Towards a Fictitious Domain Method with Optimally Smooth Solutions

Mario S. Mommer

Towards a Fictitious Domain Method with Optimally Smooth Solutions

Von der Fakultät für Mathematik, Informatik
und Naturwissenschaften
der Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
genehmigte Dissertation

vorgelegt von

Mario Salvador Mommer

aus Tübingen

Berichter: Universitätsprofessor Dr. Wolfgang Dahmen
Universitätsprofessor Dr. Karl-Henning Esser

Tag der mündlichen Prüfung: 20. Juni 2005

To my parents.

Contents

Chapter 1

Introduction

Fictitious domain methods, sometimes also called domain embedding methods, are a family of tools for the solution of boundary value problems on irregular and complex geometries. What distinguishes them from other methods is that they try to employ simple discretizations and methods which work well on regular geometries, and coerce them, in one way or another, to produce a solution of the problem on the complex geometry. They achieve this by embedding the original domain into a much simpler one (the fictitious domain), and reformulating the problem there, a step which always involves some form of extension of the data. Instead of solving the original problem directly, one obtains an extension to the fictitious domain of the solution of the original problem. The boundary conditions are usually enforced by mechanisms which do not modify the discretization on the domain, or do so only in a limited way. Prominent examples of such mechanisms are Lagrange multipliers and penalty parameters.

This type of construction produces fairly flexible methods that can cope easily with problems where the geometry changes often. A canonical application is the use as a component in shape optimization problems or free boundary problems (see for instance [24]). What makes fictitious domain methods so invaluable in these applications is their strict black box approach. Since no remeshing is necessary, they can operate on machine-generated geometry descriptions without supervision, and do so reliably.

Another possible reason to use a fictitious domain formulation is to tap the power of methods which are only available on simple geometries, a theme explored for example in [3] and the references therein. It is this point of view which shall dominate in the present thesis. We will study fictitious domain formulations as an alternative to other, more traditional formulations for standard elliptic boundary value problems, focusing on them as a vehicle to simplify the use of wavelet-Galerkin discretization schemes.

1.1 Wavelet methods and fictitious domain formulations

Wavelets, which appeared first as a tool for signal analysis, have been playing an increasingly important role in numerical algorithms for the solution of partial differential equations. For

the solution of elliptic boundary value problems, biorthogonal wavelet bases are an attractive choice. They lead easily to well conditioned discretizations of the type of operator equation that appears in these problems. This property, their good approximation power, and their clear mathematical structure have led to the development of novel methods which profit from results from related mathematical disciplines.

The adaptive wavelet methods developed with the aid of deep approximation theoretical results in [8, 7] illustrate this point quite clearly. These algorithms are capable of producing good approximations of the solutions of elliptic boundary value problems with an optimal work/accuracy balance. They are optimal in the sense that to produce an approximation of the solution to a given problem, the number of operations needed is proportional to $\epsilon^{-1/s}$, where ϵ is the desired accuracy (measured in a relevant norm, usually the Energy norm), and the parameter s depends on the smoothness of the solution, as measured by thir membership in certain Besov spaces.

Perhaps the most important property of the class of wavelets used in these methods is that they are Riesz bases for the Sobolev spaces involved. But while they are easy to construct and handle for, say, periodic domains, the situation is quite different for domains with complex geometries. And while the construction for those domains is a solved problem [13], the resulting bases are difficult to handle. The numerical properties of such bases also suffers somewhat, leading to discrete problems which are not as well conditioned as their counterparts on simple domains. Thus, a possible strategy to overcome these difficulties when dealing with complex geometries is to use a fictitious domain formulation. This approach was initiated successfully in [27].

The choice of suitable fictitious domain formulations one may consider for this endeavor is limited, however. Methods based on the introduction of penalty parameters lead to discrete problems that are not uniformly well conditioned. The same holds for any other method based on regularization techniques (see for instance [20]).

The formulation which seems to be best suited for such a purpose is the fictitious domain - Lagrange multiplier (FDLM) approach initiated by [1, 22], and used in [27]. To solve a second order elliptic boundary value problem with Dirichlet boundary conditions on a bounded domain, one extends the data (and the differential operator) to a simpler domain, and appends the boundary conditions by introducing a Lagrange multiplier. This leads to a saddle-point problem which is amenable to the discretization and solution with wavelet techniques [10].

In chapter three we will show that this approach has its limitations. While the solution of the original problem may be very smooth in either of the Sobolev or Besov scales, this does not hold in general for the extended solution obtained through the FDLM formulation. If the data was not extended in *exactly the right way*, the smoothness of the extended solution is deficient, and thus approximating it requires more degrees of freedom, and ultimately more work.

1.2 Towards a fictitious domain method with optimally smooth solutions

Correcting this deficiency in the FDLM formulation in a way that keeps the formulation practical is fairly difficult; as a matter of fact, an extensive search of the literature showed no attempt, successful or unsuccessful, to address this problem. There is one trivial way around this difficulty (take the solution of the original problem, extend it smoothly, and use the differential operator to obtain a suitable extension of the data) but it leads to a method which is hardly practical, since it needs the solution *first*.

In chapter four we will attempt to construct a method which produces optimally smooth extensions of the solution. For this we will begin by formulating on the fictitious domain a rank-deficient, but otherwise well-posed[1], least squares problem whose solutions all agree on the original domain with the original solution. Then we play with the process of solving the discrete equations to obtain a solution of the least squares problem which is also smooth.

The smooth extension is constructed by a nested iteration scheme through what amounts to emergent behavior. A proof of this property will be given subject to a few conditions on the finite dimensional problems obtained by the process of discretization. We will also construct a discretization scheme which, at least numerically, seems to satisfy these conditions.

The resulting method is fairly simple in structure. Wavelets appear in the discretization as a natural choice and, more importantly, no modification of the bases is needed. This makes our method usable as a black box. Furthermore, the method can deal in a unified way with any type of boundary conditions.

1.3 Overview

We begin in chapter two by weaving together in a uniform way the theory we will use in the following chapters. We will need some elements of approximation theory, theory of elliptic boundary value problems, and the construction of B-spline wavelets.

Chapter three is devoted to the analysis of the fictitious domain - Lagrange multiplier approach. Here we will show how the method is derived, and analyze the smoothness of the extended solutions by considering their membership in Besov and Sobolev spaces. We extend and complete first the results on smoothness in the Sobolev scale found in [21], taking an approximation theoretical point of view, and then prove new results which bound the convergence rate of nonlinear approximation schemes. We have succeeded in collecting all the difficult technical details into one lemma, which makes the discussion more transparent. The second half of chapter three is then spent proving this lemma.

The development of a fictitious domain method able to produce optimally smooth solutions takes place in chapter four. First we introduce and analyze the least-squares formulation that will serve as a foundation, and then we proceed to construct the method, and prove that under certain assumptions to the discretization, it produces optimally smooth solutions. Then we introduce a discretization scheme designed to satisfy these assumptions.

[1] in the sense that it is solvable, and that its solutions can be chosen to depend continuously on the data

In chapter five we will give numerical evidence that supports the results of chapters three and four. We will begin by showing the effects of the singularities introduced by the FDLM approach with respect to the convergence of linear and nonlinear approximation schemes. Then we will test the method developed in chapter four on a set of model problems, and observe how it succeeds in providing smooth solutions on the fictitious domain which, restricted to the original domain, solve the original problem.

A chapter with final notes can be found at the end of this thesis, summarizing our finds in a conclusions section, and discuss directions for further research.

1.4 Acknowledgements

I would like to express my deep gratitude towards prof. Wolfgang Dahmen for his kind support and his guiding input. Many thanks also go to prof. Karl-Henning Esser for kindly accepting to coreferee this thesis.

I would also like to thank prof. Angela Kunoth, prof. Silvia Bertoluzza, and prof. Peter Oswald for many interesting discussions.

I am also indebted to Dr. Ralf Massjung, Dr. Torben K. Jensen, and Dr. Daniel Castaño for invaluable mathematical discussions and moral support during these years. Also to my colegues at the "Institut für Geometrie und Praktische Mathematik" who always made me feel at home.

Chapter 2

Theoretical framework

The present chapter sets the tapestry on which the material of later chapters unfolds. Instead of presenting a loose collection of facts, we have tried to draw a map of the body of theory involved. It has been drawn in a mostly strict logical order, beginning with space interpolation and abstract approximation theory, then going on to define Besov and Sobolev spaces as approximation spaces. After reviewing the standard trace and extension theorems, and complementing them with more modern results which will be useful later, we define the class of problems we want to study: second order elliptic boundary value problems with either Dirichlet or Neumann boundary conditions. After this we introduce B-spline wavelets, and the brand of nonlinear approximation that is the foundation of adaptive wavelet methods.

2.1 Interpolation spaces

The definition of what constitutes an interpolation space requires the following steps [2]. Let A_0 and A_1 be two normed spaces. They are called *compatible* if there exists a Hausdorff topological vector space $\mathfrak{V}$ such that A_0 and A_1 are subspaces of it. A normed space A is called an *intermediate space* between the compatible spaces A_0 and A_1, if $A_0 \cap A_1 \subset A \subset A_0 + A_1$. An *interpolation space* with respect to the couple (A_0, A_1) is then any intermediate space A between A_0 and A_1 for which the following holds. Whenever a linear map $T : A_0 + A_1 \to A_0 + A_1$ is also a bounded linear map from A_0 to itself, as well as from A_1 to itself, then T maps A boundedly into itself.

To construct such spaces, we follow here the *real method* due to J. Peetre, as found in [2]. We define first the K-functional for $v \in A_0 + A_1$ by

$$K(t, v, A_0 + A_1) := \inf_{v=a_0+a_1} \left(\|a_0\|_{A_0} + t\|a_1\|_{A_1} \right),$$

where the infimum is taken over all possible representations $v = a_0 + a_1$ with $a_o \in A_0$ and $a_1 \in A_1$. For a fixed $v \in A_0 + A_1$, one can show that $K(t, v, A_0 + A_1)$ is positive, increasing, and concave.

The following observation is the key to the construction of interpolation spaces using the K-functional. Let $T : A_0 \to A_1$ be as described in the first paragraph, and let $v \in A_0 + A_1$.

Then

$$K(t, v, A_0 + A_1) \leq C\, K(t, Tv, A_0 + A_1), \tag{2.1}$$

where the constant $C \in (0, +\infty)$ is independent of t.

Now, for $0 < \theta < 1$, $0 < q \leq \infty$, let $A_{\theta,q}$ be the subspace of $A_0 + A_1$ of elements which satisfy $\|v\|_{\theta,q} < \infty$, with

$$\|v\|_{\theta,q} := \begin{cases} \left[\int_0^\infty \left\{t^{-\theta} K(t, v, A_0 + A_1)\right\}^q \frac{dt}{t}\right]^{\frac{1}{q}}, & \text{if } 0 \leq q < \infty, \\ \sup_{t\in(0,+\infty)} t^{-\theta} K(t, v, A_0 + A_1) & \text{if } q = +\infty. \end{cases}$$

From (2.1) it follows immediately that the space $A_{\theta,q}$ is an interpolation space between A_0 and A_1. But more is true. If (B_0, B_1) is another pair of compatible spaces, and $T : A_0 + A_1 \to B_0 + B_1$ is such that T maps A_0 boundedly to B_0, and A_1 boundedly to B_1, then using the same argument we see that $T : A_{\theta,q} \to B_{\theta,q}$ is also a bounded operator.

To shed light onto the relation between interpolation spaces, we include the following theorem.

Theorem 2.1.1 (The reiteration theorem). *Let $q_0, q_1 \in (0, +\infty)$, $\theta_1, \theta_2 \in (0, 1)$, and let $\theta = (1 - \eta)\theta_0 + \eta\theta_1$ for some $\eta \in (0, 1)$. Then for any $q \in (0, +\infty)$ it holds that ([2], p.50)*

$$\left((A_0, A_1)_{\theta_0,q_0}, (A_0, A_1)_{\theta_1,q_1}\right)_{\eta,q} = (A_0, A_1)_{\theta,q}$$

with equivalent norms.

2.2 Approximation spaces

Approximation spaces allow us to talk about approximation methods in an abstract setting[1]. For this, let X be a normed vector space, and let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of subsets of X satisfying the following axioms.

Axioms 2.2.1

i. $X_n \subset X_{n+1}$ for all $n \in \mathbb{N}$.

ii. $aX_n \subset X_n$ for all $x \in \mathbb{R}$.

iii. There exists a constant $c \in \mathbb{N}$ such that for every $n \in \mathbb{N}$, $X_n + X_n \subset X_{cn}$.

iv. If $f \in X$, then $\lim_{n\to+\infty} \inf_{x\in X_n} \|f - x\| \to 0$.

[1] This account follows [18]

The sequence $\{X_n\}_{n\in\mathbb{N}}$ will play the role of our approximation method.

To illustrate what these axioms mean, we consider the case when X is a separable Hilbert space, and $B = \{b_k\}_{k\in\mathbb{N}}$ is an orthonormal basis. We might choose

$$X_n = \operatorname{span}\{b_k : k \le n\},$$

and see immediately that it satisfies the above axioms. Since the X_n are linear spaces, we speak of *linear approximation*.

In contrast, consider the choice

$$X_n = \{x \in X : x = \sum_{k\in\Lambda} c_k b_k,\ \Lambda \subset \mathbb{N} \text{ with } \#\Lambda \le n,\ c_k \in \mathbb{R}\},$$

which is the nonlinear space of elements in X which are a linear combination of at most n members of B. Whenever the sequence $\{X_n\}_{n\in\mathbb{N}}$ contains sets which are not linear subspaces of X, we speak of *nonlinear approximation*. We will take a closer look at schemes of this type later on.

Note that the above are just examples, and their introduction does *not* amount to a concrete definition of the spaces X_n in a particular setting.

After having chosen an approximation method, we want to rate its performance according to the behavior of the *error of approximation*, which, for $v \in X$, is defined by

$$E_n(v) := \inf_{x\in X_n} \|v - x\|.$$

Approximation spaces classify the elements of X according to how well they can be approximated with $\{X_n\}_{n\in\mathbb{N}}$. For $0 < s < +\infty$, and $0 < q \le +\infty$, they are given by

$$\mathcal{A}_q^s(X, \{X_n\}) := \{f \in X : \|f\|_{\mathcal{A}_q^s} < +\infty\},$$

with $\|\cdot\|_{\mathcal{A}_q^s} := \|\cdot\|_X + |\cdot|_{\mathcal{A}_q^s}$, and

$$|f|_{\mathcal{A}_q^s} := \begin{cases} \left(\sum_{n=1}^{+\infty} [n^s E_n(f)]^q \frac{1}{n}\right)^{\frac{1}{q}} & \text{if } 0 < q < +\infty, \\ \sup_{n\in\mathbb{N}} n^s E_n(f) & \text{if } q = +\infty. \end{cases}$$

For an element $f \in X$, membership in $\mathcal{A}_q^s(X, \{X_n\}_{n\in\mathbb{N}})$ means above anything else that the approximation error decays at least as $\mathcal{O}(n^{-s})$. The parameter q further indicates the slightly stronger (for $q < \infty$) assertion that $\{n^s E_n(f)\}$ belongs to ℓ_q. The parameter q is of secondary nature; it is possible to prove that if $s < r$, then

$$\mathcal{A}_q^r(X, \{X_n\}_{n\in\mathbb{N}}) \subset \mathcal{A}_p^s(X, \{X_n\}_{n\in\mathbb{N}}) \qquad \forall\ 0 < q, p < +\infty \tag{2.2}$$

We obtain the same space, with an equivalent norm, if we use the following (equivalent) seminorm $|\cdot|_{\mathcal{A}_q^s(X,\{X_n\}_{n\in\mathbb{N}})}$.

$$|f|_{\mathcal{A}_q^s} := \begin{cases} \left(\sum_{n=0}^{+\infty} [2^{ns} E_{2^n}(f)]^q\right)^{\frac{1}{q}} & \text{if } 0 < q < +\infty, \\ \sup_{n\in\mathbb{N}} 2^{ns} E_{2^n}(f) & \text{if } q = +\infty. \end{cases} \tag{2.3}$$

When proving membership in a space A_q^s, it is often easier to use this last definition. The fact that (2.3) defines an equivalent norm hints at some redundancy in the sequence $\{X_n\}_{n\in\mathbb{N}}$. We shall often write $V_j := X_{2^j}$, $j = 0, 1, \ldots$, and then write

$$\mathcal{A}_q^s(X, \{V_j\}_{j\in\mathbb{N}_0}) := \mathcal{A}_q^s(X, \{X_n\}_{n\in\mathbb{N}}). \tag{2.4}$$

In this case, we will always use the seminorm defined in (2.3) for the space on the left of (2.4). Stretching things a little bit further, we will often start by defining the spaces V_j, obviating the spaces X_n with $n \neq 2^j$, and using only the space on the left of (2.4). This causes no problem, since *any* sequence $\{X_n\}$ with $V_j = X_{2^j}$, which also satisfies axioms 2.2.1, would define the same space with an equivalent norm.

A note is also in order regarding spaces of the type ℓ_p with $0 < p < 1$. The corresponding ℓ_p-"norm" is no longer a norm, but instead is only a *quasinorm*. The triangle inequality holds only in its modified form

$$\|a + b\|_{\ell_p} \leq 2^{\frac{1}{p}} \left(\|a\|_{\ell_p} + \|b\|_{\ell_p}\right).$$

To substitute the concept of Banach space we define a quasi-Banach space as a quasi-normed space $(Z, \|\cdot\|)$, where every Cauchy sequence (with respect to the quasi-norm) has a limit in Z. One can then prove that the space ℓ_p, $0 < p < 1$ is indeed a quasi-Banach space.

The same holds, *mutatis mutandis*, for L_p spaces with $0 < p < 1$.

2.2.1 approximation spaces and space interpolation

In this subsection we are going to shed some light on the relation between interpolation spaces and approximation spaces.

The first main result that is concerned with this relation states conditions under which an interpolation space is equal to an approximation space. Let $Y \subset X$ be a normed space which can be embedded continuously into X. Let $\{X_n\}_{n\in\mathbb{N}}$ be an approximation method satisfying axioms 2.2.1, and suppose that the following inequalities hold.

$$E_n(f) \leq Cn^{-r}\|f\|_Y, \forall f \in Y \qquad \text{(Jackson inequality)} \tag{2.5}$$

$$\|S\|_Y \leq Cn^r\|S\|_X, \ \forall S \in X_n \qquad \text{(Bernstein inequality)} \tag{2.6}$$

for some $r > 0$.

Theorem 2.2.1. *If the Jackson and Bernstein inequalities hold, then for every* $0 < s < r$, *and every* $0 < q \leq +\infty$,

$$\mathcal{A}_q^s(X, \{X_n\}_{n\in\mathbb{N}}) = (X, Y)_{s/r,q}$$

with equivalent norms.

An important application of this theorem is that it allows us to compare approximation spaces obtained with different approximation methods.

Corollary 2.2.2. *Let $\{X_n^1\}_{n\in\mathbb{N}}$ and $\{X_n^2\}_{n\in\mathbb{N}}$ be two sequences satisfying 2.2.1, and suppose that there exists $r > 0$ such that both satisfy the Jackson and Bernstein inequalities with respect to a space Y as described above. Then for every $0 < s < r$, $0 < q \leq +\infty$,*

$$\mathcal{A}_q^s(X, \{X_n^1\}_{n\in\mathbb{N}}) = \mathcal{A}_q^s(X, \{X_n^2\}_{n\in\mathbb{N}})$$

with equivalent norms.

As a complement of theorem 2.2.1 we have also that approximation spaces form indeed an interpolation family.

Theorem 2.2.3. *[DeVore and Popov, 1988] Let $\{X_n\}_{n\in\mathbb{N}}$ satisfy axioms 2.2.1. Then, for any $r > 0$, the sequence $\{X_n\}_{n\in\mathbb{N}}$ satisfies the Bernstein and Jackson inequalities with $Y = \mathcal{A}_q^r(X, \{X_n\}_{n\in\mathbb{N}})$, for any $0 < q \leq +\infty$. Thus, for all $0 < s < r$, and all $0 < q, t \leq +\infty$ we have*

$$\mathcal{A}_q^s(X, \{X_n\}_{n\in\mathbb{N}}) = (X, \mathcal{A}_q^r(X, \{X_n\}_{n\in\mathbb{N}}))_{s/r,q}.$$

Next we present a consequence of the reiteration theorem which characterizes what we obtain when we define an approximation space inside of an approximation space. It reads[2] as follows.

Theorem 2.2.4. *Let $0 < s < r$, and $\{X_n\}_{n\in\mathbb{N}}$ satisfy axioms 2.2.1. Then*

$$\mathcal{A}_q^{r-s}\left(\mathcal{A}_q^s(X, \{X_n\}_{n\in\mathbb{N}}), \{X_n\}_{n\in\mathbb{N}}\right) = \mathcal{A}_q^r(X, \{X_n\}_{n\in\mathbb{N}})$$

with equivalent norms.

Proof. To keep the notation from obscuring the arguments, we shall choose a fixed $0 < q \leq \infty$, and write

$$Z^\sigma = \mathcal{A}_q^\sigma(X, \{X_n\}_{n\in\mathbb{N}}) \qquad \forall\ \sigma \in (0, +\infty).$$

Given an element $v \in Z^s$, we define the error of approximation *in* Z^s by

$$\tilde{E}_n(v) := \inf_{x\in X_n} \|v - x\|_{Z^s}.$$

Suppose for the moment that we have shown that if $\rho > s$, then the Jackson inequality,

$$\tilde{E}_n(f) \lesssim n^{-(\rho-s)}\|f\|_{Z^\rho} \qquad \forall\ f \in Z^\rho, \tag{2.7}$$

and the Bernstein estimate

$$\|S\|_{Z^\rho} \lesssim n^{\rho-s}\|S\|_{Z^s} \qquad \forall\ S \in X_n, \tag{2.8}$$

hold. Then if $\rho > r > s$, we obtain from theorem 2.2.1 that

$$\mathcal{A}_q^{r-s}(Z^s, \{X_n\}_{n\in\mathbb{N}}) = (Z^s, Z^\rho)_{\left(\frac{r-s}{\rho-s}\right),q}.$$

[2] We have not found this result in the literature, and thus we prove it here.

So choose $\sigma > \rho$, and use theorem 2.2.3 to observe that

$$Z^s = (X, Z^\sigma)_{\frac{s}{\sigma},q}, \qquad Z^\rho = (X, Z^\sigma)_{\frac{\rho}{\sigma},q}.$$

The reiteration theorem now gives

$$\begin{aligned}\mathcal{A}_q^{r-s}(Z^s, \{X_n\}_{n\in\mathbb{N}}) &= \left((X, Z^\sigma)_{\frac{s}{\sigma},q}, (X, Z^\sigma)_{\frac{\rho}{\sigma},q}\right)_{(\frac{r-s}{\rho-s}),q} \\ &= (X, Z^\sigma)_{\frac{r}{\sigma},q} \\ &= \mathcal{A}_q^r(X, \{X_n\}).\end{aligned}$$

To finish the proof, it only remains to show that (2.7) and (2.8) hold. We will do so only for $0 < q < +\infty$, since the case $q = +\infty$ is straightforward. To prove (2.8), let $S \in X_n$, and $0 < q < \infty$. Since $S \in X_n$, it holds that $E_k(S) = 0$ if $k \geq n$, so that

$$|S|_{Z^\rho} = \left(\sum_{k=1}^{n-1} [E_k(S)k^\rho]^q \frac{1}{k}\right)^{\frac{1}{q}}.$$

But since $S \in Z^s$, we also have $E_k(S) \lesssim k^{-s}\|S\|_{Z^s}$, and substituting this expression above we obtain

$$|S|_{Z^\rho} \lesssim n^{\rho-s}\|S\|_{Z^s},$$

from where (2.8) follows.

To prove (2.7), we begin by observing that

$$\inf_{x\in X_k} E_j(f-x) \leq E_j(f), E_k(f), \tag{2.9}$$

which follows from the properties of the infimum. Also, since $X_n \subset X_{n+1}$, one has that $E_{j_1}(f) \geq E_{j_2}(f)$ whenever $j_2 \geq j_1$, and so we see that

$$\begin{aligned}&\inf_{x\in X_k} \left(\sum_{j=1}^{k} [E_j(f-x)j^s]^q \frac{1}{j}\right)^{\frac{1}{q}} \\ &\leq \inf_{x\in X_k} \left(\sum_{j=1}^{k} [E_1(f-x)j^s]^q \frac{1}{j}\right)^{\frac{1}{q}} \\ &\leq E_k(f)k^s.\end{aligned} \tag{2.10}$$

Now, consider the following computation,

$$\begin{aligned}\tilde{E}_k(f) &= \inf_{x\in X_k} \|f-x\|_{Z^s} \\ &= \inf_{x\in X_k} \left\{ \|f-x\|_X + \left(\sum_{j=1}^{\infty} [E_j(f-x)j^s]^q \frac{1}{j}\right)^{\frac{1}{q}} \right\} \\ &= \inf_{x\in X_k} \left\{ \|f-x\|_X + \left(\sum_{j=1}^{k} [E_j(f-x)j^s]^q \frac{1}{j} + \sum_{j=k+1}^{\infty} [E_j(f)j^s]^q \frac{1}{j}\right)^{\frac{1}{q}} \right\},\end{aligned}$$

where we have used (2.9) in the last step. Write $F(x)$ for the last expression in curly braces and note carefully that

$$\inf_{x\in X_k} F(x) \geq E_k(f) + \left(\sum_{j=1}^{k}[E_k(f)j^s]^q\frac{1}{j} + \sum_{j=k+1}^{\infty}[E_j(f)j^s]^q\frac{1}{j}\right)^{\frac{1}{q}} =: L.$$

Our next step will be to prove that in fact $\inf_{x\in X_k} F(x) = L$.

Let $\{x_n\}_{n\in\mathbb{N}} \subset X_k$ be such that $\|f - x_n\|_X - \epsilon/n \leq E_k(f) \leq \|f - x_n\|$, where $\epsilon > 0$ was chosen in such a way that $E_k(f) - \epsilon > 0$. Observe also that if $j < k$, then

$$E_j(f - x_n) \geq E_k(f) \geq E_j(f - x_n) - \frac{\epsilon}{n}. \tag{2.11}$$

These rather awkward steps are needed because we do not know enough about the sets X_k to be able to choose $x^* \in X_k$ such that $\|f - x^*\| = E_k(f)$.

Now, observe that

$$\begin{aligned}
F(x_n) &= \|f - x_n\|_X + \left(\sum_{j=1}^{k}[E_j(f - x_n)j^s]^q\frac{1}{j} + \sum_{j=k+1}^{\infty}[E_j(f)j^s]^q\frac{1}{j}\right)^{\frac{1}{q}} \\
&\geq E_k(f) + \left(\sum_{j=1}^{k}[E_k(f)j^s]^q\frac{1}{j} + \sum_{j=k+1}^{\infty}[E_j(f)j^s]^q\frac{1}{j}\right)^{\frac{1}{q}} \qquad (= L) \\
&\geq \|f - x_n\|_X - \frac{\epsilon}{n} - \left(\sum_{j=1}^{k}\left[\left(E_j(f - x_n) - \frac{\epsilon}{n}\right)j^s\right]^q\frac{1}{j} + \sum_{j=k+1}^{\infty}[E_j(f)j^s]^q\frac{1}{j}\right)^{\frac{1}{q}}.
\end{aligned}$$

Letting $n \to +\infty$ shows that indeed $\inf_{x\in X_k} F(x) = L$.

Sumarizing, we have that

$$\begin{aligned}
\tilde{E}_k(f) &= E_k(f) + \left(\sum_{j=1}^{k}[E_k(f)j^s]^q\frac{1}{j} + \sum_{j=1}^{\infty}[E_j(f)j^s]^q\frac{1}{j}\right)^{\frac{1}{q}} \\
&\leq E_k(f) + \left(E_k^q(f)k^{sq} + \sum_{j=k+1}^{\infty}[E_j(f)j^s]^q\frac{1}{j}\right)^{\frac{1}{q}},
\end{aligned}$$

where we have used (2.10). But if $f \in Z^\rho$, then $E_k(f) \lesssim k^{-(\rho-s)}\|f\|_{Z^\rho}$, so that, after some computations, we obtain from the above that

$$\tilde{E}_k(f) \lesssim k^{-(\rho-s)}\|f\|_{Z^\rho}.$$

□

2.3 Besov and Sobolev spaces as B-spline approximation spaces

As an alternative to the classical definitions, one can characterize Besov spaces, and for a useful range of parameters also Sobolev spaces, as approximation spaces. The results we cite here all refer, as they are found in the literature, to approximation using linear spaces of smooth, piecewise polynomial functions. But by corollary 2.2.2, they also apply to other types of methods. This will allow us to draw fairly general conclusions from the study of piecewise polynomial approximation alone.

2.3.1 B-splines

Let $N_1 : \mathbb{R} \to \mathbb{R}$ be given by $\chi_{[0,1)}$, where χ_Ω is the characteristic function of the set Ω. For $m \geq 2$, let

$$N_m := N_{m-1} * N_0.$$

The functions N_m, $m = 1, 2, \ldots$ are called the m-th order cardinal B-spline generator. Note that the space

$$S_j^m := \text{clos}\,(\text{span}\{N_m(2^j \cdot -k) : k \in \mathbb{Z}\}),$$

where we have used clos $(A) = \overline{A}$ as an alternative notation for closure, is a subspace of $C^{m-2}(\mathbb{R})$ if $m \geq 2$, and that $f \in S_j^m$ is a polynomial of degree $m-1$ on every interval of the form $2^{-j}[z, z+1)$, $z \in \mathbb{Z}^d$.

The spaces S_j^m reproduce locally any polynomial of degree $m-1$. That is, if $p \in \Pi_{m-1} =$ { polynomials of degree $m-1$}, and given a bounded set $X \subset \mathbb{R}^d$, there exists a function $\phi \in S_j^m$ such that $\phi_{|X} = p_{|X}$.

We extend the definition of the spaces S_j^m to $\mathbb{R}^d$ simply by letting

$$N_m^{(d)}(x_1, x_2, \ldots, x_d) := \prod_{i=1}^{d} N_m(x_i),$$

and setting $S_j^{m,(d)} := \text{clos}\,(\text{span}\{N_m^{(d)}(2^j \cdot -z) : z \in \mathbb{Z}^d\})$. In the sequel we will usually omit the index d, since it will be clear from the context.

2.3.2 Besov spaces

A common definition of Besov spaces is based on *moduli of continuity*. Since these spaces can be characterized thoroughly as approximation spaces using B-splines, and since this is the only point of view we shall take, we use this characterization as a definition instead. The remarkable connection between approximation spaces and Besov spaces was made by DeVore and Popov, see [16]. The following theorem is a version of this result which has been adapted to our needs.

Theorem 2.3.1 (DeVore and Popov, 1988). *Let $0 < p \leq +\infty$, $m \in \mathbb{N}$, and define $\sigma_{p,m,j} : L_p(\mathbb{R}^d) \to [0, +\infty)$ by*

$$\sigma_{p,m,j}(f) := \inf_{s \in S_j^m} \|f - s\|_{L_p}.$$

The following is an equivalent (quasi)-seminorm for the Besov space $B_q^s(L_p(\mathbb{R}^d))$, $0 < q \leq \infty$, $0 < s < \min\{m, m - 1 + 1/p\}$.

$$|f|_{B_q^s(L_p)} = \left(\sum_{j=-\infty}^{+\infty} [2^{js} \sigma_{p,m,j}(f)]^q \right)^{\frac{1}{q}} \tag{2.12}$$

(with the usual modification for $q = \infty$).

A Besov space $B_q^s(L_p(\mathbb{R}^d))$ is thus a collection of functions in $L_p(\mathbb{R}^d)$ which can be approximated by functions in S_k^m at a rate of $\mathcal{O}(2^{-ks})$, and such that the error of approximation $\sigma_{p,m,j}(f)$ satisfies the slightly stronger condition

$$\{2^{js} \sigma_{p,m,j}(f)\}_{j \in \mathbb{Z}} \in \ell_q(\mathbb{Z}).$$

The spaces $B_q^s(L_p(\mathbb{T}^d))$, where $\mathbb{T}^d = (\mathbb{R}/\mathbb{Z})^d$ is the $d-$dimensional torus, are defined analogously. The spaces $S_j^{m,\mathbb{T}^d}$ are now defined only for $j \geq 0$, and we define them by

$$S_j^{m,\mathbb{T}^d} = \operatorname{span} \left\{ \sum_{z \in \mathbb{Z}^d} N_m(2^j(\cdot - z) - k) : k \in \mathbb{Z}^d \right\}.$$

We also define the functionals

$$\rho_{p,m,j}(f) := \inf_{s \in S_j^{m,\mathbb{T}^d}} \|f - s\|_{L_p}, \qquad j \geq 0,$$

and then the corresponding equivalent seminorm for the space $B_q^s(L_p(\mathbb{T}^d))$ is given by

$$|f|_{B_q^s(L_p)} = \left(\sum_{j=0}^{+\infty} [2^{js} \rho_{p,m,j}(f)]^q \right)^{\frac{1}{q}}.$$

2.3.3 Besov spaces on domains

Apart from spaces defined on $\mathbb{R}^d$ and $\mathbb{T}^d$ we will also consider bounded open domains $\Omega \subset \mathbb{R}^d$ satisfying certain regularity conditions on the boundary.

Definition 2.3.2. A bounded domain $\Omega \subset \mathbb{R}^d$ is *of class* X, where $X = C^k$, $k = 0, 1, \ldots$, or $X = Lip_1$, the space of Lipschitz continuous functions, if for every $x \in \partial\Omega$ there exists $\epsilon_x > 0$, an orthogonal map $Q_x : \mathbb{R}^d \to \mathbb{R}^d$, and a function $\phi_x : \mathbb{R}^{d-1} \to \mathbb{R}$, $\phi_x \in X$, such that

$$Q^{-1}(B(x, \epsilon_x) \cap \Omega) = \{y \in Q^{-1}(B(x, \epsilon_x)) : y_d < \phi_x(y_1, \ldots, y_{d-1})\}.$$

Here we have written $B(x, \epsilon_x)$ for the open ball in $\mathbb{R}^d$ with center x and radius ϵ_x with respect to the Euclidean norm.

When Ω is of class X, we also say that $\partial\Omega$ is of class X. Often we shall also say that Ω (or $\partial\Omega$) "is X", as in "$\partial\Omega$ is C^1", since it makes the exposition easier to read and it cannot cause any confusion.

When $\partial\Omega$ is C^k, $k = 1, 2, \ldots$, then from the above discussion it follows that $\partial\Omega$ is a C^k manifold.

Since we will embed Ω into $\mathbb{T}^d$, we always assume that for some $\epsilon > 0$, the relation $\Omega \subset (\epsilon, 1-\epsilon)^d$ holds.

Given a bounded domain Ω with Lipschitz boundary, one can characterize the space $B^s_q(L_p(\Omega))$ by setting

$$\pi_{p,m,j}(f) := \inf_{s \in S^m_j} \|f - s_{|\Omega}\|_{L_p(\Omega)}.$$

and then defining a seminorm for $B^s_q(L_p(\Omega))$ as in (2.12) [17]. It is then easy to show from the above that the restriction operator

$$r_\Omega : B^s_q(L_p(\mathbb{R}^d)) \to B^s_q(L_p(\Omega)) \tag{2.13}$$

is bounded and linear for the full range of parameters.

2.3.4 Interpolation of Besov spaces

We have, for any $0 < s_1 < s_2$, $0 < q_1, q_2 \leq +\infty$, and any $0 < \theta < 1$, $0 < q \leq +\infty$, that

$$(B^{s_1}_{q_1}(L_p(\Omega)), B^{s_2}_{q_2}(L_p(\Omega)))_{\theta,q} = B^s_q(L_p(\Omega)), \tag{2.14}$$

with $s = (1-\theta)s_1 + \theta s_2$.

The above holds for Lipschitz domains as well as for $\Omega \in \{\mathbb{R}^d, \mathbb{T}^d\}$.

2.3.5 Sobolev spaces

The classical Sobolev spaces measure smoothness of functions in L_p, $p \geq 1$, by counting its number of weak derivatives in L_p. The definition is, for $1 \leq p \leq +\infty$, $m = 0, 1, \ldots$m

$$W^m_p(\Omega) = \{f \in L_p(\Omega) : \|f\|^p_{W^m_p} := \sum_{|\alpha| \leq m} \|D^\alpha f\|^p_{L_p} < +\infty\}.$$

In this thesis we shall restrict ourselves to the case $p = 2$, and write, as is customary, $H^m = W^m_2$. Sobolev spaces with positive non-integer smoothness index can be obtained simply by interpolation. After realizing that $B^m_2(L_2) = H^m$, we use (2.14) above to obtain

$$H^s = B^s_2(L_2). \tag{2.15}$$

Note, however, that this is not as simple for the spaces W^s_p, with $p \neq 2$. See again [16].

It would be quite an omission not to mention that the spaces H^s are Hilbert spaces. See [34] p.209 for instance.

Another important Sobolev space is the space $H_0^s(\Omega)$, which we define as follows.

Let $X \subset \mathbb{R}^d$ be a set, and let

$$\mathcal{D}(X) := \{f \in C^\infty(X) \text{ such that supp } f \subset K \subset X \text{ for some compact set } K\}$$

be the space of *test functions*. For $s \geq 0$, we define the space $H_0^s(\Omega)$ as the completion of $\mathcal{D}(\Omega)$ in $H^s(\Omega)$. For $0 \leq s < \frac{1}{2}$, or when Ω is either $\mathbb{R}^d$, $\mathbb{T}^d$, or a C^k manifold with $k > s$, the space $H_0^s(\Omega)$ coincides with $H^s(\Omega)$. In all other cases the space $H_0^s(\Omega)$ is a closed subspace of $H^s(\Omega)$.

The duals of the spaces $H_0^s(\Omega)$, $s \geq 0$ are denoted by $H^{-s}(\Omega)$.

The interpolation of the spaces $H_0^s(\Omega)$ is a more delicate matter. See [28] for further information.

2.4 Extension

We have already mentioned that the restriction operator (2.13) is bounded and linear for the full range of parameters. But there exist also, for the full range of parameters, operators

$$\mathcal{E} : B_q^s(L_p(\Omega)) \to B_q^s(L_p(\mathbb{T}^d)) \tag{2.16}$$

such that $r_\Omega(Eu) = u$ for all $u \in B_q^s(L_p(\Omega))$ For the case $p < 1$, however, it does not seem possible to find *linear* $\mathcal{E}$; see again [17].

Given a bounded domain Ω with Lipschitz boundary, and any $l \in \mathbb{N}$, it is possible to construct a bounded linear extension operator

$$\mathcal{F}_l : L_2(\Omega) \to L_2(\mathbb{R}^d)$$

such that

$$\mathcal{F}_l : H^l(\Omega) \to H^l(\mathbb{R}^d)$$

is also bounded [5]. By interpolation we obtain then that

$$\mathcal{F}_l : B_q^s(L_2(\Omega)) \to B_q^s(L_2(\mathbb{R}^d))$$

is a bounded linear operator for $0 < s < l$, $0 < q \leq +\infty$.

2.4.1 Traces

Given $u \in H^s(\Omega)$, and s sufficiently large, we can define and deal with quantities of the kind $u_{|\partial\Omega}$, or $\frac{\partial u}{\partial \mathbf{n}}$, where $\mathbf{n}$ denotes the outward normal at a point in $\Gamma := \partial\Omega$. Before doing so, we define Sobolev spaces on manifolds.

The family $\mathcal{U} = \{B(x, \epsilon_x)\}_{x \in \partial\Omega}$, given by definition 2.3.2, which consists of a selection of neighborhoods of x where we can parametrize $\partial\Omega$ by functions of class $X \in Lip_1, C^1, C^2, \ldots$, is an open covering of $\partial\Omega$. Thus there exist $x_i \in \partial\Omega$, $\epsilon_i > 0$, $i = 1, 2, \ldots, l$ such that

$\partial\Omega \subset \bigcup_{i=1}^{l} B(x_i, \epsilon_i)$. Remember that, associated to each pair x_i, ϵ_i, we have an orthogonal transformation Q_i and a function $\phi_i \in X$ such that

$$Q_i^{-1}(B(x_i, \epsilon_i) \cap \Omega) = \{y \in Q^{-1}(B(x_i, \epsilon_i)) : y_n < \phi_i(y_1, \ldots, y_n)\}.$$

Let $\{\gamma_i\}_{i=1,2,\ldots l}$, $\gamma_i \in \mathcal{D}(\mathbb{R}^d)$, be a partition of unity on $\partial\Omega$ subject to the covering $\mathcal{V} = \{B(x_i, \epsilon_i) : i = 1, 2, \ldots, l\}$. Given $f : \partial\Omega \to \mathbb{R}$ we have that $f(x) = \sum_{i=1}^{l} \gamma_i(x) \cdot f(x)$. One can define $\theta_i : \mathbb{R}^{d-1} \to \mathbb{R}^d$ by $\theta_i := Q_i(x, \phi_i(x))$, and $f_i : \mathbb{R}^{d-1} - \mathbb{R}$ by

$$f_i(x) = \begin{cases} \gamma_i(\theta_i(x)) \cdot f(\theta_i(x)) & \text{if } \theta_i(x) \in B(x_i, \epsilon_i) \cap \partial\Omega, \\ 0 & \text{otherwise,} \end{cases}$$

and then define $\|f\|_{H^s(\partial\Omega)}$ by

$$\|f\|^2_{H^s(\partial\Omega)} := \sum_{i=1}^{l} \|(\gamma_i \cdot f) \circ \phi_i\|^2_{H^s(\mathbb{R}^{d-1})}. \tag{2.17}$$

It can be shown that if $\partial\Omega$ is C^k, then the norms defined by (2.17) for different open coverings and partitions of unity are equivalent.

Remark 2.4.1. *It is possible to define, via local maps, piecewise polynomials on $\partial\Omega$. For this, we refer to [14]. We will not give any details here, but are content with remarking that it is possible, and that if $\partial\Omega$ is C^k, then we can define Besov and Sobolev spaces for $s < k$ using straight-forward adaptations of the results mentioned in subsection 2.3.2.*

We now continue with the main result of this section, as found in [34].

Theorem 2.4.2 (Trace theorem). *Let $r, l \in \mathbb{N}$, $s \in \mathbb{R}$ with $r \geq s > l - 1/2$. Let Ω be a domain with boundary of class C^r, and such that $\partial\Omega$ is bounded. Then there exists a continuous trace operator*

$$T_l : H^s(\Omega) \to \prod_{j=0}^{l} H^{s-j-1/2}(\partial\Omega) \tag{2.18}$$

with the property that

$$T_l\phi = \left(\phi_{|\partial\Omega}, \frac{\partial\phi}{\partial\mathbf{n}}, \ldots, \frac{\partial^l\phi}{\partial\mathbf{n}^l}\right) \tag{2.19}$$

for any $\phi \in C^\infty(\Omega)$. This operator has a continuous right inverse.

The proof of this theorem essentially extends the map given in (2.19) by continuity to the full operator T_l given in (2.18). Thus, when embedding a bounded domain Ω in a larger domain (say $X = \mathbb{T}^d$, or $X = \mathbb{R}^d$), we define the traces on $\partial\Omega$ of functions in $H^s(X)$ analogously, extending by continuity the appropriate analogon $\tilde{T}$ of (2.19). It follows from this construction that under the hypothesis of the trace theorem, if $u \in H^s(X)$, then

$$\tilde{T}u = T(r_\Omega u).$$

Note also that theorem 2.4.2 does not hold if $s \leq l - 1/2$; if this is the case then the map $\tilde{T}$ cannot be extended continuously any longer. See [28].

2.5 Second order elliptic boundary value problems

Consider the second order differential operator

$$Au = \sum_{i,j=1}^{d} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u \tag{2.20}$$

with $a_{ij}, b_i, c \in C^\infty$. We assume that A is *uniformly elliptic*, that is, that there exists $\alpha > 0$ such that

$$\sum_{i,j=1}^{\infty} v_i a_{ij}(x) v_j > \alpha \|v\|^2 \qquad \forall\, x, v \in \mathbb{R}^d.$$

It is often useful to write (2.20) in *divergence form*,

$$Au = \sum_{0 \le |\sigma|, |\gamma| \le 1} (-1)^{|\sigma|} D^\sigma \left(\tilde{a}_{\sigma\gamma}(x) D^\gamma u\right), \tag{2.21}$$

which is always possible for some $\tilde{a}_{\sigma\gamma} \in C^\infty$, $0 \le |\sigma|, |\gamma| \le 1$. A is then uniformly elliptic whenever

$$\sum_{|\sigma|,|\gamma|=1} v^\sigma \tilde{a}_{\sigma\gamma}(x) v^g \ge \theta \|v\|^2 \qquad \forall\, v, x \in \mathbb{R}^d$$

for some $\theta > 0$.

The derivatives involved in the definition of A are meant in the sense of distributions. Thus A is defined as $A : \mathcal{D}'(X) \to \mathcal{D}'(X)$, with X either $\mathbb{R}^d$, $\mathbb{T}^d$, or a bounded domain $\Omega \subset \mathbb{R}^d$. The following fact will be useful later (see [33], page 76).

Theorem 2.5.1. *The operator $A : H^s(\mathbb{T}^d) \to H^{s-2}(\mathbb{T}^d)$ is bounded and has closed range for every $s \in \mathbb{R}$. Furthermore,* $\dim(\mathcal{N}(A)) < +\infty$, *and* $\dim(\mathcal{N}(A)) = \dim(\mathcal{R}(A)^\perp)$.

We should stress that the above regularity assumptions are made for simplicity, and that they are not essential. It would be enough for the development of the theory in chapter four if we had that $A : H^s(\mathbb{T}^d) \to H^{s-2}(\mathbb{T}^d)$ is bounded and has closed range for all $s \in [s_0, 2]$ and some $s_0 > 2$ (in particular for theorem 4.3.8). But choosing the stronger assumptions alleviates us from the burden of tracking yet another parameter.

Sometimes we will place additional assumptions on the operator A, in particular when dealing with the *weak formulation*; see 2.5.2

In this thesis we are concerned with the solution of the following type of problem. Let Ω be a bounded domain. Given f, find u such that

$$Au = f \qquad \text{on } \Omega, \tag{2.22}$$

subject to one of the following boundary conditions.

Either *Neumann* boundary conditions

$$B^{\mathcal{N}} u = \frac{\partial u}{\partial \mathbf{n}} = g, \tag{2.23}$$

or *Dirichlet* boundary conditions,

$$B^{\mathcal{D}}u = u_{|\partial\Omega} = g, \tag{2.24}$$

for g given.

Equation (2.22) together with (2.23) is called a *Neumann problem*. Equation (2.22) together with (2.24) is called a *Dirichlet problem*.

2.5.1 Strong solutions

A solution u of the Dirichlet or the Neumann problem is a *strong solution*[3] if the equalities (2.22), together with (2.24) or (2.23), respectively, hold *almost everywhere*, and $Au, f \in L_2$. The situation is particularly simple when Ω has C^∞ boundary (see [28]).

Theorem 2.5.2. *Let $s \geq 0$. Then the operators $\mathcal{P}^{\mathcal{D}} : H^{s+2}(\Omega) \to H^s(\Omega) \times H^{s+3/2}(\partial\Omega)$ and $\mathcal{P}^{\mathcal{N}} : H^{s+2}(\Omega) \to H^s(\Omega) \times H^{s+1/2}(\partial\Omega)$, given by*

$$\mathcal{P}^{\mathcal{D}} = \begin{pmatrix} A \\ B^{\mathcal{D}} \end{pmatrix} \qquad\qquad \mathcal{P}^{\mathcal{N}} = \begin{pmatrix} A \\ B^{\mathcal{N}} \end{pmatrix}$$

are bounded, have finite-dimensional kernels, and their ranges are closed with finite codimension. In particular, one has that $\mathcal{P}^{\mathcal{D}}$ and $\mathcal{P}^{\mathcal{N}}$ are isomorphisms between $\mathcal{N}(\mathcal{P}^{\mathcal{D}})^\perp$ and $\mathcal{R}(\mathcal{P}^{\mathcal{D}})$, and between $\mathcal{N}(\mathcal{P}^{\mathcal{N}})^\perp$ and $\mathcal{R}(\mathcal{P}^{\mathcal{N}})$, respectively.

Above we have used the notation $\mathcal{N}(F)$ for the kernel of an operator F, and $\mathcal{R}(F)$ for its range.

2.5.2 Weak formulation

Let $u \in H^1(\Omega)$. Then the distribution Au cannot always be identified with a measurable function. The weak formulation allows us to handle this case.

For $\phi \in \mathcal{D}(\Omega)$, we have by the definition of distributional derivative that

$$\langle Au, \phi \rangle = [Au](\phi) = \sum_{0 \leq |\sigma|, |\gamma| \leq 1} \int_\Omega \tilde{a}_{\sigma\gamma}(x) D^\sigma u D^\gamma \phi d\mu.$$

We can now define the (bounded) symmetric bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$, associated with A, by

$$a(u, v) := \langle Au, v \rangle.$$

We assume also that A is *coercive*, that is, that there exists $\alpha > 0$ such that

$$a(u, u) > \alpha \|u\|^2_{H_0^1(\Omega)} \qquad \forall\, u \in H_0^1(\Omega).$$

[3] There seems to be some disagreement over the definition of a strong solution. We use here the one found in [34], p. 287.

Under these circumstances we invoke the Lax-Milgram lemma, and have then that for each $f \in H^{-1}(\Omega)$ there exists a unique $u \in H_0^1(\Omega)$ such that

$$a(u, v) = \langle f, v \rangle \qquad \forall\, v \in H_0^1(\Omega).$$

We will say that this u is a *weak solution* of the problem

$$\begin{aligned} Au &= f && \text{on } \Omega \\ u_{|\partial\Omega} &= 0. \end{aligned}$$

Given $g \in H^{1/2}(\partial\Omega)$, we can use the Trace theorem to find $u_g \in H^1(\Omega)$ such that $(u_g)_{|\partial\Omega} = g$. But then from the above discussion it follows that there exists a unique $u^* \in H_0^1(\Omega)$ such that

$$a(u^*, v) = \langle f - Au_g, v \rangle \qquad \forall\, v \in H_0^1(\Omega).$$

Now $u = u^* + u_g$ (which can be seen to be independent of the choice of u_g) satisfies $Au = f$ and also $u_{|\partial\Omega} = g$. Thus, we call it a weak solution of the problem

$$\begin{aligned} Au &= f && \text{on } \Omega \\ u_{|\partial\Omega} &= g, \end{aligned}$$

noting that a strong solution is also a weak solution.

We have the following

Theorem 2.5.3. *If A is coercive, then the operator $\mathcal{P}^{\mathcal{D}} : H^1(\Omega) \to H^{-1} \times H^{1/2}(\partial\Omega)$, given by*

$$\mathcal{P}^{\mathcal{D}} = \begin{pmatrix} A \\ B^{\mathcal{D}} \end{pmatrix}$$

is an isomorphism.

It is not possible to construct a similar theory for the Neumann problem. The operator $B^{\mathcal{N}}$ is not bounded on $H^1(\Omega)$.

2.6 B-spline wavelet bases

The type of wavelets we will use is a family of Riesz bases for Sobolev spaces and their duals. We sketch here the construction of pairs of biorthogonal wavelet bases for $L_2(\mathbb{R})$, and show how this construction can be extended to the multivariate and periodic cases. Finally, we show how to produce wavelet bases for Sobolev spaces on these domains.

Since they play no role in the rest of this thesis, we have omitted various important constructions, like wavelets on more general domains, or wavelets on manifolds. Still, we include a fairly detailed account of the construction of B-spline wavelet bases, since some of the details play a central role later on.

For a thorough introduction to the material from which the summary in this section draws, see [11].

2.6.1 Riesz bases

A Riesz basis for a (separable) Hilbert space $\mathcal{H}$ is a countable collection $F = \{f_\lambda\}$, with λ in some index set ∇, such that the map $T : \ell_2(\nabla) \to \mathcal{H}$ given by

$$T(\{x_\lambda\}) = \sum_{\ell \in \nabla} x_\lambda f_\lambda$$

is an isomorphism. It follows that there exists a *dual* Riesz basis $\tilde{F} = \{\tilde{f}_\lambda\}$ in $\mathcal{H}'$ such that for every $g \in \mathcal{H}$, and every $h \in \mathcal{H}'$, we obtain

$$g = \sum_{\lambda \in \nabla} \langle \tilde{f}_\lambda, g \rangle f_\lambda \qquad\qquad h = \sum_{\lambda \in \nabla} \langle h, f_\lambda \rangle \tilde{f}_\lambda, \tag{2.25}$$

where we have written $\langle \cdot, \cdot \rangle$ for the dual pairing between $\mathcal{H}$ and $\mathcal{H}'$. Relations (2.25) imply that $\langle f_\lambda, f_\mu \rangle = \delta_{\lambda\mu}$, where $\delta_{\lambda\mu}$ is the Kronecker delta.

Since F and $\tilde{F}$ both induce isomorphisms between ℓ_2 and $\mathcal{H}$, $\mathcal{H}'$, respectively, we obtain the norm equivalences

$$\|g\|_{\mathcal{H}} \sim \left(\sum_{\lambda \in \nabla} |\langle \tilde{f}_\lambda, g \rangle|^2 \right)^{1/2}, \qquad\qquad \|h\|_{\mathcal{H}'} \sim \left(\sum_{\lambda \in \nabla} |\langle h, f_\lambda \rangle|^2 \right)^{1/2}.$$

2.6.2 Multiresolution analysis

A *multiresolution analysis* (MRA) in $L_2(\mathbb{R})$ is a sequence of closed subspaces $\{V_j\}_{j \in \mathbb{Z}}$ that satisfies the following axioms.

Axioms 2.6.1

I. $V_j \subset V_{j+1}$, for all $j \in \mathbb{Z}$

II. $\cap_j V_j = \{0\}$

III. $\overline{\cup_j V_j} = L^2(\mathbb{R})$

IV. if $f \in V_j$, then $f(2 \cdot) \in V_{j+1}$

V. if $f \in V_0$, then $f(\cdot - k) \in V_0$ for all $k \in \mathbb{Z}$

VI. there exists $\psi^0 \in V_0$ such that the set $\{\psi^0(\cdot - k) : k \in \mathbb{Z}\}$ is a Riesz basis for V_0. This function is called the *scaling function*[4] of the MRA $\{V_j\}_{j \in \mathbb{Z}}$.

[4] Here we have taken the liberty to denote the scaling function by ψ^0, departing from the tradition which uses ϕ. It will be seen that doing so simplifies the notation greatly, in particular when handling multivariate wavelet bases.

A *pair of biorthogonal MRAs* $\{V_j\}$, $\{\tilde{V}_j\}$ is a pair of MRAs whose corresponding scaling functions ψ^0, $\tilde{\psi}^0$ satisfy $\langle \psi^0(\cdot - k), \tilde{\psi}^0(\cdot - l)\rangle = \delta_{kl}$ for all $k, l \in \mathbb{Z}$. Such a pair defines a sequence of oblique projectors $Q_j : L_2(\mathbb{R}) \to V_j$, $\tilde{Q}_j : L_2(\mathbb{R}) \to \tilde{V}_j$, given by

$$Q_j f = \sum_{k\in\mathbb{Z}} \langle \xi_j \tilde{\psi}^0(2^j \cdot - k, f\rangle \xi_j \psi^0(2^j \cdot - k),$$
$$\tilde{Q}_j f = \sum_{k\in\mathbb{Z}} \langle f, \xi_j \psi^0(2^j \cdot - k)\rangle \xi_j \tilde{\psi}^0(2^j \cdot - k),$$

where the scaling $\xi_j = 2^{-j/2}$ ensures that $\|\xi_j \psi^0(2^j \cdot - k)\|_{L_2} \sim 1$. We will say that a pair of biorthogonal MRAs is *admissible* if the projectors Q_j, $\tilde{Q}_j$ are uniformly bounded for $j \in \mathbb{Z}$.

Let $W_j = \mathcal{R}(Q_{j+1} - Q_j)$, and $\tilde{W}_j = \mathcal{R}(\tilde{Q}_{j+1} - \tilde{Q}_j)$. These spaces satisfy that

$$V_{j+1} = V_j \oplus W_j \qquad\qquad \tilde{V}_{j+1} = \tilde{V}_j \oplus \tilde{W}_j$$

while

$$V_j \perp \tilde{W}_j \qquad\qquad \tilde{V}_j \perp W_j.$$

We further have that (see [11])

$$\left(\sum_{j\in\mathbb{Z}} \|(Q_{j+1} - Q_j)f\|_{L_2}\right)^{\frac{1}{2}} \sim \left(\sum_{j\in\mathbb{Z}} \left\|(\tilde{Q}_{j+1} - \tilde{Q}_j)f\right\|_{L_2}\right)^{\frac{1}{2}} \sim \|f\|_{L_2}. \tag{2.26}$$

It turns out that it is possible to find functions $\psi^1 \in W_0$, $\tilde{\psi}^1 \in \tilde{W}_0$, such that their integer translates form a biorthogonal pair of Riesz bases for W_0, $\tilde{W}_0$, respectively. Writing $\psi^e_{jk} = \xi_j \psi^e(2^j \cdot - k)$, where $e \in \{0, 1\}$, $j, k \in \mathbb{Z}$, we can express the projectors $Q_{j+1} - Q_j$, $\tilde{Q}_{j+1} - \tilde{Q}_j$ simply through

$$(Q_{j+1} - Q_j)f = \sum_{k\in\mathbb{Z}} \langle \tilde{\psi}^1_{jk}, f\rangle \psi^1_{jk}$$
$$(\tilde{Q}_{j+1} - \tilde{Q}_j)f = \sum_{k\in\mathbb{Z}} \langle f, \psi^1_{jk}\rangle \tilde{\psi}^1_{jk}.$$

From this, and from (2.26), it follows that the collections

$$\Psi = \{\psi^1_{jk} : j, k \in \mathbb{Z}\} \qquad\qquad \tilde{\Psi} = \{\tilde{\psi}^1_{jk} : j, k \in \mathbb{Z}\}$$

constitute a pair of biorthogonal Riesz bases for L_2. The bases Ψ, $\tilde{\Psi}$, are called *wavelet bases*, and the functions ψ^1, $\tilde{\psi}^1$ are called the *mother wavelets* of theses bases.

Given a pair of (admissible) biorthogonal MRAs, we can obtain corresponding mother wavelets as follows.

First, we realize that from axioms 2.6.1, IV it follows that ψ^0, $\tilde{\psi}^0$ satisfy the equations

$$\psi^0(x) = \sum_{k\in\mathbb{Z}} a^0_k \psi^0(2 \cdot - k), \qquad\qquad \tilde{\psi}^0(x) = \sum_{k\in\mathbb{Z}} \tilde{a}^0_k \tilde{\psi}^0(2 \cdot - k),$$

for some sequences $\{a_k^0\}$, $\{\tilde{a}_k^0\}$. These sequences are called the *masks* of their respective functions. It is clear that if these functions are compactly supported, then only a finite number of entries in their masks can be nonzero. Now, let $\{a_k^1\}$, $\{\tilde{a}_k^1\}$, be the sequences whose entries are given by

$$a_k^1 = (-1)^k \tilde{a}_{1-k}^0, \qquad \tilde{a}_k^1 = (-1)^k a_{1-k}^0.$$

One possible choice for ψ^1, $\tilde{\psi}^1$, is then

$$\psi^1 = \sum_{k\in\mathbb{Z}} a_k^1 \psi^0(2\cdot - k), \qquad \tilde{\psi}^1 = \sum_{k\in\mathbb{Z}} \tilde{a}_k^1 \tilde{\psi}^0(2\cdot - k).$$

Note that whenever both ψ^0, $\tilde{\psi}^0$, are compactly supported, so are ψ^1, $\tilde{\psi}^1$.

2.6.3 B-spline wavelet bases

The spaces S_j^m, defined in 2.3.1, satisfy the definition of multiresolution analysis. To satisfy axiom VI, it is customary to choose

$$\psi^0 = N_m(x + \left\lfloor \frac{m+1}{2} \right\rfloor).$$

It is an easy exercise to compute the mask of this function. An observation which plays an important role later on is that all elements of the mask of this ψ^0 are non-negative.

The construction of the dual MRA is not at all simple. See [9] for details. Suffice it to say that for $\tilde{m} \in \mathbb{N}$, with $m + \tilde{m}$ even and $\tilde{m} \geq m$, there exists a compactly supported scaling function $\tilde{\psi}$ which reproduces polynomials of degree $\tilde{m} - 1$, and such that the spaces $V_j = S_j^m$, together with the spaces

$$\tilde{V}_j = \text{span}\,\{\tilde{\psi}^0(2^j \cdot - k) : k \in \mathbb{Z}\}$$

define an admissible pair of biorthogonal MRAs.

2.6.4 The multivariate and periodic cases

Let $\{V_j\}$, $\{\tilde{V}_j\}$ be a pair of biorthogonal MRAs, and let $d > 1$ be an integer. Write $x = (x^1, x^2, \ldots, x^d) \in \mathbb{R}^d$, let $E = \{0,1\}^d$, and consider the functions $\psi^e(x) = \psi^{e_1}(x^1)\psi^{e_2}(x^2)\cdots\psi^{e_d}(x^d)$, $\tilde{\psi}^e(x) = \tilde{\psi}^{e_1}(x^1)\tilde{\psi}^{e_2}(x^2)\cdots\tilde{\psi}^{e_d}(x^d)$ for $e \in E$. We will always use 0 to denote the element in E whose coordinates are all zero. This abuse of notation is very useful, and it never seems to cause any confusion.

The spaces $V_j^0 = \text{span}\,\{\psi^0(2^j \cdot - k) : k \in \mathbb{Z}^d\}$ form a MRA, and with the dual spaces $\{\tilde{V}_j^0\}$ (defined analogously) they form a pair of biorthogonal MRAs. The complement spaces W_j^0 such that $V_{j+1}^0 = V_j^0 \oplus W_j^0$ are spanned by the integer translates of the functions ψ^e with $e \in E \setminus \{0\}$. Using the scaling factor $\xi_j = 2^{\frac{dj}{2}}$, the functions $\{\psi_{jk}^e : j \in \mathbb{Z} \wedge k \in \mathbb{Z}^d \wedge e \in E \setminus (0,0,\ldots,0)\}$, with $\psi_{jk}^e = \xi_j \psi^e(2^j \cdot - k)$, form a Riesz basis of the space $L^2(\mathbb{R}^d)$.

Let

$$\psi_{jk}^{e(\mathbb{T}^d)}(x) = \sum_{z\in\mathbb{Z}} \psi_{jk}^{e}(x-z).$$

The spaces $V_j^{\mathbb{T}} = \text{span}\,\{\psi_{jk}^{0(\mathbb{T}^d)} : k \in \mathcal{Z}_j^d\}$, where we have written $\mathcal{Z}_j^d = \mathbb{Z}^d/2^j\mathbb{Z}^d$, satisfy all the axioms for a MRA except for axiom II, as this definition of V_j does not make sense for $j \leq 0$. Usually, this axiom is just deleted, and one contents oneself with a Riesz basis that includes the scaling functions on V_0. We still have that $\{V_j^{\mathbb{T}}\}_{j\geq 0}$, $\{\tilde{V}_j^{\mathbb{T}}\}_{j\geq 0}$ form a pair of biorthogonal MRAs, for and that the set $\{\psi_{0,0}^{0(\mathbb{T}^d)}\} \cup \{\psi_{jk}^{e(\mathbb{T}^d)} : j \in \mathbb{N} \wedge k \in \mathcal{Z}_j^d\}$ forms a Riesz basis of the space $L^2(\mathbb{T})$. We will drop the $\mathbb{T}^d$ superscript from now on, since it will become clear from the context which set of functions are used.

In the notation of 2.6.1, we have

$$\nabla = \{\lambda = (e,j,k) : e \in \{0,1\}^d, j \in \mathbb{N}, k \in \mathcal{Z}_j^d, \text{ with } e = 0 \text{ only if } j = 0\}.$$

Thus we write ψ_λ, with $\lambda = (e,k,j)$ instead of ψ_{jk}^e. We also use the notation $|\lambda| := j$ for the *level* of ψ_λ. Sometimes it is useful to consider only indices up to a certain level, or indices only on one level. We denote this by

$$\nabla_j = \{\lambda \in \nabla : |\lambda| < j\} \qquad\qquad \nabla_j^0 = \{\lambda \in \nabla : |\lambda| = j\}$$

2.6.5 Wavelet bases for Sobolev spaces

The construction of wavelet bases for Sobolev spaces from bases for L_2 amounts to rescaling. The fundamental result is the following theorem (See [11], 108-117). To avoid needless complications, we will only write it for spaces $H^s(X)$, $s \in \mathbb{R}$, defined on $X = \mathbb{R}^d$ or $X = \mathbb{T}^d$.

Theorem 2.6.1. *Consider a pair of (admissible) biorthogonal MRAs as above, together with the corresponding L_2 wavelet basis, and let*

$$\begin{aligned}
\gamma &= \sup\{s : \psi^0 \in H^s(X)\},\\
\tilde{\gamma} &= \sup\{s : \tilde{\psi}^0 \in H^s(X)\},\\
m &= \max\{r : \Pi_r \subset V_0 \ \textit{(locally)}\ \},\\
\tilde{m} &= \max\{r : \Pi_r \subset \tilde{V}_0 \ \textit{(locally)}\ \}.
\end{aligned}$$

Then, writing $r = \min\{\gamma, m\}$, $\tilde{r} = \min\{\tilde{\gamma}, \tilde{m}\}$, we obtain that for all $s \in (-\tilde{r}, r)$ the sets

$$\Psi^{(s)} = \{2^{s|\lambda|}\psi_\lambda : \lambda \in \nabla\}, \qquad\qquad \tilde{\Psi}^{(-s)} = \{2^{-s|\lambda|}\psi_\lambda : \lambda \in \nabla\},$$

form a pair of biorthogonal Riesz bases for the spaces $H^s(X)$, $H^{-s}(X)$, respectively.

When we say that Ψ is a Wavelet basis for H^s, we will assume that it has been properly scaled. That is, when we write $\psi_\lambda = \psi_{jk}^e = \xi_j \psi^e(2^j \cdot -k)$ we have

$$\xi_j = 2^{-sj}2^{jd/2}.$$

2.6.6 The fast wavelet transform

Given a pair of MRAs as above, and $f \in V_{j+1}$, $j \geq 0$, we have two representations of f available. We can either express it in terms of scaling functions in V_{j+1}, or in terms of wavelets. Here we sketch briefly how to translate from one representation to the other in the periodic case.

Given a sequence $\mathbf{x} = \{x_k\}_{k\in\mathbb{Z}^d}$, we can associate with it the matrix $M_j^{\mathbf{x}} = (m_{kl}^{\mathbf{x},j})_{k\in\mathcal{Z}_{j+1}^d, l\in\mathcal{Z}_j^d}$, whose entries are given by

$$m_{kl}^{\mathbf{x},j} = \frac{\xi_j}{\xi_{j+1}} \sum_{z\in\mathbb{Z}^d} x_{k-2l-2^{j+1}z}.$$

Note that it defines a linear map $M_j^{\mathbf{x}} : \ell_2(\mathcal{Z}_j^d) \to \ell_2(\mathcal{Z}_{j+1}^d)$

As before, let $E = \{0,1\}^d$. We will write $\mathbf{b}^e = \{b_k^e\}_{k\in\mathbb{Z}^d}$ for the sequence whose entries are

$$b_k^e = a_{k_1}^{e_1} a_{k_2}^{e_2} \cdots a_{k_d}^{e_d},$$

where $e = (e_1, e_2, \ldots, e_d)$, and $k = (k_1, k_2, \ldots, k_d)$. This sequence is just the tensor product of the corresponding 1-dimensional masks.

Note that we can write f as either

$$f = \sum_{k\in\mathcal{Z}_{j+1}^d} (c_{j+1}^0)_k \psi_{j+1,k}^0, \tag{2.27}$$

or as

$$f = \sum_{e\in E} \sum_{k\in\mathcal{Z}_j^d} (c_e^j)_k \psi_{jk}^e, \tag{2.28}$$

where the c_j^e each belong to $\ell_2(\mathcal{Z}_j^d)$.

Using the tensor product masks and the matrix mechanism defined above, we obtain that

$$c_0^{j+1} = \sum_{e\in E} M_j^{\mathbf{b}^e} c_e^j, \tag{2.29}$$

and that for $e \in E$,

$$c_e^j = \left(M_j^{\tilde{\mathbf{b}}^e}\right)^T c_0^{j+1}. \tag{2.30}$$

Relations (2.29) and (2.30) allow us to switch between the representations (2.27) and (2.28) at a cost of $\mathcal{O}(N)$ operations, with $N = 2^{(j+1)d}$. We can repeat this process for $f_j := \sum_{k\in\mathcal{Z}_j^d} (c_0^j)_k \psi_{jk}^0$, and then again analogously until $j = 0$. Then we have obtained the *wavelet representation* of f,

$$f = (c_0^0)_0 \psi_{0,0}^0 + \sum_{l=0}^{j} \sum_{e\in E\setminus\{0\}} \sum_{k\in\mathcal{Z}_j^d} (c_e^j)_k \psi_{jk}^e. \tag{2.31}$$

The cost of transforming between (2.27) and (2.31) is also $\mathcal{O}(N)$. The method we have described here is called the *fast wavelet transform*. For later use, we define

$$\nabla_j := \{(0,0,0)\} \cup \bigcup_{i=0}^{j-1} \{(j,k,e) : k \in \mathcal{Z}_j^d,\ e \in E \setminus \{0\}\},$$

which allows us to write (2.31) more succinctly as

$$f = \sum_{\lambda \in \nabla_j} c_\lambda \psi_\lambda.$$

2.6.7 Discretizing linear operator equations

The type of operator equation that we will to solve is as follows. Consider a linear, bounded operator $M : \mathcal{H}^l \to \mathcal{H}^r$ with closed range, where $\mathcal{H}^l$, $\mathcal{H}^r$ will be either Sobolev spaces, or tensor products of Sobolev spaces. We always endow the tensor product spaces with the Euclidean tensor product norm, which ensures that the resulting space is also a Hilbert space.

Given $b \in \mathcal{H}^r$, we take on the task of finding $x \in \mathcal{H}^l$ such that

$$Mx = b. \tag{2.32}$$

(Note that such a solution does not have to exist, nor does it have to be unique; we shall ignore this for the moment.)

Given a pair of isomorphisms

$$T_l : \ell_2 \to \mathcal{H}^l, \qquad T_r : \ell_2 \to \mathcal{H}^r, \tag{2.33}$$

which usually will involve wavelet bases, we can transform equation (2.32) into an equivalent system of equations by taking $\overline{M} = T_r^{-1} M T_l$, and rewrite our problem as follows. Given $b \in \mathcal{H}^r$, let $\overline{b} = T_r^{-1} b$, and find $\overline{x} \in \ell_2$ such that

$$\overline{M}\overline{x} = \overline{b}.$$

After finding $\overline{x}$, we then obtain the solution of (2.32) by taking $x = T_l \overline{x}$.

Using the fact that any isomorphism of the type (2.33) induces a Riesz basis, and that for each Riesz basis there is a biorthogonal Riesz basis, it is easy to find simple expressions for computing the entries in the matrix $\overline{M}$.

We can obtain discretizations of equation (2.32) by using pairs of biorthogonal MRAs. Suppose that $\{V_j^\sigma\}_{j\geq 0}$, $\{\tilde{V}_j^\sigma\}_{j\geq 0}$ is such a pair for $\mathcal{H}^\sigma$, $(\mathcal{H}^\sigma)'$, $\sigma = r, l$, (constructed, if needed, by taking tensor products of MRAs in the obvious way), and denote by Q_j^σ, $\tilde{Q}_j^\sigma$ their respective oblique projectors. We shall further assume that these spaces are finite dimensional. Write $M_j = Q_j^r M Q_j^l$, and consider the following discrete problem. Given an approximation $b_j \in V_j^r$ of b, find $x_j \in V_j^l$ such that

$$M_j x_j = b_j. \tag{2.34}$$

There are now two possibilities to transform (2.34) into a linear system of equations in Euclidean space. One through the scaling function representation of the elements in the respective spaces, and one using the wavelet representation. If the operator M is an invertible elliptic differential operator, then using the wavelet representation leads to a system whose condition number is uniformly bounded in j (see [11], p. 116ff).

2.7 Nonlinear approximation using Wavelets

Until now, we have considered only approximation using linear spaces. Here we will discuss in brevity approximation from *nonlinear* sets.

2.7.1 Best N-term approximation

Suppose Ψ, $\tilde{\Psi}$ are a pair of biorthogonal wavelet Riesz bases for the spaces $H^t(\mathbb{T}^d)$, $H^{-t}(\mathbb{T}^d)$, respectively, and consider the problem of approximating $f \in H^t(\mathbb{T}^d)$,

$$f = \sum_{\lambda \in \nabla} c_\lambda \psi_\lambda.$$

Let $\varphi : \mathbb{N} \to \nabla$ be a *sorting* of the coefficient vector $\{c_\lambda\}$, that is, if $n, m \in \mathbb{N}$, $m \geq n$ implies $|c_{\varphi(m)}| \leq |c_{\varphi(n)}|$. The best $N - term$ approximation of f is now defined by

$$f_{\{N\}} = \sum_{i=1}^{N} c_{\varphi(i)} \psi_{\varphi(i)}. \tag{2.35}$$

Clearly, the idea is to approximate f using only the most important coefficients of its wavelet representation, achieving, we hope, a better rate of approximation than if we approximated f by

$$f_j = \sum_{\lambda \in \nabla_j} c_\lambda \psi_\lambda \in V_j.$$

The approach (2.35) is particularly helpful when approximating functions with singularities, since the larger coefficients tend to agglomerate there.

Let us write $\Sigma_n = \{f : f = \sum_{\lambda \in A} c_\lambda \psi_\lambda, \text{with } A \subset \nabla, \#A \leq n\}$. The space $\mathcal{A}^s_\infty(H^t(\mathbb{T}^d), \{\Sigma_n\})$ consists then of all the functions $f \in H^t$ such that the convergence of its best N-term approximation is as $\mathcal{O}(N^{-s})$.

From [8] we learn that this is equivalent to the condition that the sequence $\{c_\lambda\}_{\lambda \in \nabla}$ belongs to the *weak* ℓ_τ spaces (denoted ℓ^w_τ), with $s = 1/\tau - 1/2$. That is, when

$$\#\{\lambda : |c_\lambda| \geq \epsilon\} \lesssim \epsilon^{-\tau}$$

We have the following result [8].

Theorem 2.7.1. *Let $\epsilon > 0$, and write $\tau_\epsilon = \tau + \epsilon$, $s_\epsilon = 1/\tau_\epsilon + 1/2$. Then*

$$B^{sd+t}_\tau(L_\tau) \subset \mathcal{A}^s_\infty(H^t, \{\Sigma_n\}) \subset B^{s_\epsilon d+t}_{\tau_\epsilon}(L_{\tau_\epsilon}).$$

The following characterization of ℓ_τ^w will be useful later.

Proposition 2.7.2. *Let $a > 1$. $v \in \ell_w^\tau$ if, and only if for every $j \in \mathbb{Z}$*

$$\#\{k : |v_k| \geq a^{-j}\} \lesssim a^{\tau j}$$

2.7.2 Compressible matrices, fast matrix-vector multiplication, and adaptive wavelet methods

An infinite matrix B is said to be in the class $\mathcal{B}_s$ of compressible matrices if there exist two positive summable sequences $\{\alpha_j\}_{j\in\mathbb{N}}$, $\{\beta_j\}_{j\in\mathbb{N}}$, such that for every $j \geq 0$ there exists a matrix B_j with at most $2^j\alpha_j$ nonzero entries per row and column with the property that, in the spectral norm,

$$\|B - B_j\| \leq 2^{-js}\beta_j,$$

Proposition 2.7.3. *Let $\tau = (s + \frac{1}{2})^{-1}$, with $0 < \tau < 2$. If $B \in \mathcal{B}_s$, then B maps ℓ_τ^w boundedly into itself.*

The wavelet discretizations of the regular differential operators in section 2.5 are all compressible; see [8]. The compressibility index s depends on the regularity of the primal wavelet basis and of the approximation power of the dual basis.

Another important property of a compressible matrix is that it is possible to compute its action on a sequence efficiently.

Theorem 2.7.4. *For any $v \in \ell_2$ with finite support, for any $B \in \mathcal{B}_s$, and given an accuracy $\epsilon > 0$, there exists a compactly supported sequence $w \in \ell_2$ such that*

i. $\|Bv - w\| < \epsilon$,

ii. $\|w\|_{\ell_\tau^w} \lesssim \|v\|_{\ell_\tau^w}$

iii. $\#(supp\ w) \leq C_{B,s}\epsilon^{-1/s}\|v\|_{\ell_\tau^w}^{1/s}$.

The cost of computing w stays bounded by $C_{B,s}\|v\|_{\ell_\tau^w}^{1/s}\epsilon^{-1/s} + \#\,\mathrm{supp}\ v$.

For concrete algorithms, and further information, we refer to [8].

The two last results are the key ingredients of the adaptive wavelet methods devised in [8]. We refer there and also to [7] for further details. Here we only include the following core result, which only speaks of its efficiency and convergence, and is only concerned with the problem after being transformed to a problem in ℓ_2.

Theorem 2.7.5. *Let $L : \ell_2 \to \ell_2$ be in $\mathcal{B}_s$. Assume further that L is symmetric positive definite, and consider the equation*

$$Lx = b.$$

If the solution x is in ℓ_τ^w, then given $\epsilon > 0$, the adaptive algorithm in [8] constructs a compactly supported approximation w of x such that $\|x - w\| < \epsilon$ and $\#(supp\ w) \lesssim \|v\|_{\ell_\tau^w}^{1/s}\epsilon^{-1/s}$, at a cost of at most $\mathcal{O}(\epsilon^{-1/s})$ operations.

Chapter 3

The fictitious domain Lagrange multiplier method - A case study

The fictitious domain - Lagrange multiplier (FDLM) method is a fairly popular fictitious domain method; its simplicity and good performance are appealing, and the theory behind it is very well understood. This makes it a very good example for the type of smoothness-related problems that may arise.

This is what we intend to do in this chapter: to study in depth the smoothness of the solutions obtained with the FDLM method in the fictitious domain. This solution is an extension of the solution of the original problem, and what will be shown is that, unless careful provisions are taken, this extended solution will be difficult to approximate. We will establish that the convergence rate of linear schemes based on B-splines and nonlinear schemes based on B-spline wavelets is bounded from below, independently of the order chosen. This result extends to comparable approximation schemes via corollary 2.2.2 and theorem 2.7.1, respectively.

We will begin by sketching the derivation of the FDLM method. Then we will study the results that concern linear approximation schemes, taking first a quick look at what is already known, and then extending these results to the full range of parameters. After that, we will also study the convergence rates of nonlinear schemes. In the derivation of these results, we need a central lemma which we prove in the last section, after discussing briefly how to obtain better convergence rates.

3.1 The FDLM method

Consider the following problem. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with C^1 boundary, and let $f \in [H^1(\Omega)]'$, $g \in H^{1/2}(\partial\Omega)$. We want to find $u \in H^1(\Omega)$ such that

$$\begin{aligned} Au &= f \qquad \text{on } \Omega, \\ u_{|\partial\Omega} &= g, \end{aligned} \tag{3.1}$$

where A is a uniformly elliptic second order differential operator as defined in section 2.5. We will solve problem (3.1) by embedding Ω into a larger, simpler domain Ξ, the *fictitious domain*. For simplicity we will set $\Xi = \mathbb{T}^d$.

The next step is to choose an extension $f^+ \in H^{-1}(\mathbb{T}^d)$ of f. Note that this is always possible since we have required that $f \in [H^1(\Omega)]'$. At the very least we can take $f^+ = f \circ r_\Omega$, where r_Ω is the operator which restricts functions to Ω.

A detail that needs carefull adressing is the "extension" of the differential operator defined on Ω. To this end, assume that the coefficients $\{\tilde{a}_{\gamma\nu}\}$ that define A in divergence form on Ω (see (2.21)) can be extended to $\mathbb{T}^d$ by $\{\overline{a}_{\gamma\nu}\}$, with $\overline{a}_{\sigma\gamma} \in C^\infty(\mathbb{T}^d)$, $0 \leq |\sigma|, |\gamma| \leq 1$. Now, define $A^{\mathbb{T}^d} : H^1(\mathbb{T}^d) \to H^{-1}(\mathbb{T}^d)$ by defining $A^{\mathbb{T}^d}u$ first on $C^\infty(\mathbb{T}^d)$,

$$(A^{\mathbb{T}^d}u)(\phi) := \sum_{0\leq|\sigma|,|\gamma|\leq 1} \int_{\mathbb{T}^d} \overline{a}_{\sigma\gamma}(x) D^\sigma u D^\gamma \phi d\mu, \qquad \phi \in C^\infty(\mathbb{T}^d),$$

and then extending it to $H^1(\mathbb{T}^d)$ through continuity in the usual way.

As it is a functional on $H^1(\mathbb{T}^d)$, the "restriction to Ω" of $A^{\mathbb{T}^d}u$, written $(A^{\mathbb{T}^d}u)_{|\Omega}$, is defined first for $\mathcal{D}(\Omega)$ by (see [34], page 133)

$$(A^{\mathbb{T}^d}u)_{|\Omega}(\phi) := (A^{\mathbb{T}^d}u)(\phi^0), \qquad \forall \phi \in \mathcal{D}(\Omega),$$

where we have written ϕ^0 for the extension by zero of ϕ. Afterwards, $(A^{\mathbb{T}^d}u)_{|\Omega}$ becomes a functional on $H_0^1(\Omega)$ again by continuity. Standard arguments show that it must be a bounded functional, and thus we have that $(A^{\mathbb{T}^d}u)_{|\Omega} \in H^{-1}(\Omega)$ for all $u \in H^1(\mathbb{T}^d)$.

Now, given $u \in H^1(\mathbb{T}^d)$, we observe that for each $\phi \in \mathcal{D}(\Omega)$,

$$\begin{aligned}(A^{\mathbb{T}^d}u)_{|\Omega}(\phi) &= \sum_{0\leq|\sigma|,|\gamma|\leq 1} \int_{\mathbb{T}^d} \overline{a}_{\sigma\gamma}(x) D^\sigma u D^\gamma \phi^0 \\ &= \sum_{0\leq|\sigma|,|\gamma|\leq 1} \int_{\Omega} \tilde{a}_{\sigma\gamma}(x) D^\sigma u D^\gamma \phi \\ &= [A(u_{|\Omega})](\phi).\end{aligned}$$

Thus, if we define $A^{\mathbb{T}^d}$ as above, we can write

$$(A^{\mathbb{T}^d}u)_{|\Omega} = A(u_{|\Omega}). \tag{3.2}$$

Note carefully, however, that while the restriction appearing on the right has a pointwise interpretation, the restriction on the left is in the sense of distributions. (Note also that these interpretations would agree whenever $A^{\mathbb{T}^d}u \in L_2(\mathbb{T}^d)$, which means that $A^{\mathbb{T}^d}u$ can be expressed in terms of a function in $L_2(\mathbb{T}^d)$.) This is fortunate, as we can now be sure that whenever

$$A^{\mathbb{T}^d}u^+ = f^+$$

holds, it is also true that

$$A(u^+_{|\Omega}) = f.$$

For simplicity, we introduce another slight abuse of notation and write again A for the "extension" $A^{\mathbb{T}^d}$ of A on Ω to $\mathbb{T}^d$.

Having settled this, we further assume that the bilinear form

$$a(u,v) := \langle Au, v\rangle \qquad \forall\, u, v \in H^1(\mathbb{T}^d)$$

is symmetric, and that it is *coercive* on the kernel of the trace operator $B^{\mathcal{D}} : H^1(\mathbb{T}^d) \to H^{1/2}(\partial\Omega)$. That is, that there exists a constant $\alpha > 0$ such that

$$a(u,u) \geq \alpha \|u\|^2_{H^1(\mathbb{T}^d)} \qquad \forall\, u \in \mathcal{N}(B^{\mathcal{D}}).$$

To simplify the notation, we will write B for $B^{\mathcal{D}}$ throughout the rest of this chapter.

We turn our attention now to a different problem, formulated in terms of the new extended data. We seek for the minimizer in $H^1(\mathbb{T}^d)$ of the functional

$$F(v) := \frac{1}{2}a(v,v) - \langle f, v\rangle \qquad \forall\, v \in H^1(\mathbb{T}^d), \tag{3.3}$$

subject to the additional constraint $Bv = g$. We express this constraint in the equivalent form

$$b(v,q) = \langle g, q\rangle \qquad \forall\, q \in H^{-1/2}(\partial\Omega), \tag{3.4}$$

where we have defined $b(v,q) := \langle Bv, q\rangle_{H^{1/2}\times H^{-1/2}}$.

To solve this constrained minimization problem, we append (3.4) to (3.3) using a Lagrange multiplier. Our problem now reads: find $p \in H^{-1/2}(\partial\Omega)$, $u^+ \in H^1(\mathbb{T}^d)$, such that

$$F^*(u^+, p) := \sup_{q\in H^{-1/2}(\partial\Omega)} \inf_{v\in H^1(\mathbb{T}^d)} F^*(v,q), \tag{3.5}$$

where

$$F^*(v,q) = \frac{1}{2}a(v,v) - \langle f^+, v\rangle + b(v,q) - \langle g, q\rangle$$

and p is the Lagrange multiplier.

Using standard variational arguments one concludes that $(u^+, p) \in H^1(\mathbb{T}^d) \times H^{-1/2}(\partial\Omega)$ satisfies (3.5) if and only if

$$\begin{aligned} a(u^+, v) + b(v,p) &= \langle f^+, v\rangle \qquad \forall\, v \in H^1(\mathbb{T}^d), \\ b(u^+, q) &= \langle g, q\rangle \qquad \forall\, q \in H^{-1/2}(\partial\Omega). \end{aligned} \tag{3.6}$$

We often write (3.6) in operator form. Thus (u^+, p) satisfies (3.5) if and only if

$$\begin{pmatrix} A & B^* \\ B & 0 \end{pmatrix} \begin{pmatrix} u^+ \\ p \end{pmatrix} = \begin{pmatrix} f^+ \\ g \end{pmatrix}. \tag{3.7}$$

Our new problem reads, given $(f^+, g) \in H^{-1}(\mathbb{T}^d) \times H^{1/2}(\partial\Omega)$, find $(u^+, p) \in H^1(\mathbb{T}^d) \times H^{-1/2}(\partial\Omega)$ such that (3.6), (3.7) hold. Owing to its derivation from (3.5), we call this a *saddle point problem.*

One can check (see [22], [27]) that this problem is well posed; the operator $M : H^1(\mathbb{T}^d) \times H^{-1/2}(\partial\Omega) \to H^{-1}(\mathbb{T}^d) \times H^{1/2}(\partial\Omega)$ given by

$$M = \begin{pmatrix} A & B^* \\ B & 0 \end{pmatrix} \tag{3.8}$$

is an isomorphism. Furthermore, the restriction to Ω of u^+, $u = u^+_{|\Omega}$ is the unique solution of problem (3.1).

The discretization of problem (3.1) with respect to finite dimensional subspaces of $H^1(\mathbb{T}^d) \times H^{-1/2}(\partial\Omega)$ and $H^{-1}(\mathbb{T}^d) \times H^{1/2}(\partial\Omega)$ requires some care, since otherwise the resulting discrete problem becomes unstable. We omit this discussion here, since it plays no role in the rest of this chapter, and instead refer to [4], [21], [12].

3.2 Approximating u^+

Throughout the rest of this chapter, we will work under the following assumptions. First, that Ψ, $\tilde{\Psi}$ are a pair of biorthogonal B-spline wavelet bases for $H^1(\mathbb{T}^d)$, $H^{-1}(\mathbb{T}^d)$ respectively (which means that they are already properly scaled; see subsection 2.6.5), with corresponding multiresolution analysis $\{V_j\}_{j\in\mathbb{N}_0}$, $\{\tilde{V}_j\}_{j\in\mathbb{N}_0}$. To avoid technicalities, we also assume that the members of these bases are *smooth enough*. This means, in particular, that $\{V_j\}_{j\in\mathbb{N}_0}$, $\{\tilde{V}_j\}_{j\in\mathbb{N}_0}$ satisfy appropriate Jackson and Bernstein inequalities (2.5), so that we can always write

$$\mathcal{A}^s_q(L_2, \{V_j\}_{j\in\mathbb{N}_0}) = B^s_q(L_2).$$

We prefer the notation for approximation spaces because it is a bit more flexible and to the point.

For technical reasons that will become apparent later on, we also assume that the order of the primal basis is at least $m \geq 4$. Thus, If $\psi_\lambda \in \Psi$, then $\psi_\lambda \in C^{m-2}$, and ψ_λ is at least a piecewise cubic function.

3.2.1 Approximating u^+ with linear approximation schemes

The aim of this subsection is to illustrate the effect of the Lagrange multiplier on the Sobolev smoothness of the extended solution u^+. The result we derive here states that even though f^+ and g are such that the original problem would admit a smoother solution (which could be approximated more efficiently using linear approximation schemes), a non-zero Lagrange multiplier implies that $u^+ \in H^s(\mathbb{T}^d)$ is only possible for $s \leq 3/2$.

This critical index of $3/2$, and the aim of our study, leads us to base our results on the hypothesis that $(f^+, g) \in H^{\epsilon-1/2}(\mathbb{T}^d) \times H^{\epsilon+1}(\partial\Omega)$ for some $\epsilon > 0$. If $g \notin H^{1+\epsilon}(\partial\Omega)$ for any $\epsilon > 0$, then u^+ cannot belong in any Sobolev space with an index greater than $3/2$, regardless of the value of the Lagrange multiplier. On the other hand, if $f^+ \notin H^{\epsilon-1/2}$ for any $\epsilon > 0$, then the solution may or may not be smooth, depending on the particular case at hand (see remark 3.2.6).

We begin by showing that, under certain circumstances, the Lagrange multiplier is the jump in the conormal derivatives of u^+ at $\partial\Omega$. The conormal derivatives of $v \in H^s(\Omega)$ at $\partial\Omega$ are given by

$$\mathbf{n} \cdot \tilde{a} \nabla v,$$

where $\mathbf{n}$ is the outward normal at a point in $\partial\Omega$, and $\tilde{a}$ is the coefficient matrix of the operator A in divergence form (see (2.21)).

This result has been known for quite some time. It has its origin in [1], and can be found in a slightly less general form (only for $A = -\triangle$) in [21]. The present form essentially realizes a remark in [12].

Proposition 3.2.1. *If $f^+ \in L_2(\mathbb{T}^d)$, $g \in H^{1/2}(\partial\Omega)$, and $(u^+, p) \in H^1(\mathbb{T}^d) \times H^{-1/2}(\partial\Omega)$ is the solution with this data of system (3.7), then p is the value of the jump in the conormal derivatives at $\partial\Omega$.*

Proof. Write $\tilde{\Omega} = \mathbb{T} \setminus \overline{\Omega}$. On Ω, we have that $A r_\Omega u^+ = r_\Omega f^+$, and so for any $\varphi \in C^\infty(\overline{\Omega})$, we obtain

$$\int_\Omega \varphi A u^+ \, d\mu = \int_\Omega f^+ \varphi \, d\mu.$$

Using Green's formula, we also have that

$$\int_\Omega \varphi A u^+ \, d\mu = \int_{\partial\Omega} (\mathbf{n} \cdot \tilde{a} \nabla u^+) \varphi \, d\sigma + \int_\Omega \nabla\varphi \cdot \tilde{a} \nabla u^+ \, d\mu.$$

We repeat the same argument for $\tilde{\Omega}$, and then, by adding both results, obtain that for every $\upsilon \in C^\infty(\mathbb{T}^d)$

$$\begin{aligned} \int_{\mathbb{T}^d} f^+ \upsilon \, d\mu = &\int_{\mathbb{T}^d} \nabla\upsilon \cdot \tilde{a} \nabla u^+ \, d\mu \\ &+ \int_{\partial\Omega} \upsilon (\mathbf{n} \cdot \tilde{a} \nabla u^+) \, d\sigma + \int_{\partial\tilde{\Omega}} \upsilon (\mathbf{n} \cdot \tilde{a} \nabla u^+) \, d\sigma, \end{aligned}$$

and since the outward normal at $\partial\tilde{\Omega}$ is minus the outward normal at $\partial\Omega$, we obtain that

$$\int_{\mathbb{T}^d} f^+ \upsilon \, d\mu = \int_{\mathbb{T}^d} \nabla\upsilon \cdot \tilde{a} \nabla u^+ \, d\mu + \int_{\partial\Omega} \upsilon \left[\mathbf{n} \cdot \tilde{a} \nabla u^+\right]_{\partial\Omega} \, d\sigma,$$

where we have written $[\mathbf{n} \cdot \tilde{a} \nabla u^+]_{\partial\Omega}$ for the jump in the conormal derivatives at $\partial\Omega$.

But u, p, and f^+ also satisfy the first equation in (3.6), so we see that for every $\upsilon \in C^\infty(\mathbb{T}^d)$,

$$\begin{aligned} \langle f^+, \upsilon \rangle &= a(u^+, \upsilon) + b(\upsilon, p) \\ &= a(u^+, \upsilon) + b(\upsilon, \left[\mathbf{n} \cdot \tilde{a} \nabla u^+\right]_{\partial\Omega}). \end{aligned}$$

Thus, we conclude that $p = [\mathbf{n} \cdot \tilde{a} \nabla u^+]_{\partial\Omega}$, as we wanted to prove. □

Note that the hypothesis that f^+ is in $L_2(\mathbb{T}^d)$ was mainly used when writing

$$\int_\Omega \upsilon f \, d\mu + \int_{\tilde{\Omega}} \upsilon f \, d\mu = \int_{\mathbb{T}^d} \upsilon f \, d\mu.$$

So in fact what we have used is that since υf is measurable,

$$\int_{\mathbb{T}^d \setminus (\Omega \cup \tilde{\Omega})} \upsilon f \, d\mu = 0$$

because $\mu(\mathbb{T}^d \setminus (\Omega \cup \tilde{\Omega})) = 0$. So clearly, proposition 3.2.1 should still hold under more general hypothesis. It turns out to hold form the full range of parameters we are interested in.

Proposition 3.2.2. *Suppose that, for some $\epsilon > 0$, $(f, g) \in H^{\epsilon-1/2}(\mathbb{T}^d) \times H^{\epsilon+1}(\partial\Omega)$, and let $(u^+, p) \in H^1(\mathbb{T}^d) \times H^{-1/2}(\partial\Omega)$ be the solution of system (3.7) with this data. Then p is the jump in the conormal derivative at the boundary.*

Proof. We extend proposition 3.2.1 by continuity. To that end, let $\{f_n\}_{n\in\mathbb{N}}$, $f_n \in L_2(\mathbb{T}^d)$ be such that $f_n \to f$ in $H^{\epsilon-1/2}(\mathbb{T}^d)$. Let (u_n^+, p_n) be the solutions of the system (3.7) with (f_n, g) as data.

Given any domain $\omega \subset \mathbb{T}^d$, we denote by $S_\omega : H^{\epsilon+3/2}(\omega) \to H^\epsilon(\partial\omega)$ the conormal derivative operator, defined by $S_\omega v = \mathbf{n} \cdot \nabla v$, where $\mathbf{n}$ is the outward normal at a point in $\partial\omega$. As we have done before, we also denote by r_ω the restriction to ω.

Observe that if ω has a smooth boundary, and if $\epsilon > 0$, then the operator S_ω is continuous. To see this, note that the operator $B^{\mathcal{D}} \circ \frac{\partial}{\partial x_i} : H^{\epsilon+3/2}(\omega) \to H^\epsilon(\partial\omega)$ is bounded. Furthermore, recall that if $\varphi \in C^\infty(\overline{w})$, then $v \mapsto \varphi v$ is a bounded operator from any $H^t(w)$, $t > 0$, to itself. Thus, since the coefficients $\tilde{a}$ are in $C^\infty(\mathbb{T}^d)$, the operator $G : H^{\epsilon+3/2}(\omega) \to [H^\epsilon(\partial\omega)]^d$ (where we endow the latter space with the Euclidean tensor product norm), given by

$$Gu := \tilde{a}(x) \begin{pmatrix} \frac{\partial u}{\partial x_1} \\ \frac{\partial u}{\partial x_2} \\ \vdots \\ \frac{\partial u}{\partial x_d} \end{pmatrix}$$

is bounded. Thus, $S_\omega u = \mathbf{n} \cdot \tilde{a} \nabla u = \mathbf{n} \cdot Gu$ is a bounded operator from $H^{\epsilon+3/2}(\omega)$ to $H^\epsilon(\partial\omega)$.

We will also need to define the restriction to a domain ω of a functional g in $H^{\epsilon-1/2}(\mathbb{T}^d)$. As such, this makes no sense, since g is not defined on $\mathbb{T}^d$, as it is a functional on $H^{1/2-\epsilon}(\mathbb{T}^d)$. We assume (as we can do without loss of generality) that $\epsilon < 1/2$, and given $\varphi \in C_o^\infty(\omega)$, we define $(R_\Omega g)\varphi$ as the value of g on the extension by zero of φ to $\mathbb{T}^d$. This defines, by continuity, a bounded functional on $H_0^{1/2-\epsilon}(\omega) = H^{1/2-\epsilon}(\omega)$. The map R_Ω is clearly bounded; one can check also that if g is given by $g(v) = \int_{\mathbb{T}^d} \tilde{g} v \, d\mu$, with $\tilde{g} \in L_2(\omega)$, then $(R_\Omega g)(v) = \int_\omega r_\omega \tilde{g} v \, d\mu$. We will no longer make such a fine distinction between a functional and its representation, and write, in what constitutes an abuse in notation, $r_\omega g := R_\omega g$.

Since $f_n \to f$ in $H^{\epsilon-1/2}(\mathbb{T}^d)$, we have that $r_\Omega f_n \to r_\Omega f$, and since $A u^+_{n|\Omega} = f_{n|\Omega}$, $B u^+_{n|\Omega} = g$, we also have that the sequence $\{r_\Omega u_n^+\}$ converges in $H^{\epsilon+3/2}(\Omega)$, and that it converges to $r_\Omega u^+$. An identical argument shows that $\{r_{\tilde{\Omega}} u_n^+\}$ converges to $r_{\tilde{\Omega}} u^+$.

Now, by proposition 3.2.1, $p_n = S_\Omega r_\Omega u_n^+ + S_{\tilde{\Omega}} r_{\tilde{\Omega}} u_n^+$, and so, by continuity of M^{-1} (with M defined in (3.8)), S_Ω, $S_{\tilde{\Omega}}$, and r_Ω, we obtain that $p = S_\Omega r_\Omega u^+ + S_{\tilde{\Omega}} r_{\tilde{\Omega}} u^+$, and so p is exactly the jump in the conormal derivatives of u^+ at $\partial\Omega$. □

The next question is, what does a jump in the conormal derivatives imply for the smoothness of u^+? The following lemma clears us from (almost) all doubts.

Lemma 3.2.3. *If $v \in H^{\epsilon+3/2}$ for some $\epsilon > 0$, then the jump in the conormal derivatives of v at $\partial\Omega$ vanishes.*

Proof. Let $\{\varphi_n\} \subset C^\infty(\mathbb{T}^d)$ be such that $\varphi_n \to v$ in $H^{\epsilon+3/2}(\mathbb{T}^d)$ when $n \to +\infty$. Using the same notation as in the proof of proposition 3.2.2, we have that

$$S_\Omega r_\Omega \varphi_n + S_{\tilde{\Omega}} r_{\tilde{\Omega}} \varphi_n = 0,$$

and so by continuity, we obtain that the jump in the conormal derivatives of v at $\partial\Omega$, $S_\Omega r_\Omega v + S_{\tilde{\Omega}} r_{\tilde{\Omega}} v$, must also vanish. □

We can now summarize the above results into the following.

Theorem 3.2.4. *If $(f^+, g) \in H^{\epsilon-1/2}(\mathbb{T}^d) \times H^{\epsilon+1}(\partial\Omega)$ for some $\epsilon > 0$, and the Lagrange multiplier obtained when solving (3.7) is nonzero, then $u^+ \in H^s(\mathbb{T}^d)$ implies $s \leq 3/2$.*

As a consequence, we can finally estimate the rate of approximation of u^+ by $\{V_j\}_{j\in\mathbb{N}_0}$.

Corollary 3.2.5. *If $(f^+, g) \in H^{\epsilon-1/2}(\mathbb{T}^d) \times H^{\epsilon+1}(\partial\Omega)$ for some $\epsilon > 0$, and $p \neq 0$, then*

$$u^+ \in \mathcal{A}_2^s(L_2, \{V_j\}_{j\in\mathbb{N}_0}) \qquad \text{implies } s \leq 3/2, \tag{3.9}$$

and

$$u^+ \in \mathcal{A}_2^s(H^1, \{V_j\}_{j\in\mathbb{N}_0}) \qquad \text{implies } s \leq 1/2. \tag{3.10}$$

Proof. Apply (2.15) to theorem 3.2.4, and observe that $B_2^s(L_2) = \mathcal{A}_2^s(L_2, \{V_j\}_{j\in\mathbb{N}_0})$ for the corresponding range of s. This settles (3.9). To prove (3.10), apply theorem 2.2.4. □

Remark 3.2.6. *If $f^+ \notin H^s(\mathbb{T}^d)$ for any $s > -1/2$, then it is possible that the solution of (3.7) is smooth (i.e, belongs to some Sobolev space H^t for some large t), even when the Lagrange multiplier is not zero.*

To see this, choose an arbitrary $t > 3/2$, and let $v \in H^t(\mathbb{T}^d)$. Then choose $q \in H^{-1/2}(\partial\Omega)$, $q \neq 0$, and set $f^+ := Av + B^*q$, $g := Bv$. If we solve the system (3.7) with these data, we obtain a pair (u^+, p) with $u^+ = v$, and $p = q \neq 0$. By lemma 3.2.3, it would be a contradiction if p was the jump in the conormal derivatives. But that would contradict theorem 3.2.2, unless $f^+ \notin H^s(\mathbb{T}^d)$ for any $s > -1/2$.

Under some circumstances, it is possible to rule out the case $s = 3/2$ in 3.2.4.

Theorem 3.2.7. *Suppose that $f^+ \in H^{-1/2+\epsilon}(\mathbb{T}^d)$ for some $\epsilon > 0$, and let (u^+, p) be the solution of (3.7). If there exists an open set $\mathcal{U}$, and a constant $c > 0$ such that $p(x) \geq c > 0$ almost everywhere on $\mathcal{U} \cap \partial\Omega$, or alternatively, if $p(x) \leq c < 0$ almost everywhere on $\mathcal{U} \cap \partial\Omega$ (this assumes also that p can be identified with a measurable function on that set) then $u^+ \in H^s(\mathbb{T}^d)$ implies $s < 3/2$.*

This result is based on the following lemma.

Lemma 3.2.8. *Under the hypothesis on p of theorem 3.2.7, there exists $j_0 \in \mathbb{N}$ such that for each $j \geq j_0$ we can find $G_j \subset \nabla_j^0 := \{\lambda \in \nabla : |\lambda| = j\}$ with the following properties.*

i. $\#G_j \gtrsim 2^{j(d-1)}$

ii. $\lambda \in G_j$ *implies that* $|\langle \psi_\lambda, B^* p \rangle| \gtrsim 2^{-jd/2}$.

The proof of this lemma is fairly technical, and thus we defer it for the moment.

Proof of theorem 3.2.7. We begin by directing our attention to the first equation in (3.7), and rewrite it to read

$$Au = f^+ - B^* p \tag{3.11}$$

Now whenever $u^+ \in H^s(\mathbb{T}^d)$, then $Au^+ \in H^{s-2}(\mathbb{T}^d)$, and thus by (3.11) it will be enough to show that if $f^+ - B^* p \in H^{s-2}$, then $s - 2 < -1/2$. But this reduces again to prove that if $B^* p \in H^r(\mathbb{T}^d)$, then $r < -1/2$.

Since the bases Ψ, $\tilde{\Psi}$ (chosen at the beginning of 3.2) are a pair of biorthogonal B-spline wavelet bases for $H^1(\mathbb{T}^d)$, $H^{-1}(\mathbb{T}^d)$ respectively, and thus they are Riesz bases, we can write

$$\|B^* p\|^2_{H^{-1}(\mathbb{T}^d)} \sim \sum_{\lambda \in \nabla} |\langle B^* p, \psi_\lambda \rangle|^2. \tag{3.12}$$

Given $t \geq 0$ we can compute the norm of $B^* p$ in $H^{t-1}(\mathbb{T}^d)$ by introducing an additional scaling factor in (3.12). We have that

$$\begin{aligned} \|B^* p\|^2_{H^r(\mathbb{T}^d)} &\sim \sum_{\lambda \in \nabla} 2^{t|\lambda|} |\langle B^* p, \psi_\lambda \rangle|^2 \\ &= \sum_{j \in \mathbb{N}_0} 2^{jt} \sum_{\nabla_j^0} |\langle B^* p, \psi_\lambda \rangle|^2 \end{aligned} \tag{3.13}$$

We invoke lemma 3.2.8 and see that if $j \geq j_0$,

$$\begin{aligned} \sum_{\nabla_j^0} |\langle B^* p, \psi_\lambda \rangle|^2 &\geq \sum_{\lambda \in G_j} |\langle B^* p, \psi_\lambda \rangle|^2 \\ &\gtrsim 2^{j(d-1)} \cdot 2^{-jd} = 2^{-j}, \end{aligned}$$

and thus we have that (3.13) diverges whenever $t \geq 1/2$, and thus $B^* p \in H^r$ implies that $r = t - 1 < -1/2$. □

3.2.2 Approximating u^+ with nonlinear approximation schemes based on B-spline wavelets

The only result in this subsection states (roughly speaking) that,

- if the bases we have chosen are sufficiently smooth and have enough vanishing moments,
- if the rate of convergence of the best $N-$term approximations to f^+ is higher than a certain threshold,
- and if the Lagrange multiplier obtained when solving (3.7) satisfies the hypothesis of lemma 3.2.8,

then the rate of convergence of the best $N-$ term approximations to u^+ is bounded from below. Let us be more precise.

Theorem 3.2.9. *Let $\overline{f} = \{f_\lambda\}_{\lambda\in\nabla} \in \ell_2$, $f_\lambda := \langle f^+, \psi_\lambda\rangle$, be the sequence of coefficients of f^+ with respect to the basis $\tilde{\Psi}$, and suppose that $\overline{f} \in \ell^w_\sigma$ for some $\sigma < \frac{2(d-1)}{d}$ (this is equivalent to the assumption that*

$$f^+ \in \mathcal{A}^r_\infty(H^{-1}, \Sigma_n(\tilde{\Psi}))$$

for $r = \frac{1}{\sigma} - \frac{1}{2} > \frac{1}{2(d-1)}$). If p satisfies the hypothesis of theorem 3.2.7, and if Ψ, $\tilde{\Psi}$ are sufficiently smooth and have enough vansihing moments, then the sequence $\overline{u} = \{u_\lambda\}_{\lambda\in\nabla} \in \ell_2$ of coefficients of u^+, $u_\lambda := \langle u^+, \tilde{\psi}_\lambda\rangle$, satisfies that if $\overline{u} \in \ell^w_\tau$, then $\tau \geq \frac{2(d-1)}{d}$. In other words,

$$u^+ \in \mathcal{A}^t_\infty(H^1(\mathbb{T}^d), \Sigma_n(\Psi))$$

implies that $t \leq \frac{1}{2(d-1)}$.

Proof. When we assume that Ψ, $\tilde{\Psi}$ are smooth enough and have enough vanishing moments, we mean that they where chosen such that $\overline{A}$, the matrix of A with respect to the basis Ψ, $\tilde{\Psi}$, satisfies $\overline{A} \in \mathcal{B}_s$ for some $s > \frac{1}{2(d-1)}$; see subsection 2.7.2.

Let $\overline{d} = \{d_\lambda\}_{\lambda\in\nabla} \in \ell_2$ be the coefficients of B^*p, $d_\lambda := \langle B^*p, \psi_\lambda\rangle$. From lemma 3.2.8 we obtain that if $j > j_o$,

$$G_j \subset \{\lambda \in \nabla : |d_\lambda| > C2^{-jd/2}\}.$$

From this, and again from lemma 3.2.8 we obtain that

$$\#\{\lambda \in \nabla : |d_\lambda| > 2^{-jd/2}\} \gtrsim 2^{j(d-1)},$$

which, writing $a = 2^{d/2}$, yields

$$\#\{\lambda \in \nabla : |d_\lambda| > a^{-j}\} \gtrsim a^{j\frac{2(d-1)}{d}}.$$

Using proposition 2.7.2, we have that

$$\overline{d} \in \ell^w_\tau \tag{3.14}$$

only if $\tau \geq \frac{2(d-1)}{d}$.

If $\overline{u} \in \ell^w_\tau$, and $\tau < \frac{2(d-1)}{d}$, then since $\overline{A} \in \mathcal{B}_s$ for some $s > \frac{1}{2(d-1)}$, we have that $\overline{A}\overline{u} \in \ell^w_\rho$ for some $\sigma \leq \rho < \frac{2(d-1)}{d}$. But this implies that $\overline{A}\overline{u} = \overline{f} + \overline{d} \in \ell^w_\rho$, and thus by linearity, $\overline{f} + \overline{d} - \overline{f} = \overline{d} \in \ell^w_\rho$. The theorem now follows from this contradiction. □

Thus, we have that under the hypothesis of theorem 3.2.9, the best N-term approximations of u^+ converge at best as $\mathcal{O}(N^{-\frac{1}{2(d-1)}})$. As a consequence, no adaptive method comparable with those discussed in [8], (see subsection 2.7.2), can achieve an accuracy of ϵ without spending at least $\mathcal{O}(\epsilon^{-2(d-1)})$ operations.

Note again that theorems 3.2.7 and 3.2.9 hold whenever the basis functions are *smooth enough*. Resorting to higher order B-spline wavelets is of no help.

Finally, we remark that from (3.14) it also follows that B^* is not very compressible (see proposition 2.7.3).

3.2.3 Obtaining better convergence rates

In theory, it is easy to obtain better convergence rates. This is illustrated by the following two results.

Proposition 3.2.10. *Let $V_j^\Omega = r_\Omega V_j$, and suppose that the solution u of problem (3.1) is in $\mathcal{A}_q^s(L_2(\Omega), \{V_j^\Omega\}_{j\in\mathbb{N}_0})$ for some $s \geq 1$, $0 < q \leq \infty$. Then there exists an extension f^+ of f such that the extended solution u^+ of (3.7) satisfies $u^+ \in \mathcal{A}_q^s(L_2(\mathbb{T}^d), \{V_j^\Omega\}_{j\in\mathbb{N}_0})$ and $u^+ \in \mathcal{A}_q^{s-1}(H^1(\mathbb{T}^d), \{V_j^\Omega\}_{j\in\mathbb{N}_0})$.*

Proof. Just find an appropriate extension u^* of u to $\mathbb{T}^d$ using the results of section 2.4, and take $f^+ = Au^*$. When we solve (3.7) with this right-hand side (and with g as before), we obtain that $(u^*, 0)$ is the (unique) solution, and thus $u^+ = u^* \in \mathcal{A}_q^s(L_2(\mathbb{T}^d), \{V_j^\Omega\}_{j\in\mathbb{N}_0})$. Using lemma 2.2.4, we also obtain $u^+ \in \mathcal{A}_q^{s-1}(H^1(\mathbb{T}^d), \{V_j^\Omega\}_{j\in\mathbb{N}_0})$. □

Proposition 3.2.11. *Suppose that the solution u of problem (3.1) satisfies $u \in B_\tau^{sd+1}(L_\tau(\Omega))$ for some $\tau < \frac{2(d-1)}{d}$, and where $s = \frac{1}{\tau} - \frac{1}{2}$. Then there exists an extension of f^+ of f such that the solution u^+ of 3.7 satisfies $u^+ \in B_\tau^{sd+1}(L_\tau(\mathbb{T}^d))$. That is, $u^+ \in \mathcal{A}_\infty^{s-\epsilon}(H^1, \Sigma_n(\Psi))$ for all $0 < \epsilon < s$.*

Proof. Using 2.16, we see that there exists an extension $u^* \in B_\tau^{sd+1}(L_\tau(\mathbb{T}^d))$ of u. We obtain f^+ now simply by setting $f^+ = Au^*$. □

We conclude that in order to obtain better convergence rates, we must find an adequate extension of f. Note that it is not enough to choose a smooth extension of the right hand side. It must be smooth *and* produce a smooth solution.

The naive approach to the construction of a fictitious domain method for solving problem (3.1) without these problems might follow the route proposed by propositions 3.2.11 and 3.2.10. That is, to extend the solution and then apply the differential operator. This has a major drawback from the point of view of a numerical method: it must start with a fairly accurate solution of problem (3.1), and thus renders the method pointless.

In the next chapter we will construct a method which produces smooth solutions by finding smooth extensions of u and f *simultaneously*, and without compromising accuracy. In what remains of this chapter we are going to prove lemma 3.2.8 and the auxiliary results needed.

3.3 Proof of lemma 3.2.8

To simplify a bit, we begin by assuming that j is always large enough, so that we can neglect effects caused by periodization. Specifically, we assume that there exists an $\epsilon > 0$ such that for all j considered, if $\operatorname{supp} \psi^e_{jk} \cap \Omega \neq \emptyset$, then $\operatorname{supp} \psi^e_{jk} \subset (0+\epsilon, 1-\epsilon)^d$. We will also restrict ourselves to the case $p(x) > c > 0$, since the case $p(x) < -c < 0$ is completely analogous.

Let $x_0 \in \mathcal{U} \cap \partial\Omega$, and let $\epsilon_0 > 0$, $\phi \in C^1$, and $Q : \mathbb{R}^d \to \mathbb{R}^d$ be an orthogonal transformation as in definition 2.3.2. This means that

$$Q^{-1}(B(x_0, \epsilon_0) \cap \Omega) = \{x \in Q^{-1}(B(x_0, \epsilon_0)) : x_d < \phi(x_1, x_2, \ldots, x_{d-1})\}.$$

Assume further that $B(x_0, \epsilon_0) \subset \mathcal{U}$, and let $Y \subset \mathbb{R}^{d-1}$ gather all points $y \in \mathbb{R}^d$ such that

$$\theta(y) := Q(y, \phi(y))^T \in \partial\Omega \cap B(x_0, \epsilon_0).$$

Note that θ satisfies $\|\theta(x) - \theta(y)\| > \|x - y\|$ for all $x, y \in Y$, and that Y is an open set.

Given a function $f \in H^1(\mathbb{T}^d)$ with $\operatorname{supp} f \subset B(x_0, \epsilon_0)$, we compute

$$\begin{aligned} \langle B^* p, f \rangle &= \langle p, Bf \rangle \\ &= \int_{B(x_0,\epsilon_0)\cap\partial\Omega} p(y) f(y) dS \\ &= \int_Y p(\theta(z)) f(\phi(z)) \mathcal{J}\theta(z) dz. \end{aligned} \tag{3.15}$$

Here $\mathcal{J}\theta(z)$ is the $(d-1)$-dimensional volume of the parallelogram spanned by the vectors $\{D\theta(z)e_1, \ldots, D\theta(z)e_{d-1}\}$, see for instance [19],chapter 7.

3.3.1 Index sets and banded matrices

To find the sets predicted by lemma 3.2.8, we will not use 3.15 directly on the wavelets, but will instead transform the claim of the lemma to an analogous claim on scaling functions. Before doing this, we will shed some light on the structural relationship between sets of scaling function and wavelet coefficients.

Given a level j, we can (obviously) consider the wavelet or scaling function coefficients of a function f as belonging to a vector space indexed by $\mathcal{Z}^d_j = \mathbb{Z}^d / 2^j \mathbb{Z}^d$. For instance, the scaling function representation of $B^* p$ on level j can be interpreted as $c_j \in \ell_2(\mathcal{Z}^d_j)$, with entries

$$c_{jk} = \langle B^* p, \psi^0_{jk} \rangle = \langle p, B\psi^0_{jk} \rangle, \qquad \forall k \in \mathcal{Z}^d_j. \tag{3.16}$$

This point of view is useful because it allows us to use information on the location of a basis member on $\mathbb{T}^d$. To this end we define a metric on $\mathcal{Z}^d_j$ by

$$d(k, k') = \min_{z \in Z^d} \|k + 2^j z - k'\|_\infty.$$

In this spirit, let X be some finite set, and let $W = \ell_2(X)$, $V = \ell_2(\mathcal{Z}^d_j)$. We will say that a linear map $M : W \to V$ is *banded* of *width* $d_M \in \mathbb{N}$ if for any $k \in X$ one has that if $k', k'' \in \operatorname{supp} Me_k$, then $d(k', k'') < d_M$. Here we have written e_k for the member in the canonical basis corresponding to k. That is, $(e_k)_l = \delta_{kl}$, where δ_{kl} is the Kronecker delta.

Proposition 3.3.1. *Let $A \subset \mathcal{Z}_j^d$ be such that for $a_1, a_2 \in A$, $a_1 \neq a_2$ implies $d(a_1, a_2) \geq d_M$, and suppose $v \in V$, $w \in W$ are related by $v = Mw$. If, for some $C_2 > 0$ one has $|v_a| \geq C_2$ for each $a \in A$, then there exists $C_3 > 0$, and $B \subset X$ such that $|w_b| \geq C_3$, and $\#B = \#A$.*

Proof. For each $a \in A$, write $D_a = \{k \in X : a \in Me_k\}$, and let $N = \sup_{a \in A} \#D_a$. Now, if $|w_c| < \frac{C_2}{N\|M\|_\infty}$ for all $c \in D_a$, we reach a contradiction with the hypothesis that $|v_a| \geq C_2$, since then

$$|v_a| = \left| \sum_{c \in D_a} (w_c M e_c)_a \right| < C_2.$$

Thus, we take $C_3 = \frac{C_2}{N\|M\|_\infty}$, and chose for each $a \in A$ a single $b_a \in D_a$ such that $|v_{b_q}| \geq C_3$, and collect all those b_a in the set B.

It only remains to prove that if $b_a = b_{a'}$, then $a = a'$. Indeed, if $a \neq a'$, then $a, a' \in \operatorname{supp} Me_{b_a}$, and thus $d(a, a') < d_M$, contradicting the hypothesis. □

Let $\overline{d}$ be the coefficients of $B^* p$ as above, and let us write $d_j \in \ell_2(\nabla_j^0)$ for the sequence of coefficients on level j only.

Let $\{\eta_1, \eta_2, \ldots, \eta_{2^d}\}$ be an enumeration of the set $E \setminus \{0\}$ (see 2.6.4). Then we can write $\nabla_j^0 = \prod_{i=1}^{2^d} \ell_2(\mathcal{Z}_j^d)$, and assign to each η_i a copy of $\ell_2(\mathcal{Z}_j^d)$. Then the map $M_j^1 : \nabla_j^0 \to \ell_2(\mathcal{Z}_{j+1}^d)$ given by the matrix

$$M_j^1 = \begin{pmatrix} M_j^{\eta_1} & M_j^{\eta_2} & \ldots & M_j^{\eta_{2^d}} \end{pmatrix}$$

maps the wavelet coefficients on a level j to the corresponding scaling function coefficients on level $j+1$. This map is banded in the above sense, and the bandwidth $d_{M_j^1}$ is independent of j if j is large enough. Moreover, the number $N_j = \max_{k \in \mathcal{Z}_j^d} \#\{\lambda \in \nabla_j^0 : k \in M_j^1 e_\lambda\}$ is also constant if j is large enough. A similar observation holds for $\|M\|_\infty$. We are in the position of proving lemma 3.2.8 using the following lemma.

Lemma 3.3.2. *Under the hypothesis of lemma 3.2.8, one can find $j_0 \in \mathbb{N}$ such that for each $j \geq j_0$ there exists a set $F_j \in \mathcal{Z}_j^d$ with the following properties*

$$\begin{aligned} &i. && k \in F_j \text{ implies } |(M_{j-1}^0 c_{j-1})_k| \gtrsim 2^{-\frac{jd}{2}} \text{ (see (3.16))} \\ &ii. && \#F_j \gtrsim 2^{j(d-1)} \\ &iii. && k_1, k_2 \in F_j \text{with } k_1 \neq k_2 \text{ implies } d(k_1, k_2) > d_{M_j^1} \\ &iv. && F_j \cap \operatorname{supp} c_j = \emptyset. \end{aligned} \tag{3.17}$$

Proof of lemma 3.2.8. We can write

$$c_{j+1} = M_j^1 d_j + M_j^0 c_j,$$

and thus

$$M_j^1 d_j = c_{j+1} - M_j^0 c_j.$$

If we write $v = c_{j+1} - M_j^0 c_j$, then we have that the sets F_j in (3.17), and the matrix M_j^1, both satisfy the hypothesis of proposition 3.3.1. From this, and from the observation that the constant C_3 in lemma 3.3.1 can be chosen independently of j, we infer the existence of the sets G_j for $j > j_0$, with j_0 as in lemma 3.3.2. □

3.4 Proof of lemma 3.3.2

3.4.1 Lower bounds for single integrals

We begin by introducing the notation $\square_{jk} := 2^{-j}\left([0,1]^d + k\right)$, and then associating to any set $A \subset \mathbb{T}^d$ an index set in $\mathcal{Z}_j^d$ according to the following notation.

$$\Lambda_j^0(A) := \{k \in \mathcal{Z}_j^d : \square_{jk} \cap A \neq \emptyset\},$$
$$\Lambda_j^n(A) := \{k \in \mathcal{Z}_j^d : \exists k' \in \Lambda_j^0(A) \text{ with } d(k,k') \leq n\}.$$

Let $x_0, \epsilon_0, \phi, Q, \theta$ be as chosen just before (3.15). Let $G := \partial\Omega \cap B(x_0, \epsilon/2)$, and let $Y_G := \{x \in Y : \theta(x) \in G\}$.

Proposition 3.4.1. *There exists $j_0 \in \mathbb{N}$ such that $j \geq j_0$, $k \in \Lambda_j^0(G)$ imply*

$$|c_{jk}| \gtrsim 2^{-\frac{jd}{2}}.$$

Proof. We begin by realizing that, since the primal scaling functions are B-splines of order at least 4, one has that $[0,1]^d \subset (\operatorname{supp}\psi^0)^\circ$, where A° denotes the *interior* of the set A. Thus we can find a constant $\tilde{c}$, and a $\tau > 0$ such that if $x \in B([0,1]^d, \tau)$, then $\psi^0(x) \geq \tilde{c}$.

Since θ is $C^1(\mathbb{R}^{d-1})$, we can show that θ is Lipschitz on Y. So let L be such that

$$\|\theta(x) - \theta(y)\|_2 \leq L\|x - y\|_2, \qquad \forall\, x, y \in Y, \tag{3.18}$$

write $\tau_j = \frac{1}{L}2^{-j}\tau$, and chose $j_1 \in \mathbb{N}$ such that if $j \geq j_1$, then $B(Y_G, \tau_j) \subset Y$. This is possible, by (3.18), when $\tau_j < \frac{\epsilon}{2L}$, for instance.

Let $j_0 \geq j_1$ be such that $j \geq j_0$, $k_0 \in \Lambda_j^0(G)$ implies $\operatorname{supp}\psi_{jk_0}^0 \subset B(x_0, \epsilon_0)$. Given such j, k_0, let $z \in Y_G$ be such that $\theta(z) \in \square_{jk_0} \subset \operatorname{supp}\psi_{jk_0}^0$. But then $B(z, \tau_j) \subset Y$, and also

$$\psi_{jk_0}^0(x) \geq 2^{-j}2^{\frac{jd}{2}}\tilde{c} \qquad \forall x \in B(z, \tau_j)$$

because $\theta(B(z, \tau_j)) \subset B(\square_{jk_0}, 2^{-j\tau})$, and where the powers of two come from the H^1 and L_2 normalization respectively.

Recall that $p(\theta(x)) > c$ almost everywhere on Y, and observe also that since $\|\theta(x) - \theta(y)\| \geq \|x - y\|$ for all $x, y \in Y$, we have $\mathcal{J}\theta(x) > C_4$ for some $C_4 > 0$.

From (3.15) we get

$$c_{jk_0} \gtrsim 2^{-j} \cdot 2^{\frac{jd}{2}} \int_{B(z,\tau_j)} dx \gtrsim 2^{-\frac{jd}{2}},$$

since the volume of $B(z, \tau_j)$ is larger than a constant times $2^{j(d-1)}$. □

3.4.2 Index sets and masks

To be able to satisfy requirement (iv) of (3.17) we need to obtain a better understanding of the action of the linear map M_j^0. We bring to our attention that if ψ^0 is a B-spline of order at least 4, then we have for its mask that

$$\operatorname{supp}\{a_k^0\} = \{\alpha, \alpha+1, \ldots, \beta\}^d \subset \mathbb{Z}^d \tag{3.19}$$

with

$$\alpha \leq -2, \qquad 2 \leq \beta. \tag{3.20}$$

Next we observe that if j is large enough to avoid periodization effects, we can find a constant $C_5 > 0$, independent of j, such that all nonzero entries in M_j^0 are larger that C_5. This follows from the definition of M_j^0, and from the fact that all entries in the mask of a B-spline generator are non-negative.

Let us take a look at indices $\tilde{k} \in \mathcal{Z}_j^d$ such that $\operatorname{supp} M_j^0 e_{\tilde{k}} \cap \Lambda_{j+1}^0(G) \neq \emptyset$.

Proposition 3.4.2. *For these $\tilde{k}$ it holds*

$$c_{j\tilde{k}} \gtrsim 2^{-\frac{jd}{2}}.$$

Proof. Given such an entry, we use the refinement relation to write

$$c_{j\tilde{k}} = \sum_{z \in \mathbb{Z}^d} \frac{\xi_{j+1}}{\xi_j} c_{j+1,2\tilde{k}+z} a_z. \tag{3.21}$$

If $a_z \neq 0$, then $\|z\|_\infty \leq \beta$. And if this is so, then $\operatorname{supp} \psi^0_{j+1,2\tilde{k}+z} \subset B(x_0, \epsilon_0)$, and thus from the definition of c_j, and by the hypothesis on p, we have that $c_{j+1,2\tilde{k}+z} \geq 0$, since for this index the integrand in (3.15) is non-negative. On the other hand, since $\operatorname{supp} M_j^0 e_{\tilde{k}} \cap \Lambda_{j+1}^0(G) \neq \emptyset$, there exist at least one z' such that $2\tilde{k} + z' \in \Lambda_{j+1}^0(G)$ while also $a_{z'} \neq 0$. By proposition 3.4.1, and since also the number of z such that $a_z \neq 0$ is finite (and thus there is a smallest such a_z), we have that

$$a_{z'} c_{j+1,2\tilde{k}+z'} \gtrsim 2^{-\frac{(j+1)d}{2}} = 2^{-\frac{d}{2}} 2^{-\frac{jd}{2}}.$$

Using this knowledge together with (3.21), we obtain the result. □

The elements in F_{j+1} will be chosen among those $l \in \operatorname{supp} M_j^0 e_{\tilde{k}}$ which also satisfy that $\mu_{d-1}(\operatorname{supp} \psi^0_{j+1,l} \cap \partial\Omega) = 0$, where μ_{d-1} is the Lebesgue measure on $\partial\Omega$. Those l satisfy indeed that $c_{j+1,l} = 0$ (since then the integral is defined on a set of measure zero), while also (by proposition 3.4.2) we have that $(M_j^0 c_j)_l \gtrsim 2^{-\frac{jd}{2}}$. The following lemma gives us a hint as to where to find this type of l.

Lemma 3.4.3. *Let $\{a_k^0\}_{k \in \mathbb{Z}^d}$ be the mask of ψ^0, let $\alpha, \beta \in \mathbb{Z}$ be as in (3.19), (3.20), and let $l \in \Lambda_{j+1}^{\beta+1}(\partial\Omega) \setminus \Lambda_{j+1}^{\beta}(\partial\Omega)$. Then it holds that*

i. $\mu(\operatorname{supp} \psi^0_{j+1,l} \cap \partial\Omega) = 0$

ii. There exists $k_* \in \Lambda_{j+1}^0(\partial\Omega)$ such that $d(l, k_*) = \beta + 1$.

iii. There exists $\tilde{k} \in \mathcal{Z}_j^d$ such that $l, k_* \in \operatorname{supp}\{a^0_{z-2\tilde{k}}\}_{z \in \mathbb{Z}^d}$, and thus $l, k_* \in M_j^0 e_{\tilde{k}}$.

Proof. The first two claims follow immediately from the definition of $\Lambda_{j+1}^0(\partial\Omega)$.

To prove the last one, we will show in a componentwise fashion that such a $\tilde{k}$ exists. To this end, let us write $l = (l^1, l^2, \ldots, l^d)$, $k_* = (k_*^1, k_*^2, \ldots, k_*^d)$, and $\tilde{k} = (\tilde{k}^1, \tilde{k}^2, \ldots, \tilde{k}^d)$. We

have (neglecting, as we can, effects of periodization) that the $\tilde{k}$ we are looking for satisfies that

$$l^i, k_*^i \in \{\alpha + 2\tilde{k}^i, \alpha + 2\tilde{k}^i + 1, \ldots, \beta + 2\tilde{k}^i\},$$

or in terms of inequalities, that

$$\alpha + 2\tilde{k}^i \leq l^i \leq \beta + 2\tilde{k}^i \qquad\qquad \alpha + 2\tilde{k}^i \leq k_*^i \leq \beta + 2\tilde{k}^i.$$

But this is equivalent to

$$\max\{l^i - \beta, k_*^i - \beta\} \leq 2\tilde{k}^i \leq \min\{l^i - \alpha, k_*^i - \alpha\}. \tag{3.22}$$

Such a $\tilde{k}^i$ exists, trivially, whenever $\beta - \alpha \geq 1$ (as is being assumed) and $l^i = k_*^i$.

If $l^i > k_*^i$, then (3.22) reduces to $l^i - \beta \leq 2\tilde{k}^i \leq k_*^i - \alpha$, which is equivalent to

$$l^i - k_*^i \leq 2\tilde{k}^i - k_*^i + \beta \leq \beta - \alpha. \tag{3.23}$$

Since $d(l, k_*) = \beta + 1$, we have that (3.23) can be satisfied by $\tilde{k}^i$ whenever

$$\beta + 1 \leq 2\tilde{k}^i - k_*^i + \beta \leq \beta - \alpha,$$

or simply when $1 \leq 2\tilde{k}^i - k_*^i \leq -\alpha$. We can always choose such a $\tilde{k}^i$ if, as is being assumed, $\alpha \leq -2$.

The case $k_*^i > l^i$ follows analogously. □

Having established the existence of the indices we are looking for, it only remains to show that there are enough of them.

3.4.3 Construction of the sets F_j

Let $P_m : \mathbb{R}^d \to \mathbb{R}^{d-1}$ be given by

$$P_m(x_1, \ldots, x_d) = (x_1, \ldots, x_{m-1}, x_{m+1}, \ldots, x_d).$$

We will assume for now (and prove this in the next section) that we can arrange matters to be as follows. Suppose we have found a $z_0 \in B(x_0, \epsilon_0/4)$, a $\sigma > 0$, and $m \in \{1, \ldots, d\}$ such that

$$\begin{aligned}
&i. && B(z_0, \sigma) \subset B(x_0, \epsilon_0/4) \\
&ii. && B(z_0, \sigma) \cap G = \emptyset \\
&iii. && P_m B(z_0, \sigma) \subset (P_m[G \cap B(x_0, e_0/4)])^\circ \\
&iv. && z_0 = 2^{-j^*} z_* \text{ for some } z_* \in \mathbb{Z}^d,\ j^* \in \mathbb{N}.
\end{aligned} \tag{3.24}$$

For $j > j^*$ (where we assume that j^* is larger than all previous lower bounds for j) we define the set

$$A_j := \{z \in \mathcal{Z}_j^d : (z - z^*)_m = 0 \text{ and } 2^{-j} z \in B(z_0, \sigma)\}.$$

Proposition 3.4.4. *There exists $j^{**} \geq j^*$ such that if $j \geq j^{**}$, then for each $a \in A_j$ there exists a number $r_a \in \mathbb{Z}$ such that $l_a := a + r_a e_m \in \Lambda_j^{\beta+1}(\partial\Omega) \setminus \Lambda_j^{\beta}(\partial\Omega)$, and such that $2^{-j} l_a \in B(x_0, \epsilon/4)$.*

Proof. It will be enough to choose j^{**} such that if $j > j^{**}$, then $a \in A_j$ does not belong to $\Lambda_j^{\beta+1}(\partial\Omega)$.

Since $P_m B(z_0, \sigma) \subset (P_m[G \cap B(x_0, e_0/4)])^\circ$, we have that there exists $\tilde{r}_a$ such that $a + \tilde{r}_a e_m \in \Lambda_j^0(\partial\Omega)$. We can assume, without loss of generality, that $\tilde{r}_a > 0$.

For $0 \leq i \leq \tilde{r}_a$ write $k_i := a + ie_m$, and observe that, by the convexity of the ball $B(x_0, e_0/4)$, the integer $\tilde{r}_a$ can be chosen in such a way that $2^{-j} k_i \in B(x_0, \epsilon/4)$ if $0 \leq i < \tilde{r}_a$.

Let $\mathfrak{b}(i)$ denote the smallest $n \in \mathbb{N}_0$ such that $k_i \in \Lambda_j^n(\partial\Omega)$. We see that whenever $k_i \in \Lambda_j^n(\partial\Omega)$, then $k_{i+1} \in \Lambda_j^m(\partial\Omega)$ for some $m \in \{n-1, n, n+1\}$, and thus conclude that $\mathfrak{b}(i) - 1 \leq \mathfrak{b}(i+1) \leq \mathfrak{b}(i) + 1$. From this, and since $\mathfrak{b}(0) > \beta + 1$, $\mathfrak{b}(\tilde{r}_a) = 0$, it follows that there must exist a number $r_a \in \mathbb{Z}$ (the one we are looking for) such that $\mathfrak{b}(\tilde{r}_1) = \beta + 1$. □

Another important observation is that we can choose j^{**} above in such a way that if $j > j^{**}$, $k \in \Lambda_j^{\beta+1}(\partial\Omega)$, and $2^{-j} k \in B(x_0, \epsilon_0/4)$, then $k \in \Lambda_j^{\beta-1}(G)$. Let us do just that, and let us collect all the l_a, $a \in A_j$, in the sets L_j. Note that these are precisely the l we have been looking for.

Note that if $a_1, a_2 \in A_j$, then

$$d(l_{a_1}, l_{a_2}) \geq d(a_1, a_2).$$

From this we infer that we can construct the sets F_j needed in lemma 3.3.2 if we can find sets $E_j \subset A_j$ such that

i. $\#E_j \geq 2^{j(d-1)}$

ii. $a_1, a_2 \in E_j$, $a_1 \neq a_2$ implies $d(a_1, a_2) \geq d_{M_j^0}$.

But this, thankfully, is trivial.

We are almost done. We only have to prove that we can indeed arrange matters as in (3.24).

3.4.4 A topology lemma

The problem can be reduced a bit. If we find z_0, σ, and m that satisfy the first three conditions in (3.24), then finding another pair that satisfies the last one is trivial too. But the existence of such z_0, σ, and m is a consequence of the following lemma, which will be proven at the end of this subsection.

Lemma 3.4.5. *Let $y_0 \in \mathbb{R}^{d-1}$, $\phi : \mathbb{R}^{d-1} \to \mathbb{R}$ continuous, $\delta > 0$, $Y = B(y_0, \delta)$, and let*

$$G_0 = \{(x, \phi(x)) \in \mathbb{R}^d : x \in Y\}.$$

Write $x_0 = (y_0, \phi(y_0))$. If $Q : \mathbb{R}^d \to \mathbb{R}^d$ is orthogonal, then for each $\eta > 0$ there exists $y \in B(x_0, \eta)$, $\kappa > 0$, and $m \in \{1, 2, \ldots, d\}$ such that

$$P_m Q B(y, \kappa) \subset (P_m Q G_0)^\circ$$

while $B(y, \kappa) \cap G_0 = \emptyset$.

We will need some preparations. Given two points $x_1, x_2 \in \mathbb{R}^d$, we denote by $\gamma = \gamma[x_1x_2] : [0,1] \to \mathbb{R}^d$ the function $\gamma(t) = (1-t)x_1 + tx_2$. Given $\gamma, \eta : [0,1] \to \mathbb{R}^d$, and if $\gamma(1) = \eta(0)$ we write $\iota = \gamma * \eta$ for the function $\iota : [0,1] \to \mathbb{R}^d$ given by

$$\iota(t) = \begin{cases} \gamma(2t) & \text{if } 0 \leq t < \frac{1}{2} \\ \eta(2t-1) & \text{if } \frac{1}{2} \leq t \leq 1 \end{cases}$$

We also write $\gamma[x_1x_2\cdots x_m] = \gamma[x_1x_2] * \gamma[x_2x_3] * \cdots * \gamma[x_{m-1}x_m]$.

The proof of lemma 3.4.5 is a simple consequence of the following proposition, which is just a simpler variant of it.

Proposition 3.4.6. *Let $y_0 \in \mathbb{R}^{d-1}$, $\phi : \mathbb{R}^{d-1} \to \mathbb{R}$ continuous, $\delta > 0$, $Y = B(y_0, \delta)$, and let*

$$G_0 = \{(x, \phi(x)) \in \mathbb{R}^d : x \in Y\}.$$

Write $x_0 = (y_0, \phi(y_0))$. If $Q : \mathbb{R}^d \to \mathbb{R}^d$ is orthogonal, then there exists $m \in \{1, 2, \ldots, d\}$ such that $(P_mQG_0)^\circ \neq \emptyset$.

Proof. For arbitrary n and $\epsilon > 0$, and given a point $z \in \mathbb{R}^n$, we denote by $B_\infty(z, \epsilon)$ the set $\{y \in \mathbb{R}^n : \|x-y\|_\infty < \epsilon\}$, and by $B_2(x, \epsilon)$ the set $\{y \in \mathbb{R}^n : \|x-y\|_2 < \epsilon\}$, which corresponds to the definition of $B(x, \epsilon)$ we have been using until now. Define $\alpha = \sup\{r > 0 : B_\infty(0, r) \subset B_2(0,1)\}$, and note that $\alpha \leq 1$.

Let $\sigma < \delta/2$, set $W := B_2(x_0, \sigma) \setminus G_0$, and note that this set is open and pathwise disconnected. There is, in particular, no path between the points $w_- := x_0 - \alpha\frac{\sigma}{2}e_d$ and $w_+ := x_0 + \alpha\frac{\sigma}{2}e_d$. If Q is orthogonal, then QW is also pathwise disconnected, and we cannot find a path between Qw_+ and Qw_- in QW.

Now suppose that the lemma is false for a certain orthogonal $Q_0 : \mathbb{R}^d \to \mathbb{R}^d$.

Set $W' := B_\infty(Q_0x_0, \alpha\delta) \setminus Q_0G_0$, $w'_+ := Q_0w_+$, and $w'_- := Q_0w_-$. By the way we have chosen the parameters, we have that $W' \subset Q_0W$, that W' is open, that $w'_+, w'_- \in W'$, and that there is no path between w'_+ and w'_- in W'. Let $\delta_0 > 0$ be such that $B_\infty(w'_+, \delta_0) \subset W'$ and $B_\infty(w'_-, \delta_0) \subset W'$.

Now, let $\xi_1 = w'_-$, and choose $\zeta_1 \in B_\infty(\xi_1, \frac{\delta_0}{d})$ such that $P_1\zeta_1 \notin P_1Q_0G_0$. This is possible because we assumed our lemma false, and thus $B_\infty(P_1\xi_1, \frac{\delta_0}{d}) \not\subset P_1Q_0G_0$. We next choose $\lambda_1 \in \mathbb{R}$ such that the first coordinates of $\xi_2 := \zeta_1 + \lambda_1 e_1$ and w'_+ are equal. Note that the path $\gamma[\xi_1\zeta_1\xi_2]$ lies fully in W'.

Next choose $\zeta_2 \in B_\infty(\xi_2, \frac{\delta_0}{d})$ such that $P_2\zeta_2 \notin P_2Q_0G_0$ (which again has to exist), and choose λ_2 such that the second coordinates of $\xi_3 = \zeta_2 + \lambda_2 e_2$ and w'_+ are equal. Note that, for the same reasons as above, the path $\gamma[\xi_1\zeta_1\xi_2\zeta_2\xi_2]$ lies fully in W'.

We proceed in this fashion until we have constructed ξ_d, and note immediately that $\xi_d \in B_\infty(w'_+, \delta_0)$, since each coordinate of ξ_d is at most at a distance of $\frac{(d-1)\delta_0}{d}$ of the corresponding one in w'_+. But we took care to never leave W', which implies that the path $\gamma[w'_-\zeta_1\xi_2\zeta_2\cdots\zeta_{d-1}w'_+]$ lies in W'. This contradiction finishes the proof. □

Proof of lemma 3.4.5. Without loss of generality, we can assume that $\delta < \frac{\epsilon_0}{2}$, and apply the same proof as before. The point y_0 we are looking for is the last ξ_i obtained before the process cannot be continued, and κ can be chosen as $\kappa = \frac{\delta_0}{d}$. □

Thus ends the proof of lemma 3.3.2, and thus also of lemma 3.2.8

Chapter 4

Towards a fictitious domain method with optimally smooth solutions

Introduction

In this chapter, we will introduce a fictitious domain method designed to produce optimally smooth solutions whenever the given data allows it, and which is also capable, in practice, to deliver on that promise. We also obtain, albeit with additional assumptions, a solid theoretical understanding of this method, proving convergence and reproduction of smoothness. The encouraging numerical results, to be presented in chapter five, suggest that our approach is promising, and that it should be the subject of further research.

The central idea of the approach is the division of responsibilities. Starting from our original boundary value problem on a domain, we formulate a very simple linear least-squares/fictitious-domain formulation on an extended domain whose solutions will all solve, when restricted to the original domain, the original problem. Although this extended problem does not have a unique solution, it can be seen to be solvable, and the solution can be chosen to depend continuously on the data. Instead of modifying this formulation to force it to produce smooth solutions, our approach assigns this responsibility to the solution process. We show how a simple iterative scheme is capable of recovering smoothness through what amounts to emergent behavior.

We begin in section 4.1 with a brief review of the definition and properties of the Moore-Penrose pseudoinverse. This building block is central in what follows. In section 4.2 we formulate and study the least-squares/fictitious-domain problems mentioned above. In section 4.3, starting from a sequence of discretizations of those problems, we propose a solution operator capable of recovering smoothness, and prove that it works under certain additional conditions. Finally, in section 4.4, we construct a candidate sequence of suitable discretizations.

We leave the actual implementation, and numerical experiments, to chapter five.

4.1 Moore-Penrose pseudoinverses

Let $\mathcal{H}_1$, $\mathcal{H}_2$ be two Hilbert spaces, and let $M : \mathcal{H}_1 \to \mathcal{H}_2$ be a bounded operator with closed range. Write $N := M_{|\mathcal{N}(M)^\perp}$, and recall that under these conditions, $N : \mathcal{N}(M)^\perp \to \mathcal{R}(M)$ is an isomorphism. The Moore-Penrose pseudoinverse is then defined by $M^\dagger := \mathcal{I}_{\mathcal{H}_1} N^{-1} P_{\mathcal{R}(M)}$. Here, $P_{\mathcal{R}(M)}$ denotes the orthogonal projection onto the range $\mathcal{R}(M)$ of M, and $\mathcal{I}_{\mathcal{H}_1}$ is the injection into $\mathcal{H}_1$. Given $b \in \mathcal{H}_2$, one has that $x = M^\dagger b$ is the unique minimizer of smallest norm in $\mathcal{H}_1$ of the functional $\varphi(x) := \|Mx - b\|^2_{\mathcal{H}_2}$. One also checks easily that $M^\dagger : \mathcal{H}_2 \to \mathcal{H}_1$ is a bounded operator with closed range.

The following theorem gives us a characterization of the Moore-Penrose pseudoinverse. See e.g. [15], p.182.

Theorem 4.1.1. *Let $B : \mathcal{H}_2 \to \mathcal{H}_1$ be a bounded linear operator with closed range. Then the following are equivalent*

(i). $B = M^\dagger$

(ii). $BMx = x$ *for all* $x \in \mathcal{N}(M)^\perp$, *and* $By = 0$ *for all* $y \in \mathcal{R}(M)^\perp$.

(iii). $MB = P_{\mathcal{R}(M)}$, *and* $BM = P_{\mathcal{N}(M)^\perp} = P_{\mathcal{R}(B)}$.

(iv). $(MB)^* = MB$, $(BM)^* = BM$, $MBM = M$, *and* $BMB = B$.

One has, furthermore, that if $Q : \mathcal{H}_1 \to \mathcal{H}_1$ is an orthogonal projector, then $Q^\dagger = Q$.

For the proof of these facts, and for further information, we refer to [15], chapter 8.

A remark is in order with respect to the numerical aspects of using pseudoinverses. The traditional approach to obtaining the pseudoinverse of a matrix is to use a singular value decomposition (SVD) which is rather expensive. Since we are not interested in the pseudoinverse per se, we will use instead appropriate iterative Krylov subspace methods, which have much better performance, to approximate the product of the pseudoinverse with a given vector. See subsection 4.5.

4.2 The formulation

4.2.1 Problem scope and assumptions

Consider the problem

$$\begin{aligned} Au &= f \qquad \text{on } \Omega, \\ Bu &= g, \end{aligned} \tag{4.1}$$

where A is a regular elliptic differential operator, and $B : H^2(\Omega) \to H^{\sigma(B)}(\partial\Omega)$ is either the Dirichlet or the Neumann boundary operator, with $\sigma(B) = 3/2$ resp. $\sigma(B) = 1/2$. We will assume that $\Omega \subset \mathbb{R}^d$ is a bounded domain with C^∞ boundary. The regularity assumptions on A and Ω can be relaxed, but at the price of obscuring the arguments. See remark 4.3.10.

We further assume that $f \in L_2(\Omega)$, and that $g \in H^{\sigma(B)}(\partial\Omega)$. This allows us to conclude that the solution u of problem (4.1) is at least in $H^2(\Omega)$. We shall further assume that

problem (4.1) is well posed; for each $f \in H^0(\Omega)$, $g \in H^{\sigma(B)}(\partial\Omega)$, there exists a unique solution $u \in H^2(\Omega)$ of (4.1), and this solution depends continuously on f, g.

Remark 4.2.1. *The regularity assumptions on the data restricts the applicability of the method designed here. We chose them since they simplify the theory in a few crucial aspects, and hope for further research to render the method applicable to more general settings.*

4.2.2 The formulation

We start by embedding Ω into a larger domain. Again for simplicity, we will assume that this domain is $\mathbb{T}^d$, and, of course, that Ω can be properly embedded in $\mathbb{T}^d$. That is, there exists $\epsilon > 0$ such that $\Omega \subset (0+\epsilon, 1-\epsilon)^d$. We will further assume that an extension of A to $\mathbb{T}^d$ is available, and we denote it again by A. In particular, we will use that (cf. 3.2)

$$(Au)_{|\Omega} = A(u_{|\Omega}) \qquad \forall\ u \in H^2(\mathbb{T}^d).$$

Note that this does *not* amount to a "pointwise" interpretation of the differential operator, as we are considering derivatives in the sense of distributions. What we are using here is that if $u \in H^2(\mathbb{T}^d)$, then Au can be identified in the usual way with an element of L_2.

We are looking for a way to obtain an $u^+ \in H^2(\mathbb{T}^d)$ which satisfies

$$u^+_{|\Omega} = u, \tag{4.2}$$

where u is the solution of (4.1). There are of course many elements of $H^2(\mathbb{T}^d)$ which would satisfy (4.2), but after considering the effects of smoothness on accuracy, we want to find one that is as smooth as possible. As was announced in the introduction, our approach will be to set up a minimal least squares problem whose solutions all satisfy (4.2), and then try to construct a smooth solution of said problem. Here we will concentrate on the first part of that program, addressing the second part in section 4.3.

Observe that the requirement (4.2) is equivalent to requiring that u^+ satisfies the equations

$$\begin{aligned} (Au^+)_{|\Omega} &= f, \\ Bu^+ &= g. \end{aligned}$$

Our first (prototype) least-squares/fictitious-domain problem will be as follows.

Problem **LSFD$_0$**: Given f and g as above, find $u^+ \in H^2(\mathbb{T}^d)$ such that it minimizes the functional

$$\Phi_\Omega(v) = \|r_\Omega Av - f\|^2_{H^0(\Omega)} + \|Bv - g\|^2_{H^{\sigma(B)}(\partial\Omega)}. \tag{4.3}$$

We see immediately that there is at least one drawback of this formulation: It still involves a space defined on Ω. To remove this space we introduce the operator $C_\Omega : H^0(\mathbb{T}^d) \to H^0(\mathbb{T}^d)$, defined by

$$C_\Omega f := \chi_\Omega \cdot f, \tag{4.4}$$

which assigns to each $f \in H^0(\mathbb{T}^d)$ the extension by zero of its restriction to Ω. It is easy to see that C_Ω is an orthogonal projector with respect to the canonical L_2 norm in $H^0(\mathbb{T}^d)$.

Remark 4.2.2. *The orthogonality of the operator C_Ω plays a crucial role in what follows. A suitable substitute (either for the orthogonality or for the restriction itself) would be needed to extend the method under discussion to more general settings.*

We can now reformulate a new least-squares/fictitious-domain problem, using C_Ω to avoid the space $H^0(\Omega)$.

Problem **LSFD**: Given f and g as above, and given any extension $f_1 \in H^0(\mathbb{T}^d)$ of f, find $u^+ \in H^2(\mathbb{T}^d)$ such that it minimizes the functional

$$\Phi(v) = \|C_\Omega A v - f_1\|^2_{H^0(\mathbb{T}^d)} + \|Bv - g\|^2_{H^{\sigma(B)}(\partial\Omega)}. \tag{4.5}$$

We will now check that these least-squares problems can indeed be used to solve our original problem. This involves verifying that any solution of these problems satisfies (4.2), and that we can obtain solutions whose norm is bounded by the norm of the data. We will also find out that the solutions of minimal norm of (4.3) and (4.5) are equal.

For notational simplicity, let $\mathcal{H}^l := H^2(\mathbb{T}^d)$, $\mathcal{H}^r_\Omega := H^0(\Omega) \times H^{\sigma(B)}(\partial\Omega)$, $\mathcal{H}^r := H^0(\mathbb{T}^d) \times H^{\sigma(B)}(\partial\Omega)$, and let $M_\Omega : \mathcal{H}^l \to \mathcal{H}^r_\Omega$, $M : \mathcal{H}^l \to \mathcal{H}^r$ be given by

$$M_\Omega := \begin{pmatrix} r_\Omega A \\ B \end{pmatrix} \qquad\qquad M := \begin{pmatrix} C_\Omega A \\ B \end{pmatrix}$$

where r_Ω is the restriction operator, and C_Ω is the orthonormal projector introduced above. As done before, we endow $\mathcal{H}^r_\Omega$ and $\mathcal{H}^r$ with the corresponding Euclidean tensor product norms, to ensure that they are Hilbert spaces.

With these operators, and setting $b_\Omega = (f, g)^T$, $b = (f_1, g)^T$, we rewrite the functionals appearing in problems LSFD_0 and LSFD as

$$\Phi_\Omega(v) = \|M_\Omega v - b_\Omega\|^2_{\mathcal{H}^r_\Omega} \qquad\qquad \Phi(v) = \|Mv - b\|^2_{\mathcal{H}^r}.$$

Theorem 4.2.3.

(i). The operators M_Ω and M are bounded and have closed range, (and thus have bounded pseudoinverses).

(ii). If $f_1 \in H^0(\mathbb{T}^d)$ is an extension of $f \in H^0(\Omega)$, then $u^+ := M^\dagger b$ and $w^+ := M^\dagger_\Omega b_\Omega$ both satisfy (4.2).

(iii). It holds that $u^+ = w^+$.

Proof. That these operators are bounded is obvious.

From the well-posedness of problem (4.1) it follows that M_Ω is surjective. To see this, let $h = (\phi, \gamma)^T \in \mathcal{H}^r_\Omega$ be arbitrary. Then there exists a unique $v \in H^2(\Omega)$ which satisfies (4.1), and thus any extension $v^+ \in H^2(\mathbb{T}^d)$ of v satisfies $M_\Omega v^+ = h$. Surjectivity immediately implies that the range of M_Ω is closed.

To see that the range of M is closed, we use again the well-posedness of (4.1) to prove that

$$\mathcal{R}(M) = \{(\phi, \gamma)^T \in \mathcal{H}^r : \phi_{|\Omega^c} = 0\}.$$

Now for any convergent sequence $h_n = (\phi_n, \gamma_n) \in \mathcal{R}(M)$, $n = 1, 2, \ldots$, we have that $\phi_{n|\Omega^c} = 0$. By continuity of the restriction operator it follows that for $h = (\phi, \gamma) := \lim_{n\to\infty} h_n$ it holds $\phi_{|\Omega^c} = 0$. Thus $h \in \mathcal{R}(M)$, showing that this set is closed. This finishes the proof of (i).

Back to problem LSFD_0, we conclude from the surjectivity of M_Ω that $\min \Phi_\Omega(v) = 0$. Since $w^+ = M_\Omega^\dagger b_\Omega$ is a minimizer of Φ_Ω, we have that $r_\Omega A w^+ = A r_\Omega w^+ = f$, $Bw^+ = g$, and thus that w^+ satisfies (4.2)

To see that $u^+ := M^\dagger b$ also satisfies (4.2), we begin by computing the minimum of Φ. For this, observe first that (trivially) $\Phi(v) \geq \|C_\Omega A v - f_1\|^2_{H^0(\mathbb{T}^d)}$. Since $Av \in H^0(\mathbb{T}^d)$, and since C_Ω is an orthogonal projection in this space, we see that $\Phi(v) \geq \|(C_\Omega - I) f_1\|^2_{H^0(\mathbb{T}^d)}$. A simple computation also gives us that $\Phi(w^+) = \|(C_\Omega - I) f_1\|^2_{H^0(\mathbb{T}^d)}$, showing that this last quantity is indeed the minimum of Φ.

Now observe that u^+, being the minimizer of Φ, must satisfy

$$\begin{aligned} \Phi(u^+) &= \|C_\Omega A u^+ - f_1\|^2_{H^0(\mathbb{T}^d)} + \|Bu^+ - g\|^2_{H^{\sigma(B)}(\partial\Omega)} \\ &= \|(C_\Omega - I) f_1\|^2_{H^0(\mathbb{T}^d)}. \end{aligned} \tag{4.6}$$

But one readily checks that, since C_Ω is an orthogonal projector,

$$\|C_\Omega A u^+ - f_1\|^2_{H^0(\mathbb{T}^d)} = \|C_\Omega A u^+ - C_\Omega f_1\|^2_{H^0(\mathbb{T}^d)} + \|(C_\Omega - I) f_1\|^2_{H^0(\mathbb{T}^d)},$$

and thus from (4.6) it follows that $C_\Omega A u^+ = C_\Omega f_1$, and $Bu^+ = g$. Now $C_\Omega A u^+ = C_\Omega f_1$ is possible if, and only if, $(Au^+)_{|\Omega} = f_{1|\Omega} = f$. So u^+ satisfies (4.2), finishing the proof of (ii).

Finally, let us show that $M_\Omega^\dagger b_\Omega = M^\dagger b$. The key observation here is that for any $v \in H^2(\mathbb{T}^d)$, it holds that

$$\|M_\Omega v\|_{\mathcal{H}^r_\Omega} = \|Mv\|_{\mathcal{H}^r}. \tag{4.7}$$

This follows from the fact that $\|C_\Omega h\|_{H^0(\mathbb{T}^d)} = \|h_{|\Omega}\|_{H^0(\Omega)}$ for each $h \in H^0(\mathbb{T}^d)$. As a consequence of (4.7) we have that M and M_Ω have the same kernel.

Now, for $u^+ = M^\dagger b$, and $w^+ = M_\Omega^\dagger b_\Omega$ we have that

$$\begin{aligned} &\|M_\Omega(u^+ - w^+)\|^2_{\mathcal{H}^r_\Omega} = \\ &\qquad \|r_\Omega A u^+ - r_\Omega A w^+\|^2_{H^0(\Omega)} + \|Bu^+ - Bw^+\|^2_{H^{\sigma(B)}(\partial\Omega)} = 0, \end{aligned}$$

and thus $u^+ - w^+ \in \mathcal{N}(M_\Omega)$. But since $M_\Omega^\dagger b_\Omega \perp \mathcal{N}(M_\Omega) = \mathcal{N}(M) \perp M^\dagger b$, it holds that both u^+ and w^+ are orthogonal to $\mathcal{N}(M_\Omega)$, and thus $u^+ - w^+ = 0$. This proves (iii) and finishes the proof of theorem 4.2.3. □

Remark 4.2.4. *When choosing a discretization scheme for problem LSFD, it should be kept in mind that this result depends critically on the fact that C_Ω is an orthogonal projector. On the other hand, it is important to note that theorem 4.2.3 remains valid if we change the norms of $H^2(\mathbb{T}^d)$ to any equivalent norm (the same applies to $H^{\sigma(B)}(\partial\Omega)$).*

4.3 Recovering smoothness

The method to recover smoothness we will present in this section cannot, at present, be justified completely from a theoretical point of view. The method performs quite well in practice, however, so that even though the theory we present here does not cover every aspect, we can safely conclude that our approach is promising. Further research is needed to complete the picture.

The available theory has the following form. We assume the existence of a sequence of linear discrete maps which satisfies a certain set of properties, and subsequently prove that, if such a sequence exists, and the data allows it, then we can construct a smooth solution to problem LSFD.

Let $\{V_j\}_{j\in\mathbb{N}_0}$, $\{V_j^r\}_{j\in\mathbb{N}_0}$ be nested sequences of linear spaces such that

$$\begin{aligned} \mathcal{A}_2^s(H^2(\mathbb{T}^d), \{V_j\}_{j\in\mathbb{N}_0}) &= H^{s+2}(\mathbb{T}^d), \\ \mathcal{A}_2^s(\mathcal{H}^r, \{V_j^r\}_{j\in\mathbb{N}_0}) &= H^s(\mathbb{T}^d) \times H^{\sigma(B)+s}(\partial\Omega), \end{aligned} \tag{4.8}$$

for some range $0 < s \leq s_0$. Additionally, let $\{Q_j\}_{j\in\mathbb{N}_0}$ and $\{Q_j^r\}_{j\in\mathbb{N}_0}$ be uniformly bounded sequences of projectors with $\mathcal{R}(Q_j) = V_j$, $\mathcal{R}(Q_j^r) = V_j^r$. To recover smoothness we use a sequence of linear maps $M_j : V_j \to V_j^r$ satisfying a few properties that we are going to discuss now in some depth.

It is not known, at present, whether such a sequence exists; see remark 4.3.9 for a summary of the difficulties. In section 4.4, however, we will construct a sequence of operators which, in view of the numerical evidence of chapter five, seems to us to be a strong candidate.

The first thing we would like to require from this sequence of maps is that they can be used to approximately solve problem LSFD. In particular we expect it to satisfy

(A1) $$M_j Q_j u \to Mu, \qquad M_j^\dagger Q_j^r b \to M^\dagger b,$$

in th e topology of $\mathcal{H}^r$, $H^2(\mathbb{T}^d)$, respectively, for all $u \in H^2(\mathbb{T}^d)$, and all $b \in \mathcal{H}^r$. By the uniform boundedness theorem (see e.g. [15], page 165) we have as a consequence of this assumption the existence of a finite constant $C_M > 0$ such that

$$\max\{\|M_j\|, \|M_j^\dagger\|\} \leq C_M \qquad j = 0, 1, \ldots \tag{4.9}$$

Suppose now that $b \in \mathcal{A}_2^{\tilde{s}}(\mathcal{H}^r, \{V_j\})$, for some $\tilde{s} > 0$, and write $b_j = Q_j^r b$. The next assumption is based on our hope that the solution of the problem

$$\min_{u_j \in V_j} \varphi_j(u_j) := \|M_j u_j - b_j\|_{\mathcal{H}^r}^2$$

is a good "guess" for the minimizer of φ_{j+1}. We will assume that there exists some $s_1 \in (0, s_0]$ such that

$(A2_0)$ $$\|M_{j+1} M_j^\dagger b_j - P_{\mathcal{R}(M_{j+1})} b_{j+1}\|_{\mathcal{H}^r} \lesssim 2^{-js^*} \|b\|_{\mathcal{A}_2^{s^*}(\mathcal{H}^r, \{V_j\}_{j\in\mathbb{N}_0})},$$

with $s^* = \min\{\tilde{s}, s_1\}$. While $(A2_0)$ already captures the essence of our assumption, we will ask for the (only slightly stronger)

(A2) $$\left\| \left\{ 2^{js^*} \|M_{j+1} M_j^\dagger b_j - P_{\mathcal{R}(M_{j+1})} b_{j+1}\|_{\mathcal{H}^r} \right\} \right\|_{\ell_2} \lesssim \|b\|_{\mathcal{A}_2^{s^*}(\mathcal{H}^r, \{V_j\}_{j\in\mathbb{N}_0})},$$

which will help us avoid some epsilons in the proofs that follow.

Remark 4.3.1. *Note that this is really only an epsilon, as it is easy to see that if (A2$_0$) holds for a given s_0^*, then (A2) holds for each $s^* < s_0^*$.*

Finally, we will require from the sequence $\{M_j\}$ that the kernels of these operators be nested.

(A3) $$\mathcal{N}(M_j) \subset \mathcal{N}(M_{j+1}).$$

This last assumption is what really drives the method we will introduce now.

The intuitive idea behind our method is as follows. Suppose that $\{V_j\}$ is the B-spline MRA introduced in 2.6.2. Then the minimizer $u_j = M_j^\dagger b_j$ of φ_j will have the same smoothness as any other element in V_j, and, under the right circumstances, we will have that u_j is a good approximation of some smooth solution of problem LSFD.

While we may expect u_j to converge to a solution of LSFD, we cannot expect this limit to be smooth. Looking at the kernel of M, we see that it consists of functions $\kappa \in H^2(\mathbb{T}^d)$ which are zero on Ω, and which satisfy $B\kappa = 0$. There is no reason to expect in general that an extensions of u to $\mathbb{T}^d$ with higher Sobolev smoothness than H^2 is orthogonal to this kernel.

So to obtain such a smooth extension of u using the solutions u_j of the discrete problems we may have to "grow" a component in this kernel. Our plan is to "lift" the smoothness of the finite dimensional spaces $\{V_j\}$ by collecting the components of the solutions u_j in the kernels of the operators M_{j+1}. Thus, the definition of our solution operator starts with a standard solution for some initial j (for simplicity we begin with $j = 0$),

(4.10) $$S_0 b := M_0^\dagger Q_0^r b = M_o^\dagger b_0,$$

and then define

(4.11) $$S_{j+1} b := P_{\mathcal{N}(M_{j+1})} S_j b + M_{j+1}^\dagger Q_{j+1}^r b.$$

Theorem 4.3.2. *If $\{M_j\}$ satisfies (A1), (A2$_0$), (A3), and $b \in \mathcal{A}_2^{s^*}(\mathcal{H}^r, \{V_j\})$, then $\{S_j b\}_{j\in\mathbb{N}_0}$ converges.*

Proof. We is enough to show that $\{S_j b\}_{j\in\mathbb{N}_0}$ is a Cauchy sequence.

From (4.10) and (4.11) we can derive an alternative expression for $S_j b$. We have that

$$S_j b = \sum_{i=1}^{j} P_{\mathcal{N}_j} P_{\mathcal{N}_{j-1}} \cdots P_{\mathcal{N}_i} M_{i-1}^\dagger b_{i-1} + M_j^\dagger b_j,$$

where we have written $\mathcal{N}_j := \mathcal{N}(M_j)$. Thus,

$$\begin{aligned} S_{j+1} b - S_j b &= P_{\mathcal{N}_{j+1}} S_j b + M_{j+1}^\dagger b_{j+1} - S_j b \\ &= M_{j+1}^\dagger b_{j+1} - P_{\mathcal{N}_{j+1}^\perp} S_j b \\ = M_{j+1}^\dagger b_{j+1} - \sum_{i=1}^{j} P_{\mathcal{N}_{j+1}^\perp} P_{\mathcal{N}_j} P_{\mathcal{N}_{j-1}} \cdots P_{\mathcal{N}_i} M_{i-1}^\dagger b_{i-1} &- P_{\mathcal{N}_{j+1}^\perp} M_j^\dagger b_j. \end{aligned}$$

Now, since $\mathcal{N}_j \subset \mathcal{N}_{j+1}$, we have that $\mathcal{N}_j \perp \mathcal{N}_{j+1}^{\perp}$ so that $P_{\mathcal{N}_{j+1}^{\perp}} P_{\mathcal{N}_j} = 0$. This eliminates the sum in the last expression above. Continuing with the calculations, we observe that

$$\begin{aligned} S_{j+1}b - S_j b &= P_{\mathcal{N}_{j+1}^{\perp}} \left(M_{j+1}^{\dagger} b_{j+1} - M_j^{\dagger} b_j \right) \\ &= M_{j+1}^{\dagger} M_{j+1} \left(M_{j+1}^{\dagger} b_{j+1} - M_j^{\dagger} b_j \right), \end{aligned}$$

so that

$$\begin{aligned} \|S_{j+1}b - S_j b\|_{H^2} &\leq \|M_{j+1}^{\dagger}\| \|P_{\mathcal{R}(M_{j+1})} b_{j+1} - M_{j+1} M_j^{\dagger} b_j\|_{\mathcal{H}^r} \\ &\leq C_M \|P_{\mathcal{R}(M_{j+1})} b_{j+1} - M_{j+1} M_j^{\dagger} b\|_{\mathcal{H}^r}, \end{aligned} \tag{4.12}$$

where C_M is the constant in (4.9). Using assumption $(A2_0)$, we obtain that $\|S_{j+1}b - S_j b\|_{H^2} \lesssim 2^{-js^*}$. A simple geometric sums argument now gives us that $\{S_j b\}_{j \in \mathbb{N}_0}$ is indeed a Cauchy sequence. □

The next task will be to prove that we really obtain a solution to problem LSFD from

$$Sb := \lim_{j \to +\infty} S_j b.$$

Theorem 4.3.3. *It holds that Sb is a minimizer of $\Phi(u) = \|Mu - b\|_{\mathcal{H}^r}$.*

The proof of this theorem requires some preparations.

Lemma 4.3.4. *Let $\mathcal{H}_1$, $\mathcal{H}_2$ be a pair of Hilbert spaces, and $\{A_j\}$ a sequence of bounded linear operators which is pointwise convergent. It is known that then the operator $A : \mathcal{H}_1 \to \mathcal{H}_2$ given by $Av = \lim_{j \to \infty} A_j v$ is bounded and linear. If also $A^{\dagger} w = \lim_{j \to \infty} A_j^{\dagger} w$ for all $w \in \mathcal{H}_2$, then*

$$P_{\mathcal{N}(A_j)} v \to P_{\mathcal{N}(A)} v \qquad P_{\mathcal{N}(A_j)^{\perp}} v \to P_{\mathcal{N}(A)^{\perp}} v \tag{4.13}$$

$$P_{\mathcal{R}(A_j)} w \to P_{\mathcal{R}(A)} w \qquad P_{\mathcal{R}(A_j)^{\perp}} w \to P_{\mathcal{R}(A)^{\perp}} w \tag{4.14}$$

Proof. Note that, since $P_{V^{\perp}} = (I - P_V)$, the claim on the right of (4.13) follows trivially from that on the left. Note also that since $\mathcal{R}(A_j) = \mathcal{N}(A_j^{\dagger})^{\perp}$, we obtain (4.14) from (4.13). Thus, it is enough to prove the claim on the left of (4.13).

Let $v \in \mathcal{H}_1$, and write $v = v_0 + v_1$, where $v_0 = P_{\mathcal{N}(A)} v$, and $v_1 = v - P_{\mathcal{N}(A)} v = P_{\mathcal{N}(A)^{\perp}} v$. Now, we only have to prove that $P_{\mathcal{N}(A_j)} v_0 \to v_0$ and $P_{\mathcal{N}(A_j)} v_1 \to 0$ when $j \to \infty$.

From the hypothesis on $\{A_j\}$ it follows that $A_j v_0 \to A v_0 = 0$, and so

$$P_{\mathcal{N}(A_j)^{\perp}} v_0 \to 0 \tag{4.15}$$

when $j \to \infty$. To see this, note that by the uniform boundedness theorem $\|A_j^{\dagger}\| \leq C$ for all j and some $C > 0$, and recall that $A_j^{\dagger} A_j = P_{\mathcal{N}(A_j)^{\perp}}$. Thus,

$$\|P_{\mathcal{N}(A_j)^{\perp}} v_0\|_{\mathcal{H}_1} \leq \|A_j^{\dagger}\| \|A_j v_0\|_{\mathcal{H}_2} \leq C \|A_j v_0\|_{\mathcal{H}_2} \to 0,$$

from which (4.15), as well as $P_{\mathcal{N}(A_j)} v_0 = (I - P_{\mathcal{N}(A_j)^{\perp}}) v_0 \to v_0$, follows.

Now, from $A_j v_1 - A v_1 \to 0$, we obtain that

$$\|P_{\mathcal{N}(A_j)^\perp} v_1 - A_j^\dagger A v_1\|_{\mathcal{H}_1} \le \|A_j^\dagger\| \|A_j v_1 - A v_1\|_{\mathcal{H}_2} \to 0. \tag{4.16}$$

But since $A_j^\dagger A v_1 \to v_1$, we can infer from (4.16) that $P_{\mathcal{N}(A_j)^\perp} v_1 \to v_1$, and thus $P_{\mathcal{N}(A_j)} v_1 \to 0$ when $j \to \infty$.

So,

$$\begin{aligned} P_{\mathcal{N}(A_j)} v &= P_{\mathcal{N}(A_j)} v_0 + P_{\mathcal{N}(A_j)} v_1 \\ &\to v_0 = P_{\mathcal{N}(A)} v, \end{aligned}$$

finishing the proof. □

Proof of theorem 4.3.3. Observe that

$$\begin{aligned} \|M_j S_j b - MSb\|_{\mathcal{H}^r} &= \|M_j(S_j b - Sb) + M_j Sb - MSb\|_{\mathcal{H}^r} \\ &\le C_M \|S_j b - Sb\|_{\mathcal{H}^r} + \|M_j Sb - MSb\|_{\mathcal{H}^r} \to 0 \end{aligned}$$

by theorem 4.3.2, and by (A1), so that $M_j S_j b \to MSb$.

Writing $\mathcal{R}_j := \mathcal{R}(M_j)$, and noting that $M_j S_j b = P_{\mathcal{R}_j} b_j$, we also have that

$$\begin{aligned} \|M_j S_j b - P_{\mathcal{R}(M)} b\|_{\mathcal{H}^r} &= \|P_{\mathcal{R}_j} b_j - P_{\mathcal{R}(M)} b\|_{\mathcal{H}^r} \\ &\le \|P_{\mathcal{R}_j}(b_j - b)\|_{\mathcal{H}^r} + \|P_{\mathcal{R}_j} b - P_{\mathcal{R}(M)} b\|_{\mathcal{H}^r} \to 0 \end{aligned}$$

since $b_j \to b$, and using lemma 4.3.4.

In any case, we have that $M_j S_j b \to P_{\mathcal{R}(M)} b$, and also $M_j S_j b \to MSb$, so that $MSb = P_{\mathcal{R}(M)} b$. But then

$$\begin{aligned} \min_{v \in H^2(\mathbb{T}^d)} \Phi(v) &= \min_{v \in H^2(\mathbb{T}^d)} \|Mv - b\|_{\mathcal{H}^r}^2 \\ &\ge \|P_{\mathcal{R}(M)} b - b\|_{\mathcal{H}^r}^2 = \|MSb - b\|_{\mathcal{H}^r}^2, \end{aligned}$$

finishing the proof. □

Theorem 4.3.5. *If $\{M_j\}$ satisfies (A1), (A2), and (A3), then for any $0 < s \le s_1$, the operator $S : \mathcal{A}_2^s(\mathcal{H}^r, \{V_j\}) \to \mathcal{A}_2^s(H^2(\mathbb{T}^d), \{V_j\})$ given by $b \mapsto Sb$ is linear and bounded.*

Proof. Let $b, d \in \mathcal{A}_2^s(\mathcal{H}^r, \{V_j\})$, and $\alpha, \beta \in \mathbb{R}$. Then $S(\alpha b + \beta d)$ exists, and is the limit of $S_j(\alpha b + \beta d) = \alpha S_j b + \beta S_j d$, which in turn converges to $\alpha Sb + \beta Sd$. This settles the linearity. It remains to see whether $Sb \in \mathcal{A}_2^s(H^2(\mathbb{T}^d), \{V_j\})$, and whether S is bounded.

Using (A2) and (4.12) (the s^* there amounts to our current s), we obtain

$$\left\{2^{js} \|S_j b - S_{j+1} b\|_{H^2(\mathbb{T}^d)}\right\} \in \ell_2$$

and

$$\left\| \{2^{js} \|S_j b - S_{j+1} b\|_{H^2(\mathbb{T}^d)}\} \right\|_{\ell_2} \lesssim \|b\|_{\mathcal{A}_2^s}.$$

We also have $\|S_j b - Sb\|_{H^2(\mathbb{T}^d)} \le \sum_{i \ge j} \|S_{i+1} b - S_i b\|_{H^2(\mathbb{T}^d)}$, which inspires us to borrow the following lemma, found in [16], p. 408.

Lemma 4.3.6 (Discrete Hardy Inequality). *Let $\{a_k\}_{k\in\mathbb{N}_0}$, $\{b_k\}_{k\in\mathbb{N}_0}$ be sequences of real numbers, and let $\alpha > 0$. If for some $c > 0$, $0 < \mu \leq q$,*

$$|b_k| \leq c \left(\sum_{j=k}^{\infty} |a_j|^{\mu} \right)^{1/\mu}$$

holds for all k, then

$$\left(\sum_{k=0}^{\infty} (2^{k\alpha} b_k)^q \right)^{1/q} \leq c \left(\sum_{k=0}^{\infty} (2^{k\alpha} a_k)^q \right)^{1/q}.$$

Thus, we conclude that

$$\left\| \{2^{js} \|Sb - S_j b\|\} \right\|_{\ell_2} \lesssim \left\| \{2^{js} \|S_j b - S_{j+1} b\|\} \right\|_{\ell_2} \lesssim \|b\|_{\mathcal{A}_2^s}.$$

But $\|Sb - S_j b\| \geq \|Sb - P_{V_j} Sb\|$, so that we obtain

$$\begin{aligned} \|Sb\|_{\mathcal{A}_2^s} &= \left\| \{2^{js} \|Sb - P_{V_j} Sb\|\} \right\|_{\ell_2} \\ &\lesssim \left\| \{2^{js} \|Sb - S_j b\|\} \right\|_{\ell_2} \lesssim \|b\|_{\mathcal{A}_2^s} \quad \square \end{aligned}$$

A straight-forward corollary of theorem 4.3.5 is the following.

Corollary 4.3.7. *The convergence behavior of $\{S_j b\}$ is given by*

$$\|S_j b - Sb\|_{H^2(\mathbb{T}^d)} \lesssim 2^{-js}$$

In summary, given a smooth initial extension f_1 of f, and if g is smooth too, we obtain via the linear bounded operator S a solution to problem LSFD with the same degree of smoothness. This, of course, provided the discrete operators M_j, $j \in \mathbb{N}_0$ satisfy (A1), (A2), and (A3). We summarize theorems 4.3.2, 4.3.3, and 4.3.5 as follows.

Theorem 4.3.8. *Let $\{V_j\}_{j\in\mathbb{N}_0}$, $\{V_j^r\}_{j\in\mathbb{N}_0}$ be nested sequences of linear spaces such that (4.8) holds. Let $\{M_j\}$, $M_j : V_j \to V_j^r$ be a sequence of linear maps satisfying (A1), (A2), and (A3). Let $f \in H^s(\Omega)$, $g \in H^{\sigma(B)+s}(\partial\Omega)$ for some $s_1 \geq s > 0$, and let $f_1 \in H^s(\mathbb{T}^d)$ be an extension of f to $\mathbb{T}^d$. Then*

1. *The sequence $\{S_j b\}$, with $b = (f_1, g)$ converges to Sb at a rate of $O(2^{-js})$ in the topology of $\mathcal{H}^r$.*
2. *$Sb \in H^{2+s}(\mathbb{T}^d)$*
3. *$(Sb)_{|\Omega}$ is the solution of problem (4.1).*

Thus, to obtain a smooth solution to problem LSFD, we start by choosing an arbitrary, but smooth, extension of f, and then apply S.

Remark 4.3.9. *The difficulty in finding the sequence $\{M_j\}$ is, in essence, that singular operators are hard to discretize properly. Even when the infinite dimensional problem is well posed, we cannot just use a standard Galerkin approach to obtain discrete problems. Attacking the problem via regularization is an option that does not lead too far. The above method, through assumption (A3), is based critically on the singularity of the discrete operators M_j, so eliminating it is not helpful.*

Remark 4.3.10. *Note that if A and $\partial\Omega$ do not satisfy the extreme regularity requirements imposed in subsection 4.2.1, then their regularity adds just another upper bound to s in theorem 4.3.8.*

The sequence of discrete problems we introduce in the next section seems, at least numerically, to satisfy (A1), (A2), and (A3). The author is convinced that it is possible, although not at all trivial, to prove that the sequence in question does indeed satisfy the necessary assumptions.

4.4 A sequence of discrete problems

In this section, we will discretize a simple two-dimensional family of problems using a Petrov-Galerkin approach. This sequence of discrete problems will be used in the next section to perform numerical experiments using the method outlined in the previous sections. We will go to some level of detail to explain the motivation behind each choice.

4.4.1 The model problem

Our model problem is

$$\begin{aligned} (-\triangle + \mu I)u &= f \qquad \text{on } \Omega, \\ Bu &= g, \end{aligned} \tag{4.17}$$

where $\mu \geq 0$, and B is either the Dirichlet or the Neumann[1] boundary operator. As before, we take $f \in H^0(\Omega)$, and $g \in H^{\sigma(B)}(\partial\Omega)$. We also assume that we have already an initial extension f^+ at hand. The domain $\Omega \subset \mathbb{R}^2$ is any domain with smooth boundary.

4.4.2 Norms and spaces

We want to find approximations to the minimizer u^+ of the functional

$$\Phi(v) = \|C_\Omega A v - f_1\|^2_{H^0(\mathbb{T}^d)} + \|Bv - g\|^2_{H^{\sigma(B)}(\partial\Omega)}. \tag{4.18}$$

Keeping our goal in mind (that is, to solve (4.17)), we will use the insight of remark 4.2.4 and begin by changing the involved norms.

We will approximate $u^+ \in H^2(\mathbb{T}^2)$ from the spaces V_j, $j \in \mathbb{N}_0$, which we choose to be the periodic B-spline spaces of order m, with $m \geq 3$ fixed, on dyadic grids of meshlength

[1] In this case, we assume $\mu > 0$ to ensure well-posedness.

2^{-j}. We chose an appropriate $\tilde{m}$, and let Ψ be the primal wavelet basis of $H^2(\mathbb{T}^d)$ of order m and dual order $\tilde{m}$. We will use for $H^2(\mathbb{T}^d)$ the norm induced by this basis (see section 2.6.3) since it is straight-forward to compute.

Remark 4.2.4 also warns us against changing the norm in $L_2(\mathbb{T}^d) = H^0(\mathbb{T}^d)$. There we will approximate from the spaces

$$V_j^0 = \{f \in L_2(\mathbb{T}^d) : f_{|\square_{jk}} \in \Pi_{m-1}\},$$

of discontinuous piecewise polynomials of degree $m-1$, which can be endowed easily with an orthonormal basis. To construct such a basis for V_j^0, we apply first Gram-Schmidt orthonormalization in $L_2([0,1]^2)$ to the monomials $x^i y^j$ with $i+j \leq m-1$, $i,j \geq 0$. We write $\{\phi^0, \phi^1, \ldots, \phi^n\}$ for the functions we thus obtain (here, $n = (m+1)m/2$), and note that it also is a basis for V_0^0. We write $\phi_{jk}^i(x) = 2^j \phi^i(2^j x - k)$, and observe that the set $\{\phi_{jk}^i : i = 1, \ldots, n,\ k \in \mathcal{Z}_j^2\}$, with $\mathcal{Z}_j = \mathbb{Z}/2^j\mathbb{Z}$ is an orthonormal basis for V_j^0. We use the canonical norm on $L_2(\mathbb{T}^d)$.

We identify $H^{\sigma(B)}(\partial\Omega)$ with $H^{\sigma(B)}(\mathbb{T})$ using a suitable parametrization $\Gamma : \mathbb{T} \to \partial\Omega$. For $H^{\sigma(B)}(\mathbb{T})$ we choose again the B-spline biorthogonal wavelet bases Ψ^Γ, $\tilde{\Psi}^\Gamma$, with fixed orders $m^\Gamma \geq 3$, and $\tilde{m}^\Gamma$ accordingly. But instead of spanning $H^{\sigma(B)}(\mathbb{T})$ with the primal basis, we use for that purpose the (properly rescaled) *dual* basis $\tilde{\Psi}^\Gamma$. The reason for doing this is that, from a numerical point of view, it will be far easier to compute inner products with the primal wavelets, which are piecewise polynomial, than with the duals. This implies that in $H^{\sigma(B)}(\partial\Omega)$ we approximate from the spaces $\tilde{V}_j^\Gamma$ spanned by the dual wavelets up to level j. We will write Q_j^Γ for the oblique projector onto $\tilde{V}_j^\Gamma$ associated with Ψ^Γ (again, we refer to section 2.6.3). We will also use the norm induced by these bases for $H^{\sigma(B)}$, $H^{-\sigma(B)}$.

Given an element in v in any of these spaces, we decorate it with an underscore to denote the Euclidean vector consisting of its coefficients. Thus, if $v \in V_j$, then $\underline{v} \in \ell_2(\nabla_j)$ is such that $v = \sum_{\lambda \in \nabla_j} \underline{v}_\lambda \psi_\lambda$.

4.4.3 The discrete operators

We define $A_j : V_j \to V_j^0$ by $A_j := P_j A_{|V_j}$, where $P_j := P_{V_j^0}$ is the orthogonal projector onto V_j^0, given by

$$P_{V_j^0} f = \sum_{k,i} \langle f, \phi_{jk}^i \rangle \phi_{jk}^i.$$

Given a function $v \in V_j$, we have that its trace on $\partial\Omega$ is given by $B^{\mathcal{D}} v = v \circ \Gamma \in H^{3/2}(\partial\Omega)$. If we are dealing with Neumann boundary conditions, then $B^{\mathcal{N}} v = [(\nabla v) \circ \Gamma] \cdot \mathbf{n} \in H^{1/2}(\partial\Omega)$, where $\mathbf{n}(t)$ is the outward normal of $\partial\Omega$ at the point $\Gamma(t)$. Thus, we define either $B_j^{\mathcal{D}}, B_j^{\mathcal{N}} : V_j \to \tilde{V}_j^\Gamma$, as appropriate[2], through

$$B_j^{\mathcal{D}} v := \sum_{\lambda \in \nabla_j^\Gamma} \langle v \circ \Gamma, \psi_\lambda \rangle \tilde{\psi}_\lambda \qquad (= Q_j^\Gamma B_{|V_j}^{\mathcal{D}}),$$

[2]This refers to the fact that, when considering a given problem, we will define only *one* of these two boundary operators.

or

$$B_j^{\mathcal{N}} v := \sum_{\lambda \in \nabla_j^\Gamma} \langle [(\nabla v) \circ \Gamma] \cdot \mathbf{n}, \psi_\lambda \rangle \tilde{\psi}_\lambda \qquad (= Q_j^\Gamma B_{|V_j}^{\mathcal{N}}).$$

To obtain a suitable discretization of C_Ω, some additional care is required. The obvious choice would be $\overline{C}_j : V_j^0 \to V_j^0$, $\overline{C}_j f_j = P_{V_j^0} C_\Omega f_j$, which written explicitly is given by

$$\overline{C}_j f = \sum_{i,k} \langle f_j \chi_\Omega, \phi_{jk}^i \rangle \phi_{jk}^i \qquad (= P_j C_{|V_j^0}). \tag{4.19}$$

This form has a few serious drawbacks. For one, the coefficients $\langle f_j \chi_\Omega, \phi_{jk}^i \rangle$ are, as a consequence of the non-trivial geometry of Ω, expensive to obtain, and expensive to compute accurately. But this has serious consequences, as the rank of $\overline{C}_j$ may change as the result of small errors in the computation of these coefficients, affecting the rank of the overall problem, which in turn can distort the solution in an unpredictable way. See [30], pages 335-338, for a thorough discussion.

Another possibility is to consider

$$C_j f = \sum_{i,k} \delta_{j,k,\Omega} \langle f_j, \phi_{jk}^i \rangle \phi_{jk}^i, \tag{4.20}$$

where $\delta_{j,k,\Omega}$ is given by

$$\delta_{j,k,\Omega} = \begin{cases} 1 & \text{if } \square_{jk} \cap \Omega \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

This amounts to the orthogonal projection onto V_j^0 of the restriction of $f \in V_j^0$ to

$$\Omega_j := \bigcup_{\square_{jk} \cap \Omega \neq \emptyset} \square_{jk}. \tag{4.21}$$

When writing the matrix of this map with respect to the basis $\{\phi_{jk}^i : i = 0, 1, \ldots, n,\ k \in \mathcal{Z}_j^2\}$, we obtain a section of the identity, thus reducing the possibility of numerical errors. We cannot eliminate it completely, as the computation of $\delta_{j,k,\Omega}$ itself is still subject to inaccuracies. In any case, it is much more efficient to compute, and as the numerical experiments this far sugest, it is also good enough.

Now, we define the map $M_j : V_j \to V_j^0 \times \tilde{V}_j^\Gamma$ through

$$M_j = \begin{pmatrix} C_j A_j \\ B_j \end{pmatrix},$$

where B_j is the discretized Dirichlet or Neumann boundary operator, as needed.

Let $\Xi_j = \{\psi_\lambda : \lambda \in \nabla_j\}$ be the wavelet basis for V_j, and let $\Xi_j^r = \{\phi_{jk}^i : i = 0, 1, \ldots, n,\ k \in \mathcal{Z}_j^2\} \times \{\tilde{\psi}_\lambda^\Gamma : \lambda \in \nabla_j^\Gamma\}$ be the basis for $V_j^0 \times \tilde{V}_j^\Gamma$. Let $\underline{M}_j$ be the matrix of M_j with respect to Ξ_j, Ξ_j^r, and let $f_j = P_j f^+$, $g_j = Q_j^\Gamma g$. Writing $b_j = (f_j, g_j)^T \in V_j^0 \times \tilde{V}_j^\Gamma$, $u_j \in V_j$, and $\underline{b}_j = (\underline{f}_j, \underline{g}_j)^T$, we have as a consequence of our choice of norms and spaces that

$$\left\| \underline{M}_j \underline{u}_j - \underline{b}_j \right\|_2^2 = \| M_j u_j - b_j \|_{\mathcal{H}^r}^2. \tag{4.22}$$

Thus, to find the minimizer of the quantity on the right, we compute the minimizer of the quantity on the left, which is now a simple linear least squares problem in Euclidean space.

4.4.4 Sparseness

To find a minimizer of (4.22) it would be quite helpful, for performance reasons, if given $v \in V_j$, we could evaluate $\underline{M}_j \underline{v}$ in $\mathcal{O}(\dim V_j)$ operations. The matrix $\underline{M}_j$, however, is not sparse. It is quasi-sparse, since the matrices $\underline{A}_j, \underline{B}_j$ have $\mathcal{O}(\log \dim V_j)$ entries per column, with N the number of degrees of freedom. This can be solved by factorizing these blocks using the wavelet transform; see [11], page 122.

Let $v \in V_j$. Let us write $\overline{v}$ for the coefficients of v with respect to the scaling function basis for V_j. The map $T_j : \ell_2(\nabla_j) \to \ell_2(\mathcal{Z}_j^2)$, $T_j : \underline{v} \to \overline{v}$ is simply the fast wavelet transform, and its numerical evaluation costs $\mathcal{O}(\dim V_j)$ operations. One easily sees that if $\underline{A}_j^0$ is the matrix of A_j with respect to the scaling function basis in V_j and the basis chosen for V_j^0, then $\underline{A}_j^0$ is sparse, and thus evaluating

$$\underline{A}_j \underline{v} = \underline{A}_j^0 T_j \underline{v}$$

using the factorization on the right (applying first T_j, and then $\underline{A}_j^0$) costs also $\mathcal{O}(\dim V_j)$ operations.

Similarly, let $\tilde{T}_j^\Gamma, T_j^\Gamma : \ell(\nabla_j^\Gamma) \to \ell(\mathcal{Z}_j)$ be the fast wavelet transforms $\underline{g} \to \overline{\underline{g}}$ for $g \in \tilde{V}_j^\Gamma$, $\underline{h} \to \overline{\underline{h}}$ for $h \in V_j^\Gamma$, respectively, and let $\underline{B}_j^0$ be the matrix of B_j with respect to the scaling function bases of V_j and $\tilde{V}_j^\Gamma$. Then evaluating

$$\underline{B}_j \underline{v} = (\tilde{T}_j^\Gamma)^{-1} \underline{B}_j^0 T_j \underline{v} \quad = (T_j^\Gamma)^T \underline{B}_j^0 T_j \underline{v} \tag{4.23}$$

using the factorizations on the right also costs only $\mathcal{O}(\dim V_j)$ operations. As a consequence, we obtain that through this factorization we can evaluate

$$\underline{M}_j \underline{v} = \begin{pmatrix} I & 0 \\ 0 & (T_j^\Gamma)^T \end{pmatrix} \begin{pmatrix} \underline{C}_j \underline{A}_j^0 \\ \underline{B}_j^0 \end{pmatrix} T_j \underline{v}$$

in $\mathcal{O}(\dim V_j)$ operations.

4.5 Realizing the iteration

The obtain a minimizer of

$$\Phi^*(\underline{v}_j) = \left\| \underline{M}_j \underline{v}_j - \underline{b}_j \right\|_2^2$$

we can use, for example, the conjugate gradients (CG) algorithm[26] to solve the normal equations,

$$\underline{M}_j^T \underline{M}_j \underline{v}_j = \underline{M}_j^T \underline{b}_j. \tag{4.24}$$

While this has well known disadvantages, it also has an important advantage, which is that it can give us the projection of v_{j-1} onto $\mathcal{N}(\underline{M}_j)$, needed to realize (4.11) essentially for free.

The key to that insight is obtained by taking a look at what the CG algorithm does. To find an approximate solution of the finite dimensional linear equation $Ax = d$, the CG method produces iterates x^i which are the minimizer in $W_i = x^{(0)} + \operatorname{span}\{r^0, r^{(1)}, \ldots, r^{(i-1)}\}$

of the functional $\gamma_i(y) = (y - x^*)^T A(y - x^*)$, where x^* is the exact solution of $Ax = d$, $x^{(0)}$ is some initial guess, and $r^{(k)} = A^k d$. The minimizer of γ_i in W_i exists, and is unique, only if A is symmetric positive definite on W_i. One has that $x^{(i)} = x^*$ when $W_i = W_{i+1}$ (if the algorithm is performed with exact arithmetic), but if the condition number of A is reasonable, then the $x^{(i)}$ will be a good approximation of x^* far earlier.

Suppose now that A is symmetric and positive semidefinite. If $d \perp \mathcal{N}(A)$, then $r^k \perp \mathcal{N}(A)$ for all k, and thus A is symmetric positive definite on

$$\begin{aligned} W_i &= x^{(0)} + \operatorname{span}\{r^0, r^{(1)}, \ldots, r^{(i-1)}\} \\ &= P_{\mathcal{N}(A)} x^{(0)} + P_{\mathcal{N}(A)^\perp} x^{(0)} + \operatorname{span}\{r^0, r^{(1)}, \ldots, r^{(i-1)}\} \end{aligned}$$

for all i [25]. Given an initial guess $x^{(0)}$, we will obtain at the i-th step an $x^{(i)}$ such that $P_{\mathcal{N}(A)^\perp} x^{(i)}$ is an approximation of x^*, but which also satisfies $P_{\mathcal{N}(A)} x^{(i)} = P_{\mathcal{N}(A)} x^{(0)}$. Since $\underline{M}_j^T \underline{b}_j \perp \mathcal{N}(\underline{M}_j^T \underline{M}_j)$, and since $\mathcal{N}(\underline{M}_j^T \underline{M}_j) = \mathcal{N}(\underline{M}_j)$, we can compute (see (4.11))

$$\underline{u}_{j+1} = P_{\mathcal{N}(\underline{M}_{j+1})} \underline{u}_j + \underline{M}_{j+1}^\dagger \underline{b}_{j+1}$$

by solving (4.24) with the conjugate gradient method using u_j as an initial guess.

Now write

$$\mathrm{CG}(A, d, x_0, \epsilon)$$

for the approximate solution of $Ax = d$, with $x^{(0)}$ as an initial guess, obtained by iterating until the error is smaller than ϵ. Then the numerical realization of (4.10), (4.11) is given by

$$\mathrm{SPFD}(j_0, j, \{b_j\}, \epsilon) := \tag{4.25}$$
$$\begin{cases} 0 & \text{if } j < j_0 \\ \mathrm{CG}(\underline{M}_j^T \underline{M}_j,\ \underline{M}_j^T \underline{b}_j,\ \mathrm{SPFD}(j_0, j-1, \{b_j\}, \epsilon),\ \epsilon) & \text{otherwise.} \end{cases}$$

Computing an approximation to $S_J b$ amounts to evaluate $\mathrm{SPFD}(j_0, J, \{b_j\}, \epsilon)$.

The question arises as to what effect the inexact evaluation of $\underline{M}_j^\dagger \underline{b}_j$ has on the sequence $\{S_j b\}$. In the experiments we have performed, it does not seem to play an important role; further research is needed to shed light on this issue.

Instead of using standard CG with the normal equations, one should use the mathematically equivalent but numerically superior CGLS, developed in [25]. The direct application of other Krylov subspace least-squares solvers is a delicate matter. In the case of LSQR[32], a very robust least squares solver, the problem is to implement the projections onto the kernel. Still other methods, like RRGMRES [6], assume that the system is given through a square matrix. Again, we see in further research an opportunity for improvements in performance of the method described in this chapter.

Note that if (4.9) holds, the condition number of the least-squares problems stays bounded with j, and thus, in theory, no further preconditioning is needed. We would have

$$\kappa(M_j) = \|M_j\| \|M_j^\dagger\| \leq C_M^2,$$

and if we do not avoid the normal equations, we would end up with

$$\kappa(M_j^T M_j) \leq C_M^4.$$

Chapter 5

Numerical experiments

The previous two chapters have made theoretical predictions which we would like to observe in practice. The most important reason is that we have made asymptotic predictions, and would like to know whether they are observable, and thus whether they have any relevance in practice. This is comparatively more important for the SPFD method introduced in chapter four, as it makes some strong promises, and since open questions remain, than for the results of chapter three on the smoothness of solution of the FDLM method, which concern a know method, and which are theoretically conclusive.

It is still worthwhile to check numerically the effect on smoothness of a non-zero Lagrange Multiplier. From the proofs of theorems 3.2.7, and 3.2.9 (More accurately, from the proof of lemma 3.2.8), we might be left with the impression that the convergence rate predicted can be observed only for extremely high resolutions, beyond the reach of most practical needs. These are the kinds of questions we wish to answer.

5.1 The experiments

5.1.1 Goals of the experiments

We will test both methods against a few simple examples and examine the results with the following goals.

1. Concerning the FDLM method

   (a) Observing experimentally the phenomenon predicted by theorem 3.2.7 on the convergence of linear approximation schemes.

   (b) Observing the phenomenon predicted by theorem 3.2.9, on the convergence of nonlinear approximation schemes.

2. Concerning the SPFD method

   (a) Measuring the smoothness of the solution obtained, rated through the convergence speed of linear approximation using B-splines.

(b) Observe the effects of the nested iteration on the solution. Does it really make a difference?

(c) Establish whether the method can take advantage of the approximation power of higher order B-splines.

(d) Observe the behavior of the method when faced with Neumann boundary conditions.

5.1.2 Test cases

Given $0 < r < 1/2$, we choose as a domain a simple disc

$$\Omega_r = \{x \in \mathbb{T}^2 : \|x - (0.5, 0.5)\| < r\},$$

and parametrize the boundary through $\Gamma : \mathbb{T} \to \partial\Omega$, given by

$$\Gamma(t) = (0.5, 0.5) + r(\sin(2\pi t), \cos(2\pi t)). \tag{5.1}$$

Our choice for r will be limited to $r = 0.3$, except once where we will use $r = 0.45$ to be able to better measure the convergence of nonlinear approximation schemes. As always, we embed Ω into $\mathbb{T}^2$.

We will investigate the behavior of the methods in question on the following test problems.

Problem **P1***:* Find u such that

$$\begin{aligned} (-\triangle + I)u &= 1 \qquad \text{on } \Omega, \\ B^{\mathcal{D}}u &= 0, \end{aligned}$$

with $r = 0.3$ (and only once with $r = 0.45$). We choose as the extension to $\mathbb{T}^2$ the obvious one, $f_I^+ = 1$.

The above data can be considered too canonic. Thus, we also solve the following problem, using nontrivial data.

Problem **P2** Find u such that

$$\begin{aligned} (-\triangle + I)u &= f_{II} \qquad \text{on } \Omega, \\ B^{\mathcal{D}}u &= g_{II} \end{aligned}$$

with $f_{II} = 1 + \frac{1}{2}\cos(5(x^2 + y^2))$, $g_{II} = 0.01 \cdot \sin(4\pi t)$, $r = 0.3$.

To use any of the fictitious domain methods above, we must construct an extension of f_{II} to $\mathbb{T}^d$. We could just choose the function $f(x, y) = 1 + \frac{1}{2}\cos(5(x^2 + y^2))$ on $[0, 1)^2$ as an extension of the above right-hand side, and then lift it to $\mathbb{T}^2$ by pretending f is periodic, but this has the drawback that we do not obtain a smooth function on $\mathbb{T}^d$. To find an extension for f_{II} from Ω to $[0, 1)^2$ that is smooth, and can be lifted smoothly from $[0, 1)^2$ to $\mathbb{T}^2$, we will construct an infinitely often differentiable function $\Upsilon : [0, 1]^2 \to \mathbb{R}$ which, together with all its derivatives, is zero on $\partial([0, 1]^2)$, and which is 1 on Ω. Then, we take $f_{II}^+(x, y) := \Upsilon(x, y)f(x, y)$, restrict it to $[0, 1)^2$, and finally we lift it to $\mathbb{T}^2$.

For the domain Ω_r with $r = 0.3$, a suitable function Υ can be obtained through a tensor product with itself of a one-dimensional C^∞ function $\Upsilon_0 : [0, 1] \to \mathbb{R}$ which, together with

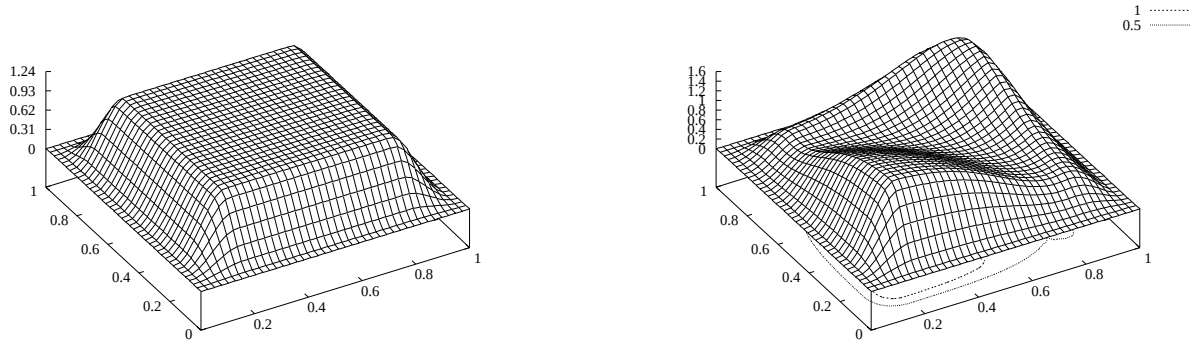
all its derivatives, is zero on $0, 1$, and is 1 on $[0.2, 0.8]$. To construct Υ_0, we will consider the standard mollifier

$$\phi_{\epsilon,y}(x) = \begin{cases} \exp\left(-\frac{\epsilon^2}{\epsilon^2 - |x-y|^2}\right) & \text{if } \|x-y\|_2 < \epsilon, \\ 0 & \text{otherwise,} \end{cases}$$

and engineer it to suit our purposes, as follows. First, we take $\Upsilon_{-2}(x) = \phi_{0.1,0.1}(x) - \phi_{0.1,0.9}(x)$. Then, we define $\Upsilon_{-1}(x) := \int_0^x \Upsilon_{-2}(y)dy$, and obtain

$$\Upsilon_0(x) = \frac{\Upsilon_{-1}(x)}{\Upsilon_{-1}(\frac{1}{2})}.$$

Now, we set $\Upsilon(x, y) = \Upsilon_0(x)\Upsilon_0(y)$ (see figure 5.1 for plots of Υ and f_{II}). In the implementation, we used a standard adaptive quadrature routine to evaluate Υ_0 at any point x.



(a) Plot of Υ.

(b) Plot of f_{II}^+ (and f_{III}^+).

Figure 5.1: *Construction of the right-hand side for problems P2 and P3.*

The next problem uses the same data as problem P2, but this time we impose Neumann boundary conditions.

Problem **P3** Find u such that

$$\begin{aligned} (-\triangle + I)u &= f_{III}, \\ B^{\mathcal{N}}u &= g_{III}, \end{aligned}$$

with $f_{II} = f_{III}$, and $g_{III} = g_{II}$, and $r = 0.3$. As the extension to $\mathbb{T}^d$ of the right hand side we use exactly the same as before, and so have $f_{III}^+ = f_{II}^+$.

5.2 Remarks on the implementation of the solvers

All the techniques used to implement the components of the solvers needed for the numerical experiments (fictitious domain - Lagrange multiplier method, and smoothness-preserving

fictitious domain method) are standard. We will briefly mention them by name but spare the reader of details which can be found in any elementary numerical analysis book. The implementation of the periodic wavelet transforms employed is also straightforward, and thus we do not discuss it here either.

We implemented the SPFD method faithfully as described in 4.4, with the following two differences. For one, we used higher precision (smaller ϵ) on lower levels, where iterations cost less; this has been documented in the iteration histories we will provide. The second difference is that we have given a higher weight to the norm on the boundary than in the discretization mentioned in 4.4. This ensures that boundary conditions where satisfied better on a lower level. Thus, instead of minimizing Φ as defined in 4.18, we minimized

$$\Phi(v) = \|C_\Omega A v - f_1\|^2_{H^0(\mathbb{T}^d)} + \rho \|Bv - g\|^2_{H^{\sigma(B)}(\partial\Omega)} \tag{5.2}$$

with $\rho = 70$. Again, see remark 4.2.4 for a justification.

The discretization of the differential operator $A = -\triangle + \mu I$ for the FDLM approach is different than that for the SPFD method described in subsection 4.4.3. For appropriate m, $\tilde{m}$ to be specified later, we consider the corresponding pair of (properly scaled) biorthogonal B-spline wavelet bases Ψ for $H^1(\mathbb{T}^2)$, $\tilde{\Psi}$ for $H^{-1}(\mathbb{T}^2)$, and the pair of biorthogonal MRAs $\{V_j\}$, $\{\tilde{V}_j\}$ of B-splines and duals, respectively, from where those bases arise. The discrete operators $\check{A}_j : V_j \to \tilde{V}_j$ are given by $\check{A}_j = \tilde{Q}_j A_{|V_j}$.

On the other hand, we have that the discretization of the Dirichlet boundary operator used for the FDLM method is almost identical to that used in the SPFD method. The only difference is in the scaling of the bases chosen, since the FDLM formulation considers $B^{\mathcal{D}} : H^1(\mathbb{T}^2) \to H^{1/2}(\partial\Omega)$, instead of $B^{\mathcal{D}} : H^2(\mathbb{T}^d) \to H^{3/2}(\partial\Omega)$. But just as before, we identify $H^{1/2}(\partial\Omega)$ with $H^{1/2}(\mathbb{T})$ via the parametrization (5.1), and instead of spanning $H^{1/2}(\mathbb{T})$ with the primal basis, we use for that purpose the (properly rescaled) *dual* basis $\tilde{\Psi}^\Gamma$, using Ψ^Γ to span $H^{-1/2}(\mathbb{T})$.

Given $f^+ \in H^{-1}(\mathbb{T}^2)$, $g \in H^{1/2}(\partial\Omega)$, we are looking for the coefficients $\underline{u}^+$, $\underline{p}$ with respect to the bases Ψ, Ψ^Γ of functions $u^+ \in H^1(\mathbb{T}^2)$, $p \in H^{-1/2}(\partial\Omega)$ such that

$$\begin{pmatrix} \check{A}_j & (B_j^{\mathcal{D}})^* \\ B_j^{\mathcal{D}} & 0 \end{pmatrix} \begin{pmatrix} u^+ \\ p \end{pmatrix} = \begin{pmatrix} f^+ \\ g \end{pmatrix},$$

or rather

$$\begin{pmatrix} \underline{\check{A}}_j & (\underline{B}_j^{\mathcal{D}})^T \\ \underline{B}_j^{\mathcal{D}} & 0 \end{pmatrix} \begin{pmatrix} \underline{u^+} \\ \underline{p} \end{pmatrix} = \begin{pmatrix} \underline{f^+} \\ \underline{g} \end{pmatrix}, \tag{5.3}$$

where the entries in the matrix $\underline{\check{A}}_j$ are given by

$$(\underline{\check{A}}_j)_{\lambda\nu} = \langle A\psi_\lambda, \psi_\nu \rangle$$

while the entries in $\underline{B}_j^{\mathcal{D}}$ are given by

$$(\underline{B}_j^{\mathcal{D}})_{\lambda v} = \langle B^{\mathcal{D}} \psi_\lambda, \psi_v^\Gamma \rangle.$$

We use the fast wavelet transform to factorize $\underline{\check{A}}_j$ in exactly the same way as done before in chapter four, subsection 4.4.4. We obtain

$$\begin{pmatrix} \underline{\check{A}}_j & (\underline{B}_j^{\mathcal{D}})^T \\ \underline{B}_j^{\mathcal{D}} & 0 \end{pmatrix} = \begin{pmatrix} \tilde{T}_j^{-1} & 0 \\ 0 & (\tilde{T}_j^{\Gamma})^{-1} \end{pmatrix} \begin{pmatrix} \underline{\check{A}}_j^0 & (\underline{B}_j^{\mathcal{D},0})^T \\ \underline{B}_j^{\mathcal{D},0} & 0 \end{pmatrix} \begin{pmatrix} T_j & 0 \\ 0 & T_j^{\Gamma} \end{pmatrix},$$

where $\underline{\check{A}^0}_j$ and $\underline{B}_j^{\mathcal{D},0}$ correspond to the representation of A and $B^{\mathcal{D}}$ in terms of scaling-functions; we will come back to this shortly.

We will use LSQR (and for comparision purposes, also CGLS) to solve the resulting system of equations (5.3).

Computing matrix coefficients

The only missing detail left is how to compute the matrix coefficients needed to set up the systems of linear equations we will solve. We shall do this here, first for the boundary operators, and then for the differential operators. The computation of the entries in $\underline{C}_j$ is straight-forward (see (4.20)), and thus we do not discuss it any further.

We explain in some detail the computation of the entries in the matrix $\underline{B}_j^0$ (see (4.23)) corresponding to the boundary operators first for the case of the Dirichlet boundary operator, and then apply the same approach to the computation of the entries corresponding to the Neumann boundary operator. Again, we always assume that the basis elements are properly scaled.

To compute

$$(\underline{B}_j^{\mathcal{D},0})_{kl} = \langle \phi_{jk} \circ \Gamma, \phi_{jl}^{\Gamma} \rangle = \int_{\mathbb{T}} [\phi_{jk} \circ \Gamma](t) \phi_{jl}^{\Gamma}(t) dt,$$

we first identify a set of pairwise disjoint open intervals $\{I_i\}$ in $\mathbb{T}$ such that, writing $\upsilon_{kl}(t) = [\phi_{jk} \circ \Gamma](t)\phi_{jl}^{\Gamma}(t)$, one has $\operatorname{supp} \upsilon_{kl} = \overline{\cup_i I_i}$, and such that υ_{kl} is C^∞ on each I_i. This obtain these intervals, it is enough to look at the intervals on which ϕ_{jl}^{Γ} is a polynomial, and intersect the cubes on which ϕ_{jk} is a polynomial with $\partial\Omega$. Finally, we compute

$$\int_{\mathbb{T}} [\phi_{jk} \circ \Gamma](t) \phi_{jl}^{\Gamma}(t) dt = \sum_i \int_{I_i} [\phi_{jk} \circ \Gamma](t) \phi_{jl}^{\Gamma}(t) dt$$

by approximating each of the integrals on the right via a high order Gauss Legendre quadrature rule. In the implementation used to perform these experiments we used one of order 10, which was deemed to be accurate enough.

To compute the entries in the matrix corresponding to the Neumann boundary operator, we simply repeated the above process, but replacing $\phi_{jk} \circ \Gamma$ with $\nabla \phi_{jk}(\Gamma(t))\mathbf{n}(t)\phi_{jl}^{\Gamma}(t)$.

To compute the entries in $\underline{\check{A}}_j^0$, given by

$$(\underline{\check{A}}_j^0)_{kl} = \langle A\phi_k, \phi_l \rangle = \int_{\mathbb{T}} \nabla \phi_k \nabla \phi_l dx,$$

we used the fact that the functions involved are piecewise polynomials, and thus we computed these entries using standard quadrature rules on each of the polynomial pieces.

The computation of the entries in the matrix $\underline{A}_j^0$ (needed for the SPFD method) were obtained by using simple quadrature rules to evaluate the inner products $(\underline{A}_j^0)_{ikl} = \langle A\phi_{jk}, \phi_{jl}^i \rangle$.

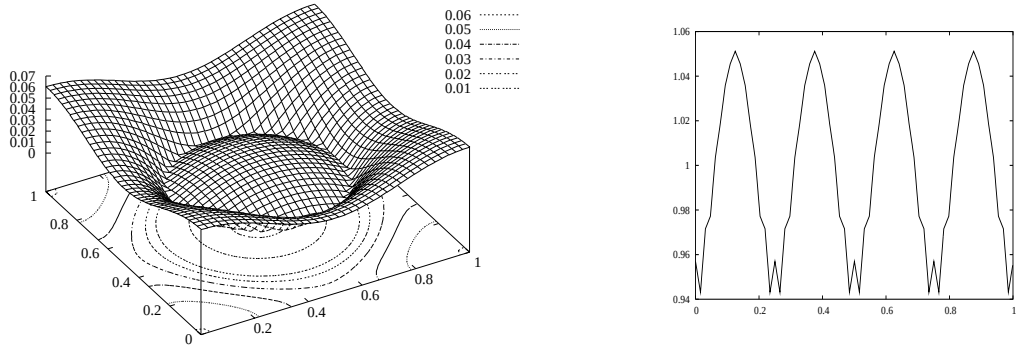
For both the SPFD and FDLM methods we have chosen $m^\Gamma = 2$, $\tilde{m}^\Gamma = 6$ for the primal and dual orders of the B-spline wavelet bases used for the boundary. For the B-spline wavelet bases occurring in the discretization of the domain, we have chosen $m = 3$, $\tilde{m} = 7$, unless otherwise stated.

5.3 Numerical results and discussion

5.3.1 Smoothness of the solutions obtained using the FDLM method

Behavior of the linear approximation error

We were able to observe the phenomenons predicted by theorems 3.2.7 for the fairly canonical problem P1, using a radius of $r = 0.3$ for Ω. We computed the solution u^+ of the FDLM with the corresponding data to level 8 on $\mathbb{T}^d$, and, following [12], we used level 6 on $\partial\Omega$ to satisfy the LBB condition and obtain better accuracy. We did let LSQR iterate until it arrived at a residual of norm smaller than 10^{-3}, which took 273 iterations[1]. A plot of the solution can be seen in figure 5.2(a), where it is also possible to appreciate optically the jump in the normal derivatives. A plot of the Lagrange multiplier can be seen in figure 5.2(b).



(a) Extended solution

(b) Lagrange multiplier

Figure 5.2: *Solution and Lagrange multiplier obtained when solving problem P1 with the FDLM method.*

We used the fast wavelet transform to obtain the wavelet coefficients of u_J^+ with respect to the basis Ψ, but this time scaled to be a basis of L_2. This gave us a representation of u_J^+ of the form

$$u_J^+ = \sum_{\lambda \in \triangle_9} c_\lambda \psi_\lambda$$

[1] the CGLS method needed 1421 iterations to reach the same accuracy, confirming its known drawbacks

with

$$\|u_J^+\|_{L_2(\mathbb{T}^d)} \sim \left(\sum_{\lambda \in \triangle_9} |c_\lambda|^2 \right)^{\frac{1}{2}}.$$

Figure 5.3 plots the errors of linear approximation in the norm induced by Ψ. That is, the quantities

$$E_j^\Psi(u_J^+) = \left(\sum_{\lambda \in \nabla_8 : \psi_\lambda \notin V_j} |c_\lambda| \right)^{\frac{1}{2}},$$

which are uniformly equivalent to the errors,

$$E_j(u_J^+) = \inf_{v \in V_j} \|u_J^+ - v\|_{L_2(\mathbb{T}^d)}$$

but easier to obtain.

Remark 5.3.1. *The phenomenon observed in figure 5.3 is the convergence rate of the linear approximation scheme when applied to the obtained solution. The error plotted should* not *be understood as the distance to the exact solution.*

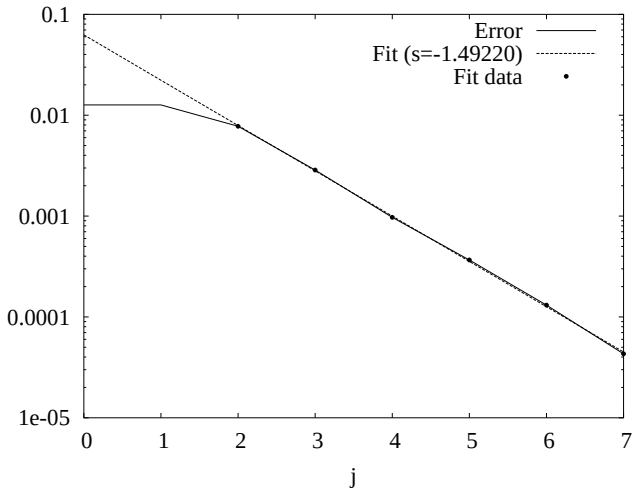


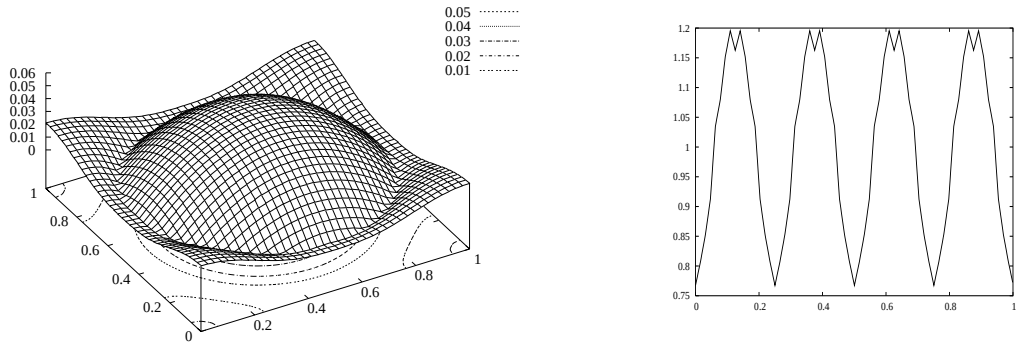
Figure 5.3: *Linear approximation errors when approximating the solution to problem P1 obtained with the FDLM method.*

After some initial irregularity, we observed the expected asymptotic behavior. To measure it, we chose a range of j where the error seemed to behave as predicted, and fitted to it the function $\eta(j) = C2^{js}$, using linear least squares in the coordinates of the plot. This gave us an estimate of the order of convergence s. We plotted the obtained η (dotted line in figure 5.3), along with marks for the data used in the fit.

Behavior of the nonlinear approximation error

To investigate the behavior of the nonlinear approximation error, it was found to be advantageous to use a larger radius (we used $r = 0.45$ for Ω). This is due to the fact that then there are more wavelet coefficients on $\mathbb{T}^2$ that intersect the boundary than if the radius is smaller.

We computed the solution u^+ of the FDLM with the corresponding data to level 8 on $\mathbb{T}^d$, and level 6 on $\partial\Omega$. We solved again the system of linear equations using LSQR with a tolerance of 10^{-3}. This time it needed 919 iterations[2]. A plot of the solution can be seen in figure 5.4, alongside the obtained Lagrange multiplier.



(a) Solution u^+ (b) Lagrange multiplier p

Figure 5.4: *Solution and Lagrange multiplier obtained when solving problem P1 with the FDLM method, this time with* $r = 0.45$.

To analyze the solution u_J^+ we used again the fast wavelet transform, this time to obtain the wavelet coefficients of u_J^+ with respect to the basis Ψ, scaled to be a basis of $H^1(\mathbb{T}^d)$. This gave us a representation of u_J^+ of the form

$$u_J^+ = \sum_{\lambda \in \nabla_8} b_\lambda \psi_\lambda$$

with

$$\|u_J^+\|_{H^1(\mathbb{T}^d)} \sim \left(\sum_{\lambda \in \triangle_9} |b_\lambda|^2 \right)^{\frac{1}{2}}.$$

Next, we sorted the 2^{16} coefficients in decreasing order of their absolute values, producing the vector of real numbers $a = (a_0, a_1, \ldots, a_{2^{16}-1})$. Thus, we still have

$$\|u_J^+\|_{H^1(\mathbb{T}^d)} \quad \left(\sum_{i=0}^{2^{16}-1} a_i^2 \right)^{\frac{1}{2}},$$

[2]In comparision, CGLS needed 1372 iterations.

while also obtaining the error of the best N-term approximation to u_J^+ from

$$E_N^{\Psi}(u_J^+) = \left(\sum_{i=N}^{2^{16}-1} a_i^2 \right)^{\frac{1}{2}}.$$

We subjected $\omega = B_J^* p$ to a similar treatment; that is, we computed the wavelet coefficients of ω with respect to the dual basis $\tilde{\Psi}$ of Ψ, which is a basis for $H^{-1}(\mathbb{T}^d)$, and proceeding analogously to how we proceeded with u_J^+.

We have plotted the convergence history of the best N-term approximation in doubly logarithmic scale, and as done in the linear approximation case, we have plotted it together with the fitted (in doubly logarithmic coordinates) $\mu(x) = CN^{-s}$ and the data points used in the fit (chosen where we believe one can observe the asymptotic behavior expected). We have done this both for u_J^+ and ω; see figure 5.5.

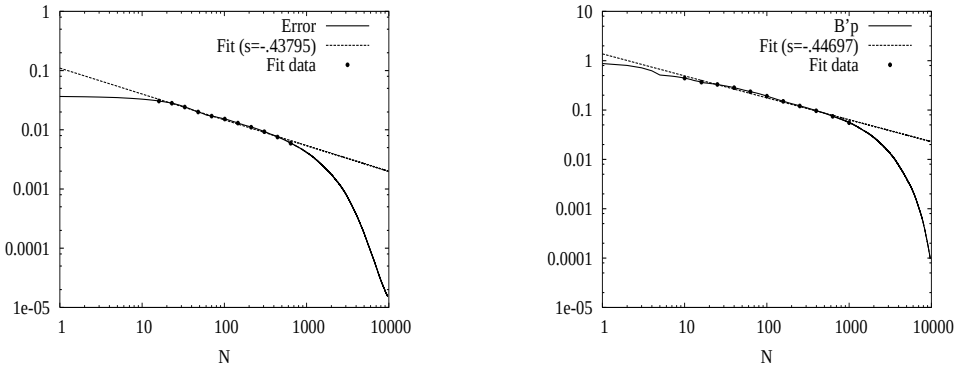


Figure 5.5: *Convergence histories of best N-term approximation to u_J^+ and $B^T p$. Idealized convergence rates have been fitted to measure actual convergence rates.*

A mixed picture emerges, which is not entirely unexpected. We are plotting the best $N-$term approximation errors with respect to the solution u_J^+ and not with respect to the solution of the infinite dimensional problem, which remains beyond our reach. The sequence of wavelet coefficients of u_J^+ is compactly supported, and thus belongs to any ℓ_τ^w. Eventually (in both figures from $N \approx 500$ onwards), the decay of the error must accelerate, as the best N-term approximation of u_J^+ is exact for $N = 2^{16}$.

Note that the acceleration is due to the exhaustion of the degrees of freedom corresponding to wavelets whose supports intersect the boundary. After around $N = 1500$, the singularity at the boundary, as reflected in the solution analized, was fully resolved. From then on, the convergence rate is due to the smoothness of the solution away from the boundary. One should not misunderstand neither the theoretical results of chapter three, nor the numerical evidence presented here. While asymptotically the convergence rate of the nonlinear approximation scheme is limited, it still yields greater accuracy with far fewer degrees of freedom than the *linear* approximation schemes.

j	Tolerance	Iterations	Initial residual
3	1.0000e-05	11	7.0711e-01
4	2.5119e-05	0	6.7104e-11
5	6.3096e-05	0	6.6714e-11
6	1.5849e-04	0	6.4922e-11
7	3.9811e-04	0	6.4058e-11
8	1.0000e-03	0	6.3563e-11

Table 5.1: *Iteration history for the SPFD method applied to problem P1*

5.3.2 Behavior of the SPFD method

To test the SPFD method, we chose the smaller radius of $r = 0.3$, which allows us to appreciate better the smooth extension of the solution. The recursion (4.25) was evaluated with $j_0 = 3$, and $J = 8$, but choosing higher precision for smaller j (where iterations are cheaper) than for higher j. We summarize the iteration history for problem P1 in table 5.1. The column labeled "initial residual" lists the errors

$$\|\underline{M}_j^T(\underline{M}_j x_j^0 - \underline{b}_j)\|,$$

where x_j^0 is the initial guess obtained from the result of the previous level (or zero, if there was no previous level). The level chosen for the discretization on the boundary was always the same as for the domain.

In this particular case we observe the promise of the SPFD method materialize in a dramatic way. Observe that the solution found for $j = 3$ was already good enough to satisfy the expected accuracy even on level 8, needing no further iterations. Find a plot of the solution in figure 5.6(a). We have also plotted the boundary values of the solution obtained in figure 5.6(b).

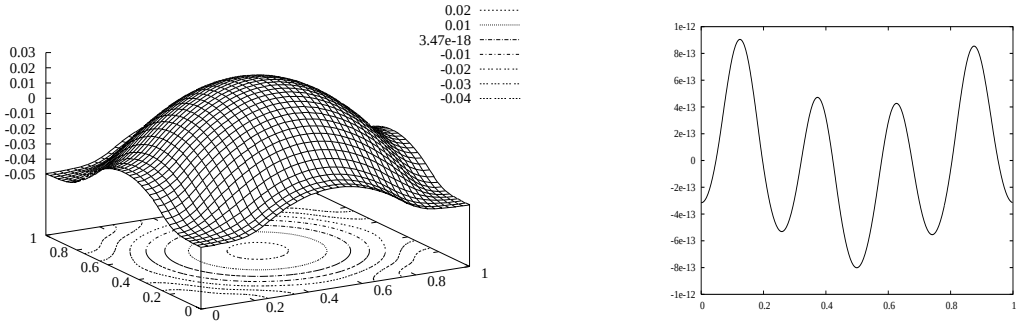
We find this experiment quite remarkable. It shows that the SPFD can indeed find very smooth solutions if that is possible. In this case, the solution on the domain is polynomial; one easily checks that the solution of the original problem is

$$u = 0.25\left(r^2 - (x - 0.5)^2 - (y - 0.5)^2\right).$$

The SPFD method is actually able to find in V_3 an *exact* extension of u to $\mathbb{T}^2$!

To test the SPFD method against more realistic data, we solved next problem P2. We have summarized the iteration history in table 5.2, and show the solution v_J^+ in figure 5.7. Using the same procedure as for the solution of the FDLM method above, we plot the linear approximation error, together with the fitted idealized convergence rate (see figure 5.7).

Since we are using piecewise quadratic C^1 functions with meshsize $h = 2^{-j}$, and since the extended right-hand side is C^∞, we expect a convergence rate of at least 2^{-3j}. The measured convergence rate is 2^{sj}, with $s \approx -3.65$, showing again that the method is able to find very good extensions for the solution.



(a) Extended solution

(b) Boundary values

Figure 5.6: *Solution and boundary values of the solution at $\partial\Omega$ obtained when solving problem P1 with the SPFD method (note the order of magnitude on the y-axis of figure 5.6(b)).*

j	Tolerance	Iterations	Initial residual
3	1.0000e-05	75	6.3801e-01
4	2.5119e-05	112	3.0223e-02
5	6.3096e-05	167	1.4061e-02
6	1.5849e-04	237	7.7306e-03
7	3.9811e-04	215	4.0875e-03
8	1.0000e-03	7	2.1018e-03

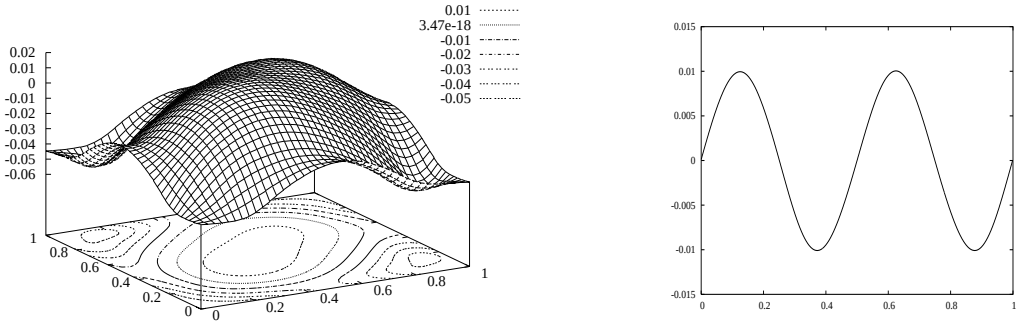
Table 5.2: *Iteration history for the SPFD method applied to problem P2*

The effect of the SPFD iteration

The next item on our checklist is to see whether we can observe the effects of the nested iteration scheme (4.11) on the solution obtained. Optically, at least, it is quite easy to spot. Contrast figure 5.7 with figure 5.9(a), where we show the solution of problem P2 on level $J = 8$ without using nested iteration. That is, we solved

$$\|\underline{M}_8 w_8^+ - \underline{b}_8\|_{\ell_2} \to \min!$$

with CGLS until the residual was smaller than 10^{-3}, which took 476 iterations. Observe also the linear approximation histories for both solutions, as seen in figures 5.8 and 5.9(b). We conclude that while the nested iteration definately drives the construction of a smooth solution, the basic SPFD formulation by itself (4.5) is quite capable of delivering better smoothness than the FDLM method.



(a) Extended solution

(b) Boundary values

Figure 5.7: *Solution and boundary values of the solution at $\partial\Omega$ obtained when solving problem P2 with the SPFD method.*

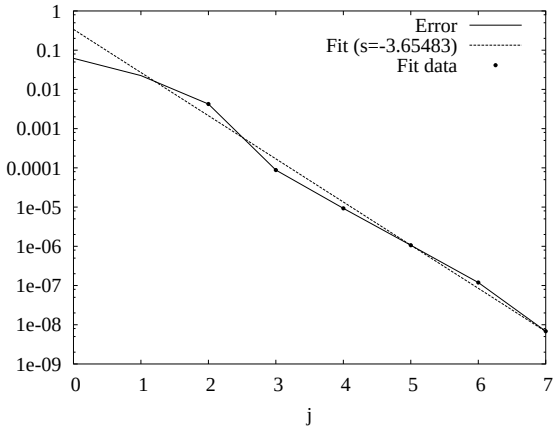


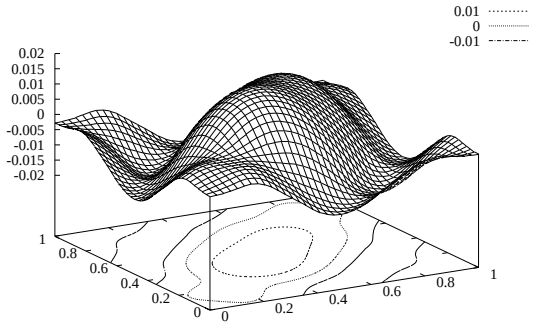
Figure 5.8: *Linear approximation error and fitted idealized convergence rate for v_J^+.*

Higher order

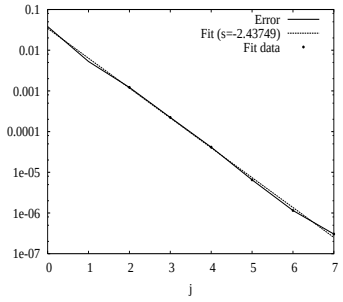
We chose $m = 5$, $\tilde{m} = 9$, and solved again problem P2. The convergence history is summarized in table 5.3, the solution can be seen in figure 5.10. We observe, as done with example I, that the solution at a lower level is good enough to satisfy the equations at a higher level to the required accuracy. The decay of the linear approximation errors is far too fast to be of any use rating the convergence.

The Neumann problem

Finally, we try out the SPFD method with the Neumann problem (problem P3). For the solution, see figure 5.11(a), while the values of the outward normal derivative at the boundary can be appreciated in figure 5.11(b). We have summarized the iteration history in table 5.4.



(a) Extended solution

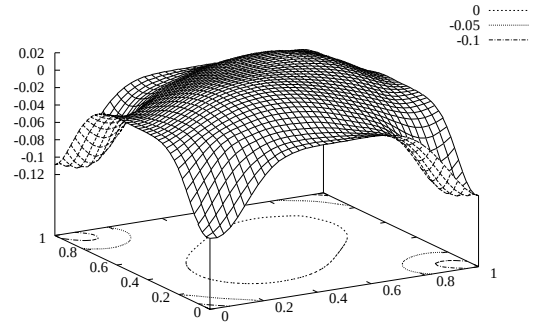


(b) Linear approximation error

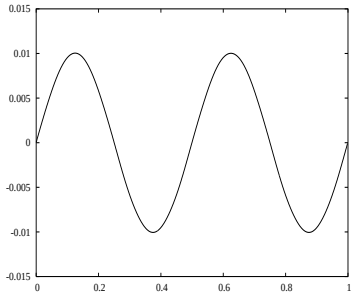
Figure 5.9: *Solution and boundary values of the solution at $\partial\Omega$ obtained when solving problem P2 with the SPFD formulation but without nested iteration.*

j	Tolerance	Iterations	Initial residual
3	1.0000e-05	225	6.3607e-01
4	2.5119e-05	858	2.0985e-02
5	6.3096e-05	926	3.9566e-04
6	1.5849e-04	0	1.0457e-04
7	3.9811e-04	0	1.0026e-04
8	1.0000e-03	0	9.6757e-05

Table 5.3: *Iteration history for the SPFD method applied to problem P2 (using higher order B-splines)*

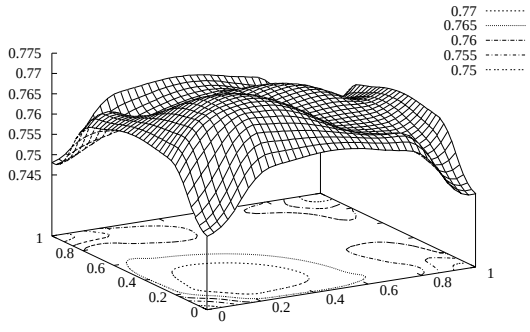


(a) Extended solution

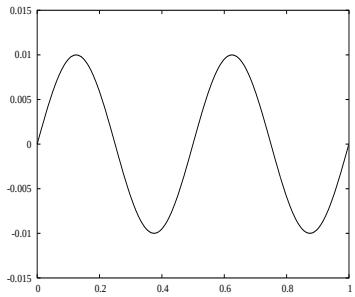


(b) Boundary values

Figure 5.10: *Solution and boundary values of the solution at $\partial\Omega$ obtained when solving problem P2 with the SPFD formulation with nested iteration, using B-splines of order* 5.



(a) Extended solution



(b) Outward normal derivative

Figure 5.11: *Solution and boundary values of the solution at $\partial\Omega$ obtained when solving problem P3 with the SPFD formulation with nested iteration.*

j	Tolerance	Iterations	Initial residual
3	1.0000e-05	118	7.4651e-01
4	2.5119e-05	204	1.1609e-01
5	6.3096e-05	282	1.9109e-02
6	1.5849e-04	271	8.6360e-03
7	3.9811e-04	201	4.1898e-03
8	1.0000e-03	14	2.0700e-03

Table 5.4: *Iteration history for the SPFD method applied to problem P3 (Neumann boundary conditions)*

Chapter 6

Final notes

6.1 Conclusions

What follows is a brief summary of the main achievements and results of this thesis.

- We generalized and complemented some results from the literature [21], and have found that the solutions obtained using the FDLM approach do, in general, suffer from a lack of regularity (see theorem 3.2.7). Whenever the Lagrange multiplier is non-zero, and if the right-hand side is in $H^{-1/2+\epsilon}$ for some $\epsilon > 0$, then the solution obtained is at best in $H^{3/2}$.

  This lack of regularity implies that the performance of linear approximation schemes (that is, in essence, approximation from uniform grids) is limited.

  In particular, it was found that if the Lagrange multiplier is non-zero, then for B-spline approximation from uniform meshes the error in the L_2 norm decays at best as $2^{-1.5j}$, where j indicates the level of resolution (that is, the meshsize is given through $h = 2^{-j}$). This behavior occurs independently of the order of the used B-spline bases.

  We were able to observe this behavior in numerical experiments.

- A similar result was obtained for standard nonlinear approximation schemes (isotropic adaptive schemes). We studied best N-term approximation using wavelet bases, and found that if the Lagrange multiplier could be identified with a measurable non-zero function on an interval, then best N-term approximation using B-spline wavelet bases converged at best as $N^{-\frac{1}{2(d-1)}}$. Again, this behavior is independent of the orders of the wavelets used.

  The above behavior was also confirmed by numerical experiments.

- A new fictitious domain method (the smoothness preserving fictitious domain method, or SPFD method) was proposed that is designed to overcome these limitations. The method constructs a smooth solution through the constructive use of fundamental principles of approximation theory.

  It was established that the solutions obtained via the SPFD method are solutions to the original elliptic boundary value problems. That is, the method is sound.

Theoretical evidence could be supplied that showed that, under certain conditions on the discretization, the solution obtained also has optimal smoothness.

A discretization scheme was introduced which promises to satisfy these requirements. Numerical experiments were provided that seem to confirm that the solutions obtained do indeed have optimal smoothness. This was evaluated by measuring the convergence rate of B-spline approximation from fixed grids, and comparing that rate with the rate predicted by standard approximation results.

- Numerical experiments with the SPFD method found that the measured approximation order was higher than the lower bounds predicted by theory.

  The numerical and theoretical results are very encouraging and suggest that the SPFD method is worth of further study.

6.2 Outlook

A lot remains to be done. In particular, we feel that the following tasks are promising routes of further research.

Analyze other linear solvers

The CGLS method is not very good. This has been known for a long time, and we were able to confirm it here, taking a look at the number of iterations needed to solve problem P1 with the fictitious domain - Lagrange multiplier approach.

However, any alternative should preserve the component in the kernel of the SPFD operator M_j to be, from a theoretic point of view, a good candidate.

Fill in the gaps of the theory

The global convergence and smoothness of the limit of the SPFD method holds, according to the provided theory, if the discrete operators satisfy assumptions A1, A2, and A3. The question is, does the sequence of operators designed in 4.4 satisfy these assumptions? We believe that it does. But if not, do such sequences of operators exist at all?

Another possibility is to explore whether requirements A1, A2, and A3 can be substituted by other requirements, that are either easier to check or easier to satisfy. We believe that there is a lot of space for variations in this formulation.

Use of other approximation spaces

For the analysis, as well as for the numerical experiments, we have used periodic splines on dyadic grids. While this choice guarantees us a lot of simplicity and approximation power, it is certainly not the only possibility.

For numerical purposes, it would be interesting to test the method with more general spline and finite-element spaces, for instance.

More general formulation

Another limitation of the SPFD method is that, due to its current formulation, it cannot deal with problems on domains that contain corners. Thus, changing the formulation to accomodate for this case is perhaps one of the most urgent directions of research that should be followed.

Adaptivity

The SPFD method as it was constructed here is not adaptive, and it is not immediately clear how to construct an adaptive strategy that still realizes the smoothness preserving behavior. It has to be noted that the point of view that has allowed us to construct and analyze this method is not too distant from the points of view taken in [8] and [7], making those articles a canonical starting point.

An adaptive SPFD solver would be a very powerful tool for dealing with problems that involve complex domains and singularities.

General elliptic boundary value problems

It is not too difficult to "upgrade" the proofs in chapter four to problems where the differential operator has higher order, and to more general boundary operators. A more interesting route of exploration are problems where different types of boundary conditions hold on different parts of the boundary.

Another interesting possibility is to try to apply the SPFD approach to other problems, as Stokes and Navier-Stokes problems.

6.3 About the software

The programs where written in Common Lisp, a modern, object oriented, ANSI standardized dialect of the second oldest programming language still in use (the oldest is Fortran). It was initially developed by John McCarthy in [29], and used mainly in the artificial intelligence community. Later it became the general purpose language it is today. Many features of the language work together to improve the productivity of the programmer at several levels.

- **Syntax:** The syntax is very regular and simple. Expressions have the form

  $$(\langle operator \rangle \ \{\langle argument \rangle\})$$

  where each of the arguments is either atomic (number, vector, symbol, etc), or another expression. A mathematical expression like $\sin(\alpha s) + Ce^x$ would be written in lisp as

```
(+ (sin (* alpha s))
   (* C (exp x)))
```

While at first this syntax strikes as hard to understand, a second inspection reveals that it contains no ambiguities. To deal with the amount of parentheses one needs the support of a good text editor. But as a side effect, syntax errors almost disappear. The number of apparent errors (which would trigger a compiler error) and subtle (which make for hard to find errors) is greatly reduced. This is a large advantage over some modern languages that suffer from an exceedingly complex syntax (most notably, and relevant to our goals, C++), a feature which has been observed to degrade programmer productivity.

- **Code generation and macros:** A side effect of the simple and regular notation is that source code itself is directly amenable to machine manipulation. What is now being called "generative metaprogramming" using C++ templates has been present in Common Lisp since far more than a decade, and, since the complete language is available at compile time, in a more mature and powerful form [23].

- **Rich environment:** Development in Common Lisp usually happens interactively. The REPL (read-eval-print loop) makes it possible to inspect immediately newly defined components of the application without needing to restart the program from scratch. The user experience is similar than that from other interactive environments, while the performance can be the same as that of monolithic programs (this depends on the implementation).

- **Mature Standard:** The ANSI Common Lisp standard was formulated at a time when ample experience on the use of all features was available. It includes The Common Lisp Object System (CLOS), and its standard library includes many facilities that are only now beginning to appear in the standard library of modern languages; hash tables are but one prominent example.

While decried as slow and hard to use, and held to be certainly not a good choice for numerical applications, we found exactly the opposite to be true, and are not alone with that appreciation; see [31]. Performance comparable to C and Fortran is available in certain implementations[1].

For our purposes, the most important advantage was that it allowed us to explore many prototypes and perform many experiments. Its interactive nature and high performance allowed us to do so with little effort. Many different discretizations and configurations were tried before arriving at the configuration presented in section 4.4. Many more than would have been possible using any other language.

[1] We used `CMUCL`, a high performance Common Lisp compiler to be found at `http://www.cons.org/cmucl`

Bibliography

[1] I. Babuska. The finite element method with lagrangian multipliers. *Numer. Math.*, 20:179–192, 1973.

[2] J. Bergh and J. Löfström. *Interpolation spaces: an introduction*, volume 223 of *Die Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, Germany / Heidelberg, Germany / London, UK / etc., 1976.

[3] C. Börgers and O. B. Widlund. On finite element domain imbedding methods. *SIAM Journal of Numerical Analysis*, 27(4):963–978, 1990.

[4] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

[5] V. I. Burenkov. Extension theory for Sobolev spaces on open sets with Lipschitz boundaries. *Nonlinear Analysis, Function Spaces, and Applications*, 6, 1999.

[6] D. Calvetti, B. Lewis, and L. Reichel. GMRES-type methods for inconsistent systems. *Linear Algebra and its Applications*, 316(1–3):157–169, 2000.

[7] A. Cohen, W. Dahmen, and R. DeVore. Adaptive wavelet methods II: Beyond the elliptic case. *IGPM Preprint Series*, 199, 2000.

[8] A. Cohen, W. Dahmen, and R. DeVore. Adaptive wavelet methods for elliptic operator equations: Convergence rates. *Mathematics of Computation*, 70(233):27–75, 2001.

[9] A. Cohen, I. Daubechies, and J. C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 45:485–560, 1992.

[10] . Dahlke, W. Dahmen, and K. Urban. Adaptive wavelet methods for saddle point problems - optimal convergence rates. *SIAM Journal of Numerical Analysis*, 40(4):1230–1262, 2002.

[11] W. Dahmen. Wavelet and multiscale methods for operator equations. *Acta Numerica*, 6:55–228, 1997.

[12] W. Dahmen and A. Kunoth. Appending boundary conditions by lagrange multipliers: Analysis of the lbb condition. *Numer. Math.*, 88:9–42, 2001.

[13] W. Dahmen and R. Schneider. Composite wavelet bases for operator equations. *Mathematics of Computation*, 68(228):1533–1567, 1999.

[14] W. Dahmen and R. Schneider. Wavelets on manifolds I: Construction and domain decomposition. *SIAM Journal on Mathematical Analysis*, 31(1):184–230, 2000.

[15] F. Deutsch. *Best approximation in inner product spaces.* Springer-Verlag, Berlin, Germany / Heidelberg, Germany / London, UK / etc., 2001.

[16] R. DeVore and V. Popov. Interpolation of besov spaces, 1988.

[17] R. DeVore and R. Sharpley. Besov spaces on domains in $\mathbb{R}^d$. *Trans. Amer. Math. Soc.*, (335):843–864, 1993.

[18] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

[19] W. H. Flemming. *Functions of severall variables.* Addison Wesley, Reading, 1965.

[20] H. Fujita, H. Kawahara, and H. Kawarada. Distribution theoretic approach to fictitious domain method for neumann problems. *East-West Journal for Numerical Analysis*, 3(2):111–126, 1995.

[21] V. Girault and R. Glowinski. Error analysis of a fictitious domain method applied to a dirichlet problem. *Japan Journal of Industrial and Applied Mathematics*, 12(3):487–514, 1995.

[22] R. Glowinski, P. Tsorng-Whay, and J. Periaux. A fictitious domain method for dirichlet problem and applications. *Comput. Methods Appl. Mech. Eng.*, 111(3-4):283–303, 1994.

[23] P. Graham. *On Lisp: Advanced Techniques for Common Lisp.* Prentice Hall, Englewood Clifs, 1994.

[24] J. Haslinger and R. A. E. Mäkinen. *Introduction to Shape Optimization: Theory, Approximation, and Computation.* Number 7 in Advances in Design and Control. SIAM, Philadelphia, 2001.

[25] M. R. Hestenes. Pseudoinverses and conjugate gradients. *Comm. of the ACM*, 18(1):40–43, 1975.

[26] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Stand.*, 49:409–436, 1952.

[27] A. Kunoth. Wavelet techniques for the fictitious domains - lagrange multiplier approach. *Numer. Algor.*, (27):291–316, 2001.

[28] J. L. Lions and E. Magenes. *Nonhomogeneous Boundary Value Problems and Applications*, volume I. Springer-Verlag, Berlin, Germany / Heidelberg, Germany / London, UK / etc., 1972.

[29] J. McCarthy. Recursive functions of symbolic expressions and their computation by machine, part i. *CACM*, 3(4):184–195, 1960.

[30] M. Z. Nashed. Perturbations and approximations for generalized inverse and linear operator equations. In M. Z. Nashed, editor, *Generalized Inverses and Applications*, pages 325–396, London, October 1973. Academic Press.

[31] N. Neuss. On using Common Lisp in scientific computing. In *Proceedings of the CISC 2002*. Springer-Verlag, 2003.

[32] C. C. Paige and M. A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions in Mathematical Software*, 8(1):43–71, 1982.

[33] S. Rempel and B. W. Schulze. *Index theory of elliptic boundary problems*. Akademie-Verlag, Berlin, 1982.

[34] M. Renardy and R. C. Rogers. *An Introduction to Partial Differential Equations*. Texts in Applied Mathematics. Springer-Verlag, Berlin, Germany / Heidelberg, Germany / London, UK / etc., 1993.

Lebenslauf:

Persönliche Daten:

Name:	Mario Salvador Mommer
Geburtsdatum:	17. Mai 1973
Geburtsort:	Tübingen
Staatsangehörigkeit:	Deutscher

Schulbildung:

1978-1984	Grundschule "Carlos Emilio Muñoz Oráa", Mérida, Venezuela.
1984-1988	Gymnasium "Arzobispo Silva", Mérida, Venezuela.
1988-1989	Gymnasium "Andrés Eloy Blanco", Mérida, Venezuela.

Studium

2/1990-5/1998	Studium zum Lizentiaten ("Licenciatura") in Mathematik, and der Universidad de los Andes, Mérida, Venezuela.
12/1999	Nach erfolgreicher Erfüllung von Auflagen, Zulassung zur Promotion an der RWTH-Aachen durch die Fakultät fur Mathematik, Informatik und Naturwissenschaften

Beruf:

05/1994-12/1997	Programmierer (C und Motif (X-Windows) unter SunOS/Irix) in SUMA(ULA) ("Sistema Unificado de coMputación Académica"), Mérida, Venezuela
02/1998-05/1998	Programmierer (VisualC++ unter Microsoft Windows) für *HACER Sistemas*, Mérida, Venezuela; Analyse und Darstellung seismischer Datensätze
11/1998-11/1999	Programmierer (C++ und Matlab unter Irix/Linux) am IGPM ("Institut für Geometrie und Praktische Mathematik"), Aachen. Implementieren und Testen numerischer Algorithmen.
12/1999-08/2005	Wissenschaftlicher Mitarbeiter am IGPM, RWTH-Aachen.