the two-dimensional advection–diffusion problem, we considered both an academic and a physical test case to assess the validity of the non-polynomial basis compared to the polynomial one.

Academic test case

Let $\Omega = [0,1] \times [0,1]$ and the flow velocity $\boldsymbol{\beta} = (\beta_1, \beta_2) = (\cos\theta, \sin\theta)$ for $0 \leq \theta \leq \pi$, $f = 0$.

Let the exact solution be given by: $u(x,y) = \begin{cases} \frac{e^{\frac{x}{\nu}}-1}{e^{\frac{1}{\nu}}-1} + y, & \text{if } \theta = 0, \\ \frac{e^{\frac{\beta_1 x}{\nu}}-1}{e^{\frac{\beta_1}{\nu}}-1} + \frac{e^{\frac{\beta_2 y}{\nu}}-1}{e^{\frac{\beta_2}{\nu}}-1}, & \text{if } 0 < \theta < \pi, \\ \frac{e^{\frac{y}{\nu}}-1}{e^{\frac{1}{\nu}}-1} + x, & \text{if } \theta = \pi. \end{cases}$

With the exponential basis at $\theta = 0$ and π, the error is at machine precision for any ν. To

test the efficiency of the non-polynomial basis, we randomly chose $\theta = \frac{\pi}{3}$ and $\nu = 10^{-2}$, In Figures 14.9 and 14.10, we plot $\epsilon_0^\nu(\mathbf{u}_h)$ and $\epsilon_1^\nu(\mathbf{u}_h)$ using both the polynomial and the non-polynomial bases:

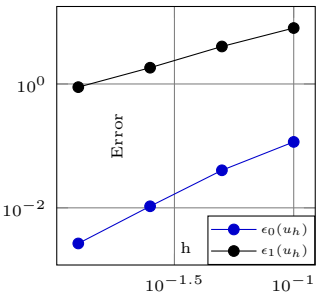


Figure 14.9 – Exponential basis

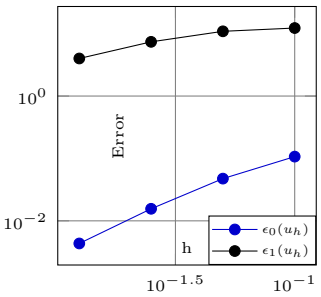


Figure 14.10 – Polynomial basis

In this simulation, the polynomial and non-polynomial bases have the same number of degrees of freedom. In the figures below 14.11, 14.12 and 14.13 we set $\nu = 10^{-6}$, which corresponds to the case of very low viscosity. It can be clearly observed that with the polynomial basis, oscillations appear throughout the domain, especially near the boundary layer. In contrast, with the non-polynomial basis, no oscillations are observed, and the resulting solution closely matches the shape of the exact one.

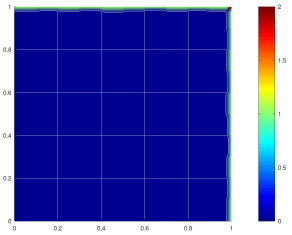


Figure 14.11 – Exact Solution

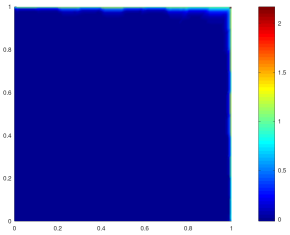


Figure 14.12 – Exponential basis

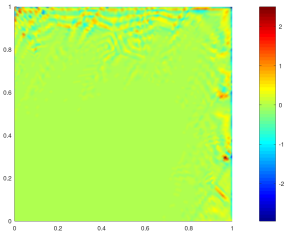


Figure 14.13 – Polynomial basis

Physical test case

In the framework of a physical test case, we consider a classical channel flow with an obstacle, in order to better capture the variation of the solution near the boundary layer.
The channel-shaped domain is defined as $\Omega = [0,1] \times [0,\frac{1}{2}] \setminus [0.425, 0.575] \times [0.175, 0.325]$, with $\Gamma_1 = \{0\} \times [0, \frac{1}{2}]$ and $\Gamma_2 = \partial\Omega \setminus \Gamma_1$. The convection field is set to $\boldsymbol{\beta} = (1,0)^T$.

The advection–diffusion problem is given by:

$$\text{Find } u \in H^1(\Omega) \text{ such that } \begin{cases} -\nu\,\Delta u + (\boldsymbol{\beta} \cdot \mathbf{grad}\,)u = 0, \\ u_{|\Gamma_1} = y(\frac{1}{2} - y), \\ u_{|\Gamma_2} = 0. \end{cases} \tag{14.4}$$

For this problem, two boundary layers appear: one near Γ_2 and another close to $\{1\} \times [0,\, \frac{1}{2}]$. To avoid the formation of two boundary layers, an open (outflow) boundary condition can be imposed on $\{1\} \times [0,\, \frac{1}{2}]$.
A uniform mesh is used for the discretization.

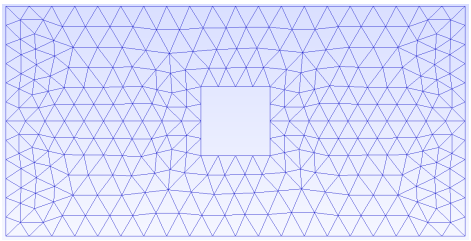


Figure 14.14 – Mesh for the channel.

In the middle of this mesh, one can clearly identify a segment composed of horizontal edges, starting from the inlet of the channel, passing through the two sides of the obstacle, and extending up to the outlet of the channel. This segment will be used to visualize the variation of the solution near the boundary layer.
Since this is a test case without an analytical solution, we rely on numerical simulation results obtained with the `FreeFEM++` code as a reference solution.
First, we present the numerical results from `FreeFEM++` computed with classical second-order Lagrange finite element method on a mesh consisting of 222,654 triangles:

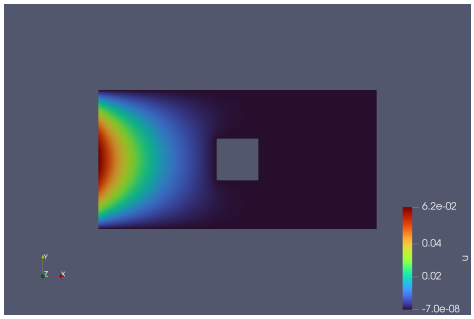


Figure 14.15 – Reference solution of problem 14.4 with $\nu = 1$.

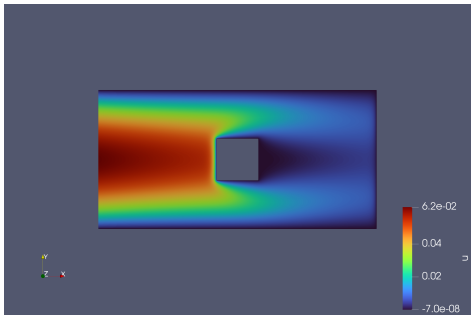


Figure 14.16 – Reference solution of problem 14.4 with $\nu = 10^{-2}$.

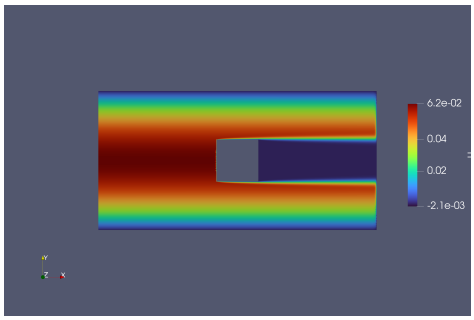


Figure 14.17 – Reference solution of problem 14.4 with $\nu = 10^{-4}$.

From these three figures, one can observe how the solution changes with respect to ν. We now present the numerical simulations obtained using both the polynomial and non-polynomial bases with, for all $\ell \in \mathcal{I}_K$, $\mathbf{b}_\ell = \boldsymbol{\beta}$. In the figures 14.18 and 14.19 shows a transversal slice of the numerical solution, from the inlet of the channel up to the obstacle, for $\nu = 10^{-4}$ and for different mesh sizes $h \in \{0.05, 0.025, 0.0125\}$:

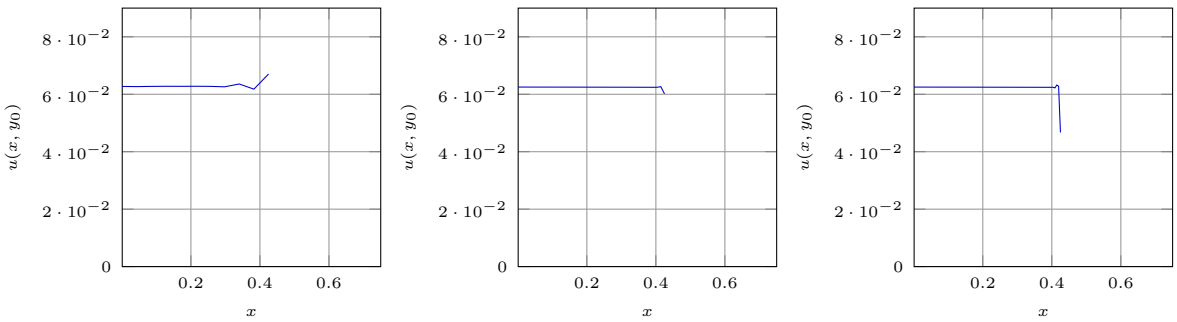


Figure 14.18 – Polynomial basis

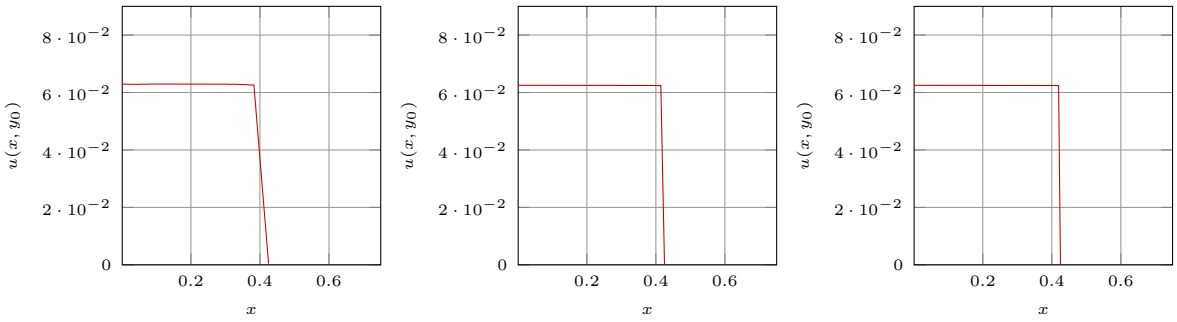


Figure 14.19 – Exponential basis

14.2. Comparison of the numerical results between the polynomial and non-polynomial bases in 2D

In the figure below, the numerical solution u_h computed with the non-polynomial basis.

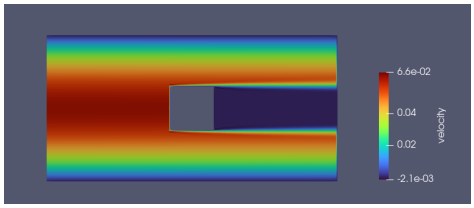


Figure 14.20 – The numerical solution of problem 14.4 with $\nu = 10^{-4}$ obtained using the non-polynomial basis with $h = 0.0125$.

In the figure below, the numerical solution u_h computed with the polynomial basis.

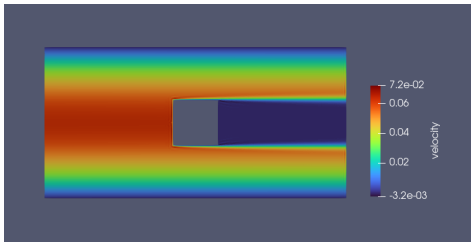


Figure 14.21 – The numerical solution of problem 14.4 with $\nu = 10^{-4}$ obtained using the polynomial basis with $h = 0.0125$.

For the advection-diffusion problem, we can see that the non-polynomial basis has a very remarkable efficiency compared to the polynomial basis. We have fewer oscillations, even on a coarse mesh and for small ν, and the non-polynomial basis accurately captures the rapid variations of the solution near the boundary layer compared to the polynomial basis.

14.2.2 Oseen problem

In this section, we present an academic test case with a manufactured solution to assess the performance of the non-polynomial basis and to analyze how the parameter **b** should be chosen in the basis construction.

Academic test case

Let $\Omega = [0,\, 1] \times [0,\, 1]$, and consider the exact solution given by:

$$\begin{cases} \mathbf{u}(x,y) &= \exp(\nu^{-1}(x+y-2))(1,\,-1), \\ p(x,y) &= \exp(\nu^{-1}(x+y-2)) - \frac{1}{|\Omega|}(\exp(\nu^{-1}(x+y-2)),\,1)_{L^2(\Omega)}. \end{cases}$$

For the non-polynomial basis, we can observe that if we take $\mathbf{b}_\ell = [1, 1]$ for all $\ell \in \mathcal{I}_K$, and if the vector $\boldsymbol{\beta}$ belongs to the approximation space spanned by the non-polynomial basis, then the numerical solution $(\mathbf{u}_h, p_h)$ coincides exactly with the exact solution $(\mathbf{u}, p)$.
However, one may question this choice, since in general we do not know the exact solution of the problem. Why, then, should we take $\mathbf{b}_\ell = [1, 1]$?
To explore this issue, we test several possible choices of $\mathbf{b}_\ell$ as defined in Section 12.4. In order to make the configuration as close as possible to the Navier–Stokes problem, we choose $\boldsymbol{\beta} = \mathbf{u}$. This choice is particularly challenging because the direction of $\boldsymbol{\beta}$ is orthogonal to the optimal choice of $\mathbf{b}_\ell$, namely $[1, 1]$.
Consider the four possible choices of $\mathbf{b}$ defined in Section 12.4. Note that with the third choice we have $\mathbf{b}_\ell = [1, 1]$ and with this choice we obtain the exact solution.
First, we fix $\nu = 1$, results are displayed the numerical results in figures 14.22 to 14.26 :

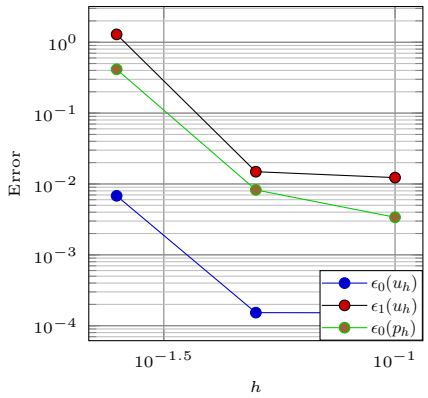


Figure 14.22 – Plots of $\epsilon_0^\nu(\mathbf{u}_h)$, $\epsilon_1^\nu(\mathbf{u}_h)$ and $\epsilon_0^\nu(p_h)$, with $\mathbf{b}_\ell = \dfrac{1}{|K_\ell|}\int_{K_\ell} \boldsymbol{\beta}$

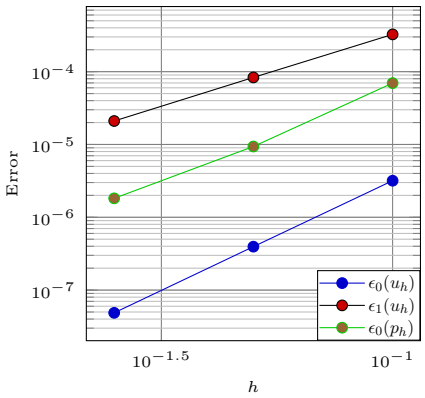


Figure 14.23 – Plots of $\epsilon_0^\nu(\mathbf{u}_h)$, $\epsilon_1^\nu(\mathbf{u}_h)$ and $\epsilon_0^\nu(p_h)$, with $\mathbf{b}_\ell$ choice (4)

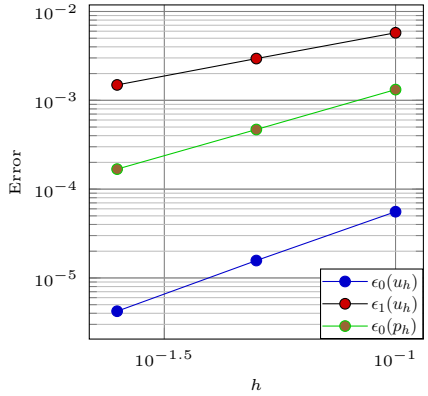


Figure 14.24 – Plots of $\epsilon_0^\nu(\mathbf{u}_h)$, $\epsilon_1^\nu(\mathbf{u}_h)$ and $\epsilon_0^\nu(p_h)$, with $\mathbf{b}_\ell = \boldsymbol{\beta}(\mathbf{x}_\ell)$.

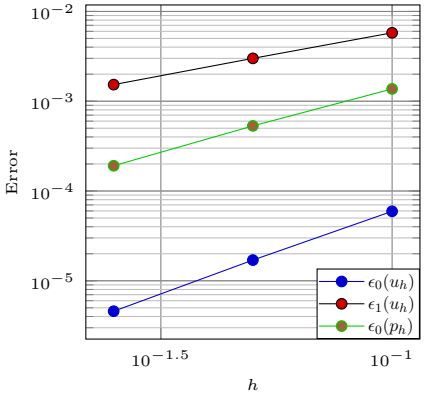


Figure 14.25 – Plots of $\epsilon_0^\nu(\mathbf{u}_h)$, $\epsilon_1^\nu(\mathbf{u}_h)$ and $\epsilon_0^\nu(p_h)$, with $\mathbf{b}_\ell = [\frac{1}{2}, \frac{1}{2}]$

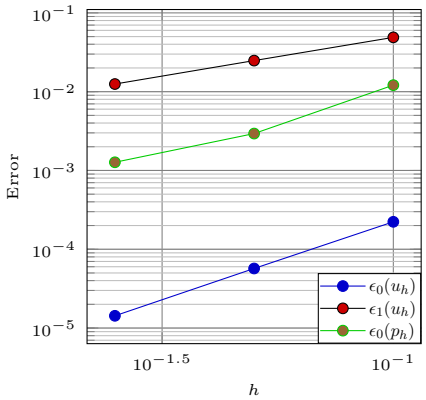


Figure 14.26 – Plots of $\epsilon_0^\nu(\mathbf{u}_h)$, $\epsilon_1^\nu(\mathbf{u}_h)$ and $\epsilon_0^\nu(p_h)$, with polynomial basis.

One can observe that the fourth choice of $\mathbf{b}_\ell$ provides better results in the $\mathbf{L}^2$ and $\mathbf{H}^1$ norms for the velocity, as well as in the L^2 norm for the pressure, compared to the other choices of $\mathbf{b}_\ell$, and also compared to the numerical result obtained with the polynomial basis. It should be noted that other choices of non-polynomial bases— except for the choice of $\mathbf{b}_\ell$ defined as an average value—yield more accurate results, meaning that the error is smaller than with the polynomial basis.

Now we fix $\nu = 10^{-2}$, the numerical results are displayed in figures 14.27 to 14.30.

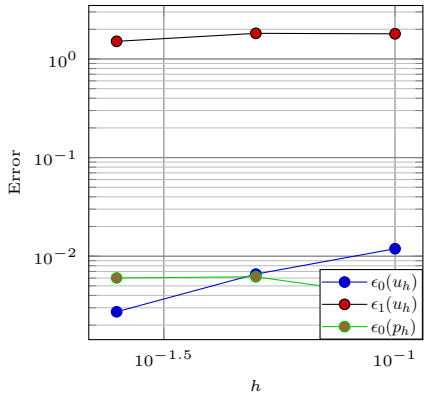


Figure 14.27 – Plots of $\epsilon_0^\nu(\mathbf{u}_h)$, $\epsilon_1^\nu(\mathbf{u}_h)$ and $\epsilon_0^\nu(p_h)$, with polynomial basis.

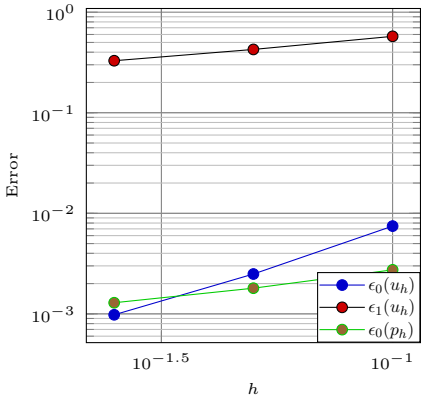


Figure 14.28 – Plots of $\epsilon_0^\nu(\mathbf{u}_h)$, $\epsilon_1^\nu(\mathbf{u}_h)$ and $\epsilon_0^\nu(p_h)$, with $\mathbf{b}_\ell = \boldsymbol{\beta}(\mathbf{x}_\ell)$

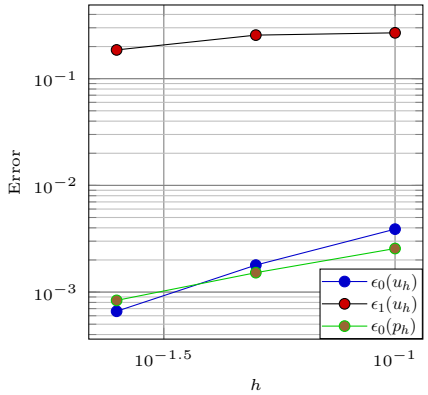


Figure 14.29 – Plots of $\epsilon_0^\nu(\mathbf{u}_h)$, $\epsilon_1^\nu(\mathbf{u}_h)$ and $\epsilon_0^\nu(p_h)$, with $\mathbf{b}_\ell = [\frac{1}{2}, \frac{1}{2}]$.

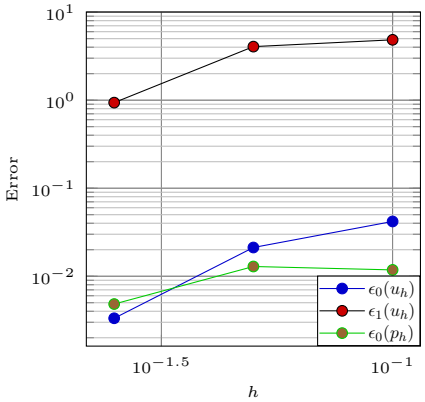


Figure 14.30 – Plots of $\epsilon_0^\nu(\mathbf{u}_h)$, $\epsilon_1^\nu(\mathbf{u}_h)$ and $\epsilon_0^\nu(p_h)$, with $\mathbf{b}_\ell$ choice (4)

We can observe that, for example, with the choice $\mathbf{b}_\ell = [0.5, 0.5]$, the pressure obtained with the non-polynomial basis is better than that obtained with the polynomial basis. We also notice that with choice 4 of $\mathbf{b}_\ell$, the convergence rate becomes higher, and with one or two additional mesh refinements, we can expect even better results compared with the other choices.

Physical test case

Let $\boldsymbol{\beta} = [1, 0]$. Under the same geometrical conditions as in the physical test case of the advection–diffusion problem 14.2.1 where we denote $\Gamma_1 = \{0\} \times [0, \frac{1}{2}]$, $\Gamma_3 = \{1\} \times [0, \frac{1}{2}]$, and

$\Gamma_2 = \partial\Omega \setminus (\Gamma_1 \cup \Gamma_3)$. The Oseen problem is given by:

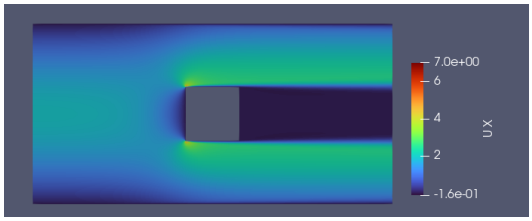
$$\text{Find } (\mathbf{u},\, p) \in \mathbf{H}^1(\Omega) \times L^2(\Omega) \text{ such that } \begin{cases} -\nu\,\Delta\mathbf{u} + (\boldsymbol{\beta}\cdot\mathbf{grad}\,)\mathbf{u} + \mathbf{grad}\,p = 0 \quad \text{in } \Omega, \\ \operatorname{div}\mathbf{u} = 0 \quad \text{in } \Omega, \\ \mathbf{u}_{|\Gamma_1} = (30\,y(\tfrac{1}{2} - y),\, 0), \\ \mathbf{u}_{|\Gamma_2} = (0,\, 0), \\ (\nu\,\mathbf{Grad}\,\mathbf{u} : \mathbf{n} - p\mathbb{I}_2)_{|\Gamma_3} = 0. \end{cases} \tag{14.5}$$

In this test case, there is a single boundary layer and an open (free) outflow for the fluid. The reference solution for the Oseen problem (see Section 14.2.1), computed with `FreeFEM++` using the continuous Galerkin scheme and the classical second-order (Taylor-Hood) Lagrange finite element method for the velocity, and first-order elements for the pressure: $\mathbf{P}^2$-P^1. The numerical results were obtained on a mesh consisting of 143,564 triangles.
The figures 14.31 and 14.32 show the comparison between the reference solution and the numerical results obtained with the polynomial and non-polynomial bases, using $\mathbf{b}_\ell = \boldsymbol{\beta}$ for all $\ell \in \mathcal{I}_K$, with the $\mathbf{P}^1_{dg}$–P^1_{dg} scheme.

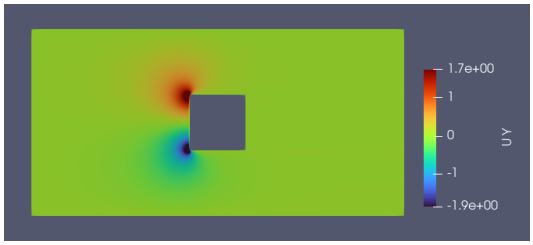
Remark 14.1 *Note that in the FreeFem++ code, a polynomial basis is used. Therefore, it is clear that with mesh refinement, the solution computed using the FEM-DG method with a polynomial basis converges to the reference solution obtained with FreeFem++. In contrast, for the exponential basis, there is no reference solution computed by software such as FreeFem++, since it does not employ an exponential-type basis. For this reason, in what follows, on refined meshes we compare the FreeFem++ results with those obtained using the non-polynomial basis, whereas on coarser meshes we compare all three approaches with each other.*

— For $\nu > 10^{-3}$, by visualizing the numerical results obtained with FreeFem++ and with the FEM-DG method using both the polynomial and exponential bases, we observe that they produce very close results.

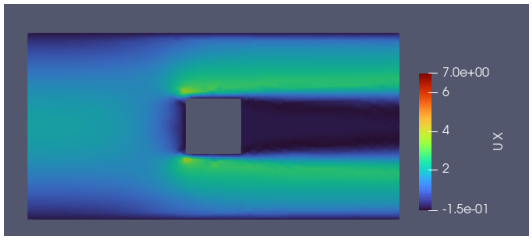
— For $\nu = 10^{-4}$, we consider two meshes: mesh 1 composed of 2336 triangles and mesh 2 composed of 9344 triangles. We compare the numerical results obtained with the non-polynomial basis to the reference solution computed with FreeFem++ in figures 14.31 and 14.32 for the velocity components.



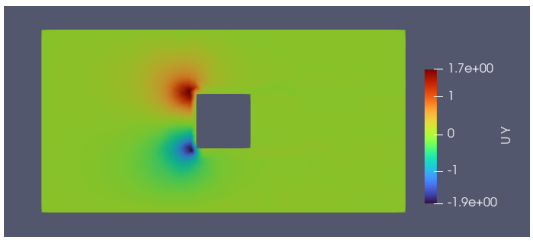
Velocity $\mathbf{u}_x$ computed with FreeFem++.



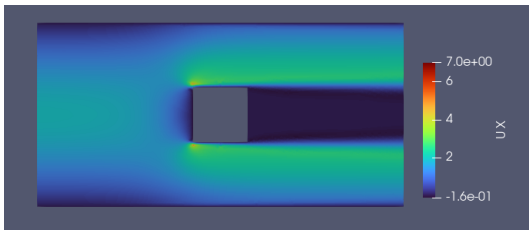
Velocity $\mathbf{u}_y$ computed with FreeFem++.



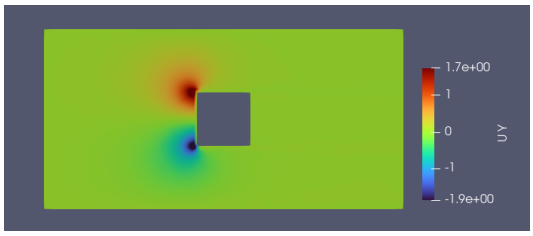
Velocity $\mathbf{u}_x$ computed with the FEM-DG method using the exponential basis (mesh 1).



Velocity $\mathbf{u}_y$ computed with the FEM-DG method using the exponential basis (mesh 1).



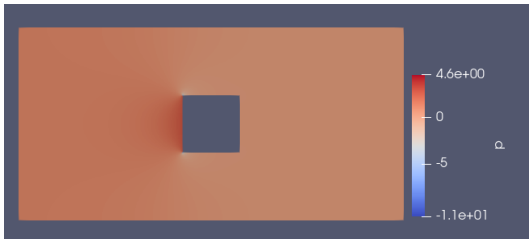
Velocity $\mathbf{u}_x$ computed with the FEM-DG method using the exponential basis (mesh 2).



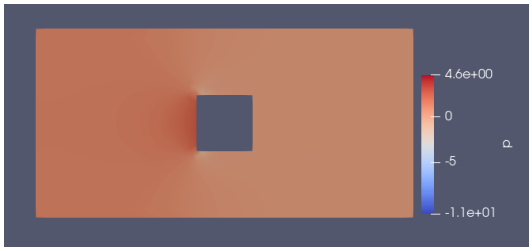
Velocity $\mathbf{u}_y$ computed with the FEM-DG method using the exponential basis (mesh 2).

Figure 14.31 – Representation of the velocities computed with FreeFem++ and the FEM-DG method (non-polynomial basis).

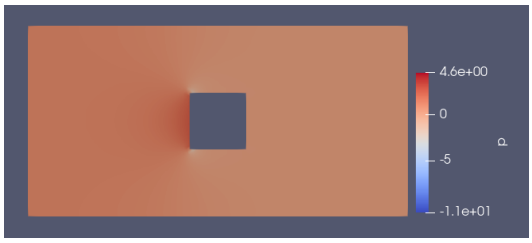
Similarly, for the pressure:



Pressure p computed with FreeFem++.



Pressure p computed with the FEM-DG method using the exponential basis (mesh 1).



Pressure p computed with the FEM-DG method using the exponential basis (mesh 2).

Figure 14.32 – Representation of the pressure computed with FreeFem++ and the FEM-DG method (non-polynomial basis).

A horizontal cross-section of the above solutions is taken along the segment AB, with $A = (0,\, 0.25)$ and $B = (1,\, 0.25)$.

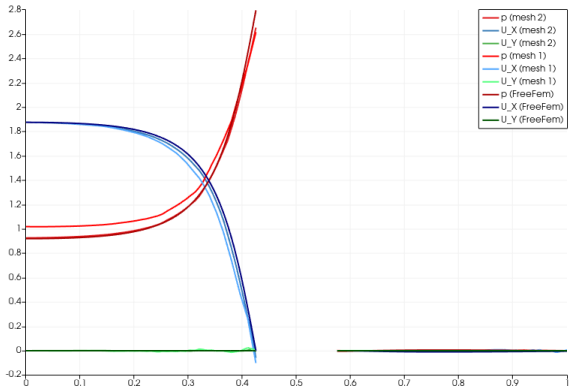


Figure 14.33 – Horizontal cross-section along the segment AB, comparing the solutions obtained with the non-polynomial basis and the reference solution for $\nu = 10^{-4}$.

We can observe that there are no oscillations on either mesh. However, for the first mesh, the non-polynomial basis does not capture well the curvature of the x-component of the velocity near the boundary layer, while with the second mesh the capture improves significantly.

Regarding the numerical solution computed with the polynomial basis on mesh 2, it is closer to the reference solution than the one obtained with the non-polynomial basis.

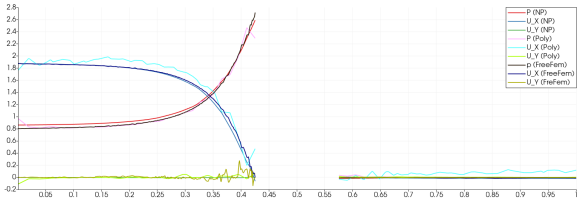


Figure 14.34 – Horizontal cross-section along the segment AB, comparing the solutions obtained with the non-polynomial basis, polynomial basis and the reference solution for $\nu = 10^{-6}$.

In Figure 14.34, we take $\nu = 10^{-6}$. To compute the reference solution with FreeFem++, I generated a refined mesh near the boundary layer, consisting of 143,564 triangles. We observe small oscillations in the x-component of $\mathbf{u}_h$ computed with FreeFem++ using the continuous Galerkin scheme $\mathbf{P}^2$–P^1.

In contrast, with the non-polynomial basis, no oscillations appear; however, the approximate solution shows a slight shift compared with the reference solution for the x-component of the velocity and the pressure.

For the polynomial basis, we notice strong oscillations in both components of $\mathbf{u}$. We also remark that the non-polynomial basis does not produce oscillations even when ν is very small and the mesh is not highly refined. Note that for both the polynomial and non-polynomial bases, the computation is performed on a uniform mesh with $h = 0.025$.

14.2.3 Navier–Stokes problem

As we observed in the Oseen problem, when ν is large, the non-polynomial basis provides better results than the polynomial one. However, when ν is small, an optimal choice of $\mathbf{b}$ has not been identified yet.

For this reason, in the Navier–Stokes problem, we consider the channel-with-obstacle test case in the laminar regime (which means we choose $\nu = 10^{-2}$). In the framework of a physical test case, we consider a classical channel flow with an obstacle in the form of a square. The channel-shaped domain is defined as $\Omega = [0, 3] \times [0, \frac{1}{2}] \setminus [0.2, 0.3] \times [0.2, 0.3]$, with $\Gamma_1 = \{0\} \times [0, \frac{1}{2}]$, $\Gamma_3 = \{3\} \times [0, \frac{1}{2}]$ and $\Gamma_2 = \partial\Omega \setminus \Gamma_1 \cup \Gamma_3$.

Note that the numerical solutions have become stationary before the final time $t = 5$ and so we plot the solutions at that time.

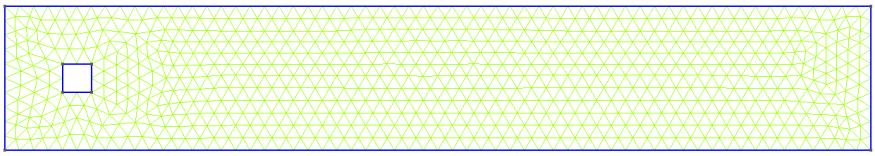


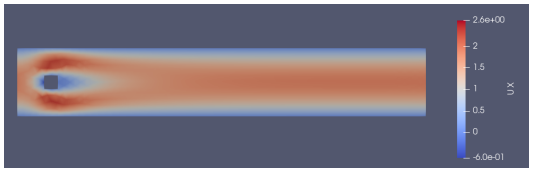
Figure 14.35 – Mesh for the channel with square obstacle.

The Navier–Stokes problem is given by: Find $(\mathbf{u}, p)$ such that

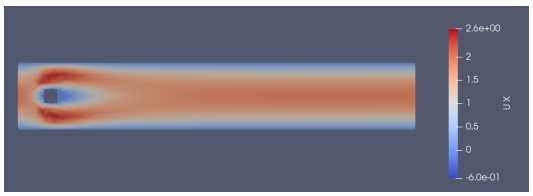
$$\begin{cases} \partial_t \mathbf{u} - \nu\,\Delta\mathbf{u} + (\mathbf{u}\cdot\mathbf{grad}\,)\mathbf{u} + \mathbf{grad}\,p = 0 \quad \text{in } \Omega \times [0,5], \\ \operatorname{div}\mathbf{u} = 0 \quad \text{in } \Omega \times [0,5], \\ \mathbf{u}_{|\Gamma_1} = (30\,y(\frac{1}{2} - y),\, 0), \\ \mathbf{u}_{|\Gamma_2} = (0,\, 0), \\ (\nu\,\mathbf{Grad}\,\mathbf{u} : \mathbf{n} - p\mathbb{I}_2)_{|\Gamma_3} = 0, \\ \mathbf{u}_0 = \tilde{\mathbf{u}}, \quad t = 0 \end{cases} \tag{14.6}$$

Where $\tilde{\mathbf{u}}$ is defined by this Stokes problem:

$$\text{Find } (\tilde{\mathbf{u}},\, \tilde{p}) \in \mathbf{H}^1(\Omega) \times L^2_{vzm}(\Omega) \text{ such that } \begin{cases} -\nu\, \Delta\tilde{\mathbf{u}} + \mathbf{grad}\, \tilde{p} = 0 \quad \in \Omega, \\ \operatorname{div} \tilde{\mathbf{u}} = 0 \quad \in \Omega, \\ \tilde{\mathbf{u}}_{|\Gamma_1} = (30\, y(\frac{1}{2} - y),\, 0), \\ \tilde{\mathbf{u}}_{|\Gamma_2} = (0,\, 0), \\ (\nu\, \mathbf{Grad}\, \mathbf{u} : \mathbf{n} - p\mathbb{I}_2)_{|\Gamma_3} = 0. \end{cases} \tag{14.7}$$

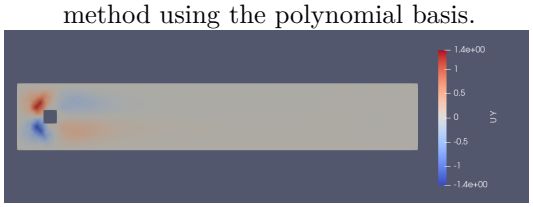


Velocity $\mathbf{u}_x$ computed with the FEM-DG method using the non-polynomial basis.

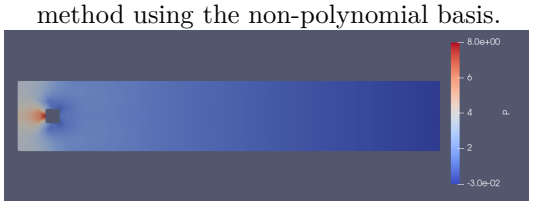


Velocity $\mathbf{u}_x$ computed with the FEM-DG method using the polynomial basis.

Velocity $\mathbf{u}_y$ computed with the FEM-DG method using the non-polynomial basis.



Velocity $\mathbf{u}_y$ computed with the FEM-DG method using the polynomial basis.



Pressure p computed with the FEM-DG method using the non-polynomial basis.

Pressure p computed with the FEM-DG method using the polynomial basis.

Figure 14.36 – Representation of the velocity components and pressure computed with FEM-DG method using polynomial and non-polynomial basis at $t = 5s$.

A horizontal cross-section of the above solutions is taken along the segment AB, with $A = (0,\, 0.25)$ and $B = (3,\, 0.25)$.

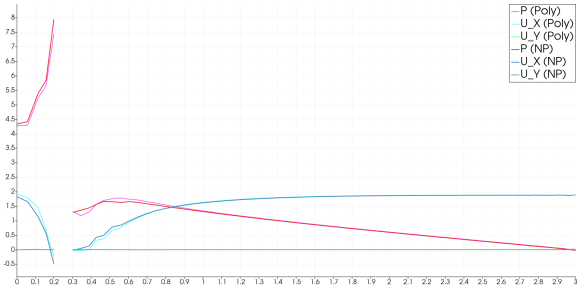


Figure 14.37 – Horizontal cross-section along the segment AB, comparing the solutions obtained with the polynomial basis and polynomial basis and $t = 5s$.

In Figures 14.36 and 14.37, we display the solution at $t = 5\,\mathrm{s}$, after sufficient time has passed for the flow to become stationary. We observe that the solutions computed using the polynomial and non-polynomial bases are close to each other, but further investigation and a reference solution are needed to draw more definitive conclusions.

In this chapter, we presented the numerical results for the advection–diffusion, Oseen, and Navier–Stokes problems, computed using both the polynomial and non-polynomial bases in order to evaluate the performance of the non-polynomial basis. We found that this basis yields very good results for the advection–diffusion and Oseen problems, and that with a more suitable choice of $\mathbf{b}$, the results could be even better. For the Navier–Stokes problem, we did not consider a sufficiently large set of test cases. However, in general, we can confirm that even on a coarse mesh, the non-polynomial basis does not produce oscillations, even when the viscosity is very small. In contrast, with the polynomial basis and very small ν, oscillations appear, especially near boundary layers.

Conclusions and Perspectives

14.3 Conclusions

This thesis was devoted to the advancement of numerical simulation of incompressible flow problems by discontinuous Galerkin methods. Motivated by the context of flows in nuclear reactor cores, we studied in particular two aspects. In the first part of the thesis, we studied the a priori error estimations that we can expect with such methods in non convex domains or with low regularity source terms. In the second part, we take advantage of the flexibility of the discontinuous methods to propose non standard (exponential) basis functions that may better capture solutions with stiff variations, which happens in particular in boundary layers.

More specifically let us detail the following points: In the first part, we introduce the SIP discretization of the Poisson problem, considering simplicial triangulations. We allow the stability parameter η to vary across mesh faces, leading to a face-dependent coefficient η_f, which seems to be new.

Then, in section 6.2, we generalize the analysis to a polytopal mesh, and we exhibit the stability constant. We detail the discretization of the Stokes problem in section 6.3, notice that when the solution of the Stokes problem has low-regularity, i.e. if there exists $s < 1$ such that $(\mathbf{u}, p) \in \mathbf{H}^{1+s}(\Omega) \times H^s(\Omega)$, we preserve the consistency of the discrete formulation, assuming that $s > \frac{1}{2}$. Moreover, in section 6.5, we exhibit the dependence of the coercivity constant for the Stokes problem with respect to the mesh regularity parameter σ.

We continued by the analysis of a priori error estimates for the discontinuous Galerkin finite element method (DGFEM) applied to Stokes problems with low-regularity solutions in chapter 7. By focusing on solutions in the space $\mathbf{H}^{1+s}(\Omega) \times H^s(\Omega)$ where $s > \frac{1}{2}$, we have demonstrated the effectiveness of this method in non-convex domains or with low-regularity source terms. The results confirm that DGFEM is a robust approach for the discretization of these problems, offering valuable flexibility for modeling complex phenomena in fluid mechanics, and, we demonstrated their efficiency and accuracy even with low-regularity

solutions. In section 6.4, we propose a new numerical scheme for which the polynomial order of the pressure is one degree higher than that of the velocity, as suggested by the proof of the a priori convergence order. We demonstrate the effectiveness of this approach in the numerical results chapter 9 when ν is very small, which is our case of interest. The $\mathbf{P}^1_{dg} - P^2_{dg}$ scheme seems more efficient than other options in this particular setting, which, to our knowledge, is a new result.

To advance further, we propose in the second part of this thesis discontinuous Galerkin methods designed to capture rapid variations or irregularities in the solution. In sections 11.1, 11.2 and 11.3, we introduce a new idea in discontinuous Galerkin finite elements, where we propose an exponential basis in one, two and three dimensions and explain the motivation behind it, based on local generic solutions of a convection diffusion 1D equation. Moreover in section 11.4 we compute the parameter C_{NP}, which is required to ensure the coercivity of the bilinear form for the Poisson problem. In the numerical results chapter 14, we have shown that the non-polynomial basis is more effective in terms of accuracy and in reducing oscillations. Even on a coarse mesh, the non-polynomial basis does not produce oscillations, whereas the polynomial basis does, as seen for the advection–diffusion problem in Section 14.1 and Subsection 14.2. For the Oseen problem in Subsection 14.2.2, we also observed that the non-polynomial basis outperforms the polynomial one, although the results could be further improved by using a more adaptive choice of $\mathbf{b}$. For the Navier–Stokes problem, we have not yet performed a sufficiently extensive numerical study, but we have included preliminary results. It would be preferable to carry out additional experiments to more thoroughly assess the performance of the non-polynomial basis.

14.4 Perspectives

The perspectives are divided into two parts corresponding to the work presented:

— Perspective 1: Following the study carried out in the first part, one may further refine the analysis by estimating more precisely the dependence of the error bound on the mesh regularity parameter σ. Since we used the averaging interpolation operator $\mathcal{I}_h^{av}$, some inequalities—such as those in Theorem 7.3 and Lemma 7.4—involve the set D_ℓ, which consists of the simplices that share at least one point with the triangle K_ℓ. Making the dependence on σ explicit via the mapping to the reference element becomes more delicate because defining the corresponding mapping for D_ℓ is not straightforward.

— Perspective 2: A natural continuation of this work is to derive a posteriori error estimates for the $\mathbf{P}^k - P^{k+1}$ scheme, together with mesh adaptation.

— Perspective 3: To go further, one may propose a new discretization for the Stokes problem, similar to that in [28], but using the FEM-DG method.

— Second part:

— Perspective 4: From the numerical results, we observed that an optimal choice of $\mathbf{b}$ for the Oseen and Navier–Stokes problems has not yet been identified. The first perspective is therefore to develop a method for determining an appropriate $\mathbf{b}$.

— Perspective 5: A second perspective is to enrich the non-polynomial basis with additional exponential terms in order to capture all possible directions of $\mathbf{b}$. As explained earlier, in dimensions 2 and 3 for the advection–diffusion problem, each direction of $\boldsymbol{\beta}$ corresponds to a solution; see Remark 11.6 for more details.

— Perspective 6: Once this enrichment is done, one may establish the necessary inequalities to prove coercivity and continuity of the bilinear forms associated with the non-polynomial basis. This would then open the way to deriving a priori and a posteriori error estimates, together with mesh adaptation strategies.

— Perspective 4: Developing a code that combines a polynomial basis with a non-polynomial basis, adapted to capture rapid variations in the solution or to handle low-regularity regimes.

Bibliography

[1] P. Acevedo et al. « Stokes and Navier-Stokes equations with Navier boundary condition ». In: *C. R. Math. Acad. Sci. Paris* 357.2 (2019), pp. 115–119.

[2] R. A. Adams and J.J.F. Fournier. *Sobolev spaces*. Vol. 140. Elsevier, 2003.

[3] C. Amrouche, H. Bouzit, and U. Razafison. « On the two and three dimensional Oseen potentials ». In: *Potential Analysis* 34 (2011), pp. 163–179.

[4] C. Amrouche and U. Razafison. « On the existence of solutions in weighted Sobolev spaces for the exterior Oseen problem ». In: *Comptes Rendus Mathématique* 341.9 (2005), pp. 587–592.

[5] C. Amrouche and M.Á. Rodríguez-Bellido. « Very weak solutions for the stationary Oseen and Navier-Stokes equations ». In: *Comptes Rendus Mathématique* 348.17–18 (2010), pp. 335–339.

[6] V. Anaya et al. « Error analysis for a vorticity/Bernoulli pressure formulation for the Oseen equations ». In: *Journal of Numerical Mathematics* 30.3 (2022), pp. 209–230.

[7] D.N. Arnold. « An interior penalty finite element method with discontinuous elements ». In: *SIAM Journal on Numerical Analysis* 19.4 (1982), pp. 742–760.

[8] F. Assous, P. Ciarlet Jr., and S. Labrunie. *Mathematical Foundations of Computational Electromagnetism*. Vol. 198. Applied Mathematical Sciences. Springer, 2018.

[9] I. Babuška. « The finite element method with Lagrange multipliers ». In: *Numerische Mathematik* 20 (1973), pp. 179–192.

[10] I. Babuška and M. Zlámal. « Nonconforming elements in the finite element method with penalty ». In: *SIAM Journal on Numerical Analysis* 10.5 (1973), pp. 863–875.

[11] G. Baker. « Finite element methods for elliptic equations using nonconforming elements ». In: *Mathematics of Computation* 31.137 (1977), pp. 45–59.

[12] F. Bassi and S. Rebay. « A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations ». In: *Journal of Computational Physics* 131 (1997), pp. 267–279.

[13] F. Bassi et al. « An implicit high-order discontinuous Galerkin method for steady and unsteady incompressible flows ». English. In: *Comput. Fluids* 36.10 (2007), pp. 1529–1546. ISSN: 0045-7930. DOI: 10.1016/j.compfluid.2007.03.012.

[14] G.K. Batchelor. *An Introduction to Fluid Dynamics*. Cambridge University Press, 1967.

[15] F. Boyer and P. Fabrie. *Mathematical Tools for the Study of the Incompressible Navier-Stokes Equations and Related Models*. Vol. 183. Applied Mathematical Sciences. Springer, 2013. URL: https://link.springer.com/book/10.1007/978-1-4614-5975-0.

[16] R.M. Brown. « The mixed problem for the Stokes system in Lipschitz domains ». In: *Advances in Differential Equations* 6.3 (2001), pp. 327–349.

[17] R.M. Brown and Z. Shen. « Estimates for the Stokes operator in Lipschitz domains ». In: *Indiana University Mathematics Journal* 51.3 (2002), pp. 871–930.

[18] C. Caussignac and R. Touzani. « A discontinuous Galerkin method for three-dimensional boundary-layer equations ». In: *Numerical Methods in Fluids* (1990).

[19] M. Čermák et al. « Adaptive inexact iterative algorithms based on polynomial-degree-robust a posteriori estimates for the Stokes problem ». In: *Numerische Mathematik* 138 (2018), pp. 1027–1065.

[20] D. Chamorro. *Introduction aux équations de Navier-Stokes incompressibles*. EDP Sciences, 2025. URL: https://www.uneautrepage.fr/product/509012/introduction-aux-equations-de-navier-stokes-incompressibles.

[21] G. Chavent and B. Cockburn. « The Local Projection $P^0 - P^1$ Discontinuous Galerkin Finite Element Method for Conservation Laws ». In: *ESAIM: Mathematical Modelling and Numerical Analysis* 23 (1989), pp. 565–592.

[22] D. Chicaud and P. Ciarlet Jr. « Analysis of time-harmonic Maxwell impedance problems in anisotropic media ». In: *SIAM Journal on Mathematical Analysis* 55 (2023), pp. 1969–2000.

[23] P.G. Ciarlet. *The finite element method for elliptic problems*. Vol. 40. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.

[24] P. Ciarlet Jr. « Analysis of the Scott–Zhang interpolation in the fractional order Sobolev spaces ». In: *Journal of Numerical Mathematics* 21.3 (2013), pp. 173–180.

[25] P. Ciarlet Jr. « On the approximation of electromagnetic fields by edge finite elements. Part 1: sharp interpolation results for low-regularity fields ». In: *Computers and Mathematics with Applications* 71 (2016), pp. 85–104.

[26] P. Ciarlet Jr. *T-coercivity: a practical tool for the study of variational formulations in Hilbert spaces*. 2025.

[27] P. Ciarlet Jr. « T-coercivity: Application to the discretization of Helmhotz-like problems ». In: *Computers & Mathematics with Applications* 64.1 (2012), pp. 22–24.

[28] P. Ciarlet Jr. and E. Jamelot. « Explicit T-coercivity for the Stokes problem: a coercive finite element discretization ». In: *Computers & Mathematics with Applications* 188 (2025), pp. 137–159. URL: https://doi.org/10.1016/j.camwa.2025.03.028.

[29] P. Ciarlet Jr. and E. Lunéville. *The Finite Element Method: From Theory to Practice*. John Wiley & Sons, 2023.

[30] B. Cockburn. « Discontinuous Galerkin methods for computational fluid dynamics ». In: *Encyclopedia of computational mechanics* 3 (2004), pp. 91–127.

[31] B. Cockburn, G. Kanschat, and D. Schötzau. « The local discontinuous Galerkin method for the Oseen equations ». In: *Mathematics of Computation* 73.246 (2004), pp. 569–593.

[32] B. Cockburn and C. W. Shu. « The local discontinuous Galerkin method for time-dependent convection-diffusion systems ». In: *SIAM Journal on Numerical Analysis* 35.6 (2000), pp. 2440–2463.

[33] B. Cockburn and C. W. Shu. « The Runge-Kutta Discontinuous Galerkin Method for Conservation Laws V: Multidimensional Systems ». In: *Journal of Computational Physics* 141 (1997), pp. 199–224.

[34] B. Cockburn and C.W. Shu. « Approximate Riemann solvers, parameterized central schemes, and the discontinuous Galerkin method for conservation laws ». In: *SIAM Journal on Numerical Analysis* 39 (2002), pp. 264–286.

[35] *Code TrioCFD*. URL: https://triocfd.cea.fr/.

[36] M. Dauge. *Elliptic Boundary Value Problems on Corner Domains*. Springer-Verlag, 1988.

[37] M. Dauge. « Stationary Stokes and Navier-Stokes systems on two- or three-dimensional domains with corners. Part I: Linearized equations ». In: *SIAM J. Math. Anal.* 20 (1989).

[38] A.-S. Bonnet-Ben Dhia and P. Ciarlet Jr. *Méthodes variationnelles pour l'analyse de problèmes non coercifs*. M.Sc. AMS lecture notes (ENSTA-IPP). 2023. URL: https://perso.ensta-paris.fr/~ciarlet/AMS303/AMS303_Presentation.html.

[39] E. Di Nezza, G. Palatucci, and E. Valdinoci. « Hitchhiker's guide to the fractional Sobolev spaces ». In: *Bull. Sci. Math.* 136 (2012), pp. 521–573.

[40] D.A. Di Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*. Vol. 69. Springer, Mathématiques & Applications, 2012.

[41] M. Dindoš and M. Mitrea. « The stationary Navier-Stokes system in nonsmooth manifolds and Lipschitz domains ». In: *Communications in Partial Differential Equations* 34.6 (2009), pp. 849–896.

[42] J. Donea and A. Huerta. *Finite Element Methods for Flow Problems*. Wiley, 2003.

[43] A.N. Douglas et al. « Unified analysis of discontinuous Galerkin methods for elliptic problems ». In: *SIAM journal on numerical analysis* 39.5 (2002), pp. 1749–1779.

[44] C. Douglas and T. Dupont. « Interior penalty procedures for elliptic and parabolic Galerkin methods ». In: *Computing Methods in Applied Sciences* 58 (1976), pp. 207–216.

[45] J. Droniou. « Non-coercive Linear Elliptic Problems ». In: *Potential Analysis* 17 (2002), pp. 181–203.

[46] A. Ern and J.-L. Guermond. « Discontinuous Galerkin methods for Friedrichs' systems. I. General theory ». In: *SIAM Journal on Numerical Analysis* 41 (2004), pp. 1020–1053.

[47] A. Ern and J.-L. Guermond. « Discontinuous Galerkin methods for Friedrichs' systems. II. Second-order elliptic PDEs ». In: *SIAM Journal on Numerical Analysis* 44 (2006), pp. 236–259.

[48] A. Ern and J.-L. Guermond. « Finite element quasi-interpolation and best approximation ». In: *ESAIM: M2AN* 51 (2017), pp. 1367–1385.

[49] A. Ern and J.-L. Guermond. *Finite elements I*. Vol. 72. Texts in Applied Mathematics. Springer, 2021.

[50] A. Ern and J.-L. Guermond. *Finite elements II*. Vol. 73. Texts in Applied Mathematics. Springer, 2021.

[51] L.C. Evans. *Partial differential equations*. 2nd ed. Vol. 19. Grad. Stud. Math. Providence, RI: American Mathematical Society (AMS), 2010.

[52] M. Fabes, O. Mendez, and M. Mitrea. « Boundary layers on Sobolev–Besov spaces and Poisson's equation for the Laplacian in Lipschitz domains ». In: *Journal of Functional Analysis* 159.2 (2002), pp. 545–588.

[53] L. Fatone, P. Gervasio, and A. Quarteroni. « Multimodels for incompressible flows: iterative solutions for the Navier-Stokes / Oseen coupling ». In: *ESAIM: Mathematical Modelling and Numerical Analysis* 35.3 (2001), pp. 549–574.

[54] C. Foias et al. *Navier-Stokes Equations and Turbulence*. Cambridge University Press, 2001.

[55] G. Galdi. *An introduction to the mathematical theory of the Navier-Stokes equations: Steady-state problems*. Springer Science & Business Media, 2011.

[56] E.H. Georgoulis, E. Hall, and P. Houston. « Discontinuous Galerkin methods for advection-diffusion-reaction problems on anisotropically refined meshes ». In: *SIAM Journal on Scientific Computing* 30.1 (2008), pp. 246–271.

[57] V. Girault and P.-A. Raviart. *Finite element methods for Navier-Stokes equations*. Springer-Verlag, 1986.

[58] V. Girault, B. Rivière, and M.F. Wheeler. « A splitting method using discontinuous Galerkin for the transient incompressible Navier-Stokes equations ». In: *ESAIM: M2AN* 39 (2005).

[59] *Graduate School of Science, Department of Physics, Kyushu University*. URL: https://ne.phys.kyushu-u.ac.jp/seminar/MicroWorld3_E/3Part3_E/3P33_E/nuclear_fission_E.htm.

[60] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Classics in Applied Mathematics. Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics, 1985.

[61] P. Grisvard. *Singularities in Boundary Value Problems*. Vol. 22. Research Notes in Applied Mathematics. Berlin, Heidelberg: Springer-Verlag, 1992.

[62] G. Grubb. « Initial value problems for the Navier-Stokes equations with Neumann conditions ». In: *The Navier–Stokes Equations II — Theory and Numerical Methods*. Vol. 1530. Lecture Notes in Mathematics. Springer, 1992, pp. 262–283.

[63] R.A. Horn and C.R. Johnson. *Matrix Analysis*. 2nd ed. Cambridge University Press, 2012.

[64] P. Houston, C. Schwab, and E. Süli. « Discontinuous hp-Finite Element Methods for Advection–Diffusion–Reaction Problems ». In: *SIAM Journal on Numerical Analysis* 39.6 (2002), pp. 2133–2163.

[65] J. Jaffré, C. Johnson, and A. Szepessy. « Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws ». In: *Mathematics of Computation* 62.206 (1995), pp. 531–553.

[66] D. S. Jerison and C. E. Kenig. « The Neumann problem on Lipschitz domains ». In: *Bull. Amer. Math. Soc.* 4 (1981), pp. 203–207.

[67] C. Johnson and J. Pitkäranta. « An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation ». In: *Mathematics of Computation* 46.173 (1986), pp. 1–26.

[68] O.A. Ladyzhenskaya. *The Mathematical Theory of Viscous Incompressible Flow*. New York: Gordon & Breach, 1969.

[69] O.A. Ladyzhenskaya, V.A. Solonnikov, and N.N. Ural'tseva. *Linear and Quasilinear Equations of Parabolic Type*. American Mathematical Society, 1968.

[70] L.D. Landau and E.M. Lifshitz. *Fluid Mechanics*. 2nd. Pergamon Press, 1987.

[71] C. Lasser and A. Toselli. « An overlapping domain decomposition preconditioner for a class of discontinuous Galerkin approximations of advection-diffusion problems ». In: *Mathematics of Computation* 72 (2003), pp. 1215–1238.

[72] P.D. Lax. « Change of Variables in Multiple Integrals ». In: *The American Mathematical Monthly* 106.6 (1999), pp. 497–501.

[73] J. Leray. « Sur le mouvement d'un liquide visqueux emplissant l'espace ». In: *Acta Mathematica* 63 (1934), pp. 193–248.

[74] P. Lesaint and P.-A. Raviart. « On a finite element method for solving the neutron transport equation ». In: *Mathematical Aspects of Finite Elements in Partial Differential Equations*. 1974, pp. 89–145.

[75] J.-L. Lions and E. Magenès. *Non-Homogeneous Boundary Value Problems and Applications*. Vol. I. Grundlehren der mathematischen Wissenschaften. Springer, 1972.

[76] A. Lozinski. « A primal discontinuous Galerkin method with static condensation on very general meshes ». In: *Numerische Mathematik* 143.3 (2019), pp. 583–604.

[77] J.-L. Marichal and M. Dorff. « Derivative Relationships between Volume and Surface Area of Compact Regions in $\mathbb{R}^d$ ». In: *Rocky Mountain Journal of Mathematics* 37.2 (2007), pp. 517–539.

[78] A. Mazaheri and H. Nishikawa. « Efficient high-order discontinuous Galerkin schemes with first-order hyperbolic advection–diffusion system approach ». In: *Journal of Computational Physics* 321 (2016), pp. 729–754.

[79] W. McLean. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, 2003.

[80] M. Mitrea and M. Wright. *Boundary value problems for the Stokes system in arbitrary Lipschitz domains*. Vol. 344. Astérisque. Société Mathématique de France, 2012, p. 247.

[81] T. Miyakawa. « The L_p approach to the Navier-Stokes equations with the Neumann boundary condition ». In: *Hiroshima Mathematical Journal* 10.3 (1980), pp. 517–537.

[82] C. Moldoveanu. « Simulation des grandes échelles de tourbillons longitudinaux soumis à une turbulence extérieure intense ». PhD thesis. Institut de Mécanique des Fluides de Toulouse, France, 2007.

[83] P. Mucha and W. Zajaczkowski. « On the existence for the Cauchy-Neumann problem for the Stokes system in the L_p-framework ». In: *Studia Math.* 143.1 (2000), pp. 75–101.

[84] F. Naddei et al. « A comparison of refinement indicators for p-adaptive simulations of steady and unsteady flows using discontinuous Galerkin methods ». In: *Journal of Computational Physics* 376 (2019), pp. 508–533.

[85] N. Nguyen, J. Peraire, and B. Cockburn. « An implicit high-order hybridizable discontinuous Galerkin method for linear convection-diffusion equations ». In: *Journal of Computational Physics* 228 (2009), pp. 8841–8855.

[86] J. Nitsche. « Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind ». In: *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg* 36 (1971), pp. 9–15.

[87] *Official website of the Canadian Mining and Energy Corporation*. URL: www.cameco.com.

[88] *Official website of the French Alternative Energies and Atomic Energy Commission*. URL: https://www.cea.fr/.

[89] M. A. Olshanskii and E. E. Tyrtyshnikov. *Iterative Methods for Linear Systems*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2014.

[90] A. Peitavy. « Schémas de discrétisation en pression et éléments finis de Crouzeix-Raviart pour les écoulements de fluides incompressibles ». Thèse de doctorat. Université Gustave Eiffel, 2024.

[91] D. Di Pietro and S. Krell. « A Hybrid High-Order method for the steady incompressible Navier–Stokes problem ». In: *Journal of Scientific Computing* 74.3 (2018), pp. 1677–1705.

[92] W. Reed and T. Hill. « Triangular mesh methods for the neutron transport equation ». In: *Los Alamos Scientific Laboratory Report* (1973).

[93] M. Renardy and R.C. Rogers. *An introduction to partial differential equations*. 2nd ed. Vol. 13. Texts Appl. Math. New York, NY: Springer, 2004.

[94] Y. Shibata and S. Shimizu. « On the L_p-L_q maximal regularity of the Neumann problem for the Stokes equations in a bounded domain ». In: *J. reine angew. Math.* 615 (2008), pp. 157–235.

[95] Y. Shibata and S. Shimizu. « On the Stokes equation with Neumann boundary condition ». In: *Banach Center Publications* 70 (2005), pp. 365–377.

[96] R. Temam. *Navier-Stokes Equations: Theory and Numerical Analysis*. North-Holland, 1977.

[97] R. Verfürth. « A Posteriori Error Estimators for the Stokes Equations ». In: *Numerische Mathematik* 55 (1989), pp. 309–325.

[98] G. Vinsard. *Modélisation/Calculs scientifiques - GME5 Leçon 3 Intégration numérique*. 2017. URL: https://gerard.vinsard.fr/Cours/MoCa/AN_Lecons_No1.pdf.

[99] T. Warburton and J.S. Hesthaven. « On the constants in hp-finite element trace inverse inequalities ». In: *Computer Methods in Applied Mechanics and Engineering* 192.25 (2003), pp. 2765–2773.

[100] M.F. Wheeler. « A Galerkin Procedure for Estimating the Flux for Two-Point Boundary Value Problems ». In: *SIAM Journal on Numerical Analysis* 11.4 (1974), pp. 764–770.

[101] K. Chen Y. Shih J. Cheng. « An exponential-fitting finite element method for convection–diffusion problems ». In: *Applied Mathematics and Computation* 217 (2011), pp. 5798–5809.