Discontinuous hp finite element methods for elliptic eigenvalue problems with singular potentials : with applications to quantum chemistry

Carlo Marcati

Sorbonne Université

Laboratoire Jacques-Louis Lions

Thèse de Doctorat de Sorbonne Université

pour l'obtention du grade de
Docteur de Sorbonne Université
Spécialité : Mathématiques Appliquées

par
Carlo Marcati

sous la direction de
Yvon Maday

Discontinuous *hp* finite element methods for elliptic eigenvalue problems with singular potentials

with applications to quantum chemistry

École Doctorale de Sciences Mathématiques
de Paris Centre ED 386

Faculté de Mathématiques
UFR 929

Cette thèse a été préparée au

Laboratoire Jacques-Louis Lions
4 place Jussieu
75005 Paris
France

Site `http://ljll.math.upmc.fr/`

Discontinuous hp finite element methods for elliptic eigenvalue problems with singular potentials

Abstract

In this thesis, we study elliptic eigenvalue problems with singular potentials, motivated by several models in physics and quantum chemistry, and we propose a discontinuous Galerkin hp finite element method for their solution. In these models, singular potentials occur naturally (associated with the interaction between nuclei and electrons). Our analysis starts from elliptic regularity in non homogeneous weighted Sobolev spaces. We show that elliptic operators with singular potential are isomorphisms in those spaces and that we can derive weighted analytic type estimates on the solutions to the linear eigenvalue problems. The isotropically graded hp method provides therefore approximations that converge with exponential rate to the solution of those eigenproblems. We then consider a wide class of nonlinear eigenvalue problems, and prove the convergence of numerical solutions obtained with the symmetric interior penalty discontinuous Galerkin method. Furthermore, when the non linearity is polynomial, we show that we can obtain the same analytic type estimates as in the linear case, thus the numerical approximation converges exponentially. We also analyze under what conditions the eigenvalue converges at an increased rate compared to the eigenfunctions. For both the linear and nonlinear case, we perform numerical tests whose objective is both to validate the theoretical results, but also evaluate the role of sources of errors not considered previously in the analysis, and to help in the design of hp/dG graded methods for more complex problems.

Keywords: hp/dG graded finite element method, discontinuous Galerkin, nonlinear eigenvalue problem, quantum chemistry, weighted Sobolev spaces, elliptic regularity

Résumé

Dans cette thèse, on étudie des problèmes aux valeurs propres elliptiques avec des potentiels singuliers, motivés par plusieurs modèles en physique et en chimie quantique, et on propose une méthode des éléments finis de type hp discontinus (dG) adaptée pour l'approximation des modes propres. Dans ces modèles, arrivent naturellement des potentiels singuliers (associés à l'interaction entre noyaux et électrons). Notre analyse commence par une étude de la régularité elliptique dans des espaces de Sobolev à poids. On montre comment un opérateur elliptique avec potentiel singulier est un isomorphisme entre espaces de Sobolev à poids non homogènes et que l'on peut développer des bornes de type analytique à poids sur les solutions des problèmes aux valeurs propres associés aux opérateurs. La méthode hp/dG graduée qu'on utilise converge ainsi de façon exponentielle. On poursuit en considérant une classe de problèmes non linéaires représentatifs des applications. On montre que, sous certaines conditions, la méthode hp/dG graduée converge et que, si la non linéarité est de type polynomiale, on obtient les mêmes estimations de type analytique que dans le cas linéaire. De plus, on étudie la convergence de la valeur propre pour voir sous quelles conditions la vitesse de convergence est améliorée par rapport à celle des vecteurs propres. Pour tous les cas considérés, on effectue des tests numériques, qui ont pour objectif à la fois de valider les résultats théoriques, mais aussi d'évaluer le rôle des sources d'erreur non considérées dans l'analyse et d'aider dans la conception de méthode hp/dG graduée pour des problèmes plus complexes.

Mots clés : méthode des éléments finis hp/dG graduée, Galerkin discontinu, problèmes aux valeurs propres non linéaires, chimie quantique, espaces de Sobolev à poids, régularité elliptique

Laboratoire Jacques-Louis Lions
4 place Jussieu – 75005 Paris – France

Contents

Abstract **v**

Contents **vii**

1 Introduction **1**
1.1 Models in quantum chemistry . . . 2
1.1.1 The Schrödinger equation . . . 3
1.1.2 The Hartree-Fock model . . . 5
1.1.3 The Gross-Pitaevskii equation . . . 6
1.2 Weighted Sobolev spaces . . . 7
1.3 Numerical methods . . . 8
1.3.1 Finite element methods . . . 8
1.3.2 A priori and a posteriori estimates . . . 10
1.4 Outline of the thesis . . . 11

2 The Mellin transform and weighted Sobolev spaces **15**
2.1 The Mellin transform in $\mathbb{R}_+$ . . . 16
2.1.1 Asymptotics near the origin . . . 20
2.2 Mellin transformation and weighted Sobolev spaces in conical domains 22
2.2.1 Mellin transform and homogeneous weighted Sobolev spaces in a cone . . . 22
2.2.2 Non homogeneous weighted Sobolev spaces . . . 24
2.2.3 Relationship between homogeneous and non homogeneous spaces 28
2.3 Elliptic operators in conical domains . . . 31
2.3.1 Regularity results for model operators . . . 31
2.3.2 Results for more general operators . . . 33

3 The hp discontinuous Galerkin method **35**
3.1 The discontinuous Galerkin method . . . 37
3.1.1 A general framework . . . 37
3.1.2 Interior penalty methods . . . 38
3.2 The discontinuous hp SIP method for problems with point singularities 39
3.2.1 Bilinear form and mesh-dependent weighted norms . . . 41

3.2.2 Continuity in weighted norms and coercivity 42
3.3 Approximation results 45

4 Regularity in weighted Sobolev spaces for linear problems 49
4.1 Introduction and presentation of the results 49
4.2 Notation and statement of the problem 51
4.2.1 Weighted Sobolev spaces 51
4.2.2 Statement of the problem 52
4.3 Regularity of the solution 54
4.4 Bounds on the Green function 58
4.5 Conjecture of pointwise convergence of the hp dG method 61
4.5.1 Introduction of g and ρ 62
4.5.2 Introduction of $\tilde{g}$ and $\tilde{\rho}$ 63
4.5.3 Local estimates 64
4.5.4 A priori estimates on the $\mathcal{D}^1_\gamma(\Omega)$ norms of $g - g_\delta$ and $\tilde{g} - \tilde{g}_\delta$ 67
4.5.5 Numerical results 73

5 Analysis of the hp dG method for elliptic linear eigenproblems 79
5.1 Statement of the problem and notation 80
5.1.1 Interior penalty method 81
5.2 Non pollution and completeness of the discrete spectrum and eigenspaces 83
5.2.1 Non pollution of the spectrum 83
5.2.2 Eigenspaces and completeness of the spectrum 85
5.3 Convergence of the eigenfunctions and eigenvalues 88
5.3.1 Convergence of the eigenvalues 88
5.3.2 Convergence of the eigenvalues for the SIP method 90

6 Numerical results for the linear eigenvalue problem 93
6.1 Two dimensional case 94
6.1.1 Analysis of the results 100
6.1.2 Detailed tables of the errors 101
6.2 Three dimensional case 105
6.2.1 Analysis of the results 107
6.2.2 Detailed error tables 109

7 Weighted analytic estimates for nonlinear eigenproblems 111
7.1 Local elliptic estimate 112
7.2 Weighted interpolation estimate 115
7.3 Nonlinear Schrödinger 116
7.4 Hartree-Fock 122

8 Analysis of the hp dG method for nonlinear eigenproblems **125**
8.1 Statement of the problem and notation . . . 127
8.1.1 Functional setting . . . 127
8.1.2 Numerical method . . . 128
8.1.3 Statement of the problem . . . 130
8.2 A priori estimates . . . 132
8.2.1 Continuity and coercivity . . . 133
8.2.2 Estimates on the adjoint problem . . . 136
8.2.3 Basic convergence . . . 137
8.2.4 Pointwise convergence . . . 140
8.2.5 Convergence revisited . . . 144
8.2.6 Exponential convergence . . . 148
8.3 Convergence of an iterative scheme . . . 149
8.4 Asymptotic analysis near the singularity . . . 152
8.4.1 Asymptotics of the solution to the Gross-Pitaevskii equation . . . 154

9 Numerical results for nonlinear eigenvalue problems **157**
9.1 Nonlinear Schrödinger . . . 157
9.1.1 Two dimensional case . . . 162
9.1.2 Detailed tables of the errors . . . 162
9.1.3 Three dimensional problem . . . 166
9.1.4 Detailed error tables . . . 168
9.2 Hartree Fock equation . . . 170
9.2.1 Practical implementation and the issue of sparsity . . . 170
9.2.2 Preliminary results for test cases . . . 173

Bibliography **177**

Chapter 1

Introduction

In this thesis, we study the approximation of linear and nonlinear eigenvalue problems with singular potentials. Many problems, arising from quantum physics and quantum chemistry, consist in the search of the *ground state* of the system, i.e., the constrained minimization of a certain energy functional E over a space X

$$\min \{E(u), u \in X\}.$$

This problem can be reformulated via Euler-Lagrange's equation as an eigenvalue problem

$$Au = \lambda u,$$

with some constraint on the norm of u and where A is a — potentially nonlinear — operator. Broadly speaking, in problems arising from quantum chemistry the function u is related to the wave function of the particles in the system; therefore, the energy always contains a kinetic term, which translates into a Laplacian in the eigenvalue problem. Furthermore, when dealing with full-electronic computations, i.e., when all the electrons are taken into account, the interaction between the nuclei and the electrons is modeled through a potential with singularities at the positions of the nuclei. For this reasons, the problems we consider belong to the class of elliptic eigenvalue problems with singular potentials. A more detailed presentation of the problems will be given later in this introduction.

With the exception of a small number of simple problems, the analytical solution to the eigenvalue problems is not known, thus there is the need to resort to numerical methods to find approximations of the energies and of the eigenfunctions. Generally speaking, a numerical method's rate of convergence depends on the regularity of the exact solution it is trying to approximate. In the context of problems with singular potential, the regularity of solution is best determined in weighted Sobolev spaces, where one can "isolate" and weight the behavior of the function at the singular point, while still exploiting the full regularity of the solution outside of singularities. The analysis of weighted Sobolev spaces has a long tradition connected to the study of elliptic problems

in non convex polyhedral and polygonal domains, and one of the goals of this thesis consists in extending it to problems with singular potentials.

The properties mentioned above of the solutions to elliptic problems with point singularities give a strong indication as to how to design a numerical method to approximate them. The basic idea is that we wish to exploit the twofold nature of the functions — smooth in parts of the domains, non-smooth in other parts, but with a controlled growth — in order to construct the numerical scheme. To do so, the method has to provide a good approximation for function with low regularity near the singularities, while different techniques can be used far from them. In the context of finite element methods, this goal is accomplished by hp finite element methods, which combine the accuracy of a classical finite element approximation near the singularity of the potential, where element sizes decrease geometrically, with the spectral convergence of a high polynomial degree approximation far from the singular points.

In practice, we construct the hp finite element approximation in a discontinuous Galerkin framework, meaning that we do not impose a continuity requirement over the functions in the discrete space that we use. In the thesis, we prove some *a priori* estimates on the convergence of the discontinuous hp method, which are particularly new in the context of nonlinear eigenvalue problem. Finally, we perform and analyze numerical tests for the different subjects of our theoretical analysis, whose goal is threefold: they confirm the theoretical results, they allow for the estimation of components of the error that are not taken into consideration in the theoretical analysis, and they provide an indication on the behavior of the method for more general cases.

In the following, we give an overview of the themes we have introduced. In Section 1.1, we introduce some *ab initio* models arising in quantum chemistry and physics on which the analysis will be based, with a brief overview of how they arise from physical models. Then, in Section 1.2 we discuss the relevance of weighted spaces in our case; this subject will be treated in more detail in Chapter 2. A brief introduction to numerical methods is given in Section 1.3 — the hp discontinuous Galerkin is fully analyzed later in Chapter 3. We then proceed to give an overview of the results presented in this thesis, in Section 1.4.

1.1 Models in quantum chemistry

We consider here models in *ab initio* quantum chemistry, i.e., models that describe the state of a system on the basis of first principles. The first of such models we introduce is the Schrödinger equation [Sch26], of capital importance from a theoretical point of view, though, as we will see, difficult to approach from the computational side. Our exposition will be limited to *non relativistic* models, which perform well when relativistic effects are negligible, e.g., for lighter atoms. We will also suppose that the Hamiltonians we consider have no effect on spin, therefore we will omit spin variables from the discussion. For a thorough presentation of electronic structure theory and its numerical treatment, see [SO12] and [CLM06; Le 03].

1.1.1 The Schrödinger equation

Description of a system

Consider a system consisting of M nuclei and N electrons: under the quantum formalism, the system is completely described at a time t by the wave function

$$\Psi(t, R_1, \dots, R_M; r_1, \dots, r_N), \tag{1.1}$$

with values in $\mathbb{C}$ and where $R_i \in \mathbb{R}^3$, $i = 1, \dots, M$ represent the positions of the nuclei and $r_i \in \mathbb{R}^3$, $i = 1, \dots, N$ represent the positions of the electrons. In order for (1.1) to be the wave function of a system, it needs to have the following properties.

Property 1. *The L^2 norm of Ψ has to be unitary, i.e.,*

$$\|\Psi(t, \cdot)\| = 1.$$

Here, the norm $\|\Psi(t, \cdot)\|^2$ is the integral over all R_i and over all r_i of the square of the wave function

$$|\Psi(t, R_1, \dots, R_M; r_1, \dots, r_N)|^2. \tag{1.2}$$

This makes (1.2) *a probability density.*

Property 2. *The wave function is symmetric with respect to the exchange of two identical bosons.*

Property 3. *The wave function is antisymmetric with respect to a change of the coordinates of the electrons, i.e., such that*

$$\Psi(t, , R_1, \dots, R_M; r_1, r_2, \dots, r_N) = \varepsilon(\sigma)\Psi(t, , R_1, \dots, R_M; r_{\sigma(1)}, r_{\sigma(2)}, \dots, r_{\sigma(N)}),$$

for any permutation σ and where $\varepsilon(\sigma)$ is the sign of the permutation.

Note that Properties 2 and 3 are consequences of the identity of bosons and fermions, respectively.

The Born-Oppenheimer approximation and the electronic Hamiltonian

The first approximation we operate on the wave function (1.1) is the Born-Oppenheimer approximation [BO27]. The mass of a proton is three orders of magnitude bigger than that of an electron, hence a classical approximation in quantum chemistry consists in fixing the position of the nuclei and on solving a minimization problem for the so called electronic Hamiltonian. This is the case in which the problem of *electronic structure calculation* is set. We will place ourselves in this setting: let us fix R_i, $i = 1, \dots, M$, and consider therefore a general time-independent wave function, that, abusing notation, we still write as $\Psi(r_1, \dots, r_N)$. Thanks to Property 1,

$$\Psi \in \bigotimes_{i=1}^{N} L^2(\mathbb{R}^3; \mathbb{C}).$$

Using Property 3, we can restrict the space to the wave functions that are antisymmetric with respect to permutations, i.e.,

$$\Psi \in \bigwedge_{i=1}^{N} L^2(\mathbb{R}^3;\mathbb{C}) = \mathcal{H}, \tag{1.3}$$

where $\bigwedge$ is the antisymmetrized tensor product.

Suppose now that the nuclei have charge Z_k, $k = 1, \dots, M$. The electronic Hamiltonian associated to the problem of electronic structure calculation (recall that we have fixed the nuclei at the positions R_k, $k = 1, \dots, N$) is given by

$$H = \sum_{i=1}^{N} \left(-\frac{1}{2}\Delta_{r_i} - \sum_{k=1}^{M} \frac{Z_k}{|r_i - R_k|} + \sum_{i<j\leq N} \frac{1}{|r_i - r_j|} \right). \tag{1.4}$$

The Hamiltonian above contains, in the order in which they are written, a kinetic term, the Coulomb interaction between electrons and nuclei, and the Coulomb interaction between electrons. The electronic minimization problem thus reads

$$E = E(R_1, \dots, R_M) = \inf \left\{ \langle \Psi, H\Psi \rangle, \Psi \in \mathcal{H} : \|\Psi\| = 1 \right\}. \tag{1.5}$$

Remark 1. *The minimization in* (1.5) *can also be written as*

$$E = \inf \left\{ \langle \Psi, H\Psi \rangle, \Psi \in \bigwedge_{i=1}^{N} H^1(\mathbb{R}^3;\mathbb{R}) : \|\Psi\| = 1 \right\}. \tag{1.6}$$

The restriction of the space to the product of H^1 spaces is required by the necessity that all terms in the energy functional be well defined, while the restriction to functions with values in $\mathbb{R}$ is due to the fact that the imaginary and real part of a wave function are treated independently by the Hamiltonian.

Taking the Euler-Lagrange equation of (1.5), we obtain the Schrödinger electronic equation

$$H\Psi = E\Psi. \tag{1.7}$$

Equation (1.7) is a linear eigenvalue problem; were one to find a solution, then Ψ would represent the electronic *ground state* of the system, and E would be its energy. From the computational point of view, this is a task of extraordinary difficulty even for problems of rather small size, due to the high dimensionality of the space where it is set. Suppose we were to subdivide a cube in the space $\mathbb{R}^d$ into a grid of n equispaced points in every direction: then we would need n^d points, and every point would be the neighbor of 2^d other points. This exponential growth, belonging to what is often called "the curse of dimensionality", indicates how approximation in higher dimensional spaces is computationally onerous. One of the main goals of quantum chemistry is therefore the development of approximations of the electronic Schrödinger equation (1.7), which

are set in the physical space: this comes at the price of the introduction of a nonlinearity. The Hartree-Fock equation is one of such models, and it is the subject of the next section.

1.1.2 The Hartree-Fock model

The basic idea of the Hartree-Fock model lies in the restriction of the space in which the search for a minimizer happens. Instead of considering the whole space

$$\bigwedge_{i=1}^{N} H^1(\mathbb{R}^d;\mathbb{R})$$

considered in (1.6), we restrict ourselves to the space of determinants

$$\Psi(x_1,\dots,x_N) = \frac{1}{\sqrt{N!}}\begin{vmatrix} \psi_1(x_1) & \psi_2(x_1) & \dots & \psi_N(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \dots & \psi_N(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(x_N) & \psi_2(x_N) & \dots & \psi_N(x_N) \end{vmatrix}, \tag{1.8}$$

where $\psi_i(x) \in H^1(\mathbb{R}^d;\mathbb{R})$. Formulation (1.8) is known as a *Slater determinant*. It is easy to see how such a wave function Ψ respects Property 3. Furthermore, we can impose that

$$\int_{\mathbb{R}^3} \psi_i\psi_j = \delta_{ij},$$

so that Property 1 is satisfied.

Using (1.8), we can compute $\langle \Psi, H\Psi\rangle$, which gives

$$\begin{aligned} E^{\mathrm{HF}}(\psi_1,\dots,\psi_N) = \sum_{i=1}^{N}\int_{\mathbb{R}^3}|\nabla\psi_i|^2 + \int_{\mathbb{R}^3} V\rho_\Psi + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho_\Psi(x)\rho_\Psi(y)}{|x-y|} \\ - \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\tau_\Psi(x,y)}{|x-y|} \end{aligned}$$

where

$$\tau_\Psi(x,y) = \sum_{i=1}^{N}\psi_i(x)\psi_i(y) \qquad \rho_\Psi = \tau_\Psi(x,x)$$

and

$$V(x) = \sum_{k=1}^{M}\frac{Z_k}{|x-R_k|}.$$

The electronic minimization problem associated with the Hartree-Fock energy thus reads

$$\inf\left\{E^{\mathrm{HF}}(\psi_1,\dots,\psi_N), \psi_i \in H^1(\mathbb{R}^3;\mathbb{R}) : \int_{\mathbb{R}^3}\psi_i\psi_j = \delta_{ij}\right\}. \tag{1.9}$$

With the usual procedure and using an orthogonal transformation to partially decouple the equations, we can write (1.9) as the nonlinear eigenvalue problem of finding $\Psi = \{\psi_1, \dots, \psi_N\}$ and $\{\varepsilon_1, \dots, \varepsilon_N\} \in \mathbb{R}^N$

$$\begin{aligned} \mathcal{F}_\Psi \psi_i &= \varepsilon_i \psi_i && \text{for all } i = 1, \dots, N \\ (\psi_i, \psi_j) &= \delta_{ij} && \text{for all } i, j = 1, \dots, N, \end{aligned} \tag{1.10}$$

where $(\cdot, \cdot)$ denotes the scalar product in $L^2(\mathbb{R}^3)$ and $\mathcal{F}_\Psi$ is the *Fock operator* defined by

$$\mathcal{F}_\Psi \varphi = -\frac{1}{2}\Delta\varphi + V\varphi + \left(\rho_\Psi \star \frac{1}{|x|}\right)\varphi - \sum_{j=1}^{N}\left((\psi_j\varphi) \star \frac{1}{|x|}\right)\psi_j.$$

Equation (1.10) is known as the *Hartree-Fock equation*.

1.1.3 The Gross-Pitaevskii equation

We conclude this section by introducing a model that does not come from the domain of quantum chemistry, but that is interesting from the point of view of theoretical analysis. The Gross-Pitaevskii equation, also known as *nonlinear Schrödinger equation* in the approximation of Bose-Einstein condensates [PS03; LSY00]. Its minimization problem reads

$$\inf\left\{E^{\mathrm{GP}}(u) : u \in H^1(\mathbb{R}^3) : \|u\|_{L^2(\mathbb{R}^3)} = 1\right\}, \tag{1.11}$$

where

$$E^{\mathrm{GP}}(u) = \frac{1}{2}\int_{\mathbb{R}^3} |\nabla u|^2 + \frac{1}{2}\int_{\mathbb{R}^3} Vu^2 + \frac{1}{4}\int_{\mathbb{R}^3} u^4.$$

The nonlinear eigenvalue problem arising from the optimality conditions of (1.11) is therefore given by: find $(u, \lambda) \in H^1(\mathbb{R}^3, \mathbb{R})$ such that

$$\begin{aligned} -\Delta u + Vu + |u|^2 u &= \lambda u \\ \|u\| &= 1, \end{aligned} \tag{1.12}$$

where $\|\cdot\|$ is the L^2 norm in the whole space.

Our interest in the Gross-Pitaevskii equations stems less from its relevance as a physical model and more from the "nonlinear Schrödinger" point of view, in that we use it as a model of the more complex systems. In our analyses, we will consider more general potentials and nonlinearities. In practice, we consider a class of elliptic problems with singular potential and local nonlinearities; in doing so, in some specific cases we end up treating the Gross-Pitaevskii equation, while still considering potentials that do not necessarily have a meaning in the context of Bose-Einstein condensates.

1.2 Weighted Sobolev spaces

In this section, we give a brief overview of weighted Sobolev spaces. This subject is treated in greater detail in Chapter 2, where weighted Sobolev spaces are derived from their Mellin characterization and where the relationship between weighted regularity and Mellin transformation is fully analyzed. Here, we only wish to give a feeling of the role of weighted spaces in problems with singular potentials and to provide some context with regards to the existing literature.

Weighted Sobolev spaces are also called *Kondrat'ev spaces*, from the seminal work [Kon67], or *Kondrat'ev-Babuška spaces*, from their application to the finite element approximation of Babuška and coworkers [GB86a; GB86b; GB86c; GB86d; GB86e].

Consider a point $\mathfrak{c}$ inside a domain $\Omega \subset \mathbb{R}^d$ or on its boundary $\partial\Omega$ and indicate by r the function of the distance from this point, i.e.,

$$r = r(x) = |x - \mathfrak{c}|,$$

where $|\cdot|$ is the euclidean norm of a vector in $\mathbb{R}^d$. Then, the classical weighted Sobolev space $\mathcal{K}^s_\gamma(\Omega)$ for integer s and real γ is composed by those functions that have bounded norm

$$\|u\|_{\mathcal{K}^s_\gamma(\Omega)} = \left(\sum_{|\alpha| \le s} \|r^{|\alpha|-\gamma} \partial^\alpha u\|^2_{L^2(\Omega)} \right)^{1/2}. \tag{1.13}$$

Here, $\alpha = (\alpha_1, \dots, \alpha_d)$ is a multi index, and we write $|\alpha| = \alpha_1 + \cdots + \alpha_d$, and $\partial^\alpha = \partial^{\alpha_1}_{x_1} \dots \partial^{\alpha_d}_{x_d}$.

We remark that the notation used for weighted Sobolev spaces is unluckily very fragmented. We detail the relationship between our notation and what is used in the literature, while taking the occasion to introduce relevant works on weighted Sobolev spaces and on the regularity of problems in domains with singularities. We use the same notation as Nistor and coworkers [AN07; HNS08; LN09; MN10; LMN10; AN15]. With respect to the work of Costabel, Dauge, Nicaise and coworkers [Nic97; CDS05; CDN10a; CDN10b; CDN14], where the space $\mathsf{K}^s_\gamma(\Omega)$ is defined, we have

$$\mathcal{K}^s_\gamma(\Omega) = \mathsf{K}^s_{-\gamma}(\Omega).$$

The same spaces are used in [SSW13b; SSW13a; SSW16], though the spaces are denoted $M^m_\beta(\Omega)$ there and defined in general for polyhedra. If we suppose that a single "corner" singularity is present, we have again

$$\mathcal{K}^s_\gamma(\Omega) = M^s_{-\gamma}(\Omega).$$

In the work by Maz'ya and coworkers, see [MP78; KMR97; KM99; MNP00; MR02; MR10] the space $V^s_{2,\beta}(\Omega)$ is used, and it compares to the one defined in (1.13) as

$$\mathcal{K}^s_\gamma(\Omega) = V^s_{2,s-\gamma}(\Omega).$$

Finally, considering [Kon67] and the space $\overset{0}{W}{}^k_\alpha(\Omega)$ defined there, we have

$$\mathcal{K}^s_\gamma(\Omega) = \overset{0}{W}{}^s_{2(s-\gamma)}(\Omega).$$

Our notation is motivated by the fact that we have

$$\mathcal{K}^{s'}_\gamma(\Omega) \subset \mathcal{K}^s_\gamma(\Omega) \text{ if } s' \geq s \qquad \text{and} \qquad \mathcal{K}^s_{\gamma'}(\Omega) \subset \mathcal{K}^s_\gamma(\Omega) \text{ if } \gamma' \geq \gamma,$$

and that, for example,

$$-\Delta : \mathcal{K}^s_\gamma(\Omega) \to \mathcal{K}^{s-2}_{\gamma-2}(\Omega)$$

so that there is a certain symmetry between the notation of the regularity and weight indices.

Alongside the space defined by the norm (1.13), that we will call *homogeneous* weighted Sobolev space, we will also consider the *non homogeneous* version, normed by

$$\|u\|_{\mathcal{J}^s_\gamma(\Omega)} = \left(\sum_{|\alpha| \leq s} \|r^{\max(|\alpha|-\gamma,\rho)} \partial^\alpha u\|^2_{L^2(\Omega)} \right)^{1/2} \quad \text{for any } \rho \in (-d/2, s-\gamma]. \tag{1.14}$$

In Chapter 2 it will be recalled how the norms are equivalent for varying ρ. The spaces normed by (1.14) have emerged in the context of the study of solutions to elliptic problems with non trivial development at non regular parts of the boundary; in this thesis we will show how they constitute appropriate spaces for the analysis of problems with singular potentials.

1.3 Numerical methods

All the eigenvalue problems presented in Section 1.1 need, in practice, to be solved via a numerical method. Generally speaking, numerical methods consist in the reduction of a continuous, infinite dimensional problem into a discrete, finite dimensional one, whose solution should be an approximation of the exact one.

In this thesis, the focus is on the hp discontinuous Galerkin (dG) method, which belongs to the wider class of finite element (FE) methods.

1.3.1 Finite element methods

Finite element methods have been developed in the second half of the 20th century, initially to solve structural mechanics problems, and have been extended to almost every domain of mathematical physics. The FE method draws largely on Galerkin's earlier work [Gal15] — and others': even though the formulation *Galerkin method* is customary, the same ideas are expressed, e.g., in [Rit09a; Rit09b]. Given a linear operator $L : X \to X'$

and a problem in its weak form of finding $u \in X$ such that

$$\langle Lu, v\rangle = \langle f, v\rangle \text{ for all } v \in X, \tag{1.15}$$

the basic principle of the Galerkin method consists in replacing (1.15) with the problem of finding a u_δ in a finite dimensional space X_δ such that

$$\langle Lu_\delta, v_\delta\rangle = \langle f, v_\delta\rangle \text{ for all } v_\delta \in X_\delta. \tag{1.16}$$

By choosing a basis for X_δ, (1.16) problem can be written as an algebraic linear system, that can be solved with the classical methods of (numerical) linear algebra. Years later, Courant, in [Cou43] — comparing finite difference schemes with what he calls the Rayleigh-Ritz method — writes:

> considering [...] only functions which are linear in the meshes of a subdivision of our net into triangles formed by diagonals of the squares of the net. [...] Such an interpretation suggests a wide generalization which provides great flexibility and seems to have considerable practical value. [...] we may consider from the outset any polyhedral surfaces with edges over an arbitrarily chosen (preferably triangular) net.

This idea is at the basis of finite element methods: the domain where the problem is set is partitioned into elements, and a basis of polynomials (linear functions in Courant's idea) is constructed into each one of them. This provides a basis for X_δ: if the basis is chosen to be continuous, we obtain the classical finite element method. Suppose that the domain where the original problem is set is $\Omega \subset \mathbb{R}^d$, that we have a collection of elements $\mathcal{T}$ that tessellates Ω, and that, for each element $K \in \mathcal{T}$, there exists an affine transformation Φ such that $\Phi(\hat{K}) = K$ for some reference element $\hat{K}$: then the discrete space of the classical FE method reads

$$X_\delta = \left\{ v_\delta \in C^0(\Omega) : v_{\delta|_K} \circ \Phi_K \in \mathbb{P}_p(\hat{K}), \forall K \in \mathcal{T} \right\}, \tag{1.17}$$

for a fixed polynomial degree p. For a thorough presentation of how those ideas go back to Euler, see [GW12]. Among the many classical references on FE methods in the literature, we point to [CL91; CL96; BS03; EG04; SF08; Qua17].

hp dG finite element methods

In the discontinuous Galerkin framework, we drop the requirement that the basis functions be continuous at inter-element interfaces: the space (1.17) is replaced by

$$X_\delta = \left\{ v_\delta \in L^2(\Omega) : v_{\delta|_K} \circ \Phi_K \in \mathbb{P}_p(\hat{K}), \forall K \in \mathcal{T} \right\}.$$

This requires that we change the discrete bilinear form (1.16): how to do so, while preserving the convergence of the numerical solution, will be detailed in Chapter 3.

Furthermore, in the hp version of the FE method, the polynomial degree is no longer uniform over the whole domain, but can vary between elements, and the same goes for the mesh size, which can go to zero in some parts of the domain, while staying bounded from below in others. Different approaches in the design of the hp spaces exist. In our case, we are close to the original idea given in [GB86a], and the refinement of the space is chosen based on information that we have from the analysis of the problem, *a priori*. We know, indeed, that the solutions to elliptic problems with singular potentials belong to the weighted Sobolev spaces we have introduced: therefore, a space isotropically refined towards the singularity, and with a polynomial degree that grows with the distance from the singularity contains function that converge with exponential rate towards the solution. Nonetheless, in a complex problem with multiple nuclei and where the approximation away from the singularities gets more complicated, one may need to chose whether to perform h- or p- refinement (i.e., whether to reduce the mesh size or to increase the polynomial degree). This is the focus of the *adaptive hp* finite element method, see [Hou05; HSW07b; GGO11; GH12; CV13], which uses *a posteriori* error estimators to choose where and how to refine the space.

We point again to Chapter 3 for an in depth presentation of the hp dG method we consider in our analysis. We continue here by giving a short overview of the concept of *a priori* and *a posteriori* estimates.

1.3.2 A priori and a posteriori estimates

In this thesis, we are mainly concerned with *a priori* estimates. We shall give here a general introduction to this concept, followed by an equally general outline of the field of *a posteriori* estimation.

A priori estimates

Given an exact solution u to the problem under analysis and a numerical approximation $u_\delta \in X_\delta$, with $\dim(X_\delta) = N$, we wish to show that, for an appropriate norm $\|\cdot\|$, there exists a $C > 0$ and a function f such that

$$\|u - u_\delta\| \leq C f(N).$$

The first thing we are interested in proving is that $f(N)$ goes to zero as N grows to infinity: this implies the convergence of the numerical method. The second estimate of interest is *how fast* $f(N)$ goes to zero. In general, we are able to show that there exists a N_0 such that for all $N > N_0$, f has a specific form. This will be influenced, among others, by the regularity of the solution and by the quality of the approximation provided by the space X_δ (which is itself strongly dependent on the regularity of the exact solution). In classical finite elements, the estimates are of the form

$$\|u - u_\delta\|_{H^1(\Omega)} \leq C N^{-k/d} \|u\|_{H^{k+1}(\Omega)}.$$

where $k \leq p$ (we have introduced p in (1.17)) and where we are supposing $u \in H^{k+1}(\Omega)$. When using an isotropically refined hp finite element method, under some conditions on the regularity in weighted Sobolev spaces of the exact solution, we obtain

$$\|u - u_\delta\|_{H^1(\Omega)} \leq Ce^{-bN^{1/(d+1)}}. \tag{1.18}$$

A posteriori estimates

A somehow complementary approach is taken when developing *a posteriori* estimates. Those are estimates that rely on computable quantities, and are generally expected to be valid even before the asymptotic regime (i.e., before the unknown N_0 used in the description of a priori estimates). A posteriori estimates can be used to adaptive refine the space, as already mentioned, or to avoid useless computations, by balancing the different error terms arising in an approximation. A posteriori error estimation is out of the scope of this thesis; nonetheless, one of the advantages of using a finite element method lies in the fact that one can draw from a great set of tools for a poteriori estimation. In the context of hp and dG methods, we point to [HSW07a; HSW07b; Wih07; GH12; GGO13; EV15; DEV16]. For eigenvalue problems approximated by finite elements, in addition to what has already been cited, see also [Can+17]. Concerning a posteriori error estimation on problems related to quantum chemistry but with non-FE basis functions, see [Can+14; Can+16; DM17].

1.4 Outline of the thesis

In this section, we give an overview of the structure of thesis, and we outline the results obtained in the different parts. As already mentioned, in Chapter 2 we introduce the Mellin transform and derive homogeneous and non homogeneous weighted Sobolev spaces starting from their Mellin characterization. Most of the properties of the Mellin transform of functions in weighted Sobolev spaces will be of great importance in the sequel, both as tools in the analysis and as means of understanding the behavior of the numerical approximations. We conclude the chapter with some classical results on regularity in homogeneous weighted Sobolev spaces, on which we will construct later results in non homogeneous spaces.

In Chapter 3 the hp dG method is introduced. We define some non standard weighted mesh dependent norms, which are well suited for the analysis in spaces of functions with isotropic singularities. The continuity of the dG formulation is then shown in this norms, in the general case of dual L^p – L^q continuity. Finally, we adapt a classical approximation results to the norms we consider. In a reduced form, we find that, in an element $K \in \mathcal{T}$ for any function $v \in \mathcal{J}^{s+1}_{\gamma'}(K)$, for any $\gamma < \gamma'$,

$$\inf_{v_\delta \in X_\delta} |||v - v_\delta|||_{\mathcal{G}^2_\gamma(K)} \leq Ch_K^{\gamma'-\gamma} p_K^{-s+1/2} \|v\|_{\mathcal{J}^{s_j+1}_{\gamma'}(K)}, \tag{1.19}$$

with the constraint that $s = 1$ if the singularity belongs to a corner of the element K.

The norm $|||\cdot|||_{\mathcal{G}^2_\gamma(K)}$ is a mesh dependent norm bounded by the norm $\|\cdot\|_{\mathcal{K}^2_\gamma(K)}$. Let us introduce a subspace of the non homogeneous weighted Sobolev spaces, given by

$$\mathcal{J}^\varpi_\gamma(\Omega) = \left\{ v \in \bigcap_k \mathcal{J}^k_\gamma(\Omega) : \exists C, A > 0 : \|v\|_{\mathcal{J}^k_\gamma(\Omega)} \leq CA^k k! \right\}.$$

The approximation estimate (1.19) crucially implies that, if we use an hp method isotropically and geometrically graded towards the singularity (see Figure 3.1) and with polynomial degrees increasing linearly, then

$$u \in \mathcal{J}^\varpi_\gamma(\Omega), \gamma > 1 \implies \inf_{v_\delta \in X_\delta} |||v - v_\delta|||_{\mathcal{G}^2_1(\mathcal{T})} \leq Ce^{-bN^{1/(d+1)}}. \tag{1.20}$$

This means that if we can control with "analytic type" estimates the norms of the solution, then the best approximation in the hp space converges exponentially. Due to the quasi optimality results we obtain in later chapters, this will also hold for numerical solutions.

In the following Chapter 4 we consider the problem of elliptic regularity in non homogeneous spaces for problems with potentials, and show that, under some conditions on the potential and on the domain (and supposing here Dirichlet boundary conditions for ease of notation), the operator

$$-\Delta + V : \mathcal{J}^k_\gamma(\Omega) \to \mathcal{J}^{k-2}_{\gamma-2}(\Omega) \times \mathcal{J}^{k-3/2}_{\gamma-3/2}(\partial\Omega)$$

is an isomorphism, for all $k \geq 1$ and for $\gamma = \begin{cases} (-1, \varepsilon) \setminus \{0\} & \text{if } d = 3 \\ [0, \varepsilon) & \text{if } d = 2 \end{cases}$, with $\varepsilon > 0$ determined by the potential. We also give some bounds on the Green function $G(x, y)$ associated with the operator: far from $\{x = y\}$, the Green function's behavior depends in the singularities.

In Chapter 5 the linear eigenvalue problem

$$(-\Delta + V)u = \lambda u$$

is considered. We recall the convergence result from [ABP06] and concentrate in particular on the symmetric version of the interior penalty method to show that, when

$$\|u - u_\delta\|_{\mathrm{DG}} \leq Ce^{-bN^{1/(d+1)}}, \tag{1.21}$$

then

$$|\lambda - \lambda_\delta| \leq Ce^{-2bN^{1/(d+1)}}, \tag{1.22}$$

where we have denoted by u_δ and λ_δ the numerical eigenfunction and eigenvalue obtained with the SIP method. The theoretical estimate is tested in Chapter 6, where, apart from validating the theory, we also investigate the effect of errors that we do not consider in the analysis. The main sources of numerical error, apart from the approximation error introduced by the method, are the quadrature error and the error arising from

the numerical linear algebra schemes. The former has mainly an effect on the rate of convergence of the eigenvalue error, while the latter affects all terms. Finally, the numerical experiments on simple test cases give an idea on how to best approximate more complex ones. We investigate in particular the polynomial slope parameter, and see how for different potentials, different slopes are optimal. This can be furthermore combined with the asymptotic analysis near the singularity that follows quite naturally from a Mellin formulation of the problem.

We then move on to non linear problems, and in Chapter 7 the regularity of the solution to the non linear eigenvalue problem

$$(-\Delta + V + |u|^{k-1})u = \lambda u \tag{1.23}$$

is investigated. After some preliminary inequalities in weighted Sobolev spaces, using techniques from [Dal+12] we are able to show that, under some conditions on the potential and if $k = 2, 3, 4$,

$$u \in \mathcal{J}^{\varpi}_{\gamma}(\Omega), \tag{1.24}$$

for the same γ as in the linear case. In addition, with a small modification to the proof, this extends to the functions in the Slater determinant of the Hartree-Fock equation: if the wave functions ψ_i, $i = 1, \dots, N$ are solutions to (1.10), under the customary hypotheses on the potential, we have

$$\psi_i \in \mathcal{J}^{\varpi}_{\gamma}(\Omega), \quad i = 1, \dots, N.$$

After this regularity result, we consider, in Chapter 8, a slightly more general equation in the form of the non linear Schrödinger equation

$$(-\Delta + V + f(u^2))u = \lambda u,$$

where we impose some conditions on the nonlinearity f, see equations (8.11a) to (8.11d). The first step in the analysis of this problem consists in proving the convergence of the solution obtained with dG methods. Compared to continuous finite elements, in the analysis in discontinuous spaces some difficulties arise due to the *non conformity* of the method (i.e., the fact that $X_\delta \not\subset X$). Furthermore, quasi optimality of the dG solution is proven. Those results do not depend specifically on the choice of the hp space. The doubling of the convergence rate, in the nonlinear case, is shown using techniques that involve the convergence of the eigenfunction in $L^\infty(\Omega)$. This is proven easily, under some regularity requirements, in the h and p versions of the finite element method, but has to be conjectured for hp methods. In addition, we note that, when the problem considered is of the type (1.23), then, thanks to (1.24) and (1.20), we can conclude with estimates of exponential convergence similar to (1.21) and (1.22). We conclude by studying an iterative scheme to treat the nonlinearity and with an asymptotic analysis of the behavior of the solution to the (linearized) problem near the singularity.

Finally, in Chapter 9, as we did in the linear case, we perform some numerical tests on the non linear eigenvalue problem, in two and three dimensions. We assess the

convergence of the method, and we test the interaction between different potentials and different discretization parameters.

Every chapter is as self-contained as possible and can be read independently, at least from the point of view of notation. When results from different parts of the thesis are used, they are often recalled.

Chapter 2

The Mellin transform and weighted Sobolev spaces

In this section, we introduce the Mellin transform and weighted Sobolev spaces. This coupling is not coincidental, as the Mellin transform is a fundamental tool in the construction of weighted Sobolev spaces, in the same way as the Fourier transform can be used to define classical Sobolev spaces in periodic domains.

Weighted Sobolev spaces are the more natural spaces in which we can define the regularity of solutions to elliptic problems in domains with singular points, as is the case, for example, of domains with corners in two dimensions, of conical domains in three dimensions, and — the focus of our interest — of equations with singular coefficients at isolated points. The role of the Mellin transform and of weighted Sobolev spaces in the analysis of elliptic problems with singular points has been introduced in the seminal paper [Kon67]; the results presented here are mostly based on [Jea79] for the general analysis of the Mellin transform, and on [KMR97; CDN10b] for the parts concerning conical domains and elliptic regularity. We also follow closely the exposition given in the latter three papers, while adapting it to our notation and trying to draw an explicit line that connects the more basic properties of the transform with advanced subjects in the analysis of elliptic problems.

We will be presenting two classes of weighted Sobolev spaces: the first and more straightforward type are the homogeneous spaces $\mathcal{K}^s_\gamma(\Omega)$, $s, \gamma \in \mathbb{R}$, whose norm for integer s and in a domain $\Omega \subset \mathbb{R}^d$ is given by

$$\|u\|_{\mathcal{K}^s_\gamma(\Omega)} = \left(\sum_{|\alpha| \leq s} \int_\Omega r^{-2\gamma+2|\alpha|} (\partial^\alpha u)^2 \right)^{1/2},$$

where r is the distance from the singular point. The second type of space we introduce

is the non homogeneous weighted Sobolev spaces, normed, *inter alia*, by

$$\|u\|_{\mathcal{J}^s_\gamma(\Omega)} = \left(\sum_{|\alpha| \leq s} \int_\Omega r^{2\max(-\gamma+|\alpha|,\rho)} (\partial^\alpha u)^2 \right)^{1/2},$$

for any $\rho \in (-d/2, s-\gamma]$. The spaces of the non homogeneous kind will be of great importance to our subsequent analysis, and one of goals of this section is that of outlining the basis on which the subsequent analysis will be built. The importance of the second kind of spaces arises from the fact that we are interested in functions with nontrivial expansion at the singularity, and the spaces $\mathcal{J}^s_\gamma(\Omega)$ are the more appropriate for this case. Consider, for example, a function that does not vanish at $r = 0$: this function belongs to $\mathcal{K}^s_\gamma(\Omega)$ only if $\gamma < d/2$; it can instead be in a space $\mathcal{J}^s_\gamma(\Omega)$ for bigger γ and this allows to better characterize its regularity.

In Section 2.1, we introduce the Mellin transform of a function in a quite general way and in $\mathbb{R}_+$. We start to outline the relationship between weighted spaces in the real space with the spaces of their Mellin transforms, and this allows for the definition of weighted spaces starting from the properties of the transformed spaces. We conclude the section by outlaying the fundamental relationship between the poles of the transform of a function and the asymptotic expansion at the singularity of the function.

In Section 2.2, we consider the case of conical domains, i.e., of domains that are the product of $\mathbb{R}_+$ and of smooth $d-1$ dimensional domains. We start by defining the homogeneous spaces in this setting; no big conceptual difference intervenes with respect to the case of $\mathbb{R}_+$, and the definitions follow similarly. The introduction of $\mathcal{J}^s_\gamma(\Omega)$ is not as straightforward, and involves the definition of a norm in the transformed space that remains bounded on poles of the transform that give rise to polynomials in the asymptotic expansion. This is done in Section 2.2.2 and followed, in Section 2.2.3, by the analysis of the relationship between the space $\mathcal{J}^s_\gamma(\Omega)$ and $\mathcal{K}^s_\gamma(\Omega)$.

We conclude with the application to elliptic problems of the tools introduced, in Section 2.3. We concentrate ourselves on the spaces $\mathcal{K}^s_\gamma(\Omega)$, since the non homogeneous case will be the focus of further analysis in the forthcoming chapters. We recall a fundamental elliptic regularity estimate, whose proof follows easily from the Mellin characterization of the spaces and of the norms.

2.1 The Mellin transform in $\mathbb{R}_+$

Consider a smooth function with compact support $u \in C^\infty_0(\mathbb{R}_+)$, with image in $\mathbb{R}$. We define the Mellin transform of u as

$$\hat{u}(\lambda) = (\mathcal{M}u)(\lambda) = \int_{\mathbb{R}_+} x^{-\lambda} u(x) \frac{dx}{x}. \tag{2.1}$$

Since u and its derivative have compact support in $(0, +\infty)$, $\hat{u}$ is an analytic function in $\mathbb{C}$. It can easily be shown, by integration by parts, that

$$\mathcal{M}(x\partial_x u)(\lambda) = -\lambda\hat{u}(\lambda). \tag{2.2}$$

Furthermore, there holds

$$\mathcal{M}(x^p u)(\lambda) = \hat{u}(\lambda - p). \tag{2.3}$$

The space of smooth compactly supported functions is too restrictive for practical applications and for the analysis of weighted Sobolev spaces, thus we wish to extend the Mellin transform, as defined in (2.1), to a wider class of functions. For any set $U \subset \mathbb{R}$, let $\mathcal{D}'(U)$ be the space of distributions on U (the dual space of the space of $C^\infty(U)$ functions with compact support). Let u be a distribution in $\mathcal{D}'(\mathbb{R}_+)$ that can furthermore be extended to a compact support distribution $\tilde{u}$ on $\mathbb{R}$. Then, the Mellin transform of u can be defined as the duality

$$(\mathcal{M}u)(\lambda) = \langle \tilde{u}, x_+^{-\lambda-1} \rangle, \tag{2.4}$$

where $x_+ = \max(x, 0)$ for $x \in \mathbb{R}$.

Remark 2. *The transformation* (2.4) *is well defined since it does not depend on the choice of* $\tilde{u}$ *and there exists an* $m \in \mathbb{N}$ *such that* $(\mathcal{M}u)(\lambda)$ *is holomorphic for all* $\operatorname{Re}\lambda < -m$*, see [Jea79].*

Remark 3. *We denote by* $\mathcal{E}_+$ *the subspace of distributions in* $\mathcal{D}'(\mathbb{R}_+)$ *that can be extended to distributions with compact support in* $\mathbb{R}$*; under definition* (2.4) *of the Mellin transform, properties* (2.2) *and* (2.3) *still hold for any* $u \in \mathcal{E}_+$*.*

The Mellin transforms of functions defined in $\mathcal{E}_+$ are therefore holomorphic in half planes of the type $\operatorname{Re}\lambda < \gamma$, $\gamma \in \mathbb{R}$. We can then introduce the space $\mathcal{H}_+$ of functions $\hat{f}$, holomorphic on a half plane and such that

$$|\hat{f}(\lambda)| \leq C(1 + |\lambda|)^m a^{\operatorname{Re}\lambda}$$

for some $m \in \mathbb{Z}$, $a > 0$ and for $\operatorname{Re}\lambda < \gamma$. Consider then $u \in \mathcal{E}_+$ such that the support of u is contained in $(0, a]$ for an $a > 0$: there exists then a continuous function f in $\mathbb{R}_+$ such that $f(x) = 0$ for $x \geq a$ and that $f^{(m)} = u$, [Jea79, Proposition 1.5], thus

$$\hat{u}(\lambda) = (-1)^m(\lambda + 1)\cdots(\lambda + m)\int_{\operatorname{supp}(f)} f(x)x^{-\lambda-m-1}dx.$$

We can then conclude that the Mellin transform of a function $u \in \mathcal{E}_+$ lies in $\mathcal{H}_+$ and define the inverse Mellin transform as

$$(\mathcal{M}^{-1}\hat{u})(x) = \frac{1}{2\pi i}\int_{\operatorname{Re}\lambda=\beta} x^\lambda \hat{u}(\lambda)d\lambda \tag{2.5}$$

where $\operatorname{Re}\lambda = \beta$ is the line parallel to the imaginary axis passing through the point $\beta \in \mathbb{R}$. We clearly suppose that $\beta < \gamma$, with γ defined as above, so that $\hat{u}$ is holomorphic on the

line $\operatorname{Re}\lambda = \beta$.

Having defined the Mellin transform and its inverse, we now restrict the space $\mathcal{E}_+$ to the functions that are relevant for the analysis of weighted Sobolev spaces. Let us introduce the norm

$$\|u\|^2_{\mathcal{K}^0_{1/2}} = \int_{\mathbb{R}_+} x^{-1} u(x)^2 dx.$$

We can then define the space

$$\mathcal{K}^0_{1/2} = \left\{ v \in \mathcal{E}_+ : \|v\|_{\mathcal{K}^0_{1/2}} < \infty \right\}.$$

The meaning of the notation we use will become clearer in the context of weighted Sobolev spaces.

As we have done before when considering functions in $\mathcal{E}_+$, we are interested in the image of the Mellin transformation applied to the space $\mathcal{K}^0_{1/2}$: we write $\mathcal{H}^0_0 = \mathcal{M}\mathcal{K}^0_{1/2}$. We recall some of the results on the relationship between $\mathcal{K}^0_{1/2}$ and $\mathcal{H}^0_0$ in the following lemma.

Lemma 1. *For any $u \in \mathcal{K}^0_{1/2}$, with $\hat{u} = \mathcal{M}u$,*

(i) $\hat{u}$ is holomorphic in the half-plane $\{\operatorname{Re}\lambda < 0\}$, and

(ii) for any $\beta < 0$ and for $a \in \mathbb{R}_+$ such that $\operatorname{supp}(u) \subset (0, a]$,

$$\int_{\operatorname{Re}\lambda = \beta} \hat{u}(\lambda)^2 d\lambda \leq C a^{-\beta}.$$

Furthermore, the converse is also true, i.e., if (i) and (ii) hold for a $\hat{u} \in \mathcal{H}^0_0$, then there exists $u \in \mathcal{E}^0_{1/2}$ such that $u = \mathcal{M}^{-1}\hat{u}$.

Finally, the function $\hat{u}_\beta : \xi \mapsto \hat{u}(\beta + i\xi)$ has a limit for $\beta \to 0^+$ and the Plancherel equality

$$\int_{\mathbb{R}_+} x^{-2\beta-1} u(x)^2 dx = \frac{1}{2\pi} \int_{\operatorname{Re}\lambda = \beta} \hat{u}(\lambda)^2 d\lambda \tag{2.6}$$

holds for all $\beta \leq 0$.

We can now consider, for a $\gamma \in \mathbb{R}$, the space

$$\mathcal{K}^0_\gamma = \left\{ v \in \mathcal{E}_+ : x^{-\gamma+1/2} v(x) \in \mathcal{K}^0_{1/2} \right\}$$

and write $\mathcal{H}^0_{\gamma-1/2} = \mathcal{M}\mathcal{K}^0_\gamma$. From (2.3) it follows that if $\hat{u} \in \mathcal{H}^0_\gamma$ then $\hat{u}(\lambda + \gamma) \in \mathcal{H}^0_0$. Then, (i) implies that $\hat{u}$ is holomorphic in the half-plane $\{\operatorname{Re}\lambda < \gamma\}$; furthermore, from (ii) we obtain that the function $\hat{u}_\beta : \xi \mapsto \hat{u}(\beta + i\xi)$ is in $L^2(\mathbb{R})$ for any $\beta < \gamma$. Finally, the norm on $\mathcal{K}^0_\gamma$ is given by

$$\|u\|^2_{\mathcal{K}^0_\gamma} = \int_{\mathbb{R}_+} x^{-2\gamma} u(x)^2 dx,$$

thus equation (2.6) shows that the Mellin transformation is an isomorphism between the spaces

$$\mathcal{K}^0_{\gamma+1/2} \to L^2(\{\mathrm{Re}\,\lambda = \gamma\}).$$

The spaces introduced thus far are, basically, weighted $L^2(\mathbb{R}_+)$ spaces of functions with compact support. Let us now introduce $\mathcal{H}^s_\gamma$ as the subspace of transforms $\hat{u} \in \mathcal{H}_+$ such that

$$(\lambda - \gamma - 1)^s \hat{u}(\lambda) \in \mathcal{H}^0_\gamma. \tag{2.7}$$

This implies that $\hat{u}$ is holomorphic on the half-plane $\{\mathrm{Re}\,\lambda < \gamma\}$ and that

$$\int_{\mathrm{Re}\,\lambda=\beta} (\lambda - \gamma - 1)^{2s} \hat{u}(\lambda)^2 d\lambda < \infty.$$

We can now define the homogeneous weighted Sobolev space

$$\mathcal{K}^s_\gamma = \mathcal{M}^{-1}\mathcal{H}^s_{\gamma-1/2}. \tag{2.8}$$

Lemma 2. *$u \in \mathcal{K}^s_\gamma$ with $s \in \mathbb{N}$ if and only if*

$$x^{j-\gamma}\partial^j_x u \in L^2(\mathbb{R}_+). \tag{2.9}$$

for all $j = 0, \dots, s$.

Proof. The above result can be proven by considering the case $\gamma = 1/2$, since the generic result with $\gamma \in \mathbb{R}$ comes from translations in the transformed spaces. Let then $u \in \mathcal{K}^s_{1/2}$ for $s \in \mathbb{N}$: we have

$$(\lambda - 1)^s \hat{u}(\lambda) \in \mathcal{H}^0_0.$$

The above equation implies that

$$|\lambda|^j \hat{u}(\lambda) \in L^2(\{\mathrm{Re}\,\lambda = \beta\})$$

for all $j = 0, \dots, s$ and $\beta < 0$, thus from (2.6) and taking an inverse Mellin transform, we obtain

$$(x\partial_x)^j u \in \mathcal{K}^0_\beta$$

for $j = 0, \dots, s$ and for all $\beta \leq 1/2$. Then, (2.9) follows by taking a linear combination of the terms above. □

We can therefore give a new characterization of the weighted spaces $\mathcal{K}^s_\gamma$ with integer s as

$$\mathcal{K}^s_\gamma = \left\{ v \in \mathcal{E}_+ : r^{j-\gamma}\partial^j v \in L^2(\mathbb{R}_+), \text{ for all } j = 0, \dots, s \right\}.$$

This definition should also clarify the classical terminology of weighted Sobolev spaces.

2.1.1 Asymptotics near the origin

Apart from being a tool to rigorously define weighted Sobolev spaces on $\mathbb{R}_+$, the Mellin transformation can be used to derive an asymptotic expansion of a function near the origin. Broadly speaking, we know that transforms of functions in $\mathcal{K}^s_\gamma$ are holomorphic on half-planes; under some conditions, they are also meromorphic on the complementary half-plane, and their poles are related to the terms arising in the generalized asymptotic expansion of the function, near the origin. In the following, we show how this can be proven more rigorously.

First, let $p \in \mathbb{R}$ and $k \in \mathbb{N}$. Given a smooth cutoff function χ defined over $\mathbb{R}_+$ such that $\chi(0) = 1$ and $\chi(a) = 0$ for some $a > 0$, we define

$$\varphi_{p,k} = x^p (\log x)^k \chi(x). \tag{2.10}$$

Since for $u \in \mathcal{E}_+$, $\mathcal{M}((\log x)u) = \partial_\lambda \hat{u}(\lambda)$,

$$\mathcal{M}\varphi_{p,k} = \partial^k_\lambda \hat{\chi}(\lambda - p)$$

with $\hat{\chi} = \mathcal{M}\chi$. It can be shown that $\hat{\chi}$ has a single pole in zero; thus, $\mathcal{M}\varphi_{p.k}$ has a pole of order $k+1$ at p.

Consider now a sequence $\{p_j\}_{j\in\mathbb{N}} \subset \mathbb{C}^{\mathbb{N}}$ such that $\operatorname{Re} p_j \to \infty$ and a sequence $\{m_j\}_{j\in\mathbb{N}} \subset \mathbb{N}^{\mathbb{N}}$ and define

$$J_\gamma = \left\{(j,k) \subset \mathbb{N}^2 : \operatorname{Re} p_j \le \gamma - 1/2, 0 \le k \le m_j\right\}.$$

for $\gamma \in \mathbb{R}$. Given $a_{j,k} \in \mathbb{C}$, for all $(j,k) \in J_\gamma$, we write

$$u \sim_{(s,\gamma)} \sum_{(j,k)\in J_\gamma} a_{j,k}\varphi_{p_j,k} \tag{2.11}$$

if

$$u - \sum_{(j,k)\in J_\gamma} a_{j,k}\varphi_{p_j,k} \in \mathcal{K}^s_\gamma.$$

The sum $\sum_{(j,k)\in J_\gamma} a_{j,k}\varphi_{p_j,k}$ will be called a generalized asymptotic expansion.

Lemma 3. *If* (2.11) *holds for all s and all γ, then $u \in C^\infty(\mathbb{R}_+)$ and the following asymptotic expansion holds in the classical sense:*

$$u \sim \sum_{(j,k)\in J} a_{j,k}\varphi_{p_j,k},$$

with $J = \bigcup_{\gamma\in\mathbb{R}} J_\gamma$.

Proof. Let

$$u_\gamma = u - \sum_{(j,k)\in J_\gamma} a_{j,k}\varphi_{p_j,k}.$$

Since (2.11) holds for all s and for all γ, we have in particular that for all $r \in \mathbb{R}$, $u_r \in \mathcal{K}_r^\infty = \bigcap_{k\in\mathbb{N}} \mathcal{K}_r^k$. This implies that for all $j \in \mathbb{N}$, $u \in C^j(\mathbb{R}_+)$ and

$$\partial_x^j u_r = o(x^{r-j+1/2}),$$

for $x \to 0^+$, see [Jea79, Proposition 3.4]. Then the generalized asymptotic expansion is a classical one, and it is infinitely differentiable. □

The next lemma will make the connection between the properties of the Mellin transform and the generalized asymptotic expansion of a function.

Lemma 4. *Let $u \in \mathcal{E}_+$ and let $\hat{u}$ be its Mellin transform. There exist $a_{j,k} \in \mathbb{C}$, $\{p_j\}_j$ and $\{m_j\}_j$ such that*

$$u \sim_{(s,\gamma)} \sum_{J_\gamma} a_{j,k}\varphi_{p_j,k}, \tag{2.12}$$

for all $s, \gamma \in \mathbb{R}$ if and only if

(i) $\hat{u}$ is meromorphic with poles of order m_j at every p_j,

(ii) there exists $a > 0$ such that for all $m \in \mathbb{N}$,

$$(1 + |\lambda|)^m \hat{u}(\lambda) a^{\operatorname{Re}\lambda}$$

is bounded outside a compact set in the half-plane $\{\operatorname{Re}\lambda \le m\}$

Proof. Suppose that

$$u \sim_{(s,\gamma)} \sum_{J_\gamma} a_{j,k}\varphi_{p_j,k},$$

holds for all $s, \gamma \in \mathbb{R}$. Define then

$$u_\gamma = u - \sum_{(j,k)\in J_\gamma} a_{j,k}\varphi_{p_j,k}.$$

For any $m \in \mathbb{N}$, by hypothesis we can take $u_{m+3/2} \in \mathcal{K}^m_{m+3/2}$, so that, by (2.8) and (2.7), $\hat{u}_{m+3/2} = \mathcal{M}u_{m+3/2} \in \mathcal{H}^m_{m+1}$, i.e., $\hat{u}_{m+3/2}(\lambda)$ is holomorphic for $\operatorname{Re}\lambda < m + 1$.

Suppose now we are given an $a > 1$ and a function $f \in C_0^\infty([0, a])$. Then, for any $k \in \mathbb{N}$

$$|\lambda^k(\mathcal{M}f)(\lambda)| \le \int_{(0,a)} \left|(x\partial_x)^k f\right| \left|x^{-\lambda-1}\right| dx \le C_k a^{-|\operatorname{Re}\lambda|} \tag{2.13}$$

Therefore, for any $m \in \mathbb{N}$, there exists a $C_m > 0$ such that for $\lambda \in \mathbb{C}$,

$$(1 + |\lambda|)^m \left|(\mathcal{M}f)(\lambda)\right| \le C_m a^{-\operatorname{Re}\lambda}. \tag{2.14}$$

Furthermore, using the definition of $\varphi_{p,k}$ given in (2.10) and denoting $\hat{\varphi}_{p,k} = \mathcal{M}\varphi_{p,k}$, we see that, for any $m \in \mathbb{N}$, there exists $C_m > 0$ such that

$$(1 + |\lambda|)^m (\lambda - p)^{k+1} \hat{\varphi}_{p,k}(\lambda) a^{\operatorname{Re}\lambda} \le C_m \tag{2.15}$$

for $\operatorname{Re}\lambda \leq m$ and where $a > 0$ is such that $\operatorname{supp}(\chi) \subset (0, a)$. Using (2.14) on u_{m+1} and (2.15) on the sum, we conclude with part (ii) of the thesis. The first part follows more directly by seeing that the poles of $\hat{u}$ are the poles of $\hat{\varphi}_{p_j,k}$ for all $0 \leq k \leq m_j$ and for all j.

Let us now suppose that (i) and (ii) hold. Since $\hat{u}$ is meromorphic, we can choose $a_{j,k} \in \mathbb{C}$ such that at every pole, the principal part of $\hat{u}$ at every p_j is given by

$$\sum_{k=0}^{m_j} a_{j,k}(-1)^k k!(\lambda - p_j)^{k+1}. \tag{2.16}$$

Fix an $s \in \mathbb{N}$ and a $\gamma \in \mathbb{R}$. Then, taking m big enough in (ii), we obtain that

$$\hat{u}_\gamma = \hat{u} - \sum_{(j,k)\in J_\gamma} a_{j,k}\hat{\varphi}_{p_j,k} \in \mathcal{H}^s_{\gamma-1/2},$$

thus $u_\gamma = \mathcal{M}^{-1}\hat{u}_\gamma \in \mathcal{K}^s_\gamma$. Notice then that the principal part at p of $\mathcal{M}\varphi_{p,k}$ is given by

$$(-1)^k k!(\lambda - p)^{k+1}.$$

Comparing this with (2.16) we see that (2.12) holds for fixed s, γ. Since for any s, γ we can choose a sufficiently large m, this implies (2.12) for all $s, \gamma \in \mathbb{R}$. □

2.2 Mellin transformation and weighted Sobolev spaces in conical domains

Consider now a domain $\Omega \subset \mathbb{R}^d$. Suppose that (e.g. through a change of variables), we can write

$$\Omega = \mathbb{R}_+ \times S,$$

where S is a subset with smooth boundary of the $d-1$ dimensional sphere $\mathbb{S}_{d-1}$. We will denote the variables as (r, ω) for $r \in \mathbb{R}_+$, $\omega \in S$, with $r = |x|$ and $\omega = x/|x|$.

We want to extend the analysis set up in the previous section to the case of conical spaces. With a slight abuse of notation, we can consider a function defined on Ω as a function $u : r \mapsto u(r, \cdot)$, i.e.,

$$u : \mathbb{R}_+ \to V,$$

where V is some space of functions defined on S that will be specified later. Applying the Mellin transformation to the function u, we therefore obtain

$$\hat{u} = \mathcal{M}_{r\to\lambda}u : \mathbb{C} \to V.$$

2.2.1 Mellin transform and homogeneous weighted Sobolev spaces in a cone

In multiple dimensions d, the same construction as in Section 2.1 can be followed, with the major difference that the metric invariant with respect to homotheties is given by

dx/x^d. The homogeneous weighted Sobolev spaces in Ω are given, for integer s and real γ, by

$$\mathcal{K}^s_\gamma(\Omega) = \left\{ v \in L^2_{\mathrm{loc}}(\Omega) : r^{|\alpha|-\gamma}\partial^\alpha v \in L^2(\Omega), \text{ for all } |\alpha| \leq s \right\}, \tag{2.17}$$

where $\alpha \in \mathbb{N}^d$ is a multi-index and $|\alpha| = \|\alpha\|_{\ell^1}$. The difference in scale manifests itself when we consider the relationship between a function and its Mellin transform: in the multidimensional conical example, we write formally

$$\mathcal{H}^s_{\gamma-d/2}(\Omega) = \mathcal{M}_{r\to\lambda}\mathcal{K}^s_\gamma(\Omega). \tag{2.18}$$

Note that this is consistent with the one-dimensional case shown above. This notation is used to make it evident that a function $v \in \mathcal{K}^s_\gamma(\Omega)$ will have a Mellin transform $\hat{v} = \mathcal{M}_{r\to\lambda}v$ holomorphic on the half-plane $\{\operatorname{Re}\lambda < \gamma - d/2\}$. We suppose that all functions we treat have compact support; we will not specify this further.

Given a function $\hat{u} : \mathbb{C} \to H^m(S)$, we define

$$\mathcal{N}^m_\beta(\hat{u}) = \left(\int_{\operatorname{Re}\lambda=\beta} \sum_{j=0}^{m} |\lambda|^{2j} \|\hat{u}\|^2_{H^{m-j}(S)} d\lambda \right)^{1/2}. \tag{2.19}$$

The following proposition holds.

Proposition 5. *The Mellin transformation is an isomorphism on the spaces*

$$\mathcal{K}^s_\gamma(\Omega) \to \mathcal{H}^s_{\gamma-d/2}(\Omega)$$

where the former space has norm

$$\|v\|^2_{\mathcal{K}^s_\gamma(\Omega)} = \sum_{|\alpha|\leq s} \|r^{|\alpha|-\gamma}\partial^\alpha v\|^2_{L^2(\Omega)} \tag{2.20}$$

and the latter is equipped with the norm $\mathcal{N}^s_{\gamma-d/2}(\cdot)$.

Proof. The differential operator ∂^α in classical cartesian coordinates can be written as

$$\partial^\alpha = r^{-|\alpha|} \sum_{j\leq|\alpha|} p_{\alpha,j}(\omega, \partial_\omega)(r\partial_r)^j, \tag{2.21}$$

where $p_{\alpha,j}(\omega, \partial_\omega)$ are differential operators of degree inferior or equal to $|\alpha| - j$ with smooth coefficients on $\mathbb{S}_{d-1}$. In addition, the converse is also true, i.e., for any $p(\omega, \partial_\omega)$ of order k with smooth coefficients on $\mathbb{S}_{d-1}$,

$$p(\omega, \partial_\omega)(r\partial_r)^j = \sum_{|\alpha|\leq k+j} a_\alpha(\omega) r^{|\alpha|}\partial^\alpha,$$

with smooth coefficients a_α. The squared norm (2.20) is therefore equivalent to

$$\int_{\mathbb{R}_+} r^{-2\gamma+d-1} \sum_{j=0}^{s} \|(r\partial_r)^j u(r)\|^2_{H^{s-j}(\Omega)} dr. \tag{2.22}$$

An application of Plancherel's inequality (2.6) and of (2.2) concludes the proof. □

Asymptotic expansions have the same behavior as in the one-dimensional case. The results obtained in the previous section can be extended without major modifications, as long as we replace the constants $a_{j,k} \in \mathbb{C}$ in (2.11) with functions $a_{j,k} : S \to \mathbb{C}$ and use the norms defined on Ω.

2.2.2 Non homogeneous weighted Sobolev spaces

The issue with homogeneous Sobolev spaces, in practical applications, lies in the fact that the regularity of a function is strongly determined by its values at the vertex of the cone. For example, any function belonging to the space $\mathcal{K}^s_\gamma(\Omega)$ for any $\gamma \geq d/2$ has to be null at the singular point. In many situations, we want our spaces to include functions with nontrivial Taylor expansion at the origin, while still being interested in the regularity of the function "modulo" this expansion. This is the situation in which non homogeneous Sobolev spaces come into play. The non homogeneous Sobolev space $\mathcal{J}^s_\gamma(\Omega)$ is defined, for integer s and real γ, as the space of functions with finite norm

$$\|u\|^2_{\mathcal{J}^s_\gamma(\Omega)} = \sum_{|\alpha|\leq s} \|r^{s-\gamma}\partial^\alpha u\|^2_{L^2(\Omega)}. \tag{2.23}$$

In the following, we will analyze, from the point of view of the Mellin transform, the properties of functions in $\mathcal{J}^s_\gamma(\Omega)$ and their relationship with the space $\mathcal{K}^s_\gamma(\Omega)$. This will also lead to a more convenient definition of the norm than (2.23). The content of the section follows closely the exposition given in [CDN10b].

Broadly speaking, the idea is to enlarge the space $\mathcal{M}_{r\to\lambda}\mathcal{K}^s_\gamma(\Omega)$, which contains functions that are holomorphic on $\{\operatorname{Re}\lambda < \gamma - d/2\}$, to a wider space of functions that are meromorphic on a wider half-plane $\{\operatorname{Re}\lambda < \gamma' - d/2\}$, with $\gamma' > \gamma$. Under some conditions on the poles and on the norms in the wider half-plane, the poles will be transformed back and represent an expansion of the function in $\mathcal{K}^s_\gamma(\Omega)$ at the singularity.

We will now show how the non homogeneous weighted Sobolev spaces can be characterized based on two weights and a set of poles of their Mellin transform. To fix ideas, consider Figure 2.1, where a schematic representation of the transform of a function in $\mathcal{J}^s_\gamma(\Omega)$ is given. Supposing that $\gamma = \eta + d/2$ lies between two integers, then the transform has to be holomorphic in the half-plane of negative $\operatorname{Re}\lambda$, and its poles in $[0, \eta]$ have to be a subset of $\mathbb{N} \cap [0, \eta]$.

Consider now a set $\mathcal{S} \subset \mathbb{N}$ containing 0 and a sequence $\Gamma = \{\gamma_i\}_{i\in\mathcal{S}}$ such that

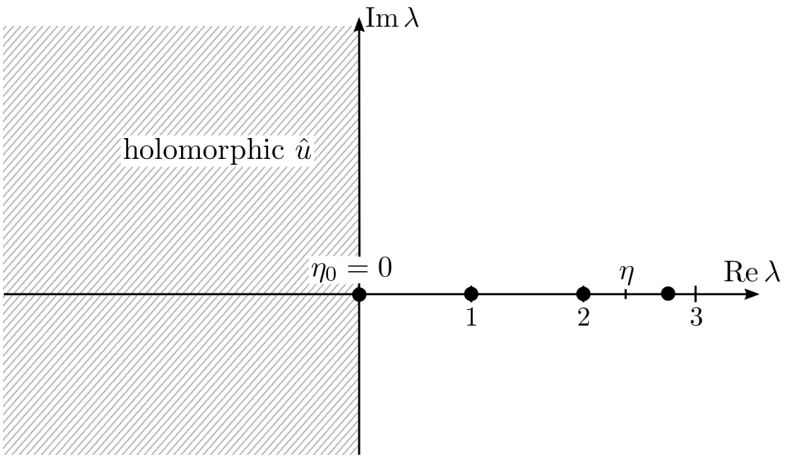


Figure 2.1 – Scheme representing the Mellin transform $\hat{u} = \mathcal{M}_{r\to\lambda}u$, for a function $u \in \mathcal{K}^s_{\gamma_0}(\Omega)$ for any $\gamma_0 < d/2$ (hence $\hat{u}$ is holomorphic in the half-plane $\{\mathrm{Re}\,\lambda < 0\}$). Furthermore, the dots represent poles of $\hat{u}$. Suppose $s \geq 3$: then, under some conditions, $u \in \mathcal{J}^s_\gamma(\Omega)$, with $\gamma = \eta + d/2$ and $s \geq 3$

$\gamma_{i+1} > \gamma_i$. We can then define a generic norm

$$\|v\|^2_{\mathcal{J}^{\mathcal{S}}_\Gamma(\Omega)} = \sum_{i\in\mathcal{S}}\sum_{|\alpha|=i} \|r^{i-\gamma_i}\partial^\alpha v\|^2_{L^2(\Omega)}. \tag{2.24}$$

Remark 4. *If we choose*

$$\mathcal{S} = \{0,\dots,s\} \qquad \gamma_i = i + \gamma - s \tag{2.25}$$

we obtain $\|\cdot\|_{\mathcal{J}^{\mathcal{S}}_\Gamma(\Omega)} = \|\cdot\|_{\mathcal{J}^s_\gamma(\Omega)}$ *as defined in* (2.23).

We can now give a description of the space $\mathcal{J}^{\mathcal{S}}_\Gamma(\Omega)$ based on the properties of its Mellin transform. Let us fix

$$\mathcal{S} = \{0,\dots,s\},$$

$\eta \in \mathbb{R}$ and $\eta_0 = \min(0,\eta)$. Consider now a set of weights $\Gamma = \{\gamma_i\}_{i\in\mathcal{S}}$ such that $\gamma_s = \eta + d/2$ and $\gamma_{i+1} > \gamma_i$. We introduce the set

$$\mathfrak{N} = \{0,\dots,s-1\} \setminus \left(\bigcup_{j\in\mathcal{S}} [j,\eta_j]\right), \tag{2.26}$$

where $\eta_j = \gamma_j - d/2$. We now affirm that the norm in $\mathcal{J}^{\mathcal{S}}_\Gamma(\Omega)$ is fully determined by η, s, and $\mathfrak{N}$. This has the remarkable consequence that, for any $\widetilde{\Gamma} = \{\tilde{\gamma}_i\}_{i\in\mathcal{S}}$ giving rise to the same $\mathfrak{N}$ and with the same last element,

$$\|\cdot\|_{\mathcal{J}^{\mathcal{S}}_\Gamma(\Omega)} \simeq \|\cdot\|_{\mathcal{J}^{\mathcal{S}}_{\widetilde{\Gamma}}(\Omega)}.$$

We outline a proof of the above statement; for the details, see [CDN10b]. Consider the norm defined in (2.19): introducing $\beta_1 > \beta_0$ and supposing that $\hat{u}$ is holomorphic in the strip of the complex plane $\{\mathrm{Re}\,\lambda \in (\beta_0, \beta_1)\}$, we define

$$\mathcal{N}^m_{[\beta_0,\beta_1]}(\hat{u}) = \sup_{\beta\in(\beta_0,\beta_1)} \mathcal{N}^m_\beta(\hat{u}), \tag{2.27}$$

for $m \in \mathbb{N}$. As illustrated in Figure 2.1, when dealing with non homogeneous spaces $\mathcal{J}^s_\gamma(\Omega)$ we need to consider functions in $\mathcal{M}_{r\to\lambda}\mathcal{J}^s_\gamma(\Omega)$, which are meromorphic in part of the complex plane. In order to do this, consider

$$\hat{u} : \mathbb{C} \to H^s(S)$$

such that $\hat{u}$ is meromorphic in the strip $\{\beta_0 < \mathrm{Re}\,\lambda < \beta_1\}$. First, we introduce a version of (2.19) that stays bounded for a subclass of meromorphic functions:

$$\mathcal{N}^m_{\beta,k}(\hat{u})^2 = \int_{\substack{\mathrm{Re}\,\lambda=\beta\\ |\,\mathrm{Im}\,\lambda|\le 1}} \|(1-P^k)\hat{u}(\lambda)\|^2_{H^m(S)}d\lambda + \int_{\substack{\mathrm{Re}\,\lambda=\beta\\ |\,\mathrm{Im}\,\lambda|\le 1}} |\lambda-k|^2\|P^k\hat{u}(\lambda)\|^2_{H^m(S)}d\lambda + \int_{\substack{\mathrm{Re}\,\lambda=\beta\\ |\,\mathrm{Im}\,\lambda|> 1}} \sum_{j=0}^{m} |\lambda|^{2j}\|\hat{u}\|^2_{H^{m-j}(S)}d\lambda d\lambda.$$

Here, P^k is a projector from $L^2(S)$ to the space of restrictions to S of homogeneous polynomials of degree k in Ω. Let now $\widetilde{\mathfrak{N}}$ be a set of integers $\{k_1, \dots, k_n\} \subset \mathbb{N} \cap (\beta_0, \beta_1)$: take $K_j = (k_j - 1/2, k_j + 1/2)$ and $K_0 = (\beta_0, \beta_1) \setminus \left(\bigcup_j K_j\right)$ and define

$$\mathcal{N}^m_{[\beta_0,\beta_1],\widetilde{\mathfrak{N}}}(\hat{u}) = \max_{i=0,\dots,n} \left(\sup_{\beta\in K_i} \mathcal{N}^m_{\beta,k}(\hat{u}) \right). \tag{2.28}$$

The relevance of the norm (2.28) is shown by the following lemma.

Lemma 6. *Let $\gamma_0, \gamma_1 \in \mathbb{R}$ such that $\gamma_0 < \gamma_1$. Furthermore, let $\beta_i = \gamma_i - d/2$, $i = 1, 2$ and let $u \in \mathcal{K}^0_{\gamma_0}(\Omega)$ and $\mathfrak{N}_s = \{0, \dots, s-1\} \cap (\beta_0, \beta_1)$. Thus,*

$$\mathcal{N}^m_{[\beta_0,\beta_1],\mathfrak{N}_s}(\hat{u}) \simeq \|u\|_{\mathcal{K}^0_{\gamma_0}(\Omega)} + |u|_{\mathcal{K}^s_{\gamma_1}(\Omega)}, \tag{2.29}$$

where $\hat{u} = \mathcal{M}_{r\to\lambda}u$.

Note now that we can rewrite definition (2.24) as

$$\|u\|^2_{\mathcal{J}^{\mathcal{S}}_\Gamma(\Omega)} = \sum_{i\in\mathcal{S}}^{s} |u|^2_{\mathcal{K}^i_{\gamma_i}}. \tag{2.30}$$

We denote by $\mathcal{H}^m_{[\eta_0,\eta_1],\mathfrak{N}}$ the space of $\hat{u}$ with finite $\mathcal{N}^m_{[\eta_0,\eta_1],\mathfrak{N}}(\hat{u})$ norm. From (2.30) and

(2.29), one sees that the Mellin transform is an isomorphism

$$\mathcal{J}^{\mathcal{S}}_{\Gamma}(\Omega) \to \bigcap_{m\in S} \mathcal{H}^{m}_{[\eta_0,\eta_j],\mathfrak{N}_j}$$

where $\mathfrak{N}_j$ is the subset of $\mathfrak{N}$ such that for all $k \in \mathfrak{N}_j$, $k \leq j-1$. Now, it can be shown that

$$\bigcap_{m\in S} \mathcal{H}^{m}_{[\eta_0,\eta_j],\mathfrak{N}_j} = \mathcal{H}^{m}_{[\eta_0,\eta],\mathfrak{N}},$$

thus the following proposition holds.

Proposition 7. *Let*

$$\mathcal{S} = \{0,\dots,s\},$$

$\eta, \eta_0 \in \mathbb{R}$ *and* $\eta_0 < \eta$. *Let then* $\Gamma = \{\gamma_i\}_{i\in\mathcal{S}}$ *such that* $\gamma_s = \eta + d/2$ *and* $\gamma_{i+1} > \gamma_i$. *Finally, let* $\mathfrak{N}$ *be defined as in* (2.26). *Then,* $\mathcal{M}_{r\to\lambda}$ *is an isomorphism between*

$$\mathcal{J}^{\mathcal{S}}_{\Gamma}(\Omega) \to \mathcal{H}^{s}_{[\eta_0,\eta],\mathfrak{N}},$$

with norms defined in (2.24) *and* (2.28).

We now consider the special choice (2.25), that gives the spaces $\mathcal{J}^{s}_{\gamma}(\Omega)$. It is evident that, if $\eta < 0$ or $\eta > s$, then $\mathfrak{N} = \varnothing$, while in the case $\eta \in [0,s]$, then $\mathfrak{N} = \{0,\dots,\lfloor\eta\rfloor\}$. Furthermore, by the definition (2.28),

$$\mathcal{N}^{m}_{[\beta_0,\beta_1],\varnothing}(\cdot) = \mathcal{N}^{m}_{[\beta_0,\beta_1]}(\cdot).$$

When $\eta < 0$ or $\eta > s$, furthermore, if $u \in \mathcal{J}^{s}_{\eta-d/2}(\Omega)$ then $\hat{u} = \mathcal{M}_{r\to\lambda}u$ is holomorphic on the half plane $\{\operatorname{Re}\lambda < \eta\}$, thus Proposition 5 implies

$$\mathcal{N}^{s}_{[\eta_0,\eta]}(\hat{u}) \simeq \|u\|_{\mathcal{K}^{s}_{\eta-d/2}(\Omega)}.$$

We can summarize this in the following statement.

Proposition 8. *Let* $s \in \mathbb{N}$, $\gamma \in \mathbb{R}$ *with* $\eta = \gamma - d/2$. *Then,*

- *if* $\eta < 0$ *or* $\eta \geq s$, *then* $\mathcal{J}^{s}_{\gamma}(\Omega) = \mathcal{K}^{s}_{\gamma}(\Omega)$
- *if* $0 \leq \eta < s$, *then*

$$\mathcal{J}^{s}_{\gamma}(\Omega) = \mathcal{M}^{-1}_{\lambda\to r}\mathcal{H}^{s}_{[\eta-s,\eta],\mathfrak{N}}, \tag{2.31}$$

  with $\mathfrak{N} = \{0,\dots,\lfloor\eta\rfloor\}$

The above proposition implies that, when $0 \leq \eta < s$, choosing in (2.24) any sequence $\Gamma = \{\gamma_0,\dots,\gamma_s\}$ giving rise to the same set $\mathfrak{N}$ and such that $\gamma_s - d/2 = \eta$, $\gamma_0 - d/2 \leq 0$ gives an equivalent $\|\cdot\|_{\mathcal{J}^{s}_{\Gamma}(\Omega)}$ norm. Let us then fix $\gamma \in \mathbb{R}$, denote $\eta = \gamma - d/2$ and fix $s_0 \in (\eta, s]$. The choice

$$\gamma_i = \begin{cases} \gamma + i - s_0 & \text{if } i \leq s_0 \\ \gamma & \text{if } i > s_0 \end{cases}$$

gives the equivalent norms

$$\|u\|^2_{\mathcal{J}^s_\gamma(\Omega)} = \sum_{|\alpha|\le s} \|r^{\max(|\alpha|-\gamma,\rho)}\partial^\alpha u\|^2_{L^2(\Omega)}, \tag{2.32}$$

for any $\rho \in (-d/2, s-\gamma]$. With this norm, it can be easily shown that $\mathcal{J}^{s+1}_\gamma(\Omega) \subset \mathcal{J}^s_\gamma(\Omega)$.

We conclude by remarking the similarity between (2.31) and the previous relations (2.8) and (2.18). We underline, however, how in this section we have always considered a regularity exponent $s \in \mathbb{N}$, and that the definition of non homogeneous spaces with non integer regularity exponent is not trivial.

2.2.3 Relationship between homogeneous and non homogeneous spaces

It remains to consider the relationship between $\mathcal{J}^s_\gamma(\Omega)$ and $\mathcal{K}^s_\gamma(\Omega)$ when $0 \le \gamma - d/2 \le s$, i.e., when the two spaces differ. We will find that, when $\eta = \gamma - d/2 \notin \mathbb{N}$, then $\mathcal{J}^s_\gamma(\Omega)$ contains $\mathcal{K}^s_\gamma(\Omega)$ as a subset of finite codimension, while when $\eta \in \mathbb{N}$, the codimension of $\mathcal{K}^s_\gamma(\Omega)$ is infinite.

Case $\eta \notin \mathbb{N}$

We start by the analysis of the spaces with $\eta \notin \mathbb{N}$. First, consider a function $\hat{u} \in \mathcal{H}^s_{[\beta_0,\beta_1],\mathfrak{N}}$. For any $\beta_0 < \beta < \beta_1$, $\beta \notin \mathfrak{N}$, by contour integration we find that

$$\int_{\mathrm{Re}\,\lambda=\beta} r^\lambda \hat{u}(\lambda)d\lambda - \int_{\mathrm{Re}\,\lambda=\beta_0} r^\lambda \hat{u}(\lambda)d\lambda = \sum_{j\in\mathfrak{N}\cap[\beta_0,\beta]} \mathrm{Res}_{\lambda=j}(r^\lambda \hat{u}(\lambda)) \tag{2.33}$$

We now have

$$\int_{\mathrm{Re}\,\lambda=\beta} r^\lambda \hat{u}(\lambda)d\lambda \in \mathcal{K}^s_{\beta+d/2}(\Omega),$$

and the second term at the left hand side above is the inverse transform of a function $u = \mathcal{M}^{-1}_{\lambda\to r}\hat{u}$, hence we can rewrite (2.33) as

$$u \sim_{(s,\gamma)} \sum_{j\in\mathfrak{N}\cap[\beta_0,\gamma-d/2]} -\mathrm{Res}_{\lambda=j}(r^\lambda \hat{u}(\lambda)),$$

as long as $u \in \mathcal{K}^s_\delta(\Omega)$ for all $\delta < \beta_0 - d/2$. The definition of the norm (2.28), then, implies that $\mathcal{N}_{[\beta_0,\beta_1],\mathfrak{N}}(\hat{u}) < \infty$ only if, for any $j \in \mathfrak{N}\cap[\beta_0,\beta_1]$,

$$\mathrm{Res}_{\lambda=j}(r^\lambda \hat{u}(\lambda)) = r^j P^j \,\mathrm{Res}_{\lambda=j}(\hat{u}(\lambda)).$$

We can therefore state that

Proposition 9. *If $u \in \mathcal{J}^s_\gamma(\Omega)$, with $\eta = \gamma - d/2$ such that $0 < \eta \leq s$ and $\eta \notin \mathbb{N}$, then*

$$u \sim_{(s,\gamma)} \sum_{|\alpha| \leq \lfloor \eta \rfloor} \partial^\alpha u(0) \frac{x^\alpha}{\alpha!}$$

This implies that for any $u \in \mathcal{J}^s_\gamma(\Omega)$, $0 < \gamma - d/2 = \eta \leq s$, there exists a $v \in \mathcal{K}^s_\gamma(\Omega)$ such that

$$u = v + \sum_{|\alpha| \leq \lfloor \eta \rfloor} (\partial^\alpha u)(0) \frac{x^\alpha}{\alpha!}.$$

Now, for any polynomial $p_k \in \mathbb{Q}_k(\Omega)$

$$p_k(x) = \sum_{\alpha \leq k} c_\alpha \frac{x^\alpha}{\alpha!},$$

$p_k \in \mathcal{K}^s_\gamma(\Omega)$ and $p_k \neq 0$ only if $\gamma < d/2 + k$. This implies that $\mathcal{K}^s_\gamma(\Omega) \cap \mathbb{Q}_{\lfloor \eta \rfloor} = \varnothing$. Denote now

$$p_\eta(u) = \sum_{|\alpha| \leq \lfloor \eta \rfloor} (\partial^\alpha u)(0) \frac{x^\alpha}{\alpha!};$$

due to the equivalency of norms in finite dimensional spaces we have shown the following result.

Theorem 1. *Consider the space $\mathcal{J}^s_\gamma(\Omega)$ with $\gamma \in \mathbb{R}$, $\eta = \gamma - d/2 \notin \mathbb{N}$ and $0 < \eta < s$. Then,*

$$\mathcal{J}^s_\gamma(\Omega) = \mathcal{K}^s_\gamma(\Omega) \oplus \mathbb{Q}_{\lfloor \eta \rfloor}$$

and for any $u \in \mathcal{J}^s_\gamma(\Omega)$ we have the equivalency of norms

$$\|u\|_{\mathcal{J}^s_\gamma(\Omega)} \simeq \|u - p_\eta(u)\|_{\mathcal{K}^s_\gamma(\Omega)} + \sum_{|\alpha| \leq \lfloor \eta \rfloor} |(\partial^\alpha u)(0)|.$$

We conclude the part where $\eta \notin \mathbb{N}$ with a result that will be useful later on, when working with elliptic operators in non homogeneous weighted Sobolev spaces.

Lemma 10. *Let $u \in \mathcal{J}^s_\gamma(\Omega)$ with $\eta = \gamma - d/2 \notin \mathbb{N}$ and $0 < \eta < s$. Then, for any $\varepsilon > 0$ and for $|\alpha| \leq \lfloor \eta \rfloor$, there exists a $C_\varepsilon > 0$ such that*

$$|(\partial^\alpha u)(0)| \leq \varepsilon \|u\|_{\mathcal{J}^s_\gamma(\Omega)} + C_\varepsilon \|u\|_{\mathcal{J}^{s-1}_{\gamma-1}(\Omega)}$$

Proof. From Theorem 1, and using definition (2.23) for the norm, we have, for any $w \in \mathcal{J}^s_\gamma(\Omega)$ and any $|\beta| \leq \lfloor \eta \rfloor$,

$$|(\partial^\beta w)(0)| \leq \sum_{|\alpha| = s} \|r^{s-\gamma} \partial^\alpha w\|_{L^2(\Omega)} + \left(\sum_{|\alpha| \leq s-1} \|r^{s-\gamma} \partial^\alpha w\|^2_{L^2(\Omega)} \right)^{1/2}$$

Now, introduce $u = w(\varepsilon x)$: from the inequality above,

$$|(\partial^\beta w)(0)| \leq \varepsilon^{-|\beta|+\gamma-d/2} \sum_{|\alpha|=s} \|r^{s-\gamma}\partial^\alpha w\|_{L^2(\Omega)} + \left(\sum_{|\alpha|\leq s-1} \varepsilon^{-|\beta|+|\alpha|-s+\gamma-d/2} \|r^{s-\gamma}\partial^\alpha w\|^2_{L^2(\Omega)} \right)^{1/2}.$$

Since $|\beta| \leq \lfloor \eta \rfloor$ implies that $-|\beta|+\gamma-d/2 > 0$, by rewriting $\varepsilon^{-|\beta|+\gamma-d/2}$ as ε and deriving the corresponding C_ε we obtain the thesis. □

Case $\eta \in \mathbb{N}$

When $\eta \in \mathbb{N}$, the situation is more complicated. This case falls out of the scope of our analyses, so we will present here only the final results for the sake of completeness.

Consider $\mathcal{J}^s_\gamma(\Omega)$, with $\eta = \gamma - d/2 \in \mathbb{N}$ and $0 < \eta \leq s$. Then, $\mathcal{J}^s_\gamma(\Omega)$ contains $\mathcal{K}^s_\gamma(\Omega) \oplus \mathbb{Q}_{\eta-1}(\Omega)$ as a strict subspace of infinite codimension. There exists an operator K_η such that

$$u = v + p_{\eta-1}(u) + K_\eta u, \tag{2.34}$$

where $v \in \mathcal{K}^s_\gamma(\Omega)$ and $K_\eta u \in \mathcal{J}^\varpi_\gamma(\Omega)$, with

$$\mathcal{J}^\varpi_\gamma(\Omega) = \left\{ v \in \bigcap_{k\in\mathbb{N}} \mathcal{J}^k_\gamma(\Omega) : \exists C, A \in \mathbb{R}_+ \text{ such that } \|v\|_{\mathcal{J}^k_\gamma(\Omega)} \leq CA^k k!, \text{ for all } k \in \mathbb{N} \right\}.$$

Therefore, the non homogeneous weighted space can be decomposed, in the critical case, into the homogeneous weighted space, the finite dimensional space of polynomial expansions at the origin, and an infinite dimensional space of "weighted analytic" functions.

The operator K_η in (2.34) is defined as

$$K_\eta = \sum_{|\alpha|=\eta} K(\chi d_\alpha)\frac{x^\alpha}{\alpha!}$$

where χ is a smooth cutoff function equal to one in the vicinity of the origin,

$$d_\alpha(r) = \langle r^{-|\alpha|}(u - p_{\eta-1}(u))(r,\cdot), \varphi_\alpha \rangle,$$

φ_α is the $L^2(S)$ dual basis to $\omega^\alpha/(\alpha!)$, and K is such that

$$\mathcal{M}_{r\to\lambda}(Kv)(\lambda) = e^{\lambda^2}(\mathcal{M}_{r\to\lambda}v)(\lambda).$$

2.3 Elliptic operators in conical domains

We now turn to the application of the results presented so far to the analysis of elliptic boundary value problems. We will focus primarily on the analysis in homogeneous weighted Sobolev spaces, which constitutes the basis for the analysis in non homogeneous spaces. The content of this section is based on [KMR97].

Before proceeding further, let us introduce a norm on the traces of $\mathcal{K}^s_\gamma(\Omega)$ functions on the boundary of the cone Ω as

$$\|u\|_{\mathcal{K}^{s-1/2}_{\gamma-1/2}(\partial\Omega)} = \inf\left\{\|v\|_{\mathcal{K}^s_\gamma(\Omega)} : v \in \mathcal{K}^s_\gamma(\Omega), v_{|\partial\Omega} = u\right\}. \tag{2.35}$$

2.3.1 Regularity results for model operators

We introduce the model operator in the cone of order k as

$$P(x,\partial_x) = r^{-k}\sum_{j=0}^{k} p_j(\omega,\partial_\omega)(r\partial_r)^j, \tag{2.36}$$

where $r \in \mathbb{R}_+$ and $\omega \in \mathbb{S}_{d-1}$ are spherical coordinates ($\mathbb{S}_{d-1}$ is the $d-1$ dimensional sphere) and $p_j(\omega,\partial_\omega)$ are operators of order $k-j$ with smooth coefficients. For example, differential operators with smooth coefficients in cartesian coordinates belong to this class, thanks to (2.21). Consider now the model problem

$$\begin{aligned} L(x,\partial_x)u &= f \quad \text{in } \Omega \\ B_k(x,\partial_x)u &= g_k \quad \text{on } \partial\Omega, \end{aligned} \tag{2.37}$$

where $L(x,\partial_x)$ is a model operator of second order and the $B_k(x,\partial_x)$ are model operators of order $k \leq 1$. We suppose that the problem is elliptic. We can now introduce

$$\begin{aligned} \mathcal{L}(\omega,\partial_\omega,r\partial_r) &= r^2 L(x,\partial_x), \\ \mathcal{B}_k(\omega,\partial_\omega,r\partial_r) &= r^k B_k(x,\partial_x) \end{aligned}$$

and define the boundary value problem

$$\begin{aligned} \mathcal{L}(\omega,\partial_\omega,\lambda)\hat{u} &= \hat{f}(\lambda-2) \quad \text{in } \Omega \\ \mathcal{B}_k(\omega,\partial_\omega,\lambda)\hat{u} &= \hat{g}_k(\lambda-2) \quad \text{on } \partial\Omega. \end{aligned} \tag{2.38}$$

We denote by $\mathcal{A}(\lambda)$ the operator associated with problem (2.38). Note that $\mathcal{A}(\lambda)$ is therefore the operator pencil associated with the operator (L, B_0, B_1), of (2.37). We denote by $\mathcal{K}^{s-\underline{k}-1/2}_{\gamma-\underline{k}-1/2}(\partial\Omega)$, for $\underline{k} = (0,1)$ the product $\mathcal{K}^{s-1/2}_{\gamma-1/2}(\partial\Omega) \times \mathcal{K}^{s-3/2}_{\gamma-3/2}(\partial\Omega)$. Under this hypotheses, the following estimate holds.

Theorem 2. *Let u be a solution of* (2.37) *and let the line $\{\operatorname{Re}\lambda = \gamma - d/2\}$ contain no eigenvalues*

of $\mathcal{A}(\lambda)$. Then,

$$\|u\|_{\mathcal{K}^s_\gamma(\Omega)} \leq C \left(\|f\|_{\mathcal{K}^{s-2}_{\gamma-2}(\Omega)} + \sum_j \|g_j\|_{\mathcal{K}^{s-k-1/2}_{\gamma-k-1/2}(\Omega)} \right) \tag{2.39}$$

and $A = (L, B_0, B_1)$ is an isomorphism between the spaces

$$\mathcal{K}^s_\gamma(\Omega)(\partial\Omega) \to \mathcal{K}^{s-2}_{\gamma-2}(\Omega) \times \mathcal{K}^{s-\underline{k}-1/2}_{\gamma-\underline{k}-1/2}(\partial\Omega).$$

Proof. Fix $\gamma \in \mathbb{R}$ such that no eigenvalues of $\mathcal{A}(\lambda)$ lie on the line $\{\operatorname{Re}\lambda = \eta\}$, with $\eta = \gamma - d/2$. Then, problem (2.38) is an isomorphism between the spaces $H^m(S)$ and $H^{m-2}(S)$. Furthermore, denoting

$$\|\hat{v}\|_{m,\lambda,S} = \sum_{j=0}^{m} |\lambda|^j \|\hat{u}\|_{H^{m-j}(S)}$$

and

$$\|\hat{v}\|_{m-1/2,\lambda,\partial S} = \|\hat{u}\|_{H^{m-1/2}(\partial S)} + |\lambda|^{m-1/2} \|\hat{u}\|_{L^2(\partial S)}$$

we have

$$\|\hat{u}\|_{s,\lambda,S} \leq C \left(\|\hat{f}\|_{s-2,\lambda-2,S} + \sum_j \|\hat{g}_j\|_{s-j-1/2,\lambda-j-1/2,\partial S} \right).$$

Integrating over the line $\{\operatorname{Re}\lambda = \eta\}$, using definition (2.19), the fact that the Mellin transform is an isomorphism between $\mathcal{K}^m_{\beta+d/2}(\Omega)$ and $\mathcal{H}^m_\beta$ stated in Proposition 5 and that an equivalent result holds for the spaces on the boundary, see [KMR97], we obtain (2.39). Furthermore, the existence of the solution is shown taking

$$u = \mathcal{M}^{-1}_{\lambda \to r} \hat{u}$$

and this completes the proof. □

As an example of the problem above, consider a two dimensional domain Ω with smooth boundary outside of a point $\mathfrak{c}$, and that in the vicinity of that point coincides with a plane sector of aperture $\alpha \in (\pi, 2\pi)$. Suppose we consider the Poisson equation with homogeneous Dirichlet boundary conditions

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega. \end{aligned}$$

Then, since we can write

$$-\Delta_x = -\frac{1}{r^2} \left((r\partial_r)^2 + \partial^2_\omega \right),$$

we have

$$\mathcal{L}(\omega, \partial_\omega, \lambda) = -(\lambda^2 + \partial^2_\omega).$$

Now, the eigenvalues μ_k of

$$\begin{aligned} \partial_\omega^2 v_k(\omega) &= \mu_k v_k(\omega) \text{ in } (0,\alpha) \\ v_k(0) &= v_k(\alpha) = 0 \end{aligned}$$

are given by $\mu_k = -(k\pi/\alpha)^2$, $k \in \mathbb{N} \setminus \{0\}$. The eigenvalues of $\mathcal{A}(\lambda)$ lie therefore at the points $\sqrt{-\mu_k} = k\pi/\alpha$ and this implies that Theorem 2 can be applied on any line $\operatorname{Re}\lambda = \gamma - 1$ with $\gamma - 1 \neq k\pi/\alpha$, $k \in \mathbb{N} \setminus \{0\}$.

2.3.2 Results for more general operators

We now consider the case of a more general class of operators, introducing

$$\widetilde{P}(x,\partial_x) = \sum_{|\alpha|\leq k} p_\alpha(x)\partial_x^\alpha \tag{2.40}$$

and supposing that

$$p_\alpha = r^{|\alpha|-k} p_\alpha^0(r,\omega),$$

with p_α^0 smooth in $\overline{S}\times\mathbb{R}_+$, continuous in $\overline{\Omega}$ and such that, for all $j \in \mathbb{N}$ and all $\zeta \in \mathbb{N}^{d-1}$,

$$(r\partial_r)^j \partial_\omega^\zeta (p_\alpha^0(r,\omega) - p_\alpha^0(0,\omega)) \to 0 \tag{2.41}$$

as $r \to 0$. For such an operator, we can define its leading part as

$$\widetilde{P}^0(x,\partial_x) = \sum_{|\alpha|\leq k} r^{|\alpha|-k} p_\alpha^0(0,\omega)\partial_x^\alpha.$$

As in the previous section, we introduce the problem

$$\begin{aligned} \widetilde{L}(x,\partial_x)u &= f \quad \text{in } \Omega \\ \widetilde{B}_k(x,\partial_x)u &= g_k \quad \text{on } \partial\Omega, \end{aligned} \tag{2.42}$$

where $\widetilde{L}$ and $\widetilde{B}_k$ are of the form (2.40). We denote by $\widetilde{L}^0$ and $\widetilde{B}_k^0$, $k \leq 2$, their leading parts and remark that $\widetilde{L}^0(x,\partial_x)$ and $\widetilde{B}_k^0(x,\partial_x)$ are model operators of the form (2.36). We denote by $\widetilde{\mathcal{A}}^0(\lambda)$ the operator pencil associated with the operator $\widetilde{A}^0 = (\widetilde{L}^0, \widetilde{B}_0^0, \widetilde{B}_1^0)$ and by $\widetilde{\mathcal{A}}$ the pencil associated to $\widetilde{A} = (\widetilde{L}, \widetilde{B}_0, \widetilde{B}_1)$, obtained in the same way as in the preceding section. In this setting, the estimate given in the following theorem holds.

Theorem 3. *Let u be a solution of* (2.42)*. If the operator pencil $\widetilde{\mathcal{A}}^0(\lambda)$ has no eigenvalues on the line $\{\operatorname{Re}\lambda = \eta\}$, $\eta = \gamma - d/2$, then the regularity estimate*

$$\|u\|_{\mathcal{K}^s_\gamma(\Omega)} \leq C\left(\|f\|_{\mathcal{K}^{s-2}_{\gamma-2}(\Omega)} + \sum_j \|g_j\|_{\mathcal{K}^{s-k-1/2}_{\gamma-k-1/2}(\Omega)} + \|u\|_{\mathcal{K}^{s-1}_{\gamma-1}(\Omega)}\right) \tag{2.43}$$

holds.

Proof. First, note that for any $\varepsilon > 0$ we can consider a function $v \in \mathcal{K}^s_\gamma(\Omega)$ with support contained in a ball $B(0,\delta)$ such that, using definition (2.22) for the norm in $\mathcal{K}^s_\gamma(\Omega)$,

$$\begin{aligned}\|(\widetilde{L} - \widetilde{L}^0)v\|_{\mathcal{K}^s_\gamma(\Omega)} &\leq C \int_{\mathbb{R}_+} r^{-2\gamma+d-1} \sum_{\alpha\leq 2} \|r^{|\alpha|-2}(r\partial_r)^j \left[(p_\alpha(r,\omega) - p_\alpha(0,\omega))v\right] \|_{H^{s-j}(S)} dr \\ &\leq \varepsilon \|v\|_{\mathcal{K}^s_\gamma(\Omega)},\end{aligned}$$

where the second inequality holds because of (2.41), by taking δ small enough. The same kind of estimate holds for $\widetilde{B}_k$, $k \leq 2$, hence it holds for the whole operator $\widetilde{A}$.

Take now a smooth cutoff function $\chi \in C^\infty(\Omega)$ such that $\operatorname{supp}\chi \subset B(0,\delta)$. Then, there exists another cutoff function ξ, with $\operatorname{supp}\xi \subset B(0,\delta')$, such that the commutator $\widetilde{A}\chi - \chi\widetilde{A} = [\widetilde{A},\chi] = 0$ in $\Omega \setminus \operatorname{supp}\xi$. From (2.42), we obtain therefore

$$\widetilde{A}^0(\chi u) = (\widetilde{A}^0 - \widetilde{A})(\chi u) + [\widetilde{A},\chi](\eta u) + \chi\widetilde{A}u. \tag{2.44}$$

First, we note that $[\widetilde{A},\chi] : \mathcal{K}^s_\gamma(\Omega) \to \mathcal{K}^{s-1}_{\gamma-1}(\Omega)$ continuously. Then, since $\widetilde{A}^0$ is an operator of type (2.36), we can apply Theorem 2 to (2.44) and obtain the estimate

$$\|\chi u\|_{\mathcal{K}^s_\gamma(\Omega)} \leq C \left(\varepsilon \|\chi u\|_{\mathcal{K}^s_\gamma(\Omega)} + \|\eta u\|_{\mathcal{K}^{s-1}_{\gamma-1}(\Omega)} + \|\chi f\|_{\mathcal{K}^{s-2}_{\gamma-2}(\Omega)} + \sum_j \|\chi g_j\|_{\mathcal{K}^{s-k-1/2}_{\gamma-k-1/2}(\Omega)} \right).$$

We can choose ε to be sufficiently small so that we can kick back the first term at the right hand side. In addition, outside of the support of χ the weighted spaces correspond to their classical Sobolev counterpart, so that classical elliptic estimates can be used to conclude with (2.43). □

All the results obtained in this section hold for any domain that is conical in the vicinity of singular points and with a smooth boundary elsewhere. Furthermore, even though we have treated the case of a single singular point, the generalization to the case of multiple isolated points is straightforward, since it suffices to construct a space which is a weighted Sobolev space around every singular point, and a classical Sobolev space elsewhere.

The case of regularity in non homogeneous weighted Sobolev spaces $\mathcal{J}^s_\gamma(\Omega)$ will be treated in the forthcoming sections, as we will show detailed results for the problems that are the focus of our analyses.

Chapter 3

The hp discontinuous Galerkin method

Discontinuous Galerkin (dG) methods are a class of finite element methods where no continuity requirement is enforced on the approximation space. Specifically, the finite element space is constructed by considering each element in the computational mesh separately, and by imposing that the finite element function is a polynomial inside the element. As such, they are a *non conforming* method, in that the approximation space X_δ is not contained in the space X where the exact solution lies and where the problem is well posed. In elliptic problems, the bilinear form associated to the method has then to be modified with respect to the continuous one, to account for the nonconformity of the method and in order to preserve consistency and stability. Discontinuous Galerkin methods allow for the straightforward treatment of discontinuities in the coefficients, of grids with hanging nodes (even containing elements of different type), and of different polynomial degrees between elements. In hyperbolic problems, the emergence of discontinuities in the solution is a classical phenomenon, and the design of an inter-element flux has a physical meaning, on top of the mathematical one. In this context, dG methods can be seen as a generalization of finite volume methods, and, for properly designed fluxes, they are conservative, differently from classical continuous finite element methods.

Historically, dG methods have been introduced for the approximation of first order steady equations in [RH73] in the context of neutron transport equations. For second order elliptic problems, the development of discontinuous Galerkin methods is based on the ideas in [Nit72], with interior penalty methods being introduced in [Whe78] and developed in [Arn82]. A wide range of different methods have been proposed throughout the years, including, among others, the local discontinuous Galerkin (LDG) method [CS98], and the already mentioned class of interior penalty (IP) methods, in its symmetric (SIP), nonsymmetric (NIP) and incomplete (IIP) versions. See [Riv08; HW08; DE12] for an overview of discontinuous Galerkin methods.

The hp version of finite element (FE) methods, introduced in [GB86a; GB86b; GB86c] in one dimension and in [GB86d; GB86e] in more dimensions, combines adaptivity in

space in low regularity regions with adaptivity in polynomial degree in high regularity regions. When applied to elliptic problems with point singularities, the numerical solutions obtained with the hp FE method can converge with exponential rate, provided that the weighted Sobolev norms of the exact solutions obey analytic-type estimates — i.e., they belong to the spaces $\mathcal{K}^{\varpi}_{\gamma}(\Omega)$ or $\mathcal{J}^{\varpi}_{\gamma}(\Omega)$ defined in Chapter 2. We also signal the recent research on hp methods in polygonal and polyhedral domains, see, among others, [CDS05; SW10; SSW13b; SSW13a; SSW16]. Finally, an estimate for the convergence of hp dG method, obtained using a slightly different functional framework, is given in [GS05].

In this setting, after a brief overview of the general framework for dG methods set in [Arn+02], we will introduce symmetric interior penalty methods directly in conjunction with hp spaces. Doing so, we can tailor the theoretical results to the regularity of our solution. In general, the classical dG analysis presumes that the continuous solutions are in $H^2(\Omega)$: this is not guaranteed in the problems we are interested in and it is therefore worthwhile to present our continuity results in weighted mesh-dependent norms. Continuity and coercivity in weighted mesh dependent norms is then shown in Section 3.2.2. The quasi optimality of the method then follows almost directly. Note that we show convergence on a couple of quite peculiar spaces; indeed, if we denote by X the continuous space, by X_δ the discrete one, and by $X(\delta) = X + X_\delta$, the continuity of a dG method is often proven as

$$a(v, v_\delta) \leq C\|v\|_{X(\delta)}\|v_\delta\|_{X_\delta},$$

where $v \in X(\delta)$ and $v_\delta \in X_\delta$. The continuity can often be extended to $X(\delta) \times X(\delta)$, but the possibility to do so depends on the regularity of the functions contained in X. In our case, we tailor our continuity result so that in can be applied to expressions of the form

$$a(u - u_\delta, v - v_\delta)$$

where $u, v \in X$, $u_\delta \in X_\delta$ are given functions and $v_\delta \in X_\delta$ is an arbitrary function we can choose. We conclude this part with some standard approximation results, applied to the weighted mesh-dependent norms we will have introduced.

We introduce some notation that will be useful throughout the analysis. First, let $\mathcal{T}$ be a shape- and contact-regular mesh. We suppose that for any $K \in \mathcal{T}$ there exists an affine transformation $\Phi : K \to \hat{K}$ to the d-dimensional cube $\hat{K}$ such that $\Phi(K) = \hat{K}$. Let then $\mathcal{E}$ be the set of the edges (for $d = 2$) or faces ($d = 3$) of the elements in $\mathcal{T}$ and $\mathcal{E}_I$ be the set of internal edges, i.e., edges not belonging to $\partial\Omega$. Let us then consider a vector of polynomial degrees $\{p_K \in \mathbb{N}, K \in \mathcal{T}\}$ and denote by h_K the diameter of $K \in \mathcal{T}$. Then, for all $e \in \mathcal{E}$, we define

$$\mathtt{h_e} = \min_{K \in \mathcal{T} : e \cap \partial K \neq \varnothing} h_K$$
$$\mathtt{p_e} = \max_{K \in \mathcal{T} : e \cap \partial K \neq \varnothing} p_K.$$

On an edge/face between two elements $K_\sharp$ and $K_\flat$, i.e., on $e \subset \partial K_\sharp \cap \partial K_\flat$, the average

$\{\!\{\cdot\}\!\}$ and jump $[\![\cdot]\!]$ operators for a function $w \in X(\delta)$ are defined by

$$\{\!\{w\}\!\} = \frac{1}{2}\left(w_{|_{K_\sharp}} + w_{|_{K_\flat}}\right), \qquad [\![w]\!] = w_{|_{K_\sharp}}\mathbf{n}_\sharp + w_{|_{K_\flat}}\mathbf{n}_\flat,$$

where $\mathbf{n}_\sharp$ (resp. $\mathbf{n}_\flat$) is the outward normal to the element $K_\sharp$ (resp. $K_\flat$). Finally, to simplify the notation, we write

$$\int_{\mathcal{T}} \cdot = \sum_{K\in\mathcal{T}} \int_K \cdot \qquad \int_{\mathcal{E}} \cdot = \sum_{e\in\mathcal{E}} \int_e \cdot \qquad \int_{\mathcal{E}_I} \cdot = \sum_{e\in\mathcal{E}_I} \int_e \cdot$$

3.1 The discontinuous Galerkin method

3.1.1 A general framework

In [Arn+02], the authors propose a unified framework for the definition and analysis of dG methods. We give a brief overview of their formalism, glossing over some of the details; the interested reader can find the full derivation in the cited paper. We consider, in a domain $\Omega \subset \mathbb{R}^d$, the Poisson problem with homogeneous Dirichlet boundary conditions

$$-\Delta u = f \text{ in } \Omega, \qquad u = 0 \text{ on } \partial\Omega \tag{3.1}$$

and rewrite it in mixed form as

$$\sigma = \nabla u \qquad -\nabla\cdot\sigma = f \qquad u_{|_{\partial\Omega}} = 0. \tag{3.2}$$

Let us now introduce the discrete spaces

$$V_\delta = \left\{v_\delta \in L^2(\Omega) : (v_{|_K} \circ \Phi^{-1}) \in \mathbb{Q}_{p_K}(\hat{K}),\ \forall K \in \mathcal{T}\right\}$$
$$W_\delta = \left\{v_\delta \in (L^2(\Omega))^d : (v_{|_K} \circ \Phi^{-1}) \in (\mathbb{Q}_{p_K}(\hat{K}))^d,\ \forall K \in \mathcal{T}\right\},$$

where $\mathbb{Q}_{p_K}$ is the space of polynomials of maximal degree p_K in any variable.

Writing the weak form of (3.2) inside an element $K \in \mathcal{T}$ and summing over the whole domain, we obtain the general formulation of the discrete problem, i.e., that of finding $u_\delta \in X_\delta$, $\sigma_\delta \in W_\delta$ such that

$$\int_\Omega \sigma_\delta\cdot\tau_\delta = -\int_{\mathcal{T}} u_\delta \nabla\cdot\tau_\delta + \sum_{K\in\mathcal{T}}\int_{\partial K} \hat{u}_K n_K\cdot\tau_\delta$$
$$\int_{\mathcal{T}} \sigma_\delta\cdot\nabla v_\delta = \int_\Omega f v_\delta + \sum_{K\in\mathcal{T}}\int_{\partial K} \hat{\sigma}_K\cdot n_K v_\delta,$$

for all $(\tau_\delta, v_\delta) \in V_\delta \times W_\delta$. Here n_K is the outward normal of element K and $\hat{u}$ and $\hat{\sigma}$ are numerical fluxes, whose determination will define the specific dG method employed. In

the general case, the components of numerical fluxes belong to $\Pi_{K\in\mathcal{T}} L^2(\partial K)$ and are therefore double-valued on every internal edge $e \in \mathcal{E}_I$. With some manipulation of the sums of boundary terms, it can be shown that the two equations above are equivalent to

$$\int_\Omega \sigma_\delta \cdot \tau_\delta = -\int_\mathcal{T} u_\delta \nabla \cdot \tau_\delta + \int_\mathcal{E} [\![\hat{u}]\!] \{\!\{\tau_\delta\}\!\} \int_{\mathcal{E}_I} \{\!\{\hat{u}\}\!\} [\![\tau_\delta]\!] \tag{3.3}$$

$$\int_\mathcal{T} \sigma_\delta \cdot \nabla v_\delta - \int_\mathcal{E} \{\!\{\hat{\sigma}\}\!\} [\![v_\delta]\!] - \int_{\mathcal{E}_I} [\![\hat{\sigma}]\!] \{\!\{v_\delta\}\!\} = \int_\Omega f v_\delta. \tag{3.4}$$

At this point, we want to go back to the so called "primal formulation", i.e., the formulation in primal variables. In order to do this, we use the first equation to express σ_δ in function of u_δ, and use this definition in the second equation. We introduce the lifting operators $r : (L^2(\mathcal{E}))^d \to W_\delta$ and $l : L^2(\mathcal{E}_\mathcal{I}) \to W_\delta$ such that

$$\int_\Omega r(\psi) \cdot \tau = -\int_\mathcal{E} \psi \cdot \{\!\{\tau\}\!\} \qquad \int_\Omega l(\varphi) \cdot \tau = -\int_{\mathcal{E}_I} \varphi [\![\tau]\!].$$

Through some manipulations, equation (3.3) can then be rewritten as

$$\sigma_\delta = \nabla u_\delta - r([\![\hat{u} - u_\delta]\!]) - l(\{\!\{\hat{u} - u_\delta\}\!\})$$

where the gradient has to be interpreted as an elementwise operator. Injecting the relation above into (3.4), we obtain

$$\int_\mathcal{T} \nabla u_\delta \cdot \nabla v_\delta + \int_\mathcal{E} \{\!\{\nabla v_\delta\}\!\} [\![\hat{u} - u_\delta]\!] - \{\!\{\hat{\sigma}\}\!\} [\![v_\delta]\!] + \int_{\mathcal{E}_I} [\![\nabla v_\delta]\!] \{\!\{\hat{u} - u_\delta\}\!\} - [\![\hat{\sigma}]\!] \{\!\{v_\delta\}\!\}. \tag{3.5}$$

Different choices of numerical fluxes $\hat{u}$ and $\hat{\sigma}$ lead to different dG methods. In our case, we are interested in the class of interior penalty methods, in particular in the symmetric version of interior penalty methods. We present it in the next section.

3.1.2 Interior penalty methods

We introduce a weight $\alpha : \mathcal{E} \to \mathbb{R}$ such that for all $e \in \mathcal{E}$, $\alpha_{|_e} = \alpha_e \geq 0$. We choose the fluxes such that they have the same value on both sides of every edge $e \in \mathcal{E}$, and in particular such that

$$\hat{u} = \{\!\{u_\delta\}\!\} \text{ in } \mathcal{E}_I, \quad \hat{u} = 0 \text{ in } \partial\Omega$$

$$\hat{\sigma} = \{\!\{\nabla u_\delta\}\!\} - \alpha \frac{\mathtt{p_e}^2}{\mathtt{h_e}} [\![u_\delta]\!] \text{ in } \mathcal{E},$$

in (3.5), the bilinear form of the dG method reads

$$d_\delta(u_\delta, v_\delta) = \int_\mathcal{T} \nabla u_\delta \cdot \nabla v_\delta - \int_\mathcal{E} [\![u_\delta]\!] \{\!\{\nabla v_\delta\}\!\} - \int_{\mathcal{E}_I} \{\!\{\nabla u_\delta\}\!\} [\![v_\delta]\!] + \int_\mathcal{E} \alpha \frac{\mathtt{p_e}^2}{\mathtt{h_e}} [\![u_\delta]\!] [\![v_\delta]\!].$$

This is the classical bilinear form associated with the SIP method for the approximation of a Laplacian. The numerical approximation to problem (3.1) would then consist in finding a $u_\delta \in V_\delta$ such that

$$d_\delta(u_\delta, v_\delta) = (f, v_\delta), \tag{3.6}$$

for all $v_\delta \in X_\delta$. In the classical case of smooth domains and regular exact solutions, the bilinear for d_δ is consistent, stable, continuous, and a quasi optimality result can be proven on a mesh dependent norm of the error of the numerical method.

A remarkable property of the SIP method is that it is *adjoint consistent*. This implies, from a theoretical point of view, that proofs that involve an Aubin-Nitsche duality method can generally be replicated for the SIP method. Furthermore, from the computational point of view, if the continuous operator is self adjoint, then the SIP method gives rise to symmetric matrices, which are in general easier to treat than non symmetric ones. On the other side, the stability of the method is assured only when α defined above is uniformly bigger than some $\alpha_{\min}$, which is not always possible to estimate in theory, especially on meshes with a wide variability of shapes and sizes.

We now turn to the introduction of the hp discontinuous Galerkin method for problems with point singularities. We present it from a point of view that tightly couples the weighted spaces and the finite element space on which the numerical solution is computed.

3.2 The discontinuous hp SIP method for problems with point singularities

Consider the approximation of the elliptic problem given by

$$\begin{aligned} Lu = (-\Delta + V)u = f \text{ in } \Omega \\ Bu = 0 \text{ on } \partial\Omega. \end{aligned} \tag{3.7}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded domain. We suppose that there is a set of isolated points $\mathfrak{C} \subset \Omega$, and that for all $\mathfrak{c} \in \mathfrak{C}$ there exists $\varepsilon > 0$ such that for all multi index $\alpha \in \mathbb{N}^d$,

$$\| |x - \mathfrak{c}|^{2-\varepsilon+|\alpha|} \partial^\alpha V(x) \|_{L^\infty(\Omega)} \leq C$$

and

$$\| |x - \mathfrak{c}|^{2-\varepsilon+|\alpha|} \partial^\alpha f(x) \|_{L^\infty(\Omega)} \leq C,$$

with $C = C(|\alpha|)$. We suppose that B is an operator of order $k \leq 1$. Furthermore, the potential V is strictly positive, i.e., $V > 0$ in Ω. Finally, we suppose that (3.7) has a unique solution in the space $X = H^1(\Omega)$.

We introduce on a set $D \subset \mathbb{R}^d$ the homogeneous weighted norm

$$\|u\|^p_{\mathcal{K}^{k,p}_\gamma(D)} = \sum_{j=0}^{k} \sum_{|\alpha|=j} \|r^{j-\gamma} \partial^\alpha u\|^p_{L^p(D)} \tag{3.8}$$

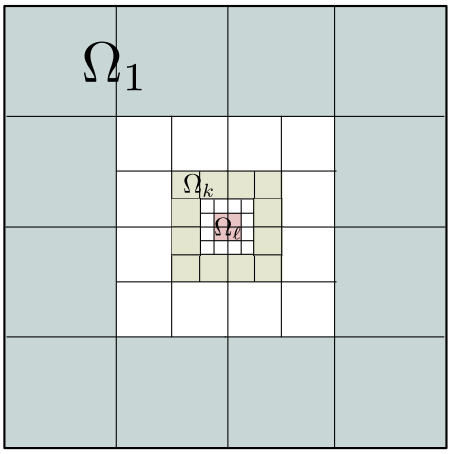


Figure 3.1 – Mesh with domains Ω_k shown, for $d = 2$.

with seminorm

$$|u|^p_{\mathcal{K}^{k,p}_\gamma(D)} = \sum_{|\alpha|=k} \|r^{k-\gamma}\partial^\alpha u\|^p_{L^p(D)}$$

and denote the space $\mathcal{K}^{k,p}_\gamma(\Omega)$ as the space of all functions with bounded $\mathcal{K}^{k,p}_\gamma(\Omega)$ norm. We also introduce the inhomogeneous norm

$$\|u\|^p_{\mathcal{J}^{m,p}_\gamma(\Omega)} = \sum_{j=0}^{m} \sum_{|\alpha|=j} \|r^{\max(-\gamma+|\alpha|,\rho)}\partial^\alpha u\|^p_{L^p(\Omega)}, \tag{3.9}$$

for $\gamma - d/p < m$ and $\rho \in (-d/p, -\gamma+m]$, if $1 \le p < \infty$, $\rho \in [0, -\gamma+m]$ if $p = \infty$. We write $\mathcal{K}^k_\gamma(\Omega) = \mathcal{K}^{k,2}_\gamma(\Omega)$ and $\mathcal{J}^k_\gamma = \mathcal{J}^{k,2}_\gamma$. We postpone all consideration about the regularity of the solution to Chapter 4; in the following, to fix ideas, one can assume that

$$u \in \mathcal{J}^{2,p}_\gamma(\Omega)$$

for any $p \ge 2$ and $\gamma < d/p + \varepsilon$.

We now specify some further condition on our mesh and space. Let $\mathcal{T}$ be a mesh isotropically and geometrically graded around the points in $\mathfrak{C}$. We indicate by Ω_j, $j = 1, \dots, \ell$, the set of elements and edges at the same level of refinement, see Figure 3.1. We introduce on this mesh an hp space with refinement ratio σ and linear polynomial slope s, i.e., for an element $K \in \mathcal{T}$ such that $K \in \Omega_j$,

$$h_K \simeq h_j := \sigma^j \text{ and } p_K = p_j := p_0 + \mathfrak{s}(\ell - j),$$

where h_K is the diameter of the element K. $\Phi : K \to \hat{K}$ is, again, the affine transformation of K into the d-dimensional cube $\hat{K}$ such that $\Phi(K) = \hat{K}$, and introduce the discrete hp

space

$$X_\delta = \left\{ v_\delta \in L^2(\Omega) : (v_{|_K} \circ \Phi^{-1}) \in \mathbb{Q}_{p_K}(\hat{K}) \forall K \in \mathcal{T} \right\}. \tag{3.10}$$

In the following, for an $S \subset \Omega$, we denote by $(\cdot,\cdot)_S$ the $L^2(S)$ scalar product and by $\|\cdot\|_S$ the $L^2(S)$ norm.

3.2.1 Bilinear form and mesh-dependent weighted norms

The hp symmetric interior penalty dG approximation of the solution of (3.7) is given by $u_\delta \in X_\delta$ such that

$$a_\delta(u_\delta, v_\delta) = (f, v_\delta) \qquad \forall v_\delta \in X_\delta, \tag{3.11}$$

with

$$a_\delta(u_\delta, v_\delta) = a(u_\delta, v_\delta) - \sum_{e\in\mathcal{E}_I} (\{\!\{\nabla u_\delta\}\!\}, [\![v_\delta]\!])_e - \sum_{e\in\mathcal{E}} (\{\!\{\nabla v_\delta\}\!\}, [\![u_\delta]\!])_e + \sum_{e\in\mathcal{E}} \alpha_e \frac{\mathtt{p_e}^2}{\mathtt{h_e}} ([\![u_\delta]\!], [\![v_\delta]\!])_e, \tag{3.12}$$

is the bilinear form associated to problem (3.7)

$$a(u_\delta, v_\delta) = \sum_{K\in\mathcal{T}} (\nabla u_\delta, \nabla v_\delta)_K + (V u_\delta, v_\delta)_K.$$

In the following, we introduce the mesh dependent norms that will be used throughout the chapter. First,

$$\|u\|^2_{\mathrm{DG}(D)} = \sum_{K\in\mathcal{T}\cap D} \|u\|^2_{\mathcal{J}_1^{1,2}(K)} + \sum_{e\in\mathcal{E}\cap D} \mathtt{h_e}^{-1}\mathtt{p_e}^2 \|[\![u]\!]\|^2_{L^2(e)}.$$

Then, let $\gamma \geq 2 - d - d/p$; if $1 \leq p < \infty$, we introduce

$$|||u|||^p_{\mathcal{D}^p_\gamma(D)} = \sum_{K\in\mathcal{T}\cap D} \|u\|^p_{\mathcal{J}_\gamma^{1,p}(K)} + \sum_{e\in\mathcal{E}\cap D} \mathtt{p_e}^2 \mathtt{h_e}^{1-p\gamma} \|[\![u]\!]\|^p_{L^p(e)} + \sum_{e\in\mathcal{E}\cap D} \mathtt{p_e}^{2(1-p)} \|r^{1+1/p-\gamma} \{\!\{\nabla u\}\!\}\|^p_{L^p(e)} \tag{3.13}$$

and

$$|||u|||^p_{\mathcal{G}^p_\gamma(D)} = \sum_{K\in\mathcal{T}\cap D} \|u\|^p_{\mathcal{K}_\gamma^{1,p}(K)} + \sum_{e\in\mathcal{E}\cap D} \mathtt{p_e}^2 \mathtt{h_e}^{1-p\gamma} \|[\![u]\!]\|^p_{L^p(e)} + \sum_{e\in\mathcal{E}\cap D} (1+|\log(\mathtt{h_e})|)^{2\mathfrak{d}(1-p)}\, \mathtt{p_e}^{2(1-p)} \|(1+|\log(r)|)^{2\mathfrak{d}/q} r^{1+1/p-\gamma} \{\!\{\nabla u\}\!\}\|^p_{L^p(e)} \tag{3.14}$$

where q is the Hölder conjugate to p and

$$\mathfrak{d} = \begin{cases} 1 & \text{if } \gamma = 2 - d + d/p \\ 0 & \text{if } \gamma > 2 - d + d/p. \end{cases}$$

When $p = \infty$ we define

$$\begin{aligned} |||u|||_{\mathcal{D}^\infty_\gamma(D)} = \max_{K\in\mathcal{T}\cap D} \|u\|_{\mathcal{J}^{1,\infty}_\gamma(K)} + \max_{e\in\mathcal{E}\cap D} \mathtt{h_e}^{-\gamma} \|[\![u]\!]\|_{L^\infty(e)} \\ + \max_{e\in\mathcal{E}\cap D} \mathtt{p_e}^{-2} \|r^{1-\gamma} \{\!\{\nabla u\}\!\}\|_{L^\infty(e)}, \end{aligned} \tag{3.15}$$

and

$$\begin{aligned} |||u|||_{\mathcal{G}^\infty_\gamma(D)} = \max_{K\in\mathcal{T}\cap D} \|u\|_{\mathcal{K}^{1,\infty}_\gamma(K)} + \max_{e\in\mathcal{E}\cap D} \mathtt{h_e}^{-\gamma} \|[\![u]\!]\|_{L^\infty(e)} \\ + \max_{e\in\mathcal{E}\cap D} (1 + |\log(\mathtt{h_e})|)^{-2\mathfrak{d}} \mathtt{p_e}^{-2} \|(1 + |\log(r)|)^{2\mathfrak{d}} r^{1-\gamma} \{\!\{\nabla u\}\!\}\|_{L^\infty(e)}, \end{aligned} \tag{3.16}$$

where $\mathfrak{d} = 1$ if $\gamma = 2 - d$, $\mathfrak{d} = 0$ if $\gamma > 2 - d$. Let us now consider the broken spaces

$$\mathcal{K}^{s,p}_\gamma(\mathcal{T}) = \left\{ v \in L^p(\Omega) : v \in \mathcal{K}^{s,p}_\gamma(K), \forall K \in \mathcal{T} \right\}$$

and

$$\mathcal{J}^{s,p}_\gamma(\mathcal{T}) = \left\{ v \in L^p(\Omega) : v \in \mathcal{J}^{s,p}_\gamma(K), \forall K \in \mathcal{T} \right\}.$$

Then, for two functions $u \in \mathcal{J}^{s,p}_\gamma(\mathcal{T})$ and $v \in \mathcal{K}^{s,p}_\gamma(\mathcal{T})$, $\gamma > 2-d+d/p$, the norms $|||u|||_{\mathcal{D}^p_\gamma(\Omega)}$ and $|||v|||_{\mathcal{G}^p_\gamma(\Omega)}$ are bounded.

The norms (3.13) and (3.14) are slightly different from those usually considered in the analysis of dG methods, in that edge terms are weighted.

The classic $|||\cdot|||_{\mathrm{DG}}$ norm for the problem considered is thus a special case of (3.13) and we write

$$|||u|||_{\mathrm{DG}(D)} = |||u|||_{\mathcal{D}^2_1(D)}.$$

Remark 5. *Note that on X_δ and for $d \leq 3$, the two norms $|||\cdot|||_{\mathrm{DG}}$ and $\|\cdot\|_{\mathrm{DG}}$ are equivalent, thanks to the discrete trace inequality [DE12]*

$$h_e^{(1-d)/p+d/2} \|w_\delta\|_{L^p(e)} \leq C_{d,p} \|w_\delta\|_{L^2(K)}, \tag{3.17}$$

on a mesh element K, for $e \in \partial K$ and for all $w_\delta \in X_\delta$. The constant $C_{d,p}$ is bounded by $\mathtt{p_e}^2$ if $p = 2$.

3.2.2 Continuity in weighted norms and coercivity

In the following lemma we prove the coercivity and continuity of the bilinear form and the quasi optimality of the discrete solution. We introduce the space

$$Y = \mathcal{J}^{2,p}_{\gamma_p}(\Omega) + X_\delta \tag{3.18}$$

and

$$Z = \left(\mathcal{J}^{2,q}_{\gamma_q}(\Omega) + X_\delta\right) \cap \mathcal{K}^{2,q}_{\gamma_q}(\mathcal{T}) \tag{3.19}$$

where $\gamma_s = d/s + \varepsilon$. The choice of the spaces Y and Z in Lemma 11 stems from the fact that we will have to bound $a_\delta(u - u_\delta, v - v_\delta)$, for fixed $u \in \mathcal{J}^{2,p}_{d/p+\varepsilon}(\Omega)$, $v \in \mathcal{J}^{2,q}_{d/q+\varepsilon}(\Omega)$, $u_\delta \in X_\delta$ and for an arbitrary $v_\delta \in X_\delta$. It is then evident how $u - u_\delta \in Y$ and how we can choose v_δ such that $v - v_\delta \in Z$.

Lemma 11. *There exists $\alpha_{\min} > 0$ such that if $\min_e \alpha_e \geq \alpha_{\min}$, the bilinear form a_δ is coercive: for $v_\delta \in X_\delta$*

$$a_\delta(v_\delta, v_\delta) \geq C\|v_\delta\|^2_{\mathrm{DG}}. \tag{3.20}$$

If $u \in Y$ and $v \in Z$, as defined in (3.18) *and* (3.19),

$$|a_\delta(u, v)| \leq C\, |\!|\!| u |\!|\!|_{\mathcal{D}^p_\gamma(\mathcal{T})}\, |\!|\!| v |\!|\!|_{\mathcal{G}^q_{2-\gamma}(\mathcal{T})}\,, \tag{3.21}$$

for $2 - d/q - \varepsilon < \gamma \leq d/p$ and where p and q are Hölder conjugates. Furthermore, let u be the exact solution to problem (3.7) *and let u_δ satisfy* (3.11)*. Then,*

$$\|u - u_\delta\|_{\mathrm{DG}} \leq C \inf_{v_\delta \in X_\delta} |\!|\!| u - v_\delta |\!|\!|_{\mathcal{G}^2_1(\mathcal{T})}\,. \tag{3.22}$$

Proof. The coercivity of the bilinear form on $X_\delta \times X_\delta$ is classical — see, e.g., [DE12]. Let $v_\delta \in X_\delta$ and note that, by multiple applications of the Cauchy-Schwartz inequality,

$$\left|\sum_{e\in\mathcal{E}} \int_e \{\!\{\nabla v_\delta\}\!\}[\![v_\delta]\!]\right| \leq C \left(\sum_{K\in\mathcal{T}} \frac{h_K}{p_K^2}\|\nabla v_\delta\|^2_{L^2(\partial K)}\right)^{1/2} \left(\sum_{e\in\mathcal{E}} \frac{\mathtt{p_e}^2}{\mathtt{h_e}}\|[\![v_\delta]\!]\|^2_{L^2(e)}\right)^{1/2}. \tag{3.23}$$

Thanks to (3.17), we have

$$\frac{h_K}{p_K^2}\|\nabla v_\delta\|^2_{L^2(\partial K)} \leq C\|\nabla v_\delta\|^2_{L^2(K)}. \tag{3.24}$$

Furthermore,

$$a(v_\delta, v_\delta) \simeq \sum_{K\in\mathcal{T}} \|v_\delta\|^2_{H^1(K)}. \tag{3.25}$$

Using an Hölder inequality on (3.23) and combining it with (3.24) and (3.25), we can then conclude that

$$a_\delta(v_\delta, v_\delta) \gtrsim \sum_{K\in\mathcal{T}} \|v_\delta\|^2_{H^1(K)} + \sum_{e\in\mathcal{E}} \alpha_e \frac{\mathtt{p_e}^2}{\mathtt{h_e}}\|[\![v_\delta]\!]\|^2_{Lr(e)} - \frac{1}{2}\sum_{K\in\mathcal{T}} \|v_\delta\|^2_{H^1(K)} - \tilde{C}\sum_{e\in\mathcal{E}} \frac{\mathtt{p_e}^2}{\mathtt{h_e}}\|[\![v_\delta]\!]\|^2_{L^2(e)}$$

for some appropriate $\widetilde{C} > 0$. Then, if $\min_e \alpha_e \geq \alpha_{\min} > \tilde{C}$, (3.20) follows.

Let now $u \in Y$, $v \in Z$. We can then decompose $u = \tilde{u} + u_\delta$ and $v = \tilde{v} + v_\delta$, where $\tilde{u}, \tilde{v} \in C^{0,\varepsilon}(\Omega)$ and $u_\delta, v_\delta \in X_\delta$. Consider an edge/face $e \in \Omega_j$, for $j \in \{1, \ldots, \ell\}$. Then, $[\![u]\!]_{|e} = [\![u_\delta]\!]_{|e}$ and $[\![v]\!]_{|e} = [\![v_\delta]\!]_{|e}$. If $j \neq \ell$, then $\mathtt{h_e} \simeq r$; if instead $j = \ell$, $[\![u_\delta]\!]_{|e} \in \mathbb{Q}_{p_0}(e)$

and $[\![v_\delta]\!]_{|e} \in \mathbb{Q}_{p_0}(e) \cap \mathcal{K}^{2-1/p,p}_{(d-1)/p+\varepsilon}(e)$. Therefore the norms

$$\mathtt{h_e}^{1-p\gamma}(1+|\log(\mathtt{h_e})|)^{2\mathfrak{d}(1-p)}\|[\![\cdot]\!]\|^p_{L^p(e)} \simeq \|(1+|\log(r)|)^{-2\mathfrak{d}q} r^{1/p-\gamma}[\![\cdot]\!]\|^p_{L^p(e)} \tag{3.26}$$

are equivalent on Y for any $\gamma \le d/p$ and

$$\mathtt{h_e}^{1-p\gamma}\|[\![\cdot]\!]\|^p_{L^p(e)} \simeq \|r^{1/p-\gamma}[\![\cdot]\!]\|^p_{L^p(e)} \tag{3.27}$$

are equivalent on Z for any $\gamma < d/p + \varepsilon$.

The continuity estimate (3.21) can be obtained through from multiple applications of Hölder's inequality: we consider the terms in (3.12) separately. First, since $V \in \mathcal{K}^{0,\infty}_{\varepsilon-2}(\Omega)$, i.e., $\|V\|_{L^\infty(\Omega)} \lesssim r^{-2+\varepsilon}$,

$$|a(u,v)| \lesssim \sum_K \|r^{1-\gamma}\nabla u\|_{L^p(K)}\|r^{\gamma-1}\nabla v\|_{L^q(K)} + \|r^{-\gamma+\varepsilon}u\|_{L^p(K)}\|r^{-2+\gamma}v\|_{L^q(K)}.$$

Secondly,

$$\begin{aligned}
\left|\sum_e (\{\!\{\nabla u\}\!\}, [\![v]\!])_e\right| &\lesssim \sum_e \mathtt{p_e}^{-2/q}\|r^{1+1/p-\gamma}\{\!\{\nabla u\}\!\}\|_{L^p(e)}\mathtt{p_e}^{2/q}\|r^{-2+1/q+\gamma}[\![v]\!]\|_{L^q(e)} \\
&\lesssim \left(\sum_e \mathtt{p_e}^{2(1-p)}\|r^{1+1/p-\gamma}\{\!\{\nabla u\}\!\}\|^p_{L^p(e)}\right)^{1/p}\left(\sum_e \mathtt{p_e}^2\|r^{-2+1/q+\gamma}[\![v]\!]\|^q_{L^q(e)}\right)^{1/q} \\
&\lesssim \left(\sum_e \mathtt{p_e}^{2(1-p)}\|r^{1+1/p-\gamma}\{\!\{\nabla u\}\!\}\|^p_{L^p(e)}\right)^{1/p}\left(\sum_e \mathtt{p_e}^2\mathtt{h_e}^{1-q(2-\gamma)}\|[\![v]\!]\|^q_{L^q(e)}\right)^{1/q}
\end{aligned}$$

where the last inequality follows from (3.27) as long as $2-\gamma < d/q+1$. Then, if $\gamma \le d/p$,

$$\begin{aligned}
\left|\sum_e (\{\!\{\nabla v\}\!\}, [\![u]\!])_e\right| &\lesssim \sum_e \Big(\mathtt{p_e}^{-2/p}\|(1+\log(r))^{2\mathfrak{d}/p}r^{1+1/q-(2-\gamma)}\{\!\{\nabla v\}\!\}\|_{L^q(e)} \times \\
&\qquad \mathtt{p_e}^{2/p}\|(1+\log(r))^{-2\mathfrak{d}/p}r^{1/p-\gamma}[\![u]\!]\|_{L^p(e)}\Big) \\
&\lesssim \sum_e \Big(\mathtt{p_e}^{-2/p}\|(1+\log(r))^{2\mathfrak{d}/p}r^{1+1/q-(2-\gamma)}\{\!\{\nabla v\}\!\}\|_{L^q(e)} \times \\
&\qquad \mathtt{p_e}^{2/p}(1+\log(\mathtt{h_e}))^{-2\mathfrak{d}/p}\mathtt{h_e}^{1/p-\gamma}\|[\![u]\!]\|_{L^p(e)}\Big) \\
&\lesssim \left(\sum_e \mathtt{p_e}^{2(1-q)}(1+\log(\mathtt{h_e}))^{-2\mathfrak{d}q/p}\|(1+\log(r))^{2\mathfrak{d}/p}r^{1+1/q-(2-\gamma)}\{\!\{\nabla v\}\!\}\|^q_{L^q(e)}\right)^{1/q} \times \\
&\qquad \left(\sum_e \mathtt{p_e}^2\mathtt{h_e}^{1-p\gamma}\|[\![u]\!]\|^p_{L^p(e)}\right)^{1/p}
\end{aligned}$$

Finally,

$$\left|\sum_{e\in\mathcal{E}}\alpha_e\frac{\mathtt{p_e}^2}{\mathtt{h_e}}([\![u]\!],[\![v]\!])_e\right| \lesssim C\sum_e\left(\mathtt{p_e}^{2/p}\mathtt{h_e}^{1/p-\gamma}\|[\![u]\!]\|_{L^p(e)}\right)\times \left(\mathtt{p_e}^{2/q}\mathtt{h_e}^{1/q+2-\gamma}\|[\![v]\!]\|_{L^q(e)}\right)$$
$$\lesssim C\left(\sum_e \mathtt{p_e}^2\mathtt{h_e}^{1-p\gamma}\|[\![u]\!]\|^p_{L^p(e)}\right)^{1/p}\left(\sum_e \mathtt{p_e}^2\mathtt{h_e}^{1-q(2-\gamma)}\|[\![v]\!]\|^q_{L^q(e)}\right)^{1/q}.$$

The terms at the right hand sides of the last four equations are part of their respective norms and give (3.21).

We conclude with (3.22) which is a consequence of the triangle inequality

$$\|u-u_\delta\|_{\mathrm{DG}}\le\|u-v_\delta\|_{\mathrm{DG}}+\|v_\delta-u_\delta\|_{\mathrm{DG}}$$

where the second term can be estimated using (3.20), (3.21) for $p=q=2$ and $\gamma=1$, and Galerkin orthogonality. □

Remark 6. *Equations* (3.22) *and* (3.17)*, combined with the triangle inequality give*

$$|\!|\!|u-u_\delta|\!|\!|_{\mathrm{DG}}\le C\inf_{v_\delta\in X_\delta}|\!|\!|u-v_\delta|\!|\!|_{\mathrm{DG}}\,.$$

3.3 Approximation results

Lemma 12. *The following approximation result holds. Let $v\in\mathcal{J}^{s_j+1}_{\gamma'}(\Omega_j)$ and for $K\in\Omega_j$ and for any $\gamma<\gamma'$,*

$$\inf_{v_\delta\in X_\delta}|\!|\!|v-v_\delta|\!|\!|_{\mathcal{G}^2_\gamma(K)}\le Ch^{\gamma'-\gamma}p_j^{-s_j+1/2}\|v\|_{\mathcal{J}^{s_j+1}_{\gamma'}(K)}. \tag{3.28}$$

with $s_\ell=1$. Furthermore, if $v\in\mathcal{J}^{s_j+1,\infty}_{\beta'}(\Omega_j)$ and for $K\in\Omega_j$ and $\beta<\beta'$

$$\inf_{v_\delta\in X_\delta}|\!|\!|v-v_\delta|\!|\!|_{\mathcal{G}^\infty_\beta(K)}\le Ch^{\beta'-\beta}p_j^{-s_j}\|v\|_{\mathcal{J}^{s_j+1,\infty}_{\beta'}(K)}, \tag{3.29}$$

with $s_\ell=0$.

Proof. We introduce on a reference element

$$\hat K=\left\{x\in\mathbb{R}^d:1/2<x_i<1,i=1,\dots,d\right\},$$

the affine transformation $\Phi:\mathbb{R}^d\to\mathbb{R}^d$ such that $\Phi(K)=\hat K$ and denote with a hat the rescaled quantities on the reference element (e.g., $\hat v=v\circ\Phi^{-1}$). Let then $v\in H^s(\hat K)$, for

integer $s \geq 2$, and $k = 0, 1$. On $\hat{K}$,

$$\inf_{v_\delta \in \mathbb{Q}_p} \|v - v_\delta\|_{H^k(\hat{K})} \leq C p^{-s+k} \|v\|_{H^s(\hat{K})} \tag{3.30}$$

and, for an edge/face $\hat{e}$ of $\hat{K}$,

$$\inf_{v_\delta \in \mathbb{Q}_p} \|v - v_\delta\|_{H^k(\hat{e})} \leq C p^{-s+1/2+k} \|v\|_{H^s(\hat{K})}, \tag{3.31}$$

We decompose $v \in \mathcal{J}^{s_j+1}_{\gamma'}(\Omega_j)$ as

$$v = w + p_v \tag{3.32}$$

where $w \in \mathcal{K}^{s_j+1}_{\gamma'}(\Omega_j)$ and $p_v \in \mathbb{Q}_{\lfloor \gamma' - d/2 \rfloor}(\Omega_j)$. We also suppose that $p_0 > \lfloor \gamma - d/2 \rfloor$. Then, denoting $\eta = w - w_\delta$ for a $w_\delta \in \mathbb{Q}_{p_j}(K)$, and $K \in \Omega_j$

$$\begin{aligned} |||w - w_\delta|||^2_{\mathcal{G}^2_\gamma(K)} &\lesssim h_j^{d-2\gamma} \left(\|\hat{\eta}\|^2_{H^1(\hat{K})} + \sum_{\hat{e} \subset \partial \hat{K}} p_j^2 \|\hat{\eta}\|^2_{\hat{e}} + \|\hat{\nabla} \hat{\eta}\|^2_{\hat{e}} \right) \\ &\lesssim p_j^{-2s+3} h_j^{d-2\gamma} \|\hat{w}\|^2_{H^s(\hat{K})}, \end{aligned}$$

where the second inequality is a consequence of (3.30) and (3.31). Inserting the weights and scaling back,

$$\begin{aligned} |||w - w_\delta|||^2_{\mathcal{G}^2_\gamma(K)} &\lesssim p_j^{-2s+3} h_j^{d-2\gamma} h_j^{-d+2\gamma'} \|w\|^2_{\mathcal{K}^s_{\gamma'}(K)} \\ &\lesssim h^{2(\gamma'-\gamma)} p_j^{-2s_j+3} \|v\|^2_{\mathcal{J}^{s_j}_{\gamma'}(K)}. \end{aligned}$$

Since $p_j \geq \lfloor \gamma' - d/2 \rfloor$, the choice $v_\delta = w_\delta + p_v$ in (3.28) gives the desired estimate for any $K \in \Omega_j$, $j \neq \ell$. If $K \in \Omega_\ell$, the choice $v_\delta = p_v$ give

$$|||v - v_\delta|||_{\mathcal{G}^2_\gamma(K)} \leq h^{\gamma'-\gamma} \|v\|_{\mathcal{J}^2_{\gamma'}(K)},$$

for any $\gamma' > \gamma$ such that $v \in \mathcal{J}^2_{\gamma'}(K)$.

Let us prove (3.29). For $K \in \Omega_j$

$$\begin{aligned} |||v - v_\delta|||_{\mathcal{G}^\infty_\beta(K)} &\leq \|v - v_\delta\|_{\mathcal{K}^{1,\infty}_\beta(K)} \\ &\quad + \sum_{e \in \partial K} \|r^{-\beta}(v - v_\delta)\|_{L^\infty(e)} \\ &\quad + \sum_{e \in \partial K} \mathtt{p_e}^{-2} (1 + \log(\mathtt{h_e}))^{-2\mathfrak{d}} \|(1 + \log(r))^{2\mathfrak{d}} r^{1-\beta} \nabla(v - v_\delta)\|_{L^\infty(e)} \\ &\lesssim \|r^{-\beta}(v - v_\delta)\|_{L^\infty(K)} \\ &\quad + (1 + \log(\mathtt{h_e}))^{-2\mathfrak{d}} \|(1 + \log(r))^{2\mathfrak{d}} r^{1-\beta} \nabla(v - v_\delta)\|_{L^\infty(K)} \end{aligned} \tag{3.33}$$

with $\mathfrak{d} = 1$ if $\beta = 2 - d$, $\mathfrak{d} = 0$ if $\beta > 2 - d$. The second inequality holds because of the regularity of u in $\Omega \setminus \mathfrak{C}$.

We decompose v as in (3.32) and start by considering a domain Ω_j not belonging to the terminal layer (i.e., $j < \ell$). On $\hat{K}$ we can generalize the approximation result introduced in [AK99]: for every function $\hat{w} \in W^{s,\infty}(\hat{K})$, there exists a polynomial $\hat{w}_\delta \in \mathbb{Q}_{p_j}(\hat{K})$ such that

$$\|\hat{w} - \hat{w}_\delta\|_{W^{1,\infty}(\hat{K})} \leq C(1 + \log(p_j))^d p_j^{-s} \|\hat{w}\|_{W^{s,\infty}(\hat{K})}. \tag{3.34}$$

Let $K \in \Omega_j$: rescaling (3.33) to the reference element gives

$$|||w - w_\delta|||_{\mathcal{G}^\infty_\beta(K)} \lesssim h_j^{-\beta} \|\hat{r}^{-\beta}(\hat{w} - \hat{w}_\delta)\|_{L^\infty(\hat{K})} + h_j^{-\beta} \|\hat{r}^{1-\beta} \hat{\nabla}(\hat{w} - \hat{w}_\delta)\|_{L^\infty(\hat{K})}.$$

Choosing $v_\delta = w_\delta + p_v$, using (3.34), considering that $\hat{r} \simeq 1$, and rescaling back to the element K we obtain

$$\begin{aligned} |||v - v_\delta|||_{\mathcal{G}^\infty_\beta(K)} &\leq C h_j^{-\beta} p_j^{-s} (1 + \log(p_j))^d \sum_{|\alpha| \leq s} \|\hat{r}^{-\beta'} \partial^\alpha \hat{w}\|_{L^\infty(\hat{K})} \\ &\leq C h_j^{-\beta} p_j^{-s} (1 + \log(p_j))^d \sum_{i=0}^{s} h_j^{\beta' - i} \sum_{|\alpha| = i} \|r^{-\beta'} \partial^\alpha w\|_{L^\infty(K)} \end{aligned}$$

Thus, inserting the weights,

$$|||v - v_\delta|||_{\mathcal{G}^\infty_\beta(K)} \leq C h_j^{\beta' - \beta} p_j^{-s} (1 + \log(p_j))^d \|v - p_v\|_{\mathcal{K}^{s,\infty}_{\beta'}(K)},$$

for $K \in \Omega_j$, $j = 1, \ldots, \ell - 1$. Inequality (3.29) is then a consequence of the equivalence

$$\|v - v(\mathfrak{c})\|_{\mathcal{K}^m_\gamma(\Omega)} + |v(\mathfrak{c})| \simeq \|v\|_{\mathcal{J}^m_\gamma(\Omega)},$$

which holds for $m \geq 1$ and $0 < \gamma - d/2 < 1$.

Let us now consider an element in the terminal layer $K \in \Omega_\ell$. Choosing $v_\delta = p_v$ we have, for $\beta < \beta'$,

$$|||v - v_\delta|||_{\mathcal{G}^\infty_\beta(K)} \leq h^{\beta' - \beta} \|w\|_{\mathcal{K}^{1,\infty}_{\beta'}(K)} \lesssim h^{\beta' - \beta} \|v\|_{\mathcal{J}^{1,\infty}_{\beta'}(K)},$$

and this concludes the proof.

□

Chapter 4

Regularity in weighted Sobolev spaces for linear elliptic problems with singular points

4.1 Introduction and presentation of the results

In this section we consider the issues related to the regularity of solutions to linear elliptic problem with singular points. We are mainly interested by singular points as a consequence of singular potentials, but we place ourselves in the more general case of a conical domain. The analysis therefore applies also to corner domains in two and three dimensions, a situation that has been widely studied, see, among the others, [CDN12; ES97; KMR97; MR10].

While most of the literature is concerned with the analysis in homogeneous weighted Sobolev spaces, denoted here as $\mathcal{K}^{s,p}_\gamma(\Omega)$, here we focus on non homogeneous spaces, denoted as $\mathcal{J}^{s,p}_\gamma(\Omega)$. The latter spaces have been studied mainly as the domain of solutions to elliptic problems in corner domains with Neumann boundary conditions. The similarity arises from the fact that problems with singular potential and Neumann boundary problems in domains with conical points share solutions that have nontrivial Taylor expansions at the singular points.

The reason why a regularity result in non homogeneous weighted spaces is more relevant than its homogeneous counterpart lies in the fact that, by taking wider spaces — in general, $\mathcal{K}^{s,p}_\gamma(\Omega) \subset \mathcal{J}^{s,p}_\gamma(\Omega)$ — we can obtain an estimate with a bigger weight γ. This is relevant since, for example, $\mathcal{J}^{2,p}_{d/p+\alpha}(\Omega) \subset L^\infty(\Omega)$ for any $\alpha > 0$ (see Lemma 15), while $\mathcal{K}^{2,p}_{d/p}(\Omega) \not\subset L^\infty(\Omega)$.

From the point of view of the Mellin transformation, working in non homogeneous spaces consists in isolating some singularities of the transform of the solution, bounding the rest of the function using the theory of homogeneous spaces, and finally bounding the terms in the expansion of the solution corresponding to the singulari-

ties via embeddings in higher order non homogeneous spaces. To illustrate this, consider the Mellin symbol $\mathfrak{L}$ related to a Laplacian operator in a conical domain given by $\mathfrak{L}(\omega, \partial_\omega, \lambda) = -((\lambda + d - 2)\lambda + \Delta_U)$, with $-\Delta_U$ representing the Laplace-Beltrami operator on $U \subset \mathbb{S}_{d-1}$, in the case where Δ_U has a null eigenvalue (corresponding to spherically symmetric functions). The symbol has a single (resp. double) zero for $\lambda = 0$ in three (resp. two) dimensions. In three dimension, this zero corresponds to a constant in the asymptotic expansion of the solution near the singularity; in two dimensions, we would have a constant and a logarithmic term $\log(|x|)$, but the latter would not be in $H^1(\Omega)$. In the asymptotic expansion of the solution near the singularity, we will therefore find a constant, followed by a term due to the potential or the geometry of the domain. The former case depends on the asymptotic expansion of the potential near the singularity, while the latter depends on the eigenvalues of Δ_U. In the following sections, we will suppose that the term following the constant in the expansion goes as $|x|^\varepsilon$ for an $0 < \varepsilon < 1$. As an example of a potential that would generate such a behavior, consider $V(x) = |x|^{-2+\varepsilon}$. A geometry causing an expansion containing $|x|^\varepsilon$ would instead be one such that $\varepsilon(\varepsilon + d - 2) \in \sigma(-\Delta_U, B_{\partial U})$, i.e., there exists a function $\hat{u} : \mathbb{R}^+ \to H^s(U)$, $s \geq 2$ such that

$$(\mathfrak{L}\hat{u})(\lambda) = -((\lambda + d - 2)\lambda - (\varepsilon + d - 2)\varepsilon)\,\hat{u}(\lambda).$$

As it can easily be seen, ε is indeed a zero of the symbol above. More practically, this happens if we consider a two dimensional wedge with aperture π/ε, as it will be outlined later in Section 4.2.2.

Returning to weighted Sobolev spaces, in light of the analysis of the operator given above, we can consider a simple case by neglecting the higher order terms, and consider a function $v(x) = v(0) + |x|^\varepsilon$, with $v(0) \neq 0$ as a model of our solution. As long as $|x| \ll 1$, those terms are indeed the predominant ones. The norm $\|r^{-d/p} v\|_{L^p(\Omega)}$ is clearly unbounded, thus $v \notin \mathcal{K}^{s,p}_\gamma(\Omega)$ for any $s \in \mathbb{N}$ and any $\gamma \geq d/p$. It is easy to see, though, that the statement $v \in \mathcal{K}^{s,p}_\gamma(\Omega)$ for any $s \in \mathbb{N}$ and $\gamma < d/p$ does not tell the whole story, since $v - v(0) \in \mathcal{K}^{s,p}_\gamma(\Omega)$ also when $\gamma \in (d/p, d/p + \varepsilon)$. The non homogeneous weighted spaces give therefore a framework where functions such as v can be treated more naturally than in homogeneous spaces.

We define the spaces treated above in more detail and outline the relationships between the homogeneous and non homogeneous ones in the following Section 4.2.1. Then, in Section 4.2.2 we specify the class of operators of interest. The main regularity result for those operators is then given in Section 4.3. Specifically, we give an elliptic regularity result in non homogeneous weighted Sobolev spaces for operators with singular potential, and we follow with an observation on how this can be used as a basis to obtain "analytic regularity" in weighted spaces – see Corollary 17.

In the following Section 4.4 some bounds on the Green functions of our operators with singular potential are introduced. While in regular problems the values of the Green function $G(x, y)$ and of its derivatives are bounded by a (negative) power of the distance from the diagonal $\{x = y\}$, when dealing with singular points the distance from the singular point comes into play. Results in this domain have been obtained in the literature mainly by Maz'ya and coworkers – see [MP85]. The proof we give follows

the same techniques, while giving explicit constants and using the estimates from the preceding section in non homogeneous spaces. This results may be used to obtain local estimates on the norms of the solution of the elliptic problem considered; once again, working in non homogeneous weighted Sobolev spaces we obtain results that exploit the full regularity of the solution.

We conclude in Section 4.5 with an analysis on the convergence in $L^\infty(\Omega)$-type norms. We are not able to fully prove the convergence of the numerical solution to the exact one in those norms, due to the lack of local estimates for the class of hp refinements considered. Nonetheless, we propose a strategy to prove this result based on a specific assumption on the local norms, and conjecture therefore the weighted $L^\infty(\Omega)$ convergence. The results on this type of convergence will be useful in the sequel to prove some improved results on the convergence of the eigenvalues for the nonlinear eigenproblems. We test the method on some test cases obtaining results in line with our conjecture.

4.2 Notation and statement of the problem

Let us consider a domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ which will be specified later and let $\mathcal{C}$ be a set of isolated points in Ω; for the sake of simplicity we consider the case of a single point $\mathcal{C} = \{\mathfrak{c}\}$; the theory can be trivially extended to the case of a finite number of points. We then denote by $r = r(x)$ the distance $|x - \mathfrak{c}|$ where $|\cdot|$ is the euclidean norm of $\mathbb{R}^d$ (it would be a smooth function representing the distance from the nearest point in $\mathcal{C}$ if there were more than one).

We denote by $C \in \mathbb{R}^+$ a generic constant subject to change at any point, and write $A \lesssim B$ (resp. $A \gtrsim B$) if $A \leq CB$ (resp. $A \geq CB$) and $A \simeq B$ if both $A \lesssim B$ and $A \gtrsim B$ hold.

4.2.1 Weighted Sobolev spaces

Recall that on a set $\Omega \subset \mathbb{R}^d$ we have defined the homogeneous weighted norm

$$\|u\|^p_{\mathcal{K}^{k,p}_\gamma(\Omega)} = \sum_{j=0}^{k} \sum_{|\alpha|=j} \|r^{j-\gamma}\partial^\alpha u\|^p_{L^p(\Omega)} \tag{4.1}$$

with seminorm

$$|u|^p_{\mathcal{K}^{k,p}_\gamma(\Omega)} = \sum_{|\alpha|=k} \|r^{k-\gamma}\partial^\alpha u\|^p_{L^p(\Omega)}$$

and that we denote the space $\mathcal{K}^{k,p}_\gamma(\Omega)$ as the space of all functions with bounded $\mathcal{K}^{k,p}_\gamma(\Omega)$ norm. We have also introduced the non homogeneous norm

$$\|u\|^p_{\mathcal{J}^{m,p}_\gamma(\Omega)} = \sum_{j=0}^{m} \sum_{|\alpha|=j} \|r^{\max(-\gamma+|\alpha|,\rho)}\partial^\alpha u\|^p_{L^p(\Omega)}, \tag{4.2}$$

for $\gamma - d/p < m$ and $\rho \in (-d/p, -\gamma+m]$, if $1 \leq p < \infty$, $\rho \in [0, -\gamma+m]$ if $p = \infty$. We write $\mathcal{K}^k_\gamma(\Omega) = \mathcal{K}^{k,2}_\gamma(\Omega)$ and $\mathcal{J}^k_\gamma(\Omega) = \mathcal{J}^{k,2}_\gamma(\Omega)$. We also remark that, for $m \geq 1$ and $\gamma - d/2 < 0$, $\mathcal{K}^m_\gamma(\Omega) = \mathcal{J}^m_\gamma(\Omega)$. Furthermore, if $v \in \mathcal{J}^m_\gamma(\Omega)$ for $m \geq 1$ and $0 < \gamma - d/2 < 1$ (condition under which $|v(c)| \lesssim \|v\|_{\mathcal{J}^m_\gamma(\Omega)}$),

$$\|v - v(\mathfrak{c})\|_{\mathcal{K}^m_\gamma(\Omega)} + |v(\mathfrak{c})| \simeq \|v\|_{\mathcal{J}^m_\gamma(\Omega)}. \tag{4.3}$$

On the boundary, for integer $s \geq 1$, we introduce the space $\mathcal{K}^{s-1/p,p}_{\gamma-1/p}(\partial\Omega)$ (resp. $\mathcal{J}^{s-1/p,p}_{\gamma-1/p}(\partial\Omega)$) of traces of functions from $\mathcal{K}^{s,p}_\gamma(\Omega)$ (resp. $\mathcal{J}^{s,p}_\gamma(\Omega)$) with norm

$$\|u\|_{\mathcal{K}^{s-1/p,p}_{\gamma-1/p}(\partial\Omega)} = \inf\{\|v\|_{\mathcal{K}^{s,p}_\gamma(\Omega)}, v_{|\partial\Omega} = u\},$$

and

$$\|u\|_{\mathcal{J}^{s-1/p,p}_{\gamma-1/p}(\partial\Omega)} = \inf\{\|v\|_{\mathcal{J}^{s,p}_\gamma(\Omega)}, v_{|\partial\Omega} = u\}.$$

Note that on portions of the boundary not touching the singularity $\mathfrak{c}$, the weighted trace spaces coincide with Sobolev trace spaces.

The dual space to $\mathcal{K}^s_\gamma(\Omega) \cap H^1_0(\Omega)$ is denoted $\mathcal{K}^{-s}_{-\gamma}(\Omega)$ and has norm

$$\|v\|_{\mathcal{K}^{-s}_{-\gamma}(\Omega)} = \sup_{\substack{\psi \in C^\infty_0(\Omega) \\ \|\psi\|_{\mathcal{K}^s_\gamma(\Omega)} = 1}} (v, \psi)_\Omega. \tag{4.4}$$

Finally, we introduce the spaces

$$\mathcal{J}^{\varpi,p}_\gamma(\Omega) = \left\{ v \in \mathcal{J}^{\infty,p}_\gamma(\Omega) : \exists A, C \in \mathbb{R} \text{ s.t. } \|v\|_{\mathcal{J}^{k,p}_\gamma(\Omega)} \leq C A^k k!, \forall k \in \mathbb{N} \right\},$$

and

$$\mathcal{K}^{\varpi,p}_\gamma(\Omega) = \left\{ v \in \mathcal{K}^{\infty,p}_\gamma(\Omega) : \exists A, C \in \mathbb{R} \text{ s.t. } |v|_{\mathcal{K}^{k,p}_\gamma(\Omega)} \leq C A^k k!, \forall k \in \mathbb{N} \right\}.$$

In the following, for an $S \subset \Omega$, we denote by $(\cdot,\cdot)_S$ the $L^2(S)$ scalar product and by $\|\cdot\|_S$ the $L^2(S)$ norm.

4.2.2 Statement of the problem

Let us now assume that in a neighborhood of $\mathfrak{c}$, the domain $\Omega \subset \mathbb{R}^d$ is conical, i.e., there exists a ball $B_\zeta(\mathfrak{c})$ centered in $\mathfrak{c}$ with radius $\zeta > 0$ such that

$$\Omega \cap B_\zeta(\mathfrak{c}) = (0, \zeta) \times U$$

where $U \subset \mathbb{S}_{d-1}$ the $d-1$ dimensional sphere, and ∂U is smooth.

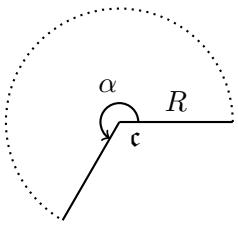


Figure 4.1 – Two dimensional wedge

In this domain we set the problem

$$\begin{aligned} L(x,\partial_x)u = -\Delta_x u + V(x)u = f \quad & \text{in } \Omega \\ B(x,\partial_x)u = 0 \quad & \text{on } \partial\Omega \end{aligned} \tag{4.5}$$

where $B(x,\partial_x)$ is a boundary operator with analytic coefficients of order $m \leq 1$ covering $L(x,\partial_x)$, i.e., such as the problem defined by (L,B) is elliptic. Furthermore, $V : \mathbb{R}^d \to \mathbb{R}_+$ is a potential such that $V \in \mathcal{K}^{\varpi,\infty}_{\varepsilon-2}(\Omega)$ for some $0 < \varepsilon \leq 1$, and $f \in \mathcal{K}^{\varpi,2}_{d/2+\varepsilon-2}(\Omega)$. We will omit the dependence of L and B on x and ∂_x when it will not be strictly necessary.

We recall the definition of the Mellin transformation

$$\hat{u}(\lambda) = (\mathcal{M}_{r\to\lambda})\, u = \int_{\mathbb{R}_+} u(r,\omega) r^{-\lambda-1} dr$$

where $(r,\omega) \in \mathbb{R}_+ \times \mathbb{S}_{d-1}$ are spherical coordinates. We denote the leading part of the operator L in (4.5) by L^0 and introduce the Mellin symbol $\mathfrak{L}(\omega,\partial_\omega,\lambda)$ of the leading part L^0, such as

$$r^{-2}\mathfrak{L}(\omega,\partial_\omega,r\partial_r) = L^0(x,\partial_x). \tag{4.6}$$

We also suppose, for ease of notation, that the smallest nonzero eigenvalue of the Laplace-Beltrami operator $-\Delta_U$ on U with boundary operator $B_{\partial U}$ on ∂U is bigger than $\varepsilon(\varepsilon+d-2)$, i.e.,

$$\min\{\mu > 0 : \mu \in \sigma(-\Delta_U, B_{\partial U})\} \geq \varepsilon(\varepsilon+d-2). \tag{4.7}$$

Since $\mathfrak{L}(\omega,\partial_\omega,\lambda) = -\left((\lambda+d-2)\lambda + \Delta_U\right)$, this condition, combined with $V \in \mathcal{K}^{\varpi,\infty}_{\varepsilon-2}(\Omega)$ guarantees that the positive pole with smallest real part of the Mellin transform of the solution $\hat{u}(\lambda)$ lies in the half-space $\{\mathrm{Re}(\lambda) \geq \varepsilon\}$. As an example of condition (4.7), consider a two dimensional domain that coincides near the origin with the wedge with angle of aperture $\alpha \in (0, 2\pi)$

$$\{0 < r < R, \vartheta \in (0,\alpha)\},$$

where r and ϑ are polar coordinates, as in Figure 4.1. On the boundary we impose either homogeneous Dirichlet or homogeneous Neumann boundary conditions. Then, (4.7) is

equivalent to

$$\alpha \leq \frac{\pi}{\varepsilon}.$$

4.3 Regularity of the solution

The first lemma concerns the regularity of the solution of (4.5): we specialize here the results of [KMR97]. We also introduce the set I_d as

$$I_d = \begin{cases} (-1, \varepsilon) \setminus \{0\} & \text{if } d = 3 \\ [0, \varepsilon) & \text{if } d = 2. \end{cases}$$

In what follows, let us analyze the set of weighted spaces where the operator (L, B) is an isomorphism. We place ourselves in the Hilbertian setting ($p = 2$). In general, we avoid considering the cases where $\gamma - d/2 \in \mathbb{N}$, since for those γ the operator is not Fredholm, the exception being $\gamma = 1$ when $d = 2$, since $\mathcal{J}^1_1(\Omega) = H^1(\Omega)$.

The idea of the proof is then to start from in homogeneous weighted spaces spaces and then to extend the results to the non homogeneous ones, by function decomposition.

Lemma 13. *The operator (L, B) is an isomorphism between the spaces*

$$\mathcal{J}^k_\gamma(\Omega) \to \mathcal{J}^{k-2}_{\gamma-2}(\Omega) \times \mathcal{J}^{k-m-1/2}_{\gamma-m-1/2}(\partial\Omega) \tag{4.8}$$

for $\gamma - d/2 \in I_d$, $k \geq 1$.

Proof. Let $\mathcal{F}_\gamma = (L_\gamma, B_\gamma) : \mathcal{K}^2_\gamma(\Omega) \to \mathcal{K}^0_{\gamma-2}(\Omega) \times \mathcal{K}^{3/2-m}_{\gamma-m-1/2}(\partial\Omega)$. The operator $\mathcal{F}_\gamma$ is Fredholm for all $\gamma - d/2 \notin \mathbb{N}$ [KMR97]; its index is defined as

$$\operatorname{ind} \mathcal{F}_\gamma = \dim(\ker \mathcal{F}_\gamma) - \dim(\ker \mathcal{F}^*_\gamma).$$

In the case $d = 3$ the index is given by

$$\operatorname{ind} \mathcal{F}_\gamma = \begin{cases} 0 & \text{if } \gamma - 3/2 \in (-1, 0) \\ -1 & \text{if } \gamma - 3/2 \in (0, \varepsilon). \end{cases}$$

When $d = 2$, instead,

$$\operatorname{ind} \mathcal{F}_\gamma = \begin{cases} 1 & \text{if } \gamma - 1 \in (-1, 0) \\ -1 & \text{if } \gamma - 1 \in (0, \varepsilon). \end{cases}$$

Let us first consider the case $\gamma - 3/2 \in (0, 1)$ and $d = 3$. The operator $\mathcal{F}_\gamma$ is coercive on $H^1(\Omega) = \mathcal{J}^1_1(\Omega) = \mathcal{K}^1_1(\Omega)$. It is then an isomorphism between the spaces $\mathcal{K}^1_1(\Omega)$ and $\mathcal{K}^{-1}_{-1}(\Omega) \times \mathcal{K}^{1/2-m}_{1/2-m}(\partial\Omega)$. Therefore, $\mathcal{F}_\gamma$ is an isomorphism between the spaces (4.8) for all $-1 < \gamma - 3/2 < 0$, see [KMR97, Corollary 6.3.3].

In the case where $\gamma = 1$ and $d = 2$, the uniqueness of the solution in $H^1(\Omega)$ implies that the operator is an isomorphims between the spaces (4.8).

Let us now consider the case $\gamma - d/2 \in (0, \varepsilon)$ and go back to the generic case $d = 2, 3$. We introduce β such that $\beta - d/2 \in (-1, 0)$ and consider a solution $u \in \mathcal{K}^s_\beta(\Omega) \cap H^1(\Omega)$, $s \geq 2$, to

$$\begin{aligned} L_\beta u &= f \quad \text{in } \Omega \\ B_\beta u &= g \quad \text{on } \partial\Omega \end{aligned}$$

for $(f, g) \in \mathcal{J}^{s-2}_{\gamma-2}(\Omega) \times \mathcal{J}^{s-1/2-m}_{\gamma-1/2-m}(\partial\Omega)$. The Mellin transform $\mathfrak{L}(\lambda)$ of the principal part of L has a single zero at $\lambda = 0$ if $d = 3$ and a double zero if $d = 2$. We can decompose u as

$$u = w + u(\mathfrak{c})$$

where $w \in \mathcal{K}^s_\gamma(\Omega)$. This is straightforward for $d = 3$; for $d = 2$ there could be a term proportional to $\log(r)$, but this term would not belong to $H^1(\Omega)$. Then, w is solution to

$$\begin{aligned} L_\gamma w &= f - Vu(\mathfrak{c}) \quad \text{in } \Omega \\ B_\gamma w &= g - B_\gamma u(\mathfrak{c}) \quad \text{on } \partial\Omega \end{aligned}$$

In this case $\operatorname{ind} \mathcal{F}_\gamma = -1$ but the right hand side in the above equation belongs to the image of $\mathcal{F}_\gamma$, by definition. Furthermore, $f - Vu(\mathfrak{c}) \in \mathcal{K}^{s-2}_{\gamma-2}(\Omega)$ and $g - B_\gamma u(\mathfrak{c}) \in \mathcal{K}^{s-1/2-m}_{\gamma-1/2-m}(\partial\Omega)$. Therefore,

$$\|w\|_{\mathcal{K}^s_\gamma(\Omega)} \leq C \left(\|f\|_{\mathcal{K}^{s-2}_{\gamma-2}(\Omega)} + \|g - Bu(\mathfrak{c})\|_{\mathcal{K}^{s-1/2-m}_{\gamma-1/2-m}(\partial\Omega)} + |u(\mathfrak{c})| \right)$$

We now conclude as in [KMR97, Theorem 7.1.1]: since for any $\delta > 0$, there exists a C_δ such that

$$|u(\mathfrak{c})| \leq \delta \|u\|_{\mathcal{J}^2_\gamma(\Omega)} + C_\delta \|u\|_{\mathcal{J}^1_{\gamma-1}(\Omega)},$$

we can write

$$\begin{aligned} \|u\|_{\mathcal{J}^s_\gamma(\Omega)} &\leq C \left(\|w\|_{\mathcal{K}^s_\gamma(\Omega)} + |u(\mathfrak{c})| \right) \\ &\leq C \left(\delta \|u\|_{\mathcal{J}^2_\gamma(\Omega)} + \|f\|_{\mathcal{K}^{s-2}_{\gamma-2}(\Omega)} + C_\delta \|u\|_{\mathcal{J}^1_{\gamma-1}(\Omega)} + \|g\|_{\mathcal{J}^{s-1/2-m}_{\gamma-1/2-m}(\partial\Omega)} \right). \end{aligned}$$

Since $\gamma - 1 \leq d/2$, by the arguments of the first part of the proof

$$\begin{aligned} \|u\|_{\mathcal{J}^1_{\gamma-1}(\Omega)} \leq \|u\|_{\mathcal{J}^2_{\gamma-1}(\Omega)} &\leq C \left(\|f\|_{\mathcal{J}^0_{\gamma-3}(\Omega)} + \|g\|_{\mathcal{J}^{3/2-m}_{\gamma-3/2-m}(\Omega)} \right) \\ &\leq C \left(\|f\|_{\mathcal{K}^{s-2}_{\gamma-2}(\Omega)} + \|g\|_{\mathcal{J}^{s-1/2-m}_{\gamma-1/2-m}(\partial\Omega)} \right) \end{aligned}$$

for all $s \geq 2$. The choice of a sufficiently small δ then concludes the proof. □

In the following lemma we extend the estimates for corner domains developed in [CDN12, Theorem 3.7] to the case of an operator with singular potential in three dimensions. The proof follows directly from the one in the cited reference.

Lemma 14. *Let $\gamma \in I_d$ and let $Lg \in \mathcal{K}^{\infty}_{\gamma-2}(\Omega)$. We consider a dyadic decomposition of Ω given by*

$$\Omega_n = \left\{ x \in \Omega : 2^{-n-1} < \|x\|_{\ell^\infty} < 2^{-n} \right\}, \quad n \geq 1$$

and denote Ω'_n as the interior of $\overline{\Omega}_{n-1} \cup \overline{\Omega}_n \cup \overline{\Omega}_{n+1}$. The estimate

$$|g|_{\mathcal{K}^{s,2}_{\gamma}(\Omega_n)} \leq C^s s! \left(\sum_{j=1}^{s-2} \frac{1}{j!} |Lg|_{\mathcal{K}^{j,2}_{\gamma-2}(\Omega'_n)} + \|g\|_{\mathcal{K}^{1,2}_{\gamma}(\Omega'_n)} \right) \tag{4.9}$$

holds, with $n \geq 2$.

Proof. Consider the reference annuli

$$\widehat{\Gamma} = \left\{ x \in \mathbb{R}^d : 1/2 < \|x\|_{\ell^\infty} < 1 \right\}$$

and

$$\widehat{\Gamma}' = \left\{ x \in \mathbb{R}^d : 1/4 < \|x\|_{\ell^\infty} < 2 \right\}.$$

In this domain, elliptic regularity [CDN10a] gives

$$|\widehat{g}|_{H^s(\widehat{\Gamma})} \leq C^s s! \left(\sum_{j=0}^{s-2} \frac{1}{j!} |\widehat{L}\widehat{g}|_{H^j(\widehat{\Gamma}')} + \|\widehat{g}\|_{H^1(\widehat{\Gamma}')} \right). \tag{4.10}$$

We can insert the weight $\widehat{r}$ since $\widehat{r} \simeq 1$ in $\widehat{\Gamma}$

$$\|\widehat{r}^{s-\gamma} \partial^s \widehat{g}\|_{\widehat{\Gamma}} \leq C^s s! \left(\sum_{j=1}^{s-2} \frac{1}{j!} |\widehat{r}^{j+2-\gamma} \partial^j \widehat{L}\widehat{g}|_{\widehat{\Gamma}'} + \|\widehat{g}\|_{\mathcal{K}^{1,2}_{\gamma}(\widehat{\Gamma}')} \right)$$

and rescale $\widehat{\Gamma}$ to Ω_k and $\widehat{\Gamma}'$ to Ω'_k via the transformation $x \simeq 2^{-k}\widehat{x}$, obtaining the thesis

$$|g|_{\mathcal{K}^{s,2}_{\gamma}(\Omega_k)} \leq C^s s! \left(\sum_{j=1}^{s-2} \frac{1}{j!} |Lg|_{\mathcal{K}^{j,2}_{\gamma-2}(\Omega'_k)} + \|g\|_{\mathcal{K}^{1,2}_{\gamma}(\Omega'_k)} \right).$$

□

We now prove an embedding result that bounds $L^\infty(\Omega)$ norms in weighted spaces with norms of higher derivatives for $p = 2$. This is simply the weighted version of the classical embedding of $H^s(\Omega)$ into $L^\infty(\Omega)$ for $s > d/2$, and the proof follows almost directly via dyadic decomposition.

Lemma 15. *Let $u \in \mathcal{J}^t_\gamma(\Omega)$ for $t > s + d/2$. Then*

$$\|u\|_{\mathcal{J}^{s,\infty}_{\gamma-d/2}(\Omega)} \leq C\|u\|_{\mathcal{J}^t_\gamma(\Omega)} \tag{4.11}$$

for any $\gamma - d/2 \notin \mathbb{N}$.

Proof. To prove (4.11) we use the fact that $\mathcal{J}^t_\gamma(\Omega) = \mathcal{K}^t_\gamma(\Omega) \oplus \mathbb{Q}_{\lfloor \gamma - d/2 \rfloor}(\Omega)$ and decompose $u = v + w$ such that

$$v \in \mathcal{K}^{s,2}_\gamma(\Omega) \quad \text{and} \quad w \in \mathbb{Q}_{\lfloor \gamma - d/2 \rfloor}(\Omega).$$

Furthermore, we have

$$\|u\|_{\mathcal{J}^t_\gamma(\Omega)} \simeq \|v\|_{\mathcal{K}^t_\gamma(\Omega)} + \|w\|_{\mathbb{Q}_{\lfloor \gamma - d/2 \rfloor}(\Omega)}$$

for any chosen norm $\|\cdot\|_{\mathbb{Q}_{\lfloor \gamma - d/2 \rfloor}(\Omega)}$, thanks to the equivalency of norms in finite dimensional spaces, see [CDN10b] and [KMR97, Theorem 7.1.1]. Then, by the triangle inequality and the definition of the norms in the weighted spaces,

$$\begin{aligned}\|u\|_{\mathcal{J}^{s,\infty}_{\gamma-d/2}(\Omega)} &\leq \|v\|_{\mathcal{J}^{s,\infty}_{\gamma-d/2}(\Omega)} + \|w\|_{\mathcal{J}^{s,\infty}_{\gamma-d/2}(\Omega)} \\ &\leq \|v\|_{\mathcal{K}^{s,\infty}_{\gamma-d/2}(\Omega)} + \|w\|_{\mathcal{J}^{s,\infty}_{\gamma-d/2}(\Omega)},\end{aligned}$$

and we consider separately the two terms at the right hand side. Consider the annuli

$$\Gamma_j = \left\{x \in \Omega : 2^{-j} < \|x\|_\infty < 2^{-j+1}\right\}, \, j \in \mathbb{N}$$

and let $\widehat{\Gamma} = \Gamma_0$. Then, scaling and using a Sobolev inequality,

$$\begin{aligned}\|v\|_{\mathcal{K}^{s,\infty}_{\gamma-d/2}(\Gamma_j)} &\simeq 2^{j(\gamma-d/2)}\|\hat{v}\|_{W^{s,\infty}(\widehat{\Gamma})} \\ &\lesssim 2^{j(\gamma-d/2)}\|\hat{v}\|_{H^t(\widehat{\Gamma})} \\ &\lesssim 2^{j(\gamma-d/2)}\|\hat{v}\|_{\mathcal{K}^t_\gamma(\widehat{\Gamma})} \\ &\simeq \|v\|_{\mathcal{K}^t_\gamma(\Gamma_j)} \\ &\lesssim \|u\|_{\mathcal{J}^t_\gamma(\Omega)},\end{aligned}$$

where the quantities with a hat are rescaled on $\widehat{\Gamma}$. Therefore

$$\|v\|_{\mathcal{K}^{s,\infty}_{\gamma-d/2}(\Omega)} = \sup_j \|v\|_{\mathcal{K}^{s,\infty}_{\gamma-d/2}(\Gamma_j)} \leq C\|u\|_{\mathcal{J}^t_\gamma(\Omega)}.$$

Since w lies in the finite dimensional space of polynomials of degree $\lfloor \gamma - d/2 \rfloor$, we can conclude with (4.11), where the constant C can depend on the domain Ω, on the dimension d and on γ. □

The weighted analytic estimates then follow for $p = \infty$. Lemma 15 directly implies the following statement.

Corollary 16. *Let $\gamma - d/2 \notin \mathbb{N}$. If $u \in \mathcal{J}^{\varpi,2}_{\gamma}(\Omega)$, then $u \in \mathcal{J}^{\varpi,\infty}_{\gamma-d/2}(\Omega)$.*

It is now evident that, using Lemmas 13 and 14, we can prove that when the right hand side and the potential of (4.5) obey analytic growth estimates on the weighted norms of the derivatives, the solution u is in the same regularity class. This is the content of the following corollary.

Corollary 17. *If u is solution to* (4.5) *with $V : \mathbb{R}^d \to \mathbb{R}_+$ such that $V \in \mathcal{K}^{\varpi,\infty}_{\varepsilon-2}(\Omega)$ for some $0 < \varepsilon \leq 1$, and $f \in \mathcal{K}^{\varpi,2}_{d/2+\varepsilon-2}(\Omega)$, then $u \in \mathcal{J}^{\varpi,2}_{\gamma}(\Omega)$ for any $\gamma < d/2 + \varepsilon$.*

Proof. Since by Lemma 13 we have that $u \in \mathcal{J}^2_{\gamma}(\Omega)$ for $\gamma \in I_d$ we can decompose $u = (u - u(\mathfrak{c})) + u(\mathfrak{c})$ and apply (4.9) to $g = u - u(\mathfrak{c})$, remarking that since $f \in \mathcal{K}^{\varpi}_{\gamma-2}(\Omega)$, $V \in \mathcal{K}^{\varpi,\infty}_{-2+\varepsilon}(\Omega)$, and $|u(\mathfrak{c})| \leq C$ by Lemma 15, then $Lg = f - Vu(\mathfrak{c}) \in \mathcal{K}^{\varpi}_{\gamma-2}(\Omega)$. Furthermore $\|u - u(\mathfrak{c})\|_{\mathcal{K}^1_{\gamma}(\Omega)} \lesssim \|u\|_{\mathcal{J}^1_{\gamma}(\Omega)} + |u(\mathfrak{c})|$. Summing the left and right hand sides of (4.9) over all Ω_k gives the existence of $C, A \in \mathbb{R}^+$ such that

$$|u|_{\mathcal{K}^s_{\gamma}(\Omega)} \leq CA^s s!,$$

for all $s \geq 2$ and $\gamma \in I_d$, thus $u \in \mathcal{J}^{\varpi}_{\gamma}(\Omega)$. □

4.4 Bounds on the Green function

We provide a bound on the Green function of problem (4.5). The proof mainly follows [MP85]; here we consider non homogeneous weighted Sobolev spaces and give explicit constants. In the following lemmas, we will denote $r_x = |x - \mathfrak{c}|$ and $r_y = |y - \mathfrak{c}|$.

Lemma 18. *Let $G(x, y)$ be the Green function of the operator L, i.e.,*

$$\begin{aligned} L(x, \partial_x)G(x, y) &= \delta(x - y) && \text{in } \Omega \\ B(x, \partial_x)G(x, y) &= 0 && \text{on } \partial\Omega, \end{aligned}$$

for $y \in \Omega$ and where $\delta(\cdot)$ is the Dirac delta distribution. Let $\tau \in (0, 1)$ and $\gamma - d/2 \in I_d$. Then, for $x \neq y$ and $r_y > 0$,

(a) if $|x - y| > \frac{\tau}{2} r_y$ and for $|\alpha| + |\beta| + d > 2$

$$|\partial^{\alpha}_y \partial^{\beta}_x G(x, y)| \leq C^{|\alpha|+|\beta|} r_x^{\min(-|\beta|+\gamma-d/2,0)} r_y^{2-d/2-\gamma-|\alpha|} |\alpha|!|\beta|! \tag{4.12}$$

If $d = 2$,

$$|G(x, y)| \leq C {r_x}^{\min(\gamma-1,0)} {r_y}^{1-\gamma} |\log(r_y)|. \tag{4.13}$$

(b) If $|x - y| \leq \frac{\tau}{2} r_y$, then

$$|\partial^{\alpha}_y \partial^{\beta}_x G(x, y)| \leq C^{|\alpha|+|\beta|} |x - y|^{2-d-|\beta|-|\alpha|} |\alpha|!|\beta|! \tag{4.14}$$

when $|\alpha| + |\beta| + d > 2$ *and*

$$|G(x,y)| \leq C|\log(|x-y|)| \tag{4.15}$$

when $d = 2$.

The constants C *depends on the dimension* d*, on the operator* L *and on* τ *but not on* r_x, r_y, α *and* β*.*

Proof. For the sake of simplicity, we will suppose in the proof that $\mathfrak{c} = 0$. Let us fix $y \in \Omega$ and $\tau \in (0,1)$. Let us furthermore introduce the fundamental solution to the Dirichlet problem for the operator L in a ball around y, i.e., the function $\Gamma(x,y)$, solution to

$$\begin{aligned} L(x,\partial_x)\Gamma(x,y) &= \delta(x-y) && \text{in } B_{\tau|y|}(y) \\ B(x,\partial_x)\Gamma(x,y) &= 0 && \text{on } \overline{B}_{\tau|y|}(y) \cap \partial\Omega. \end{aligned}$$

Then, as the coefficients of L are analytic in $B_{\tau|y|}(y)$, for any $x \in B_{\tau|y|}(y)$, $x \neq y$,

$$\begin{aligned} |\partial_x^\alpha \partial_y^\beta \Gamma(x,y)| &\leq C^{|\alpha|+|\beta|}|\alpha|!|\beta|!|x-y|^{2-d-|\alpha|-|\beta|} && \text{if } d = 3 \text{ or } d = 2 \text{ and } |\alpha|+|\beta| > 0 \\ |\Gamma(x,y)| &\leq C|\log(|x-y|)| && \text{if } d = 2, \end{aligned} \tag{4.16}$$

with constants C not depending on α or β, see [Joh50]. We define

$$R_\alpha(x,y) = \eta(x,y)\partial_y^\alpha \Gamma(x,y) - \partial_y^\alpha G(x,y),$$

where $\eta(x,y) = \widetilde{\eta}\left(\dfrac{x-y}{|y|}\right)$ and $\widetilde{\eta}$ is a cutoff function such that

$$\widetilde{\eta} \in C_0^\infty(\Omega), \quad \widetilde{\eta}(x) = 0 \text{ if } |x| > \tau/2, \quad \widetilde{\eta}(x) = 1 \text{ if } |x| < \tau/4.$$

Let then

$$\begin{aligned} f_\alpha(x,y) &= L(x,\partial_x)R_\alpha(x,y) \\ &= -\left[-\Delta_x, \eta(x,y)\right]\partial_y^\alpha \Gamma(x,y) \end{aligned} \tag{4.17}$$

It can be shown by inspection of the above equation that

$$\operatorname{supp}(f_\alpha(\cdot,y)) \subset \left(B_{\frac{\tau}{2}|y|}(y) \setminus B_{\frac{\tau}{4}|y|}(y)\right)$$

and

$$|\partial_x^\beta f_\alpha(x,y)| \leq C^{|\alpha|+|\beta|}|y|^{-d-|\beta|-|\alpha|}|\beta|!|\alpha|!. \tag{4.18}$$

if $d + |\alpha| + |\beta| > 2$ and

$$|f_0(x,y)| \leq C|\log(|y|)||y|^{-2} \tag{4.19}$$

if $d = 2$.

Lemma 13 implies that, for $\gamma - d/2 \in I_d$,

$$\|R_\alpha(\cdot, y)\|_{\mathcal{J}^2_\gamma(\Omega)} \leq C^{|\alpha|}|y|^{2-d/2-|\alpha|-\gamma}|\alpha|!. \tag{4.20}$$

when $d + |\alpha| > 2$ and

$$\|R_0(\cdot, y)\|_{\mathcal{J}^2_\gamma(\Omega)} \leq C|y|^{1-\gamma}|\log(|y|)|.$$

when $d = 2$. The weighted elliptic regularity estimate (4.9) (recalling that the norm given by the infinite sum over all Ω_j is equivalent to the norm in the full domain Ω) then implies

$$\|R_\alpha(\cdot, y)\|_{\mathcal{J}^s_\gamma(\Omega)} \leq C^{|\alpha|+s}|y|^{2-d/2-|\alpha|-\gamma}s!|\alpha|!, \tag{4.21}$$

if $d + |\alpha| + s > 4$. Thanks to Lemma 15, if $s > |\beta| + d/2$,

$$\|R_\alpha(\cdot, y)\|_{\mathcal{J}^{|\beta|,\infty}_{\gamma-d/2}(\Omega)} \lesssim \|R_\alpha(\cdot, y)\|_{\mathcal{J}^s_\gamma(\Omega)},$$

i.e., by definition of the norm,

$$\sum_{j=0}^{|\beta|}\sum_{|\delta|=j} \||x|^{\max(j-\gamma+d/2,0)}\partial^\delta_x R_\alpha(\cdot, y)\|_{L^\infty(\Omega)} \lesssim \|R_\alpha(\cdot, y)\|_{\mathcal{J}^{|\beta|+2}_\gamma(\Omega)}$$

which implies

$$|\partial^\beta_x R_\alpha(x, y)| \lesssim |x|^{\min(\gamma-d/2-|\beta|,0)}\|R_\alpha(\cdot, y)\|_{\mathcal{J}^{|\beta|+2}_\gamma(\Omega)}. \tag{4.22}$$

We can now analyze separately the cases $|x - y| < \frac{\tau}{2}|y|$ and $|x - y| > \frac{\tau}{2}|y|$.
Case $|\mathbf{x} - \mathbf{y}| \geq \frac{\boldsymbol{\tau}}{\mathbf{2}}|\mathbf{y}|$. In this case, $R_\alpha(x, y) = \partial^\alpha_y G(x, y)$. Therefore, (4.21) and (4.22) give

$$|\partial^\alpha_y\partial^\beta_x G(x, y)| \leq C^{|\alpha|+|\beta|}|x|^{\min(-|\beta|+\gamma-d/2,0)}|y|^{2-d/2-\gamma-|\alpha|}|\alpha|!|\beta|!$$

Case $|\mathbf{x} - \mathbf{y}| < \frac{\boldsymbol{\tau}}{\mathbf{2}}|\mathbf{y}|$. In this case combining (4.21), (4.22) and (4.16) gives

$$|\partial^\alpha_y\partial^\beta_x G(x, y)| \leq C^{|\alpha|+|\beta|}|x - y|^{2-d-|\beta|-|\alpha|}|\alpha|!|\beta|!$$

with the usual modification for $|\alpha| + |\beta| + d = 2$. □

Remark 7. *Since the operator is real valued, the Green function G of Lemma 18 is also solution to the adjoint problem*

$$\begin{aligned} L^*(y, \partial_y)G(x, y) &= \delta(x - y), \\ B^*(y, \partial_y)G(x, y) &= 0, \end{aligned} \tag{4.23}$$

see [MP85].

4.5 Conjecture of pointwise convergence of the hp dG method

In this section, we consider the issue of pointwise convergence for problems of the form

$$\begin{aligned} Lu &= f \quad \text{in } \Omega \\ Bu &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{4.24}$$

where L is an elliptic operator with a singular potential. Maximum norm estimates for discontinuous Galerkin methods have been obtained, among others, in [CC04; Guz06], and are based on the technique developed in [Sch98; Sch01]. Convergence estimates have been given for problems with point singularities in two dimensions in [SW78; ARS09]. All those estimates involve graded meshes but are set in the context of the h version of the finite element method. Pointwise convergence on similar graded meshes is also considered in [Dem+11], for two and three dimensions and in convex domains.

In our framework, a quasi optimality result for the maximum weighted norm of the error of the hp dG FE approximation is a result of the form

$$\|u - u_\delta\|_{L^\infty(\Omega)} + \|r^{1-\gamma}\nabla(u - u_\delta)\|_{L^\infty(\Omega)} \leq C \exp\left(-bN^{1/(d+1)}\right), \tag{4.25}$$

where N is the number of degrees of freedom of the approximation, $d = 2, 3$ is the dimension of the space, and r is the distance from the point singularity. We consider the problem introduced in Section 4.2.2 and the dG method as in Chapter 3 and adopt the notation used there. We also indicate by a prime $\cdot'$ the set comprised by a subdomain and its neighbors , i.e., for $S \subset \Omega$,

$$S' = \left\{K \in \mathcal{T} : \overline{K} \cap \overline{S} \neq \varnothing\right\},$$

with $S^{(n)} = S^{'\cdots'(n \text{ times})}$.

In the following lemma we show an inverse inequality for polynomial functions that will be useful in the sequel.

Lemma 19. *Let $w_\delta \in \mathbb{Q}_p(K)$, $K \in \mathcal{T}$. Then,*

$$h_K \|w_\delta\|_{L^\infty(K)} \leq C p_K^{d+2} h_K^{-d/2} \|w_\delta\|_{\mathcal{K}_{-1}^{-1}(K)}. \tag{4.26}$$

Proof. Let $K \in \mathcal{T}$. Consider a reference element $\hat{K}$ such that $K \simeq h_K \hat{K}$. Denoting with a hat the quantities scaled on a reference element,

$$\|w_\delta\|_{L^\infty(K)} \leq \|\hat{w}_\delta\|_{L^\infty(\hat{K})} \leq C p_K^{d+2} \|\hat{w}_\delta\|_{H^{-1}(\hat{K})}, \tag{4.27}$$

see, e.g., [Geo08]. Now, for any $w \in \mathcal{K}_{-1}^{-1}(K)$ and using the equivalence between the

$\mathcal{K}^1_1(\hat{K})$ and $H^1(\hat{K})$ norms on $C^\infty_0(\hat{K})$,

$$\begin{aligned}\|\hat{w}\|_{H^{-1}(\hat{K})} &= \sup_{\hat{\psi}\in C^\infty_0(\hat{K})} \frac{(\hat{w},\hat{\psi})_{\hat{K}}}{\|\hat{\psi}\|_{H^1(\hat{K})}} \\ &\lesssim \sup_{\hat{\psi}\in C^\infty_0(\hat{K})} \frac{(\hat{w},\hat{\psi})_{\hat{K}}}{\|\hat{\psi}\|_{\mathcal{K}^1_1(\hat{K})}} \\ &\lesssim \sup_{\psi\in C^\infty_0(K)} \frac{h_K^{-d}(w,\psi)_K}{h_K^{-d/2+1}\|\psi\|_{\mathcal{K}^1_1(K)}} \\ &\lesssim h_K^{-d/2-1}\|w\|_{\mathcal{K}^{-1}_{-1}(K)}.\end{aligned}$$

Combining the last inequality with (4.27) we obtain (4.26). □

In the following two sections, we consider separately the terms $\|u-u_\delta\|_{L^\infty(\Omega)}$ and $\|r^{1-\gamma}\nabla(u-u_\delta)\|_{L^\infty(\Omega)}$ from the left hand side of (4.25).

4.5.1 Introduction of g and ρ

To treat $\|u-u_\delta\|_{L^\infty(\Omega)}$, we start by noting that for an element $K\in\mathcal{T}$

$$\begin{aligned}\|u-u_\delta\|_{L^\infty(K)} &\le \|u-v_\delta\|_{L^\infty(K)} + \|v_\delta-u_\delta\|_{L^\infty(K)} \\ &\lesssim \|u-v_\delta\|_{L^\infty(K)} + \frac{p_K^d}{h_K^{d/2}}\|v_\delta-u_\delta\|_{L^2(K)} \\ &\lesssim \left(1+p_K^d\right)\|u-v_\delta\|_{L^\infty(K)} + \frac{p_K^d}{h_K^{d/2}}\|u-u_\delta\|_{L^2(K)}.\end{aligned} \tag{4.28}$$

We now consider the second term in the last inequality. Without loss of generality, we can restrict ourselves to the case of a domain Ω of unitary radius. Let then ρ be

$$\rho = h^{-d/2}\frac{u-u_\delta}{\|u-u_\delta\|_K}\mathbb{1}_K \tag{4.29}$$

and let us introduce $q\in\{1,\dots,\ell\}$ so that

$$\operatorname{supp}(\rho)\subset K\in\Omega_q. \tag{4.30}$$

We have therefore that

$$h_K^{-d/2}\|u-u_\delta\|_K = (\rho, u-u_\delta)_\Omega. \tag{4.31}$$

We now introduce g, solution to

$$\begin{aligned} Lg &= \rho \quad \text{in } \Omega, \\ Bg &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{4.32}$$

and define g_δ as the hp dG finite elements approximation to g.

4.5.2 Introduction of $\tilde{g}$ and $\tilde{\rho}$

We proceed similarly in order to treat $\|r^{1-\alpha}\nabla(u-u_\delta)\|_{L^\infty(K)}$: let $K \in \mathcal{T}$. Then,

$$\begin{aligned}\|r^{1-\alpha}\nabla(u-u_\delta)\|_{L^\infty(K)} &\leq \|r^{1-\alpha}\nabla(u-v_\delta)\|_{L^\infty(K)} + \|r^{1-\alpha}\nabla(v_\delta-u_\delta)\|_{L^\infty(K)} \\ &\lesssim \|r^{1-\alpha}\nabla(u-v_\delta)\|_{L^\infty(K)} + h_K^{1-\alpha}\|\nabla(v_\delta-u_\delta)\|_{L^\infty(K)}\end{aligned}$$

Thanks to the inverse inequality recalled in Lemma 19,

$$\|r^{1-\alpha}\nabla(u-u_\delta)\|_{L^\infty(K)} \lesssim \|r^{1-\alpha}\nabla(u-v_\delta)\|_{L^\infty(K)} + p_K^{2+d}h_K^{-d/2-\alpha}\|\nabla(v_\delta-u_\delta)\|_{\mathcal{K}_{-1}^{-1}(K)}.$$

Since

$$\begin{aligned}\|\nabla(u-v_\delta)\|_{\mathcal{K}_{-1}^{-1}(K)} &= \sup_{\substack{\psi\in C_0^\infty(K)\\ \|\psi\|_{\mathcal{K}_1^1(K)}=1}} (\nabla(u-v_\delta),\psi)_{L^2(K)} \\ &\lesssim \sup_{\substack{\psi\in C_0^\infty(K)\\ \|\psi\|_{\mathcal{K}_1^1(K)}=1}} (r^{1-\alpha}\nabla(u-v_\delta), r^{-1+\alpha}\psi)_{L^2(K)} \\ &\lesssim \|r^{1-\alpha}\nabla(u-v_\delta)\|_{L^\infty(K)}h_K^{d/2+\alpha} \sup_{\substack{\psi\in C_0^\infty(K)\\ \|\psi\|_{\mathcal{K}_1^1(K)}=1}} \|r^{-1}\psi\|_{L^2(K)},\end{aligned}$$

where in the last inequality, the supremum is equal to one due to the definition of the $\mathcal{K}_1^1(K)$ norm. Using a triangular inequality, we can conclude that

$$\begin{aligned}\|r^{1-\alpha}\nabla(u-u_\delta)\|_{L^\infty(K)} \lesssim (1+p_K^{2+d})\|r^{1-\alpha}\nabla(u-v_\delta)\|_{L^\infty(K)} \\ + p_K^{2+d}h_K^{-d/2-\alpha}\|\nabla(u-u_\delta)\|_{\mathcal{K}_{-1}^{-1}(K)}.\end{aligned} \tag{4.33}$$

We now define $\tilde{\rho}$ as

$$\tilde{\rho} = h_K^{-d/2}\partial_{x_i}\psi, \quad i \in \{1,\dots,d\} \tag{4.34}$$

for a function ψ such that

$$\|\psi\|_{\mathcal{K}_1^1(K)} = 1 \text{ and } \psi \in C_0^\infty(K). \tag{4.35}$$

Then, by integration by parts,

$$\begin{aligned}h_K^{-d/2}\|\partial_{x_i}(u-u_\delta)\|_{K_{-1}^{-1}(K)} &= \sup_{\substack{\psi\in C_0^\infty(K)\\ \|\psi\|_{\mathcal{K}_1^1(K)}=1}} (u-u_\delta, h_K^{-d/2}\partial_{x_i}\psi)_K \\ &= \sup_{\tilde{\rho}}(u-u_\delta,\tilde{\rho}),\end{aligned} \tag{4.36}$$

where the last supremum is over all $\tilde{\rho}$ defined as in (4.34), with ψ subject to (4.35). Similarly to what we did before, we introduce $\tilde{g}$, solution to

$$\begin{aligned} L\tilde{g} &= \tilde{\rho} \quad \text{in } \Omega \\ B\tilde{g} &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{4.37}$$

4.5.3 Local estimates

Lemma 20. *Let g be solution to* (4.32)*, with $\rho \in L^2(\Omega)$, $\mathrm{supp}(\rho) \subset \Omega_q$ and*

$$\|\rho\|_\Omega = h_q^{-d/2}.$$

Then, if $K \in \Omega_j$, with $j \notin \{q-1, q, q+1\}$, the following estimate holds

$$|g|_{\mathcal{K}^s_\gamma(K)} \leq C^s h_q^{\min(\eta-d/2,0)} h_j^{2-\gamma-\eta} s!, \tag{4.38}$$

for any η such that $\eta - d/2 \in I_d$, $\eta < 2-\gamma$ if $j = \ell$, and with multiplication by $|\log(h_j)|$ if $s = 0$ and $d = 2$. If instead $K \in \Omega'_q$,

$$\|g\|_{\mathcal{J}^2_\gamma(K)} \leq C h_q^{2-d/2-\gamma}. \tag{4.39}$$

Proof. Let $K \in \Omega_j$ where $j \notin \{q-1, q, q+1\}$. Let furthermore G be the Green function of the operator (L, B) defined in Lemma 18. Then, for $x \in K$, using the estimates on the Green function obtained in Lemma 18 and the symmetry of L,

$$\begin{aligned} \left|\partial_y^\alpha g(y)\right| &\leq \int_{\mathrm{supp}(\rho)} \left|\partial_y^\alpha G(x,y)\rho(x)\right| dx \\ &\leq C^{|\alpha|} r_y^{2-d/2-\eta-|\alpha|} |\alpha|! \int_{\Omega_q} r_x^{\min(\eta-d/2,0)} |\rho(x)| dx \\ &\leq C^{|\alpha|} r_y^{2-d/2-\eta-|\alpha|} |\alpha|! h_q^{\min(\eta-d/2,0)}. \end{aligned}$$

with multiplication by $|\log(r_y)|$ if $d = 2$ and $\alpha = 0$. Integrating over K therefore gives

$$\|r^{-\gamma+|\alpha|}\partial^\alpha g\|_K \leq C^{|\alpha|} h_q^{\min(\eta-d/2,0)} h_j^{2-\eta-\gamma} |\alpha|! \tag{4.40}$$

with multiplication by $|\log(h_j)|$ if $d = 2$ and $\alpha = 0$. This shows (4.38). Let us now consider Ω'_q. Lemma 13 gives, for $K \in \Omega'_q$,

$$\|g\|_{\mathcal{J}^2_\gamma(K)} \leq C h_q^{2-d/2-\gamma}$$

and this completes the proof. □

Lemma 21. *Let g be solution to* (4.32) *with $\rho \in \mathcal{K}^0_{\gamma-2}(\Omega)$ for any $\gamma < d/2 + \varepsilon$, $\mathrm{supp}(\rho) \subset \Omega_q$ and $\|\rho\|_\Omega = h_q^{-d/2}$. Furthermore, let g_δ be the dG finite element approximation to g, i.e.,*

$a_\delta(g - g_\delta, v_\delta) = 0$, for all $v_\delta \in X_\delta$. Then there exists $C > 0$ not depending on ℓ nor on q such that

$$|||g - g_\delta|||_{\mathrm{DG}} \leq C h_q^{1-d/2}. \tag{4.41}$$

Proof. Let us denote $Q = \mathrm{supp}(\rho) \subset \Omega_q$. From (3.22) and (3.28) we have

$$\begin{aligned}
|||g - g_\delta|||_{\mathrm{DG}}^2 &\leq C \inf_{v_\delta \in X_\delta} \sum_{K \cap Q' = \varnothing} |||g - v_\delta|||_{\mathcal{G}_1^2(K)}^2 + |||g - v_\delta|||_{\mathcal{G}_1^2(Q')}^2 \\
&\leq C \left(\sum_{K \cap Q' = \varnothing} p_K^{-2s_K+1} \|g\|_{\mathcal{J}_1^{s_K+1}(K)}^2 + p_q^{-1} \|g\|_{\mathcal{J}_1^2(Q')}^2 \right) \\
&\leq \left(\sum_{j=1}^{\ell} C^{2s_j} h_q^{2-d} s_j!^2 p_j^{-2s_j+1} + C p_q^{-1} h_q^{2-d} \right)
\end{aligned}$$

where the second inequality comes from Lemma 12 and the third is a consequence of Lemma 20 (via the choice $\eta = 1$ in (4.38)). Since for all $j = 1, \ldots, \ell$ there exists $b > 0$ such that $\inf_{s_j} C^{s_j} s_j! p_j^{-s_j} \lesssim e^{-bp_j}$,

$$\begin{aligned}
|||g - g_\delta|||_{\mathrm{DG}}^2 &\leq C \left(h_q^{2-d} \sum_{j=1}^{\ell} e^{-2bp_j} + h_q^{2-d} p_q^{-1} \right) \\
&\lesssim \left(h_q^{2-d} + h_q^{2-d} p_q^{-1} \right)
\end{aligned}$$

which gives (4.41). □

Lemma 22. *Let $\varphi \in C_0^\infty(\Omega_p')$ for some $p \in \{1, \ldots, \ell\}$ be such that*

$$\|\varphi\|_\Omega = 1 \tag{4.42}$$

Let Φ be the solution to

$$\begin{aligned}
L^*\Phi &= \varphi \quad \text{in } \Omega \\
B^*\Phi &= 0 \quad \text{on } \partial\Omega
\end{aligned}$$

Suppose that $K \in \Omega_j$, with $\Omega_j \cap \Omega_p = \varnothing$. Then, for any $s \in \mathbb{N}$, $s - d > 2$

$$\|\Phi\|_{\mathcal{J}_\gamma^{s,\infty}(K)} \leq C^s h_j^{\min(\eta-\gamma-d/2, s-\gamma)} h_p^{2-\eta} s! \tag{4.43}$$

for all $\eta - d/2 \in I_d$, $\eta - d/2 > \gamma$ if $j = \ell$. When $s = 0$ and $d = 2$ the right hand side of the inequality is multiplied by $|\log(h_j)|$.

Proof. The proof is similar to that of Lemma 20. We suppose $|\alpha| + d > 2$. Consider $K \in \Omega_j$ for $j \notin \{p-1, p, p+1\}$. In this case we use Remark 7 and write the Green

function G as the solution to

$$\begin{aligned} L^*(x,\partial_x)G(y,x) &= \delta(x-y),\\ B^*(x,\partial_x)G(y,x) &= 0. \end{aligned}$$

Then, for $y \in K$,

$$\begin{aligned} \left|\partial_y^\alpha \Phi(y)\right| &\leq \int_{\mathrm{supp}(\varphi)} \left|\partial_y^\alpha G(y,x)\varphi(x)\right| dx\\ &\leq C^{|\alpha|} r_y^{\min(\eta-d/2-|\alpha|,0)} |\alpha|! \int_{\Omega_p'} r_x^{2-d/2-\eta} |\varphi(x)| dx\\ &\leq C^{|\alpha|} r_y^{\min(\eta-d/2-|\alpha|,0)} |\alpha|! h_p^{2-\eta}. \end{aligned}$$

The weighted seminorm is then bounded by

$$\|r^{|\alpha|-\gamma}\partial^\alpha g\|_{L^\infty(K)} \leq C^{|\alpha|} h_p^{2-\eta} h_j^{\min(\eta-d/2-\gamma,|\alpha|-\gamma)} |\alpha|!.$$

Multiplication by $|\log(h_j)|$ when $\alpha = 0$ and $d = 2$ follows as usual. □

We now introduce the equivalent of Lemma 20 for $\tilde{g}$. The proof will follow along the same lines, but with a slight difference that permits to account for the term $h_K^{-\alpha}$ in (4.33).

Lemma 23. *Let $\tilde{g}$ be solution to*

$$\begin{aligned} L\tilde{g} &= \tilde{\rho} \quad \text{in } \Omega\\ B\tilde{g} &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

where $\tilde{\rho} = h_q^{-d/2}\partial_{x_i}\psi$ and, for $K \in \Omega_q$

$$\psi \in C_0^\infty(K') \qquad \|\psi\|_{\mathcal{K}_1^1(\Omega)} = 1$$

Then, if $K \in \Omega_j$ such that $\Omega_j' \cap \Omega_q = \varnothing$,

$$|\tilde{g}|_{\mathcal{K}_\gamma^s(K)} \leq C^s h_j^{2-\eta-\gamma} h_q^{\eta-d/2} s!, \tag{4.44}$$

for any $\eta \in I_d$, with $\eta < 2-\gamma$ if $j = \ell$. If instead $K \in \Omega_q'$,

$$\|\tilde{g}\|_{\mathcal{J}_\gamma^2(K)} \leq C h_q^{2-d/2-\gamma}. \tag{4.45}$$

Proof. Let $K \in \Omega_j$ be such that $\Omega_j \cap \Omega_q'' = \varnothing$ and $\eta - d/2 \in I_d$. Then, for $y \in K$, using

the estimates on the Green function obtained in Lemma 18 and integrating by parts,

$$\begin{aligned}
|\partial^\alpha \tilde{g}(y)| &\leq \left| \int_{\mathrm{supp}(\tilde{\rho})} \partial_y^\alpha G(x,y) h_q^{-d/2} \partial_{x_i} \psi(x) dx \right| \\
&\leq \int_{\mathrm{supp}(\tilde{\rho})} \left| \partial_{x_i} \partial_y^\alpha G(x,y) h_q^{-d/2} \psi(x) \right| dx \\
&\leq C^{|\alpha|} r_y^{2-|\alpha|-d/2-\eta} |\alpha|! \int_{K'} r_x^{\eta-d/2} h_q^{-d/2} r_x^{-1} |\psi(x)| dx \\
&\leq C^{|\alpha|} r_y^{2-|\alpha|-d/2-\eta} h_q^{\eta-d/2} |\alpha|!
\end{aligned}$$

with multiplication by $|\log(|y|)|$ if $d = 2$ and $\alpha = 0$. Integrating over K therefore gives

$$\|r^{-\gamma+|\alpha|} \partial^\alpha \tilde{g}\|_K \leq C^{|\alpha|} h_q^{\eta-d/2} h_j^{2-\eta-\gamma} |\alpha|! \tag{4.46}$$

while for $\alpha = 0$ and $d = 2$ we find

$$\|r^{-\gamma} \tilde{g}\|_K \leq C h_q^{\eta-1} h_j^{2-\gamma-\eta} |\log(h_j)|$$

and the logarithmic term can be absorbed into the preceding one, provided that $2-\gamma-\eta > 0$. Let us now consider Ω_q'. Lemma 13 implies, for $K \in \Omega_q'$,

$$\|\tilde{g}\|_{\mathcal{J}_\gamma^2(K)} \leq C h_q^{2-d/2-\gamma}$$

and this completes the proof. □

We also note here that $\tilde{\rho}$ fulfills the hypotheses of Lemma 21, thus

$$|||\tilde{g} - \tilde{g}_\delta|||_{\mathrm{DG}} \leq C h_q^{1-d/2}, \tag{4.47}$$

where $\tilde{g}_\delta$ is such that $a_\delta(\tilde{g} - \tilde{g}_\delta, v_\delta) = 0$, for all $v_\delta \in X_\delta$.

4.5.4 A priori estimates on the $\mathcal{D}_\gamma^1(\Omega)$ norms of $g - g_\delta$ and $\tilde{g} - \tilde{g}_\delta$

In this section we give some estimates on the functions g and $\tilde{g}$.

Let g be the solution to (4.32) with ρ defined in (4.29) and let $\tilde{g}$ be the solution to (4.37) with $\tilde{\rho}$ defined in (4.34). Let g_δ and $\tilde{g}_\delta$ be the hp dG approximations to g and $\tilde{g}$, respectively. We introduce the main assumption on which the proof is based.

Assumption 1. *There exists $m \in \mathbb{N}$ such that for any $j > m$, there exists a $\widetilde{C} = \widetilde{C}(h_j, p_j)$ and a fixed $n \in \mathbb{N}$ such that*

$$|||g - g_\delta|||_{\mathrm{DG}(\Omega_j)} \lesssim \inf_{v_\delta \in X_\delta} |||g - v_\delta|||_{\mathrm{DG}(\Omega_j')} + \widetilde{C}(h_j, p_j) \|g - g_\delta\|_{L^2(\Omega_j^{(n)})}$$

and

$$|||\tilde{g} - \tilde{g}_\delta|||_{\mathrm{DG}(\Omega_j)} \lesssim \inf_{v_\delta \in X_\delta} |||\tilde{g} - v_\delta|||_{\mathrm{DG}(\Omega_j')} + \widetilde{C}(h_j, p_j) \|\tilde{g} - \tilde{g}_\delta\|_{L^2(\Omega_j^{(n)})},$$

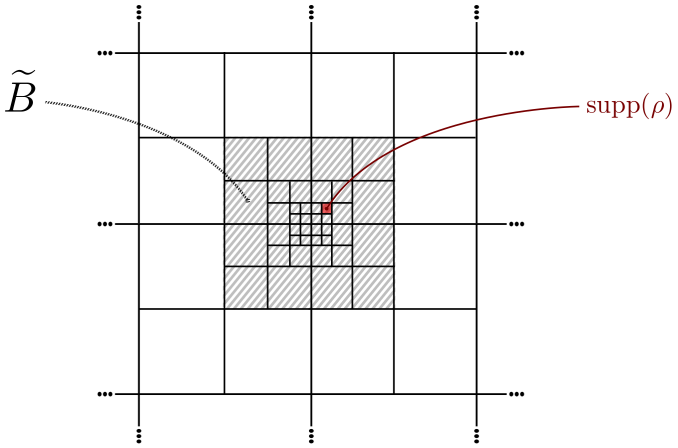


Figure 4.2 – An illustration of $\widetilde{B}$ (striped part of the domain) for $m = 2$, at a refinement step where $q = \ell - 1$.

with

$$\widetilde{C}(h_j, p_j) h_j p_j^{1/2} \to 0$$

as ℓ goes to infinity.

Lemma 24. *Let g satisfy* (4.32) *and let g_δ be the hp dG finite element approximation to g. Suppose that Assumption 1 holds and that ℓ is big enough. Then, for $\alpha \in [0, \varepsilon)$, it holds*

$$|||g - g_\delta|||_{\mathcal{D}^1_{2-\alpha}(\Omega)} \leq C \left(h_q^\alpha + e^{-b\ell} \right)$$

where the constants C and b do not depend on q.

Proof. We denote $e = g - g_\delta$. Let us consider a ball $\widetilde{B}$ centered on the singularity and containing Ω_{q-m} for a fixed m to be specified later, i.e., a ball of radius $C\sigma^{-m} h_q$, where C depends only on σ and on the dimension d. See Figure 4.2 for an illustration in two dimensions. The Cauchy-Schwartz inequality gives

$$\begin{aligned} |||e|||_{\mathcal{D}^1_{2-\alpha}(\widetilde{B})} &\leq C' \sigma^{-m} h_q^{d/2-1+\alpha} \, |||e|||_{\mathrm{DG}(\widetilde{B})} \\ &\leq C' \sigma^{-m} h_q^{d/2-1+\alpha} \, |||e|||_{\mathrm{DG}(\Omega)} , \end{aligned}$$

where we have used the fact that

$$\left(\sum_{j=q-m}^{\ell} h_j^{1+2\alpha} \right)^{1/2} \leq C \sigma^{-m} h_q^{1/2+\alpha}.$$

Using Lemma 21

$$|||e|||_{\mathcal{D}^1_{2-\alpha}(\widetilde{B})} \leq C \sigma^{-m} h_q^\alpha.$$

We now consider the elements in Ω_j, for $j < q - m$, i.e., the elements outside $\widetilde{B}$. As before, we have

$$|||e|||_{\mathcal{D}^1_{2-\alpha}(\Omega_j)} \leq C h_j^{d/2-1+\alpha} |||e|||_{\mathrm{DG}(\Omega_j)}. \tag{4.48}$$

In the following, we will estimate the term at the right hand side of this inequality. Assumption 1 gives

$$|||e|||_{\mathrm{DG}(\Omega_j)} \lesssim \inf_{v_\delta \in X_\delta} |||g - v_\delta|||_{\mathrm{DG}(\Omega'_j)} + \widetilde{C}(h_j, p_j) \|e\|_{\Omega_j^{(n)}}. \tag{4.49}$$

The estimate for the first term comes from (3.28) and (4.38): we have, for $j < q - m$,

$$\begin{aligned} \inf_{v_\delta \in X_\delta} |||g - v_\delta|||_{\mathrm{DG}(\Omega'_j)} &\leq C \|g\|_{\mathcal{J}_1^{s_j+1}(\Omega'_j)} p_j^{-s_j+1/2} \\ &\leq C^{s_j} s_j! p_j^{-s_j+1/2} h_q^{\min(\eta-d/2,0)} h_j^{1-\eta}. \end{aligned}$$

and the choice $\eta = d/2 + \alpha - \zeta$, with $0 < \zeta < \alpha$ gives

$$h_j^{d/2-1+\alpha} |||e|||_{\mathrm{DG}(\Omega_j)} \leq C h_j^{\zeta} e^{-bp_j} + \widetilde{C}(h_j, p_j) h_j^{d/2-1+\alpha} \|e\|_{\Omega_j^{(n)}}. \tag{4.50}$$

We can thus consider the second term on the right hand side of (4.50). By definition, for any $v_\delta \in X_\delta$,

$$\begin{aligned} \|e\|_{\Omega_j^{(n)}} = \sup_{\substack{\varphi \in C_0^\infty(\Omega_j^{(n)}) \\ \|\varphi\|=1}} (e, \varphi) &= a_\delta(e, \Phi - v_\delta) \\ &= a_{\delta, \Omega \setminus \Omega_j^{(n+1)}}(e, \Phi - v_\delta) + a_{\delta, \Omega_j^{(n+1)}}(e, \Phi - v_\delta), \end{aligned} \tag{4.51}$$

where Φ is the solution to the adjoint problem with right hand side $\varphi \in C_0^\infty(\Omega_j^{(n)})$

$$\begin{aligned} L^*\Phi &= \varphi \quad \text{in } \Omega \\ B^*\Phi &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

and $a_{\delta,S}(\cdot,\cdot)$ is the restriction of the bilinear form $a_\delta(\cdot,\cdot)$ to the set S. By elliptic regularity, as in the proof of Lemma 20, $\|\Phi\|_{\mathcal{J}^2_\gamma(\Omega_j^{(n+1)})} \leq C h_j^{2-\gamma}$ and

$$\begin{aligned} \inf_{v_\delta \in X_\delta} a_{\delta, \Omega_j^{(n+1)}}(e, \Phi - v_\delta) &\leq C \inf_{v_\delta \in X_\delta} |||\Phi - v_\delta|||_{\mathcal{G}^2_1(\Omega_j^{(n+1)})} |||e|||_{\mathrm{DG}(\Omega_j^{(n+1)})} \\ &\leq C p_j^{-1/2} \|\Phi\|_{\mathcal{J}^2_1(\Omega_j^{(n+1)})} |||e|||_{\mathrm{DG}(\Omega_j^{(n+1)})} \\ &\leq C p_j^{-1/2} h_j |||e|||_{\mathrm{DG}(\Omega_j^{(n+1)})}. \end{aligned} \tag{4.52}$$

Furthermore, the first term in (4.51) can be estimated by

$$\begin{aligned}\inf_{v_\delta\in X_\delta} a_{\delta,\Omega\setminus\Omega_j^{(n+1)}}(e,\Phi-v_\delta) &\le C\sum_{K\in\Omega\setminus\Omega_j^{(n+1)}} |||e|||_{\mathcal{D}^1_{2-\alpha}(K)} \inf_{v_\delta\in X_\delta} |||\Phi-v_\delta|||_{\mathcal{G}^\infty_\alpha(K)}\\ &\le C\,|||e|||_{\mathcal{D}^1_{2-\alpha}(\Omega)} \max_{k\in\{1,\dots,j-2,j+2,\dots,\ell\}} \inf_{v_\delta\in X_\delta} |||\Phi-v_\delta|||_{\mathcal{G}^\infty_\alpha(\Omega_k)}\end{aligned} \tag{4.53}$$

We introduce $\bar{k}=\bar{k}(j)$ such that

$$\inf_{v_\delta\in X_\delta} |||\Phi-v_\delta|||_{\mathcal{G}^\infty_\alpha(\Omega_{\bar{k}})} = \max_{k\in\{1,\dots,j-2,j+2,\dots,\ell\}} \inf_{v_\delta\in X_\delta} |||\Phi-v_\delta|||_{\mathcal{G}^\infty_\alpha(\Omega_k)}\,.$$

Combining equations (4.50) to (4.53) gives

$$\begin{aligned}\sum_{j=1}^{q-m} h_j^{d/2-1+\alpha}\,|||e|||_{\mathrm{DG}(\Omega_j)} &\le C\sum_{j=1}^{q-m} h_j^\zeta e^{-bp_j} + C\sum_{j=1}^{q-m}\widetilde{C}(h_j,p_j)h_j^{d/2+\alpha}p_j^{-1/2}\,|||e|||_{\mathrm{DG}(\Omega_j^{(n+1)})}\\ &\quad+\sum_{j=1}^{q-m} h_j^{d/2-1+\alpha}\widetilde{C}(h_j,p_j)\,|||e|||_{\mathcal{D}^1_{2-\alpha}(\Omega)}\inf_{v_\delta\in X_\delta}|||\Phi-v_\delta|||_{\mathcal{G}^\infty_\alpha(\Omega_{\bar{k}})}\,.\end{aligned} \tag{4.54}$$

Using again Assumption 1 and supposing ℓ is big enough, we can choose m to be big enough so that the second term of the right hand side of (4.54) is arbitrarily small. Using a first kickback argument on (4.54) gives

$$\begin{aligned}\sum_{j=1}^{q-m} h_j^{d/2-1+\alpha}\,|||e|||_{\mathrm{DG}(\Omega_j)} &\le C\sum_{j=1}^{q-m} h_j^\zeta e^{-bp_j} + C\sigma^{-m}h_q^\alpha\\ &\quad+\sum_{j=1}^{q-m} h_j^{d/2-1+\alpha}\widetilde{C}(h_j,p_j)\,|||e|||_{\mathcal{D}^1_{2-\alpha}(\Omega)}\inf_{v_\delta\in X_\delta}|||\Phi-v_\delta|||_{\mathcal{G}^\infty_\alpha(\Omega_{\bar{k}})}\,.\end{aligned} \tag{4.55}$$

Equations (4.48) to (4.55) thus give

$$\begin{aligned}|||e|||_{\mathcal{D}^1_{2-\alpha}(\Omega)} &= |||e|||_{\mathcal{D}^1_{2-\alpha}(\widetilde{B})} + \sum_{j=1}^{q-m}|||e|||_{\mathcal{D}^1_{2-\alpha}(\Omega_j)}\\ &\le C\sigma^{-m}h_q^\alpha + C\sum_{j=1}^{q-m} h_j^\zeta e^{-bp_j}\\ &\quad+|||e|||_{\mathcal{D}^1_{2-\alpha}(\Omega)}\sum_{j=1}^{q-m} h_j^{d/2-2+\alpha}\inf_{v_\delta\in X_\delta}|||\Phi-v_\delta|||_{\mathcal{G}^\infty_\alpha(\Omega_{\bar{k}})}\,.\end{aligned} \tag{4.56}$$

The sum in the last term can be bounded using the approximation result stated in Lemma

12 and the estimates of Lemma 22. Indeed,

$$\inf_{v_\delta \in X_\delta} |||\Phi - v_\delta|||_{\mathcal{G}^\infty_\alpha(\Omega_{\bar{k}})} \leq C^{s_{\bar{k}}} p_{\bar{k}}^{-s_{\bar{k}}} \|\Phi\|_{\mathcal{J}^{s_{\bar{k}}+1,\infty}_\alpha(\Omega_{\bar{k}})}.$$

Since $\bar{k} \notin \{j-1, j, j+1\}$, equation (4.43) with $\eta = d/2 + \alpha$ gives

$$\|\Phi\|_{\mathcal{J}^{s_{\bar{k}}+1,\infty}_\alpha(\Omega_{\bar{k}})} \leq C^{s_{\bar{k}}+1} h_j^{2-d/2-\alpha} s_{\bar{k}}!. \tag{4.57}$$

Therefore,

$$\inf_{v_\delta \in X_\delta} |||\Phi - v_\delta|||_{\mathcal{G}^\infty_\alpha(\Omega_{\bar{k}})} \leq C e^{-bp_{\bar{k}}} h_j^{2-d/2-\alpha}$$

and

$$\sum_{j=1}^{q-m} h_j^{d/2-2+\alpha} |||\Phi - v_\delta|||_{\mathcal{G}^\infty_\alpha(\Omega_{\bar{k}})} \leq C \sum_{j=1}^{q-m} e^{-b(\ell-j)} \leq C e^{-bm}.$$

Provided once again that m is big enough, by a second kickback argument in (4.56) we obtain

$$|||e|||_{\mathcal{D}^1_{2-\alpha}(\Omega)} \leq C\sigma^{-m} h_q^\alpha + C \sum_{j=1}^{q-m} h_j^\zeta e^{-bp_j}$$

and the thesis follows. □

We proceed now to the equivalent of Lemma 24, for $\tilde{g}$.

Lemma 25. *Let $\tilde{g}$ satisfy* (4.37) *and let $\tilde{g}_\delta$ be the hp dG finite element approximation to $\tilde{g}$. Let Assumption 1 hold. Then, for $\alpha \in [0, \varepsilon)$, it holds*

$$|||\tilde{g} - \tilde{g}_\delta|||_{\mathcal{D}^1_{2-\alpha}(\Omega)} \leq C h_q^\alpha \tag{4.58}$$

where the constant C does not depend on q.

Proof. We denote $\tilde{e} = \tilde{g} - \tilde{g}_\delta$ and use the same notation as in the proof of Lemma 24. We have then

$$|||\tilde{e}|||_{\mathcal{D}^1_{2-\alpha}(\widetilde{B})} \leq C\sigma^{-m} h_q^\alpha,$$

while in Ω_j, $j < q - m$,

$$|||\tilde{e}|||_{\mathcal{D}^1_{2-\alpha}(\Omega_j)} \leq C h_j^{d/2-1+\alpha} |||\tilde{e}|||_{\mathrm{DG}(\Omega_j)}. \tag{4.59}$$

and

$$|||\tilde{e}|||_{\mathrm{DG}(\Omega_j)} \lesssim \inf_{v_\delta \in X_\delta} |||\tilde{g} - v_\delta|||_{\mathrm{DG}(\Omega'_j)} + \widetilde{C}(h_j, p_j) \|\tilde{e}\|_{\Omega'_j}. \tag{4.60}$$

As before we use the approximation result (3.28) for the first term. In this case, though,

we use Lemma 23: for $j < q - m$,

$$\begin{aligned}\inf_{v_\delta \in X_\delta} |||\tilde{g} - v_\delta|||_{\mathrm{DG}(\Omega_j')} &\leq C\|\tilde{g}\|_{\mathcal{J}_1^{s_j+1}(\Omega_j')} p_j^{-s_j+1/2} \\ &\leq C^{s_j} s_j! p_j^{-s_j+1/2} h_j^{1-\eta} h_q^{\eta-d/2}.\end{aligned}$$

and the choice $\eta = d/2 + \alpha$ gives

$$h_j^{d/2-1+\alpha} |||\tilde{e}|||_{\mathrm{DG}(\Omega_j)} \leq C h_q^\alpha e^{-bp_j} + \widetilde{C}(h_j, p_j) h_j^{d/2-1+\alpha} \|\tilde{e}\|_{\Omega_j^{(n)}}. \tag{4.61}$$

The estimate of $h_j^{d/2-2+\alpha}\|\tilde{e}\|_{\Omega_j^{(n)}}$ is done exactly as in the proof of Lemma 24. After the second kickback, in this case we find

$$|||\tilde{e}|||_{\mathcal{D}^1_{2-\alpha}(\Omega)} \leq C\sigma^{-m} h_q^\alpha + C h_q^\alpha \sum_{j=1}^{q-m} e^{-bp_j}$$

and the sum at the right hand side of the inequality is bounded by a constant, thus giving (4.58). □

The last result allows for the conclusion of the proof of the uniform convergence of the hp dG approximation, given Assumption 1.

Conjecture 1. *Let u be the solution of* (4.5) *and let u_δ satisfy* (3.11). *Let furthermore Assumption 1 hold. Then, for $\gamma < \varepsilon$,*

$$p_{\max}^2 \|u - u_\delta\|_{L^\infty(\Omega)} + \|r^{1-\gamma}\nabla(u - u_\delta)\|_{L^\infty(\Omega)} \leq C(p_{\max}^{d+2}) \inf_{v_\delta \in X_\delta} \|u - v_\delta\|_{\mathcal{K}^{1,\infty}_\gamma(\Omega)}. \tag{4.62}$$

Furthermore, there exist $C > 0, b > 0$ such that

$$\|u - u_\delta\|_{\mathcal{J}^{1,\infty}_\gamma(\Omega)} \leq C e^{-bN^{1/(d+1)}}, \tag{4.63}$$

where N is the total number of degrees of freedom of the approximation.

Proof. Inequality (4.31), Galerkin orthogonality, Lemma 11, and Lemma 24 give, for $K \in \mathcal{T}$ and for any $v_\delta \in X_\delta$,

$$\begin{aligned}h_K^{-d/2}\|u - u_\delta\|_K &\leq C a_\delta(u - u_\delta, g) \\ &\leq C a_\delta(u - u_\delta, g - g_\delta) \\ &\leq C a_\delta(u - v_\delta, g - g_\delta) \\ &\leq C\, |||g - g_\delta|||_{\mathcal{D}^1_{2-\alpha}(\mathcal{T})}\, |||u - v_\delta|||_{\mathcal{G}^\infty_\alpha(\mathcal{T})},\end{aligned}$$

where g is solution to (4.32) and g_δ is its Galerkin projection. From (4.28) we have then

$$\begin{aligned}\|u-u_\delta\|_{L^\infty(\Omega)} &\lesssim \max_K \left(\left(1+p_K^d\right)\|u-v_\delta\|_{L^\infty(K)} + \frac{p_K^d}{h_K^{d/2}}\|u-u_\delta\|_K \right) \\ &\lesssim \max_K \left(1+p_K^d\right)\|u-v_\delta\|_{L^\infty(K)} \\ &\quad + \max_K p_K^d \left(h_K^\alpha + e^{-b\ell}\right) |||u-v_\delta|||_{\mathcal{G}_\alpha^\infty(\Omega)} .\end{aligned}$$

This concludes the estimate on $\|u-u_\delta\|_{L^\infty(\Omega)}$.

The estimate on $\|r^{1-\alpha}\nabla(u-u_\delta)\|_{L^\infty(K)}$, $\alpha<\varepsilon$ then follows the same idea: we need to estimate $h_K^{-d/2-\alpha}\|\nabla(u-u_\delta)\|_{\mathcal{K}_{-1}^{-1}(K)}$ from equation (4.33). First, given the solution $\tilde{g}$ to (4.37) with its Galerkin projection $\tilde{g}_\delta$, for any $v_\delta \in X_\delta$

$$\begin{aligned}h_K^{-d/2}\|\nabla(u-u_\delta)\|_{\mathcal{K}_{-1}^{-1}(K)} &\le Ca_\delta(u-u_\delta,\tilde{g}) \\ &\le Ca_\delta(u-u_\delta,\tilde{g}-\tilde{g}_\delta) \\ &\le Ca_\delta(u-v_\delta,\tilde{g}-\tilde{g}_\delta) \\ &\le C\,|||\tilde{g}-\tilde{g}_\delta|||_{\mathcal{D}^1_{2-\alpha}(\mathcal{T})}\,|||u-v_\delta|||_{\mathcal{G}_\alpha^\infty(\mathcal{T})} .\end{aligned}$$

Then, using the result of Lemma 25 and (4.33),

$$\begin{aligned}\|r^{1-\alpha}\nabla(u-u_\delta)\|_{L^\infty(\Omega)} &\lesssim (1+p_{\max}^{2+d})\|r^{1-\alpha}\nabla(u-v_\delta)\|_{L^\infty(\Omega)} \\ &\qquad + p_{\max}^{2+d}h_q^{-\alpha}\,|||\tilde{g}-\tilde{g}_\delta|||_{\mathcal{D}^1_{2-\alpha}(\mathcal{T})}\,|||u-v_\delta|||_{\mathcal{G}_\alpha^\infty(\mathcal{T})} \qquad (4.64) \\ &\lesssim p_{\max}^{2+d}\,|||u-v_\delta|||_{\mathcal{G}_\alpha^\infty(\mathcal{T})} .\end{aligned}$$

Using (3.29) for the function $u \in \mathcal{J}_\gamma^{\varpi,\infty}(\Omega)$, absorbing the algebraic term in $p_{\max}$ into the exponential term and using $\ell \simeq N^{1/(d+1)}$, we obtain the exponential convergence estimate (4.63).

□

4.5.5 Numerical results

In this section we test the convergence of the hp dG method on two and three dimensional test cases. All simulation are performed with a code based on the `deal.ii` library [Arn+17]. The maximum norms are approximated by the values at tensor product equispaced point inside each element. We will show estimates of the constants b_X, defined as

$$\|u-u_\delta\|_X \lesssim \exp(-b_X N^{1/(d+1)}) \qquad (4.65)$$

for $X = L^2(\Omega)$, DG, $L^\infty(\Omega)$, $\mathcal{J}_0^\infty(\Omega)$. The estimates are obtained through a linear regression of the logarithms of the errors over $N^{1/(d+1)}$, after the exclusion of the preasympotic part of the curves.

(a) (b)

Figure 4.3 – Left: Fichera corner domain. Γ_1 corresponds to the thicker boundary part, Γ_2 to the rest of the boundary. Right: computational mesh.

	b_{L^2}	b_{DG}	b_{L^∞}	$b_{\mathcal{J}_0^\infty}$
Dirichlet b.c.	0.79	0.711	0.767	0.748
Mixed b.c.	0.786	0.714	0.767	0.751

Table 4.1 – Estimated constants for problems (4.66) and (4.67).

Two dimensions

In two dimensions, we consider the Fichera domain $\Omega = (-1/2, 1/2)^2 \setminus (0, 1/2)^2$, with two boundary portions Γ_1 and Γ_2 defined as in Figure 4.3a. We introduce two problems, one with Dirichlet boundary conditions and the other with homogeneous Neumann boundary conditions on the edges next to the concave angle, and Dirichlet boundary conditions elsewhere. Let the first problem be

$$\begin{cases} -\Delta u = 0 \text{ in } \Omega \\ u_{|\partial\Omega} = g \text{ on } \partial\Omega \end{cases} \tag{4.66}$$

where g is chosen so that the exact solution is $u_{\mathrm{ex}} = r^{2/3}\sin\left(\frac{2}{3}(\vartheta - \frac{\pi}{2})\right)$, with $r = \sqrt{x^2+y^2}$ and $\vartheta = \arg(x+iy)$. In order to test Neumann boundary conditions at the corner we consider instead

$$\begin{cases} -\Delta v = 0 \text{ in } \Omega \\ \partial_n v_{|\Gamma_1} = 0 \\ v_{|\Gamma_2} = v_{\mathrm{ex}|\Gamma_2} \end{cases} \tag{4.67}$$

where $v_{\mathrm{ex}} = 1 + r^{2/3}\cos\left(\frac{2}{3}(\vartheta - \frac{\pi}{2})\right)$. The rate of convergence of the computed $L^\infty(\Omega)$ and $\mathcal{J}_0^{1,\infty}(\Omega)$ errors is compared to that of the computed $L^2(\Omega)$ and DG errors in Figure 4.4. All solutions are computed with a mesh refinement ratio $\sigma = 1/2$, polynomial degrees $p_j = \lceil \frac{1}{4} j \rceil + 2$, and a mesh refined around the corner at the origin, as in Figure 4.3b. In Figure 4.5 we show the behavior of the $\mathcal{J}_\alpha^{1,\infty}(\Omega)$ norm of the error for different α. It is evident how the condition $\alpha < \varepsilon$ in Theorem 1 is sharp. In Table 4.1 we give an

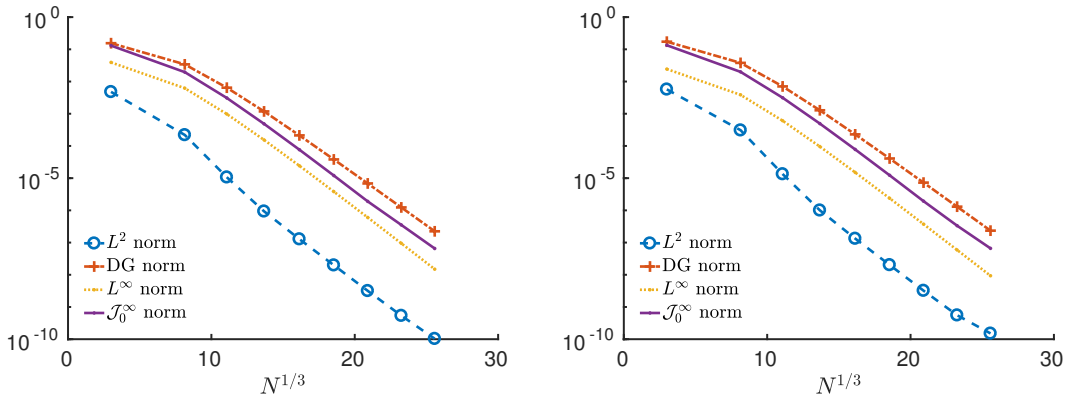


Figure 4.4 – Different convergence rates as a function of $N^{1/3}$, where N is the number of degrees of freedom, for problem (4.66), left, and for problem (4.67), right. Semilogarithmic scale.

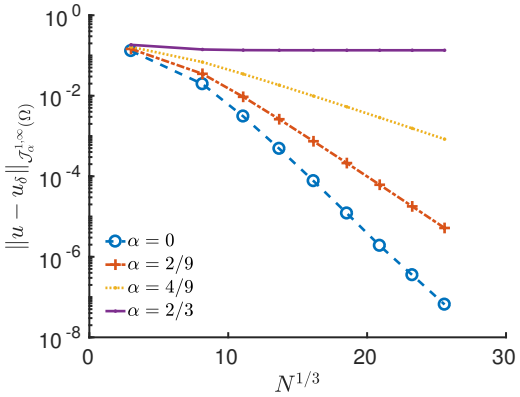


Figure 4.5 – Convergence rates in the $\mathcal{J}_\alpha^{1,\infty}(\Omega)$ norm as a function of $N^{1/3}$ for problem (4.66) and for varying α. Logarithmic scale on the y-axis, linear scale on the x-axis.

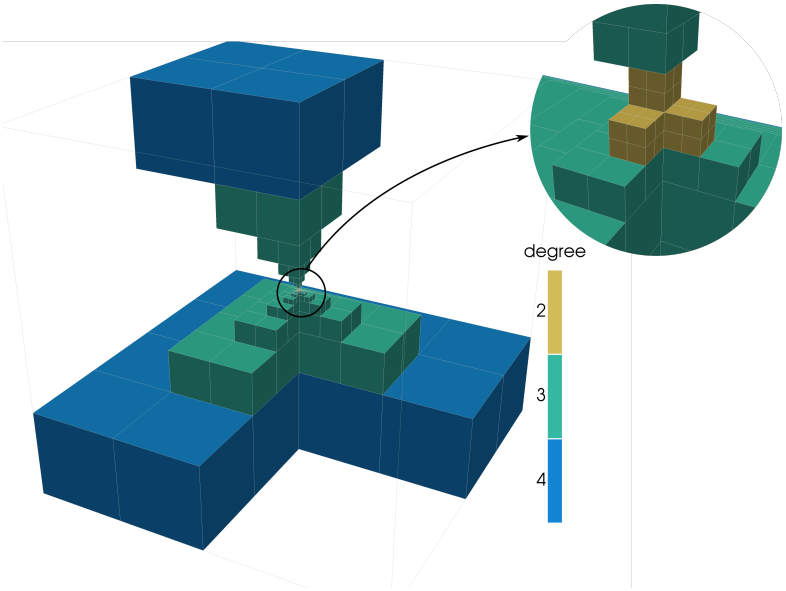


Figure 4.6 – Some of the elements of the mesh for $d = 3$.

b_{L^2}	b_{DG}	b_{L^∞}	$b_{\mathcal{J}_0^\infty}$
0.58	0.508	0.651	0.538

Table 4.2 – Estimated constants for problem (4.68).

estimate of the constants b_X defined as in (4.65).

Three dimensions

In three dimensions, we consider the problem

$$\begin{cases} (-\Delta + V)u = f \text{ in } \Omega \\ u_{|\partial\Omega} = g \end{cases} \tag{4.68}$$

set in a cube $\Omega = (-1/2, 1/2)^3$. We choose $V(x) = |x|^{-3/2}$ and f and g such that the exact solution is given by $u_{\mathrm{ex}} = 1 + |x|^{1/2}$. The approximation is done with $\sigma = 1/2$, $p_j = \lceil \frac{1}{6} j \rceil + 2$; the mesh is isotropically refined around the origin, as shown in Figure 4.6.

The convergence of the errors is shown in Figure 4.7; an estimate of the constants b_X defined in (4.65) is given in Table 4.2.

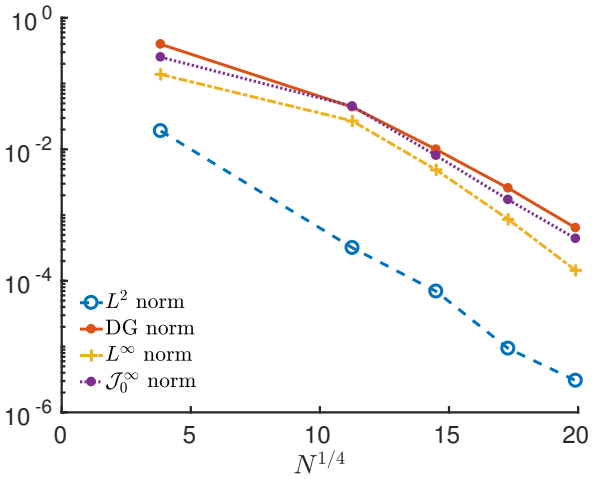


Figure 4.7 – Errors over $N^{1/4}$ for problem (4.68), with N degrees of freedom. Logarithmic scale on the y-axis, linear scale on the x-axis.

Chapter 5

Analysis of the hp discontinuous Galerkin method for elliptic linear eigenvalue problems

In this section, we consider the approximation of a linear elliptic eigenvalue problem obtained through a discontinuous Galerkin hp method. The contents of the section are largely based on [ABP06], where the convergence of the discontinuous Galerkin method is proven for linear elliptic eigenvalue problems. The result obtained in that paper is an extension to discontinuous Galerkin methods of the theory developed almost three decades earlier, see [DNR78a; DNR78b]. A thorough presentation of the approximation of eigenvalue problems is also given in [CL91, Chapter II].

Our problem fits into the assumptions made in [ABP06], thus the results shown there can be applied almost directly. The only minor differences will be due to the presence of a potential and to the specificity of approximation in isotropically refined hp finite element spaces.

We start by defining our problem and by giving some context from the functional point of view, in Section 5.1. We also introduce, for the sake of self containedness of the chapter, the interior penalty discontinuous Galerkin methods that will be taken into consideration. We will be dealing with a symmetric operator and a coercive linear form, thus the spectrum is composed of real isolated eigenvalues of ascent one. The analysis can still be partially extended to non-symmetric problems, but it has to be taken into account that the operators are not self-adjoint. We conclude the section by introduction the "solution operators" T, for the continuous problem, and T_δ, for the discrete approximation. T and T_δ are continuous and invertible operators, with the same eigenspaces as the ones of the original problems and with reciprocal eigenvalues. The analysis will center around those operators, and the final results can easily be applied back to the original problems. Finally, we will need a way to measure a "distance" between eigenspaces: this is the role of $\delta(\cdot,\cdot)$ and $\hat{\delta}(\cdot,\cdot)$ defined in (5.8).

The interest of the analysis of the approximation of an eigenvalue problem lies not only in the convergence of the numerical eigenpairs to the exact ones, but also in the non pollution and completeness of eigenfunctions and eigenvalues. Basically, a good approximation of an eigenproblem should not introduce any spurious numerical eigenvalue or eigenvector (non-pollution) and should approximate all eigenpairs (completeness). In Theorem 4 we show that the spectrum is not polluted, while in Theorem 5 the completeness of the approximation is shown (more precisely, Theorem 5 gives both completeness and convergence for finite dimensional eigenspaces, while simple completeness is a consequence of Lemma 32). Note that, in practice, some techniques may still introduce spurious eigenvalues in the approximation: consider for example the "strong" imposition of boundary conditions in a numerical code, where the matrix resulting from the approximation of the operator is modified in order to set the degrees of freedom at the boundary, see, e.g., the documentation of [Arn+17]. This is out of the scope of the present analysis; furthermore, the spurious eigenvalues can often be easily identified and filtered out.

We conclude with Section 5.3, where the focus is on the rate of convergence of the numerical eigenpairs. We consider finite dimensional exact eigenspaces and we introduce a projector from the exact to the numerical eigenspace, thus obtaining an algebraic problem, at least in the relationship between the eigenvalues and the (projected) operators $\widehat{T}$ and $\widehat{T}_\delta$ (the latter can be seen as a tensor in the finite dimensional eigenspace). We obtain the expected quasi optimal estimates on the difference between exact and numerical eigenfunctions. The eigenvalue error, additionally, can be shown to converge with a higher rate of convergence — quadratically with respect to the eigenfunctions — if the method is adjoint consistent (symmetric, in our case).

Let us then introduce the problem under consideration.

5.1 Statement of the problem and notation

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ be a bounded and regular domain. Consider the eigenvalue problem

$$\begin{aligned} &Lu = (-\Delta + V)u = \lambda u \text{ in } \Omega \\ &Bu = 0 \text{ on } \partial\Omega \end{aligned} \tag{5.1}$$

where B is a boundary operator inducing a well defined problem. We suppose that the potential $V \geq \eta > 0$ is singular in a set of points $\mathfrak{C} \subset \Omega$ and that for an $\varepsilon > 0$ and a $p > d/2$,

$$V \in \mathcal{K}^{\varpi,\infty}_{\varepsilon-2}(\Omega) \cap L^p(\Omega). \tag{5.2}$$

Recall that the homogeneous weighted norm is given by

$$\|u\|^p_{\mathcal{K}^{k,p}_\gamma(\Omega)} = \sum_{j=0}^{k} |u|^p_{\mathcal{K}^{j,p}_\gamma(\Omega)}, \qquad |u|^p_{\mathcal{K}^{j,p}_\gamma(\Omega)} = \sum_{|\alpha|=j} \|r^{j-\gamma}\partial^\alpha u\|^p_{L^p(\Omega)}$$

and the inhomogeneous one reads

$$\|u\|^p_{\mathcal{J}^{k,p}_\gamma(\Omega)} = \sum_{j=0}^{k} \sum_{|\alpha|=j} \|r^{\max(-\gamma+|\alpha|,\rho)} \partial^\alpha u\|^p_{L^p(\Omega)}, \text{ for any } \rho \in (-d/p, -\gamma + k],$$

for $1 \le p \le \infty$, with the usual modification when $p = \infty$. The spaces $\mathcal{K}^{k,p}_\gamma(\Omega)$ and $\mathcal{J}^{k,p}_\gamma(\Omega)$ can then by defined starting from their norms; the analytic class $\mathcal{K}^{\varpi,p}_\gamma(\Omega)$ is given by

$$\mathcal{K}^{\varpi,p}_\gamma(\Omega) = \left\{ v \in \bigcap_{k\in\mathbb{N}} \mathcal{K}^{k,p}_\gamma(\Omega) : \exists C, A > 0 \text{ such that } |v|_{K^{k,p}_\gamma(\Omega)} \le C A^k k!, \forall k \in \mathbb{N} \right\},$$

and $\mathcal{J}^{\varpi,p}_\gamma(\Omega)$ is defined similarly.

We recall two regularity results from Chapter 4.

Lemma 26. *If* (5.2) *holds, the operator L is an isomorphism between the spaces $\mathcal{J}^k_\gamma(\Omega) \to \mathcal{J}^{k-2}_{\gamma-2}(\Omega)$ for all $k \ge 2$ and*

$$\gamma \in I_d = \begin{cases} [1, 1+\varepsilon) & \text{if } d = 2 \\ (1/2, 3/2) \cup (3/2 + \varepsilon) & \text{if } d = 3, \end{cases}$$

Lemma 27. *If* (5.2) *holds, then $u \in \mathcal{J}^\varpi_\gamma(\Omega)$, for $\gamma \in I_d$.*

We write $(\cdot,\cdot)$ for the scalar product in $L^2(\Omega)$ and $\|\cdot\|$ for the $L^2(\Omega)$ norm. Finally, we denote by $a(\cdot,\cdot)$ be the bilinear form associated to L, i.e.

$$a(u, v) = (\nabla u, \nabla v) + (Vu, v).$$

We now introduce the discontinuous Galerkin interior penalty method.

5.1.1 Interior penalty method

Let $\mathcal{T}$ be a mesh isotropically and geometrically graded around the points in $\mathfrak{C}$. We assume that the mesh is shape- and contact-regular and we indicate by Ω_j, $j = 1, \dots, \ell$, the set of elements and edges at the same level of refinement.We introduce on this mesh the hp space with refinement ratio σ and linear polynomial slope $\mathfrak{s}$, i.e., for an element $K \in \mathcal{T}$ such that $K \in \Omega_j$,

$$h_K \simeq h_j = \sigma^j \text{ and } p_K \simeq p_j = p_0 + \mathfrak{s}(\ell - j),$$

where h_K is the diameter of the element K and p_K is the polynomial order whose role will be specified in (5.3). We suppose that for any $K \in \mathcal{T}$ there exists an affine transformation $\Phi : K \to \hat{K}$ to the d-dimensional cube $\hat{K}$ such that $\Phi(K) = \hat{K}$, and introduce the discrete space

$$X_\delta = \left\{ v_\delta \in L^2(\Omega) : (v_{|_K} \circ \Phi^{-1}) \in \mathbb{Q}_{p_K}(\hat{K}) \, \forall K \in \mathcal{T} \right\}, \tag{5.3}$$

where $\mathbb{Q}_p$ is the space of polynomials of maximal degree p in any variable. Let then $\mathcal{E}$ be the set of the edges (for $d = 2$) or faces ($d = 3$) of the elements in $\mathcal{T}$ and

$$\mathtt{h_e} = \min_{K\in\mathcal{T}:e\cap\partial K\neq\varnothing} h_K$$
$$\mathtt{p_e} = \max_{K\in\mathcal{T}:e\cap\partial K\neq\varnothing} p_K.$$

On an edge/face between two elements $K_\sharp$ and $K_\flat$, i.e., on $e \subset \partial K_\sharp \cap \partial K_\flat$, the average $\{\!\{\cdot\}\!\}$ and jump $[\![\cdot]\!]$ operators for a function $w \in X(\delta)$ are defined by

$$\{\!\{w\}\!\} = \frac{1}{2}\left(w_{|_{K_\sharp}} + w_{|_{K_\flat}}\right), \qquad [\![w]\!] = w_{|_{K_\sharp}}\mathbf{n}_\sharp + w_{|_{K_\flat}}\mathbf{n}_\flat,$$

where $\mathbf{n}_\sharp$ (resp. $\mathbf{n}_\flat$) is the outward normal to the element $K_\sharp$ (resp. $K_\flat$). In the following, for an $S \subset \Omega$, we denote by $(\cdot,\cdot)_S$ the $L^2(S)$ scalar product and by $\|\cdot\|_S$ the $L^2(S)$ norm.

Given $\vartheta \in \{-1, 0, 1\}$, we indicate by $a_\delta(\cdot,\cdot) : X_\delta \times X_\delta \to \mathbb{R}$ the interior penalty bilinear form, given by

$$\begin{aligned} a_\delta(u_\delta, v_\delta) = (\nabla u_\delta, \nabla v_\delta)_{\mathcal{T}} - (\{\!\{\nabla u_\delta\}\!\}, [\![v_\delta]\!])_{\mathcal{E}_I} - \vartheta(\{\!\{\nabla v_\delta\}\!\}, [\![u_\delta]\!])_{\mathcal{E}} \\ + \sum_{e\in\mathcal{E}} \alpha_e \frac{\mathtt{p_e}^2}{\mathtt{h_e}} ([\![u_\delta]\!], [\![v_\delta]\!])_e + \int_\Omega V u_\delta v_\delta. \end{aligned} \tag{5.4}$$

Here, $\mathcal{E}_I$ is the set of internal edges such that for all $e \in \mathcal{E}_I$, $e \cap \partial\Omega = \varnothing$, and we have written

$$(\cdot,\cdot)_{\mathcal{T}} = \sum_{K\in\mathcal{T}} (\cdot,\cdot)_K \qquad (\cdot,\cdot)_{\mathcal{E}} = \sum_{e\in\mathcal{E}} (\cdot,\cdot)_e.$$

The discrete eigenvalue problem then reads: find $(\lambda_\delta, u_\delta) \in \mathbb{C} \times X_\delta$

$$a_\delta(u_\delta, v_\delta) = \lambda_\delta(u_\delta, v_\delta) \text{ for all } v_\delta \in X_\delta. \tag{5.5}$$

Choosing $\vartheta = 1$ in (5.4) gives the symmetric interior penalty (SIP) method, while $\vartheta = -1$ gives the non-symmetric interior penalty (NIP) method, and $\vartheta = 0$ gives the incomplete interior penalty method (IIP). We remark that the choice $\vartheta = 1$ is the only one that preserves the symmetry of the bilinear form; the SIP method is *adjoint consistent*.

We write $X = H^1(\Omega)$, $X(\delta) = X + X_\delta$ and introduce the mesh dependent norms

$$\|v\|_{\mathrm{DG}}^2 = \sum_{K\in\mathcal{T}} \|v\|_{H^1(K)}^2 + \sum_{e\in\mathcal{E}} \mathtt{p_e}^2\mathtt{h_e}^{-1}\|[\![v]\!]\|_{L^2(e)}^2$$

and

$$|||v|||_{\mathrm{DG}}^2 = \|v\|_{\mathrm{DG}}^2 + \sum_{e\in\mathcal{E}} \mathtt{h_e}\mathtt{p_e}^{-2}\|\nabla v\|_{L^2(e)}^2.$$

Note that $\|\cdot\|_{\mathrm{DG}}$ is defined on $X(\delta)$, while $|||\cdot|||_{\mathrm{DG}}$ is defined only on the broken space

$$X(\delta) \cap H^{d/2}(\mathcal{T}) = \left\{ v \in X(\delta) : v \in H^{d/2}(K) \text{ for all } K \in \mathcal{T} \right\},$$

due to the presence of the boundary gradient term. We introduce the continuous solution operator

$$T : L^2(\Omega) \to X \tag{5.6}$$

such that

$$a(Tu, v) = (u, v), \text{for all } v \in X$$

and its discrete counterpart, given by

$$T_\delta : L^2(\Omega) \to X_\delta \tag{5.7}$$

such that

$$a_\delta(T_\delta u, v) = (u, v), \text{for all } v_\delta \in X_\delta.$$

The analysis of the relation between the spectra associated to the operator L in (5.1) and to the discrete bilinear form a_δ can be transformed into the analysis of the spectra of T and T_δ. In the following, the spectrum of T will be denoted by $\sigma(T)$ and its resolvent set by $\rho(T)$. Similarly, $\sigma(T_\delta)$ and $\rho(T_\delta)$ will be respectively the spectrum and resolvent set of T_δ. Let then

$$R_z(T) = (z - T)^{-1}$$

be the resolvent operator associated with T, and

$$R_z(T_\delta) = (z - T_\delta)^{-1}$$

be the resolvent operator associated with T_δ, both defined for $z \in \mathbb{C}$. Finally, we introduce a measure of the gap between subspaces of $X(\delta)$: let Y and Z be close subspaces of $X(\delta)$; then for an $x \in X$ we define

$$\begin{aligned} &\delta(x, Y) = \inf_{y \in Y} \|x - y\|_{\mathrm{DG}}, \qquad \delta(Y, Z) = \sup_{y \in Y : \|y\|_{\mathrm{DG}} = 1} \delta(y, Z) \\ &\hat{\delta}(Y, Z) = \max(\delta(Y, Z), \delta(Z, Y)). \end{aligned} \tag{5.8}$$

5.2 Non pollution and completeness of the discrete spectrum and eigenspaces

5.2.1 Non pollution of the spectrum

In this section we detail the technique used in [ABP06] to prove the non-pollution of the discrete spectrum. Note that, thus far, $R_z(T_\delta)$ has only been defined formally. We will now show its existence and continuity, together with the existence and continuity of its inverse. This will imply the non pollution of the discrete spectrum and guarantee that,

for a sufficient number of degrees of freedom, the discrete spectrum lies in the vicinity of the continuous one.

We start by introducing a lemma, whose proof we postpone to the end of the section.

Lemma 28. *Let $z \in \rho(T)$ such that $z \neq 0$ and $u \in X(\delta)$. Then,*

$$\|(z - T)u\|_{\mathrm{DG}} \geq C\|u\|_{\mathrm{DG}}$$

where C depends on L, on Ω, and on $|z|$.

By the triangle inequality, then,

$$\|(z - T_\delta)u\|_{\mathrm{DG}} \geq \|(z - T)u\|_{\mathrm{DG}} - \|(T - T_\delta)u\|_{\mathrm{DG}}. \tag{5.9}$$

Now, the second term at the right hand side is the classical error of the method; by the coercivity and continuity of the discrete bilinear form, Lemma 26 and the approximation properties of the hp space, we have that

$$\|(T - T_\delta)u\|_{\mathrm{DG}} \to 0 \text{ as } N \to \infty$$

where N is the dimension of X_δ. Using Lemma 28 and the above estimate in (5.9), we obtain that, for a sufficient number of degrees of freedom,

$$\|(z - T_\delta)u\|_{\mathrm{DG}} \geq C\|u\|_{\mathrm{DG}} \tag{5.10}$$

for $0 \neq z \in \rho(T)$. For a fixed z and for a sufficient number of degrees of freedom (depending on z), thus, $z - T_\delta$ is invertible and $R_z(T_\delta)$ is well defined. Furthermore, Lemma 28 implies that $R_z(T)$ is well defined and bounded as an operator on the spaces $X(\delta) \to X(\delta)$. We have therefore shown that $R_z(T_\delta)$ is bounded as a linear operator from $X(\delta)$ to $X(\delta)$, and that the spectrum is not polluted; in the following we summarize this results. Denoting by $\|\cdot\|_{\mathcal{L}(V,W)}$ the classical operator norm

$$\|F\|_{\mathcal{L}(V,W)} = \sup_{v \in V : \|v\|_V = 1} \|Fv\|_W,$$

from (5.10) we conclude that

Lemma 29. *Let $A \subset \rho(T)$ be a closed set. Then, for all $z \in A$, there exists a constant C such that*

$$\|R_z(T_\delta)\|_{\mathcal{L}(X(\delta),X(\delta))} \leq C.$$

The non-pollution of the spectrum follows directly, taking the complementary of the set A above.

Theorem 4. *Let $B \supset \sigma(T)$ be an open set. Then, for a sufficient number of degrees of freedom,*

$$\sigma(T_\delta) \subset B.$$

We conclude the section with the proof of Lemma 28.

Proof of Lemma 28. Consider $u \in X(\delta)$ and $0 \neq z \in \rho(T)$. Then, by the triangle inequality,

$$|z|\|u\|_{\mathrm{DG}} \leq \|zTu\|_{\mathrm{DG}} + \|(z-T)u\|_{\mathrm{DG}}. \tag{5.11}$$

Let now $v = zTu$. Then, by the definition of T, $zu = Lv$ and

$$Lv - \frac{1}{z}v = (z-T)u$$

with the associated boundary conditions. Since $z \in \rho(T)$, the operator $L-1/z$ is invertible, and

$$\|zTu\|_{\mathrm{DG}} = \|v\|_X \leq C\|(z-T)u\|_{L^2(\Omega)}.$$

The constant C clearly depends on z, on the operator L, and on Ω. Inserting the above inequality into (5.11) one obtains the thesis. □

5.2.2 Eigenspaces and completeness of the spectrum

Consider a smooth curve $\Gamma \subset \rho(T)$. We introduce the spectral projectors

$$E = \frac{1}{2\pi i}\int_\Gamma R_z(T)dz \quad \text{and} \quad E_\delta = \frac{1}{2\pi i}\int_\Gamma R_z(T_\delta)dz \tag{5.12}$$

Clearly, both projectors depend on Γ, we omit that in our notation as is customary: suppose that Γ is fixed and that it encloses a single eigenvalue of T. The discrete projector E_δ is, once again, well defined provided that the space X_δ contains a sufficient number of degrees of freedom. Suppose that Γ contains an eigenvalue of T; then, E is the projector on the eigenspace associated to the eigenvalue. The same holds for the discrete version.

We now wish to prove the convergence of the discrete projector to the continuous one, in the operator norm. We start by noting that

$$(z-T)^{-1} - (z-T_\delta)^{-1} = (z-T_\delta)^{-1}(T-T_\delta)(z-T)^{-1},$$

therefore,

$$\begin{aligned}\|R_z(T) - R_z(T_\delta)\|_{\mathcal{L}(L^2(\Omega),X(\delta))} &= \|R_z(T_\delta)(T-T_\delta)R_z(T)\|_{\mathcal{L}(L^2(\Omega),X(\delta))} \\ &\leq \|R_z(T_\delta)\|_{\mathcal{L}(X(\delta),X(\delta))} \\ &\quad \times \|(T-T_\delta)\|_{\mathcal{L}(L^2(\Omega),X(\delta))}\|R_z(T)\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega)}.\end{aligned}$$

Due to the boundedness of the continuous – see [ABP06] – and discrete – see Lemma 29 – resolvent operators, we conclude that

$$\|E - E_\delta\|_{\mathcal{L}(L^2(\Omega),X(\delta))} \leq C\|(T-T_\delta)\|_{\mathcal{L}(L^2(\Omega),X(\delta))}. \tag{5.13}$$

Lemma 30. *Given the definitions of E and E_δ in* (5.12)*, if X_δ has a sufficient number of degrees*

of freedom, there holds

$$\|E - E_\delta\|_{\mathcal{L}(L^2(\Omega),X(\delta))} \to 0.$$

Consider now the definitions given in (5.8). The convergence of the projectors allows for the proof of the convergence to zero of some "distances" between eigenspaces. The first almost direct result is in the following lemma.

Lemma 31. *Let $\delta(\cdot,\cdot)$ be defined as in* (5.8)*. Then,*

$$\delta(E_\delta(X_\delta), E(X)) \to 0$$

Proof. For any $x_\delta \in E_\delta(X_\delta)$, $E_\delta(x_\delta) = x_\delta$. We remark that, due to the regularity result given in Lemma 13, $E(L^2(\Omega)) = E(X)$. Thus, for any $x_\delta \in E_\delta(X_\delta)$ such that $\|x_\delta\|_{\mathrm{DG}} = 1$,

$$\begin{aligned}\inf_{x\in E(X)} \|x_\delta - x\|_{\mathrm{DG}} &= \inf_{x\in E(L^2(\Omega))} \|x_\delta - x\|_{\mathrm{DG}} \\ &= \inf_{y\in L^2(\Omega)} \|E_\delta x_\delta - Ey\|_{\mathrm{DG}} \\ &\leq \|E_\delta - E\|_{\mathcal{L}(L^2(\Omega),X(\delta))}.\end{aligned}$$

Taking the supremum over all $x_\delta \in E_\delta(X_\delta)$ one obtains the thesis. □

This is a proof of the non pollution of the eigenspaces: we have indeed shown that all numerical eigenfunction converges to an exact one. We continue by showing the completeness of the eigenspaces. This involves proving that all exact eigenfunction is approximated by a numerical one.

Lemma 32. *For any $x \in E(X)$,*

$$\delta(x, E_\delta(X_\delta)) \to 0$$

Proof. Let $x \in E(X)$ and $x_\delta \in X_\delta$. Then,

$$\|E_\delta x_\delta - x\|_{\mathrm{DG}} \leq \|E\|_{\mathcal{L}(X(\delta),X(\delta))} \|\!|x_\delta - x|\!\|_{\mathrm{DG}} + \|E - E_\delta\|_{\mathcal{L}(X(\delta),X(\delta))} \|x_\delta\|_{\mathrm{DG}}.$$

Taking x_δ as the projection of x in X_δ and thanks to the convergence of E_δ towards E, we obtain the thesis. □

We now restrict our focus to finite dimensional eigenspaces. Let then $n = \dim(E(X))$ and $n_\delta = \dim(E_\delta(X_\delta))$: if $n = \infty$, then $n_\delta \to \infty$; we consider the case where n is finite. If n is finite, the above lemma implies that

$$\delta(E(X), E_\delta(X_\delta)) \to 0.$$

We now analyse the action of the resolvent of the continuous operator on the eigenspace.

Lemma 33. *Let $0 \neq z \in \rho(T)$ and $x \in E(X)$. Then, for $s \in \mathbb{N}$ and $\gamma \in I_d$,*

$$\|(z - T)^{-1}x\|_{\mathcal{J}^s_\gamma(\Omega)} \lesssim \|x\|_{\mathcal{J}^s_\gamma(\Omega)}$$

Proof. Let $v = R_z(T)x$. Then,

$$Lv - \frac{1}{z}v = Lx.$$

Since z is not in the spectrum of T, $1/z$ does not belong to the spectrum of L and the operator $L - 1/z$ is invertible. Furthermore, $x \in E(X)$ implies $x \in \mathcal{J}^{\varpi}_{\gamma}(\Omega)$, for $\gamma \in I_d$, see Corollary 17. Hence, for all $\gamma \in I_d$ and any $s \in \mathbb{N}$, there exists a constant C depending on $|z|$ such that

$$\|v\|_{\mathcal{J}^s_\gamma(\Omega)} \leq C\|x\|_{\mathcal{J}^s_\gamma(\Omega)}.$$

□

Consider then an $x \in E(X)$: we have

$$\inf_{x_\delta \in X_\delta} \|E_\delta x_\delta - x\|_{\mathrm{DG}} \leq \|E_\delta\|_{\mathcal{L}(X(\delta),X(\delta))} \inf_{x_\delta \in X_\delta} |||x_\delta - x|||_{\mathrm{DG}} + \|(E - E_\delta)\, x\|_{\mathrm{DG}} \tag{5.14}$$

Due to the approximation properties of X_δ there exist $C, b > 0$ such that

$$\inf_{x_\delta \in X_\delta} |||x - x_\delta|||_{\mathrm{DG}} \leq Ce^{-bN^{1/(d+1)}}, \tag{5.15}$$

with $N = \dim(X_\delta)$. In addition,

$$\begin{aligned}\sup_{x\in E(X)} |||(R_z(T) - R_z(T_\delta))\, x|||_{\mathrm{DG}} &= \sup_{x\in E(X)} |||(R_z(T_\delta)(T - T_\delta)R_z(T))\, x|||_{\mathrm{DG}} \\ &\leq C\|R_z(T_\delta)\|_{\mathcal{L}(X(\delta),X(\delta))} \\ &\quad \times \sup_{x\in Y} |||(T - T_\delta)x|||_{\mathrm{DG}}\, \|R_z(T)\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega))},\end{aligned}$$

where Y is the space of all functions $v \in \mathcal{J}^{\varpi}_{\gamma}(\Omega)$ such that for all $s \in \mathbb{N}$, $\|v\|_{\mathcal{J}^s_\gamma(\Omega)} \leq C \sup_{x\in E(X)} \|x\|_{\mathcal{J}^s_\gamma(\Omega)}$. We obtain

$$\sup_{x\in E(X)} |||(E - E_\delta)x|||_{\mathrm{DG}} \leq C \sup_{x\in Y} |||(T - T_\delta)x|||_{\mathrm{DG}} \tag{5.16}$$

Thanks to Lemma 33, the right hand side of the above equation is the error of the numerical method for a problem with source term belonging to $\mathcal{J}^{\varpi}_{\gamma}(\Omega)$: by Lemma 13 and the approximation properties of the hp space, there exist $C, b > 0$ such that

$$\sup_{x\in E(X)} |||(E - E_\delta)x|||_{\mathrm{DG}} \leq Ce^{-bN^{1/(d+1)}}. \tag{5.17}$$

Combining (5.14), (5.15) and (5.17), we have then the explicit rate

$$\delta(E(X), E_\delta(X_\delta)) \leq Ce^{-bN^{1/(d+1)}}.$$

We summarize this in the following statement.

Theorem 5. *If* $\dim(E(X)) < \infty$ *and for a sufficient number of degrees of freedom, there exist* $C, b > 0$ *such that*

$$\delta(E(X), E_\delta(X_\delta)) \leq Ce^{-bN^{1/(d+1)}}.$$

5.3 Convergence of the eigenfunctions and eigenvalues

In this section we consider the convergence of the numerical eigenfunctions and eigenvalues obtained through the hp approximation. As far as the eigenspaces are concerned, Lemma 31 proves that they are not polluted and Lemma 32 proves that they are complete. As a direct consequence of Theorem 5, furthermore, we have that for any $u \in E(X)$ with $\dim(E(X)) < \infty$ there exists $u_\delta \in E_\delta(X_\delta)$ such that

$$\|u - u_\delta\|_{\mathrm{DG}} \leq Ce^{-bN^{1/(d+1)}}.$$

We now consider the convergence of the eigenvalues; we will do so in the case of a symmetric numerical scheme.

5.3.1 Convergence of the eigenvalues

We are mainly interested in the analysis of the convergence of the eigenvalues for the symmetric interior penalty method, obtained by choosing $\vartheta = 1$ in (5.4). The first part of the section will, nonetheless, hold for non-symmetric methods, but we will signal when the hypothesis of symmetry of the numerical method will become necessary. The final result obtained for the SIP method will be stronger that what can be obtained in the case of non-symmetric methods, since they lack the property of adjoint consistency.

We start by considering the operator $\Lambda_\delta = E_{\delta|_{E(X)}} : E(X) \to E_\delta(X_\delta)$. For a sufficient number of degrees of freedom, the operator is invertible. For any $u \in E(X)$,

$$\|u\|_{\mathrm{DG}} \leq \|(E - E_\delta)u\|_{\mathrm{DG}} + \|E_\delta u\|_{\mathrm{DG}}$$

and the convergence of $E - E_\delta$ in the operator norm implies that for a sufficient number of degrees of freedom, ${\Lambda_\delta}^{-1}$ is bounded. Let us then introduce the operators

$$\widehat{T} = T_{|_{E(X)}} \quad \text{and} \quad \widehat{T}_\delta = {\Lambda_\delta}^{-1} T_\delta \Lambda_\delta, \tag{5.18}$$

both defined on the spaces $E(X) \to E(X)$. We consider the case where Γ contains a single eigenvalue μ of T, with multiplicity n and where $\mu_{\delta i}$, $i = 1, \ldots, n$ are the eigenvalues of T_δ. There exists, then, an $x \in E(X)$ such that

$$\widehat{T}_\delta x = \mu_{\delta j} x.$$

Let now T' and ${T_\delta}'$ be the adjoint operators to T and T_δ, and let E' and ${E_\delta}'$ be the associated spectral projectors. Furthermore, consider a $y \in E'(X)$ such that $(x, y) = 1$:

since for all $x \in E(X)$, $(T-\mu)x = 0$ (since all eigenvalues have ascent one), we have

$$\begin{aligned}\mu - \mu_{\delta j} &= \langle (\mu - \widehat{T}_\delta)x, y\rangle \\ &= \langle (T - \widehat{T}_\delta)x, y\rangle \\ &= \langle (T - {\Lambda_\delta}^{-1} T_\delta E_\delta)x, y\rangle\end{aligned}$$

Note now that ${\Lambda_\delta}^{-1} E_{\delta|_{E(X)}} = I$ and that T_δ and E_δ commute on $E(X)$, thus

$$\begin{aligned}\mu - \mu_{\delta j} &= \langle ({\Lambda_\delta}^{-1} E_\delta)(T - T_\delta)x, y\rangle \\ &= \langle (T - T_\delta)x, y\rangle + \langle ({\Lambda_\delta}^{-1} E_\delta - I)(T - T_\delta)x, y\rangle.\end{aligned}$$

We remark that $\ker(({\Lambda_\delta}^{-1} E_\delta - I)_{|_{E(X)}}) = \ker(E_\delta)^\perp$, hence

$${\Lambda_\delta}^{-1} E_\delta - I : E(X) \to \mathrm{im}({E_\delta}')^\perp.$$

Using also the fact that $E'y = y$, the second term at the right hand side above can be written as

$$\langle ({\Lambda_\delta}^{-1} E_\delta - I)(T - T_\delta)x, y\rangle = \langle ({\Lambda_\delta}^{-1} E_\delta - I)(T - T_\delta)x, (E' - {E_\delta}')y\rangle.$$

As already shown ${\Lambda_\delta}^{-1}$ is bounded for a sufficient number of degrees of freedom, and so is E_δ; thus,

$$\left|\langle ({\Lambda_\delta}^{-1} E_\delta - I)(T - T_\delta)x, y\rangle\right| \leq C \|T - T_\delta\|_{\mathcal{L}(Y, X(\delta))} \|T' - {T_\delta}'\|_{\mathcal{L}(\widetilde{Y}, X(\delta))} \|x\| \|y\|$$

where $\widetilde{Y} = \{v \in \mathcal{J}_\gamma^\varpi(\Omega) : \|v\|_{\mathcal{J}_\gamma^s(\Omega)} \lesssim \sup_{x \in E'(X)} \|x\|_{\mathcal{J}_\gamma^s(\Omega)}\}$ and we have used (5.13) for the adjoint spectral projectors. Let us now choose $\|x\| = \|y\| = 1$, and introduce two bases $\{\varphi_i\}_i$ and $\{\varphi'_j\}_j$ for $E(X)$ and $E'(X)$ respectively. Since the spaces are finite dimensional, i.e., $n = \dim(E(X)) = \dim(E'(X)) < \infty$, there exists a constant $C > 0$ such that

$$\begin{aligned}\langle (T - T_\delta)\, x, y\rangle &\leq \sup_{\|x\|=\|y\|=1} |\langle (T - T_\delta)\, x, y\rangle| \\ &\leq C \sum_{i,j=1}^{n} \left|\langle (T - T_\delta)\, \varphi_i, \varphi'_j\rangle\right|,\end{aligned}$$

where C depends on n. We conclude that

$$|\mu - \mu_{\delta j}| \leq C \left(\sum_{i,j=1}^{n} \langle (T - T_\delta)\varphi_i, \varphi'_j\rangle + \|T - T_\delta\|_{\mathcal{L}(Y, X(\delta))} \|T' - {T_\delta}'\|_{\mathcal{L}(\widetilde{Y}, X(\delta))} \right), \quad (5.19)$$

Remark 8. *The above estimate* (5.19) *holds since we have considered a case where all eigenvalues*

have ascent one. If this were not the case, one would find that

$$|\mu - \frac{1}{n}\sum_{j=1}^{n}\mu_{\delta j}| \leq C\left(\sum_{i,j=1}^{n}\langle (T-T_\delta)\varphi_i, \varphi_j'\rangle + \|T-T_\delta\|_{\mathcal{L}(Y,X(\delta))}\|T'-{T_\delta}'\|_{\mathcal{L}(\widetilde{Y},X(\delta))}\right),$$

and

$$|\mu - \frac{1}{n}\sum_{j=1}^{n}\mu_{\delta j}| \leq C\left(\sum_{i,j=1}^{n}\langle (T-T_\delta)\varphi_i, \varphi_j'\rangle + \|T-T_\delta\|_{\mathcal{L}(Y,X(\delta))}\|T'-{T_\delta}'\|_{\mathcal{L}(\widetilde{Y},X(\delta))}\right)^{1/\alpha},$$

α being the ascent of the eigenvalue μ, see [DNR78b; CL91].

5.3.2 Convergence of the eigenvalues for the SIP method

We now restrict ourselves to the symmetric interior penalty method and consider the fact that our operator is self-adjoint: then, $T' = T$, ${T_\delta}' = T_\delta$, and (5.19) reads

$$|\mu - \mu_{\delta j}| \leq C\left(\sum_{i,j=1}^{n}\langle (T-T_\delta)\varphi_i, \varphi_j\rangle + \|T-T_\delta\|^2_{\mathcal{L}(Y,X(\delta))}\right). \tag{5.20}$$

At this stage, the goal is in bounding the first term at the right hand side of the inequality by something quadratic in nature, to show that it converges as fast as the second term. This is where the adjoint consistency of the SIP method is crucial. Let $y \in E(X)$ with $\|y\| = 1$ and let $\psi \in X$ be the solution to the adjoint problem

$$\begin{aligned} L\psi &= y \text{ in } \Omega \\ B\psi &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{5.21}$$

Now, recall that thanks to Lemma 27 we have that, for all $x \in E(X)$, $x \in \mathcal{J}_\gamma^\varpi(\Omega)$, for any $\gamma \in I_d$ and $k \in \mathbb{N}$. We can then fix an $A > 0$ such that for all $x \in E(X)$, with $\|x\| = 1$ and for all $\gamma \in I_d$,

$$\|x\|_{\mathcal{J}_\gamma^k(\Omega)} \leq CA^k k!.$$

Without loss of generality, we take $A \geq C$. Lemma 26 applied to problem (5.21) implies therefore that, for all $k \in \mathbb{N}$, $\gamma \in I_d$,

$$\|\psi\|_{\mathcal{J}_\gamma^k(\Omega)} \leq CA^k k!.$$

This implies that

$$\inf_{v_\delta \in X_\delta} |||\psi - v_\delta|||_{\mathrm{DG}} \leq C \max_{\substack{u \in E(X) \\ \|u\|=1}} \inf_{v_\delta \in X_\delta} |||u - v_\delta|||_{\mathrm{DG}}. \tag{5.22}$$

Consider then $x \in E(X)$, with $\|x\| = 1$:

$$\begin{aligned}\langle (T - T_\delta)x, y\rangle &= \langle (T - T_\delta)x, L\psi\rangle \\ &= a_\delta((T - T_\delta)x, \psi) \\ &= a_\delta((T - T_\delta)x, \psi - v_\delta)\end{aligned}$$

Finally, by the continuity of the bilinear form, the quasi optimality of the discontinuous Galerkin method, and using (5.22), we conclude that

$$\begin{aligned}|\langle (T - T_\delta)x, y\rangle| &\leq C \,|||(T - T_\delta)x|||_{\mathrm{DG}} \,|||\psi - v_\delta|||_{\mathrm{DG}} \\ &\leq C \sup_{\substack{u\in E(X)\\ \|u\|=1}} \inf_{v_\delta \in X_\delta} |||u - v_\delta|||^2_{\mathrm{DG}} \,.\end{aligned}$$

Since clearly

$$\|T - T_\delta\|^2_{\mathcal{L}(Y,X(\delta))} \leq C \max_{\substack{u\in E(X)\\ \|u\|=1}} \inf_{v_\delta \in X_\delta} |||u - v_\delta|||^2_{\mathrm{DG}} \,,$$

from (5.20) we conclude that

$$\max_{j=1,\ldots,n} |\mu - \mu_{\delta j}| \leq C \max_{\substack{u\in E(X)\\ \|u\|=1}} \inf_{v_\delta \in X_\delta} |||u - v_\delta|||^2_{\mathrm{DG}} \,.$$

Since for every eigenvalue μ of T, $1/\mu$ is an eigenvalue of (5.1), we have proven the following theorem.

Theorem 6. *Let λ be an eigenvalue of problem* (5.1) *with associated eigenspace $U = \mathrm{span}(u_1, \ldots, u_n)$, with $\|u_i\| = 1$ for $i = 1, \ldots, n$ and $n < \infty$. Then, there exist n eigenvalue-eigenfunction pairs $\{(\lambda_{\delta j}, u_{\delta j})\}_j$ of the finite dimensional problem* (5.5) *such that for all $j = 1, \ldots, n$*

$$\begin{aligned}\min_{u\in U} \|u - u_{\delta j}\|_{\mathrm{DG}} &\lesssim \sup_{u\in U} \inf_{v_\delta \in X_\delta} |||u - v_\delta|||_{\mathrm{DG}} \\ |\lambda - \lambda_{\delta j}| &\lesssim \sup_{u\in U} \inf_{v_\delta \in X_\delta} |||u - v_\delta|||_{\mathrm{DG}} \,.\end{aligned}$$

Furthermore, if the numerical solutions are obtained with the SIP method,

$$|\lambda - \lambda_{\delta j}| \lesssim \sup_{u\in U} \inf_{v_\delta \in X_\delta} |||u - v_\delta|||^2_{\mathrm{DG}} \,.$$

Finally, there are no spurious numerical eigenvalues or eigenvectors.

Given the approximation properties of the hp method and considering that all eigenfunctions of (5.1) belong to the space $\mathcal{J}^{\varpi}_{\gamma}(\Omega)$ for a $\gamma > d/2$, we can also provide the following corollary.

Corollary 34. *Let λ, u, U, $\lambda_{\delta j}$, and $u_{\delta j}$ be defined as in Theorem 6 and let $N = \dim(X_\delta)$. Then,*

there exist $C, b > 0$ *such that , for all* $j = 1, \ldots, n$

$$\min_{u \in U} \|u - u_{\delta j}\|_{\mathrm{DG}} \leq C e^{-bN^{1/(d+1)}}$$
$$|\lambda - \lambda_{\delta j}| \leq C e^{-bN^{1/(d+1)}}$$

Furthermore, if the numerical solutions are obtained with the SIP method,

$$|\lambda - \lambda_{\delta j}| \leq C e^{-2bN^{1/(d+1)}}.$$

Chapter 6

Numerical results for the linear eigenvalue problem

In this section, we perform some numerical experiments on the linear eigenvalue problem of finding $(\lambda, u) \in \mathbb{R} \times H^1(\Omega)$ such that $\|u\|_{L^2(\Omega)} = 1$ and

$$\begin{aligned} (-\Delta + V)u &= \lambda u \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{6.1}$$

The domain Ω is the d-dimensional cube with unitary edge $(-1/2, 1/2)^d$, and V is a potential with a singularity at the origin that will be specified in the different cases. Since no exact solution is available, every numerical solution is compared with the solution obtained at a higher degree of refinement than those presented.

In all cases, the mesh is isotropically and geometrically refined around the origin, with a geometric refinement ratio $\sigma = 1/2$. All elements are axiparallel d-dimensional cubes. This means that, introducing the refinement layers Ω_j, $j = 1, \dots, \ell$, such that for all $K \in \Omega_j$,

$$\inf_{x \in K} \|x\|_{\ell^\infty} = \sigma^{j+1} \quad j = 1, \dots, \ell - 1$$

we have

$$|K| = h_K^d = \sigma^{(j+1)d}.$$

Furthermore, the elements in Ω_ℓ have a vertex on the singularity. The polynomial slope $\mathfrak{s}$, defined as the parameter such that for all $v_\delta \in X_\delta$, on an element $K \in \Omega_j$

$$v_{\delta|_K} \in \mathbb{Q}_{p_j}(K),$$

with

$$p_j = p_0 + \lfloor \mathfrak{s}(\ell - j) \rfloor$$

is instead variable between experiments, and it is one of the main parameters whose role

in the approximation we investigate. The base polynomial degree is fixed at $p_0 = 1$.

All the simulations are obtained with C++ code based on the library `deal.II` [Arn+17]. Furthermore, we use PETSc [Bal+17] for the solution of algebraic linear systems, and SLEPc [HRV05] for the solution of the algebraic eigenvalue problem. The actual methods used will vary between the two and the three dimensional cases, and will be specified in the respective sections. The boundary conditions are imposed weakly, as is customary in the framework of discontinuous Galerkin methods, so no spurious eigenvalue is introduced, as shown in Chapter 5 holds.

The results we will shown in the following concern the estimation of the DG, $L^2(\Omega)$ and $L^\infty(\Omega)$ norms of the error, and of the difference between the computed and the "exact" eigenvalue. Furthermore, we will try to estimate the constants b_X such that

$$\|u - u_\delta\|_X \leq C_X \exp(-b_X N^{1/(d+1)}),$$

for $X = \mathrm{DG}, L^2(\Omega), L^\infty(\Omega)$, and

$$|\lambda - \lambda_\delta| \leq C_\lambda \exp(-b_\lambda N^{1/(d+1)}).$$

Here, $u_\delta \in X_\delta$ (resp. $\lambda_\delta \in \mathbb{R}$) is the numerical eigenfunction (resp. eigenvalue) computed with $\dim(X_\delta) = N$ and u (resp. λ) is the exact one.

We start by illustrating the results obtained in the framework of a two dimensional approximation.

6.1 Two dimensional case

We solve problem (6.1) with $d = 2$ on a mesh built as shown in Figure 6.1. An example of a numerically computed eigenfunction is shown in Figure 6.2a. We can see the combination of the effect of the laplacian with homogeneous Dirichlet boundary conditions and of the potential. The cusp introduced by the potential is partially hidden by the rest of the solution; in Figure 6.2b, where a close up of the solution over a line is represented, we can see it more clearly.

We consider three different potentials, given by $V(x) = r^{-\alpha}$, with $\alpha \in \{1/2, 1, 3/2\}$. Clearly, the bigger the exponent α, the lower the regularity of the exact solution. In particular, from the point of view of classical Sobolev spaces, denoting u_α as the solution of

$$\begin{aligned}(-\Delta + r^{-\alpha})u_\alpha &= \lambda_\alpha u_\alpha \text{ in } \Omega \\ u_\alpha &= 0 \text{ on } \partial\Omega,\end{aligned}$$

we have $u_\alpha \in H^{3-\alpha-\xi}(\Omega)$, for any $\xi > 0$. In particular, the problem with $\alpha = 3/2$ roughly corresponds to a two dimensional elliptic problem in a domain with a crack, see [CD02]. When considering weighted Sobolev spaces, we have

$$u_\alpha \in \mathcal{J}^{\varpi}_{3-\alpha-\xi}(\Omega), \tag{6.2}$$

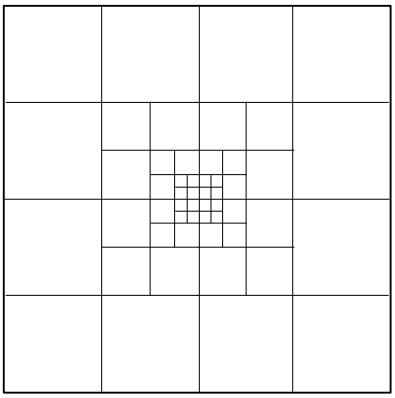


Figure 6.1 – Example of a two dimensional mesh, with $\ell = 5$

Table 6.1 – Estimated coefficients. Potential: $r^{-1/2}$

$\mathfrak{s}$	b_{L^2}	b_{DG}	b_{L^∞}	b_λ
0.125	0.73	0.78	0.78	1.34
0.25	0.9	0.89	0.86	1.12
0.5	1.07	1	1	1.19

again for any $\xi > 0$.

From the algebraic point of view, the eigenpairs are computed using a Krylov-Schur method [Ste02]. Furthermore, a shift and invert spectral transformation is used to precondition and speed up computations. Due to the relatively small size of the problems we consider here, the linear system introduced by the shift and invert spectral transformation is solved via an LU decomposition. When considering the problem set in three dimensions, we will see how to deal with problems with more degrees of freedom, where memory availability becomes a concern.

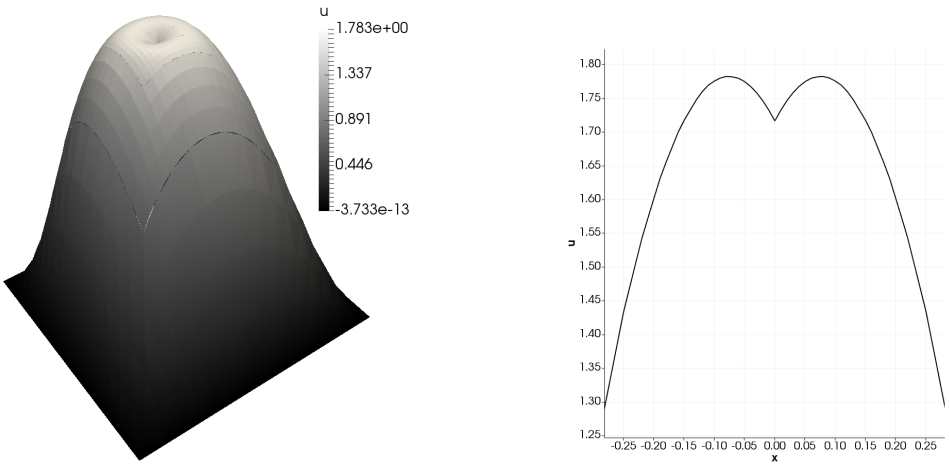
We conclude by remarking that the estimate on the eigenvalue we give here have the form

$$|\lambda - \lambda_\delta| \tag{6.3}$$

while in Chapter 5 the estimates we obtained are scaled as

$$\frac{|\lambda - \lambda_\delta|}{\lambda\lambda_\delta}. \tag{6.4}$$

This does not change anything from the point of view of the analysis; note nonetheless that (6.4) is between two and three orders of magnitude smaller than (6.3).



(a) (b)

Figure 6.2 – Numerical solution to (6.1) with $V(x) = r^{-1}$. Figure a: representation vertically not to scale; the separation between some elements is an artifact of the visualization on grids with hanging nodes. Figure b: close up around the singularity of the function $u(\cdot, 0)$, i.e., of u on the line $\{y = 0\}$.

Table 6.2 – Estimated coefficients. Potential: r^{-1}

$\mathfrak{s}$	b_{L^2}	b_{DG}	b_{L^∞}	b_λ
0.125	0.76	0.75	0.71	1.37
0.25	0.87	0.85	0.84	1.12
0.5	0.72	0.72	0.63	0.64

Table 6.3 – Estimated coefficients. Potential: $r^{-3/2}$

$\mathfrak{s}$	b_{L^2}	b_{DG}	b_{L^∞}	b_λ
0.062	0.49	0.48	0.47	0.85
0.125	0.61	0.59	0.64	1.09
0.25	0.6	0.53	0.42	0.48

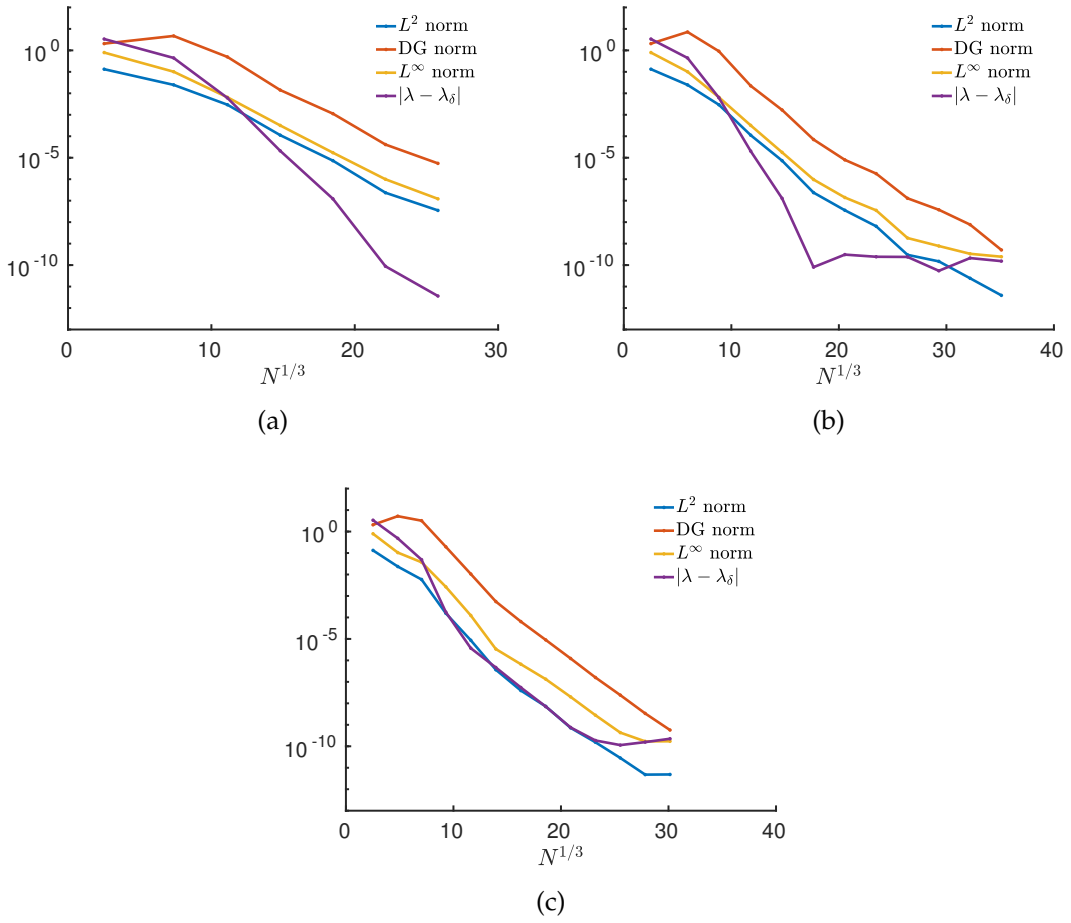


Figure 6.3 – Errors for the numerical solution with potential $V(x) = r^{-1/2}$. Polynomial slope: $\mathfrak{s} = 1/8$ in Figure a; $\mathfrak{s} = 1/4$ in Figure b and $\mathfrak{s} = 1/2$ in Figure c.

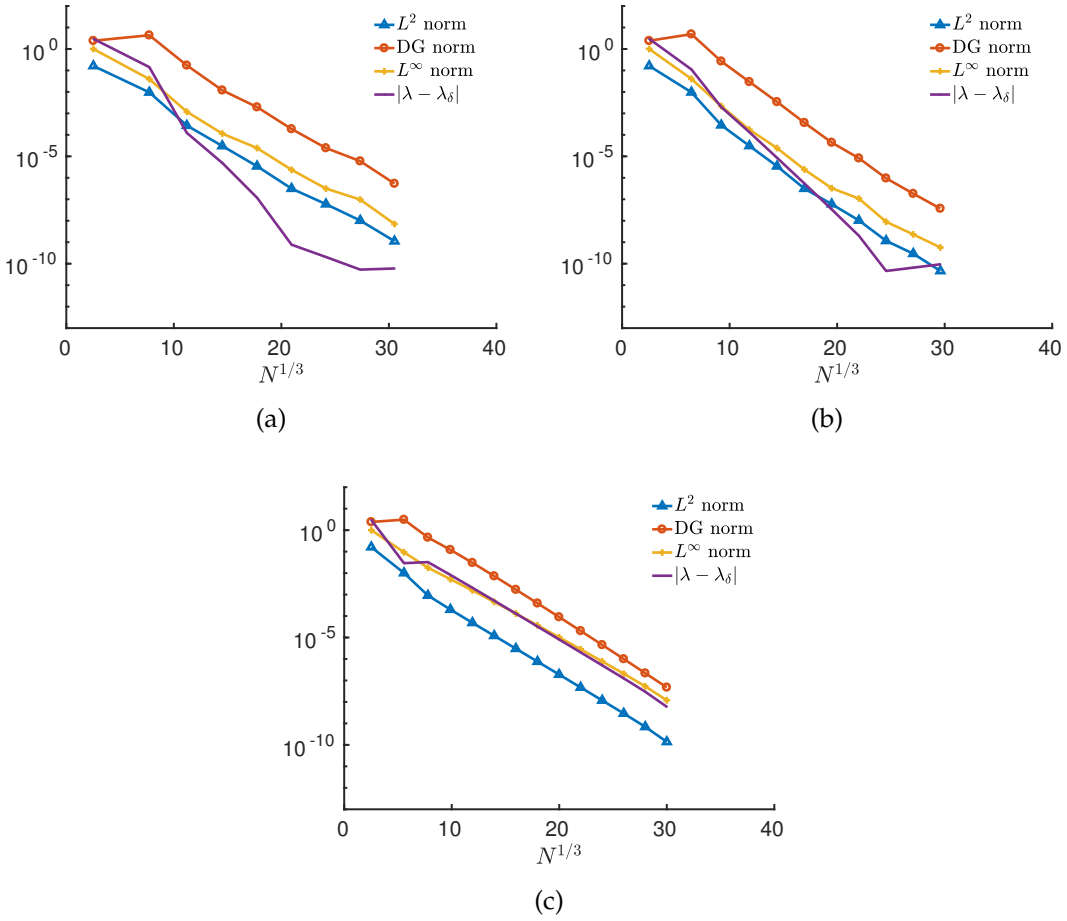


Figure 6.4 – Errors for the numerical solution with potential $V(x) = r^{-1}$. Polynomial slope: $\mathfrak{s} = 1/8$ in Figure a; $\mathfrak{s} = 1/4$ in Figure b and $\mathfrak{s} = 1/2$ in Figure c.

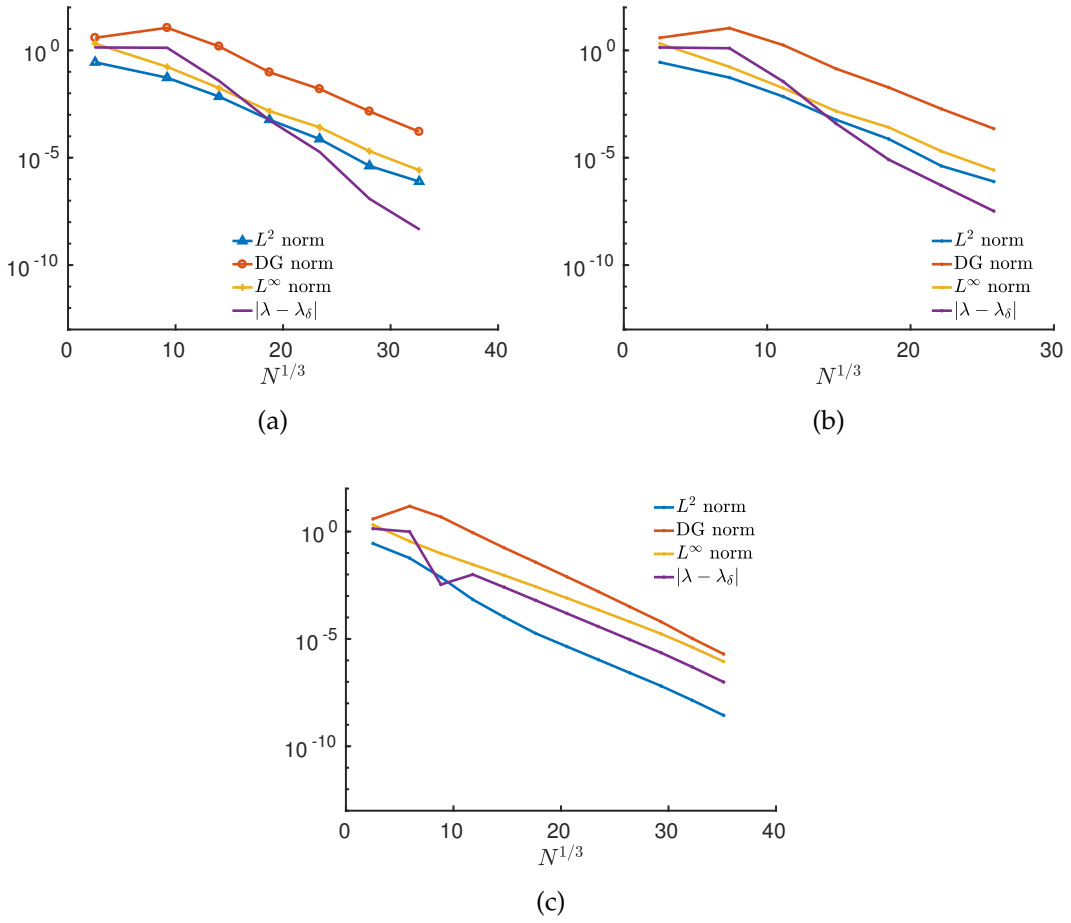


Figure 6.5 – Errors for the numerical solution with potential $V(x) = r^{-3/2}$. Polynomial slope: $\mathfrak{s} = 1/16$ in Figure a; $\mathfrak{s} = 1/8$ in Figure b and $\mathfrak{s} = 1/4$ in Figure c.

Table 6.4 – Estimated coefficients. Potential: r^{-1}, high degree quadrature formula

$\mathfrak{s}$	b_{L^2}	b_{DG}	b_{L^∞}	b_λ
0.25	0.89	0.81	0.85	1.5

6.1.1 Analysis of the results

The results on the error for the potential $V(x) = r^{-1/2}$ are shown in Figure 6.3, and the estimated coefficients are given in Table 6.1. Similarly, when the potential is given by $V(x) = r^{-1}$ the error curves are in Figure 6.4, with coefficients b_X in Table 6.2, and the case $V(x) = r^{-3/2}$ is reported in Figure 6.5 and Table 6.3.

We can clearly see that in many cases the error reaches at some point a plateau; we estimate the coefficients b_X by linear regression on the points before the plateau. This will be done for all subsequent potentials. Furthermore, as expected, the less regular the potential, the slowest the convergence of the numerical solution.

Two phenomena are less expected from the point of view of the theory. The first one is the emergence of a plateau at relatively high values compared to the machine epsilon. Through the choice of different algebraic scheme, we can see that we get a lower plateau: this is an indication that the dominating error at the points where it is not converging to zero is the algebraic one. The fact that matrices arising from the hp method are ill conditioned explains the size of the algebraic error. In practical applications, the fact that a relative error of approximately 10^{-12} can be reached should be sufficient.

The second "unexpected phenomenon" is evident when looking at Figures 6.3c, 6.4b, 6.5b, and 6.5c. We remark that, after an initial part where the eigenvalue converges faster than the other norms of the error, its rate of convergence then stabilizes to the same rate of the other norms. This can be shown [CCM10] to be dependent on the quadrature formula employed. When using a higher degree quadrature formula, the highest rate for the eigenvalue error is recovered, see Figure 6.6 and Table 6.4, obtained with a higher quadrature formula and compare them with Figure 6.4b and Table 6.2. As a side effect of a higher quadrature order, the plateau is raised.

In practice, one has to quite carefully balance computational cost, conditioning of the matrix, and speed of convergence. The usefulness of this numerical experiments lies therefore not only in the fact that we verify our theoretical results and we see the impact of components of the error we did not account for in the theoretical analysis, but also in the fact that we see, practically, how the parameters affect the simulation for different exact solutions. Since by asymptotic analysis we can see, locally and *a priori*, how the solution of a problem behaves, this gives an indication on how to construct and locally *a priori* optimize the hp spaces.

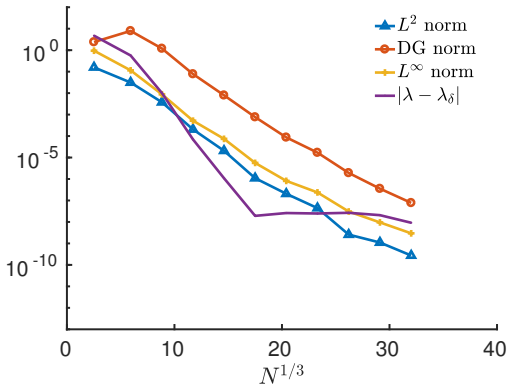


Figure 6.6 – Errors of the numerical solution for $V(x) = r^{-1}$ and a high degree quadrature formula. Polynomial slope $\mathfrak{s} = 0.25$.

6.1.2 Detailed tables of the errors

In this section, we show the exact values for the plots shown above. This makes more explicit the number of degrees of freedom used, which is someway hidden in the previous exposition.

Writing $V(x) = r^{-\alpha}$, we have the results for $\alpha = 1/2$ in Tables 6.5 to 6.7, those for $\alpha = 1$ in Tables 6.8 to 6.10 and those for $\alpha = 3/2$ in Tables 6.11 to 6.13.

The values of the degrees of freedom for the steps of the approximation differ depending on the slope $\mathfrak{s}$; this is due to the fact that every reported approximation is a multiple of $1/\mathfrak{s}$ refinements away from the previous one. The reason for this is that using those points makes it easier to have smooth error curves, and the estimation of the coefficients b_X is therefore easier.

Table 6.5 – Errors. Potential: $r^{-0.5}$, polynomial slope $\mathfrak{s} = 0.125$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.13	2.06	0.79	3.38
400	$2.47{\cdot}10^{-2}$	4.69	0.1	0.44
1,374	$2.94{\cdot}10^{-3}$	0.5	$6.53{\cdot}10^{-3}$	$6.09{\cdot}10^{-3}$
3,242	$1.12{\cdot}10^{-4}$	$1.42{\cdot}10^{-2}$	$3.19{\cdot}10^{-4}$	$2.1{\cdot}10^{-5}$
6,306	$7.3{\cdot}10^{-6}$	$1.13{\cdot}10^{-3}$	$1.72{\cdot}10^{-5}$	$1.22{\cdot}10^{-7}$
10,857	$2.35{\cdot}10^{-7}$	$4.11{\cdot}10^{-5}$	$9.92{\cdot}10^{-7}$	$8.77{\cdot}10^{-11}$
17,179	$3.52{\cdot}10^{-8}$	$5.45{\cdot}10^{-6}$	$1.22{\cdot}10^{-7}$	$3.62{\cdot}10^{-12}$

Table 6.6 – Errors. Potential: $r^{-0.5}$, polynomial slope $\mathfrak{s} = 0.25$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.13	2.06	0.79	3.38
208	$2.47{\cdot}10^{-2}$	7.18	0.1	0.44
690	$2.94{\cdot}10^{-3}$	0.91	$6.58{\cdot}10^{-3}$	$6.26{\cdot}10^{-3}$
1,642	$1.12{\cdot}10^{-4}$	$2.24{\cdot}10^{-2}$	$3.22{\cdot}10^{-4}$	$2.04{\cdot}10^{-5}$
3,190	$7.3{\cdot}10^{-6}$	$1.67{\cdot}10^{-3}$	$1.77{\cdot}10^{-5}$	$1.28{\cdot}10^{-7}$
5,493	$2.37{\cdot}10^{-7}$	$7.04{\cdot}10^{-5}$	$9.65{\cdot}10^{-7}$	$8.01{\cdot}10^{-11}$
8,686	$3.57{\cdot}10^{-8}$	$7.96{\cdot}10^{-6}$	$1.4{\cdot}10^{-7}$	$3.08{\cdot}10^{-10}$
12,928	$6.43{\cdot}10^{-9}$	$1.82{\cdot}10^{-6}$	$3.54{\cdot}10^{-8}$	$2.44{\cdot}10^{-10}$
18,369	$3{\cdot}10^{-10}$	$1.29{\cdot}10^{-7}$	$1.82{\cdot}10^{-9}$	$2.4{\cdot}10^{-10}$
25,144	$1.49{\cdot}10^{-10}$	$3.77{\cdot}10^{-8}$	$7.74{\cdot}10^{-10}$	$5.42{\cdot}10^{-11}$
33,402	$2.43{\cdot}10^{-11}$	$7.67{\cdot}10^{-9}$	$3.36{\cdot}10^{-10}$	$2.12{\cdot}10^{-10}$
43,299	$3.89{\cdot}10^{-12}$	$5.06{\cdot}10^{-10}$	$2.45{\cdot}10^{-10}$	$1.53{\cdot}10^{-10}$

Table 6.7 – Errors. Potential: $r^{-0.5}$, polynomial slope $\mathfrak{s} = 0.5$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.13	2.06	0.79	3.38
112	$2.35{\cdot}10^{-2}$	5.2	0.11	0.48
348	$5.86{\cdot}10^{-3}$	3.21	$3.72{\cdot}10^{-2}$	$4.9{\cdot}10^{-2}$
802	$1.57{\cdot}10^{-4}$	0.2	$2.66{\cdot}10^{-3}$	$1.77{\cdot}10^{-4}$
1,566	$8.64{\cdot}10^{-6}$	$1.06{\cdot}10^{-2}$	$1.24{\cdot}10^{-4}$	$3.69{\cdot}10^{-6}$
2,710	$3.67{\cdot}10^{-7}$	$5.51{\cdot}10^{-4}$	$3.38{\cdot}10^{-6}$	$4.66{\cdot}10^{-7}$
4,305	$3.95{\cdot}10^{-8}$	$6.43{\cdot}10^{-5}$	$6.65{\cdot}10^{-7}$	$5.47{\cdot}10^{-8}$
6,415	$7.2{\cdot}10^{-9}$	$9.01{\cdot}10^{-6}$	$1.34{\cdot}10^{-7}$	$7.11{\cdot}10^{-9}$
9,120	$7.25{\cdot}10^{-10}$	$1.23{\cdot}10^{-6}$	$2{\cdot}10^{-8}$	$7.53{\cdot}10^{-10}$
12,502	$1.53{\cdot}10^{-10}$	$1.59{\cdot}10^{-7}$	$2.79{\cdot}10^{-9}$	$1.86{\cdot}10^{-10}$
16,619	$2.85{\cdot}10^{-11}$	$2.41{\cdot}10^{-8}$	$4.33{\cdot}10^{-10}$	$1.14{\cdot}10^{-10}$
21,558	$4.79{\cdot}10^{-12}$	$3.39{\cdot}10^{-9}$	$1.69{\cdot}10^{-10}$	$1.56{\cdot}10^{-10}$
27,386	$4.87{\cdot}10^{-12}$	$5.7{\cdot}10^{-10}$	$1.71{\cdot}10^{-10}$	$2.24{\cdot}10^{-10}$

Table 6.8 – Errors. Potential: r^{-1}, polynomial slope $\mathfrak{s} = 0.125$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.16	2.41	1	2.97
460	$9.53{\cdot}10^{-3}$	4.37	$3.95{\cdot}10^{-2}$	0.14
1,408	$2.73{\cdot}10^{-4}$	0.17	$1.2{\cdot}10^{-3}$	$1.24{\cdot}10^{-4}$
3,052	$3.04{\cdot}10^{-5}$	$1.2{\cdot}10^{-2}$	$1.15{\cdot}10^{-4}$	$4.94{\cdot}10^{-6}$
5,584	$3.45{\cdot}10^{-6}$	$1.96{\cdot}10^{-3}$	$2.43{\cdot}10^{-5}$	$1.18{\cdot}10^{-7}$
9,196	$3.14{\cdot}10^{-7}$	$1.88{\cdot}10^{-4}$	$2.36{\cdot}10^{-6}$	$7.67{\cdot}10^{-10}$
14,080	$5.94{\cdot}10^{-8}$	$2.44{\cdot}10^{-5}$	$3.15{\cdot}10^{-7}$	$2.05{\cdot}10^{-10}$
20,428	$1.02{\cdot}10^{-8}$	$5.96{\cdot}10^{-6}$	$9.68{\cdot}10^{-8}$	$5.3{\cdot}10^{-11}$
28,432	$1.12{\cdot}10^{-9}$	$5.37{\cdot}10^{-7}$	$7.07{\cdot}10^{-9}$	$5.98{\cdot}10^{-11}$

Table 6.9 – Errors. Potential: r^{-1}, polynomial slope $\mathfrak{s} = 0.25$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.16	2.41	1	2.97
268	$9.62{\cdot}10^{-3}$	4.9	$4.05{\cdot}10^{-2}$	0.11
784	$2.83{\cdot}10^{-4}$	0.27	$2.19{\cdot}10^{-3}$	$1.97{\cdot}10^{-3}$
1,660	$3.07{\cdot}10^{-5}$	$2.95{\cdot}10^{-2}$	$1.75{\cdot}10^{-4}$	$1.33{\cdot}10^{-4}$
2,992	$3.46{\cdot}10^{-6}$	$3.44{\cdot}10^{-3}$	$2.42{\cdot}10^{-5}$	$8.54{\cdot}10^{-6}$
4,876	$3.14{\cdot}10^{-7}$	$3.67{\cdot}10^{-4}$	$2.42{\cdot}10^{-6}$	$5.43{\cdot}10^{-7}$
7,408	$5.94{\cdot}10^{-8}$	$4.4{\cdot}10^{-5}$	$3.28{\cdot}10^{-7}$	$3.38{\cdot}10^{-8}$
10,684	$1.03{\cdot}10^{-8}$	$8.18{\cdot}10^{-6}$	$1.08{\cdot}10^{-7}$	$2{\cdot}10^{-9}$
14,800	$1.16{\cdot}10^{-9}$	$9.54{\cdot}10^{-7}$	$8.97{\cdot}10^{-9}$	$4.59{\cdot}10^{-11}$
19,852	$2.93{\cdot}10^{-10}$	$1.83{\cdot}10^{-7}$	$2.31{\cdot}10^{-9}$	$6.54{\cdot}10^{-11}$
25,936	$4.61{\cdot}10^{-11}$	$3.7{\cdot}10^{-8}$	$5.7{\cdot}10^{-10}$	$9.37{\cdot}10^{-11}$

Table 6.10 – Errors. Potential: r^{-1}, polynomial slope $\mathfrak{s} = 0.5$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.16	2.41	1	2.97
172	$1.01{\cdot}10^{-2}$	3.09	$9.33{\cdot}10^{-2}$	$2.91{\cdot}10^{-2}$
472	$9.07{\cdot}10^{-4}$	0.46	$1.78{\cdot}10^{-2}$	$3.25{\cdot}10^{-2}$
964	$2.01{\cdot}10^{-4}$	0.12	$5.2{\cdot}10^{-3}$	$8.22{\cdot}10^{-3}$
1,696	$4.85{\cdot}10^{-5}$	$3.04{\cdot}10^{-2}$	$1.56{\cdot}10^{-3}$	$2.07{\cdot}10^{-3}$
2,716	$1.21{\cdot}10^{-5}$	$7.27{\cdot}10^{-3}$	$4.57{\cdot}10^{-4}$	$5.2{\cdot}10^{-4}$
4,072	$3.03{\cdot}10^{-6}$	$1.7{\cdot}10^{-3}$	$1.31{\cdot}10^{-4}$	$1.3{\cdot}10^{-4}$
5,812	$7.56{\cdot}10^{-7}$	$3.94{\cdot}10^{-4}$	$3.7{\cdot}10^{-5}$	$3.25{\cdot}10^{-5}$
7,984	$1.89{\cdot}10^{-7}$	$9{\cdot}10^{-5}$	$1.03{\cdot}10^{-5}$	$8.13{\cdot}10^{-6}$
10,636	$4.72{\cdot}10^{-8}$	$2.04{\cdot}10^{-5}$	$2.83{\cdot}10^{-6}$	$2.03{\cdot}10^{-6}$
13,816	$1.18{\cdot}10^{-8}$	$4.55{\cdot}10^{-6}$	$7.71{\cdot}10^{-7}$	$5.06{\cdot}10^{-7}$
17,572	$2.91{\cdot}10^{-9}$	$1{\cdot}10^{-6}$	$2.07{\cdot}10^{-7}$	$1.25{\cdot}10^{-7}$
21,952	$6.92{\cdot}10^{-10}$	$2.18{\cdot}10^{-7}$	$5.35{\cdot}10^{-8}$	$2.98{\cdot}10^{-8}$
27,004	$1.39{\cdot}10^{-10}$	$4.77{\cdot}10^{-8}$	$1.19{\cdot}10^{-8}$	$5.95{\cdot}10^{-9}$

Table 6.11 – Errors. Potential: $r^{-1.5}$, polynomial slope $\mathfrak{s} = 0.0625$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.28	3.86	2.07	1.36
400	$5.41{\cdot}10^{-2}$	10.9	0.17	1.26
1,374	$7.06{\cdot}10^{-3}$	1.77	$1.75{\cdot}10^{-2}$	$3.55{\cdot}10^{-2}$
3,242	$5.91{\cdot}10^{-4}$	0.14	$1.48{\cdot}10^{-3}$	$3.91{\cdot}10^{-4}$
6,306	$7.47{\cdot}10^{-5}$	$1.87{\cdot}10^{-2}$	$2.61{\cdot}10^{-4}$	$8.17{\cdot}10^{-6}$
10,857	$4.15{\cdot}10^{-6}$	$1.89{\cdot}10^{-3}$	$2.02{\cdot}10^{-5}$	$5.15{\cdot}10^{-7}$
17,194	$7.79{\cdot}10^{-7}$	$2.27{\cdot}10^{-4}$	$2.65{\cdot}10^{-6}$	$3.26{\cdot}10^{-8}$

Table 6.12 – Errors. Potential: $r^{-1.5}$, polynomial slope $\mathfrak{s} = 0.125$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.28	3.86	2.07	1.36
400	$5.41{\cdot}10^{-2}$	10.9	0.17	1.26
1,374	$7.06{\cdot}10^{-3}$	1.77	$1.75{\cdot}10^{-2}$	$3.55{\cdot}10^{-2}$
3,242	$5.91{\cdot}10^{-4}$	0.14	$1.48{\cdot}10^{-3}$	$3.91{\cdot}10^{-4}$
6,306	$7.47{\cdot}10^{-5}$	$1.87{\cdot}10^{-2}$	$2.61{\cdot}10^{-4}$	$8.17{\cdot}10^{-6}$
10,857	$4.15{\cdot}10^{-6}$	$1.89{\cdot}10^{-3}$	$2.02{\cdot}10^{-5}$	$5.15{\cdot}10^{-7}$
17,194	$7.79{\cdot}10^{-7}$	$2.27{\cdot}10^{-4}$	$2.65{\cdot}10^{-6}$	$3.26{\cdot}10^{-8}$

Table 6.13 – Errors. Potential: $r^{-1.5}$, polynomial slope $\mathfrak{s} = 0.25$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.28	3.86	2.07	1.36
208	$5.8{\cdot}10^{-2}$	15.1	0.35	0.99
690	$7.38{\cdot}10^{-3}$	4.83	$9.58{\cdot}10^{-2}$	$3.41{\cdot}10^{-3}$
1,642	$6.84{\cdot}10^{-4}$	0.89	$2.95{\cdot}10^{-2}$	$9.99{\cdot}10^{-3}$
3,190	$1.05{\cdot}10^{-4}$	0.18	$9.11{\cdot}10^{-3}$	$2.54{\cdot}10^{-3}$
5,493	$1.81{\cdot}10^{-5}$	$3.74{\cdot}10^{-2}$	$2.72{\cdot}10^{-3}$	$6.25{\cdot}10^{-4}$
8,686	$4.4{\cdot}10^{-6}$	$7.8{\cdot}10^{-3}$	$7.96{\cdot}10^{-4}$	$1.54{\cdot}10^{-4}$
12,928	$1.08{\cdot}10^{-6}$	$1.61{\cdot}10^{-3}$	$2.27{\cdot}10^{-4}$	$3.79{\cdot}10^{-5}$
18,369	$2.64{\cdot}10^{-7}$	$3.19{\cdot}10^{-4}$	$6.35{\cdot}10^{-5}$	$9.35{\cdot}10^{-6}$
25,144	$6.48{\cdot}10^{-8}$	$6.36{\cdot}10^{-5}$	$1.74{\cdot}10^{-5}$	$2.3{\cdot}10^{-6}$
33,437	$1.39{\cdot}10^{-8}$	$1.05{\cdot}10^{-5}$	$4.11{\cdot}10^{-6}$	$4.91{\cdot}10^{-7}$
43,354	$2.75{\cdot}10^{-9}$	$1.97{\cdot}10^{-6}$	$8.82{\cdot}10^{-7}$	$9.76{\cdot}10^{-8}$

6.2 Three dimensional case

In the three dimensional case, we replicate the setting introduced in Section 6.1. In this case, $\Omega = (-1/2, 1/2)^3$. Note that the regularity of the solution of

$$\begin{aligned}(-\Delta + r^{-\alpha})u_\alpha &= \lambda_\alpha u_\alpha \text{ in } \Omega\\ u_\alpha &= 0 \text{ on } \partial\Omega,\end{aligned}$$

scales differently with respect to α, if compared to the two dimensional case. Specifically, we have

$$u_\alpha \in H^{7/2-\alpha-\xi}(\Omega)$$

and

$$u_\alpha \in \mathcal{J}^{\varpi}_{7/2-\alpha-\xi}(\Omega),$$

for any $\xi > 0$.

The mesh is built in a tensor product way as in Section 6.1, with refinement ratio $\sigma = 1/2$. A representation of a mesh is given in Figure 6.7. The numerical solution for $V(x) = r^{-1}$ is shown in Figure 6.8.

From the algebraic point of view, the assembled matrices are bigger in size and less sparse, thus a direct LU method is less feasible than in the previous case (up to completely unfeasible for the simulations with a high number of degrees of freedom). Hence, we turn to iterative methods, and try to employ an algebraic eigenvalue method that is not too sensible to the error introduced by the linear solver. Therefore, the search for the eigenvalues is done with a Jacobi-Davidson method [SV96]. Internally, we employ a biconjugate gradient stabilized method (BiCGS, [Vor92; SVF94]) as a linear solver, with simple Jacobi preconditioner. The tolerance for the linear solver is set at 10^{-6}, while the

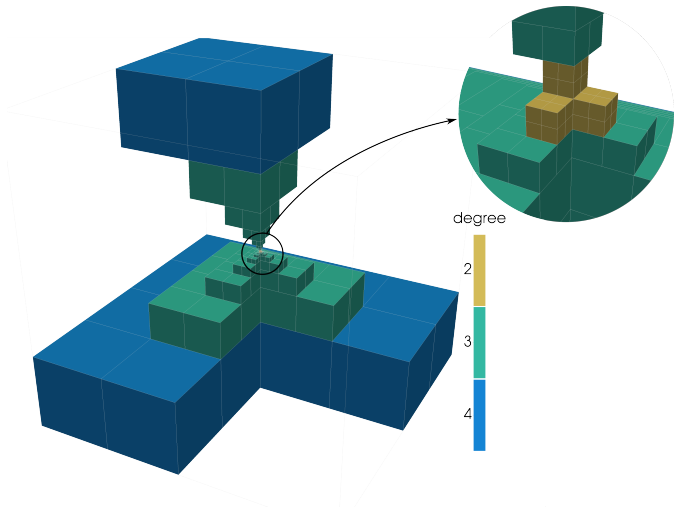


Figure 6.7 – Example mesh for the three dimensional approximation

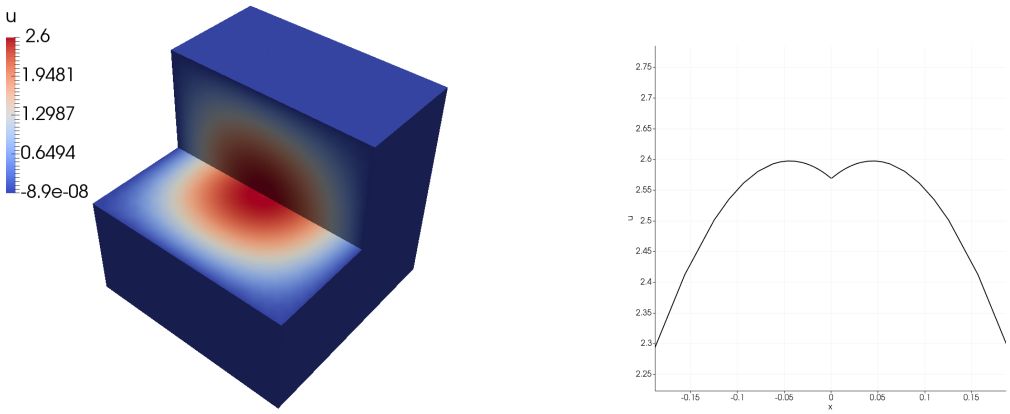


Figure 6.8 – Numerical solution in the three dimensional case: solution in the cube, left, and close up near the origin of the restriction to the line $\{y = z = 0\}$, right

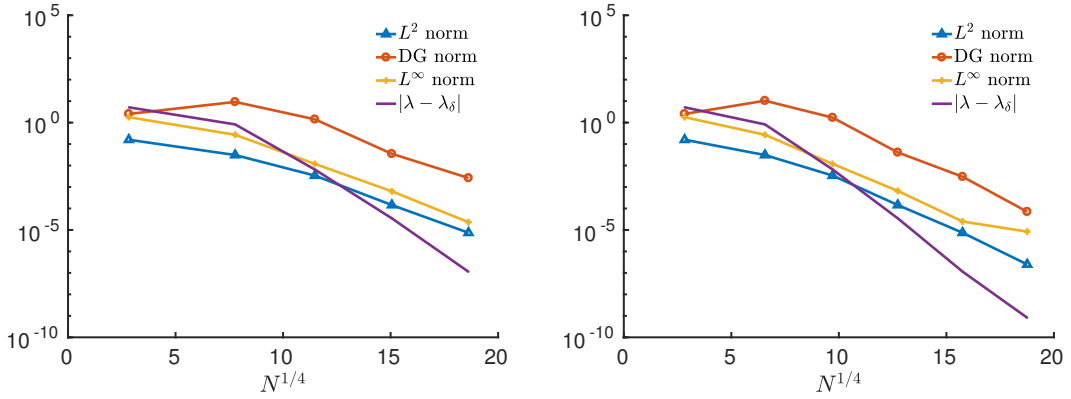


Figure 6.9 – Errors of the numerical solution for $V(x) = r^{-1/2}$. Polynomial slope $\mathfrak{s} = 1/8$, left and $\mathfrak{s} = 1/4$, right.

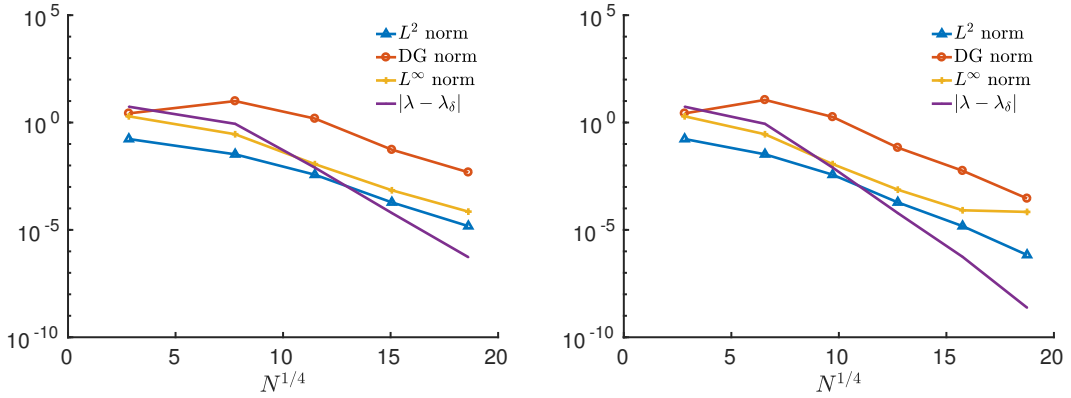


Figure 6.10 – Errors of the numerical solution for $V(x) = r^{-1}$. Polynomial slope $\mathfrak{s} = 1/8$, left and $\mathfrak{s} = 1/4$, right.

tolerance of the Jacobi-Davidson method is set at 10^{-8}.

6.2.1 Analysis of the results

Results for $V(x) = r^{-1/2}$ are given in Figure 6.9 and Table 6.14, while the case $V(x) = r^{-1}$ is analyzed in Figure 6.10 and Table 6.15 and the errors and estimates when $V(x) = r^{-3/2}$ are shown in Figure 6.11 and Table 6.16. The three dimensional approximation has far more degrees of freedom than the two dimensional one for a given level of refinement ℓ, thus the results we show have lower levels of refinement than the two dimensional ones. This is partially balanced by the fact that the solutions are more regular, but the errors are still obviously higher than those of the two dimensional case, at the same number of degrees of freedom, compare the tables in Sections 6.1.2 and 6.2.2.

In the three dimensional case, we do not see a great effect neither of the algebraic

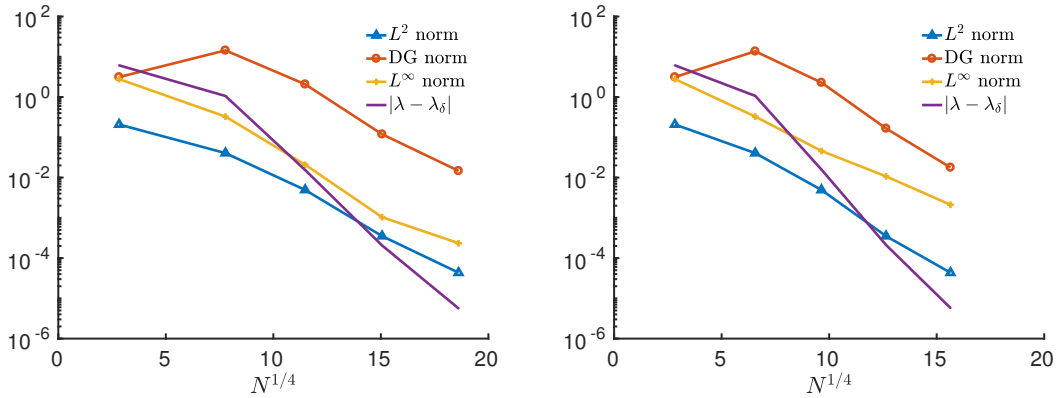


Figure 6.11 – Errors of the numerical solution for $V(x) = r^{-3/2}$. Polynomial slope $\mathfrak{s} = 1/8$, left and $\mathfrak{s} = 1/4$, right.

Table 6.14 – Estimated coefficients. Potential: $r^{-1/2}$

$\mathfrak{s}$	b_{L^2}	b_{DG}	b_{L^∞}	b_λ
0.125	0.78	0.78	0.86	1.46
0.25	0.97	0.99	0.89	1.72

Table 6.15 – Estimated coefficients. Potential: r^{-1}

$\mathfrak{s}$	b_{L^2}	b_{DG}	b_{L^∞}	b_λ
0.125	0.72	0.73	0.77	1.32
0.25	0.89	0.88	0.71	1.61

Table 6.16 – Estimated coefficients. Potential: $r^{-3/2}$

$\mathfrak{s}$	b_{L^2}	b_{DG}	b_{L^∞}	b_λ
0.125	0.64	0.65	0.69	1.13
0.25	0.77	0.75	0.55	1.35

error nor of the quadrature formulas. The coefficients b_λ listed in Tables 6.14 to 6.16 are almost double the respective coefficients b_{DG}; thus, if the effect of the quadrature error is present, it is nonetheless negligible compared to other sources of error for the quite comprehensive potentials and polynomial slopes considered in this experiments.

6.2.2 Detailed error tables

As we have done before in Section 6.1.2, we list here explicitly, in Tables 6.17 to 6.22, the values plotted in Figures 6.9 to 6.11 and from which we obtained the estimate in Tables 6.14 to 6.16. As already mentioned, this allows for a direct comparison with the bi-dimensional case. Furthermore, we can appreciate the difference in order of magnitude between the different approximations: for the highest number of degrees of freedom, the eigenvalue error is between 4 and 5 orders of magnitude smaller the DG norm of the error.

Table 6.17 – Errors. Potential: $r^{-0.5}$, polynomial slope $\mathfrak{s} = 0.125$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
64	0.16	2.53	1.79	5.14
3,648	$3.08{\cdot}10^{-2}$	9.19	0.27	0.82
17,359	$3.4{\cdot}10^{-3}$	1.42	$1.19{\cdot}10^{-2}$	$6.51{\cdot}10^{-3}$
51,419	$1.44{\cdot}10^{-4}$	$3.53{\cdot}10^{-2}$	$6.31{\cdot}10^{-4}$	$3.48{\cdot}10^{-5}$
$1.2 \cdot 10^5$	$7.41{\cdot}10^{-6}$	$2.64{\cdot}10^{-3}$	$2.31{\cdot}10^{-5}$	$1.13{\cdot}10^{-7}$

Table 6.18 – Errors. Potential: $r^{-0.5}$, polynomial slope $\mathfrak{s} = 0.25$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
64	0.16	2.53	1.79	5.14
1,856	$3.08{\cdot}10^{-2}$	10.5	0.27	0.82
8,892	$3.4{\cdot}10^{-3}$	1.71	$1.19{\cdot}10^{-2}$	$6.53{\cdot}10^{-3}$
26,340	$1.45{\cdot}10^{-4}$	$4.08{\cdot}10^{-2}$	$6.58{\cdot}10^{-4}$	$3.48{\cdot}10^{-5}$
61,585	$7.41{\cdot}10^{-6}$	$3.02{\cdot}10^{-3}$	$2.47{\cdot}10^{-5}$	$1.15{\cdot}10^{-7}$
$1.24 \cdot 10^5$	$2.48{\cdot}10^{-7}$	$7.07{\cdot}10^{-5}$	$8.38{\cdot}10^{-6}$	$8.17{\cdot}10^{-10}$

Table 6.19 – Errors. Potential: r^{-1}, polynomial slope $\mathfrak{s} = 0.125$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
64	0.17	2.67	1.95	5.45
3,648	$3.31{\cdot}10^{-2}$	9.99	0.29	0.87
17,359	$3.71{\cdot}10^{-3}$	1.53	$1.15{\cdot}10^{-2}$	$7.94{\cdot}10^{-3}$
51,419	$1.91{\cdot}10^{-4}$	$5.44{\cdot}10^{-2}$	$6.98{\cdot}10^{-4}$	$6.1{\cdot}10^{-5}$
$1.2 \cdot 10^5$	$1.49{\cdot}10^{-5}$	$4.83{\cdot}10^{-3}$	$7.09{\cdot}10^{-5}$	$5.36{\cdot}10^{-7}$

Table 6.20 – Errors. Potential: r^{-1}, polynomial slope $\mathfrak{s} = 0.25$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
64	0.17	2.67	1.95	5.45
1,856	$3.31{\cdot}10^{-2}$	11.3	0.29	0.87
8,892	$3.71{\cdot}10^{-3}$	1.82	$1.16{\cdot}10^{-2}$	$7.94{\cdot}10^{-3}$
26,340	$1.91{\cdot}10^{-4}$	$6.78{\cdot}10^{-2}$	$7.46{\cdot}10^{-4}$	$6.08{\cdot}10^{-5}$
61,585	$1.5{\cdot}10^{-5}$	$5.69{\cdot}10^{-3}$	$8.19{\cdot}10^{-5}$	$5.55{\cdot}10^{-7}$
$1.24 \cdot 10^5$	$6.79{\cdot}10^{-7}$	$2.91{\cdot}10^{-4}$	$6.87{\cdot}10^{-5}$	$2.38{\cdot}10^{-9}$

Table 6.21 – Errors. Potential: $r^{-1.5}$, polynomial slope $\mathfrak{s} = 0.125$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
64	0.21	3.14	2.79	6.13
3,648	$4.05{\cdot}10^{-2}$	14.4	0.33	1.06
17,359	$4.93{\cdot}10^{-3}$	2.08	$2.09{\cdot}10^{-2}$	$1.54{\cdot}10^{-2}$
51,400	$3.51{\cdot}10^{-4}$	0.12	$1.04{\cdot}10^{-3}$	$2.06{\cdot}10^{-4}$
$1.2 \cdot 10^5$	$4.34{\cdot}10^{-5}$	$1.47{\cdot}10^{-2}$	$2.35{\cdot}10^{-4}$	$5.62{\cdot}10^{-6}$

Table 6.22 – Errors. Potential: $r^{-1.5}$, polynomial slope $\mathfrak{s} = 0.25$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
64	0.21	3.14	2.79	6.13
1,856	$4.05{\cdot}10^{-2}$	13.7	0.33	1.07
8,645	$4.92{\cdot}10^{-3}$	2.29	$4.59{\cdot}10^{-2}$	$1.58{\cdot}10^{-2}$
25,574	$3.52{\cdot}10^{-4}$	0.17	$1.07{\cdot}10^{-2}$	$2.12{\cdot}10^{-4}$
59,857	$4.33{\cdot}10^{-5}$	$1.78{\cdot}10^{-2}$	$2.11{\cdot}10^{-3}$	$5.78{\cdot}10^{-6}$

Chapter 7

Weighted analytic regularity estimates for nonlinear eigenvalue problems

In this section we concern ourselves with the proof of analytic-type estimates on the norms of the solution to nonlinear elliptic problems. Specifically, we consider the nonlinear Schrödinger and Hartree-Fock equations and prove that, under some conditions on the coefficients of the operator, the solution belongs to $\mathcal{J}^{\varpi}_{\gamma}(\Omega)$, for the same γ as in the linear case. Since the singularities we consider are internal to the domain, we suppose that the domain Ω is a compact domain without without boundary, e.g., $\Omega = (-1,1)^d/2\mathbb{Z}$. The extension of the theory to the case of a bounded domain with smooth boundary can be done using the classical tools used in the analysis of elliptic problems in Sobolev spaces, as long as $r_{|\partial\Omega} \simeq 1$, i.e., the singularity is bounded away from the boundary.

First, in Section 7.1 we prove the local elliptic estimate in weighted Sobolev spaces that will allow for the derivation of the bounds on higher order derivatives from those obtained on lower order ones. Then, in order to estimate the norms of the nonlinear terms, we follow the proof technique used in [Dal+12]. The idea is to procede by induction and to consider L^p norms in nested balls and with a big enough p. Let L_{lin} be an elliptic linear operator and consider for example an operator $L_u u = L_{\text{lin}} u + u^{\delta}$, where $\delta = 2, 3, 4$: the L^p norms of the nonlinear terms can then be broken up into products of $L^{p\delta}$ norms by a Cauchy-Schwartz inequality. In order to get back to L^p norms, in [Dal+12] the authors use an interpolation inequality where $L^{\delta p}$ is seen as the interpolation between L^p and $W^{1,p}$. Since in our case we need to deal with weighted spaces, in Section 7.2 we derive the weighted version of this inequality, via a dyadic decomposition of the domain near the singular points.

The proof of the analytic bound on a nonlinear scalar elliptic eigenvalue problem is then given in Section 7.3, in the case of the nonlinear Schrödinger equation up to a quartic nonlinear term. Starting from a basic regularity assumption, we are able to treat

the potential and the nonlinear term thanks to the results presented in the preceding sections.

We then turn to the analysis of the Hartree-Fock equation: the nonlinear term in this problem is nonlocal, but cubic in nature. We remark that we can get rid of the convolution term by rewriting the Hartree-Fock equation as a system of nonlinear elliptic equations, as is done in (7.19). The analysis can then be carried out similarly as in the scalar case with cubic nonlinearity, with an additional step related to the nonlocality of the coefficients and given by Lemma 42. The analysis of the Hartree-Fock system is carried out in Section 7.4.

7.1 Local elliptic estimate

We start by proving a local seminorm estimate in weighted Sobolev spaces. This has been already established in [CDN12], as an intermediate estimate leading to the proof of another regularity result. We restate it here fully, in the specific form that will be needed in the sequel. The goal is to control the weighted norm of a higher order derivative of a function with the weighted norm of its laplacian and of lower order derivatives in a bigger domain, while giving an explicit dependence of the constants on the distance between the domains.

From now on, we denote the commutator by square brackets, i.e., we write

$$[A, B] = AB - BA.$$

Proposition 35. *Let $1 < p < \infty$, $R > 0$, $k \in \mathbb{N}$ and $\rho \in (0, \frac{R}{2(k+1)})$. Furthermore, let $\gamma \in \mathbb{R}$ and $j \in \mathbb{N}$ such that $1 \leq j \leq k$. Then,*

$$\sum_{|\alpha|=k+1} \|r^{k+1-\gamma}\partial^\alpha u\|_{L^p(B_{R-(j+1)\rho})} \lesssim \sum_{|\beta|=k-1} \|r^{k+1-\gamma}\partial^\beta(\Delta u)\|_{L^p(B_{R-j\rho})} + \sum_{|\alpha|=k} \rho^{-1}\|r^{|\alpha|-\gamma}\partial^\alpha u\|_{L^p(B_{R-j\rho})} + \sum_{|\alpha|=k-1} \rho^{-2}\|r^{|\alpha|-\gamma}\partial^\alpha u\|_{L^p(B_{R-j\rho})}. \quad (7.1)$$

In order to prove this proposition we introduce a smooth cutoff function $\psi \in C_0^\infty(B_{R-j\rho})$ such that for $\alpha \in \mathbb{N}^d$, $|\alpha| \leq 2$

$$0 \leq \psi \leq 1, \qquad \psi = 1 \text{ on } B_{R-(j+1)\rho}, \qquad |\partial^\alpha \psi| \leq C\rho^{-|\alpha|}, \quad (7.2)$$

and we derive an auxiliary estimate.

Lemma 36. *Let $\beta \in \mathbb{N}^d$, $1 < p < \infty$, $R > 0$, and $\rho \in (0, \frac{R}{2(|\beta|+2)})$. Then, for any $j \in \mathbb{N}$ such that $1 \leq j \leq |\beta| + 1$,*

$$\sum_{|\alpha|=2} \| \left[\partial^\alpha, r^{|\beta|+2-\gamma}\right] \psi\partial^\beta u\|_{L^p(B_{R-j\rho})} \leq C \sum_{|\alpha|\leq 1} \rho^{-2+|\alpha|}\|r^{|\beta|+|\alpha|-\gamma}\partial^{\alpha+\beta}u\|_{L^p(B_{R-j\rho})} \quad (7.3)$$

Proof. First, let us fix $i,k \in \{1,\dots,d\}$ such that $\partial^\alpha = \partial_i\partial_k$. Then, writing $(\star) = \| \left[\partial^\alpha, r^{|\beta|+2-\gamma}\right] \psi\partial^\beta u\|_{L^p(B_{R-j\rho})}$, we have that

$$\begin{aligned}(\star) &\lesssim \| \left(\partial_{ik} r^{|\beta|+2-\gamma}\right) \psi\partial^\beta u\|_{L^p(B_{R-j\rho})} + \| \left(\partial_i r^{|\beta|+2-\gamma}\right) \partial_k \left(\psi\partial^\beta u\right) \|_{L^p(B_{R-j\rho})} \\ &\lesssim \left(|\beta|^2 + \delta_{ik}|\beta|\right) \|r^{|\beta|-\gamma}\psi\partial^\beta u\|_{L^p(B_{R-j\rho})} + |\beta|\|r^{|\beta|+1-\gamma}\partial_k\left(\psi\partial^\beta u\right)\|_{L^p(B_{R-j\rho})},\end{aligned}$$

where $\delta_{ik}=1$ if $i=k$, $\delta_{ik}=0$ otherwise. Now,

$$\begin{aligned}|\beta|\|r^{|\beta|+1-\gamma}\partial_k\left(\psi\partial^\beta u\right)\|_{L^p(B_{R-j\rho})} &\lesssim |\beta|\|r^{|\beta|+1-\gamma}\psi\partial_k\partial^\beta u\|_{L^p(B_{R-j\rho})} \\ &\quad + |\beta|\|r^{|\beta|+1-\gamma}\left[\psi,\partial_k\right]\partial^\beta u\|_{L^p(B_{R-j\rho})} \\ &\lesssim |\beta|\|r^{|\beta|+1-\gamma}\psi\partial_k\partial^\beta u\|_{L^p(B_{R-j\rho})} \\ &\quad + |\beta|\|r^{|\beta|+1-\gamma}\left(\partial_k\psi\right)\partial^\beta u\|_{L^p(B_{R-j\rho})}.\end{aligned}$$

Denoting by $e_j \in \mathbb{N}^d$ the multi index with 1 at the jth position and 0 elsewhere, by the definition of ψ given in (7.2), we obtain

$$(\star) \lesssim \left(|\beta|^2 + \delta_{ij}|\beta| + |\beta|\rho^{-1}\right)\|r^{|\beta|-\gamma}\psi\partial^\beta u\|_{L^p(B_{R-j\rho})} + |\beta|\|r^{|\beta|+1-\gamma}\partial^{\beta+e_j}u\|_{L^p(B_{R-j\rho})}.$$

Summing over all multi indices $|\alpha|=2$,

$$\begin{aligned}\sum_{|\alpha|=2}\| \left[\partial^\alpha, r^{|\beta|+2-\gamma}\right]\psi\partial^\beta u\|_{L^p(B_{R-j\rho})} &\lesssim \left(|\beta|^2+|\beta|\rho^{-1}\right)\|r^{|\beta|-\gamma}\partial^\beta u\|_{L^p(B_{R-j\rho})} \\ &\quad + \sum_{|\alpha|=1}|\beta|\|r^{|\beta|+1-\gamma}\partial^{\beta+\alpha}u\|_{L^p(B_{R-j\rho})}.\end{aligned}$$

Since $\rho \in (0, \frac{R}{2(|\beta|+2)})$ implies $|\beta| \le \rho^{-1}$, we can conclude with (7.3). □

We can now prove estimate (7.1).

Proof of Proposition 35. Let us consider a multiindex β. First,

$$\begin{aligned}\sum_{|\alpha|=2}\|r^{|\beta|+2-\gamma}\partial^{\alpha+\beta}u\|_{L^p(B_{R-(j+1)\rho})} \le \sum_{|\alpha|=2}\Big\{&\|\partial^\alpha\left(r^{|\beta|+2-\gamma}\partial^\beta u\right)\|_{L^p(B_{R-(j+1)\rho})} \\ &+\| \left[\partial^\alpha, r^{|\beta|+2-\gamma}\right]\partial^\beta u\|_{L^p(B_{R-(j+1)\rho})}\Big\}. \quad (7.4)\end{aligned}$$

We consider the first term at the right hand side: using (7.2)

$$\sum_{|\alpha|=2}\|\partial^\alpha\left(r^{|\beta|+2-\gamma}\partial^\beta u\right)\|_{L^p(B_{R-(j+1)\rho})} \le \sum_{|\alpha|=2}\|\partial^\alpha\left(r^{|\beta|+2-\gamma}\psi\partial^\beta u\right)\|_{L^p(B_{R-j\rho})}$$

and by elliptic regularity and using the triangular inequality

$$\begin{aligned}\sum_{|\alpha|=2} \|\partial^\alpha \left(r^{|\beta|+2-\gamma}\psi\partial^\beta u\right)\|_{L^p(B_{R-j\rho})} &\leq C\|\Delta\left(r^{|\beta|+2-\gamma}\psi\partial^\beta u\right)\|_{L^p(B_{R-j\rho})} \\ &\leq C\|r^{|\beta|+2-\gamma}\psi\Delta\partial^\beta u\|_{L^p(B_{R-j\rho})} \\ &\quad + C\|\left[\Delta, r^{|\beta|+2-\gamma}\right]\psi\partial^\beta u\|_{L^p(B_{R-j\rho})} \\ &\quad + C\|r^{|\beta|+2-\gamma}\left[\Delta,\psi\right]\partial^\beta u\|_{L^p(B_{R-j\rho})}.\end{aligned}$$

Combining the last inequality with (7.4) we obtain

$$\begin{aligned}\sum_{|\alpha|=2} \|r^{|\beta|+2-\gamma}\partial^{\alpha+\beta}u\|_{L^p(B_{R-(j+1)\rho})} &\lesssim \|r^{|\beta|+2-\gamma}\psi\partial^\beta\left(\Delta u\right)\|_{L^p(B_{R-j\rho})} \\ &\quad + \sum_{i=1}^{d}\Big\{\|r^{|\beta|+2-\gamma}\left(\partial_{ii}\psi\right)\partial^\beta u\|_{L^p(B_{R-j\rho})} \\ &\qquad + \|r^{|\beta|+2-\gamma}\left(\partial_i\psi\right)\partial^\beta\partial_i u\|_{L^p(B_{R-j\rho})}\Big\} \\ &\quad + \sum_{|\alpha|=2}\|\left[\partial^\alpha, r^{|\beta|+2-\gamma}\right]\partial^\beta u\|_{L^p(B_{R-j\rho})}.\end{aligned}$$

The bounds on the derivatives of ψ given in (7.2) and the estimate of Lemma 36 then give

$$\begin{aligned}\sum_{|\alpha|=2} \|r^{|\beta|+2-\gamma}\partial^{\alpha+\beta}u\|_{L^p(B_{R-(j+1)\rho})} &\lesssim \|r^{|\beta|+2-\gamma}\psi\partial^\beta\left(\Delta u\right)\|_{L^p(B_{R-j\rho})} \\ &\quad + \sum_{|\alpha|\leq 1}\rho^{-2+|\alpha|}\|r^{|\beta|+|\alpha|-\gamma}\partial^{\alpha+\beta}u\|_{L^p(B_{R-j\rho})}.\end{aligned}$$

We can now sum over all multi indices β such that $|\beta| = k-1$ to obtain the thesis (7.1). □

7.2 Weighted interpolation estimate

Lemma 37. *Let $R > 0$ such that $B_R \in \Omega$, $u \in \mathcal{K}_\gamma^{|\beta|+1,p}(B_R)$, $\delta > 1$, $\gamma - d/p \geq 2/(1-\delta)$, and $p \geq d(1-1/\delta)$. Then, the following "interpolation" estimate holds*

$$\|r^{\frac{2-\gamma}{\delta}+|\beta|}\partial^\beta u\|_{L^{\delta p}(B_R)} \leq C\|r^{|\beta|-\gamma}\partial^\beta u\|_{L^p(B_R)}^{1-\vartheta}\left\{(|\beta|+1)^\vartheta\|r^{|\beta|-\gamma}\partial^\beta u\|_{L^p(B_R)}^\vartheta + \sum_{i=1}^d \|r^{|\beta|+1-\gamma}\partial^{\beta+e_i}u\|_{L^p(B_R)}^\vartheta\right\}, \quad (7.5)$$

with $\vartheta = \frac{d}{p}\left(1-\frac{1}{\delta}\right)$.

Proof. Consider a dyadic decomposition of Ω given by the sets

$$V^j = \left\{x \in \Omega : 2^{-j} \leq |x| \leq 2^{-j+1}\right\},\ j = 1, 2, \dots$$

and decompose the ball B_R into its intersections with the sets belonging to the decomposition, i.e., into $B^j = B_R \cap V^j$. Let us introduce the linear maps $\varphi_j : V^1 \to V^j$ and write with a hat the pullback of functions by φ_j^{-1}, e.g, $\hat{r} = r \circ \varphi_j^{-1}$ and $\widehat{B}^j = \varphi_j^{-1}(B^j)$. Then,

$$\|r^{\frac{2-\gamma}{\delta}+|\beta|}\partial^\beta u\|_{L^{\delta p}(B^j)} \leq 2^{\frac{j}{\delta}(\gamma-2-d/p)}\|\hat{r}^{\frac{2-\gamma}{\delta}+|\beta|}\hat{\partial}^\beta \hat{u}\|_{L^{\delta p}(\widehat{B}^j)}$$

We can now use the interpolation inequality

$$\|v\|_{L^{\delta p}(B)} \leq C\|v\|_{L^p(B)}^{1-\vartheta}\|v\|_{W^{1,p}(B)}^\vartheta,$$

for $B \subset \mathbb{R}^d$, $v \in W^{1,p}(B)$ and with ϑ defined as above, see [Dal+12]. Therefore,

$$\|r^{\frac{2-\gamma}{\delta}+|\beta|}\partial^\beta u\|_{L^{\delta p}(B^j)} \leq C2^{\frac{j}{\delta}(\gamma-2-d/p)}\|\hat{r}^{\frac{2-\gamma}{\delta}+|\beta|}\hat{\partial}^\beta \hat{u}\|_{L^p(\widehat{B}^j)}^{1-\vartheta}\sum_{|\alpha|=1}\|\hat{\partial}^\alpha \hat{r}^{\frac{2-\gamma}{\delta}+|\beta|}\hat{\partial}^\beta \hat{u}\|_{L^p(\widehat{B}^j)}^\vartheta. \quad (7.6)$$

Let us now consider the first norm in the product above. Since $\hat{r} \in (1/2, 1)$, we can inject in the norm a term $\hat{r}^{\gamma(1-1/\delta)} \leq \max(1, 2^{\gamma(1-1/\delta)}) = C(\gamma, \delta)$, i.e.,

$$\|\hat{r}^{\frac{2-\gamma}{\delta}+|\beta|}\hat{\partial}^\beta \hat{u}\|_{L^p(\widehat{B}^j)}^{1-\vartheta} \leq C\|\hat{r}^{|\beta|-\gamma}\hat{\partial}^\beta \hat{u}\|_{L^p(\widehat{B}^j)}^{1-\vartheta}.$$

We now compute more explicitely the second norm in the product in (7.6):

$$\sum_{|\alpha|=1}\|\hat{\partial}^\alpha \hat{r}^{\frac{2-\gamma}{\delta}+|\beta|}\hat{\partial}^\beta \hat{u}\|_{L^p(\widehat{B}^j)}^\vartheta \leq \left(|\beta| + \frac{2-\gamma}{\delta}\right)^\vartheta \|\hat{r}^{\frac{2-\gamma}{\delta}+|\beta|-1}\hat{\partial}^\beta \hat{u}\|_{L^p(\widehat{B}^j)}^\vartheta + \sum_{i=1}^d \|\hat{r}^{\frac{2-\gamma}{\delta}+|\beta|}\hat{\partial}^{\beta+e_i}\hat{u}\|_{L^p(\widehat{B}^j)}^\vartheta$$

and we may adjust the exponents of $\hat{r}$ and the term in $\frac{2-\gamma}{\delta}$ introducing a constant that

depends on γ, δ, d and p, obtaining

$$\sum_{|\alpha|=1} \|\hat{\partial}^\alpha \hat{r}^{\frac{2-\gamma}{\delta}+|\beta|}\hat{\partial}^\beta \hat{u}\|^\vartheta_{L^p(\widehat{B}^j)} \leq C\,(|\beta|+1)^\vartheta\, \|\hat{r}^{|\beta|-\gamma}\hat{\partial}^\beta \hat{u}\|^\vartheta_{L^p(\widehat{B}^j)} + \sum_{i=1}^d \|\hat{r}^{|\beta|-\gamma+1}\hat{\partial}^{\beta+e_i}\hat{u}\|^\vartheta_{L^p(\widehat{B}^j)}.$$

Scaling everything back to B^j and adjusting the exponents,

$$\|r^{\frac{2-\gamma}{\delta}+|\beta|}\partial^\beta u\|_{L^{\delta p}(B^j)} \leq C 2^{\frac{j}{\delta}((\gamma-d/p)(1-\delta)-2)} \|r^{|\beta|-\gamma}\partial^\beta u\|^{1-\vartheta}_{L^p(B^j)} \left\{ (|\beta|+1)^\vartheta\, \|r^{|\beta|-\gamma}\partial^\beta u\|^\vartheta_{L^p(B^j)} + \sum_{i=1}^d \|r^{|\beta|-\gamma+1}\partial^{\beta+e_i} u\|^\vartheta_{L^p(B^j)} \right\}.$$

If $\gamma - d/p \geq 2/(1-\delta)$ then we can sum over all $j = 1, 2, \ldots$ thus obtaining the estimate (7.1) on the whole ball B_R. □

7.3 Nonlinear Schrödinger

We now consider the nonlinear Schrödinger eigenvalue problem, given by

$$Lu = -\Delta u + Vu + |u|^{\delta-1}u = \lambda u. \tag{7.7}$$

We suppose that the potential V is singular on a finite set of discrete points and consider the case of an up-to-quartic nonlinearity (i.e., $\delta \in \mathbb{N}$ and $\delta \leq 4$). We show, in the following theorem, that the results on the regularity of the solution that can be obtained in the linear case can be extended to the nonlinear one. We fix $\varepsilon \in (0,1)$ to specify the regularity of the potential in weighted spaces.

Theorem 7. *Let u be the solution to* (7.7) *with $V \in \mathcal{K}^{\varpi,\infty}_{\varepsilon-2}(\Omega)$, and $\delta = 2, 3, 4$. Then,*

$$u \in \mathcal{J}^{\varpi,p}_\gamma(\Omega) \tag{7.8}$$

for any $\gamma < \min(d/p + \varepsilon, 2)$.

In order to prove the analiticity in weighted spaces of the function u we need to bound the nonlinear term. We will introduce some preliminary lemmas and proceed by induction: let us specify the induction hypothesis. We suppose, here and in the sequel, that we have fixed a nonempty ball $B_R \subset \Omega$.

Induction Assumption 1. *u satisfies Induction Assumption 1, up to $k \in \mathbb{N}$ and for $1 < p < \infty$, if $u \in \mathcal{J}^{k,p}_\gamma(\Omega)$ for any $\gamma \leq \widehat{\gamma} < d/p + \varepsilon$ and*

$$|u|_{\mathcal{K}^{|\alpha|,p}_\gamma(B_{R-k\rho})} \leq \widetilde{C} A^{|\alpha|} (k\rho)^{-|\alpha|} |\alpha|^{|\alpha|} \tag{7.9}$$

for all $\alpha \in \mathbb{N}^d$ *such that* $1 < |\alpha| \le k$, $\gamma \le \widehat{\gamma}$, *and* $\rho \in (0, R/(2k))$.

From now on, we suppose that $\delta \in \{2, 3, 4\}$.

Lemma 38. *Let* u *satisfy Induction Assumption 1 up to* $k \in \mathbb{N}$ *and for* p, γ *such that* $2/(1-\delta) \le \gamma - d/p < \min(\varepsilon, 2 - d/p)$ *and* $p \ge d(1 - 1/\delta)$. *Then,*

$$\|r^{\frac{2-\gamma}{\delta}+|\beta|}\partial^\beta u\|_{L^{\delta p}(B_{R-k\rho})} \le CA^{|\beta|+\vartheta}(k\rho)^{-|\beta|-\vartheta}|\beta|^{|\beta|}(|\beta|+1)^\vartheta \tag{7.10}$$

for $0 < |\beta| \le k-1$ *and with* $\vartheta = \frac{d}{p}\left(1 - \frac{1}{\delta}\right)$.

Proof. First, we use (7.5) in order to go back to integrals in L^p:

$$\|r^{\frac{2-\gamma}{\delta}+|\beta|}\partial^\beta u\|_{L^{\delta p}(B_{R-k\rho})} \le C\|r^{|\beta|-\gamma}\partial^\beta u\|^{1-\vartheta}_{L^p(B_{R-k\rho})} \left\{ (|\beta|+1)^\vartheta \|r^{|\beta|-\gamma}\partial^\beta u\|^\vartheta_{L^p(B_{R-k\rho})} + \sum_{i=1}^d \|r^{|\beta|+1-\gamma}\partial^\beta \partial_i u\|^\vartheta_{L^p(B_{R-k\rho})} \right\}.$$

Then, hypothesis (7.9) implies

$$\|r^{|\beta|-\gamma}\partial^\beta u\|^{1-\vartheta}_{L^p(B_{R-k\rho})} \le CA^{|\beta|(1-\vartheta)}\rho^{-|\beta|(1-\vartheta)}\left(\frac{|\beta|}{k}\right)^{|\beta|(1-\vartheta)}$$

and

$$(|\beta|+1)^\vartheta \|r^{|\beta|-\gamma}\partial^\beta u\|^\vartheta_{L^p(B_{R-k\rho})} + \sum_{i=1}^d \|r^{|\beta|+1-\gamma}\partial^\beta\partial_i u\|^\vartheta_{L^p(B_{R-k\rho})} \le C(|\beta|+1)^\vartheta A^{|\beta|\vartheta}\rho^{-|\beta|\vartheta}\left(\frac{|\beta|}{k}\right)^{|\beta|\vartheta} + CA^{(|\beta|+1)\vartheta}\rho^{-(|\beta|+1)\vartheta}\left(\frac{|\beta|+1}{k}\right)^{(|\beta|+1)\vartheta}.$$

Therefore, multyplying the right hand sides of the last two inequalities,

$$\|r^{\frac{2-\gamma}{\delta}+|\beta|}\partial^\beta u\|_{L^{\delta p}(B_{R-k\rho})} \le CA^{|\beta|+\vartheta}(k\rho)^{-|\beta|-\vartheta}|\beta|^{|\beta|(1-\vartheta)}(|\beta|+1)^{(|\beta|+1)\vartheta}.$$

We finally need to bound the last two terms in the multiplication above. By Stirling's formula,

$$|\beta|^{|\beta|(1-\vartheta)}(|\beta|+1)^{(|\beta|+1)\vartheta} \le C|\beta|!|\beta|^{-1/2}e^{|\beta|}(j+1)^{\vartheta/2}j^{\vartheta/2}$$

and another application of Stirling's formula gives the thesis. □

In order to estimate the L^p weighted norms of derivatives of u^δ we will use Leibniz's rule and break the L^p norms into multiple $L^{\delta p}$ norms. Lemma 38 then allows to go back to the induction hypothesis. We continue by estimating the weighted norms of u^2 through the procedure we just outlined. For two multi indices $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$

and $\beta = (\beta_1, \dots, \beta_d) \in \mathbb{N}^d$, we write $\alpha! = \alpha_1! \cdots \alpha_d!$, $\alpha + \beta = (\alpha_1 + \beta_1, \dots, \alpha_d + \beta_d)$, and

$$\binom{\alpha}{\beta} = \frac{\alpha!}{\beta!(\alpha - \beta)!}.$$

Furthermore, recall [Kat96] that

$$\sum_{\substack{|\beta|=n \\ \beta \le \alpha}} \binom{\alpha}{\beta} = \binom{|\alpha|}{n}.$$

Lemma 39. *Let u satisfy Induction Assumption 1 up to $|\alpha|$ and for p, γ such that $2/(1-\delta) \le \gamma - d/p < \min(\varepsilon, 2 - d/p)$ and $p \ge d(1 - 1/\delta)$. Then,*

$$\|r^{2\frac{2-\gamma}{\delta}+|\alpha|}\partial^\alpha(u^2)\|_{L^{\delta p/2}(B_{R-k\rho})} \le CA^{|\alpha|+2\vartheta}\rho^{-|\alpha|-2\vartheta}\left(\frac{|\alpha|}{k}\right)^{|\alpha|}|\alpha|^{1/2}. \tag{7.11}$$

Proof. By Leibniz's rule and the Cauchy-Schwartz inequality,

$$\begin{aligned}
&\|r^{2\frac{2-\gamma}{\delta}+|\alpha|}\partial^\alpha(u^2)\|_{L^{\delta p/2}(B_{R-k\rho})} \\
&\qquad \le \sum_{0<\beta<\alpha}\binom{\alpha}{\beta}\|r^{\frac{2-\gamma}{\delta}+|\beta|}\partial^\beta u\|_{L^{\delta p}(B_{R-k\rho})}\|r^{\frac{2-\gamma}{\delta}+|\alpha|-|\beta|}\partial^{\alpha-\beta}u\|_{L^{\delta p}(B_{R-k\rho})} \\
&\qquad\quad + 2\|r^{2\frac{2-\gamma}{\delta}+|\alpha|}\partial^\alpha u\|_{L^{\delta p/2}(B_{R-k\rho})}\|u\|_{L^\infty(B_{R-k\rho})}
\end{aligned} \tag{7.12}$$

Considering the sum over $0 < \beta < \alpha$, Lemma 38 and Stirling's inequality give

$$\begin{aligned}
&\sum_{0<\beta<\alpha}\binom{\alpha}{\beta}\|r^{\frac{2-\gamma}{\delta}+|\beta|}\partial^\beta u\|_{L^{\delta p}(B_{R-k\rho})}\|r^{\frac{2-\gamma}{\delta}+|\alpha|-|\beta|}\partial^{\alpha-\beta}u\|_{L^{\delta p}(B_{R-k\rho})} \\
&\quad \le CA^{|\alpha|+2\vartheta}\rho^{-|\alpha|-2\vartheta}\sum_{j=1}^{|\alpha|-1}\binom{|\alpha|}{j}j!(|\alpha|-j)!e^{|\alpha|}\frac{(j+1)^\vartheta(|\alpha|-j+1)^\vartheta}{k^{|\alpha|+2\vartheta}}\frac{1}{\sqrt{j(|\alpha|-j)}} \\
&\quad \le CA^{|\alpha|+2\vartheta}\rho^{-|\alpha|-2\vartheta}\frac{|\alpha|!e^{|\alpha|}}{k^{|\alpha|}} \\
&\quad \le CA^{|\alpha|+2\vartheta}\rho^{-|\alpha|-2\vartheta}\left(\frac{|\alpha|}{k}\right)^{|\alpha|}|\alpha|^{1/2}.
\end{aligned}$$

The second term at the right hand side of (7.12) is controlled using Lemma 38, as long as $\gamma \le 2$, and the injection $\mathcal{J}^2_{d/2+\eta}(\Omega) \hookrightarrow L^\infty(\Omega)$, valid for any $\eta > 0$ [KMR97]. □

With the same proof as above, we can deal with a cubic nonlinear term, as we show in the following lemma.

Lemma 40. *Under the same hypotheses as in Lemma 39,*

$$\|r^{3\frac{2-\gamma}{\delta}+|\alpha|}\partial^\alpha(u^3)\|_{L^{\delta p/3}(B_{R-k\rho})} \le CA^{|\alpha|+3\vartheta}\rho^{-|\alpha|-3\vartheta}\left(\frac{|\alpha|}{k}\right)^{|\alpha|}|\alpha| \tag{7.13}$$

Proof. We have

$$\begin{aligned}&\|r^{3\frac{2-\gamma}{\delta}+|\alpha|}\partial^\alpha(u^3)\|_{L^{\delta p/3}(B_{R-k\rho})}\\ &\qquad\le C\sum_{\beta\le\alpha}\binom{\alpha}{\beta}\|r^{\frac{2-\gamma}{\delta}+|\beta|}\partial^\beta u\|_{L^{\delta p}(B_{R-k\rho})}\|r^{2\frac{2-\gamma}{\delta}+|\alpha|-|\beta|}\partial^{\alpha-\beta}(u^2)\|_{L^{\delta p/2}(B_{R-k\rho})}.\end{aligned}$$

Using (7.11) we follow the same procedure as in the proof of Lemma 39. When $0<\beta<\alpha$ in the sum above,

$$\begin{aligned}&\sum_{0<\beta<\alpha}\binom{\alpha}{\beta}\|r^{\frac{2-\gamma}{\delta}+|\beta|}\partial^\beta u\|_{L^{\delta p}(B_{R-k\rho})}\|r^{2\frac{2-\gamma}{\delta}+|\alpha|-|\beta|}\partial^{\alpha-\beta}(u^2)\|_{L^{\delta p/2}(B_{R-k\rho})}\\ &\qquad\le CA^{|\alpha|+3\vartheta}\rho^{-|\alpha|-3\vartheta}\sum_{j=1}^{|\alpha|-1}\binom{|\alpha|}{j}j!(|\alpha|-j)!e^{|\alpha|}\frac{(j+1)^\vartheta(|\alpha|-j+1)^\vartheta}{k^{|\alpha|+2\vartheta}}\frac{\sqrt{|\alpha|-j}}{\sqrt{j(|\alpha|-j)}}\\ &\qquad\le CA^{|\alpha|+3\vartheta}\rho^{-|\alpha|-3\vartheta}\frac{|\alpha|!e^{|\alpha|}\sqrt{|\alpha|}}{k^{|\alpha|}}\\ &\qquad\le CA^{|\alpha|+3\vartheta}\rho^{-|\alpha|-3\vartheta}\left(\frac{|\alpha|}{k}\right)^{|\alpha|}|\alpha|.\end{aligned}$$

As before, the terms in the sum where $\beta=0$ and $\beta=\alpha$ give the same bound. □

The proof of the next lemma, in which we control a quartic term, amounts to a repetition of the arguments above; we show its proof for completeness.

Lemma 41. *Under the same hypotheses as in Lemma 39,*

$$\|r^{2-\gamma+|\alpha|}\partial^\alpha(u^4)\|_{L^p(B_{R-k\rho})} \le CA^{|\alpha|+4\vartheta}\rho^{-|\alpha|-4\vartheta}\left(\frac{|\alpha|}{k}\right)^{|\alpha|}|\alpha|^{3/2}. \tag{7.14}$$

Proof. There holds

$$\begin{aligned}&\|r^{2-\gamma+|\alpha|}\partial^\alpha(u^4)\|_{L^p(B_{R-k\rho})}\\ &\qquad\le C\sum_{\beta\le\alpha}\binom{\alpha}{\beta}\|r^{\frac{2-\gamma}{4}+|\beta|}\partial^\beta u\|_{L^{4p}(B_{R-k\rho})}\|r^{3\frac{2-\gamma}{4}+|\alpha|-|\beta|}\partial^{\alpha-\beta}(u^3)\|_{L^{4p/3}(B_{R-k\rho})}.\end{aligned}$$

We can now use the result of Lemma 40 with $\delta = 4$. When $0 < \beta < \alpha$ in the sum above,

$$\begin{aligned}
&\sum_{0<\beta<\alpha} \binom{\alpha}{\beta} \|r^{\frac{2-\gamma}{4}+|\beta|}\partial^\beta u\|_{L^{4p}(B_{R-k\rho})} \|r^{3\frac{2-\gamma}{4}+|\alpha|-|\beta|}\partial^{\alpha-\beta}(u^3)\|_{L^{4p/3}(B_{R-k\rho})} \\
&\quad \leq CA^{|\alpha|+4\vartheta}\rho^{-|\alpha|-4\vartheta} \sum_{j=1}^{|\alpha|-1} \binom{|\alpha|}{j} j!(|\alpha|-j)! e^{|\alpha|} \frac{(j+1)^\vartheta (|\alpha|-j+1)^\vartheta}{k^{|\alpha|+2\vartheta}} \frac{\sqrt{|\alpha|-j}}{\sqrt{j(|\alpha|-j)}} \\
&\quad \leq CA^{|\alpha|+4\vartheta}\rho^{-|\alpha|-4\vartheta} \left(\frac{|\alpha|}{k}\right)^{|\alpha|} |\alpha|.
\end{aligned}$$

The direct application of Lemmas 38 and 40 let us obtain the estimate for the terms in the sum where $\beta = \alpha$ and $\beta = 0$, respectively. □

The proof of (7.8) is now complete: we just need to bring the estimates together.

Proof of Theorem 7. The operator $L_{\text{lin}} = -\Delta + V$ is an isomorphism

$$\mathcal{J}^{k+2}_\gamma(\Omega) \to \mathcal{J}^k_{\gamma-2}(\Omega)$$

for any $0 < \gamma - d/2 < \varepsilon$ and all $k \in \mathbb{N}$, and, since we can also show that $u \in L^\infty(\Omega)$ [Sta65], the solution to (7.7) is such that $u \in \mathcal{J}^2_\gamma(\Omega)$. Iterating this line of reasoning, we can show that $u \in \mathcal{J}^3_\gamma(\Omega)$ for all $0 < \gamma - d/2 < \varepsilon$, thus, by injection, $u \in \mathcal{J}^{1,p}_\gamma(\Omega)$, for all $p > 1$, $0 < \gamma - d/p < \varepsilon$. The induction assumption is therefore verified up to $k = 2$, for all $p > 1$ and all $0 < \gamma - d/p < \varepsilon$.

We proceed by induction and impose a restriction on p; specifically,

$$p \geq 2d\frac{\delta - 1}{5 - \delta}. \tag{7.15}$$

The role of this condition on p will be clearer in the sequel. Let us now fix $\gamma \in (d/p, d/p+\varepsilon)$, suppose without loss of generality that $\widetilde{C} \geq \|u\|_{\mathcal{J}^{1,p}_\gamma(\Omega)}$ and suppose that the Induction Assumption 1 holds up until $k \in \mathbb{N}$, with p subject to (7.15). Then, let $0 < \rho \leq \frac{R}{2(k+1)}$: we will show that

$$\|r^{|\alpha|-\gamma}\partial^\alpha u\|_{L^p(B_{R-k\rho})} \leq \widetilde{C}A^{|\alpha|}(k\rho)^{-|\alpha|}|\alpha|^{|\alpha|} \tag{7.16}$$

holds for $|\alpha| = k+1$ under condition (7.15). From (7.1) and (7.7),

$$\begin{aligned}
\sum_{|\alpha|=k+1} \|r^{k+1-\gamma}\partial^\alpha u\|_{L^p(B_{R-(k+1)\rho})} \lesssim &\sum_{|\beta|=k-1} \|r^{k+1-\gamma}\partial^\beta(Vu + |u|^{\delta-1}u + \lambda u)\|_{L^p(B_{R-k\rho})} \\
&+ \sum_{|\alpha|=k-1,k} \rho^{|\alpha|-k-1}\|r^{|\alpha|-\gamma}\partial^\alpha u\|_{L^p(B_{R-|\alpha|\rho})}.
\end{aligned} \tag{7.17}$$

We fix a β such that $|\beta| = k-1$ and consider the term containing the potential V:

$$\begin{aligned}\|r^{k+1-\gamma}\partial^\beta(Vu)\|_{L^p(B_{R-k\rho})} \le \sum_{0<\zeta<\beta}\binom{\beta}{\zeta}\|r^{2-\varepsilon+|\zeta|}\partial^\zeta V\|_{L^\infty(B_{R-k\rho})}\|r^{\varepsilon-\gamma+|\beta|-|\zeta|}\partial^{\beta-\zeta}u\|_{L^p(B_{R-k\rho})}\\ + \|r^{2-\varepsilon}V\|_{L^\infty(B_{R-k\rho})}\|r^{\varepsilon-\gamma+|\beta|}\partial^\beta u\|_{L^p(B_{R-k\rho})}\\ + \|r^{2-\varepsilon+|\beta|}\partial^\beta V\|_{L^\infty(B_{R-k\rho})}\|r^{\varepsilon-\gamma}u\|_{L^p(B_{R-k\rho})}\end{aligned} \tag{7.18}$$

Since $V \in \mathcal{K}^{\varpi,\infty}_{\varepsilon-2}(\Omega)$, and using $\sum_{\zeta:|\zeta|=j}\binom{\beta}{\zeta} = \binom{|\beta|}{j}$,

$$\begin{aligned}\sum_{0<\zeta<\beta}\binom{\beta}{\zeta}\|r^{2-\varepsilon+|\zeta|}\partial^\zeta V\|_{L^\infty(B_{R-k\rho})}\|r^{\varepsilon-\gamma+|\beta|-|\zeta|}\partial^{\beta-\zeta}u\|_{L^p(B_{R-k\rho})}\\ \le C\sum_{0<\zeta<\beta}\binom{\beta}{\zeta}A^{|\beta|}|\zeta|!(k\rho)^{|\zeta|-|\beta|}\left(|\beta|-|\zeta|\right)^{|\beta|-|\zeta|}\\ \le CA^{|\beta|}(k\rho)^{-|\beta|}|\beta|!e^{|\beta|}\sum_{j=1}^{|\beta|-1}\frac{(k\rho/e)^j}{\sqrt{|\beta|-j}}\\ \le A^{|\beta|+1}(k\rho)^{-|\beta|}|\beta|^{|\beta|}\end{aligned}$$

where we have concluded supposing that $A \ge C$ and $k\rho/e \le 1$. The bound on the second to last term in (7.18) is straightforward, while for the last term we note that $-\gamma+\varepsilon > -d/p$ thus $\|r^{\varepsilon-\gamma}u\|_{L^p(\Omega)} \le C$.

We now consider the nonlinear term: in the lemmas above we have shown that, for $\delta = 2,3,4$

$$\|r^{2-\gamma+|\beta|}\partial^\beta u^\delta\|_{L^p(B_{R-k\rho})} \le CA^{|\beta|+\delta\vartheta}\rho^{-|\beta|-\delta\vartheta}\left(\frac{|\beta|}{k}\right)^{|\beta|}|\beta|^{(\delta-1)/2}.$$

In addition, $|\beta| \le C\rho^{-1}$, therefore

$$\|r^{2-\gamma+|\beta|}\partial^\beta u^\delta\|_{L^p(B_{R-k\rho})} \le CA^{|\beta|+\delta\vartheta}\rho^{-|\beta|-\delta\vartheta-(\delta-1)/2}\left(\frac{|\beta|}{k}\right)^{|\beta|}.$$

If (7.15) holds, then $\delta\vartheta + (\delta-1)/2 \le 2$, hence

$$\|r^{2-\gamma+|\beta|}\partial^\beta u^\delta\|_{L^p(B_{R-k\rho})} \le \widetilde{C}A^{|\beta|+2}\rho^{-|\beta|-2}\left(\frac{|\beta|}{k}\right)^{|\beta|},$$

where we have supposed that $A^{2-\delta\vartheta} \ge C/\widetilde{C}$. Note that for all d and δ considered, (7.15) is stronger than the hypothesis $p \ge d(1-1/\delta)$ of Lemma 37. The bound on the term in λu and on the second sum of the right hand side of (7.17) can be obtained straightforwardly from the induction hypothesis.

We have shown that (7.16) holds for all $k \in \mathbb{N}$; furthermore, since $R - k\rho \geq R/2$, we can find a covering of Ω that gives

$$\|r^{|\alpha|-\gamma}\partial^\alpha u\|_{L^p(\Omega)} \leq \widetilde{C}A^{|\alpha|}|\alpha|^{|\alpha|},$$

for $|\alpha| \geq 3$. Thanks to Stirling's formula, this is equivalent to (increasing the constant A in order to absorb the esponential and square root terms)

$$\|r^{|\alpha|-\gamma}\partial^\alpha u\|_{L^p(\Omega)} \leq \widetilde{C}A^{|\alpha|}|\alpha|!,$$

from which we infer (7.8) □

7.4 Hartree-Fock

Consider now the Hartree-Fock equations, which can be rewritten as a nonlinear elliptic system as

$$\begin{aligned} (-\Delta + V)\,\varphi_i + \sum_j \sum_{a<b} c_{ab}^{ij} u_{ab}\varphi_j = \sum_j \lambda_{i,j}\varphi_j \qquad & i = 1, \dots, K \\ -\Delta u_{ab} = 4\pi\varphi_a\varphi_b \qquad & 1 \leq a, b \leq K. \end{aligned} \tag{7.19}$$

with $c_{ab}^{ij} \in \mathbb{R}$. The analyticity of the wave functions away from the positions of the nuclei (i.e., the singularities of V) is classical, see, e.g., [Lew04]. In this setting we consider instead the parts of the domain containing the nuclei, in order to deduce the weighted estimates we obtained for the nonlinear Schrödinger equation. All wieghted spaces can be generalized to $\mathbb{R}^d$ by supposing that the weight r behaves as a the distance from the singularity near it, while it is bounded away from it.

The Hartree-Fock equation has a cubic nonlinearity; nonetheless, we cannot apply directly the results of Section 7.3 because of the nonlocal dependence of u_{ab} on $\varphi_a\varphi_b$ given by the second equation in (7.19). We introduce therefore a Lemma to control the terms depending on u_{kl}. In order to control the jth order derivative in the ball $B_{r-k\rho}$ we consider $j+1$ concentric balls, the first and smallest one being $B_{R-k\rho}$ and the biggest one being B_R. Using (7.1), we can go from the jth derivative of u_{ab} to smaller order derivatives of u_{ab} and $\varphi_a\varphi_b$ in bigger balls. The estimates of $\varphi_a\varphi_b$ are given by Lemma 39 (notice that the domain we consider never gets bigger than B_R), while on u_{ab} we can iterate until we reach a point where we can use global elliptic estimates.

Lemma 42. *Let u_{ab} be the solution in $\mathbb{R}^d$ to*

$$-\Delta u_{ab} = 4\pi\varphi_a\varphi_b \tag{7.20}$$

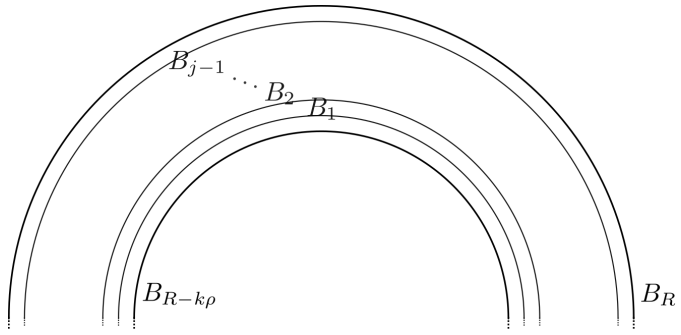


Figure 7.1 – Concentric balls B_i.

where φ_a and φ_b satisfy Assumption 1. Then,

$$\sum_{|\alpha|=j} \|r^{j-\tilde{\gamma}}\partial^\alpha u_{ab}\|_{L^{3p/2}(B_{R-k\rho})} \leq CA^{j+2\vartheta}\rho^{-j-2\vartheta}\left(\frac{j}{k}\right)^j j^{1/2}, \tag{7.21}$$

where $\tilde{\gamma} = 2\frac{2-\gamma}{3}$.

Proof. Suppose $j \geq 3$. Having fixed j and k, we start by considering $j+1$ concentric balls

$$B_i = B_{R-k\frac{j-i}{j}\rho}, \quad i = 0, \ldots, j,$$

see Figure 7.1. Clearly, $B_{R-k\rho} = B_0 \subset B_1 \subset \cdots \subset B_j = B_R$. Now, using Proposition 35 twice and equation (7.20) we find

$$|u_{ab}|_{\mathcal{K}^{j,3p/2}_{\tilde{\gamma}}(B_0)} \leq C|\varphi_a\varphi_b|_{\mathcal{K}^{j-2,3p/2}_{\tilde{\gamma}-2}(B_2)} + C\left(\rho\frac{k}{j}\right)^{-1}|u_{ab}|_{\mathcal{K}^{j-1,3p/2}_{\tilde{\gamma}}(B_1)} + C\left(\rho\frac{k}{j}\right)^{-2}|u_{ab}|_{\mathcal{K}^{j-2,3p/2}_{\tilde{\gamma}}(B_2)}.$$

We now iterate on the last two terms of the above equation, obtaining

$$|u_{ab}|_{\mathcal{K}^{j,3p/2}_{\tilde{\gamma}}(B_0)} \leq C\sum_{i=0}^{j-2} C^{i+1}\left(\rho\frac{k}{j}\right)^{-i}|\varphi_a\varphi_b|_{\mathcal{K}^{j-i-2,3p/2}_{\tilde{\gamma}-2}(B_i)} + C^{j-1}\left(\rho\frac{k}{j}\right)^{-j+1}\|u_{ab}\|_{\mathcal{K}^{1,3p/2}_{\tilde{\gamma}}(B_{j-1})}. \tag{7.22}$$

We consider the first term at the right hand side of the above equation: φ_a and φ_b satisfy the hypothesis of Lemma 39, thus

$$|\varphi_a\varphi_b|_{\mathcal{K}^{j-i-2,3p/2}_{\tilde{\gamma}-2}(B_i)} \leq CA^{j-i-2+2\vartheta}\left(k\frac{j-i}{j}\rho\right)^{-j+i+2}\rho^{-2\vartheta}(j-i-2)^{j-i-2}(j-i-2)^{1/2}.$$

and, supposing $A \geq C$

$$\sum_{i=0}^{j-2} C^{i+1} \left(\rho \frac{k}{j}\right)^{-i} |\varphi_a \varphi_b|_{\mathcal{K}^{j-i-2,3p/2}_{\tilde{\gamma}-2}(B_i)} \leq A^{j+2\vartheta} \rho^{-j+2-2\vartheta} j^{1/2} \sum_{i=0}^{j-2} \left(\frac{j(j-i-2)}{k(j-i)}\right)^{j-i-2}$$
$$\leq A^{j+2\vartheta} \rho^{-j+2-2\vartheta} j^{1/2} \left(\frac{j}{k}\right)^{j}.$$

We still need to bound the second term at the right hand side of (7.22). This is done by considering that, due to elliptic regularity, $\varphi_a \varphi_b \in L^\infty(\mathbb{R}^d)$ so that we can bound $\|u_{ab}\|_{\mathcal{K}^{1,3p/2}_{\tilde{\gamma}}(B_{j-1})}$ by elliptic regularity. In the same way we bound (7.21) when $j \leq 2$ and this concludes the proof. □

Using this last lemma we extend the result obtained on the nonlinear Schrödinger equation to the Hartree-Fock system.

Theorem 8. *Let $\{\varphi_i\}_{i=1}^K$ be the solution to the Hartree-Fock equation* (7.19) *with $V \in \mathcal{K}^{\varpi,\infty}_{\varepsilon-2}(\mathbb{R}^d)$. Then,*

$$\varphi_i \in \mathcal{J}^{\varpi}_{\gamma}(\mathbb{R}^d), \quad i = 1, \dots, K \tag{7.23}$$

for any $\gamma < \min(d/2 + \varepsilon, 2)$.

Proof. The proof follows the same lines as the proof of Theorem 7 with $\delta = 3$. Firstly, we remark that $\varphi_i, \varphi_j \in H^1(\mathbb{R}^d)$ implies $u_{ij} \in W^{2,3}(\mathbb{R}^d)$ via the second equation of (7.19) and that, in turn, this implies $\varphi_i \in \mathcal{J}^{1,p}_{\gamma}(\mathbb{R}^d)$ for all $p > 1$, $\gamma < d/p + \varepsilon$, and $i = 1, \dots, K$. We can then assume that the $\{\varphi_i\}_{i=1}^K$ satisfy the Induction Assumption 1 and we can follow the proof of Theorem 7 with the difference that instead of the estimate of Lemma 40 (in the case $\delta = 3$) we use

$$\|r^{2-\gamma+|\alpha|} \partial^\alpha (u_{ab}\varphi_c)\|_{L^p(B_{R-k\rho})} \leq C A^{|\alpha|+3\vartheta} \rho^{-|\alpha|-3\vartheta} \left(\frac{|\alpha|}{k}\right)^{|\alpha|} |\alpha|.$$

This is proven by replacing Lemma 39 with Lemma 42 in the proof of Lemma 40. The thesis follows. □

Chapter 8

Analysis of the hp discontinuous Galerkin method for elliptic nonlinear eigenvalue problems

In this section we consider the approximation of an elliptic nonlinear eigenvalue problem obtained with an hp discontinuous Galerkin finite element method. Specifically, we consider the problem of looking, in a domain Ω, for an eigenvalue-eigenfunction couple (λ, u) such that $\|u\|_{L^2(\Omega)} = 1$ and

$$(-\Delta + V + f(u^2))u = \lambda u, \tag{8.1}$$

where the hypotheses on V and f will be specified later. The interest in a problem of this kind lies in the fact that it corresponds to the Euler-Lagrange equation of a minimization problem, in which one looks for the minimizer of a nonlinear energy. Such problems are widely present in physics and chemistry, though they often involve nonlocal nonlinearities, while here we only treat the local case. Equations of the form (8.1) are also often referred to as nonlinear Schrödinger equations.

Our analysis is centered mainly on potentials that are singular at a set of isolated this points; this includes the electric attraction generated by a Coulomb potential, i.e., $V(x) = 1/d(x, \mathfrak{c})$ for a fixed point $\mathfrak{c} \in \Omega$, but applies more generally to any potential that, in the vicinity of the singular point, behaves as

$$V(x) \sim \frac{1}{d(x, \mathfrak{c})^\alpha},$$

for an $\alpha < 2$. Clearly, V is not regular in Sobolev spaces, thus we cannot expect the solution to be regular in those spaces either. Nonetheless, we can work in weighted Kondratiev-Babuška spaces, and, if the solution is sufficiently regular, the hp approximation converges exponentially, see [SSW13a]. This is the case treated in Section 8.2.6.

Taking a step back from what has been outlined in the previous paragraph, we stress that, even though our focus is on hp methods, most of the proofs are more general. Suppose we consider a simpler h-type finite element method: the proof of Theorem 9 – i.e., convergence and quasi optimality of the numerical solution – holds, since we do not use any specific feature of hp refinement. The proof of convergence of the discontinuous Galerkin method for a nonlinear eigenvalue problem of the form (8.1) is a new result as far as we are aware. Previous results include the convergence of the discontinuous Galerkin method for linear eigenproblems [ABP06] and the convergence of conforming methods for the nonlinear problem [CCM10]. The latter paper has been a major source of inspiration for the present work. The main difference is that the discontinuous Galerkin method is not conforming, thus some relations between exact and numerical quantities, e.g., between the exact eigenvalue λ and the numerical one λ_δ, are less straightforward. In general, Theorem 9 should be readily extendable to any nonconforming symmetric method such that the thesis of Lemma 44, akin to coercivity and continuity of the numerical bilinear form, holds. The requirement of symmetry in the bilinear form of the numerical method is a strong one, and will be used without explicit mention throughout the proofs. This could be seen as a limit; nonetheless, from a practical point of view, there is little interest in the approximating a symmetric eigenvalue problem with a non symmetric numerical method. Apart from the properties of the method itself, that, even in the linear case, would show a lower order of convergence, the finite dimensional problem would be more problematic, since the solution of a finite dimensional eigenvalue problem can be done more accurately and efficiently for symmetric matrices.

In Section 8.1, we start by introducing the notation of this part and the infinite dimensional functional spaces that will be necessary in the following. Then, we introduce the hp finite element space and the mesh dependent norms associated to it. In Section 8.1.3, we proceed by presenting problem (8.1) in its weak formulation, and by introducing the associated energy minimization problem. The finite element approximation is then easily derived by replacing the continuous Laplacian bilinear form with the classical SIP bilinear form. We introduce our basic assumptions on F, which are approximately the same as those introduced in [CCM10] and on the potential V. As the analysis progresses, we will introduce more restrictive hypotheses; the assumptions we make at the beginning will nonetheless be sufficient to prove convergence and to assure the that solution u lies in $H^{d/2+\alpha}(\Omega) \cap \mathcal{J}^2_{d/2+\beta}(\Omega)$ for some $\alpha, \beta > 0$.

In Section 8.2 we then prove the convergence estimates on the eigenvalue and eigenfunction. The hypotheses get gradually more restrictive as we progress in the Section, and the results get stronger. We start by the hypotheses introduced in the previous section, and show that those are sufficient to prove convergence and quasi optimality for the DG norm of the eigenfunction error. We also show that the eigenvalue converges as fast as the eigenfunction. For the sake of simplicity we do not specify it, but this simple results do not need the full regularity in weighted Sobolev spaces that we have assumed for V. Imposing further hypotheses on the nonlinearity (and, if we had not done it before, on the potential) and assuming convergence for an associated problem, we can then proceed to an improved proof of convergence for the eigenvalues, given in

Section 8.2.5. We conclude by considering the case where the nonlinearity is polynomial, and weighted analyticity can be proven. In this instance, as we have mentioned before, the hp space converges exponentially towards the exact solution, thus we can specialize the previous convergence results.

Section 8.3 is concerned with the introduction of an iterative scheme that deals with the nonlinearity. This scheme is shown, up to extraction, to converge to an eigenstate of the fully nonlinear equation. This results has interesting theoretical consequences: through a transformation of the original problem, we can derive an asymptotic expansion near the singularity for the solution. This is done in Section 8.4, where we are able to fully determine this asymptotic expansion, provided we fix a nonlinear function and a potential. We consider the case of a regular nonlinearity, corresponding to the case where the solution is analytic in weighted spaces. In this context, we can prove higher order convergence estimates, which lead to the determination of the asymptotic expansion of the eigenfunction of the nonlinear problem.

8.1 Statement of the problem and notation

8.1.1 Functional setting

Let $\Omega = (\mathbb{R}/L)^d$ be a periodic d-cube of edge $L < 1$.We use the standard notation for Sobolev spaces $W^{k,p}(\Omega)$, with $W^{k,2}(\Omega) = H^k(\Omega)$ and $W^{0,p}(\Omega) = L^p(\Omega)$. We denote the scalar product in $L^2(\Omega)$ as $(\cdot,\cdot)$ and the norm as $\|u\| = (u,u)$. For a given triangulation $\mathcal{T}$, we denote

$$(u,v)_{\mathcal{T}} = \sum_{K\in\mathcal{T}} (u,v)_K$$

and similarly, for a set of edges $\mathcal{E}$,

$$(u,v)_{\mathcal{E}} = \sum_{e\in\mathcal{E}} (u,v)_e.$$

We now recall the definition of weighted spaces given in Chapter 2. Given a set of isolated points $\mathfrak{C} \subset \Omega$, the homogeneous Kontratiev-Babuška space $\mathcal{K}^{k,p}_\gamma(\Omega,\mathfrak{C})$ is defined as

$$\mathcal{K}^{k,p}_\gamma(\Omega,\mathfrak{C}) = \{u : r^{|\alpha|-\gamma}\partial^\alpha u \in L^p(\Omega)\ \forall \alpha \in \mathbb{N}^d : |\alpha| \le k\},$$

where $r = r(x)$ is a smooth function which is, in the vicinity of every point $\mathfrak{c} \in \mathfrak{C}$, equal to the euclidean distance $d(x,\mathfrak{c})$ from the point. The nonhomogeneous Kondratiev-Babuška space is defined by

$$\mathcal{J}^{k,p}_\gamma(\Omega,\mathfrak{C}) = \{u : r^{k-\gamma}\partial^\alpha u \in L^p(\Omega)\ \forall \alpha \in \mathbb{N}^d : |\alpha| \le k\},$$

We define the associated seminorm as $|u|_{\mathcal{J}_\gamma^{|\alpha|,p}} = |u|_{\mathcal{K}_\gamma^{|\alpha|,p}} = \|r^{|\alpha|-\gamma}\partial^\alpha u\|_{L^p(\Omega)}$. We also introduce the spaces of regular function with weighted analytic type estimates as

$$\mathcal{K}_\gamma^{\varpi,p}(\Omega,\mathfrak{C}) = \{v \in \mathcal{K}_\gamma^{\infty,p}(\Omega,\mathfrak{C}) : |v|_{\mathcal{K}_\gamma^{k,p}} \leq CA^k k!\ \forall k\},$$

and

$$\mathcal{J}_\gamma^{\varpi,p}(\Omega,\mathfrak{C}) = \{v \in \mathcal{J}_\gamma^{\infty,p}(\Omega,\mathfrak{C}) : |v|_{\mathcal{J}_\gamma^{k,p}} \leq CA^k k!\ k > \gamma - d/p\},$$

where $\mathcal{K}_\gamma^{\infty,p} = \bigcap_k \mathcal{K}_\gamma^{k,p}$, $\mathcal{J}_\gamma^{\infty,p}(\Omega)$ defined similarly. To simplify the notation, we will suppose that the there is only one singular point, i.e., $\mathfrak{C} = \{\mathfrak{c}\}$ and omit $\mathfrak{C}$ from the notation of the spaces. Furthermore, we write $\mathcal{K}_\gamma^k(\Omega) = \mathcal{K}_\gamma^{k,2}(\Omega)$, $\mathcal{J}_\gamma^k(\Omega) = \mathcal{J}_\gamma^{k,2}(\Omega)$, $\mathcal{K}_\gamma^{\varpi,2}(\Omega) = \mathcal{K}_\gamma^{\varpi,2}(\Omega)$, and $\mathcal{J}_\gamma^{\varpi,2}(\Omega) = \mathcal{J}_\gamma^{\varpi,2}(\Omega)$. Note that the results obtained in the following can be trivially extended to the case where $\mathfrak{C}$ contains more than one point, as long as $\mathfrak{C}$ is a finite set of isolated points.

Finally, let $X = H^1(\Omega) \cap \mathcal{J}_\gamma^2(\Omega,\mathcal{C})$, for $\gamma \in (d/2, d/2+\varepsilon)$, where $0 < \varepsilon < 1$ will be specified later, namely in hypothesis (8.12b).

8.1.2 Numerical method

In this section we introduce the hp discontinuous Galerkin method. Concerning the design of the hp space, the setting is the one from [GB86d; GB86e]: imagine a situation where the refinement happens around the singular point and, at every refinement step, the innermost elements – i.e., those who have the singular point as one of their corners – are subdivided into elements smaller by a ratio $\sigma \in (0, 1/2)$. Additionally, the refinement step consists in the update of the polynomial degree over all elements, so that is has a linear slope, i.e., it grows linearly from the "central" elements towards the exterior. The requirements we will introduce on the mesh and space mainly impose that they do not deviate too much from this model. In the following we will do so more rigorously.

Let $\mathcal{T}$ be a mesh isotropically and geometrically graded around the points in $\mathfrak{C}$. We assume that the mesh is shape- and contact-regular and we indicate by Ω_j, $j = 1,\dots,\ell$, the set of elements and edges at the same level of refinement.We introduce on this mesh the hp space with refinement ratio σ and linear polynomial slope s, i.e., for an element $K \in \mathcal{T}$ such that $K \in \Omega_j$,

$$h_K \simeq h_j = \sigma^j \text{ and } p_K \simeq p_j = p_0 + \mathfrak{s}(\ell - j),$$

where h_K is the diameter of the element K and p_K is the polynomial order whose role will be specified in (8.2). We suppose that for any $K \in \mathcal{T}$ there exists an affine transformation $\Phi : K \to \hat{K}$ to the d-dimensional cube $\hat{K}$ such that $\Phi(K) = \hat{K}$, and introduce the discrete space

$$X_\delta = \left\{ v_\delta \in L^2(\Omega) : (v_{|_K} \circ \Phi^{-1}) \in \mathbb{Q}_{p_K}(\hat{K})\ \forall K \in \mathcal{T} \right\}, \tag{8.2}$$

where $\mathbb{Q}_p$ is the space of polynomials of maximal degree p in any variable. Let then $\mathcal{E}$ be

the set of the edges (for $d = 2$) or faces ($d = 3$) of the elements in $\mathcal{T}$ and

$$\begin{aligned} \mathtt{h_e} &= \min_{K\in\mathcal{T}:e\cap\partial K\neq\varnothing} h_K \\ \mathtt{p_e} &= \max_{K\in\mathcal{T}:e\cap\partial K\neq\varnothing} p_K. \end{aligned}$$

On an edge/face between two elements $K_\sharp$ and $K_\flat$, i.e., on $e \subset \partial K_\sharp \cap \partial K_\flat$, the average $\{\!\{\cdot\}\!\}$ and jump $[\![\cdot]\!]$ operators for a function $w \in X(\delta)$ are defined by

$$\{\!\{w\}\!\} = \frac{1}{2}\left(w_{|K_\sharp} + w_{|K_\flat}\right), \qquad [\![w]\!] = w_{|K_\sharp}\mathbf{n}_\sharp + w_{|K_\flat}\mathbf{n}_\flat,$$

where $\mathbf{n}_\sharp$ (resp. $\mathbf{n}_\flat$) is the outward normal to the element $K_\sharp$ (resp. $K_\flat$). In the following, for an $S \subset \Omega$, we denote by $(\cdot,\cdot)_S$ the $L^2(S)$ scalar product and by $\|\cdot\|_S$ the $L^2(S)$ norm.

We introduce the mesh dependent norms that will be used in this section. First, for a $v \in X(\delta)$,

$$\|v\|_{\mathrm{DG}}^2 = \sum_{K\in\mathcal{T}} \|v\|_{\mathcal{J}_1^{1,2}(K)}^2 + \sum_{e\in\mathcal{E}} \mathtt{h_e}^{-1}\mathtt{p_e}^2\|[\![v]\!]\|_{L^2(e)}^2. \tag{8.3}$$

Remark that on X, this norm is equivalent to the $\mathcal{J}_1^1(\Omega) = H^1(\Omega)$ norm, since functions in X are continuous. Then, on $X(\delta) = X + X_\delta$ again, we introduce the norm

$$|||u|||_{\mathrm{DG}}^2 = \sum_{K\in\mathcal{T}} \|u\|_{\mathcal{J}_1^{1,2}(K)}^2 + \sum_{e\in\mathcal{E}} \mathtt{p_e}^2\mathtt{h_e}^{-1}\|[\![u]\!]\|_{L^2(e)}^2 + \sum_{e\in\mathcal{E}} \mathtt{p_e}^{-2}\|r^{1/2}\frac{w_d(r)}{w_d(\mathtt{h_e})}\{\!\{\nabla u\}\!\}\|_{L^2(e)}^2 \tag{8.4}$$

where

$$w_d(x) = \begin{cases} 1 + |\log(x)| & \text{if } d = 2 \\ 1 & \text{if } d = 3. \end{cases}$$

This is slightly different from the classical norm introduced for interior penalty discontinuous Galerkin methods, where the last term of (8.4) would be replaced by the sum of $\mathtt{p_e}^{-2}\mathtt{h_e}\|\nabla v\|_e^2$. For a classical isotropically refined mesh, where $r < h_K$ on all elements, the norms we use is smaller, thus all the approximation properties are conserved, while it makes for a tighter bound in the continuity estimate (8.16a). Furthermore, in the framework of weighted Sobolev spaces, it can be dealt with quite naturally, as shown in the following remark.

Remark 9. *Let us introduce the broken space*

$$\mathcal{J}_\gamma^{s,p}(\Omega,\mathcal{T}) = \left\{v : v \in \mathcal{J}_\gamma^{s,p}(K), \forall K \in \mathcal{T}\right\}.$$

Since for $e \in \partial K$

$$\|r^{1-\gamma}\nabla u\|_{L^2(e)} \leq \|u\|_{\mathcal{J}_\gamma^{1,2}(\partial K)} \leq C\|u\|_{\mathcal{J}_\gamma^{2,2}(K)},$$

we remark that if $v \in \mathcal{J}_\gamma^2(\Omega,\mathcal{T})$*, for* $\gamma > 2 - d/2$*, then* $|||v|||_{\mathrm{DG}}$ *is bounded. Since furthermore* $X(\delta) \subset \mathcal{J}_\gamma^2(\Omega,\mathcal{T})$*,* (8.4) *is bounded on* $X(\delta)$*.*

The next remark concerns the equivalency of (8.3) and (8.4) for finite element functions.

Remark 10. *Note that on X_δ and for $d \leq 3$, the two norms* (8.3) *and* (8.4) *are equivalent, since for any $K \in \mathcal{T}$, $r_{|_K} \lesssim h_K$ and thanks to the discrete trace inequality [DE12]*

$$h_e^{(1-d)/p+d/2} \|w_\delta\|_{L^p(e)} \leq C_{d,p} \|w_\delta\|_{L^2(K)}, \tag{8.5}$$

valid for $e \in \partial K$ and for all $w_\delta \in X_\delta$. The constant $C_{d,p}$ is bounded by $\mathtt{p_e}^2$ if $p = 2$.

8.1.3 Statement of the problem

In this section, we introduce the problem under consideration. From the "physical" point of view, it consists in a minimization of an energy consisting in a kinetic energy term, an interaction with a singular potential V and a nonlinear self-interaction term. The minimization is constrained by fixing the norm of the minimizer; taking the Lagrangian, the energy minimization problem translates into a nonlinear elliptic eigenvalue problem. This is the form under which most of the analysis will be carried out.

Apart from introducing the exact solution u and the solution to the nonlinear given by the numerical method u_δ, we will also introduce the solution given by the numerical method when one "freezes" the solution u in the operator A^u. Note that while a conforming method (e.g., continuous finite elements) would always give a numerical eigenvalue bounded from below by the exact one, this is not the case for the class of nonconforming methods, to which the discontinuous Galerkin finite element method belongs. The usefulness of the eigenvalue λ_δ^* introduced in the following stems therefore from the fact that it bounds from below the quadratic form induced by A_δ^u over $X_\delta \times X_\delta$, i.e., for all $v_\delta \in X_\delta$ such that $\|v_\delta\| = 1$

$$\langle A_\delta^u v_\delta, v_\delta \rangle \geq \lambda_\delta^*.$$

We start therefore by introducing the bilinear form over $X \times X$

$$a(u,v) = (\nabla u, \nabla v)_\Omega + \int_\Omega Vuv \tag{8.6}$$

and the bilinear form over $X_\delta \times X_\delta$

$$\begin{aligned} a_\delta(u_\delta, v_\delta) = (\nabla u_\delta, \nabla v_\delta)_\mathcal{T} - (\{\!\{\nabla u_\delta\}\!\}, [\![v_\delta]\!])_{\mathcal{E}_I} - (\{\!\{\nabla v_\delta\}\!\}, [\![u_\delta]\!])_\mathcal{E} \\ + \sum_{e \in \mathcal{E}} \alpha_e \frac{\mathtt{p_e}^2}{\mathtt{h_e}} ([\![u_\delta]\!], [\![v_\delta]\!])_e + \int_\Omega V u_\delta v_\delta. \end{aligned}$$

Remark 11. *By proving continuity as in Lemma 44 (see the part of the proof referring to inequality* (8.16a)*) and thanks to Remark 9, it can be shown that this form can be extended over $X(\delta) \times X_\delta$.*

We introduce a function $F : \mathbb{R}^+ \to \mathbb{R}$, whose properties will be listed later in this

section for the sake of clarity. Let

$$E(v) = \frac{1}{2}a(v,v) + \frac{1}{2}\int_\Omega F(v^2) \tag{8.7}$$

and

$$E_\delta(v_\delta) = \frac{1}{2}a_\delta(v_\delta, v_\delta) + \frac{1}{2}\int_\Omega F(v_\delta^2). \tag{8.8}$$

Suppose u is the unique minimizer of (8.7) over the space $\{v \in X : \|v\| = 1\}$: then, for $\lambda \in \mathbb{R}$, u is the solution of

$$_{X'}\langle A^u u - \lambda u, v\rangle_X = 0 \quad \forall v \in X \tag{8.9}$$

where

$$_{X'}\langle A^u v, w\rangle_X = a(u,v) + \int_\Omega f(u^2)vw,$$

with $f = F'$. Similarly, let u_δ be a minimizer of (8.8). Then, for an eigenvalue $\lambda_\delta \in \mathbb{R}$ we have

$$\langle A_\delta^{u_\delta} u_\delta - \lambda_\delta u_\delta, v_\delta\rangle = 0 \quad \forall v_\delta \in X_\delta \tag{8.10}$$

where

$$\langle A_\delta^{u_\delta} v_\delta, w_\delta\rangle = a_\delta(v_\delta, w_\delta) + \int_\Omega f(u_\delta^2)v_\delta w_\delta.$$

We introduce also

$$\langle E''(u)v, w\rangle = \langle A^u v, w\rangle + 2\int_\Omega f'(u^2)u^2 vw,$$

with the δ-version defined on $X_\delta \times X_\delta$ and obtained by replacing A^u with A_δ^u. The properties of the function F will be similar to those in [CCM10], namely we suppose that

$$F \in C^1([0,+\infty),\mathbb{R}) \cap C^\infty((0,+\infty),\mathbb{R}) \text{ and } F'' > 0 \text{ in } (0,+\infty), \tag{8.11a}$$
$$\exists q \in [0,2), \exists C \in \mathbb{R} : \forall t \geq 0, |F'(t)| \leq C(1+t^q), \tag{8.11b}$$
$$F''(t)t \text{ locally bounded in } [0,+\infty), \tag{8.11c}$$

and we suppose that $\forall R > 0, \exists C_R \in \mathbb{R}_+ : \forall t_1 \in (0,R], \forall t_2 \in \mathbb{R}$,

$$|F'(t_2^2)t_2 - F'(t_1^2)t_2 - 2F''(t_1^2)(t_1^2)(t_2 - t_1)| \leq C_R(1+|t_2|^s)|t_2 - t_1|^r \tag{8.11d}$$

for $r \in (1,2]$ and $s \in [0, 5-r)$. We will impose additional conditions on F in order to obtain some improved convergence estimates: those conditions will be specified when necessary. Finally, we suppose that the potential V is such that

$$V \in L^{p_V}(\Omega) \tag{8.12a}$$

with $p_V > \max(1, d/2)$ and that there exists $0 < \varepsilon < 1$ such that

$$V \in \mathcal{K}^{\varpi,\infty}_{-2+\varepsilon}(\Omega, \mathfrak{C}). \tag{8.12b}$$

For $d = 2, 3$, (8.12b) implies (8.12a) as long as $p_V < d/(2-\varepsilon)$. A consequence of (8.12a) is, in particular, that for $u, v \in H^1(\Omega)$,

$$(Vu, v)_\Omega \leq C\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}.$$

where the constant C depends on V and on the domain. We have also the following regularity result, which follows from (8.11b) and (8.12b).

Lemma 43. *The solution u to* (8.9) *belongs to the space*

$$u \in \mathcal{J}^2_{d/2+\alpha}(\Omega) \tag{8.13}$$

for any $0 < \alpha < \varepsilon$.

We conclude this section by introducing the discrete approximation to the solution of the linear problem, i.e. the function $u^*_\delta \in X_\delta$ such that

$$\langle A^u_\delta u^*_\delta - \lambda^*_\delta u^*_\delta, v_\delta\rangle = 0 \quad \forall v_\delta \in X_\delta \tag{8.14}$$

for an eigenvalue λ^*_δ. Note that, since u is an eigenfunction of A^u and the associated eigenspace is of dimension 1 [CCM10], we have that

$$\begin{aligned} \|u^*_\delta - u\|_{\mathrm{DG}} &\lesssim \inf_{v_\delta \in X_\delta} |||u - v_\delta|||_{\mathrm{DG}}\,, \\ |\lambda^*_\delta - \lambda| &\lesssim \inf_{v_\delta \in X_\delta} |||u - v_\delta|||^2_{\mathrm{DG}}\,, \end{aligned} \tag{8.15}$$

and the eigenspace associated with u^*_δ is of dimension one, for a sufficient number of degrees of freedom [ABP06].

8.2 A priori estimates

In this section we prove some a priori estimates on the convergence of the numerical eigenfunction and eigenvalue. We start by giving some continuity and coercivity estimates, then we provide with an auxiliary estimate on a scalar product where we construct an adjoint problem, and we conclude by proving convergence and quasi optimality for the eigenfunctions. The rate of convergence proven for the eigenvalues is smaller than what is obtained in the linear case: in the following it will be shown that under some additional hypothesis we can recover the rate typically obtained in the linear case.

Since our main focus here is on isotropically refined hp methods, the approach we take uses the assumption that finite element space and the underlying mesh are those of an hp discontinuous Galerkin method, as described in the previous sections. It is important to remark, nonetheless, that the results of this section can be extended, with

minimal effort, to the analysis of a general discontinuous Galerkin approximation. The novelty of the approach we use in this section lies, indeed, more into the treatment of the nonconformity of the method than in the aspects related to the hp space. The modification that would be necessary to get a proof that applies to a classical h-type discontinuous Galerkin finite element method, for example, would be related to the continuity and coercivity estimates, since those would need not to use the hypothesis that $r \simeq h$.

For the aforementioned reason, and for the sake of generality, we prove our results for an F as general as possible, even though the hp method shows its full power (i.e., exponential rate of convergence) only in a less general setting.

To conclude, we mention the fact that we will mainly write our proofs so that they work for $d = 3$, even though this sometimes means using a suboptimal strategy for the case $d \leq 2$. Consider for example the bound

$$\|v\|_{L^p(\Omega)} \leq C\|v\|_{H^1(\Omega)},$$

for a $v \in H^1(\Omega)$: we will always impose $p \leq 6$, even if for $d = 2$ any $1 \leq p < \infty$ would be acceptable.

8.2.1 Continuity and coercivity

We start with an auxiliary lemma, where we prove the continuity, positivity and coercivity of some operators. As mentioned before, we use the numerical eigenvalue λ_δ^* obtained from the numerical approximation of the linear problem as a lower bound of the operators over the discrete space X_δ.

Lemma 44. *Given the definition of the operators A_δ^u and $E_\delta''(u)$, of the spaces X_δ and $X(\delta)$, and of λ_δ^* provided in Section 8.1, the following results hold*

$$|\langle (A_\delta^u - \lambda_\delta^*)\, v, v_\delta\rangle| \lesssim |||v|||_{\mathrm{DG}}\, \|v_\delta\|_{\mathrm{DG}} \qquad \forall v \in X(\delta), v_\delta \in X_\delta \tag{8.16a}$$

$$\langle (A_\delta^u - \lambda_\delta^*)\, v_\delta, v_\delta\rangle \geq 0 \qquad \forall v_\delta \in X_\delta. \tag{8.16b}$$

Furthermore,

$$\langle (A_\delta^u - \lambda_\delta^*)\, (u_\delta - u_\delta^*), (u_\delta - u_\delta^*)\rangle \gtrsim \|u_\delta - u_\delta^*\|_{\mathrm{DG}}^2 \tag{8.17}$$

and

$$\langle \left(E_\delta''(u) - \lambda_\delta^*\right) v_\delta, v_\delta\rangle \gtrsim \|v_\delta\|_{\mathrm{DG}}^2 \qquad \forall v_\delta \in X_\delta \tag{8.18a}$$

$$|\langle \left(E_\delta''(u) - \lambda_\delta^*\right) v, v_\delta\rangle| \lesssim |||v|||_{\mathrm{DG}}\, \|v_\delta\|_{\mathrm{DG}} \qquad \forall v \in X(\delta), v_\delta \in X_\delta. \tag{8.18b}$$

Proof. We now turn our attention to the continuity inequality (8.16a). Consider a function $v \in X(\delta)$. We can decompose $v = \tilde{v} + \tilde{v}_\delta$, where $\tilde{v} \in C^{0,\alpha}(\Omega)$ for any $\alpha < \varepsilon$ and $\tilde{v}_\delta \in X_\delta$. Consider an edge/face $e \in \mathcal{E}$. Then, $[\![v]\!]_{|e} = [\![\tilde{v}_\delta]\!]_{|e}$. If $\mathfrak{C} \cap \bar{e} = \varnothing$, then $\mathtt{h_e} \simeq r$; if instead there exists a $\mathfrak{c} \in \mathfrak{C}$ such that $\mathfrak{c}$ is one of the vertices of e, then $[\![\tilde{v}_\delta]\!]_{|e} \in \mathbb{Q}_{p_0}(e)$, which is a

finite dimensional space of fixed size. Therefore on $X(\delta)$ we have the equivalency

$$\mathtt{h_e}^{-1/2}(1+|\log(\mathtt{h_e})|)^{-1}\|[\![\cdot]\!]\|_{L^2(e)} \simeq \|(1+|\log(r)|)^{-1}r^{-1/2}[\![\cdot]\!]\|_{L^2(e)} \tag{8.19}$$

if $d=2$ and

$$\mathtt{h_e}^{-1}\|[\![\cdot]\!]\|^2_{L^2(e)} \simeq r^{-1/2}\|[\![\cdot]\!]\|^2_{L^2(e)} \tag{8.20}$$

if $d=3$. The continuity estimate (8.16a) can be obtained through multiple applications of Hölder's inequality: we consider the terms in the bilinear form separately. First, on the broken space $H^1(\mathcal{T}) := \{v : v_{|_K} \in H^1(K),\, \forall K \in \mathcal{T}\}$ we exploit the fact that, as shown in [LS03],

$$\|v\|_{L^q(\Omega)} \lesssim \|v\|_{\mathrm{DG}} \quad \forall v \in H^1(\mathcal{T}) \tag{8.21}$$

with $q \le 2d/(d-2)$ if $d \ge 3$ and $q \in [1,\infty)$ if $d=2$. Note that $X(\delta) \subset H^1(\Omega,\mathcal{T})$, thus

$$|a(v,v_\delta)| \lesssim \|v\|_{\mathrm{DG}}\|v_\delta\|_{\mathrm{DG}}$$

Secondly,

$$\begin{aligned}
\left|\sum_e (\{\!\{\nabla v\}\!\},[\![v_\delta]\!])_e\right| &\lesssim \sum_e \mathtt{p_e}^{-1}\|r^{1/2}w_d(r)\{\!\{\nabla v\}\!\}\|_{L^2(e)}\mathtt{p_e}\|r^{-1/2}w_d(r)^{-1}[\![v_\delta]\!]\|_{L^2(e)} \\
&\lesssim \sum_e \mathtt{p_e}^{-1}w_d(\mathtt{h_e})^{-1}\|r^{1/2}w_d(r)\{\!\{\nabla v\}\!\}\|_{L^2(e)}\mathtt{p_e}\mathtt{h_e}^{-1/2}\|[\![v_\delta]\!]\|_{L^2(e)} \\
&\lesssim \left(\sum_e \mathtt{p_e}^{-2}\|r^{1/2}\frac{w_d(r)}{w_d(\mathtt{h_e})}\{\!\{\nabla v\}\!\}\|^2_{L^2(e)}\right)^{1/2}\left(\sum_e \mathtt{p_e}^2\mathtt{h_e}^{-1}\|[\![v_\delta]\!]\|^2_{L^2(e)}\right)^{1/2}
\end{aligned}$$

where the second inequality follows from (8.19) and (8.20). Similarly,

$$\begin{aligned}
\left|\sum_e (\{\!\{\nabla v_\delta\}\!\},[\![v]\!])_e\right| &\lesssim \left(\sum_e \mathtt{p_e}^{-2}\mathtt{h_e}\|\{\!\{\nabla v_\delta\}\!\}\|^2_{L^2(e)}\right)^{1/2}\left(\sum_e \mathtt{p_e}^2\mathtt{h_e}^{-1}\|[\![v]\!]\|^2_{L^2(e)}\right)^{1/2} \\
&\lesssim \left(\sum_K \|\nabla v_\delta\|^2_{L^2(K)}\right)^{1/2}\left(\sum_e \mathtt{p_e}^2\mathtt{h_e}^{-1}\|[\![v]\!]\|^2_{L^2(e)}\right)^{1/2},
\end{aligned}$$

using (10) in the second line. Then,

$$\begin{aligned}
\left|\sum_{e\in\mathcal{E}} \alpha_e \frac{\mathtt{p_e}^2}{\mathtt{h_e}}([\![v]\!],[\![v_\delta]\!])_e\right| &\lesssim C\sum_e \left(\mathtt{p_e}\mathtt{h_e}^{-1/2}\|[\![u]\!]\|_{L^2(e)}\right)\left(\mathtt{p_e}\mathtt{h_e}^{-1/2}\|[\![v_\delta]\!]\|_{L^2(e)}\right) \\
&\lesssim C\left(\sum_e \mathtt{p_e}^2\mathtt{h_e}^{-1}\|[\![v]\!]\|^2_{L^2(e)}\right)^{1/2}\left(\sum_e \mathtt{p_e}^2\mathtt{h_e}^{-1}\|[\![v_\delta]\!]\|^2_{L^2(e)}\right)^{1/2}.
\end{aligned}$$

Thanks to the Hölder inequality, Sobolev embeddings, hypothesis (8.11b), and (8.21),

$$\begin{aligned}\left|\int_\Omega f(u^2)vv_\delta\right| &\lesssim \|1+u^{2q}\|_{L^{3/2}(\Omega)}\|v\|_{L^6(\Omega)}\|v_\delta\|_{L^6(\Omega)}\\ &\lesssim \|u\|^{2q}_{H^1(\Omega)}\|v\|_{H^1(\Omega)}\|v_\delta\|_{\mathrm{DG}}\\ &\lesssim \|v\|_{H^1(\Omega)}\|v_\delta\|_{\mathrm{DG}}.\end{aligned}$$

Since then $\lambda^*_\delta \to \lambda$, we have that $|\lambda^*_\delta(v,v_\delta)| \le C\|v\|\|v_\delta\|$ and this, combined with the above inequalities, proves (8.16a).

We now consider (8.16b). As already stated, λ^*_δ is a simple eigenvalue for a sufficient number of degrees of freedom and therefore $A^u_\delta - \lambda^*_\delta$ is coercive on the subspace of X_δ L^2-orthogonal to u^*_δ. Hence, since $\|u^*_\delta\| = 1$ and A^u_δ is symmetric,

$$\begin{aligned}\langle (A^u_\delta - \lambda^*_\delta)\, v_\delta, v_\delta\rangle &= \langle (A^u_\delta - \lambda^*_\delta)\,(v_\delta - (v_\delta,u^*_\delta)_\Omega u^*_\delta), v_\delta - (v_\delta,u^*_\delta)_\Omega u^*_\delta\rangle\\ &\gtrsim \|v_\delta\|^2 - (u^*_\delta,v_\delta)^2 \ge 0,\end{aligned} \tag{8.22}$$

for all $v_\delta \in X_\delta$. We may then prove (8.17) following the same reasoning as in [CCM10]. We recall it here for ease of reading. From the above inequality we have (recall that $\|u^*_\delta\| = \|u_\delta\| = 1$)

$$\begin{aligned}\langle (A^u_\delta - \lambda^*_\delta)\,(u_\delta - u^*_\delta), (u_\delta - u^*_\delta)\rangle &\gtrsim \|u_\delta - u^*_\delta\|^2 - (u^*_\delta, u_\delta - u^*_\delta)^2\\ &= \|u_\delta - u^*_\delta\|^2 - (1 + (u^*_\delta,u_\delta)^2 - 2(u^*_\delta,u_\delta))\\ &= 1 - (u^*_\delta,u_\delta)^2\\ &\ge \frac{1}{2}\|u_\delta - u^*_\delta\|^2,\end{aligned} \tag{8.23}$$

and this proves (8.17). To prove (8.18a), we note that

$$\langle \big(E''_\delta(u) - \lambda^*_\delta\big)\, v_\delta, v_\delta\rangle \ge \langle (A^u_\delta - \lambda^*_\delta)\, v_\delta, v_\delta\rangle + \int_\Omega f'(u^2)u^2v_\delta^2. \tag{8.24}$$

Suppose we negate (8.18a): then, there has to be a series $\{v^j_\delta\}_j \subset X_\delta$ such that $\|v^j_\delta\| = 1$ and $\langle (E''_\delta(u) - \lambda^*_\delta)\, v^j_\delta, v^j_\delta\rangle \to 0$. Since $\int_\Omega f'(u^2)u^2(v^j_\delta)^2 > 0$, from (8.22) we have that

$$\begin{aligned}\frac{1}{2}\|v^j_\delta - u^*_\delta\|^2 &= \|v^j_\delta\| - (v^j_\delta,u^*_\delta)^2\\ &\lesssim \langle (E''(u) - \lambda^*_\delta)v^j_\delta, v^j_\delta\rangle,\end{aligned}$$

thus, $v^j_\delta \to u^*_\delta$ in $L^2(\Omega)$. Now, since u^*_δ converges towards u in the DG norm, and using (8.11c) and the positivity of f', we can show that there exists an $\alpha > 0$ such that, for a sufficient number of degrees of freedom,

$$\int_\Omega f'(u^2)u^2(u^*_\delta)^2 > \alpha.$$

This negates the contradiction hypothesis that $\langle (E''_\delta(u) - \lambda^*_\delta)\, v^j_\delta, v^j_\delta\rangle \to 0$, hence

$$\langle \big(E''_\delta(u) - \lambda^*_\delta\big)\, v_\delta, v_\delta\rangle \geq C\|v_\delta\|^2 \tag{8.25}$$

for all $v_\delta \in X_\delta$. Then, using the classical result that

$$(\nabla v_\delta, \nabla v_\delta)_{\mathcal{T}} - (\{\!\{\nabla v_\delta\}\!\}, [\![v_\delta]\!])_{\mathcal{E}_I} - (\{\!\{\nabla v_\delta\}\!\}, [\![v_\delta]\!])_{\mathcal{E}} + \sum_{e\in\mathcal{E}} \alpha_e \frac{\mathtt{p_e}^2}{\mathtt{h_e}} ([\![v_\delta]\!], [\![v_\delta]\!])_e \geq \|v_\delta\|^2_{\mathrm{DG}},$$

combined with the estimate from the proof of [CCM10, Lemma 1], we can show that

$$\langle (A^u_\delta - \lambda^*_\delta)\, v_\delta, v_\delta\rangle \geq \alpha\|v_\delta\|^2_{\mathrm{DG}} - C\|v_\delta\|^2. \tag{8.26}$$

The coercivity estimate (8.18a) then follows from (8.25) and (8.26).

Finally, (8.18b) follows directly from the definition of $E''_\delta(u)$, the continuity estimate (8.16b) and the fact that $|f'(u^2)u^2| \leq C$. □

8.2.2 Estimates on the adjoint problem

In this section we develop an estimate on the scalar product between a function and the error $u - u_\delta$, whose interest lies mainly in the $L^2(\Omega)$ convergence estimate given in Theorem 9. The estimate is based on the introduction of the adjoint problem (8.27).

Lemma 45. *Let $u^{\star\perp}_\delta = \{v_\delta \in X_\delta : (v_\delta, u^*_\delta) = 0\}$ be the space of functions $L^2(\Omega)$-orthogonal to u^*_δ and let ψ_{w_δ} be the solution to the problem*

$$\begin{aligned} &\textit{find } \psi_{w_\delta} \in u^{\star\perp}_\delta \textit{ such that} \\ &\langle \big(E''_\delta(u) - \lambda^*_\delta\big)\, \psi_{w_\delta}, v_\delta\rangle = \langle w_\delta, v_\delta\rangle, \forall v_\delta \in u^{\star\perp}_\delta \end{aligned} \tag{8.27}$$

Then, if hypotheses (8.11a) *to* (8.11d) *hold,*

$$\begin{aligned} |\langle w, u_\delta - u^*_\delta\rangle| \lesssim \|u_\delta - u^*_\delta\|^r_{L^{6r/(5-s)}}\|\psi_{w_\delta}\|_{\mathrm{DG}} + |\lambda_\delta - \lambda^*_\delta|\|u_\delta - u^*_\delta\|\|\psi_{w_\delta}\| + \|u - u^*_\delta\|\|\psi_{w_\delta}\| \\ + \|u_\delta - u^*_\delta\|^2\|\psi_{w_\delta}\| + \|u_\delta - u^*_\delta\|^2\|w_\delta\|, \end{aligned} \tag{8.28}$$

Proof. We break $u_\delta - u^*_\delta$ into two parts, one parallel to u^*_δ and one perpendicular to it. Those are given respectively by

$$(u_\delta - u^*_\delta, u^*_\delta)u^*_\delta = -\frac{1}{2}\|u_\delta - u^*_\delta\|^2 u^*_\delta \quad \text{and} \quad u_\delta - (u_\delta, u^*_\delta)u^*_\delta \in u^{\star\perp}_\delta.$$

Then,

$$\begin{aligned}\langle w_\delta, u_\delta - u_\delta^*\rangle &= (w_\delta, u_\delta - (u_\delta, u_\delta^*)u_\delta^*) - \frac{1}{2}\|u_\delta - u_\delta^*\|^2(w_\delta, u_\delta^*)\\ &= \langle \left(E_\delta''(u) - \lambda_\delta^*\right)\psi_{w_\delta}, u_\delta - (u_\delta, u_\delta^*)u_\delta^*\rangle - \frac{1}{2}\|u_\delta - u_\delta^*\|^2(w_\delta, u_\delta^*)\\ &= \langle \left(E_\delta''(u) - \lambda_\delta^*\right)(u_\delta - u_\delta^*), \psi_{w_\delta}\rangle - \frac{1}{2}\|u_\delta - u_\delta^*\|^2\langle \left(E_\delta''(u) - \lambda\right)u_\delta^*, \psi_{w_\delta}\rangle\\ &\quad - \frac{1}{2}\|u_\delta - u_\delta^*\|^2(w_\delta, u_\delta^*)\\ &= \langle \left(E_\delta''(u) - \lambda_\delta^*\right)(u_\delta - u_\delta^*), \psi_{w_\delta}\rangle - \|u_\delta - u_\delta^*\|^2\int_\Omega f'(u^2)u^2u_\delta^*\psi_{w_\delta}\\ &\quad - \frac{1}{2}\|u_\delta - u\|^2(w_\delta, u_\delta^*).\end{aligned} \tag{8.29}$$

We consider the first term:

$$\begin{aligned}\langle \left(E_\delta''(u) - \lambda_\delta^*\right)(u_\delta - u_\delta^*), \psi_{w_\delta}\rangle &= \langle (A_\delta^u - \lambda_\delta^*)u_\delta, \psi_{w_\delta}\rangle + 2\int_\Omega f'(u^2)u^2\psi_{w_\delta}(u_\delta - u_\delta^*)\\ &= -\int_\Omega \left[f({u_\delta}^2)u_\delta - f(u^2)u_\delta - 2f'(u^2)u^2(u_\delta - u)\right]\psi_{w_\delta}\\ &\quad + (\lambda_\delta - \lambda_\delta^*)(u_\delta - u_\delta^*, \psi_{w_\delta})\\ &\quad + 2\int_\Omega f'(u^2)u^2\psi_{w_\delta}(u - u_\delta^*).\end{aligned} \tag{8.30}$$

Thanks to (8.11d), combining (8.29) and (8.30) we can infer that

$$\begin{aligned}|\langle w_\delta, u_\delta - u_\delta^*\rangle| \lesssim \|u_\delta - u_\delta^*\|_{L^{6r/(5-s)}}^r\|\psi_{w_\delta}\|_{\mathrm{DG}} + |\lambda_\delta - \lambda_\delta^*|\|u_\delta - u_\delta^*\|\|\psi_{w_\delta}\| + \|u - u_\delta^*\|\|\psi_{w_\delta}\|\\ + \|u_\delta - u_\delta^*\|^2\int_\Omega \left|f'(u^2)u^2u_\delta^*\psi_{w_\delta}\right| + \|u_\delta - u_\delta^*\|^2\left|(w_\delta, u_\delta^*)\right|,\end{aligned}$$

which gives the thesis. □

8.2.3 Basic convergence

At this stage, we are able to prove the first convergence result for the numerical eigenfunction and eigenvalue. We work mainly in the discrete setting, in order to avoid the issues due to the nonconformity of the method. The analysis is carried out for the symmetric interior penalty discontinuous Galerkin method, but it holds for any nonconforming symmetric method, as long as the results of Lemma 44 hold for such a method. Furthermore, the remark made at the beginning of Section 8.2 still holds, in that the result can be adapted with few modifications to a classical h-type discontinuous Galerkin finite element method.

In general, the goal is to prove that the numerical eigenvalue-eigenfunction couple obtained as solution to the nonlinear problem converges as fast as in the linear case. In

this section, we obtain this result for the eigenfunction, which is shown to converge quasi optimally. The hypotheses on the function F are instead not strong enough to prove that the eigenvalue converges twice as fast as the eigenfunction in the $\| \ \|_{\mathrm{DG}}$. We can nonetheless prove that the eigenvalue converges at least as fast as the eigenfunction; the doubling of the rate of convergence is deferred to the later Theorem 11, where we will have introduced additional hypotheses on F.

The following theorem gives then the above mentioned estimates on the convergence of the eigenfunction and eigenvalue. We start by showing the convergence to zero of the error, and use this result to show that the estimate is quasi optimal. We then show that the eigenvalue convergence, with the basic rate mentioned above, and conclude by showing an estimate on the $L^2(\Omega)$ norm of the error.

Theorem 9. *If the hypotheses* (8.11a) *to* (8.11d) *on F hold and the hypotheses on the potential V* (8.12a), (8.12b) *hold, then*

$$\|u - u_\delta\|_{\mathrm{DG}} \to 0. \tag{8.31}$$

In particular, we have the quasi-optimal convergence

$$\|u - u_\delta\|_{\mathrm{DG}} \lesssim \inf_{v_\delta \in X_\delta} |||u - v_\delta|||_{\mathrm{DG}}\,. \tag{8.32}$$

Furthermore,

$$|\lambda - \lambda_\delta| \lesssim \inf_{v_\delta \in X_\delta} |||u - v_\delta|||_{\mathrm{DG}} \tag{8.33}$$

and

$$\|u - u_\delta\| \lesssim \|u - u_\delta^*\|^r_{L^{6r/(5-s)}} + \|u - u_\delta\|^r_{L^{6r/(5-s)}} + \|u - u_\delta^*\|. \tag{8.34}$$

where r is defined in (8.11d) *and u_δ^* is the solution of the linear eigenvalue problem defined in* (8.14).

Proof. We start by proving (8.31), i.e. the convergence of the numerical solution towards the exact one. We have

$$\begin{aligned}
2\left(E_\delta(u_\delta) - E(u)\right) &= \langle A_\delta^u u_\delta, u_\delta\rangle - \langle A^u u, u\rangle + \int_\Omega \left(F(u_\delta^2) - F(u^2) - f(u^2)(u_\delta^2 - u^2)\right) \\
&= \langle (A_\delta^u - \lambda_\delta^*)\,(u_\delta - u_\delta^*), u_\delta - u_\delta^*\rangle - \lambda + \lambda_\delta^* \\
&\quad + \int_\Omega \left(F(u_\delta^2) - F(u^2) - f(u^2)(u_\delta^2 - u^2)\right) \\
&\gtrsim \|u_\delta - u_\delta^*\|^2_{\mathrm{DG}} - |\lambda - \lambda_\delta^*| + \int_\Omega \left(F(u_\delta^2) - F(u^2) - f(u^2)(u_\delta^2 - u^2)\right).
\end{aligned}$$

Therefore, exploiting the convexity of F and the convergence of λ towards λ_δ^*, we have that

$$\begin{aligned}
\|u_\delta - u_\delta^*\|^2_{\mathrm{DG}} &\lesssim E_\delta(u_\delta) - E(u) + |\lambda - \lambda_\delta^*| \\
&\le E_\delta(\Pi_\delta u) - E_\delta(u) + |\lambda - \lambda_\delta^*| \to 0.
\end{aligned} \tag{8.35}$$

Considering that u_δ^* converges towards u in the DG norm, (8.35) implies (8.31). Note

then that

$$\begin{aligned}\lambda_\delta - \lambda_\delta^* &= \langle A_\delta^u u_\delta, u_\delta\rangle - \lambda_\delta^* + \int_\Omega \left[f(u_\delta^2) - f(u^2)\right] u_\delta^2 \\ &= \langle (A_\delta^u - \lambda_\delta^*)\,(u_\delta - u_\delta^*)\,, u_\delta - u_\delta^*\rangle + \int_\Omega \left[f(u_\delta^2) - f(u^2)\right] u_\delta^2.\end{aligned} \tag{8.36}$$

Remarking, as in [CCM10, Proof of Theorem 1], that

$$\int_\Omega \left[f(u_\delta^2) - f(u^2)\right] u_\delta^2 \le \|1 + {u_\delta}^{2q+1}\|_{L^{6/(2q+1)}(\Omega)} \|u - u_\delta\|_{\mathrm{DG}}$$

and using (8.31) we can conclude that

$$|\lambda - \lambda_\delta| \lesssim |\lambda - \lambda_\delta^*| + \|u_\delta - u_\delta^*\|_{\mathrm{DG}} + \|u - u_\delta\|_{\mathrm{DG}}. \tag{8.37}$$

Now, from (8.18a) we have

$$\begin{aligned}\|u_\delta - u_\delta^*\|_{\mathrm{DG}}^2 &\lesssim \langle \left(E_\delta''(u) - \lambda_\delta^*\right)(u_\delta - u_\delta^*), u_\delta - u_\delta^*\rangle \\ &= \langle (A_\delta^u - \lambda_\delta^*)\,(u_\delta - u_\delta^*), u_\delta - u_\delta^*\rangle + 2\int_\Omega f'(u^2)u^2(u_\delta - u_\delta^*)^2 \\ &= \langle (A_\delta^u - \lambda_\delta)\,u_\delta, u_\delta - u_\delta^*\rangle + (\lambda_\delta - \lambda_\delta^*)\|u_\delta - u_\delta^*\|^2 + 2\int_\Omega f'(u^2)u^2(u_\delta - u_\delta^*)^2 \\ &= \int_\Omega \left[\left(f(u^2) - f({u_\delta}^2)\right) u_\delta + 2f'(u^2)u^2(u_\delta - u_\delta^*)\right](u_\delta - u_\delta^*) + (\lambda_\delta - \lambda_\delta^*)\|u_\delta - u_\delta^*\|^2.\end{aligned}$$

Consider the first term: hypothesis (8.11c) gives

$$\int_\Omega f'(u^2)u^2(u_\delta - u_\delta^*)^2 \lesssim \int_\Omega f'(u^2)u^2(u_\delta - u)(u_\delta - u_\delta^*) + \|u - u_\delta^*\|\|u_\delta - u_\delta^*\|.$$

The two above equations and (8.11d) thus give

$$\|u_\delta - u_\delta^*\|_{\mathrm{DG}}^2 \lesssim \|1 + |u_\delta|^s\|_{L^{6/s}(\Omega)} \|u_\delta - u\|_{L^{6r/(5-s)}(\Omega)}^r \|u_\delta - u_\delta^*\|_{\mathrm{DG}} + |\lambda_\delta - \lambda_\delta^*|\|u_\delta - u_\delta^*\|^2 + \|u - u_\delta^*\|\|u_\delta - u_\delta^*\|$$

and, since $r > 1$ and $6r/(5-s) \le 6$, we can conclude that

$$\|u - u_\delta\|_{\mathrm{DG}} \lesssim \|u - u_\delta^*\|_{\mathrm{DG}}.$$

The quasi optimality of u_δ^* then implies (8.32). Additionally, we can use this estimate in (8.37) and, considering that

$$|\lambda - \lambda_\delta^*| \lesssim |||u - u_\delta^*|||_{\mathrm{DG}}^2 \lesssim \inf_{v_\delta \in X_\delta} |||u - v_\delta|||_{\mathrm{DG}}^2\,,$$

we conclude that

$$|\lambda - \lambda_\delta| \lesssim \inf_{v_\delta \in X_\delta} |||u - v_\delta|||_{\mathrm{DG}}\,.$$

Note that this result can be a bit sharper if q in (8.11b) is significantly smaller than 2; we write it this way for ease of reading. As already mentioned, we will prove a sharper result under some additional conditions in the following sections.

We finish by showing the estimate for the L^2 norm of the error. This follows from Lemma 45, since (8.28) implies

$$\begin{aligned}\|u_\delta - u_\delta^*\|^2 \lesssim \|u_\delta - u_\delta^*\|^r_{L^{6r/(5-s)}}\|\psi_{u_\delta-u_\delta^*}\|_{\mathrm{DG}} + |\lambda_\delta - \lambda_\delta^*|\|u_\delta - u_\delta^*\|\|\psi_{u_\delta-u_\delta^*}\| + \|u - u_\delta^*\|\|\psi_{u_\delta-u_\delta^*}\| \\ + \|u_\delta - u_\delta^*\|^2\|\psi_{u_\delta-u_\delta^*}\| + \|u_\delta - u_\delta^*\|^3\end{aligned} \tag{8.38}$$

for $\psi_{u_\delta-u_\delta^*} \in X_\delta$ defined as in (8.27), with $w_\delta = u_\delta - u_\delta^*$. Now, the coercivity of $\langle (E''(u) - \lambda_\delta^*)\cdot, \cdot\rangle$ over X_δ shown in (8.18a) and a Cauchy-Schwartz inequality imply

$$\|\psi_{u_\delta-u_\delta^*}\|_{\mathrm{DG}} \lesssim \|u_\delta - u_\delta^*\|. \tag{8.39}$$

Hence, from the combination of (8.38), (8.39), and the convergences of λ_δ towards λ_δ^* and of u_δ towards u_δ^* in the $L^2(\Omega)$ norm, we derive

$$\|u_\delta - u_\delta^*\| \lesssim \|u_\delta - u_\delta^*\|^r_{L^{6r/(5-s)}} + \|u - u_\delta^*\|.$$

Noting that

$$\|u - u_\delta\| \leq \|u - u_\delta^*\| + \|u_\delta - u_\delta^*\|$$

we conclude the proof. □

8.2.4 Pointwise convergence

We now wish to recover the doubling of the convergence rate normally obtained for the eigenvalue error, with respect to the eigenfunction. We therefore introduce a result on the convergence of the error

$$\|u - u_\delta\|_{L^\infty(\Omega)}$$

that will be instrumental in the following. In order to do this, we introduce some additional hypotheses that, in the context of hp methods, impose a higher degree of regularity on the solution u. Having an exact solution u which is regular away from the singular points is the motivation for using isotropically refined hp methods, so this should not come as surprising.

We introduce therefore condition (8.40) – note that (8.40) is a sufficient condition for

$$\|g - g_\delta\|_{\mathcal{J}_2^{1,1}(\Omega)} \leq C$$

where g is the solution to $(E''(u) - \lambda_\delta^*)g = \rho$, with ρ defined below. We will then be able to bound the $L^\infty(\Omega)$ error using the errors in the eigenvalue and eigenfunction errors and the $L^\infty(\Omega)$ bound for the Galerkin projection $\widetilde{u}_\delta$ of u defined in (8.41).

Let us introduce a function ρ such that

$$\operatorname{supp}(\rho) = \widetilde{K} \text{ for a } \widetilde{K} \in \mathcal{T}$$
$$\|\rho\|_{L^p(\Omega)} = h_{\widetilde{K}}^{d\frac{1-p}{p}} \text{ for } p \in [1,2]$$

and suppose that, given the solution $g_\delta \in X_\delta$ of

$$\langle (E''(u) - \lambda_\delta^*) g_\delta, v_\delta \rangle = (\rho, v_\delta)$$

for all $v_\delta \in X_\delta$, then

$$\sum_{j=1}^{\ell} h_j^{1/2} \|g_\delta\|_{\mathrm{DG}(\Omega_j)} \leq C, \tag{8.40}$$

where the constant C does not depend on $h_{\widetilde{K}}$.

We furthermore introduce $\widetilde{u}_\delta \in X_\delta$ as the Galerkin projection of u for the operator A^u, i.e., such that

$$\langle A_\delta^u \widetilde{u}_\delta, v_\delta \rangle = \langle A_\delta^u u, v_\delta \rangle \text{ for all } v_\delta \in X_\delta. \tag{8.41}$$

Denote also

$$p_{\max} = \max_{K \in \mathcal{T}} p_K. \tag{8.42}$$

Theorem 10. *Suppose that the hypotheses of Theorem 9 hold. Furthermore, suppose that* (8.40) *holds and that at least one of the following is true: either*

$$p_{\max}^d \|u - u_\delta\|_{\mathrm{DG}}^{r-1} \to 0, \tag{8.43}$$

or

$$s < 4 - r, \tag{8.44}$$

where s and r are defined in (8.11d)*. Then,*

$$\|u - u_\delta\|_{L^\infty(\Omega)} \lesssim p_{\max}^d \left(\|u - u_\delta\|_{\mathrm{DG}}^r + \|u - u_\delta\|_{L^2(\Omega)} + |\lambda - \lambda_\delta| + |\lambda - \lambda_\delta^*| + \|u - \widetilde{u}_\delta\|_{L^\infty(\Omega)} \right),$$

where $\widetilde{u}_\delta \in X_\delta$ is defined as in (8.41)*.*

Proof. We prove the theorem assuming that (8.43) holds; at the end we will delineate the necessary modifications in case only (8.44) holds. The $L^\infty(\Omega)$ error between u_δ and u can be split in two parts, as

$$\|u - u_\delta\|_{L^\infty(\Omega)} \leq \|u - \widetilde{u}_\delta\|_{L^\infty(\Omega)} + \|u_\delta - \widetilde{u}_\delta\|_{L^\infty(\Omega)} \tag{8.45}$$

The first term of the right hand side of the inequality above is the $L^\infty(\Omega)$ norm of the error for a linear problem. We now consider the second part of the right hand side of (8.45),

$$\|\widetilde{u}_\delta - u_\delta\|_{L^\infty(\Omega)} = \|\widetilde{u}_\delta - u_\delta\|_{L^\infty(\widetilde{K})}$$

for a $\widetilde{K} \in \mathcal{T}$. An inverse inequality gives

$$\begin{aligned} \|\widetilde{u}_\delta - u_\delta\|_{L^\infty(\Omega)} &\lesssim h_{\widetilde{K}}^{-d/2} p_{\widetilde{K}}^d \|\widetilde{u}_\delta - u_\delta\|_{L^2(\widetilde{K})} \\ &= p_{\widetilde{K}}^d (\rho, \widetilde{u}_\delta - u_\delta), \end{aligned} \tag{8.46}$$

where we have chosen ρ as

$$\rho = h_{\widetilde{K}}^{-d/2} \frac{\widetilde{u}_\delta - u_\delta}{\|\widetilde{u}_\delta - u_\delta\|_{L^2(\widetilde{K})}} \mathbb{1}_{\widetilde{K}}. \tag{8.47}$$

We now introduce the finite element function g_δ as the solution of an adjoint problem with right hand side ρ. Let

$$\langle \left(E''(u) - \lambda_\delta^*\right) g_\delta, v_\delta \rangle = (\rho, v_\delta) \text{ for all } v_\delta \in X_\delta \tag{8.48}$$

for all $v_\delta \in X_\delta$. Then, we have

$$(\rho, \widetilde{u}_\delta - u_\delta) = \langle (A_\delta^u - \lambda_\delta^*) g_\delta, \widetilde{u}_\delta - u_\delta \rangle + 2 \int_\Omega f'(u^2) u^2 g_\delta (\widetilde{u}_\delta - u_\delta). \tag{8.49}$$

Due to the definition of $\widetilde{u}_\delta$ and the symmetry of the bilinear form,

$$\langle (A_\delta^u - \lambda_\delta^*) g_\delta, \widetilde{u}_\delta - u_\delta \rangle = \lambda(u - \widetilde{u}_\delta, g_\delta) + (\lambda - \lambda_\delta^*)(\widetilde{u}_\delta, g_\delta) - \langle (A_\delta^u - \lambda_\delta^*) g_\delta, u_\delta \rangle \tag{8.50}$$

We can treat the second term by noting that

$$- \langle (A_\delta^u - \lambda_\delta^*) u_\delta, g_\delta \rangle = \int_\Omega \left[f({u_\delta}^2) - f(u^2) \right] g_\delta u_\delta + (\lambda_\delta^* - \lambda_\delta)(u_\delta, g_\delta) \tag{8.51}$$

We want to use (8.11d) on the integrals containing f and its derivative in (8.49) and (8.51). We start by showing that

$$\begin{aligned} 2 \int_\Omega f'(u^2) u^2 g_\delta (\widetilde{u}_\delta - u_\delta) &= 2 \int_\Omega f'(u^2) u^2 g_\delta (u - u_\delta) + 2 \int_\Omega f'(u^2) u^2 g_\delta (\widetilde{u}_\delta - u) \\ &\leq \int_\Omega f'(u^2) u^2 g_\delta (u - u_\delta) + C \|g_\delta\|_{L^2(\Omega)} \|u - \widetilde{u}_\delta\|_{L^2(\Omega)}. \end{aligned}$$

Therefore,

$$\begin{aligned} \left| \int_\Omega \left[f({u_\delta}^2) - f(u^2) \right] g_\delta u_\delta + 2 \int_\Omega f'(u^2) u^2 g (\widetilde{u}_\delta - u_\delta) \right| \lesssim \int_\Omega (1 + |u_\delta|^s) |u - u_\delta|^r |g_\delta| \\ + \|g_\delta\|_{L^2(\Omega)} \|u - \widetilde{u}_\delta\| \end{aligned}$$

Combining (8.49), (8.50), and (8.51) with the above equation gives

$$(\rho, \widetilde{u}_\delta - u_\delta) \lesssim \int_\Omega (1 + |u_\delta|^s)|u - u_\delta|^r |g_\delta| + \|g_\delta\|_{L^2(\Omega)} \|u - \widetilde{u}_\delta\| + (\lambda_\delta^* - \lambda_\delta)(u_\delta, g_\delta) + \lambda(\widetilde{u}_\delta - u_\delta, g_\delta). \quad (8.52)$$

A Hölder inequality and the condition $s < 5 - r$ imply that there exists an

$$0 < \alpha \leq \frac{15 - 3(s + r)}{7 - s - r}$$

such that

$$(\rho, \widetilde{u}_\delta - u_\delta) \lesssim \left(1 + \|u_\delta\|_{L^6(\Omega)}^s\right) \|u - u_\delta\|_{L^6(\Omega)}^{r-1} \|g_\delta\|_{L^{3-\alpha(\Omega)}} \|u - u_\delta\|_{L^\infty(\Omega)} + \|g_\delta\|_{L^2(\Omega)} \left(\|u - \widetilde{u}_\delta\|_{L^2(\Omega)} + \|u - u_\delta\|_{L^2(\Omega)} + |\lambda_\delta^* - \lambda_\delta|\right).$$

Consider now that

$$\mathcal{J}^1_{1/2}(\Omega) \hookrightarrow H^{1/2-\alpha}(\Omega) \hookrightarrow L^{3-\alpha}(\Omega), \quad (8.53)$$

see [Nic97] for the first embedding; the second one is classical in Sobolev spaces. The double embedding (8.53) then implies

$$\begin{aligned} \|g_\delta\|_{L^{3-\alpha}(\Omega)} &\leq C \|g_\delta\|_{\mathcal{J}^1_{1/2}(\Omega)} \\ &\leq C \left(\sum_j \|g_\delta\|^2_{\mathcal{J}^1_{1/2}(\Omega_j)} \right)^{1/2} \\ &\leq C \left(\sum_j h_j \|g_\delta\|^2_{H^1(\Omega_j)} \right)^{1/2} \end{aligned}$$

where the second inequality follows from the fact that $r_{|\Omega_j}/h_j \leq C$ for all $j = 1, \ldots, \ell$. Therefore, using (8.40) and noting that the $\ell^2(\{1, \ldots, \ell\})$ norm is bounded by the $\ell^1(\{1, \ldots, \ell\})$ norm with constants that do not depend on ℓ, we conclude that $\|g_\delta\|_{L^{3-\alpha}(\Omega)} \leq C$ for any positive α, thus,

$$(\rho, \widetilde{u}_\delta - u_\delta) \lesssim \|u - u_\delta\|_{L^6(\Omega)}^{r-1} \|u - u_\delta\|_{L^\infty(\Omega)} + \|u - \widetilde{u}_\delta\|_{L^2(\Omega)} + \|u - u_\delta\|_{L^2(\Omega)} + |\lambda - \lambda_\delta| + |\lambda - \lambda_\delta^*|.$$

If hypothesis (8.43) holds, we can conclude with the thesis. If (8.43) does not hold, then hypothesis (8.44) is necessary: the proof follows the same lines, though at (8.52) we use the inequality

$$\int_\Omega (1 + |u_\delta|^s)|u - u_\delta|^r |g_\delta| \leq C \|u_\delta\|_{L^6(\Omega)}^s \|u - u_\delta\|_{L^6(\Omega)}^r \|g_\delta\|_{L^{3-\alpha}(\Omega)}$$

for an

$$0 < \alpha \leq \frac{12 - 3s - 3r}{6 - s - r}.$$

Note that such an α exists thanks to (8.44). In this case we conclude

$$(\rho, \widetilde{u}_\delta - u_\delta) \lesssim \|u - u_\delta\|^r_{L^6(\Omega)} + \|u - \widetilde{u}_\delta\|_{L^2(\Omega)} + \|u - u_\delta\|_{L^2(\Omega)} + |\lambda - \lambda_\delta| + |\lambda - \lambda^*_\delta|,$$

hence the thesis. □

8.2.5 Convergence revisited

In this section we finally show that, if the solution converges in the $L^\infty(\Omega)$ norm, we can prove that the eigenvalue converges with the same rate as the square of the eigenfunction. We therefore suppose that the following hold

$$p^d_{\max} \inf_{v_\delta \in X_\delta} \|u - v_\delta\|_{\mathrm{DG}} \to 0 \tag{8.54a}$$

$$p^d_{\max} \|u - \widetilde{u}_\delta\|_{L^\infty(\Omega)} \to 0, \tag{8.54b}$$

which is sufficient for $\|u - u_\delta\|_{L^\infty(\Omega)} \to 0$. Note that (8.54a) depends on the regularity of u (which in turn depends on the regularity of F). Note also that, in the case of isotropically refined hp methods, assuming (8.54b) is pleonastic, as it is a consequence of (8.40).

When dealing with h type discontinuous finite element methods, $p_{\max}$ is bounded by some global constant, thus (8.54b) translates into the requirement of simple $L^\infty(\Omega)$ convergence for the linear problem. Furthermore, if the mesh is globally regular, (8.54b) and (8.40) can be proven if the solution is sufficiently regular, see [CC04]. Therefore, the following theorem can be extended to h-type finite element methods too, as long as the solution is sufficiently regular. We will not treat this case, as it it outside the focus of our analysis.

Under the additional hypotheses (8.56), we prove that the eigenvalues converge as the square of the eigenfunction.

We introduce another adjoint problem: let $\psi \in u^\perp$ such that

$$\langle \left(E''(u) - \lambda\right) \psi, v \rangle = (f'(u^2)u^3, v) \tag{8.55}$$

for all $v \in \{v \in X(\delta) : (u, v) = 0\}$.

Theorem 11. *Suppose that the hypotheses of Theorems 9 and 10, and conditions* (8.54a) *and* (8.54b) *hold. Furthermore, suppose that* (8.11d) *holds with $r = 2$. and that*

$$F \in C^3((0, +\infty), \mathbb{R}), \text{ and } F'''(t)t^2 \text{ is locally bounded in } [0, +\infty). \tag{8.56}$$

Then

$$|\lambda - \lambda_\delta| \lesssim |||u - u_\delta|||_{\mathrm{DG}} \left(\inf_{v_\delta \in X_\delta} |||\psi - v_\delta|||_{\mathrm{DG}} + |||u - u_\delta|||_{\mathrm{DG}} \right), \tag{8.57}$$

where ψ is defined in (8.55) *above.*

Proof. The proof begins similarly to (8.36):

$$\begin{aligned}\lambda_\delta - \lambda &= \langle (A^u_\delta - \lambda)\,(u-u_\delta)\,, u-u_\delta\rangle + \int_\Omega \left[f(u_\delta^2) - f(u^2)\right] {u_\delta}^2 \\ &= \langle (A^u_\delta - \lambda)\,(u_\delta-u)\,, u_\delta-u\rangle + \int_\Omega \left[f(u_\delta^2) - f(u^2) - f'(u^2)({u_\delta}^2-u^2)\right] u_\delta^2 \end{aligned} \tag{8.58a}$$

$$+ \int_\Omega f'(u^2)\left[{u_\delta}^2(u+u_\delta) - 2u^3\right](u-u_\delta) \tag{8.58b}$$

$$+ \int_\Omega 2f'(u^2)u^3(u-u_\delta). \tag{8.58c}$$

We consider the three integrals in the last equation separately. Firstly, considering term (8.58b), we have

$$\begin{aligned}\int_\Omega f'(u^2)\left[{u_\delta}^2(u+u_\delta) - 2u^3\right](u-u_\delta) &= \int_\Omega f'(u^2)(u^2+2uu_\delta+2{u_\delta}^2)(u-u_\delta)^2 \\ &\lesssim \int_\Omega \left(1 + \frac{u_\delta}{u} + \frac{{u_\delta}^2}{u^2}\right)(u-u_\delta)^2.\end{aligned}$$

Thanks to the Cauchy-Schwartz inequality, to the assumed $L^\infty(\Omega)$ convergence of u_δ towards u, and to the fact that there exists $u_{\min}$ such that $u \geq u_{\min} > 0$, see [CCM10], the above inequality implies that (after a certain level of refinement)

$$\left|\int_\Omega f'(u^2)\left[{u_\delta}^2(u+u_\delta) - 2u^3\right](u-u_\delta)\right| \lesssim \|u-u_\delta\|^2. \tag{8.59}$$

Integral (8.58c) is then treated by using (8.55)

$$\begin{aligned}\int_\Omega f'(u^2)u^3(u-u_\delta) &= (f'(u^2)u^3, (u-u_\delta)^{u^\perp}) + (f'(u^2)u^3, (u-u_\delta, u)u) \\ &= \langle \left(E''(u) - \lambda\right)\psi, (u-u_\delta)^{u^\perp}\rangle + \frac{1}{2}\|u-u_\delta\|^2\|f'(u^2)u^3\|_{L^2}\end{aligned} \tag{8.60}$$

Consider the first term above: for any $\tilde{v}_\delta \in u_\delta^\perp$,

$$\begin{aligned}\langle \left(E''(u) - \lambda\right)\psi, (u-u_\delta)^{u^\perp}\rangle &= \langle \left(E''_\delta(u) - \lambda\right)(\psi - \tilde{v}_\delta), (u-u_\delta)^{u^\perp}\rangle \\ &\quad + \langle \left(E''_\delta(u) - \lambda\right)\tilde{v}_\delta, (u-u_\delta)^{u^\perp}\rangle,\end{aligned} \tag{8.61}$$

Now,

$$\begin{aligned}\langle \left(E_\delta''(u)-\lambda\right)\tilde{v}_\delta, (u-u_\delta)^{u^\perp}\rangle &= -\langle (A_\delta^u-\lambda)\,u_\delta, \tilde{v}_\delta\rangle + 2\int_\Omega f'(u^2)u^2(u-u_\delta)^{u^\perp}\tilde{v}_\delta \\ &= \int_\Omega \left(f({u_\delta}^2)-f(u^2)\right)u_\delta\tilde{v}_\delta + (\lambda-\lambda_\delta)(u_\delta,\tilde{v}_\delta) \\ &\quad + 2\int_\Omega f'(u^2)u^2(u-u_\delta)\tilde{v}_\delta + \|u-u_\delta\|^2\int_\Omega f'(u^2)u^3\end{aligned}$$

thus

$$\langle \left(E_\delta''(u)-\lambda\right)\tilde{v}_\delta, (u-u_\delta)^{u^\perp}\rangle \lesssim \|u-u_\delta\|^2\|\tilde{v}_\delta\|_{L^\infty(\Omega)} + \|u-u_\delta\|^2,$$

where we have used the fact that (8.11d) holds with $r=2$, the orthogonality between $\tilde{v}_\delta$ and u_δ, condition (8.11c), and the $L^\infty(\Omega)$ convergence of u_δ towards u. We now turn to the first term at the right hand side of (8.61). We have

$$\langle \left(E_\delta''(u)-\lambda\right)(\psi-\tilde{v}_\delta), (u-u_\delta)^{u^\perp}\rangle \lesssim \left(|||u-u_\delta|||_{\mathrm{DG}} + \|u-u_\delta\|^2\,|||u|||_{\mathrm{DG}}\right)|||\psi-\tilde{v}_\delta|||_{\mathrm{DG}}\,.$$

Note that the term in $\|u-u_\delta\|^2$ is of higher order, so we can omit it from the following estimates. We have therefore, from (8.61),

$$\begin{aligned}\langle \left(E''(u)-\lambda\right)\psi, (u-u_\delta)^{u^\perp}\rangle &\lesssim \inf_{\tilde{v}_\delta\in u_\delta^\perp}\left[\|u-u_\delta\|^2\left(\|\tilde{v}_\delta\|_{L^\infty(\Omega)}+1\right) + |||u-u_\delta|||_{\mathrm{DG}}\,|||\psi-\tilde{v}_\delta|||_{\mathrm{DG}}\right] \\ &\lesssim \inf_{v_\delta\in X_\delta}\Big[\|u-u_\delta\|^2\left(\|v_\delta-(v_\delta,u_\delta)u_\delta\|_{L^\infty(\Omega)}+1\right) \\ &\qquad + |||u-u_\delta|||_{\mathrm{DG}}\,|||\psi-v_\delta+(v_\delta,u_\delta)u_\delta|||_{\mathrm{DG}}\Big],\end{aligned} \tag{8.62}$$

where we have replaced $\tilde{v}_\delta$ by $v_\delta-(v_\delta,u_\delta)u_\delta$, thus being able to extend the $\inf$ over all v_δ in X_δ. Now,

$$\begin{aligned}|||\psi-v_\delta+(v_\delta,u_\delta)u_\delta|||_{\mathrm{DG}} &\le |||\psi-v_\delta|||_{\mathrm{DG}} + |||(\psi,u_\delta)u_\delta|||_{\mathrm{DG}} + |||(\psi-v_\delta,u_\delta)u_\delta|||_{\mathrm{DG}} \\ &\lesssim |||\psi-v_\delta|||_{\mathrm{DG}} + \|\psi-v_\delta\| + \|\psi\|\|u-u_\delta\|.\end{aligned} \tag{8.63}$$

Furthermore,

$$\|v_\delta-(v_\delta,u_\delta)u_\delta\|_{L^\infty(\Omega)} \le \|v_\delta\|_{L^\infty(\Omega)} + \|u_\delta\|_{L^\infty(\Omega)}\|v_\delta\|. \tag{8.64}$$

The best approximation v_δ to ψ in the $|||\cdot|||_{\mathrm{DG}}$ norm is such that $\|\psi-v_\delta\|_{L^\infty\Omega}\to 0$; furthermore, the norm $\|\psi\|_{L^\infty(\Omega)}$ can be bounded by a constant depending on u by elliptic regularity, hence $\|v_\delta-(v_\delta,u_\delta)u_\delta\|_{L^\infty(\Omega)}\le C$. Using these remarks, (8.63), and (8.64), inequality (8.61) can be rewritten as

$$\langle \left(E_\delta''(u)-\lambda\right)\psi, (u-u_\delta)^{u^\perp}\rangle \lesssim |||u-u_\delta|||_{\mathrm{DG}}\left(\inf_{v_\delta\in X_\delta}|||\psi-v_\delta|||_{\mathrm{DG}} + \|u-u_\delta\|\right),$$

where, once again, we have omitted the higher order terms. Going back to (8.60) we obtain

$$\left|\int_\Omega f'(u^2)u^3(u-u_\delta)\right| \lesssim |||u-u_\delta|||_{\mathrm{DG}} \left(\inf_{v_\delta\in X_\delta} |||\psi - v_\delta|||_{\mathrm{DG}} + \|u-u_\delta\|\right) \tag{8.65}$$

We finally consider the second term of line (8.58a). Under the additional hypotheses $F \in C^3$ and $t^2F'''(t)$ locally bounded in $[0,\infty)$, denoting $w = [f(u_\delta^2) - f(u^2) - f'(u^2)({u_\delta}^2 - u^2)]$ and recalling that $u \geq u_{\min} > 0$,

$$\begin{aligned}
\int_\Omega w{u_\delta}^2 &= \int_\Omega \left(\int_0^1 tf''(u^2 + t({u_\delta}^2 - u^2))dt\right) {u_\delta}^2({u_\delta}^2 - u^2)^2 \\
&\lesssim \int_\Omega \left(\int_0^1 \frac{t}{(u^2 + t({u_\delta}^2 - u^2))^2}dt\right) {u_\delta}^2({u_\delta}^2 - u^2)^2 \\
&= \int_\Omega \left|{u_\delta}^2 \log\left(\frac{u^2}{{u_\delta}^2}\right) + {u_\delta}^2 - u^2\right| \left|{u_\delta}^2 - u^2\right|
\end{aligned}$$

Under the hypothesis of $L^\infty(\Omega)$ convergence given in Theorem 10, then,

$$\left|\int_\Omega \left[f(u_\delta^2) - f(u^2) - f'(u^2)({u_\delta}^2 - u^2)\right] u_\delta^2\right| \lesssim \|u-u_\delta\|^2. \tag{8.66}$$

The thesis follows from (8.58a)–(8.58c), (8.59), (8.65), and (8.66). □

We conclude this section by remarking that ψ satisfies the equation

$$\left(A^u + 2f'(u^2)u^2 - \lambda\right)\psi = 2\left(\int_\Omega f'(u^2)u^3\psi\right)u + f'(u^2)u^3 - (f'(u^2)u^3, u)u. \tag{8.67}$$

The regularity of ψ depends then on the regularity of f, f', and u. In particular, if $u \in \mathcal{J}^s_\gamma(\Omega)$ for a certain $s > 2$, and

$$\|f'(u^2)u^2\|_{\mathcal{J}^j_\gamma(\Omega)} \leq C\|f(u^2)\|_{\mathcal{J}^j_\gamma(\Omega)} \tag{8.68}$$

for all $j \leq s-2$, then the line of reasoning used to prove the regularity of u can be used for ψ, using the above inequality to derive the estimates on u and ψ by elliptic estimates in weighted Sobolev spaces. Then, (8.57) means that the eigenvalues converge at a rate which is approximately double that of the eigenfunction. In the next section, we will see this in the special case of a polynomial f, which implies, under some hypotheses, $u \in \mathcal{J}^\varpi_\gamma(\Omega)$, $\psi \in \mathcal{J}^\varpi_\gamma(\Omega)$, and exponential convergence of the numerical solution.

8.2.6 Exponential convergence

In this section, as mentioned above, we restrict further the hypothesis made on F, in that we consider the concrete case where F is a polynomial. Let then

$$f(u^2) = u^k \tag{8.69}$$

for $1 \leq k \leq 3$ (the case $k = 0$ is the linear one). Remark that this class of functions satisfies (8.11a) to (8.11d), with in particular $r = 2$ in (8.11d). Furthermore, remark that (8.68) is an equality with $C = k/2$. We recall here from Chapter 7 a result on the regularity of the solution u.

Theorem 12. *Let u be the solution to* (8.9) *with $V \in \mathcal{K}^{\varpi,\infty}_{\varepsilon-2}(\Omega)$ and f defined as in* (8.69)*, with $k = 1, 2, 3$. Then,*

$$u \in \mathcal{J}^{\varpi,p}_{\gamma}(\Omega) \tag{8.70}$$

for any $\gamma < \min(d/p + \varepsilon, 2)$.

Furthermore, we have that for a function $v \in \mathcal{J}^{\varpi}_{\gamma}(\Omega)$, for a $\gamma > d/2$, there exists two constants $C, b > 0$ such that

$$\inf_{v_\delta \in X_\delta} |||v - v_\delta|||_{\mathrm{DG}} \leq Ce^{-b\ell}.$$

Here X_δ is a isotropically refined hp finite element space, as described in Section 8.1.2, ℓ is the number of refinement steps, and $\ell = N^{1/(d+1)}$, with N denoting the number of degrees of freedom of X_δ. We finally remark that since in this instance $f'(u^2)u^2 = Cf(u^2)$, the non scalar coefficients in (8.67) are the same that we find in (8.9). Hence, using elliptic regularity in weighted Sobolev spaces and the proof of Theorem 12, we obtain

$$\psi \in \mathcal{J}^{\varpi,p}_{\gamma}(\Omega) \tag{8.71}$$

for any $\gamma < \min(d/p + \varepsilon, 2)$, and in particular for all $s \geq 2$, $\|\psi\|_{\mathcal{J}^s_\gamma(\Omega)} \leq C\|u\|_{\mathcal{J}^s_\gamma(\Omega)}$. Therefore,

$$\inf_{v_\delta \in X_\delta} |||\psi - v_\delta|||_{\mathrm{DG}} \leq Ce^{-b\ell}. \tag{8.72}$$

We can regroup the results of the previous sections, applied to the case where (8.69) holds, in the following theorem.

Theorem 13. *Let u, λ be the solution to* (8.9) *and u_δ, λ_δ be the solution to* (8.10)*. Suppose that* (8.12a), (8.12b)*, and* (8.69) *hold. Then, for a space X_δ with N degrees of freedom, there exists $b > 0$ such that*

$$\|u - u_\delta\|_{\mathrm{DG}} \leq Ce^{-bN^{1/(d+1)}} \tag{8.73}$$

and

$$|\lambda - \lambda_\delta| \leq Ce^{-bN^{1/(d+1)}}. \tag{8.74}$$

Furthermore, if (8.40) *holds, then,*

$$|\lambda - \lambda_\delta| \leq Ce^{-2bN^{1/(d+1)}}. \tag{8.75}$$

8.3 Convergence of an iterative scheme

Following [CL02], we introduce the *level shifting* iterative scheme: given u_n, we compute u_{n+1} as the eigenfunction corresponding to the smallest eigenvalue λ_{n+1} in

$$-\Delta u_{n+1} + Vu_{n+1} + f(u_n^2)u_{n+1} - b(u_{n+1}, u_n)u_n = \lambda_{n+1}u_{n+1}, \tag{8.76}$$

where $b > 0$ is a shift parameter. Note that, given a sufficiently regular initial function u_0, the positive solution u_{n+1} to (8.76) is strictly positive [Sta65] in the open domain, i.e., $u_{n+1} > 0$ in Ω, for all n. Furthermore, we define the "energy" $\widetilde{E}$ as

$$\widetilde{E}(u, v) = \frac{1}{2}\left(\|\nabla u\|^2 + \|\nabla v\|^2 + (Vu, u) + (Vv, v) + (f(v^2), u^2)\right) + \frac{b}{2}\left(1 - (u, v)^2\right).$$

The solution u_{n+1} to (8.76) then satisfies the relation

$$\widetilde{E}(u_{n+1}, u_n) \leq \widetilde{E}(u, u_n), \quad \forall u \in X,\ \|u\| = 1.$$

This implies in particular $\widetilde{E}(u_{n+1}, u_n) \leq \widetilde{E}(u_n, u_n)$, or equivalently

$$E(u_{n+1}) + \frac{b}{8}\|u_{n+1} - u_n\|^2\|u_{n+1} + u_n\|^2 \leq E(u_n) + \frac{1}{2}\int_\Omega F(u_{n+1}^2) - F(u_n^2) - f(u_n^2)(u_{n+1}^2 - u_n^2). \tag{8.77}$$

Thanks to the convexity of F,

$$\int_\Omega F(u_{n+1}^2) - F(u_n^2) - f(u_n^2)(u_{n+1}^2 - u_n^2) \leq 0.$$

Furthermore, note that for two functions $v, w \in X$, considering the case $d = 3$

$$\begin{aligned}\int_\Omega (\nabla v)^2 + \int_\Omega Vv^2 + \int_\Omega f(w^2)v^2 &\geq \|\nabla v\|^2 - \|V\|_{L^{p_V}(\Omega)}\|v\|_{L^2(\Omega)}^{(2p-3)/p}\|v\|_{L^6(\Omega)}^{3/p} + f(0)\|v\|^2 \\ &\geq \frac{1}{2}\|v\|_{H^1(\Omega)}^2 - C\|v\|_{L^2(\Omega)}^2\end{aligned}$$

where p_V is defined in (8.12a) and the constant C depends on $f(0)$, on V and on p_V, and on the constant related to the Sobolev embedding of $H^1(\Omega)$ in $L^6(\Omega)$. When $d = 1$ or

$d = 2$ the same thing can be proven similarly. Therefore, since $\|u_n\| = \|u_{n+1}\| = 1$

$$\begin{aligned}\widetilde{E}(u_{n+1}, u_n) &\geq \frac{1}{2}\left(\int_\Omega (\nabla u_{n+1})^2 + \int_\Omega V u_{n+1}^2 + \int_\Omega (\nabla u_n)^2 + \int_\Omega V u_n^2 + \int_\Omega f(u_n^2)(u_{n+1})^2\right) \\ &\gtrsim \|u_{n+1}\|^2_{H^1(\Omega)} + \|u_n\|^2_{H^1(\Omega)} - C\end{aligned}$$

This provides a lower bound for $\widetilde{E}$; thus, if $b > 0$ we infer from (8.77) that

i) $\|u_n\|_{H^1(\Omega)}$ is uniformly bounded and

ii) $\sum_n \|u_{n+1} - u_n\|^2 \|u_{n+1} + u_n\|^2 < \infty$.

Take then a subsequence $\{u_{n_k}\}_k$ such that

$$u_{n_k} \to v \qquad \text{and} \qquad u_{n_k - 1} \to w$$

both weakly in $H^1(\Omega)$ and strongly in $L^2(\Omega)$. Thanks to item ii), we have $v = \pm w$. We suppose without loss of generality that $v = w$. Furthermore, By the lower semicontinuity of the norm, the uniform boundedness of $\|u_n\|_{H^1(\Omega)}$ extends to w. Since (8.76) holds, using elliptic regularity, the uniform boundedness of λ_n, and item i) above, we can affirm that $\{u_n\}_n$ is uniformly bounded in $L^\infty(\Omega)$. We affirm now that

Lemma 46. *If $\|v_n - v\| \to 0$ and $\|v_n\|_{L^\infty(\Omega)}$ is uniformly bounded, then $\|v\|_{L^\infty(\Omega)}$ is bounded.*

Proof. Suppose v is not bounded in $L^\infty(\Omega)$. Then, for any $k > 0$, there exists a subset $E_k \subset \Omega$ such that $|v_{|_{E_k}}| \geq k$ and the measure of E_k is strictly positive. Let us introduce C such that, for all n, $\|v_n\|_{L^\infty(\Omega)} \leq C$. Take a $\tilde{k} > C$ and note that

$$\left|(v_n - v)_{|_{E_{\tilde{k}}}}\right| \geq \tilde{k} - C$$

for any n. Therefore, denoting $|\cdot|$ the measure of a set,

$$\|v_n - v\|_{L^2(\Omega)} \geq \|v_n - v\|_{L^2(E_{\tilde{k}})} \geq (\tilde{k} - C)|E_{\tilde{k}}|^{1/2}.$$

Since the measure of $E_{\tilde{k}}$ is positive, this contradicts the hypothesis that v_n converges towards v in $L^2(\Omega)$, thus proving that v is bounded in $L^\infty(\Omega)$. □

This implies that $\|w\|_{L^\infty(\Omega)} \leq C$. We now show that $f(u_{n_k-1}^2)u_{n_k} \xrightarrow{L^2} f(w^2)w$. Consider a function φ such that $\|\varphi\|_{L^2(\Omega)} = 1$; then,

$$\begin{aligned}\int_\Omega f(u_{n_k-1}^2)u_{n_k}\varphi - f(w^2)w\varphi = \int_\Omega \left[f(u_{n_k-1}^2) - f(w^2)\right] u_{n_k-1}\varphi + \int_\Omega f(u_{n_k-1}^2)(u_{n_k} - u_{n_k-1})\varphi \\ + \int_\Omega f(w^2)(u_{n_k-1} - w)\varphi. \quad (8.78)\end{aligned}$$

Let us now consider

$$h(a,b) = a\frac{f(a^2) - f(b^2)}{a-b}.$$

Suppose $a > b > 0$. Then,

$$|h(a,b)| \leq \begin{cases} \max\limits_{\xi\in[0,2\|b\|_{L^\infty(\Omega)}]} 6|f'(\xi^2)|\xi^2 & \text{if } a \leq 2b \\ f(a^2) + f(b^2) & \text{if } a > 2b. \end{cases}$$

The requirement $a > b > 0$ can be easily dropped, thus, since u_{n_k-1} and w are bounded in $L^\infty(\Omega)$ and using (8.11b) and (8.11c),

$$\|h(u_{n_k-1}, w)\|_{L^\infty(\Omega)} \leq C$$

and this implies

$$\begin{aligned} \int_\Omega \left[f(u_{n_k-1}^2) - f(w^2)\right] u_{n_k-1}\varphi &\leq C\int_\Omega |u_{n_k-1} - w||\varphi| \\ &\leq C\|\varphi\|\|u_{n_k-1} - w\| \end{aligned} \tag{8.79}$$

From (8.78) we can conclude that

$$\|f(u_{n_k-1}^2)u_{n_k} - f(w^2)w\| \lesssim \|u_{n_k-1} - u_{n_k}\| + \|u_{n_k-1} - w\| \to 0. \tag{8.80}$$

taking the limit in (8.76) and denoting $\mu = \lim \lambda_{n_k}$,

$$A^w w = \mu w.$$

The weak lower semicontinuity of the norms can then be used to prove that $\|w\| = 1$, see the proof of Theorem 7 in [CL02]. In addition, we find that

$$\begin{aligned} \lim \|\nabla u_{n_k}\|^2 &= \lim \left(\lambda_{n_k} - \int_\Omega V u_{n_k}^2 - \int_\Omega f(u_{n_k-1}^2)u_{n_k}^2\right) \\ &= \mu - \int_\Omega V w^2 - \int_\Omega f(w^2)w^2 \\ &= \|\nabla w\|^2, \end{aligned}$$

hence, $u_{n_k} \xrightarrow{H^1} w$. Finally, by embedding and elliptic regularity,

$$\begin{aligned} \|u_{n_k} - w\|_{L^\infty(\Omega)} &\leq C\|A^w(u_{n_k} - w)\| \\ &\lesssim \|\lambda_{n_k}u_{n_k} - \mu w\| + \|\left[f(u_{n_k-1}^2) - f(w^2)\right] u_{n_k}\| \\ &\lesssim |\lambda_{n_k} - \mu| + |\mu|\|u_{n_k} - w\| + \|\left[f(u_{n_k-1}^2) - f(w^2)\right] u_{n_k}\| \to 0. \end{aligned}$$

We summarize the results.

Proposition 47. *Let $(u_n, \lambda_n)_n \subset X^{\mathbb{N}} \times \mathbb{R}^{\mathbb{N}}$ be defined by (8.76), for a smooth u_0 and $b > 0$. Then, $\{u_n\}_n$ converges in $H^1(\Omega)$ and $L^\infty(\Omega)$ to an eigenfunction w of the nonlinear problem (8.9), with eigenvalue $\mu = \lim_n \lambda_n$, i.e.,*

$$A^w w = \mu w.$$

Furthermore,

$$\sum_{n\in\mathbb{N}} \|u_n - u_{n-1}\|^2 < \infty$$

and the energy $E(u_n)$ decreases towards that of a stationary state.

8.4 Asymptotic analysis near the singularity

In this section, we perform the analysis of the asymptotic expansion of the solution to (8.9) and (8.76) near the singular point. Apart from the theoretical interest of such a result, this can be used to construct hp spaces that are *a priori* optimized for the approximation of the exact solution; in addition, it could also be a starting point for the construction of an extended finite elements method in which some of the singular functions that show up in the asymptotic expansion of the exact solution are directly added to the finite element basis. This may or may not be advisable, depending on the nonlinearity, on the potential and on a number of computational observations; the discussion of such a subject is out of the scope of the present analysis.

Under general hypotheses on the nonlinearity and the on the potential, only the lowest order of the expansion can be obtained; if instead we fully specify V and f, we can derive a full asymptotic expansion by an iterative procedure. This is done in Section 8.4.1, where we choose a simple radially symmetric potential and a polynomial nonlinearity.

In order to carry out the general asymptotic analysis, we introduce a space of functions with set asymptotic expansion: consider a vector of exponents $\mathbf{p} = \{p_j\}_j \in \mathbb{R}^n$, sorted in increasing order, and a vector $\mathbf{m} = \{m_j\}_j \in \mathbb{N}^n$. Here we allow both $n \in \mathbb{N}$ and $n = \infty$; in the latter case, we impose the additional restriction that $p_j \to \infty$ when $j \to \infty$. We then introduce the space $\mathcal{E}_{\mathbf{p},}$

$as the space of functions u such that there exist a_{jk} \in C^\infty(\mathbf{S}^{d-1})$ such that

$$u - \sum_{j=1}^{n}\sum_{k=0}^{m_j} a_{jk} r^{p_j} (\log r)^k \in \mathcal{K}^\infty_{\sup \mathbf{p}+d/2}(\Omega). \tag{8.81}$$

This can also be written

$$u \sim_{(\infty,\gamma)} \sum_{j=1}^{n}\sum_{k=0}^{m_j-1} a_{jk} r^{p_j} (\log r)^k$$

with $\gamma = \sup \mathbf{p} + d/2$, using the notation introduced in Section 2.1. Note that in the case

of multiple points inside $\mathfrak{C}$, we would simply assign an asymptotic class (i.e., $\mathbf{p}$ and $\mathbf{m}$) to every point in $\mathfrak{C}$, separately. We are interested in the asymptotic expansion of the solution near the singularity, so the analysis is eminently local.

We make an additional hypothesis on the potential and suppose that we can develop it as a series of powers of r around every singularity, i.e., there exists an infinite vector of real numbers $\mathbf{p^V} = \{p_j^V\}_{j\in\mathbb{N}}$, sorted in increasing order, such that $p_j^V \to \infty$, and

$$V \in \mathcal{E}_{\mathbf{p^V},\mathbf{0}} \tag{8.82}$$

We indicate by $\hat{u}$ the Mellin transform of u, defined with its inverse by

$$\hat{u}(z) = \mathcal{M}_{r\to z}u(r) = \int_0^\infty r^{-z-1}u(r)dr, \qquad \mathcal{M}^{-1}_{z\to r}\hat{u}(z) = \int_{\mathrm{Re}\, z=\gamma} r^z\hat{u}(z)dz,$$

where $\mathrm{Re}\, z = \gamma$ is the straight line through $\gamma \in \mathbb{R}$ parallel to the imaginary axis. For a thorough analysis of the Mellin transform see Section 2.1.

Equation (8.76) can be transformed and written as

$$\begin{aligned}(-z(z+d-2) - \Delta_{S^{d-1}})\,\hat{u}_{n+1}(z) = \sum_j a_{jk}^V \hat{u}_{n+1}(z-2-p_j^V) + b(u_{n+1},u_n)\hat{u}_n(z-2)\\ - \mathcal{M}_{r\to z}\left(r^2 f(u_n^2)u_{n+1}\right) + \lambda_{n+1}\hat{u}_{n+1}(z+2).\end{aligned} \tag{8.83}$$

The eigenvalues of $-\Delta_{\mathbb{S}^{d-1}}$ are $\mu_k = k(k+d-2)$, $k \in \mathbb{N}$; the inverse of the operator symbol $P(z) = -z(z+d-2) - \Delta_{\mathbb{S}^{d-1}}$ in the vicinity of μ_k is given by [KM99, Theorem A.10.2]

$$P^{-1}(z) = \frac{1}{z-\mu_k}\Pi_{\mu_k} + \Gamma(z),$$

where Π_{μ_k} is the projector on the eigenspace of $-\Delta_{\mathbb{S}^{d-1}}$ associated with the eigenvalue μ_k and $\Gamma(z)$ is an holomorphic operator function. Both u_n and u_{n+1} are in $L^\infty(\Omega)$. Suppose the Mellin transform of a function has a pole for $z = \zeta$: then, there would be a term in its expansion proportional to r^ζ. Thus, $\hat{u}_n(z)$ and $\hat{u}_{n+1}(z)$ have to be holomorphic for $z < 0$. Similarly, we can affirm that the Mellin transform of $r^2 f(u_n^2)u_{n+1}$ has no poles for $z < 2$. We start by considering the solutions of

$$-z(z+d-2) + k(k+d-2) = 0,$$

i.e., $z = -k$ or $z = k+d-2$. When $d = 2$ and for $k = 0$, the equation above reduces to $z^2 = 0$, thus $\hat{u}_{n+1}$ could have a pole with double multiplicity at $z = 0$. This would give rise to an asymptotic expansion

$$u(r) = a_1 + a_2\log(r) + \dots,$$

where $a_{1,2}$ are functions from $\mathbb{S}^1$ to $\mathbb{R}$ and the omitted terms are of higher order in r. Such a function would not belong to $H^1(\Omega)$ if $a_2 \neq 0$, hence it is necessary that $a_2 = 0$. This implies that the pole in $z = 0$ has single multiplicity even when $d = 2$.

Iterating on (8.83), we see that the last three terms at the right hand side are holomorphic for $z < 2$; the term coming from the transformation of the potential has a pole for $z = 2 + p_0^V$. Suppose that $p_0^V = -2 + \varepsilon$, for an $0 < \varepsilon < 1$ (this is consistent with (8.12b)): then $z = \varepsilon$ is a pole of $\hat{u}_{n+1}$, corresponding to a term proportional to r^ε in the asymptotic expansion of u_{n+1}.

To summarize, we have found that $\hat{u}_{n+1}$ has

i) no poles for $z < 0$,

ii) a single pole at $z = 0$,

iii) a single pole at $z = \varepsilon$ if the potential's most singular term is proportional to $r^{-2+\varepsilon}$.

The following (in the direction of increasing $\operatorname{Re} z$) poles depend on the explicit form of the potential V and of the function f. We consider a special case, where a full asymptotic expansion of the function can be obtained, in the next section.

8.4.1 Asymptotics of the solution to the Gross-Pitaevskii equation

In this section we consider the special case of the Gross-Pitaevskii equation, where $f(u^2) = u^2$. Furthermore, we suppose that the potential $V : \mathbb{R}^d \to \mathbb{R}$ is given by

$$V(x) = -\frac{Z}{|x|^\beta}, \tag{8.84}$$

with $\beta < 2$ if $d = 2, 3$ and $\beta < 1$ if $d = 1$. Note that for $\beta = 1$ the potential is Coulombian.

We will need to differentiate between the cases $d = 1$ and $d = 2, 3$. To do so, we introduce

$$\rho_d = \begin{cases} 1 & \text{if } d = 1 \\ 2 & \text{if } d = 2, 3, \end{cases}$$

and consider the asymptotic class vectors

$$\begin{aligned} \bar{\mathbf{p}} &= \{\rho_d j - k_j \beta\}_{j,k_j} \text{ for } j \in \mathbb{N}, k_j = 0, \dots, \lfloor j\rho_d/2 \rfloor \\ \overline{\mathbf{m}} &= \mathbf{0} \end{aligned}$$

and consider the functions $v \in \mathcal{E}_{\bar{\mathbf{p}},\overline{\mathbf{m}}}$. This corresponds to functions with asymptotic expansion

$$v \sim \sum_{j \in \mathbb{Z}} \sum_{k=0}^{\lfloor \rho_d j/2 \rfloor} a_{jk} r^{\rho_d j - k\beta}$$

in the classical sense near the singularity, see Lemma 3, with $a_{jk} \in C^\infty(\mathbb{S}_{d-1})$.

Suppose now that $u_n \in \mathcal{E}_{\bar{\mathbf{p}},\overline{\mathbf{m}}}$: then $u_n^2 \in \mathcal{E}_{\bar{\mathbf{p}},\overline{\mathbf{m}}}$. We consider then equation (8.83) and derive, as already discussed at the end of the previous section, that the poles with biggest

real part of $\hat{u}(z)$ are given by

$$z \in \mathfrak{P}_0 = \begin{cases} \{0,1\} & \text{if } d = 1 \\ \{0\} & \text{if } d = 2,3, \end{cases}$$

with multiplicity one. We can then iterate by seeing that the right hand side of (8.83) has singularities for

$$\mathfrak{P}_1 = \{z \in \mathbb{R} : z - 2 + \beta \in \mathfrak{P}_0 \\ \vee\, z - 2 - \rho_d j + k\beta \in \mathfrak{P}_0, j \in \mathbb{N}, k = 0, \dots, \lfloor \rho_d j/2 \rfloor\},$$

which in turn implies that the poles of $\hat{u}_{n+1}$ are in $\mathfrak{P}_0 \cup \mathfrak{P}_1$. Applying this step iteratively, we obtain we obtain $\mathfrak{P}_{k+1}$ from $\mathfrak{P}_k$. We can therefore conclude that the set $\mathfrak{P} = \cup_k \mathfrak{P}_k$ of all poles of $\hat{u}_{n+1}$ is given by

$$\mathfrak{P} = \{z \in \mathbb{R} : z = \rho_d j - k\beta, j \in \mathbb{N}, k = 0, \dots, \lfloor \rho_d j/2 \rfloor\}$$

and this implies that $u_{n+1} \in \mathcal{E}_{\overline{\mathbf{p}},\overline{\mathbf{m}}}$.

We now prove that, when a polynomial nonlinearity is considered, the convergence of the sequence happens in higher order norms. We start by remarking that Theorem 12 directly implies $w \in \mathcal{J}^{\varpi}_{\gamma}(\Omega)$, for any $\gamma < d/2 + 2 - \beta$. In addition, if we inspect the proof of this statement from Section 7.3, we see that we have shown that, if u is an eigenfunction of the nonlinear Schrödinger equation,

$$\text{if } \|u\|_{\mathcal{J}^{s,p}_{\gamma}(\Omega)} \leq CA^s s!, \text{ for all } s = 0, \dots, k-1 \text{ then } \|u\|_{\mathcal{J}^{k,p}_{\gamma}(\Omega)} \leq CA^k k!$$

for all integer $k \geq 2$, $\gamma < d/p + 2 - \beta$, and sufficiently large p. The same proof can therefore be used to show that, for u_{n+1} and u_n as in (8.76),

$$\text{if } \begin{cases} \|u_{n+1}\|_{\mathcal{J}^{s,p}_{\gamma}(\Omega)} \leq CA^s s! \\ \|u_n\|_{\mathcal{J}^{s,p}_{\gamma}(\Omega)} \leq CA^s s! \end{cases} \text{ for all } s = 0, \dots, k-1 \text{ then } \|u_{n+1}\|_{\mathcal{J}^{k,p}_{\gamma}(\Omega)} \leq CA^k k! \quad (8.85)$$

for all integer $k \geq 2$, $\gamma < d/p + 2 - \beta$, and sufficiently large p. Furthermore, the uniform bound

$$\|u_n\|_{\mathcal{J}^2_{\gamma}(\Omega)} \leq C \quad (8.86)$$

holds for all n and with C independent of $n \in \mathbb{N}$. We wish to construct an inductive proof that

$$\|u_n\|_{\mathcal{J}^{k,p}_{\gamma}(\Omega)} \leq CA^k k! \quad (8.87)$$

for all $k \in \mathbb{N}$, sufficiently large p, $\gamma \leq \widetilde{\gamma} < d/p + 2 - \beta$, and with constants C, A independent of n. Assume we choose u_0 such that

$$\|u_0\|_{\mathcal{J}^{k,p}_{\gamma}(\Omega)} \leq C_0 A_0^k k!$$

for all $k \in \mathbb{N}$, sufficiently large p, and $\gamma \leq \widetilde{\gamma}$: then, (8.85) implies that (8.87) holds for u_1, with constants C, A that depend on the equation and on A_0, C_0. At this point, crucially, suppose (8.87) holds for u_n: using (8.85) and (8.86) we can show that (8.87) holds for u_{n+1} *with the same constants* as those used for u_n. This implies that $\{u_n\}$ is uniformly bounded in the $\mathcal{J}^k_\gamma(\Omega)$ norms, for $\gamma < d/2 + 2 - \beta$ and for all $k \in \mathbb{N}$.

We now consider that

$$\begin{aligned}\|u_{n_k} - w\|_{\mathcal{J}^{j+2}_\gamma(\Omega)} &\leq C\|A^w(u_{n_k} - w)\|_{\mathcal{J}^j_{\gamma-2}(\Omega)} \\ &\lesssim |\lambda_{n_k} - \mu|\|w\|_{\mathcal{J}^j_{\gamma-2}(\Omega)} + |\lambda_{n_k}|\|u_{n_k} - w\|_{\mathcal{J}^j_{\gamma-2}(\Omega)} + \|(u^2_{n_k-1} - w^2)u_{n_k}\|_{\mathcal{J}^j_{\gamma-2}(\Omega)}\end{aligned} \tag{8.88}$$

The first two terms at the right hand side can be shown to converge due to the convergence of the eigenvalue and using an induction hypothesis; we consider the last one. We have

$$\|(u^2_{n_k-1} - w^2)u_{n_k}\|_{\mathcal{J}^j_{\gamma-2}(\Omega)} \leq C\|(u_{n_k-1} + w)u_{n_k}\|_{\mathcal{J}^{j,\infty}_{\gamma-d/2}(\Omega)}\|u_{n_k-1} - w\|_{\mathcal{J}^j_\gamma(\Omega)}$$

Using an embedding inequality and the uniform boundedness of $\{u_n\}_n$ in the $\mathcal{J}^k_\gamma(\Omega)$ norms, then,

$$\|(u^2_{n_k-1} - w^2)u_{n_k}\|_{\mathcal{J}^j_{\gamma-2}(\Omega)} \lesssim \|u_{n_k-1} - w\|_{\mathcal{J}^j_\gamma(\Omega)}$$

Injecting this inequality into (8.88), using the convergence in $H^1(\Omega)$ as the basis of an inductive proof, we obtain that

$$u_{n_k} \xrightarrow{\mathcal{J}^j_\gamma(\Omega)} w,$$

for all j, and for any $\gamma < d/2 + 2 - \beta$. This implies that the limit w belongs to the same asymptotic class as the functions in the sequence $\{u_{n_k}\}_k$. We have thus proved the following proposition.

Proposition 48. *Let $\{u_n\}_n$ be the sequence generated by* (8.76)*, with $f(u^2) = u^2$, V defined as in* (8.84)*, and β such that conditions* (8.12a), (8.12b) *are satisfied. Then, if $u_0 \in \mathcal{J}^\varpi_{d/2+2-\beta}(\Omega) \cap \mathcal{E}_{\overline{\mathbf{p}},\overline{\mathbf{m}}}$,*

$$u_{n+1} \in \mathcal{J}^\varpi_\gamma(\Omega) \cap \mathcal{E}_{\overline{\mathbf{p}},\overline{\mathbf{m}}},$$

for all $n \in \mathbb{N}$, $\gamma < d/2 + 2 - \beta$. Furthermore, let w be the limit, up to extraction, of $\{u_n\}_n$. Then,

$$w \in \mathcal{J}^\varpi_{d/2+2-\beta}(\Omega) \cap \mathcal{E}_{\overline{\mathbf{p}},\overline{\mathbf{m}}},$$

Remark 12. *We have considered for simplicity the Gross-Pitaevskii equation with $f(u^2) = u^2$; nevertheless, the analysis can be extended to any function $f(u^2) = u^k$ with integer k such that* (8.11a)–(8.11d) *hold. As a matter of fact, the only additional requirements on f we introduced here are that*

$$v \in \mathcal{E}_{\overline{\mathbf{p}},\overline{\mathbf{m}}} \text{ implies } f(v^2) \in \mathcal{E}_{\overline{\mathbf{p}},\overline{\mathbf{m}}}.$$

and that the solution to the problem belongs to the weighted analyticity class $\mathcal{J}^\varpi_\gamma(\Omega)$.

Chapter 9

Numerical results for nonlinear eigenvalue problems

9.1 Nonlinear Schrödinger

We consider, in this section, the problem presented and analyzed in Chapter 8, from the numerical point of view. In its continuous form, the problem reads: find the eigenpair $(\lambda, u) \in \mathbb{R} \times H^1(\Omega)$ such that $\|u\|_{L^2(\Omega)} = 1$ and

$$\begin{aligned} -\Delta u + Vu + f(u^2)u &= \lambda u \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{9.1}$$

In particular, we focus on the computation of the *lowest eigenvalue* and of its associated eigenvector, corresponding, from a phisical point of view, with the ground state of the system. The domain is given by the d-dimensional cube of unitary edge $(-1/2, 1/2)^d$.

We take potentials of the form $V(x) = -r^{-\alpha}$, for $\alpha = -1/2, -1, -3/2$. The approximation is done as in the linear case: see Chapter 6 for the details on the computational mesh and on the space. We use a SIP method, and solve the nonlinearity by a fixed point method. The stopping criterion on the nonlinear iterations is residual based, i.e., we stop iterating when

$$\langle (A^{u_\delta} - \lambda_\delta) u_\delta, u_\delta \rangle \leq \varepsilon_{\text{tol}}$$

for a given computed solution $u_\delta \in X_\delta$ and a given tolerance ε_{tol} We will indicate the tolerance we use, on a case by case basis, in the following sections.

From the point of view of the analysis, we find results similar to those arising in the linear case: in particular, we see the effect of algebraic and quadrature errors, mainly in the context of the two dimensional approximation.

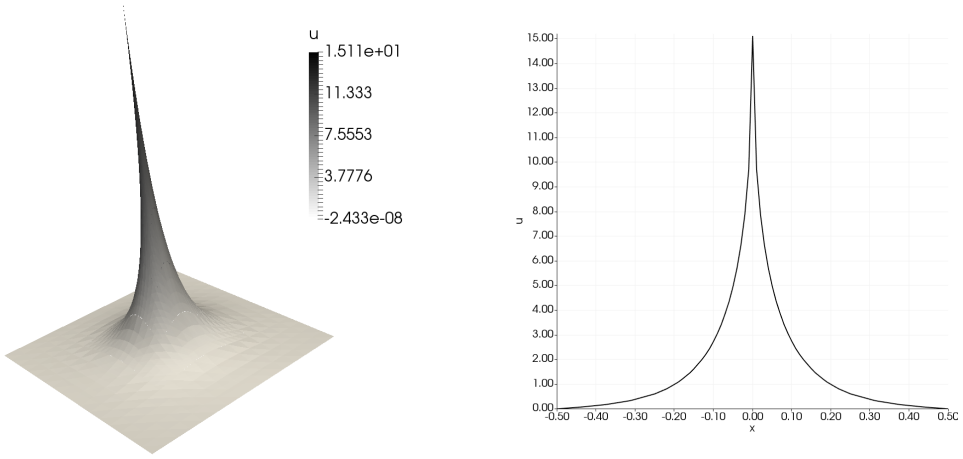


Figure 9.1 – Numerical solution to (9.1) with $V(x) = -r^{-3/2}$ and $f(u^2) = u^2$. Left: representation on the plane, vertically not to scale. Right: restriction of u to the line $\{y = 0\}$.

Table 9.1 – Estimated coefficients. Potential: $-r^{-1/2}$, $f(u^2) = u^2$

$\mathfrak{s}$	b_{L^2}	b_{DG}	b_{L^∞}	b_λ
0.125	0.72	0.73	0.74	1.24
0.25	0.92	0.94	0.94	1.5
0.5	1.06	1	0.98	1.25

Table 9.2 – Estimated coefficients. Potential: $-r^{-1}$, $f(u^2) = u^2$

$\mathfrak{s}$	b_{L^2}	b_{DG}	b_{L^∞}	b_λ
0.125	0.59	0.6	0.58	1.05
0.25	0.72	0.72	0.7	1.01
0.5	0.68	0.68	0.58	0.65

Table 9.3 – Estimated coefficients. Potential: $-r^{-3/2}$, $f(u^2) = u^2$

$\mathfrak{s}$	b_{L^2}	b_{DG}	b_{L^∞}	b_λ
0.062	0.43	0.45	0.5	0.8
0.125	0.56	0.52	0.65	0.76
0.25	0.48	0.47	0.43	0.47

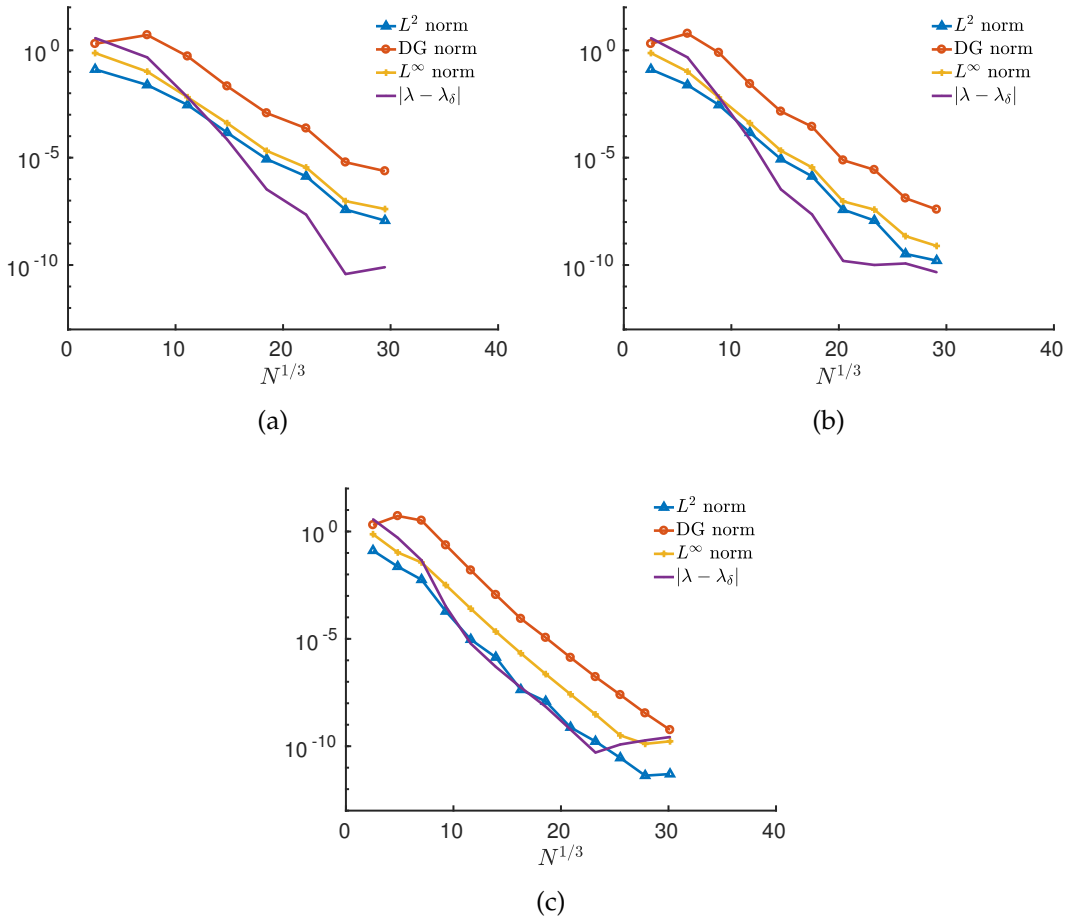


Figure 9.2 – Errors for the numerical solution with potential $V(x) = -r^{-1/2}$ and nonlinearity $f(u^2) = u^2$. Polynomial slope: $\mathfrak{s} = 1/8$ in Figure a; $\mathfrak{s} = 1/4$ in Figure b and $\mathfrak{s} = 1/2$ in Figure c.

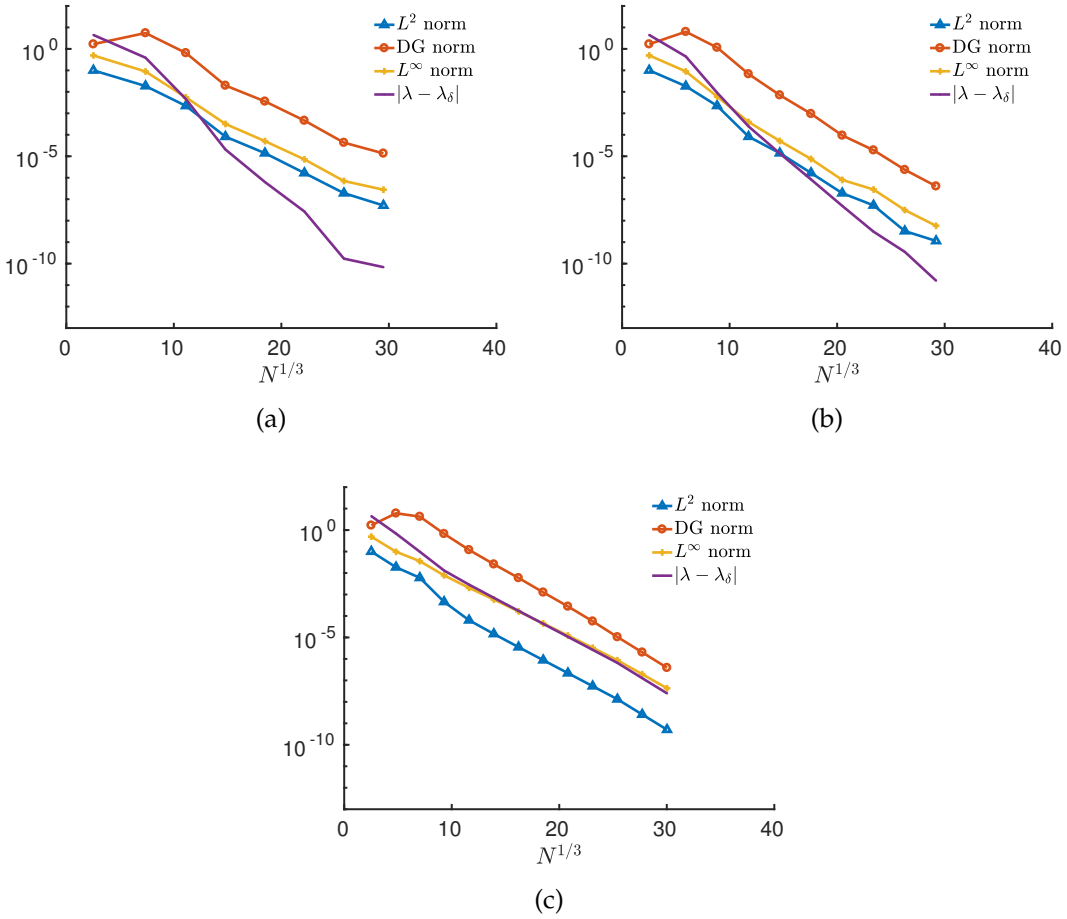


Figure 9.3 – Errors for the numerical solution with potential $V(x) = -r^{-1}$ and nonlinearity $f(u^2) = u^2$. Polynomial slope: $\mathfrak{s} = 1/8$ in Figure a; $\mathfrak{s} = 1/4$ in Figure b and $\mathfrak{s} = 1/2$ in Figure c.

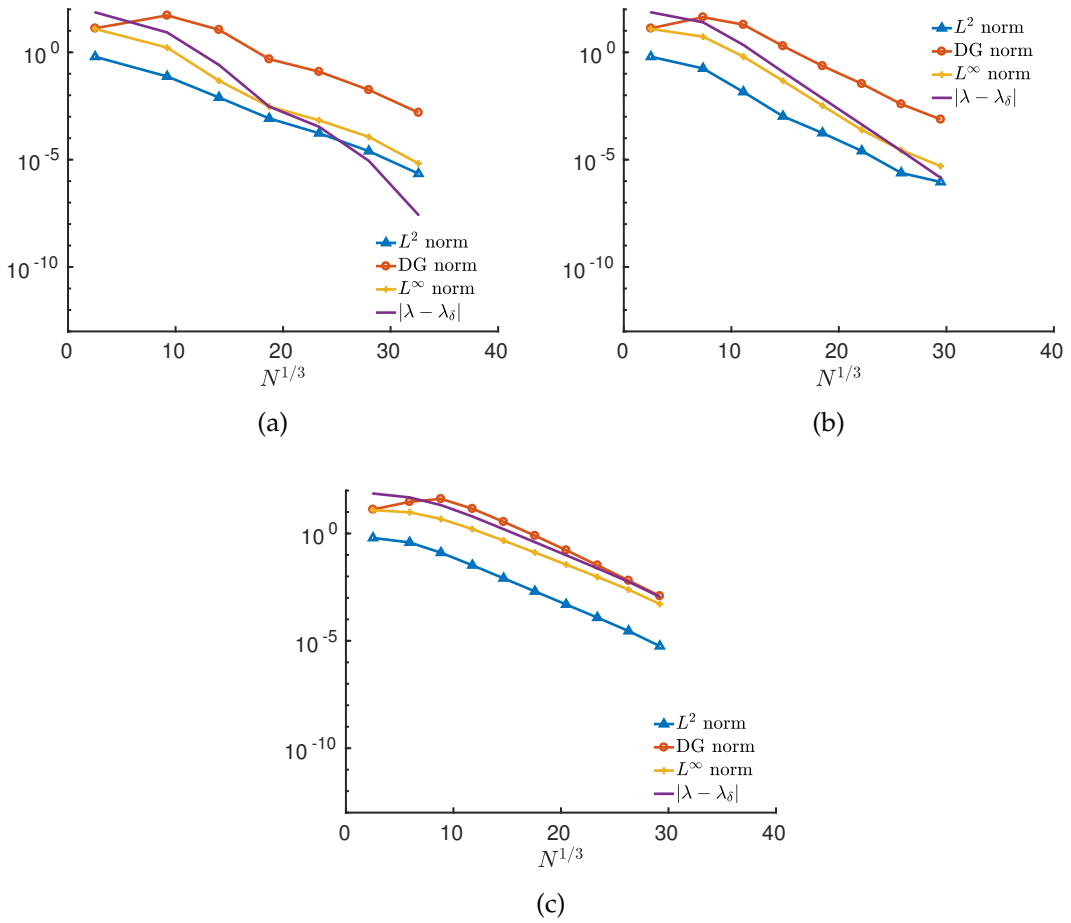


Figure 9.4 – Errors for the numerical solution with potential $V(x) = -r^{-3/2}$ and nonlinearity $f(u^2) = u^2$. Polynomial slope: $\mathfrak{s} = 1/16$ in Figure a; $\mathfrak{s} = 1/8$ in Figure b and $\mathfrak{s} = 1/4$ in Figure c.

9.1.1 Two dimensional case

In the two dimensional case, we compute the numerical solutions on meshes built as in the linear case, see Figure 6.1. A visualization of the solution (in the most singular problem we analyse) is given in Figure 9.1.

We consider mainly the case where $f(u^2) = u^2$, corresponding to the cubic nonlinearity of the Gross-Pitaevskii equation. Having set this nonlinearity, writing $V(x) = -r^{-\alpha}$, we plot the curves of the errors in Figures 9.2 ($\alpha = 1/2$), 9.3 ($\alpha = 1$), and 9.4 ($\alpha = 3/2$). The numerical approximation exhibits the same properties of the linear case and we the same phenomena arising. Namely, in the case of the approximations with low polynomial slopes, all errors converge exponentially in the number of refinement steps, with the eigenvalue error converging faster than the norms of the eigenfunction error. A plateau due to the algebraic error is evident around 10^{-10}. When the polynomial slopes are higher, the quadrature error manifests itself more strongly and causes, in extreme cases, the total loss of the doubling of the convergence rate.

The coefficients b_X, for $X = L^2(\Omega), \mathrm{DG}, L^\infty(\Omega)$ and λ are shown in Tables 9.1 to 9.3. As already outlined, the higher the slope, the biggest the quadrature error and the furthest the estimated coefficients b_λ is from the double of the one for the DG norm.

9.1.2 Detailed tables of the errors

In the following Tables 9.4 to 9.12 we give the computed values for the error on which the analysis in the preceding section is based.

Table 9.4 – Errors. Potential: $r^{-0.5}$, polynomial slope $\mathfrak{s} = 0.125$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.13	2.01	0.75	3.71
400	$2.43{\cdot}10^{-2}$	5.15	0.1	0.47
1,374	$2.86{\cdot}10^{-3}$	0.53	$6.53{\cdot}10^{-3}$	$6.55{\cdot}10^{-3}$
3,242	$1.46{\cdot}10^{-4}$	$2.13{\cdot}10^{-2}$	$4.05{\cdot}10^{-4}$	$7.02{\cdot}10^{-5}$
6,306	$8.54{\cdot}10^{-6}$	$1.19{\cdot}10^{-3}$	$2.08{\cdot}10^{-5}$	$3.34{\cdot}10^{-7}$
10,857	$1.35{\cdot}10^{-6}$	$2.34{\cdot}10^{-4}$	$3.5{\cdot}10^{-6}$	$2.22{\cdot}10^{-8}$
17,194	$3.81{\cdot}10^{-8}$	$6.1{\cdot}10^{-6}$	$9.48{\cdot}10^{-8}$	$3.79{\cdot}10^{-11}$
25,618	$1.18{\cdot}10^{-8}$	$2.37{\cdot}10^{-6}$	$4.04{\cdot}10^{-8}$	$7.93{\cdot}10^{-11}$

Table 9.5 – Errors. Potential: $r^{-0.5}$, polynomial slope $\mathfrak{s} = 0.25$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.13	2.01	0.75	3.71
208	$2.43{\cdot}10^{-2}$	6.03	0.1	0.47
690	$2.86{\cdot}10^{-3}$	0.78	$6.37{\cdot}10^{-3}$	$6.77{\cdot}10^{-3}$
1,612	$1.46{\cdot}10^{-4}$	$2.73{\cdot}10^{-2}$	$4.05{\cdot}10^{-4}$	$7.09{\cdot}10^{-5}$
3,118	$8.54{\cdot}10^{-6}$	$1.43{\cdot}10^{-3}$	$2.12{\cdot}10^{-5}$	$3.34{\cdot}10^{-7}$
5,362	$1.35{\cdot}10^{-6}$	$2.78{\cdot}10^{-4}$	$3.5{\cdot}10^{-6}$	$2.27{\cdot}10^{-8}$
8,487	$3.81{\cdot}10^{-8}$	$7.6{\cdot}10^{-6}$	$9.19{\cdot}10^{-8}$	$1.55{\cdot}10^{-10}$
12,649	$1.18{\cdot}10^{-8}$	$2.73{\cdot}10^{-6}$	$3.82{\cdot}10^{-8}$	$1.01{\cdot}10^{-10}$
17,983	$3.28{\cdot}10^{-10}$	$1.29{\cdot}10^{-7}$	$2.22{\cdot}10^{-9}$	$1.18{\cdot}10^{-10}$
24,633	$1.61{\cdot}10^{-10}$	$3.88{\cdot}10^{-8}$	$7.73{\cdot}10^{-10}$	$4.62{\cdot}10^{-11}$

Table 9.6 – Errors. Potential: $r^{-0.5}$, polynomial slope $\mathfrak{s} = 0.5$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.13	2.01	0.75	3.71
112	$2.32{\cdot}10^{-2}$	5.3	0.11	0.51
348	$5.7{\cdot}10^{-3}$	3.26	$3.61{\cdot}10^{-2}$	$4.56{\cdot}10^{-2}$
802	$1.92{\cdot}10^{-4}$	0.23	$3.2{\cdot}10^{-3}$	$3.22{\cdot}10^{-4}$
1,566	$9.32{\cdot}10^{-6}$	$1.57{\cdot}10^{-2}$	$2.53{\cdot}10^{-4}$	$5.99{\cdot}10^{-6}$
2,710	$1.38{\cdot}10^{-6}$	$1.13{\cdot}10^{-3}$	$2.18{\cdot}10^{-5}$	$4.96{\cdot}10^{-7}$
4,305	$4.29{\cdot}10^{-8}$	$8.73{\cdot}10^{-5}$	$2.13{\cdot}10^{-6}$	$5.46{\cdot}10^{-8}$
6,415	$1.22{\cdot}10^{-8}$	$1.14{\cdot}10^{-5}$	$2.28{\cdot}10^{-7}$	$6.93{\cdot}10^{-9}$
9,120	$7.51{\cdot}10^{-10}$	$1.34{\cdot}10^{-6}$	$2.61{\cdot}10^{-8}$	$5.95{\cdot}10^{-10}$
12,502	$1.67{\cdot}10^{-10}$	$1.69{\cdot}10^{-7}$	$3.04{\cdot}10^{-9}$	$5.07{\cdot}10^{-11}$
16,619	$2.85{\cdot}10^{-11}$	$2.47{\cdot}10^{-8}$	$3.17{\cdot}10^{-10}$	$1.2{\cdot}10^{-10}$
21,558	$4.26{\cdot}10^{-12}$	$3.46{\cdot}10^{-9}$	$1.29{\cdot}10^{-10}$	$1.87{\cdot}10^{-10}$
27,386	$5.05{\cdot}10^{-12}$	$5.76{\cdot}10^{-10}$	$1.69{\cdot}10^{-10}$	$2.67{\cdot}10^{-10}$

Table 9.7 – Errors. Potential: r^{-1}, polynomial slope $\mathfrak{s} = 0.125$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.1	1.67	0.5	4.46
400	$1.86{\cdot}10^{-2}$	5.48	$8.87{\cdot}10^{-2}$	0.39
1,374	$2.22{\cdot}10^{-3}$	0.65	$5.56{\cdot}10^{-3}$	$4.63{\cdot}10^{-3}$
3,242	$8.17{\cdot}10^{-5}$	$1.97{\cdot}10^{-2}$	$3.2{\cdot}10^{-4}$	$2.06{\cdot}10^{-5}$
6,306	$1.39{\cdot}10^{-5}$	$3.6{\cdot}10^{-3}$	$5.07{\cdot}10^{-5}$	$6.29{\cdot}10^{-7}$
10,857	$1.66{\cdot}10^{-6}$	$4.61{\cdot}10^{-4}$	$7.16{\cdot}10^{-6}$	$2.68{\cdot}10^{-8}$
17,214	$1.91{\cdot}10^{-7}$	$4.32{\cdot}10^{-5}$	$7.04{\cdot}10^{-7}$	$1.7{\cdot}10^{-10}$
25,661	$5.11{\cdot}10^{-8}$	$1.36{\cdot}10^{-5}$	$2.77{\cdot}10^{-7}$	$6.86{\cdot}10^{-11}$

Table 9.8 – Errors. Potential: r^{-1}, polynomial slope $\mathfrak{s} = 0.25$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.1	1.67	0.5	4.46
208	$1.83{\cdot}10^{-2}$	6.31	$8.97{\cdot}10^{-2}$	0.44
690	$2.24{\cdot}10^{-3}$	1.16	$6.06{\cdot}10^{-3}$	$9.12{\cdot}10^{-3}$
1,627	$8.22{\cdot}10^{-5}$	$6.78{\cdot}10^{-2}$	$3.95{\cdot}10^{-4}$	$2.45{\cdot}10^{-4}$
3,154	$1.39{\cdot}10^{-5}$	$7.05{\cdot}10^{-3}$	$5.17{\cdot}10^{-5}$	$1.35{\cdot}10^{-5}$
5,425	$1.66{\cdot}10^{-6}$	$9.52{\cdot}10^{-4}$	$7.47{\cdot}10^{-6}$	$8.32{\cdot}10^{-7}$
8,583	$1.91{\cdot}10^{-7}$	$9.37{\cdot}10^{-5}$	$7.91{\cdot}10^{-7}$	$4.89{\cdot}10^{-8}$
12,784	$5.17{\cdot}10^{-8}$	$1.93{\cdot}10^{-5}$	$2.83{\cdot}10^{-7}$	$3.08{\cdot}10^{-9}$
18,178	$3.27{\cdot}10^{-9}$	$2.35{\cdot}10^{-6}$	$3.13{\cdot}10^{-8}$	$3.52{\cdot}10^{-10}$
24,900	$1.14{\cdot}10^{-9}$	$4.03{\cdot}10^{-7}$	$5.77{\cdot}10^{-9}$	$1.6{\cdot}10^{-11}$

Table 9.9 – Errors. Potential: r^{-1}, polynomial slope $\mathfrak{s} = 0.5$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.1	1.67	0.5	4.46
112	$1.86{\cdot}10^{-2}$	6.07	$9.8{\cdot}10^{-2}$	0.69
348	$5.99{\cdot}10^{-3}$	4.24	$3.56{\cdot}10^{-2}$	$9.79{\cdot}10^{-2}$
802	$4.55{\cdot}10^{-4}$	0.67	$7.76{\cdot}10^{-3}$	$1.31{\cdot}10^{-2}$
1,566	$6.33{\cdot}10^{-5}$	0.12	$2.04{\cdot}10^{-3}$	$2.87{\cdot}10^{-3}$
2,695	$1.46{\cdot}10^{-5}$	$2.58{\cdot}10^{-2}$	$5.85{\cdot}10^{-4}$	$7.02{\cdot}10^{-4}$
4,264	$3.54{\cdot}10^{-6}$	$5.91{\cdot}10^{-3}$	$1.63{\cdot}10^{-4}$	$1.73{\cdot}10^{-4}$
6,345	$8.76{\cdot}10^{-7}$	$1.26{\cdot}10^{-3}$	$4.49{\cdot}10^{-5}$	$4.3{\cdot}10^{-5}$
9,010	$2.18{\cdot}10^{-7}$	$2.78{\cdot}10^{-4}$	$1.23{\cdot}10^{-5}$	$1.07{\cdot}10^{-5}$
12,341	$5.4{\cdot}10^{-8}$	$5.67{\cdot}10^{-5}$	$3.3{\cdot}10^{-6}$	$2.65{\cdot}10^{-6}$
16,409	$1.33{\cdot}10^{-8}$	$1.05{\cdot}10^{-5}$	$8.66{\cdot}10^{-7}$	$6.52{\cdot}10^{-7}$
21,293	$2.58{\cdot}10^{-9}$	$2.03{\cdot}10^{-6}$	$1.94{\cdot}10^{-7}$	$1.27{\cdot}10^{-7}$
27,060	$5.07{\cdot}10^{-10}$	$3.89{\cdot}10^{-7}$	$4.3{\cdot}10^{-8}$	$2.49{\cdot}10^{-8}$

Table 9.10 – Errors. Potential: $r^{-1.5}$, polynomial slope $\mathfrak{s} = 0.0625$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.62	13.1	12.2	72.7
400	0.18	43.3	5.35	24.9
1,374	$1.42{\cdot}10^{-2}$	19.4	0.63	2.2
3,242	$1.04{\cdot}10^{-3}$	1.96	$4.68{\cdot}10^{-2}$	0.12
6,306	$1.74{\cdot}10^{-4}$	0.23	$3.27{\cdot}10^{-3}$	$7.2{\cdot}10^{-3}$
10,842	$2.54{\cdot}10^{-5}$	$3.42{\cdot}10^{-2}$	$2.4{\cdot}10^{-4}$	$4.33{\cdot}10^{-4}$
17,173	$2.4{\cdot}10^{-6}$	$3.85{\cdot}10^{-3}$	$2.77{\cdot}10^{-5}$	$2.57{\cdot}10^{-5}$
25,591	$9.12{\cdot}10^{-7}$	$7.52{\cdot}10^{-4}$	$5.02{\cdot}10^{-6}$	$1.42{\cdot}10^{-6}$

Table 9.11 – Errors. Potential: $r^{-1.5}$, polynomial slope $\mathfrak{s} = 0.125$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.62	13.1	12.2	72.7
400	0.18	43.3	5.35	24.9
1,374	$1.42{\cdot}10^{-2}$	19.4	0.63	2.2
3,242	$1.04{\cdot}10^{-3}$	1.96	$4.68{\cdot}10^{-2}$	0.12
6,306	$1.74{\cdot}10^{-4}$	0.23	$3.27{\cdot}10^{-3}$	$7.2{\cdot}10^{-3}$
10,842	$2.54{\cdot}10^{-5}$	$3.42{\cdot}10^{-2}$	$2.4{\cdot}10^{-4}$	$4.33{\cdot}10^{-4}$
17,173	$2.4{\cdot}10^{-6}$	$3.85{\cdot}10^{-3}$	$2.77{\cdot}10^{-5}$	$2.57{\cdot}10^{-5}$
25,591	$9.12{\cdot}10^{-7}$	$7.52{\cdot}10^{-4}$	$5.02{\cdot}10^{-6}$	$1.42{\cdot}10^{-6}$

Table 9.12 – Errors. Potential: $r^{-1.5}$, polynomial slope $\mathfrak{s} = 0.25$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
16	0.62	13.1	12.2	72.7
208	0.38	29.8	9.71	47.8
690	0.13	41	4.74	21
1,627	$3.3{\cdot}10^{-2}$	14.3	1.6	6.2
3,154	$8.18{\cdot}10^{-3}$	3.5	0.47	1.59
5,430	$2.01{\cdot}10^{-3}$	0.79	0.13	0.4
8,590	$4.95{\cdot}10^{-4}$	0.17	$3.57{\cdot}10^{-2}$	$9.74{\cdot}10^{-2}$
12,793	$1.21{\cdot}10^{-4}$	$3.32{\cdot}10^{-2}$	$9.53{\cdot}10^{-3}$	$2.38{\cdot}10^{-2}$
18,174	$2.85{\cdot}10^{-5}$	$6.44{\cdot}10^{-3}$	$2.43{\cdot}10^{-3}$	$5.62{\cdot}10^{-3}$
24,877	$5.67{\cdot}10^{-6}$	$1.23{\cdot}10^{-3}$	$5.16{\cdot}10^{-4}$	$1.12{\cdot}10^{-3}$

9.1.3 Three dimensional problem

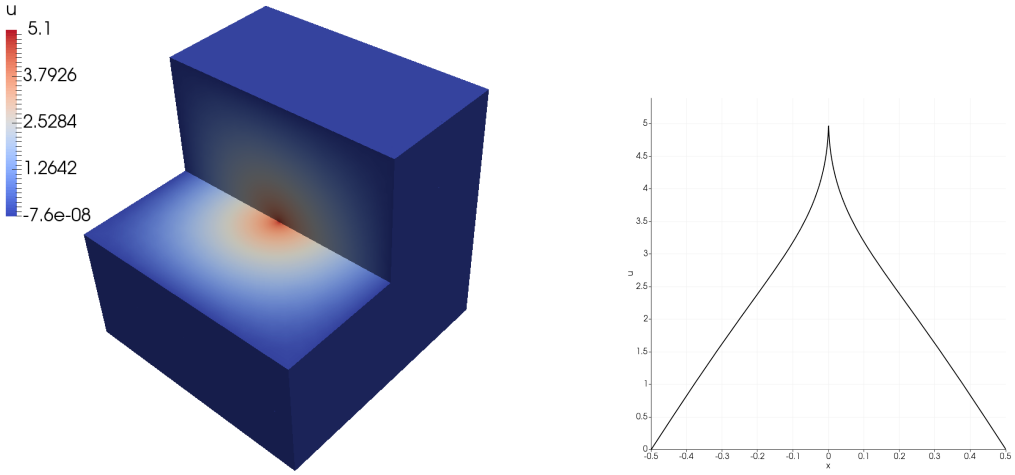


Figure 9.5 – Numerical solution in the three dimensional case: solution in the cube, left, and close up near the origin of the restriction to the nonline $\{y = z = 0\}$, right

In the three dimensional setting, we consider the domain $(-1/2, 1/2)^3$, and a mesh exemplified in Figure 6.7. The numerical solution of the problem with $V(x) = r^{-3/2}$ and $f(u^2) = u^2$ is shown in Figure 9.5. The solution shown is obtained at one of the highest degrees of refinement. As in the linear case, the algebraic eigenproblem solver uses the Jacobi-Davidson method [SV96], with a biconjugate gradient method [Vor92; SVF94] as the linear algebraic system solver. The fixed point nonlinear iteration are set to a tolerance $\varepsilon_{\text{tol}} = 10^{-7}$; everything else is left unchanged from the linear case.

The results we obtain are perfectly in line with those presented in Chapter 6. Therefore, the same conclusions can be drawn: the algebraic and quadrature error are not as evident as in the two dimensional case, and it can clearly be seen that an optimal slope can be chose to better approximate the eigenvalue. The nonlinearity does not seem to influence the rate of convergence; this is expected, since the source of the loss of regularity — the factor that most influences the rate of convergence — is primarily due to the potential.

Table 9.13 – Estimated coefficients. Potential: $r^{-1/2}$, $f(u^2) = u^2$

$\mathfrak{s}$	b_{L^2}	b_{DG}	b_{L^∞}	b_λ
0.125	0.73	0.74	0.81	1.28
0.25	0.82	0.82	0.85	1.3

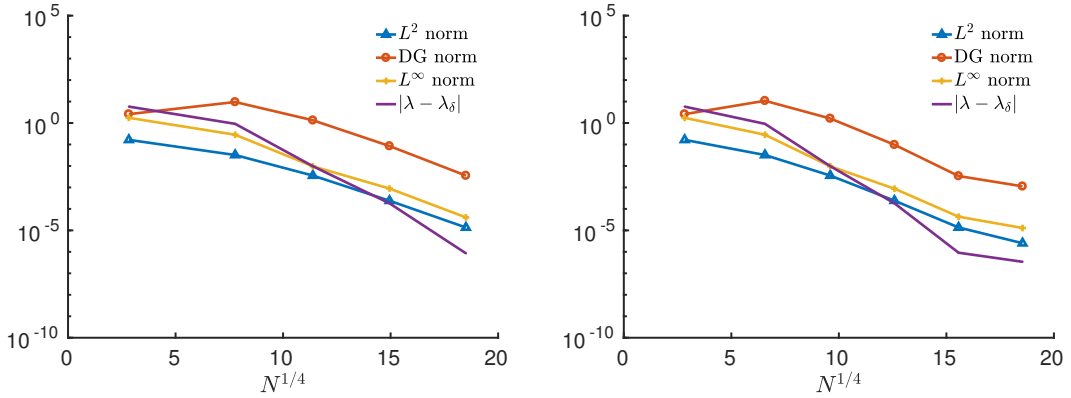


Figure 9.6 – Errors of the numerical solution for $V(x) = r^{-1/2}$, $f(u^2) = u^2$. Polynomial slope $\mathfrak{s} = 1/8$, left and $\mathfrak{s} = 1/4$, right.

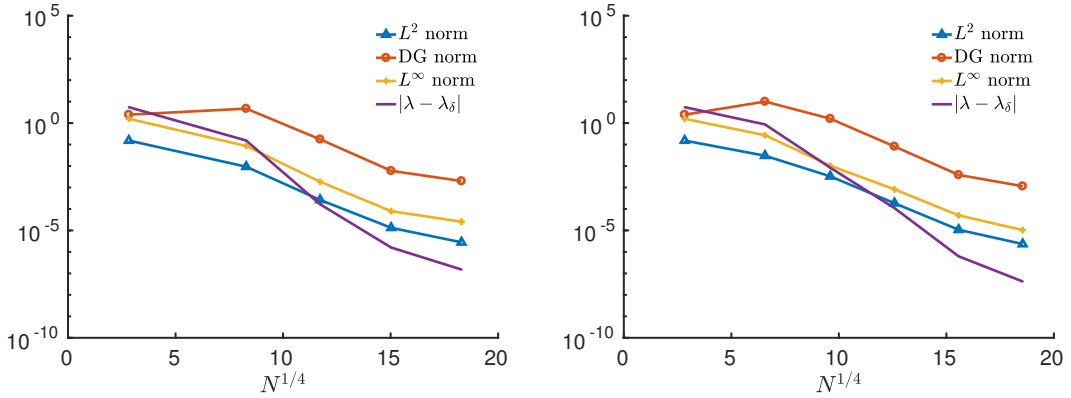


Figure 9.7 – Errors of the numerical solution for $V(x) = r^{-1}$, $f(u^2) = u^2$. Polynomial slope $\mathfrak{s} = 1/8$, left and $\mathfrak{s} = 1/4$, right.

Table 9.14 – Estimated coefficients. Potential: r^{-1}, $f(u^2) = u^2$

$\mathfrak{s}$	b_{L^2}	b_{DG}	b_{L^∞}	b_λ
0.125	0.82	0.8	0.83	1.39
0.25	0.82	0.81	0.86	1.44

Table 9.15 – Estimated coefficients. Potential: $r^{-3/2}$, $f(u^2) = u^2$

$\mathfrak{s}$	b_{L^2}	b_{DG}	b_{L^∞}	b_λ
0.125	0.69	0.67	0.71	1.29
0.25	0.8	0.73	0.52	1.3

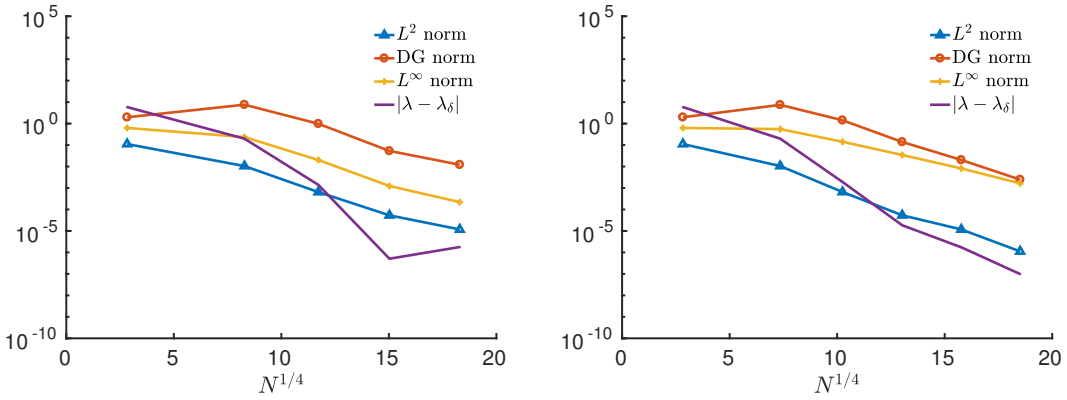


Figure 9.8 – Errors of the numerical solution for $V(x) = r^{-3/2}$, $f(u^2) = u^2$. Polynomial slope $\mathfrak{s} = 1/8$, left and $\mathfrak{s} = 1/4$, right.

9.1.4 Detailed error tables

We give here the detailed errors, in Tables 9.16 to 9.21. It is interesting to compare the figures shown here with those given in Tables 6.17 to 6.22, given in Section 6.2.2.

Table 9.16 – Errors. Potential: $r^{-0.5}$, polynomial slope $\mathfrak{s} = 0.125$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
64	0.16	2.59	1.75	5.81
3,648	$3.23{\cdot}10^{-2}$	9.52	0.29	0.92
16,808	$3.57{\cdot}10^{-3}$	1.34	$9.63{\cdot}10^{-3}$	$9.74{\cdot}10^{-3}$
49,947	$2.4{\cdot}10^{-4}$	$8.47{\cdot}10^{-2}$	$8.87{\cdot}10^{-4}$	$1.82{\cdot}10^{-4}$
$1.17 \cdot 10^5$	$1.36{\cdot}10^{-5}$	$3.56{\cdot}10^{-3}$	$4.02{\cdot}10^{-5}$	$8.53{\cdot}10^{-7}$

Table 9.17 – Errors. Potential: $r^{-0.5}$, polynomial slope $\mathfrak{s} = 0.25$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
64	0.16	2.59	1.75	5.81
1,856	$3.23{\cdot}10^{-2}$	10.9	0.29	0.92
8,493	$3.57{\cdot}10^{-3}$	1.62	$9.64{\cdot}10^{-3}$	$9.77{\cdot}10^{-3}$
25,088	$2.41{\cdot}10^{-4}$	$9.72{\cdot}10^{-2}$	$8.75{\cdot}10^{-4}$	$1.82{\cdot}10^{-4}$
58,733	$1.37{\cdot}10^{-5}$	$3.4{\cdot}10^{-3}$	$4.34{\cdot}10^{-5}$	$9.11{\cdot}10^{-7}$
$1.18 \cdot 10^5$	$2.52{\cdot}10^{-6}$	$1.13{\cdot}10^{-3}$	$1.29{\cdot}10^{-5}$	$3.44{\cdot}10^{-7}$

Table 9.18 – Errors. Potential: r^{-1}, polynomial slope $\mathfrak{s} = 0.125$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
64	0.15	2.46	1.58	5.51
4,712	$9.28{\cdot}10^{-3}$	4.72	$8.59{\cdot}10^{-2}$	0.15
18,880	$2.64{\cdot}10^{-4}$	0.18	$1.85{\cdot}10^{-3}$	$1.66{\cdot}10^{-4}$
50,968	$1.33{\cdot}10^{-5}$	$6{\cdot}10^{-3}$	$7.89{\cdot}10^{-5}$	$1.6{\cdot}10^{-6}$
$1.12 \cdot 10^5$	$2.8{\cdot}10^{-6}$	$2{\cdot}10^{-3}$	$2.55{\cdot}10^{-5}$	$1.51{\cdot}10^{-7}$

Table 9.19 – Errors. Potential: r^{-1}, polynomial slope $\mathfrak{s} = 0.25$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
64	0.15	2.46	1.58	5.51
1,856	$3.02{\cdot}10^{-2}$	10.1	0.27	0.87
8,493	$3.3{\cdot}10^{-3}$	1.6	$1.03{\cdot}10^{-2}$	$8.35{\cdot}10^{-3}$
25,088	$1.85{\cdot}10^{-4}$	$8.16{\cdot}10^{-2}$	$8.21{\cdot}10^{-4}$	$1.1{\cdot}10^{-4}$
58,733	$1.08{\cdot}10^{-5}$	$3.87{\cdot}10^{-3}$	$5.04{\cdot}10^{-5}$	$6.25{\cdot}10^{-7}$
$1.18 \cdot 10^5$	$2.33{\cdot}10^{-6}$	$1.15{\cdot}10^{-3}$	$1.04{\cdot}10^{-5}$	$4.21{\cdot}10^{-8}$

Table 9.20 – Errors. Potential: $r^{-1.5}$, polynomial slope $\mathfrak{s} = 0.125$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
64	0.11	1.97	0.63	5.88
4,712	$1.05{\cdot}10^{-2}$	7.58	0.24	0.2
18,880	$6.47{\cdot}10^{-4}$	0.98	$2.01{\cdot}10^{-2}$	$1.4{\cdot}10^{-3}$
50,968	$5.35{\cdot}10^{-5}$	$5.33{\cdot}10^{-2}$	$1.24{\cdot}10^{-3}$	$5.09{\cdot}10^{-7}$
$1.12 \cdot 10^5$	$1.18{\cdot}10^{-5}$	$1.21{\cdot}10^{-2}$	$2.21{\cdot}10^{-4}$	$1.79{\cdot}10^{-6}$

Table 9.21 – Errors. Potential: $r^{-1.5}$, polynomial slope $\mathfrak{s} = 0.25$, $p_0 = 1$

Ndof	$\|u - u_\delta\|_{L^2(\Omega)}$	$\|u - u_\delta\|_{\mathrm{DG}}$	$\|u - u_\delta\|_{L^\infty(\Omega)}$	$\|\lambda - \lambda_\delta\|$
64	0.11	1.97	0.63	5.88
2,920	$1.06{\cdot}10^{-2}$	7.44	0.55	0.2
11,040	$6.52{\cdot}10^{-4}$	1.43	0.14	$1.91{\cdot}10^{-3}$
28,792	$5.36{\cdot}10^{-5}$	0.14	$3.45{\cdot}10^{-2}$	$1.84{\cdot}10^{-5}$
61,888	$1.18{\cdot}10^{-5}$	$2.03{\cdot}10^{-2}$	$8.16{\cdot}10^{-3}$	$1.73{\cdot}10^{-6}$
$1.17 \cdot 10^5$	$1.12{\cdot}10^{-6}$	$2.46{\cdot}10^{-3}$	$1.64{\cdot}10^{-3}$	$9.83{\cdot}10^{-8}$

9.2 Hartree Fock equation

We now discuss briefly the implementation details of the numerical approximation of the Hartree Fock equation. We recall that the continuous problem consists in minimising the energy

$$\inf \left\{ E^{\mathrm{HF}}(\psi_1, \dots, \psi_N), \psi_i \in H^1(\mathbb{R}^3; \mathbb{R}) : \int_{\mathbb{R}^3} \psi_i \psi_j = \delta_{ij} \right\},$$

with

$$E^{\mathrm{HF}}(\psi_1, \dots, \psi_{N_e}) = \sum_{i=1}^{N_e} \int_{\mathbb{R}^3} |\nabla \psi_i|^2 + \int_{\mathbb{R}^3} V \rho_\Psi + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho_\Psi(x)\rho_\Psi(y)}{|x-y|} - \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\tau_\Psi(x,y)}{|x-y|}.$$

We have written

$$\tau_\Psi(x,y) = \sum_{i=1}^{N_e} \psi_i(x)\psi_i(y), \qquad \rho_\Psi = \tau_\Psi(x,x)$$

and

$$V(x) = \sum_{k=1}^{M} \frac{Z_k}{|x - R_k|}.$$

In practice, this is done by looking for N_e functions $\Psi = \{\psi_i\}$, $i = 1, \dots, N_e$, such that

$$\mathcal{F}_\Psi \psi_i = -\frac{1}{2}\Delta\psi_i + V\psi_i + \left(\rho_\Psi \star \frac{1}{|x|}\right)\psi_i - \sum_{j=1}^{N_e} \left((\psi_j\psi_i) \star \frac{1}{|x|}\right)\psi_j = \lambda_i \psi_i, \tag{9.2}$$

where the λ_i are the lowest eigenvalues of the Fock operator.

9.2.1 Practical implementation and the issue of sparsity

Since problem (9.2) is nonlinear, an iterative scheme will be used to numerically compute its solution. Different methods are used and have been analysed: see, for example, the Roothaan and level-shifting algorithms, analysed in [CL02], and the direct inversion in the iterative subspace method (DIIS) [Pul80].

In the sequel, we restict ourselves to closed shell restricted Hartree-Fock approximations: write $N_c = N_e/2$ and consider the problem of finding $\Psi = \{\psi_i\}_{i=1}^{N_c}$ such that

$$\mathcal{F}^{\mathrm{R}}_\Psi \psi_i = -\frac{1}{2}\Delta\psi_i + V\psi_i + 2\left(\rho_\Psi \star \frac{1}{|x|}\right)\psi_i - \sum_{j=1}^{N_c} \left((\psi_j\psi_i) \star \frac{1}{|x|}\right)\psi_j = \lambda_i \psi_i,$$

for $i = 1, \dots, N_c$ and where the λ_i are the N_c lowest eigenvalues of the operator. See

[SO12] for a derivation of this equation. We will present the basic ideas of the implementation of the finite element approximation of the Hartree-Fock equation in the context of the Roothaan alogrithm, in which the nonlinearity is approximated by fixed point interations: given $\Psi^{\text{old}} = \{\psi_1^{\text{old}}, \dots, \psi_{N_c}^{\text{old}}\}$, find the lowest N_c eigenvalues $\lambda_i, \dots, \lambda_{N_c}$ and the associated eigenfunctions $\psi_1, \dots, \psi_{N_c}$ such that

$$\mathcal{F}_{\Psi^{\text{old}}}\psi_i = -\frac{1}{2}\Delta\psi_i + V\psi_i + 2\left(\rho_{\Psi^{\text{old}}} \star \frac{1}{|x|}\right)\psi_i - \sum_{j=1}^{N_c}\left((\psi_j\psi_i^{\text{old}}) \star \frac{1}{|x|}\right)\psi_j = \lambda_i\psi_i. \quad (9.3)$$

The equation is set in $\mathbb{R}^3$. In practice, we will have to consider a bounded domain $\Omega \subset \mathbb{R}^3$, on which we construct the usual graded mesh $\mathcal{T}$. This implies that we also have to impose some artificial boundary conditions on $\partial\Omega$: those will be either homogeneous Dirichlet or Neumann boundary conditions; the numerical investigation of the effect of this choice is still ongoing.

We introduce the bilinear form

$$h_\delta(u_\delta, v_\delta) = \sum_{K\in\mathcal{T}}(\nabla u_\delta, \nabla v_\delta)_K + (Vu_\delta, v_\delta)_K - \sum_{e\in\mathcal{E}_I}(\{\!\{\nabla u_\delta\}\!\}, [\![v_\delta]\!])_e - \sum_{e\in\mathcal{E}}(\{\!\{\nabla v_\delta\}\!\}, [\![u_\delta]\!])_e + \sum_{e\in\mathcal{E}}\alpha_e\frac{\mathtt{p_e}^2}{\mathtt{h_e}}([\![u_\delta]\!], [\![v_\delta]\!])_e.$$

The Galerkin formulation of (9.3) thus reads: given $\Psi_\delta^{\text{old}} = (\psi_{\delta 1}^{\text{old}}, \dots, \psi_{\delta N_c}^{\text{old}}) \in {X_\delta}^{Ne}$,

find $\Psi_\delta = \{\psi_{\delta i}, , \lambda_{\delta i}\}_i \in (X_\delta \times \mathbb{R})^{N_c}$, such that

$$F_\delta(\psi_{\delta i}, v_\delta) = h_\delta(\psi_{\delta i}, v_\delta) + 2\int_\Omega\left(\rho_{\Psi_\delta^{\text{old}}} \star \frac{1}{|x|}\right)\psi_{\delta i}v_\delta - \sum_{j=1}^{N_c}\int_\Omega\left((\psi_{\delta i}\psi_{\delta j}^{\text{old}}) \star \frac{1}{|x|}\right)\psi_{\delta j}^{\text{old}}v_\delta = \lambda_{\delta i}\int_\Omega\psi_{\delta i}v_\delta, \quad (9.4)$$

for all $v_\delta \in X_\delta$.

The nonlocal terms need a special consideration: a naive implementation of those terms would give rise to a dense matrix, which is practically untreatable in the context of a finite element implementation. Let us then introduce a basis for the discrete space $\{\chi_1, \dots, \chi_N\}$ such that

$$X_\delta = \text{span}(\chi_1, \dots, \chi_N).$$

The linear part of the operator can be treated in the classical way: we introduce the matrix $\mathbf{h} \in \mathbb{R}^{N\times N}$ such that

$$\mathbf{h}_{ij} = h_\delta(\chi_i, \chi_j).$$

Consider now the Hartree potential term in (9.3). If we write

$$U(x) = \int_{\mathbb{R}^3} \frac{\rho_{\Psi^{\text{old}}}(y)}{|x-y|} dy,$$

then

$$-\Delta U = 4\pi \rho_{\Psi^{\text{old}}} \text{ in } \mathbb{R}^3. \tag{9.5}$$

We introduce the bilinear form associated with the Poisson problem

$$l_\delta(u_\delta, v_\delta) = \frac{1}{4\pi} \left(\sum_{K\in\mathcal{T}} (\nabla u_\delta, \nabla v_\delta)_K - \sum_{e\in\mathcal{E}_I} (\{\!\{\nabla u_\delta\}\!\}, [\![v_\delta]\!])_e - \sum_{e\in\mathcal{E}} (\{\!\{\nabla v_\delta\}\!\}, [\![u_\delta]\!])_e \right.$$
$$\left. + \sum_{e\in\mathcal{E}} \alpha_e \frac{\mathtt{p_e}^2}{\mathtt{h_e}} ([\![u_\delta]\!], [\![v_\delta]\!])_e \right)$$

and define $U_\delta \in X_\delta$ as the finite element function such that

$$l_\delta(U_\delta, v_\delta) = \int_\Omega \rho_{\Psi^{\text{old}}} v_\delta, \tag{9.6}$$

for all $v_\delta \in X_\delta$ and with either Dirichlet or Neumann boundary conditions on $\partial\Omega$. The linear and Poisson potential terms can therefore be explicitely written as sparse matrices. Note that we do not claim that these matrices need to be explicitely assembled, especially in parts of the domain where a high polynomial order is used.

To treat the exchange term, let us define the matrix $\mathbf{L}$ such that

$$\mathbf{L}_{ij} = l_\delta(\chi_i, \chi_j).$$

and the matrices $\mathbf{M}^n$, $n = 1, \ldots, N_c$, with elements

$$\mathbf{M}^n_{ij} = \int_{\mathbb{R}^d} \psi_n^{\text{old}} \chi_i \chi_j.$$

Write then, for all $n = 1, \ldots, N_c$,

$$\psi_{\delta n} = \sum_{j=1}^{N} c_{nj} \chi_j.$$

Using the same approximation of the convolution term used for the Hartree potential, and denoting $\mathbf{c}_n = (c_{n1}, \ldots, c_{nN})^T$, we obtain

$$\sum_{n=1}^{N_c} \int_\Omega \left((\psi_{\delta j} \psi_{\delta n}^{\text{old}}) \star \frac{1}{|x|} \right) \psi_{\delta n}^{\text{old}} \chi_i \approx \sum_{n=1}^{N_c} \left[\mathbf{M}^n \mathbf{L}^{-1} \mathbf{M}^n \mathbf{c}_j \right]_i.$$

The computation of the exchange term can therefore be written as two matrix vector product and a solution of a linear system, and this preserves the sparsity of the problem. In practice, since this computation happens at the innermost loop, one has to factorize the matrix $\mathbf{L}$. This can be done in parallel, see for example the MUMPS solver [Ame+01]. Furthermore, the matrix $\mathbf{L}$ is does not depend on the fixed point step, so the factorization can be done only once.

To summarize, denote by $\mathbf{M}$ the mass matrix, i.e.,

$$\mathbf{M}_{ij} = \int_\Omega \chi_i \chi_j$$

and by $\mathbf{M}^{U_\delta}$ the matrix with entries

$$\mathbf{M}_{ij}^{U_\delta} = \int_\Omega U_\delta \chi_i \chi_j,$$

with U_δ defined in (9.6). Then, the finite element approximation of problem (9.2) is given by the alebraic problem of finding the N_c smallest eigenvalues $\lambda_{\delta 1}, \dots, \lambda_{\delta N_c}$ and the corresponding eigenvectors $\mathbf{c}_1, \dots, \mathbf{c}_{N_c}$ of the generalized algebraic eigenvalue problem

$$\left(\mathbf{h} + 2\mathbf{M}^{U_\delta} + \sum_{n=1}^{N_c} \mathbf{M}^n L^{-1} \mathbf{M}^n \right) \mathbf{c}_k = \lambda_{\delta k} \mathbf{M} \mathbf{c}_k$$

As already mentioned, if the eigenvalues are computed with an iterative scheme such as the Jacobi-Davidson method, the matrices do not need to be explixitely assembled.

9.2.2 Preliminary results for test cases

We now show some extremely basic results obtained with the hp dG approximation of the closed shell Hartree-Fock equations.

Helium hydride ion

In Figure 9.9 we show the computed eigenfunction for the molecule HeH^+ with an equilibrium bond length of $R_0 = 0.772$ Å. Using atomic units, we place the two nuclei at $x_1 = (-0.73, 0, 0)$ and $x_2 = (0.73, 0, 0)$ and the potential reads

$$V(x) = -\frac{1}{|x - x_1|} - \frac{2}{|x - x_2|}.$$

Since the molecule has two electrons of opposite spins, $N_c = 1$. The simulation is performed in a cube of edge approximately 30 times the bond length. The numerical wave function shown in Figure 9.9 is obtained after twelve refinements of the initial

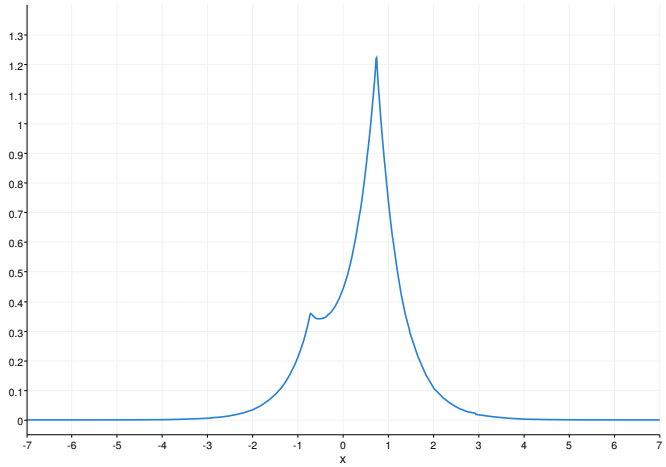


Figure 9.9 – Wave function ψ for HeH^+, plotted over the line $\{-7 \leq x \leq 7, y = 0, z = 0\}$, as a function of x, in atomic units.

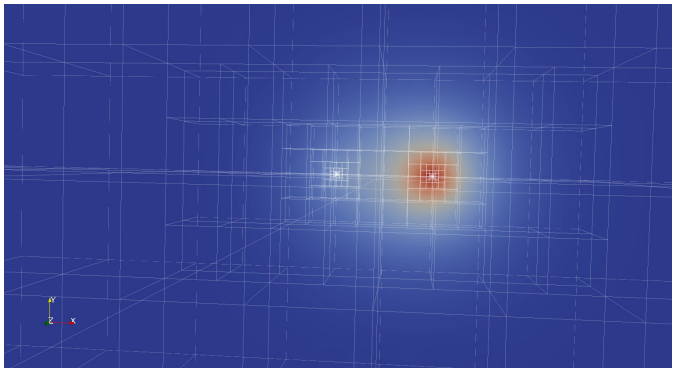


Figure 9.10 – Section of the solution for the HeH^+ molecule, with outline of the three dimensional mesh.

coarse mesh, with a polynomial slope $\mathfrak{s} = 1/8$, and corresponds to a total energy

$$E_\delta^{\mathrm{RHF}} = 2\sum_{i=1}^{N_c} h_\delta(\psi_{\delta i}, \psi_{\delta i}) + 2\sum_{i=1}^{N_c} \int_\Omega \left(\rho_{\Psi_\delta} \star \frac{1}{|x|}\right) \psi_{\delta i}^2 - \sum_{i,j=1}^{N_c} \int_\Omega \left(\psi_{\delta i}\psi_{\delta j} \star \frac{1}{|x|}\right) \psi_{\delta i}\psi_{\delta j} + \frac{2}{R_0}$$
$$\simeq -2.963280 \text{ Hartree},$$

i.e., an error of approximately $1.5 \cdot 10^{-2}$ Hartree compared to the values in [BC79]. A section of the solution is shown, together with a close up of the three dimensional computational mesh, in Figure 9.10.

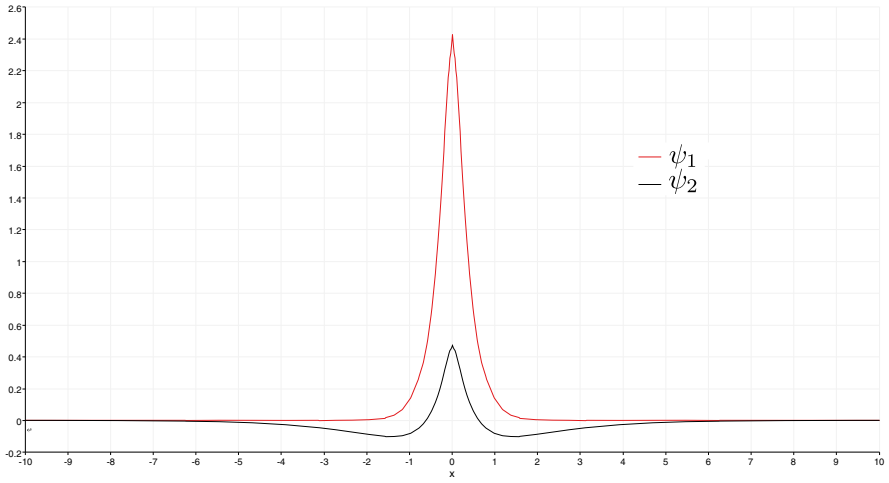


Figure 9.11 – Wavefunctions of the Beryllium atom, plotted on the line $\{-10 \leq x \leq 10, y = 0, z = 0\}$ as a function of x, in atomic units.

Beryllium

In the second case we consider a computation performed on the Beryllium atom, with the nucleus located at $(0, 0, 0)$ and $Z_1 = 4$, $N_c = 2$. The eigenfunctions shown in Figure 9.11 correspond to a computation obtained after 24 refinements of the initial coarse mesh, with a polynomial slope $\mathfrak{s} = 1/8$, and gives an error of approximately 0.18 Hartree on the most negative eigenvalue and of approximately $5 \cdot 10^{-3}$ Hartree on the other eigenvalue, with respect to the values reported in [BBB93].

Perspectives

The method converges to the exact quantities in both cases shown; nonetheless, a careful analysis needs to be done, in order to separate the different components of the error. Specifically, we wish to separate the components due to the approximation of the Poisson problem (9.5), from those due to the finite element approximation, and from those due to the nonlinear and algebraic iterations. In particular, considering the role of the error in the approximation of the Poisson problem (9.5), both the size of the computational domain and the boundary conditions imposed on the solution U appear to be of great relevance in order to fully exploit the accuracy provided by the hp method and they are the most difficult to control. These analyses are the subject of ongoing work.

Bibliography

[ABP06] P. F. Antonietti, A. Buffa, and I. Perugia. "Discontinuous Galerkin approximation of the Laplace eigenproblem". In: *Computer Methods in Applied Mechanics and Engineering* 195.25-28 (2006), pp. 3483–3503.

[AK99] M. Ainsworth and D. Kay. "The approximation theory for the p-version finite element method and application to non-linear elliptic PDEs". In: *Numerische Mathematik* 82 (1999), pp. 351–388.

[Ame+01] P. R. Amestoy, I. S. Duff, J. Koster, and J.-Y. L'Excellent. "A Fully Asynchronous Multifrontal Solver Using Distributed Dynamic Scheduling". In: *SIAM Journal on Matrix Analysis and Applications* 23.1 (2001), pp. 15–41.

[AN07] B. Ammann and V. Nistor. "Weighted Sobolev spaces and regularity for polyhedral domains". In: *Computer Methods in Applied Mechanics and Engineering* 196.37-40 SPEC. ISS. (2007), pp. 3650–3659.

[AN15] J. Adler and V. Nistor. "Graded mesh approximation in weighted Sobolev spaces and elliptic equations in 2D". In: *Mathematics of Computation* 84.295 (2015), pp. 2191–2220.

[Arn+02] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. "Unified Analysis of Discontinuous Galerkin Methods for Elliptic Problems". In: *SIAM Journal on Numerical Analysis* 39.5 (2002), pp. 1749–1779.

[Arn+17] D. Arndt, W. Bangerth, D. Davydov, T. Heister, L. Heltai, M. Kronbichler, M. Maier, J.-P. Pelteret, B. Turcksin, and D. Wells. "The `deal.II` Library, Version 8.5". In: *Journal of Numerical Mathematics* 25.3 (2017), pp. 137–145.

[Arn82] D. N. Arnold. "An Interior Penalty Finite Element Method with Discontinuous Elements". In: *SIAM Journal on Numerical Analysis* 19.4 (1982), pp. 742–760.

[ARS09] T. Apel, A. Rösch, and D. Sirch. "L^∞-Error Estimates on Graded Meshes with Application to Optimal Control". In: *SIAM Journal on Control and Optimization* 48.3 (2009), pp. 1771–1796.

[Bal+17] S. Balay, S. Abhyankar, M. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. Gropp, D. Kaushik, M. Knepley, D. May, L. C. McInnes, K. Rupp, B. Smith, S. Zampini, H. Zhang, and H. Zhang. *PETSc Web page*. `http://www.mcs.anl.gov/petsc`. 2017.

[BBB93] C. Bunge, J. Barrientos, and A. Bunge. "Roothaan-Hartree-Fock Ground-State Atomic Wave Functions: Slater-Type Orbital Expansions and Expectation Values for Z = 2-54". In: *Atomic Data and Nuclear Data Tables* 53.1 (1993), pp. 113–162.

[BC79] D. M. Bishop and L. M. Cheung. "A theoretical investigation of HeH+". In: *Journal of Molecular Spectroscopy* 75.3 (1979), pp. 462–473.

[BO27] M. Born and R. Oppenheimer. "Zur Quantentheorie der Molekeln". In: *Annalen der Physik* 389.20 (1927), pp. 457–484.

[BS03] S. Brenner and L. R. Scott. *The Mathematical theory of finite element methods*. Vol. 46. 2-3. 2003, pp. 512–513.

[Can+14] É. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralık. "A perturbation-method-based a posteriori estimator for the planewave discretization of nonlinear Schrödinger equations". In: *Comptes Rendus Mathematique* 352.11 (2014), pp. 941–946.

[Can+16] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralık. "A perturbation-method-based post-processing for the planewave discretization of Kohn-Sham models". In: *Journal of Computational Physics* 307 (2016), pp. 446–459.

[Can+17] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralık. "Guaranteed and Robust a Posteriori Bounds for Laplace Eigenvalues and Eigenvectors: Conforming Approximations". In: *SIAM Journal on Numerical Analysis* 55.5 (2017), pp. 2228–2254.

[CC04] Z. Chen and H. Chen. "Pointwise error estimates of discontinuous Galerkin methods with penalty for second-order elliptic problems". In: *SIAM Journal on Numerical Analysis* 151.3712 (2004), pp. 859–60.

[CCM10] E. Cancès, R. Chakir, and Y. Maday. "Numerical Analysis of Nonlinear Eigenvalue Problems". In: *Journal of Scientific Computing* 45.1-3 (2010), pp. 90–117.

[CD02] M. Costabel and M. Dauge. "Crack Singularities for General Elliptic Systems". In: *Mathematische Nachrichten* 235.1 (2002), pp. 29–49.

[CDN10a] M. Costabel, M. Dauge, and S. Nicaise. *Corner singularities and analytic regularity for linear elliptic systems*. Book in preparation. 2010.

[CDN10b] M. Costabel, M. Dauge, and S. Nicaise. "Mellin Analysis of Weighted Sobolev Spaces with Nonhomogeneous Norms on Cones". In: *Around the Research of Vladimir Maz'ya I*. Springer New York, 2010, pp. 105–136.

[CDN12] M. Costabel, M. Dauge, and S. Nicaise. "Analytic Regularity for Linear Elliptic Systems in Polygons and Polyhedra". In: *Mathematical Models and Methods in Applied Sciences* 22.08 (2012), p. 1250015.

[CDN14] M. Costabel, M. Dauge, and S. Nicaise. "Weighted analytic regularity in polyhedra". In: *Computers and Mathematics with Applications* 67.4 (2014), pp. 807–817.

[CDS05] M. Costabel, M. Dauge, and C. Schwab. "Exponential convergence of hp-FEM for Maxwell equations with weighted regularization in polygonal domains". In: *Mathematical Models and ...* 15.4 (2005), pp. 575–622.

[CL02] E. Cancès and C. Le Bris. "On the convergence of SCF algorithms for the Hartree-Fock equations". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 34.4 (2002), pp. 749–774.

[CL91] P. G. Ciarlet and J.-L. Lions, eds. *Handbook of numerical analysis. Vol. II*. Handbook of Numerical Analysis, II. Finite element methods. Part 1. North-Holland, Amsterdam, 1991, pp. x+928.

[CL96] P. G. Ciarlet and J. L. Lions, eds. *Handbook of numerical analysis. Vol. IV*. Handbook of Numerical Analysis, IV. Finite element methods. Part 2. Numerical methods for solids. Part 2. North-Holland, Amsterdam, 1996, pp. x+974.

[CLM06] E. Cancès, C. Le Bris, and Y. Maday. *Méthodes mathématiques en chimie quantique : une introduction*. Springer, 2006, p. 409.

[Cou43] R. Courant. "Variational Methods for the Solution of Problems of Equilibrium and Vibrations". In: *Bulletin of the American Mathematical Society* 49.1 (1943), pp. 1–24.

[CS98] 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 (1998), pp. 2440–2463.

[CV13] C. Canuto and M. Verani. "On the Numerical Analysis of Adaptive Spectral/hp Methods for Elliptic Problems". In: *Analysis and Numerics of Partial Differential Equations*. Vol. 4. 2013.

[Dal+12] A. Dall'Acqua, S. Fournais, T. Østergaard Sørensen, and E. Stockmeyer. "Real analyticity away from the nucleus of pseudorelativistic Hartree–Fock orbitals". In: *Analysis & PDE* 5.3 (2012), pp. 657–691.

[DE12] D. A. Di Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*. Vol. 69. Mathématiques et Applications. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[Dem+11] A. Demlow, D. Leykekhman, A. H. Schatz, and L. B. Wahlbin. "Best approximation property in the W^1_∞ norm for finite element methods on graded meshes". In: *Mathematics of Computation* 81.278 (2011), pp. 743–764.

[DEV16] V. Dolejšı, A. Ern, and M. Vohralık. "hp-Adaptation Driven by Polynomial-Degree-Robust A Posteriori Error Estimates for Elliptic Problems". In: *SIAM Journal on Scientific Computing* 38.5 (2016), A3220–A3246.

[DM17] G. Dusson and Y. Maday. "A posteriori analysis of a nonlinear Gross–Pitaevskii-type eigenvalue problem". In: *IMA Journal of Numerical Analysis* 37.1 (2017), pp. 94–137.

[DNR78a] J. Descloux, N. Nassif, and J. Rappaz. “On spectral approximation. I. The problem of convergence”. In: *RAIRO Analyse Numérique* 12.2 (1978), pp. 97–112, iii.

[DNR78b] J. Descloux, N. Nassif, and J. Rappaz. “On spectral approximation. II. Error estimates for the Galerkin method”. In: *RAIRO Analyse Numérique* 12.2 (1978), pp. 113–119, iii.

[EG04] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*. Vol. 159. Applied Mathematical Sciences. New York, NY: Springer New York, 2004.

[ES97] Y. V. Egorov and B.-W. Schulze. *Pseudo-Differential Operators, Singularities, Applications*. Basel: Birkhäuser Basel, 1997, pp. 349+XII.

[EV15] A. Ern and M. Vohralik. “Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous galerkin, and mixed discretizations”. In: *SIAM Journal on Numerical Analysis* 53.2 (2015), pp. 1058–1081.

[Gal15] B. G. Galerkin. “Rods and plates. Series occurring in various questions concerning the elastic equilibrium of rods and plates”. In: *Engineers Bulletin (Vestnik Inzhenerov)* 19 (1915), pp. 897–908.

[GB86a] W. Gui and I. Babuška. “The h, p and h-p versions of the finite element method in 1 dimension. Part I. The Error Analysis of the p-Version”. In: *Numerische Mathematik* 612 (1986), pp. 577–612.

[GB86b] W. Gui and I. Babuška. “The h, p and h-p versions of the finite element method in 1 dimension. Part II. The Error analysis of the $h-$ and $h-p$ versions.” In: *Numerische Mathematik* 49.6 (1986), pp. 613–657.

[GB86c] W. Gui and I. Babuška. “The h, p and h-p versions of the finite element method in 1 dimension. Part III. The Adaptive h-p Version”. In: *Numerische Mathematik* 683 (1986), pp. 659–683.

[GB86d] B. Guo and I. Babuška. “The h-p version of the finite element method - Part 1: The basic approximation results”. In: *Computational Mechanics* 1.1 (1986), pp. 21–41.

[GB86e] B. Guo and I. Babuška. “The h-p version of the finite element method - Part 2: General results and applications”. In: *Computational Mechanics* 1.3 (1986), pp. 203–220.

[Geo08] E. H. Georgoulis. “Inverse-type estimates on hp-finite element spaces and applications”. In: *Mathematics of Computation* 77.261 (2008), pp. 201–219.

[GGO11] S. Giani, L. Grubišić, and J. Ovall. “Reliable a-posteriori error estimators for hp-adaptive finite element approximations of eigenvalue/eigenvector problems”. In: *arXiv preprint arXiv:1112.0436* (2011).

[GGO13] S. Giani, L. Grubišić, and J. Ovall. “Error control for hp-adaptive approximations of semi-definite eigenvalue problems”. In: *Computing* (2013), pp. 1–31.

[GH12] S. Giani and E. J. C. Hall. "An a posteriori error estimator for hp-adaptive discontinuous Galerkin methods for elliptic eigenvalue problems". In: *Mathematical Models and Methods in Applied Sciences* 22.10 (2012), p. 1250030.

[GS05] E. Georgoulis and E. Süli. "Optimal error estimates for the hp-version interior penalty discontinuous Galerkin finite element method". In: *IMA Journal of Numerical Analysis* 03 (2005), pp. 1–17.

[Guz06] J. Guzmán. "Pointwise error estimates for discontinuous Galerkin methods with lifting operators for elliptic problems". In: *Mathematics of Computation* 75.255 (2006), pp. 1067–1085.

[GW12] M. J. Gander and G. Wanner. "From Euler, Ritz, and Galerkin to Modern Computing". In: *SIAM Review* 54.4 (2012), pp. 627–666.

[HNS08] E. Hunsicker, V. Nistor, and J. O. Sofo. "Analysis of periodic Schrödinger operators: Regularity and approximation of eigenfunctions". In: *Journal of Mathematical Physics* 49.8 (2008), p. 083501.

[Hou05] P. Houston. "Discontinuous Galerkin finite element approximation of quasi-linear elliptic boundary value problems I: the scalar case". In: *IMA Journal of Numerical Analysis* 25.4 (2005), pp. 726–749.

[HRV05] V. Hernandez, J. E. Roman, and V. Vidal. "SLEPc: A scalable and flexible toolkit for the solution of eigenvalue problems". In: *ACM Transactions on Mathematical Software* 31.3 (2005), pp. 351–362.

[HSW07a] P. Houston, E. Suli, and T. P. Wihler. "A posteriori error analysis of hp-version discontinuous Galerkin finite-element methods for second-order quasi-linear elliptic PDEs". In: *IMA Journal of Numerical Analysis* 28.2 (2007), pp. 245–273.

[HSW07b] P. Houston, D. Schötzau, and T. P. Wihler. "Energy norm a posteriori error estimation of hp-adaptive discontinuous Galerkin methods for elliptic problems". In: *Mathematical Models and Methods in Applied Sciences* 17.01 (2007), pp. 33–62.

[HW08] J. S. Hesthaven and T. Warburton. *Nodal Discontinuous Galerkin Methods*. Vol. 54. Texts in Applied Mathematics. Springer New York, 2008.

[Jea79] P. Jeanquartier. "Tranformation de Mellin et développements". In: *L'Enseignement Mathématique* 25 (1979), pp. 285–308.

[Joh50] F. John. "The fundamental solution of linear elliptic differential equations with analytic coefficients". In: *Communications on Pure and Applied Mathematics* 3.3 (1950), pp. 273–304.

[Kat96] K. Kato. "New idea for proof of analyticity of solutions to analytic nonlinear elliptic equations". In: *SUT Journal of Mathematics* 32.2 (1996), pp. 157–161.

[KM99] V. Kozlov and V. Maz'ya. *Differential Equations with Operator Coefficients*. Springer Monographs in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999.

[KMR97] V. Kozlov, V. G. Mazya, and J. Rossmann. *Elliptic boundary value problems in domains with point singularities*. American Mathematical Society, 1997, p. 414.

[Kon67] V. A. Kondrat'ev. "Boundary value problems for elliptic equations in domains with conical or angular points". In: *Trudy Moskovskogo Matematičeskogo Obščestva* 16 (1967), pp. 209–292.

[Le 03] C. Le Bris, ed. *Handbook of numerical analysis. Vol. X*. Handbook of Numerical Analysis, X. Special volume: computational chemistry. North-Holland, Amsterdam, 2003, pp. xvi+899.

[Lew04] M. Lewin. "Solutions of the Multiconfiguration Equations in Quantum Chemistry". In: *Archive for Rational Mechanics and Analysis* 171.1 (2004), pp. 83–114.

[LMN10] H. Li, A. Mazzucato, and V. Nistor. "Analysis of the finite element method for transmission/mixed boundary value problems on general polygonal domains". In: *Electronic Transactions Numerical Analysis* 37 (2010), pp. 41–69.

[LN09] H. Li and V. Nistor. "Analysis of a modified Schrödinger operator in 2D: Regularity, index, and FEM". In: *Journal of Computational and Applied Mathematics* 224.1 (2009), pp. 320–338.

[LS03] A. Lasis and E. Suli. "Poincaré-type inequalities for broken Sobolev spaces". In: *Technical Report 03/10, Oxford University Computing Laboratory* (2003), pp. 1–20.

[LSY00] E. H. Lieb, R. Seiringer, and J. Yngvason. "Bosons in a trap: A rigorous derivation of the Gross-Pitaevskii energy functional". In: *The Stability of Matter: From Atoms to Stars*. Berlin/Heidelberg: Springer-Verlag, 2000, pp. 759–771.

[MN10] A. Mazzucato and V. Nistor. "Well-posedness and regularity for the elasticity equation with mixed boundary conditions on polyhedral domains and domains with cracks". In: *Archive for Rational Mechanics and Analysis* (2010), pp. 1–45.

[MNP00] V. G. Maz'ya, S. Nazarov, and B. Plamenevskij. *Asymptotic Theory Elliptic Boundary Value Problems in Singularly Perturbed Domains: Vol. 1*. Birkhäuser, 2000.

[MP78] V. G. Maz'ya and B. A. Plamenevskii. "Estimates of Green's functions and Schauder estimates for solutions of elliptic boundary problems in a dihedral angle". In: *Siberian Mathematical Journal* 19.5 (1978), pp. 752–764.

[MP85] V. G. Maz'ya and B. A. Plamenevskii. "On the Asymptotics of the Fundamental Solutions of Elliptic Boundary—Value Problems in Regions with Conical Points". In: *Selecta Mathematica Sovietica* 4.4 (1985), pp. 363–397.

[MR02] V. G. Maz'ya and J. Rossmann. "Point estimates for Green's matrix to boundary value problems for second order elliptic systems in a polyhedral cone". In: *ZAMM Zeitschrift fur Angewandte Mathematik und Mechanik* 82.5 (2002), pp. 291–316.

[MR10] V. G. Maz'ya and J. Rossmann. *Elliptic Equations in Polyhedral Domains*. Vol. 162. Mathematical Surveys and Monographs. American Mathematical Society, 2010.

[Nic97] S. Nicaise. "Regularity of the solutions of elliptic systems in polyhedral domains". In: *Bulletin of the Belgian Mathematical Society - Simon Stevin* 4.3 (1997), pp. 411–429.

[Nit72] J. Nitsche. "On Dirichlet problems using subspaces with nearly zero boundary conditions". In: *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*. Elsevier, 1972, pp. 603–627.

[PS03] L. P. Pitaevskii and S. Stringari. *Bose-Einstein condensation*. Clarendon Press, 2003.

[Pul80] P. Pulay. "Convergence acceleration of iterative sequences. the case of scf iteration". In: *Chemical Physics Letters* 73.2 (1980), pp. 393–398.

[Qua17] A. Quarteroni. *Numerical Models for Differential Problems*. Vol. 16. MS&A. Springer International Publishing, 2017.

[RH73] W. H. Reed and T. Hill. *Triangular mesh methods for the neutron transport equation*. Tech. rep. Los Alamos Scientific Lab., (USA), 1973.

[Rit09a] W. Ritz. "Theorie der Transversalschwingungen einer quadratischen Platte mit freien Rändern". In: *Annalen der Physik* 333.4 (1909), pp. 737–786.

[Rit09b] W. Ritz. "Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik." In: *Journal für die reine und angewandte Mathematik* 135 (1909), pp. 1–61.

[Riv08] B. Rivière. *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations*. Society for Industrial and Applied Mathematics, 2008.

[Sch01] A. H. Schatz. "Pointwise Error Estimates and Asymptotic Error Expansion Inequalities for the Finite Element Method on Irregular Grids: Part II. Interior Estimates". In: *SIAM Journal on Numerical Analysis* 38.4 (2001), pp. 1269–1293.

[Sch26] E. Schrödinger. "An Undulatory Theory of the Mechanics of Atoms and Molecules". In: *Physical Review* 28.6 (1926), pp. 1049–1070.

[Sch98] A. H. Schatz. "Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids: Part I. Global estimates". In: *Mathematics of Computation* 67.223 (1998), pp. 877–900.

[SF08] G. Strang and G. Fix. *An analysis of the finite element method*. Second. Wellesley-Cambridge Press, Wellesley, MA, 2008, pp. x+402.

[SO12] A. Szabo and N. Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.

[SSW13a] D. Schötzau, C. Schwab, and T. P. Wihler. "hp-dGFEM for second order elliptic problems in polyhedra. II: Exponential convergence". In: *SIAM Journal on Numerical Analysis* 51.4 (2013), pp. 2005–2035.

[SSW13b] D. Schötzau, C. Schwab, and T. Wihler. "hp-dGFEM for Second-Order Elliptic Problems in Polyhedra I: Stability on Geometric Meshes". In: *SIAM Journal on Numerical Analysis* 51.3 (2013), pp. 1610–1633.

[SSW16] D. Schötzau, C. Schwab, and T. P. Wihler. "hp-dGFEM for second-order mixed elliptic problems in polyhedra". In: *Mathematics of Computation* 85.299 (2016), pp. 1051–1083.

[Sta65] G. Stampacchia. "Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus". In: *Université de Grenoble. Annales de l'Institut Fourier* 15.1 (1965), pp. 189–258.

[Ste02] G. W. Stewart. "A Krylov–Schur Algorithm for Large Eigenproblems". In: *SIAM Journal on Matrix Analysis and Applications* 23.3 (2002), pp. 601–614.

[SV96] G. L. Sleijpen and H. A. Van der Vorst. "A Jacobi–Davidson Iteration Method for Linear Eigenvalue Problems". In: *SIAM Journal on Matrix Analysis and Applications* 17.2 (1996), pp. 401–425.

[SVF94] G. L. G. Sleijpen, H. A. van der Vorst, and D. R. Fokkema. "BiCGstab(l) and other hybrid Bi-CG methods". In: *Numerical Algorithms* 7.1 (1994), pp. 75–109.

[SW10] B. Stamm and T. P. Wihler. "hp-optimal discontinuous Galerkin methods for linear elliptic problems". In: *Mathematics of Computation* 79 (2010), pp. 2117–2133.

[SW78] A. H. Schatz and L. B. Wahlbin. "Maximum Norm Estimates in the Finite Element Method on Plane Polygonal Domains. Part 1". In: *Mathematics of Computation* 32.141 (1978), pp. 73–109.

[Vor92] H. A. van der Vorst. "Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of Nonsymmetric Linear Systems". In: *SIAM Journal on Scientific and Statistical Computing* 13.2 (1992), pp. 631–644.

[Whe78] M. F. Wheeler. "An Elliptic Collocation-Finite Element Method with Interior Penalties". In: *SIAM Journal on Numerical Analysis* 15.1 (1978), pp. 152–161.

[Wih07] T. Wihler. "Weighted L2-norm a posteriori error estimation of FEM in polygons". In: *International Journal of Numerical Analysis and Modeling* 4.1 (2007), pp. 100–115.

Discontinuous hp finite element methods for elliptic eigenvalue problems with singular potentials

Abstract

In this thesis, we study elliptic eigenvalue problems with singular potentials, motivated by several models in physics and quantum chemistry, and we propose a discontinuous Galerkin hp finite element method for their solution. In these models, singular potentials occur naturally (associated with the interaction between nuclei and electrons). Our analysis starts from elliptic regularity in non homogeneous weighted Sobolev spaces. We show that elliptic operators with singular potential are isomorphisms in those spaces and that we can derive weighted analytic type estimates on the solutions to the linear eigenvalue problems. The isotropically graded hp method provides therefore approximations that converge with exponential rate to the solution of those eigenproblems. We then consider a wide class of nonlinear eigenvalue problems, and prove the convergence of numerical solutions obtained with the symmetric interior penalty discontinuous Galerkin method. Furthermore, when the non linearity is polynomial, we show that we can obtain the same analytic type estimates as in the linear case, thus the numerical approximation converges exponentially. We also analyze under what conditions the eigenvalue converges at an increased rate compared to the eigenfunctions. For both the linear and nonlinear case, we perform numerical tests whose objective is both to validate the theoretical results, but also evaluate the role of sources of errors not considered previously in the analysis, and to help in the design of hp/dG graded methods for more complex problems.

Keywords: hp/dG graded finite element method, discontinuous Galerkin, nonlinear eigenvalue problem, quantum chemistry, weighted Sobolev spaces, elliptic regularity

---

Résumé

Dans cette thèse, on étudie des problèmes aux valeurs propres elliptiques avec des potentiels singuliers, motivés par plusieurs modèles en physique et en chimie quantique, et on propose une méthode des éléments finis de type hp discontinus (dG) adaptée pour l'approximation des modes propres. Dans ces modèles, arrivent naturellement des potentiels singuliers (associés à l'interaction entre noyaux et électrons). Notre analyse commence par une étude de la régularité elliptique dans des espaces de Sobolev à poids. On montre comment un opérateur elliptique avec potentiel singulier est un isomorphisme entre espaces de Sobolev à poids non homogènes et que l'on peut développer des bornes de type analytique à poids sur les solutions des problèmes aux valeurs propres associés aux opérateurs. La méthode hp/dG graduée qu'on utilise converge ainsi de façon exponentielle. On poursuit en considérant une classe de problèmes non linéaires représentatifs des applications. On montre que, sous certaines conditions, la méthode hp/dG graduée converge et que, si la non linéarité est de type polynomiale, on obtient les mêmes estimations de type analytique que dans le cas linéaire. De plus, on étudie la convergence de la valeur propre pour voir sous quelles conditions la vitesse de convergence est améliorée par rapport à celle des vecteurs propres. Pour tous les cas considérés, on effectue des tests numériques, qui ont pour objectif à la fois de valider les résultats théoriques, mais aussi d'évaluer le rôle des sources d'erreur non considérées dans l'analyse et d'aider dans la conception de méthode hp/dG graduée pour des problèmes plus complexes.

Mots clés : méthode des éléments finis hp/dG graduée, Galerkin discontinu, problèmes aux valeurs propres non linéaires, chimie quantique, espaces de Sobolev à poids, régularité elliptique

---

Laboratoire Jacques-Louis Lions
4 place Jussieu – 75005 Paris – France