Aalto University
School of Science

Master’s programme in Mathematics and Operations Research

# $L^2$ Convergence of the $p$- Finite Element Method for the 2D Poisson Problem with a Dirac Delta Load

**Joonas Laaksonen**

Master’s Thesis
2024

© 2024

This work is licensed under a Creative Commons “Attribution-NonCommercial-ShareAlike 4.0 International” license.

Aalto University
School of Science

**Author** Joonas Laaksonen

**Title** $L^2$ Convergence of the $p$- Finite Element Method for the 2D Poisson Problem with a Dirac Delta Load

**Degree programme** Mathematics and Operations Research

**Major** Applied Mathematics

**Supervisor and advisor** D.Sc. (Tech.) Harri Hakula

**Date** 29 April 2024 **Number of pages** 88 **Language** English

**Abstract**

In this thesis, we study the convergence properties of the $p$-version of the finite element method when it is applied to the two-dimensional Poisson problem with a Dirac delta load term and with Dirichlet or Neumann boundary conditions. Based on identical existing results for the $h$-version of the finite element method, we prove that the approximations from the $p$-version converge to the exact solution in $L^2$ when the domain satisfies certain convexity assumptions, which does not seem to be covered by the existing literature. Before proving this result, we provide detailed descriptions of the solvability of the Dirac delta problem and of the $p$- finite element method along with the needed approximation theorems to prove the main convergence result. Finally, we consider the accuracy of the obtained error bound through various numerical experiments, and we conclude that the predicted convergence rate of the error can be fairly close to the actual observed rate when the Dirac delta load is located either on a side or in the interior of an element of the mesh. When the load is located at a vertex of the mesh, the predicted rate of convergence is overly pessimistic, however.

**Keywords** $p$-FEM, Poisson's equation, Dirac delta function, $L^2$ convergence

Aalto-yliopisto
Perustieteiden
korkeakoulu

**Tekijä** Joonas Laaksonen

**Työn nimi** Elementtimenetelmän $p$-version $L^2$-suppeneminen pistekuormitetun 2D Poissonin ongelman tapauksessa

**Koulutusohjelma** Matematiikka ja operaatioanalyysi

**Pääaine** Sovellettu matematiikka

**Työn valvoja ja ohjaaja** TkT Harri Hakula

**Päivämäärä** 29.4.2024 **Sivumäärä** 88 **Kieli** englanti

**Tiivistelmä**

Tässä diplomityössä tarkastellaan elementtimenetelmän $p$-version suppenemista kaksiulotteisen Poissonin yhtälön tapauksessa, kun yhtälön kuormana on Diracin deltafunktio ja määrittelyjoukon reunoille on asetettu Dirichlet'n tai Neumannin reunaehdot. Elementtimenetelmän $h$-version tiedetään suppenevan ongelman tarkkaan ratkaisuun $L^2$:ssa, kun ongelman määrittelyjoukko toteuttaa tietyt konveksisuusehdot. Tämä tulos laajennetaan tässä työssä $p$-versiolle, mitä ei ilmeisesti ole tarkasteltu aiemmin. Ensin kuitenkin tarkastellaan ongelman tarkan ratkaisun olemassaoloa ja sen yksikäsitteisyyttä. $L^2$-suppenemistulos seuraa pitkälti $p$-versiolle ominaisten approksimaatiotulosten pohjalta, jotka käydään myös perusteellisesti läpi. Saatua virhe-estimaattia verrataan tietokoneella numeerisesti laskettuihin virheisiin. Numeeristen tulosten pohjalta voidaan päätellä, että virheen ennustettu suppenemisnopeus on kohtalaisen lähellä todellista suppenemisnopeutta, kun deltakuorma on elementin reunalla tai elementin sisällä. Kun deltakuorma on elementin solmussa, todellinen suppenemisnopeus on kuitenkin huomattavasti suurempi kuin ennustettu nopeus.

**Avainsanat** $p$-FEM, Poissonin yhtälö, Diracin deltafunktio, $L^2$-suppeneminen

# Preface

This thesis is inspired by a similar project which my supervisor and advisor Dr. Harri Hakula collaborated in during his visit to the University of Texas at Austin in 2013. The results from that project helped shape the body of this thesis.

I would like to extend my sincere thanks to Dr. Hakula for his guidance and encouragement which have been invaluable to me. I also wish to thank Professor Nuutti Hyvönen for helping me find a topic for my thesis.

My special thanks go to my parents, my brother and my late grandparents for their unwavering support throughout my educational journey.

Joonas Laaksonen
Otaniemi, 29 April 2024

# Contents

# 1 Introduction

A partial differential equation (PDE) is an equation that consists of an unknown function of two or more variables and its partial derivatives of arbitrary order [1]. Such equations describe how a function, possibly corresponding to a physical quantity of interest, behaves over its domain of definition which in practical applications typically corresponds to a geometric shape or time or both. For example, many fundamental laws of physics can be elegantly expressed as partial differential equations, e.g. Maxwell's equations in electromagnetism. The set of all possible partial differential equations is incredibly vast and complex, which makes it unwieldy, if not impossible, to come up with a general PDE theory that could be productively used for any kind of problem. Instead, the study of PDEs focuses on important families and instances of partial differential equations arising from different fields of science.

A solution to a partial differential equation is said to be classical if it can be differentiated at least as many times as the formulation of the equation requires and the equation holds pointwise everywhere in the domain of the problem. The solution may also need to satisfy some conditions on the boundary of its domain, turning the problem into a boundary value problem. However, it turns out that rather few partial differential equations have classical solutions [1]. Thus, the modern theory of partial differential equations typically abandons the quest of searching for a classical solution and instead reformulates the problem in a more generalized form which relaxes the requirements that the solution must fulfil regarding e.g. differentiability. The solution space is essentially expanded to contain less smooth functions, which may even be favorable for a modeled phenomenon for which the solution is expected to be non-differentiable at some points. A solution to the generalized problem is typically called a generalized solution or a weak solution, and its existence can be guaranteed for a large set of problems. Whether a weak solution is also a classical solution can then be assessed separately, and such results belong to the regularity theory of partial differential equations.

In practical applications, numerical methods are used to approximate the solutions of partial differential equations. One such method that has been extremely successful is the finite element method (FEM). Oden [2] presents a review of its history. Without too stringent a viewpoint, some attributes of the finite element method can be traced back a couple of centuries, but serious interest in the method started to accumulate during the mid-1950s and 1960s especially in the engineering community. During this time, the finite element method also gained its name. The mathematical foundations were established somewhat later during the 1970s, and ever since, different variants of the finite element method have been developed, with the $h$-, $p$- and $hp$-versions being among the most popular variants.

Oden also discusses some of the factors leading to the success of the finite element method. The method is based on the generalized, weak formulation of a partial differential equation, which will be discussed in more detail in later sections, but for now it suffices to say that it is a crucial factor for the reason why the finite element method merits its success. Being based on the weak formulation, the finite element method is essentially geometry-agnostic, which means that it can be used to solve problems over almost any kind of shape. Combined with the rich modern theory of partial differential equations, the finite element method has solid mathematical foundations which offer optimal estimates for the convergence of the approximations. Moreover, from a computational standpoint, the implementation of the method in computer code lends itself extremely well to parallelization.

Decision-making based on computed information requires that the computed information is reliable. Szabó and Babuška [3] discuss this systematically in the context of finite element analysis, i.e. the process of using the finite element method to solve a problem. In finite element analysis, the typical workflow is to first create a mathematical model, i.e. a set of partial differential equations and constraints that represent the physical system of interest, then find an approximate solution to the model by using the finite element method and, finally, extract the desired information from the computed approximate solution. There are two critical factors contributing to the reliability of such computed information: the suitability of the mathematical model as a representation of idealized reality and the accuracy of the approximate solution with respect to the exact solution. Szabó and Babuška term the processes of assessing these qualities validation and verification, respectively. The validation process may consist of, for example, comparing the results of real-life experiments to predictions obtained from the mathematical model. The verification process leans on the well-understood approximation properties of the finite element method.

Sometimes a physical phenomenon is modeled sufficiently well by a mathematical model for which the approximation error estimates from the standard theory of the finite element method do not directly apply. If possible, one option to still be able to perform the verification process would be to modify the model so that the error estimates readily apply. However, there could be some tradeoffs involved in the choice between the models, which could still make the initial model more favorable, e.g. the initial model is a crude simplification but requires much less effort to solve. For example, Babuška et al. [4] study the effects of replacing holes having extremely small radii with singular points, i.e. holes with the radii equal to zero, which simplifies the approximation process but essentially renders the model as incorrect.

In this thesis, we study a problem similar in vein to the problem mentioned above studied by Babuška et al. which may simplify modeling but for which the convergence of the finite element approximations is not obvious. We study the Poisson problem with a Dirac delta load term in a two-dimensional domain, and it can be expressed as finding a real-valued function $u$ defined over a domain $\Omega \subset \mathbb{R}^2$ with some predefined boundary values such that

$$-\Delta u = \delta_{x_0} \quad \text{in } \Omega, \tag{1.1}$$

where $\Delta u$ is the Laplacian of $u$, i.e. the sum of the second partial derivatives of $u$ with respect to each independent variable, and $\delta_{x_0}$ is the Dirac delta for a given point $x_0 \in \Omega$ which can be loosely regarded as a function that is concentrated at the point $x_0$ such that

$$\delta_{x_0}(x) = \begin{cases} \infty, & x = x_0 \\ 0, & x \neq x_0 \end{cases}$$

for all $x \in \Omega$ and the integral of $\delta_{x_0}$ over $\Omega$ is one. The adverb "loosely" shall be emphasized because the standard Lebesgue integration theory forbids the existence of such a function, but the above characterization still provides useful intuition. This also means that the partial differential equation (1.1) should be understood in some specific sense, namely in the weak sense. Poisson's equation can be used to model, for example, the electrostatic potential caused by an electric charge, and with the Dirac delta load term, the equation can be used to model a point charge located at the point $x_0$, i.e. an idealized charge with no area or volume [5].

More specifically, we study the problem (1.1) in bounded polygonal two-dimensional domains with some additional convexity assumptions and with Dirichlet and Neumann boundary conditions. The first objective of this thesis is to consider the unique solvability of these boundary value problems. Casas [6] proves this for the Dirichlet problem with homogeneous boundary values, and we extend this result to the other boundary value problems with non-homogeneous boundary values.

The second objective is to study the convergence of the $p$-version of the finite element method when it is applied to the problem (1.1). The convergence of the $h$-version when applied to the same problem has already been studied quite extensively in the existing literature. For example, Casas [6] and Scott [7] prove convergence in $L^2$ when applied to the homogeneous Dirichlet problem. Moreover, Schatz and Wahlbin [8] prove estimates for pointwise convergence, and more recent results have been obtained by Millar et al. [9] who obtain convergence by approximating the Dirac delta and by Araya et al. [10] who deduce a posteriori error estimates.

The $p$-version has not received as much attention for the same problem, however. We extend the $L^2$ convergence result by Casas to the $p$-version resulting in the bound

$$\|u - u_S\|_{L^2(\Omega)} \leq Cp^{-1}\left(1 + \sqrt{\ln(p+1)}\right),$$

where $u$ and $u_S$ are the exact solution and the finite element solution, respectively, and $C > 0$ is a constant independent of $p$. The value of the constant $C$ is not known, however. The bound above is the main result of this thesis, and all the details will be covered in the subsequent sections. We also assess the convergence numerically for a Neumann problem for which the exact solution is known.

The remainder of this thesis is structured as follows. Section 2 presents the preliminary mathematical concepts that are essential in the weak formulation of partial differential equations and in the theory of the finite element method. In Section 3, we first consider how a classically formulated boundary value problem for Poisson's equation can be transformed into the weak form, and then we consider its solvability. In particular, we consider the weak formulation and solvability of the problem (1.1), reaching the first objective of this thesis. Section 4 is devoted to the theory of the finite element method with emphasis on the $p$-version. A substantial portion of this section deals with the approximation properties of high-order piecewise polynomials which form the basis for the convergence properties of the $p$-version. In Section 5, we prove the $L^2$ convergence of the $p$-version for the problem (1.1) and compare the obtained error bound to numerical results, reaching the second objective of this thesis. Finally, Section 6 contains a summary of the results, and we discuss possible future lines of study.

# 2 Preliminaries

In this section, we review the following preliminary content. We begin by covering properties related to open subset of $\mathbb{R}^n$ which correspond to the geometric domains where the boundary value problems shall be defined. We will see in a later section that the solvability of many boundary value problems follows from abstract results in functional analysis. In anticipation of this, we go through the definitions and basic properties of Banach spaces, Hilbert spaces, basic types of linear operators and dual spaces. Finally, we consider Lebesgue spaces and Sobolev spaces that correspond to the function spaces where we shall search for the solutions of the problems.

## 2.1 Domains and Boundaries

Throughout the rest of this thesis, $\Omega$ denotes a subset of $\mathbb{R}^n$, the $n$-dimensional Euclidean space. We start restricting ourselves to the case $n = 2$ towards the end of this section, as that is the setting for the boundary value problems we are looking to study. The set $\Omega$ is typically assumed to be non-empty, open and connected, and such a set is called a domain. Being connected simply means that $\Omega$ cannot be given as the union of two disjoint non-empty open sets. Disconnected sets could be handled by considering the disjoint parts separately. We typically also assume that $\Omega$ is bounded, i.e. it can be enclosed within a ball of finite radius.

Convergence results in finite element analysis typically rely on regularity results that guarantee higher-order smoothness of the solutions of the boundary value problems. Many of these regularity results, in turn, rely on the properties of the domain. One such property is convexity which will be one of our key assumptions later on.

**Definition 2.1** (Convex set)**.** A set $\Omega \subset \mathbb{R}^n$ is said to be convex if for every pair of points $x, y \in \Omega$ it holds that $tx + (1-t)y \in \Omega$ for all $t \in (0, 1)$.

Another important property of a domain is the regularity of its boundary. The boundary of a domain $\Omega$ is denoted by $\partial\Omega$. Analogously to functions in analysis, boundaries can be defined to be continuous, Lipschitz continuous, continuously differentiable, etc. The definition of a domain with Lipschitz boundary is given below, and it borrows notation from Grisvard [11] and Schwab [12]. See also Figure 1 below.

**Definition 2.2** (Lipschitz boundary)**.** Assume that $n \geq 2$, and let $\Omega \subset \mathbb{R}^n$ be a domain. The boundary $\partial\Omega$ is said to be Lipschitz if for every $x \in \partial\Omega$ there exist a Lipschitz function $g : \mathbb{R}^{n-1} \to \mathbb{R}$, a local coordinate system $\{\tilde{e}_1, \dots, \tilde{e}_n\} \subset \mathbb{R}^n$ and constants $d > 0$ and $h > 0$ such that whenever a point $y = \sum_{i=1}^n \tilde{y}_i \tilde{e}_i \in \mathbb{R}^n$ satisfies $|\tilde{y}_i| < d$ for all $i = 1, \dots, n-1$, then the following hold.

(i) If $g(\tilde{y}_1, \dots, \tilde{y}_{n-1}) = \tilde{y}_n$, then $y \in \partial\Omega$.

(ii) If $g(\tilde{y}_1, \dots, \tilde{y}_{n-1}) < \tilde{y}_n < g(\tilde{y}_1, \dots, \tilde{y}_{n-1}) + h$, then $y \in \Omega$.

(iii) If $g(\tilde{y}_1, \dots, \tilde{y}_{n-1}) > \tilde{y}_n > g(\tilde{y}_1, \dots, \tilde{y}_{n-1}) - h$, then $y \notin \Omega$.

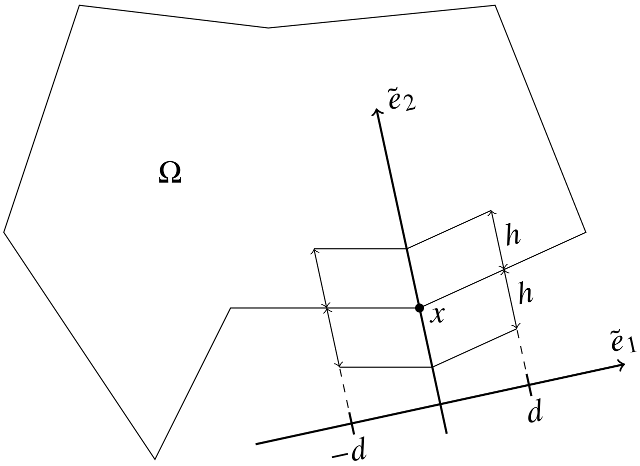


**Figure 1:** A domain with Lipschitz boundary.

In other words, the requirement (i) means that a Lipschitz boundary is locally the image of a Lipschitz function. The requirements (ii) and (iii) mean that the set $\Omega$ is not allowed to be on both sides of its boundary. Other types of boundary regularity have analogous definitions by just replacing the word “Lipschitz” with e.g. “continuously differentiable”.

In finite element analysis, the domain is typically modeled as a union of polygons. A polygon consists of vertices and line segments that connect the vertices. In general, a polygonal domain can have holes, that are of course polygonal as well, but a convex polygonal domain cannot have any holes. Assuming that the angle between adjacent line segments is never 0 nor $2\pi$, a polygonal domain has Lipschitz boundary [11]. This is an important property, as having Lipschitz boundary is needed for many later results. Clearly, the boundary of a polygonal domain is not differentiable.

## 2.2 Functional Analysis

We describe here some of the most fundamental concepts and results from functional analysis. For a more complete reference, see for example [13] and [14].

### 2.2.1 Vector Spaces, Normed Spaces and Inner Product Spaces

We begin with the definition of a vector space over the real numbers.

**Definition 2.3** (Real vector space)**.** A vector space over the real numbers is a non-empty set $X$ equipped with the following two operations.

(i) Addition: $x + y \in X$ for all $x, y \in X$.

(ii) Scalar multiplication: $ax \in X$ for all $a \in \mathbb{R}$ and $x \in X$.

In addition, the following axioms hold for all $x, y, z \in X$ and $a, b \in \mathbb{R}$.

(i) Commutativity: $x + y = y + x$.

(ii) Associativity: $(x + y) + z = x + (y + z)$ and $(ab)x = a(bx)$.

(iii) Existence of additive identity: there exists a zero vector $0 \in X$ such that $x+0 = x$.

(iv) Existence of additive inverse: for every $x \in X$, there exists a $y \in X$ such that $x + y = 0$.

(v) Multiplicative identity: $1x = x$.

(vi) Distributivity: $a(x + y) = ax + ay$ and $(a + b)x = ax + bx$.

By a vector space, we shall always mean a real vector space. A subset of a vector space is said to be a vector subspace, or simply a subspace, if it is closed under the operations of addition and scalar multiplication of the ambient vector space.

**Definition 2.4** (Vector subspace)**.** A non-empty subset $S$ of a vector space $X$ is said to be a vector subspace of $X$ if $ax + by \in S$ for all $a, b \in \mathbb{R}$ and $x, y \in S$.

Next, we define a norm on a vector space which can be thought to be a measure of the magnitude of a vector.

**Definition 2.5** (Norm)**.** A norm on a vector space $X$ is a function $\|\cdot\| : X \longrightarrow \mathbb{R}$ that satisfies the following properties for all $x, y \in X$ and $a \in \mathbb{R}$.

(i) $\|x + y\| \leq \|x\| + \|y\|$.

(ii) $\|ax\| = |a|\|x\|$.

(iii) $\|x\| \geq 0$, and $\|x\| = 0$ if and only if $x = 0$.

A vector space equipped with a norm is called a normed space. For example, $\mathbb{R}^n$ is a normed space with the Euclidean norm. We shall always assume that a normed space is equipped with the usual norm topology. This means that given a normed space $X$, a subset $S \subset X$ is open if and only if for every $x \in S$ there exists a ball $B(x, \varepsilon) = \{y \in X : \|x - y\| < \varepsilon\}$ with a positive radius $\varepsilon > 0$ such that $B(x, \varepsilon) \subset S$. If a function $s : X \longrightarrow \mathbb{R}$ does not satisfy the condition that $s(x) = 0$ implies $x = 0$ but otherwise satisfies Definition 2.5, then $s$ is said to be a seminorm.

A norm enables us to consider the convergence of a sequence of vectors.

**Definition 2.6** (Convergence of a sequence)**.** A sequence of vectors $(x_i)_{i=1}^{\infty}$ in a normed space $X$ is said to converge to the limit $x \in X$ if for every $\varepsilon > 0$ there exists a number $N > 0$ such that $\|x - x_i\| < \varepsilon$ whenever $i \geq N$.

The condition in Definition 2.6 can be written succinctly as $\lim_{i\to\infty}\|x - x_i\| = 0$. We can use convergence to define the concept of a closure point of a set.

**Definition 2.7** (Closure point)**.** Let $S$ be a subset of a normed space $X$. A point $x \in X$ is said to be a closure point of $S$ if there exists a sequence $(x_i)_{i=1}^{\infty}$ in $S$ that converges to $x$. The set of all closure points of $S$ is the closure of $S$, and it is denoted by $\overline{S}$.

The definition of the closure has some useful consequences. It is a standard result that $\overline{S}$ is closed. In fact, $S$ is closed precisely when $S = \overline{S}$. Since obviously $S \subset \overline{S}$, this means that to prove closedness of $S$ we only need to show that every closure point of $S$ belongs to $S$, i.e. the limit belongs to $S$. If $\overline{S} = X$, it is said that the subset $S$ is dense in the normed space $X$. If $S$ is dense in $X$ and $S$ has a well-known structure with useful properties, then these properties can often be extended to $X$ via so-called density arguments.

The closure and density essentially allow one to pick a converging sequence for a given point. Compactness, on the other hand, allows one to pick a limit point for a given (sub)sequence.

**Definition 2.8** (Compact set)**.** A subset $S$ of a normed space $X$ is said to be compact if every sequence in $S$ has a subsequence that has a limit in $S$. Moreover, a subset $S$ is called precompact if $\overline{S}$ is compact.

Definition 2.8 is more generally referred to as sequential compactness. Compactness generally means that every open cover of the set $S$ has a finite subcover, but, in normed spaces, the more general definition of compactness and sequential compactness are equivalent.

An important class of normed spaces is Banach spaces for which we need to introduce two key concepts: Cauchy sequences and completeness.

**Definition 2.9** (Cauchy sequence)**.** A sequence $(x_i)_{i=1}^{\infty}$ in a normed space $X$ is said to be a Cauchy sequence if for every $\varepsilon > 0$ there exists a number $N > 0$ such that $\|x_i - x_j\| < \varepsilon$ whenever $i, j \geq N$.

**Definition 2.10** (Completeness)**.** A normed space $X$ is said to be complete if every Cauchy sequence in $X$ converges to a limit in $X$.

A Banach space is simply a normed space that is complete, i.e. it satisfies Definition 2.10. It is easy to see that a closed subspace $S$ of a Banach space $X$ is also a Banach space: a Cauchy sequence in $S$ is also a Cauchy sequence in $X$, which means that it converges to some point $x \in X$, but this implies that $x$ is a closure point of $S$, which then implies that $x \in S$ since $S$ is closed and, thus, $S$ is complete.

Hilbert spaces are an important special case of Banach spaces for which we need the concept of an inner product. An inner product can be thought to be a measure of the orthogonality between two vectors.

**Definition 2.11** (Inner product)**.** An inner product on a vector space $X$ is a function $\langle \cdot, \cdot \rangle : X \times X \longrightarrow \mathbb{R}$ that satisfies the following properties for all $x, y, z \in X$ and $a \in \mathbb{R}$.

(i) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$.

(ii) $\langle ax, y \rangle = a \langle x, y \rangle$.

(iii) $\langle x, y \rangle = \langle y, x \rangle$.

(iv) $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0$ if and only if $x = 0$.

A vector space equipped with an inner product is called an inner product space. An inner product also induces a norm. Namely, given an inner product space $X$, the mapping $x \mapsto \sqrt{\langle x, x \rangle}$ defines a norm on $X$, which means that $X$ is also a normed space. An inner product space that is complete with respect to the induced norm is said to be a Hilbert space.

An important result for inner product spaces is the Cauchy-Schwarz inequality.

**Theorem 2.1** (Cauchy-Schwarz inequality)**.** *Let $X$ be an inner product space with the inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|\cdot\|$. Then for all $x, y \in X$, it holds that*

$$|\langle x, y \rangle| \leq \|x\|\|y\|.$$

*Proof.* Let $x, y \in X$. Then

$$\begin{aligned} 0 &\leq \|x + y\|^2 \\ &= \langle x + y, x + y \rangle \\ &= \langle x, x \rangle + \langle y, y \rangle + 2 \langle x, y \rangle \\ &= \|x\|^2 + \|y\|^2 + 2 \langle x, y \rangle \end{aligned}$$

so that $- \langle x, y \rangle \leq (\|x\|^2 + \|y\|^2)/2$. An identical argument with $x - y$ instead of $x + y$ yields $\langle x, y \rangle \leq (\|x\|^2 + \|y\|^2)/2$. Combining these two inequalities yields

$$|\langle x, y \rangle| \leq (\|x\|^2 + \|y\|^2)/2. \tag{2.1}$$

If $x = 0$ or $y = 0$, the claim is obviously true. Let us thus assume that $x \neq 0 \neq y$. Now

$$\begin{aligned} |\langle x, y \rangle| &= \|x\|\|y\| \left| \left\langle \frac{x}{\|x\|}, \frac{y}{\|y\|} \right\rangle \right| \\ &\leq \|x\|\|y\|(1 + 1)/2 \\ &= \|x\|\|y\|, \end{aligned}$$

where we used the inequality (2.1). □

### 2.2.2 Linear Operators

In functional analysis, the noun operator is typically a synonym for a function. Operators that are linear are arguably among the most important entities in functional analysis. We consider here some basic types of linear operators that we will need later. Specifically, we cover continuous, bounded and compact linear operators.

**Definition 2.12** (Linear operator)**.** An operator $A : X \longrightarrow Y$ between two vector spaces $X$ and $Y$ is said to be linear if it satisfies the following properties for all $x, y \in X$ and $a \in \mathbb{R}$.

(i) $A(x + y) = A(x) + A(y)$.

(ii) $A(ax) = aA(x)$.

An obvious corollary to the above definition is that $A(0) = 0$ since $A(0) = A(x - x) = A(x) - A(x) = 0$. When an operator is linear, we adopt the common practice of dropping the parentheses around a singular argument. That is, $Ax$ is used to mean $A(x)$.

The definition of a continuous linear operator between two normed spaces is identical to the definition of a continuous function typically encountered in e.g. calculus.

**Definition 2.13** (Continuous linear operator)**.** A linear operator $A : X \to Y$ between two normed spaces $X$ and $Y$ is said to be continuous if for every $x \in X$ and for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $\|Ax - Ay\|_Y < \varepsilon$ for all $y \in X$ satisfying $\|x - y\|_X < \delta$.

Next, we define a bounded linear operator which should not be confused with the usual definition of boundedness of a function.

**Definition 2.14** (Bounded linear operator)**.** A linear operator $A : X \to Y$ between two normed spaces $X$ and $Y$ is said to be bounded if there exists a constant $C > 0$ such that $\|Ax\|_Y \le C\|x\|_X$ for all $x \in X$.

In normed spaces, there is an important connection between continuous linear operators and bounded linear operators. Namely, they are precisely the same concept. We prove this basic but fundamental result.

**Theorem 2.2.** *Let $A : X \to Y$ be a linear operator between normed spaces $X$ and $Y$. Then $A$ is continuous if and only if it is bounded.*

*Proof.* Assume first that $A$ is continuous. $A$ is then continuous at the origin $0 \in X$, which means that for $\varepsilon = 1$ there exists a $\delta > 0$ such that $\|Ay\|_Y < 1$ for all $y \in X$ satisfying $\|y\|_X < \delta$. Let now $x \in X$, and assume that $x \neq 0$. Define

$$y = \frac{\delta}{2}\frac{x}{\|x\|_X}$$

for which clearly $\|y\|_X = \delta/2 < \delta$. Now by using the linearity and continuity of $A$, we get

$$\frac{\delta}{2\|x\|_X}\|Ax\|_Y = \left\| A\left(\frac{\delta}{2}\frac{x}{\|x\|_X}\right)\right\|_Y = \|Ay\|_Y < 1.$$

Multiplying each side by $2\|x\|_X/\delta$ yields

$$\|Ax\|_Y \le \frac{2}{\delta}\|x\|_X.$$

If $x = 0$, this inequality obviously holds as well since $\|Ax\|_Y = 0 = \|x\|_X$. Thus, $A$ is bounded with the constant $C = 2/\delta > 0$.

Then assume that $A$ is bounded. Let $x \in X$, $\varepsilon > 0$, and set $\delta = \varepsilon/C$, where $C$ is the constant in the definition of boundedness. Now for all $y \in X$ satisfying $\|x - y\|_X < \delta$, it holds that

$$\|Ax - Ay\|_Y = \|A(x - y)\|_Y \le C\|x - y\|_X < C\delta = \varepsilon,$$

where we used the linearity and boundedness of $A$. Thus, $A$ is continuous. □

A compact linear operator is defined as follows.

**Definition 2.15** (Compact linear operator)**.** A linear operator $A : X \to Y$ between two normed spaces $X$ and $Y$ is said to be compact if $A$ maps every bounded set in $X$ to a precompact set in $Y$.

A precompact set was defined in Definition 2.8, and a bounded set $S$ in $X$ is naturally understood as the existence of a radius $r > 0$ such that $S \subset B(0, r)$.

### 2.2.3 Dual Spaces

The concept of a dual space and, in particular, the so-called Riesz representation theorem will be essential tools when we study the existence and uniqueness of solutions to weakly formulated boundary value problems.

**Definition 2.16** (Dual space)**.** The dual space of a normed space $X$, denoted by $X'$, is the vector space of all real-valued continuous linear operators defined on the space $X$. The vector space operations are the usual pointwise operations: for all $\varphi, \psi \in X'$, $a \in \mathbb{R}$ and $x \in X$, the addition and scalar multiplication are defined by

(i) $(\varphi + \psi)(x) = \varphi(x) + \psi(x)$,

(ii) $(a\varphi)(x) = a\varphi(x)$.

Real-valued operators defined on a vector space are more commonly referred to as functionals. As an example, we now interpret the Dirac delta $\delta_{x_0}$ as an element of the dual space of $C(\overline{\Omega})$, where $C(\overline{\Omega})$ is the normed space of bounded uniformly continuous functions defined on a domain $\Omega \subset \mathbb{R}^n$. The norm is given by

$$\|u\|_{C(\overline{\Omega})} = \sup_{x\in\Omega} |u(x)|.$$

Let $x_0 \in \Omega$. We define the Dirac delta as a functional $\delta_{x_0} : C(\overline{\Omega}) \to \mathbb{R}$ that evaluates its argument at the point $x_0$, that is, $\delta_{x_0}(u) = u(x_0)$ for all $u \in C(\overline{\Omega})$. To show that $\delta_{x_0} \in C(\overline{\Omega})'$, we need to show that it is linear and continuous. Linearity is easy: for all $u, v \in C(\overline{\Omega})$ and $a, b \in \mathbb{R}$, we have

$$\delta_{x_0}(au + bv) = (au + bv)(x_0) = au(x_0) + bv(x_0) = a\delta_{x_0}(u) + b\delta_{x_0}(v).$$

Boundedness is also easy: for all $u \in C(\overline{\Omega})$, we have

$$|\delta_{x_0}(u)| = |u(x_0)| \leq \sup_{x\in\Omega} |u(x)| = \|u\|_{C(\overline{\Omega})}.$$

By Theorem 2.2, boundedness and continuity mean the same thing so $\delta_{x_0}$ is continuous. Thus, $\delta_{x_0} \in C(\overline{\Omega})'$.

Given an inner product space $X$ and a vector $y \in X$, it immediately follows from the definition of an inner product (Definition 2.11) and the Cauchy-Schwarz inequality (Theorem 2.1) that the functional $x \mapsto \langle y, x\rangle$ defined on $X$ belongs to the dual space $X'$. When $X$ is also a Hilbert space, it turns out that every functional in $X'$ is of this form. This is the message of the Riesz representation theorem.

**Theorem 2.3** (Riesz representation theorem)**.** *Let $X$ be a Hilbert space with the inner product $\langle \cdot, \cdot \rangle$. Let $\varphi \in X'$. Then there exists a unique $y \in X$ such that $\varphi(x) = \langle y, x \rangle$ for all $x \in X$.*

For a proof, see for example [14, Theorem 4.12 on p. 81].

## 2.3 Lebesgue Spaces

This subsection is concerned with the properties of Lebesgue integrable functions. We skip the complete definitions of the Lebesgue measure and Lebesgue integration, for which the reader is referred to [15]. At the end of this subsection, we do, however, consider the problem of defining an integral over the boundary of a domain with Lipschitz boundary.

### 2.3.1 Lebesgue Integral

The Lebesgue measure over $\mathbb{R}^n$ is essentially a real-valued function that returns the $n$-dimensional volume of practically any set in $\mathbb{R}^n$, excluding some pathological sets, and that has properties one would typically expect such a function to have, e.g. invariance under rigid motion, such as translation and rotation, and that the measure of the union of two disjoint sets is the sum of their individual measures. We denote the Lebesgue measure of a measurable set $\Omega \subset \mathbb{R}^n$ by $|\Omega|$.

The notation for the Lebesgue integral of a measurable function $f : \Omega \to \mathbb{R}$ over a measurable set $\Omega \subset \mathbb{R}^n$ is the usual

$$\int_\Omega f \, dx, \tag{2.2}$$

given that the integral exists. When $\Omega = [a, b] \in \mathbb{R}$, we simply write $\int_a^b f \, dx$. The function $f$ is said to be integrable if its integral (2.2) exists and is finite. Integrability underlays the definition of Lebesgue spaces.

**Definition 2.17** (Lebesgue spaces)**.** Let $\Omega \subset \mathbb{R}^n$ be open, and let $1 \leq s < \infty$ be a real number. The Lebesgue space $L^s(\Omega)$ is the normed space of functions $u : \Omega \to \mathbb{R}$ that satisfy

$$\int_\Omega |u|^s \, dx < \infty, \tag{2.3}$$

and the norm is given by

$$\|u\|_{L^s(\Omega)} = \left( \int_\Omega |u|^s \, dx \right)^{1/s}.$$

Moreover, the space $L^2(\Omega)$ is defined to be an inner product space with the inner product

$$\langle u, v \rangle_{L^2(\Omega)} = \int_\Omega uv \, dx.$$

In fact, the Lebesgue spaces $L^s(\Omega)$ are Banach spaces, and $L^2(\Omega)$ is a Hilbert space.

We defined $u \in L^s(\Omega)$ as a regular function $u : \Omega \to \mathbb{R}$. However, this is not strictly speaking valid because such $u$ is not uniquely defined. To see why, assume that $v$ is another function that is equal to $u$ everywhere in $\Omega$ except in a non-empty set whose Lebesgue measure is zero, e.g. a countable set of points. Now $u - v$ is zero everywhere except in a non-empty set with Lebesgue measure zero, but then clearly $\|u - v\|_{L^s(\Omega)} = 0$, which implies that $u = v$ in the normed space $L^s(\Omega)$ even though they are technically different functions. The proper way to define $u \in L^s(\Omega)$ would be to consider it as an equivalence class of functions where two functions are considered to be equivalent if they satisfy the integrability condition (2.3) and they are equal almost everywhere, i.e. everywhere except in a set with Lebesgue measure zero. However, for simplicity, we stick to treating $u \in L^s(\Omega)$ as a function and say that $u = v$ if they are equal almost everywhere in $\Omega$. An obvious caveat regarding this choice is that the pointwise values of $u$ are not necessarily well-defined.

The space $L^\infty(\Omega)$ can be defined as the normed space of functions (again, equivalence classes to be pedantic) that are essentially bounded over $\Omega$. A function $u : \Omega \to \mathbb{R}$ is said to be essentially bounded over $\Omega$ if there exists a constant $M > 0$ such that $|u(x)| \leq M$ for almost every $x \in \Omega$. That is, $L^\infty(\Omega)$ is the normed space of functions $u : \Omega \to \mathbb{R}$ that satisfy

$$\|u\|_{L^\infty(\Omega)} = \operatorname*{ess\,sup}_{x\in\Omega} |u(x)| < \infty,$$

where

$$\operatorname*{ess\,sup}_{x\in\Omega} u(x) = \inf\{M \in \mathbb{R} : |\{x \in \Omega : u(x) > M\}| = 0\}$$

is the essential supremum of $u$ over $\Omega$. The space $L^\infty(\Omega)$ is a Banach space as well.

Different $L^s$-norm inequalities will be the backbone of many applications later. An important inequality is Hölder's inequality.

**Theorem 2.4** (Hölder's inequality)**.** *Let $1 \leq s \leq \infty$ and $1 \leq s' \leq \infty$ be such that $1/s + 1/s' = 1$ (when $s = \infty$ or $s' = \infty$, set $s' = 1$ and $s = 1$, respectively). Let $u \in L^s(\Omega)$ and $v \in L^{s'}(\Omega)$. Then $uv \in L^1(\Omega)$ and*

$$\|uv\|_{L^1(\Omega)} \leq \|u\|_{L^s(\Omega)} \|v\|_{L^{s'}(\Omega)}.$$

For a reference, see [15, Theorem 6.2 on p. 182] for the case $1 < s, s' < \infty$ and [15, Theorem 6.8 on p. 184] for the case $s = \infty$ or $s' = \infty$. Numbers $1 \leq s, s' \leq \infty$ that satisfy $1/s + 1/s' = 1$ are called conjugate exponents. Theorem 2.4 actually holds in an arbitrary measure space. In particular, it holds for the counting measure, which implies that for any two real-valued sequences $(x_i)_{i=1}^\infty$ and $(y_i)_{i=1}^\infty$ it holds that

$$\sum_{i=1}^{\infty} |x_i y_i| \leq \left(\sum_{i=1}^{\infty} |x_i|^s\right)^{\frac{1}{s}} \left(\sum_{i=1}^{\infty} |y_i|^{s'}\right)^{\frac{1}{s'}}.$$

When $\Omega$ is bounded, Hölder's inequality implies the imbedding $L^r(\Omega) \subset L^s(\Omega)$ when $1 \leq s \leq r \leq \infty$, which we prove next.

**Theorem 2.5.** *Assume that $\Omega$ is bounded, that is, $|\Omega| < \infty$. Let $1 \leq s \leq r \leq \infty$. Then there exists a constant $C > 0$ such that*

$$\|u\|_{L^s(\Omega)} \leq C\|u\|_{L^r(\Omega)}$$

*for all $u \in L^r(\Omega)$. In other words, $L^r(\Omega) \subset L^s(\Omega)$.*

*Proof.* The case $s = r$ is trivial so assume that $s < r$. Let $u \in L^r(\Omega)$, and assume first that $r < \infty$.

Define $v \equiv 1$ on $\Omega$. Since $\Omega$ is bounded, we have for all $1 \leq q < \infty$ that

$$\|v\|^q_{L^q(\Omega)} = \int_\Omega 1\,dx = |\Omega| < \infty \quad \text{and} \quad \|v\|_{L^\infty(\Omega)} = 1 < \infty.$$

That is, $v \in L^q(\Omega)$ for all $1 \leq q \leq \infty$.

Since $u \in L^r(\Omega)$, clearly $|u|^s \in L^{r/s}(\Omega)$. Note that $r/s > 1$, and its conjugate exponent is given by $r/(r-s)$. Now by applying Hölder's inequality to $|u|^s \in L^{r/s}(\Omega)$ and $v \in L^{r/(r-s)}(\Omega)$, we get

$$\begin{aligned}\|u\|^s_{L^s(\Omega)} &= \int_\Omega |u|^s\,dx \\ &= \int_\Omega v|u|^s\,dx \\ &\leq \|v\|_{L^{\frac{r}{r-s}}(\Omega)} \left(\int_\Omega |u|^r\,dx\right)^{\frac{s}{r}} \\ &= |\Omega|^{\frac{r-s}{r}} \|u\|^s_{L^r(\Omega)}.\end{aligned}$$

Taking the $s$th root from both sides proves the claim for $r < \infty$.

If $u \in L^\infty(\Omega)$, then clearly $|u|^s \in L^\infty(\Omega)$ as well with $\||u|^s\|_{L^\infty(\Omega)} = \|u\|^s_{L^\infty(\Omega)}$. Now by applying Hölder's inequality to $|u|^s \in L^\infty(\Omega)$ and $v \in L^1(\Omega)$, we get

$$\begin{aligned}\|u\|^s_{L^s(\Omega)} &= \int_\Omega v|u|^s\,dx \\ &\leq \|v\|_{L^1(\Omega)}\||u|^s\|_{L^\infty(\Omega)} \\ &= |\Omega|\|u\|^s_{L^\infty(\Omega)}.\end{aligned}$$

Taking the $s$th root from both sides proves the claim for $r = \infty$. □

The next theorem is the dominated convergence theorem which we will use later to verify the integrability of certain functions. For a proof, see [15, Theorem 2.24 on p. 54].

**Theorem 2.6** (The dominated convergence theorem)**.** *Let $\Omega \subset \mathbb{R}^n$ be open. Let $(f_i)_{i=1}^\infty$ be a sequence of functions in $L^1(\Omega)$ that converges pointwise almost everywhere in $\Omega$ to a function $f$. Assume that there exists a function $g \in L^1(\Omega)$ such that $|f_i| \leq g$ almost everywhere in $\Omega$ for all $i = 1, 2, \ldots$. Then $f \in L^1(\Omega)$ and*

$$\lim_{i\to\infty} \int_\Omega f_i\,dx = \int_\Omega f\,dx.$$

We will also need the following result which essentially states that the average values of a continuous function over balls or spheres with a fixed center point converge to the value of the function at that center point as the radius tends to zero.

**Theorem 2.7.** *Let $\Omega \subset \mathbb{R}^n$ be open, and let $f : \Omega \to \mathbb{R}$ be a continuous function. Then for all $x \in \Omega$, it holds that*

$$\lim_{r\to 0^+} \frac{1}{|B(x,r)|} \int_{B(x,r)} f(y)\, dy = f(x)$$

*and*

$$\lim_{r\to 0^+} \frac{1}{|\partial B(x,r)|} \int_{\partial B(x,r)} f(y)\, dS = f(x).$$

*Proof.* Let $x \in \Omega$ and $\varepsilon > 0$. The function $f$ is continuous at $x$, which means that there exists an $r_0 > 0$ such that $|f(x) - f(y)| < \varepsilon$ whenever $y \in B(x, r_0) \subset \Omega$. Now whenever $r \leq r_0$, it holds that

$$\begin{aligned}
\left| \frac{1}{|B(x,r)|} \int_{B(x,r)} f(y)\, dy - f(x) \right| &= \left| \frac{1}{|B(x,r)|} \int_{B(x,r)} f(y) - f(x)\, dy \right| \\
&\leq \frac{1}{|B(x,r)|} \int_{B(x,r)} |f(y) - f(x)|\, dy \\
&\leq \frac{1}{|B(x,r)|} \int_{B(x,r)} \varepsilon\, dy \\
&= \varepsilon.
\end{aligned}$$

Thus, by the definition of a limit, it holds that

$$\lim_{r\to 0^+} \frac{1}{|B(x,r)|} \int_{B(x,r)} f(y)\, dy = f(x).$$

The proof with the spheres $\partial B(x,r)$ is identical. □

### 2.3.2 Boundary Integral

This short section presents how to define integration over the boundary of a bounded domain $\Omega \subset \mathbb{R}^n$, $n \geq 2$, with Lipschitz boundary. We give the same definition that is given by e.g. Schwab [12] and Nečas [16].

Let $Q_d$ denote an $(n-1)$-dimensional hypercube centered at the origin and with side length $2d > 0$. By Definition 2.2, for every $x \in \partial\Omega$, there exists a Lipschitz function $g : \mathbb{R}^{n-1} \to \mathbb{R}$ and an open set

$$U = \left\{ y = \sum_{i=1}^{n} \tilde{y}_i \tilde{e}_i \in \mathbb{R}^n : \tilde{y}' = (\tilde{y}_1, \ldots, \tilde{y}_{n-1}) \in Q_d,\ |\tilde{y}_n - g(\tilde{y}')| < h \right\}$$

that covers a part of the boundary $\partial\Omega$. The collection of the sets $U$ for all $x \in \partial\Omega$ is an open cover of $\partial\Omega$. Since $\Omega$ is bounded, its boundary is compact. This means that the open cover has a finite subcover which we denote by $\{U_j\}_{j=1}^m$. The corresponding Lipschitz functions are denoted by $g_j$ for $1 \leq j \leq m$. Similarly, we denote $Q_{d_j}$ and $\{\tilde{e}_i^j\}_{i=1}^n$.

The functions $g_j$ are the graphs of $\partial\Omega \cap U_j$. Thus, it makes sense to define the integral of a function $f : \partial\Omega \to \mathbb{R}$ over $\partial\Omega \cap U_j$ as

$$\int_{\partial\Omega\cap U_j} f\, dS = \int_{Q_{d_j}} f\left(\sum_{i=1}^{n-1} \tilde{y}_i \tilde{e}_i^j + g_j(\tilde{y}')\tilde{e}_n^j\right)\sqrt{1 + |\nabla g_j|^2}\, d\tilde{y}', \tag{2.4}$$

which is essentially a line integral when $n = 2$ and a surface integral when $n = 3$. For simplicity, we assume in (2.4) that the local coordinate system $\{\tilde{e}_1, \dots, \tilde{e}_n\}$ is obtained from the standard basis with rotation and translation but with no scaling. Note that since $g_j$ is Lipschitz, the gradient $\nabla g_j$ exists almost everywhere and is essentially bounded by Rademacher's theorem [12].

To then define integration over the whole boundary, note that since $\{U_j\}_{j=1}^m$ is an open cover of $\partial\Omega$, there exists a partition of unity of $\partial\Omega$ subordinate to $\{U_j\}_{j=1}^m$. That is, there exist functions $\eta_j \in C_0^\infty(U_j)$, $1 \leq j \leq m$, such that

$$0 \leq \eta_j \leq 1 \quad \text{and} \quad \sum_{j=1}^m \eta_j(x) = 1 \text{ for all } x \in \partial\Omega.$$

The space $C_0^\infty(U_j)$ is the space of infinitely differentiable functions whose support is a compact subset of $U_j$. We now define the integral of $f : \partial\Omega \to \mathbb{R}$ over $\partial\Omega$ by

$$\int_{\partial\Omega} f\, dS = \sum_{j=1}^m \int_{\partial\Omega\cap U_j} \eta_j f\, dS, \tag{2.5}$$

where the right-hand side integrals are defined by (2.4). This definition does not depend on the cover $\{U_j\}_{j=1}^m$ or the partition of unity $\{\eta_j\}_{j=1}^m$ [16].

The normed spaces $L^s(\partial\Omega)$ for $1 \leq s \leq \infty$ are defined analogously to the spaces $L^s(\Omega)$. Two functions in $L^s(\partial\Omega)$ are again identified if they are equal everywhere in $\partial\Omega$ except possibly in a non-empty set with zero $(n-1)$-dimensional measure over $\partial\Omega$. Similarly, the definition of the space $L^\infty(\partial\Omega)$ uses an $(n-1)$-dimensional measure to define essentially bounded functions. We skip the full construction of a measure over $\partial\Omega$ and merely characterize the measure of a measurable set $\Gamma \subset \partial\Omega$ as the integral of its indicator function over $\partial\Omega$ via (2.5).

Hölder's inequality can be extended to the spaces $L^s(\partial\Omega)$, which we prove next.

**Theorem 2.8.** *Let $1 \leq s \leq \infty$ and $1 \leq s' \leq \infty$ be such that $1/s + 1/s' = 1$. Let $u \in L^s(\partial\Omega)$ and $v \in L^{s'}(\partial\Omega)$. Then $uv \in L^1(\partial\Omega)$ and*

$$\|uv\|_{L^1(\partial\Omega)} \leq \|u\|_{L^s(\partial\Omega)}\|v\|_{L^{s'}(\partial\Omega)}.$$

*Proof.* Assume first that $1 < s < \infty$ and $1 < s' < \infty$. Since $\eta_j \in C_0^\infty(U_j)$, we may extend it by zero to the whole $\mathbb{R}^n$. Then

$$\int_{\partial\Omega} |uv|\, dS = \sum_{j=1}^m \int_{\partial\Omega\cap U_j} \eta_j |uv|\, dS$$

$$
\begin{aligned}
&= \sum_{j=1}^{m} \int_{Q_{d_j}} \eta_j |uv| \sqrt{1 + |\nabla g_j|^2} \, d\tilde{y}' \\
&= \int_{\mathbb{R}^{n-1}} \sum_{j=1}^{n} \eta_j |uv| \sqrt{1 + |\nabla g_j|^2} \, d\tilde{y}' \\
&= \int_{\mathbb{R}^{n-1}} \sum_{j=1}^{n} \left( \eta_j^{\frac{1}{s}} |u| \left( 1 + |\nabla g_j|^2 \right)^{\frac{1}{2s}} \right) \left( \eta_j^{\frac{1}{s'}} |v| \left( 1 + |\nabla g_j|^2 \right)^{\frac{1}{2s'}} \right) d\tilde{y}' \\
&\leq \int_{\mathbb{R}^{n-1}} \left( \sum_{j=1}^{n} \eta_j |u|^s \sqrt{1 + |\nabla g_j|^2} \right)^{\frac{1}{s}} \left( \sum_{j=1}^{n} \eta_j |v|^{s'} \sqrt{1 + |\nabla g_j|^2} \right)^{\frac{1}{s'}} d\tilde{y}' \\
&\leq \left( \int_{\mathbb{R}^{n-1}} \sum_{j=1}^{n} \eta_j |u|^s \sqrt{1 + |\nabla g_j|^2} \, d\tilde{y}' \right)^{\frac{1}{s}} \left( \int_{\mathbb{R}^{n-1}} \sum_{j=1}^{n} \eta_j |v|^{s'} \sqrt{1 + |\nabla g_j|^2} \, d\tilde{y}' \right)^{\frac{1}{s'}} \\
&= \left( \int_{\partial\Omega} |u|^s \, dS \right)^{\frac{1}{s}} \left( \int_{\partial\Omega} |v|^{s'} \, dS \right)^{\frac{1}{s'}} \\
&= \|u\|_{L^s(\partial\Omega)} \|v\|_{L^{s'}(\partial\Omega)}.
\end{aligned}
$$

Both inequalities above follow from Hölder's inequality. For the second application of Hölder's inequality, we used the assumptions $u \in L^s(\partial\Omega)$ and $v \in L^{s'}(\partial\Omega)$ and the definition of a boundary integral as a regular Lebesgue integral over $\mathbb{R}^{n-1}$.

Let us then consider the case $s = \infty$ or $s' = \infty$. Without loss of generality, let $s = \infty$. Then clearly

$$
\int_{\partial\Omega} |uv| \, dS \leq \int_{\partial\Omega} \|u\|_{L^\infty(\partial\Omega)} |v| \, dS = \|u\|_{L^\infty(\partial\Omega)} \|v\|_{L^1(\partial\Omega)}.
$$

□

Since $\Omega$ is assumed to be bounded, $\partial\Omega$ has finite $(n-1)$-dimensional measure. Thus, with a proof more or less identical to the proof of Theorem 2.5, the imbedding $L^r(\partial\Omega) \subset L^s(\partial\Omega)$ holds when $1 \leq s \leq r \leq \infty$.

## 2.4 Sobolev Spaces

A Sobolev space is a Banach space consisting of functions where the functions along with their weak partial derivatives up to a given order belong to a given Lebesgue space. The Sobolev spaces are the vector spaces where we shall search for the solutions of the boundary value problems. The following review of the Sobolev spaces is based on [16] and [17].

### 2.4.1 Weak Derivatives

Let us first clarify the notation used for the spaces of classically differentiable functions on an open set $\Omega \subset \mathbb{R}^n$. The space of $k$ times continuously differentiable functions is denoted by $C^k(\Omega)$, and $C(\Omega) = C^0(\Omega)$ is the space of continuous functions. The intersection of $C^k(\Omega)$ for all $k \in \mathbb{N}_0$ is the space of infinitely differentiable functions $C^\infty(\Omega)$, and $C_0^\infty(\Omega)$ denotes the space of infinitely differentiable functions with compact support.

Let us then motivate weak differentiability with a classically differentiable function. Let $u \in C^1(\Omega)$. For all $\phi \in C_0^\infty(\Omega)$, integration by parts gives

$$\int_\Omega u \frac{\partial \phi}{\partial x_j}\, dx = -\int_\Omega \frac{\partial u}{\partial x_j} \phi \, dx \tag{2.6}$$

for all $1 \leq j \leq n$. There is no boundary term because $\phi$ vanishes near the boundary. The identity (2.6) essentially defines the meaning of a weak partial derivative: a function $v_j$ is a weak partial derivative of $u$ with respect to the variable $x_j$ if it satisfies

$$\int_\Omega u \frac{\partial \phi}{\partial x_j}\, dx = -\int_\Omega v_j \phi \, dx \tag{2.7}$$

for all $\phi \in C_0^\infty(\Omega)$.

Clearly, a classical derivative is also a weak derivative, but the reverse need not be true. The integrals in the identity (2.7) exist with the modest assumption that $u \in L^1_{\text{loc}}(\Omega)$ and $v_j \in L^1_{\text{loc}}(\Omega)$, where $L^1_{\text{loc}}(\Omega)$ is the space of functions that are integrable over all compact subsets of $\Omega$. Indeed, weak derivatives exist for a larger class of functions than just classically differentiable functions. If a weak partial derivative exists, it is unique up to sets of measure zero [17].

Extending the above to higher orders of weak differentiability is straightforward. For this, we use the standard multi-index notation. Let $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$ be a multi-index, and define $|\alpha| = \alpha_1 + \cdots + \alpha_n$. The elements of $\alpha$ correspond to the numbers of times a function is differentiated with respect to each variable. Now for a $u \in C^k(\Omega)$, $k \geq 1$, and for a multi-index $\alpha$ satisfying $|\alpha| \leq k$, we denote the $\alpha$th partial derivative of $u$ by

$$D^\alpha u = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}} u = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}} u.$$

Naturally, $D^{(0,\ldots,0)} u = u$. Successive application of integration by parts gives

$$\int_\Omega u D^\alpha \phi \, dx = (-1)^{|\alpha|} \int_\Omega D^\alpha u \phi \, dx$$

for all $\phi \in C_0^\infty(\Omega)$, and $v_\alpha = D^\alpha u$ is also the $\alpha$th weak partial derivative of $u$.

Let us synthesize the above discussion into a self-contained definition of a weak partial derivative of arbitrary order.

**Definition 2.18** (Weak partial derivative)**.** Let $\Omega \subset \mathbb{R}^n$ be open, $u \in L^1_{\text{loc}}(\Omega)$, and let $\alpha \in \mathbb{N}_0^n$ be a multi-index. If there exists a $v_\alpha \in L^1_{\text{loc}}(\Omega)$ such that

$$\int_\Omega u D^\alpha \phi \, dx = (-1)^{|\alpha|} \int_\Omega v_\alpha \phi \, dx$$

for all $\phi \in C_0^\infty(\Omega)$, then $v_\alpha$ is said to be the $\alpha$th weak partial derivative of $u$, and it is denoted by $D^\alpha u$.

Weak derivatives behave much like classical derivatives. For example, weak differentiation is commutative and linear:

$$D^\alpha(D^\beta u) = D^\beta(D^\alpha u) \quad \text{and} \quad D^\alpha(au + bv) = aD^\alpha u + bD^\alpha v, \quad a, b \in \mathbb{R}.$$

We tend to use the classical notation $\partial u/\partial x_j$ or $D_j u$ for weak partial derivatives as well, and this is extended in an obvious manner to other differential operators, such as the gradient $\nabla u$ and the Laplacian $\Delta u$.

We are now ready to define the Sobolev spaces.

**Definition 2.19** (Sobolev spaces)**.** Let $\Omega \subset \mathbb{R}^n$ be open. Let $k$ be a positive integer and $1 \le s \le \infty$. The Sobolev space $W^{k,s}(\Omega)$ is the normed space of functions $u \in L^s(\Omega)$ that have all weak partial derivatives of orders $0 \le |\alpha| \le k$, and the weak partial derivatives satisfy $D^\alpha u \in L^s(\Omega)$ for all $0 \le |\alpha| \le k$. The norm is given by

$$\begin{aligned} \|u\|_{W^{k,s}(\Omega)} &= \left( \sum_{0\le|\alpha|\le k} \|D^\alpha u\|^s_{L^s(\Omega)} \right)^{\frac{1}{s}} && \text{for } 1 \le s < \infty, \\ \|u\|_{W^{k,\infty}(\Omega)} &= \max_{0\le|\alpha|\le k} \|D^\alpha u\|_{L^\infty(\Omega)} && \text{for } s = \infty. \end{aligned}$$

Moreover, the space $W^{k,2}(\Omega)$, commonly denoted by $H^k(\Omega)$, is an inner product space with the inner product

$$\langle u, v\rangle_{H^k(\Omega)} = \sum_{0\le|\alpha|\le k} \langle D^\alpha u, D^\alpha v\rangle_{L^2(\Omega)} .$$

For $1 \le s < \infty$, it is also common to consider the $L^s$-norms of the weak derivatives of some fixed order $|\alpha| = k$:

$$|u|_{W^{k,s}(\Omega)} = \left( \sum_{|\alpha|=k} \|D^\alpha u\|^s_{L^s(\Omega)} \right)^{\frac{1}{s}} .$$

The above defines a seminorm in $W^{k,s}(\Omega)$.

The Sobolev spaces inherit many of their properties from the Lebesgue spaces. Once again, two functions in a Sobolev space are identified if they are equal almost everywhere in their domain. The spaces $W^{k,s}(\Omega)$ are Banach spaces for all $k$ and $s$, and $H^k(\Omega)$ is a Hilbert space for all $k$. When $k$ and $m$ are positive integers such that $k \le m$, it is obvious that the imbedding $W^{m,s}(\Omega) \subset W^{k,s}(\Omega)$ holds for all $1 \le s \le \infty$. When $\Omega$ is bounded, Theorem 2.5 can also be extended to the Sobolev spaces, which we prove next.

**Theorem 2.9.** *Assume that $\Omega$ is bounded. Let $k$ be a positive integer, and let $1 \le s \le r \le \infty$. Then there exists a constant $C > 0$ such that*

$$\|u\|_{W^{k,s}(\Omega)} \le C\|u\|_{W^{k,r}(\Omega)}$$

*for all $u \in W^{k,r}(\Omega)$. In other words, $W^{k,r}(\Omega) \subset W^{k,s}(\Omega)$.*

*Proof.* The case $s = r$ is trivial so assume that $s < r$. Let $u \in W^{k,r}(\Omega)$, and assume first that $r < \infty$. Then by Theorem 2.5, we get

$$\begin{aligned}
\|u\|^s_{W^{k,s}(\Omega)} &= \sum_{0\le|\alpha|\le k} \|D^\alpha u\|^s_{L^s(\Omega)} \\
&\le C_1 \sum_{0\le|\alpha|\le k} \|D^\alpha u\|^s_{L^r(\Omega)} \\
&= C_1 \sum_{0\le|\alpha|\le k} \|D^\alpha u\|^{r\frac{s}{r}}_{L^r(\Omega)} \qquad (2.8)
\end{aligned}$$

for some constant $C_1 > 0$ independent of $u$. The function $t \mapsto t^{s/r}$ is increasing on $[0, \infty)$, which implies that

$$\|D^\alpha u\|^{r\frac{s}{r}}_{L^r(\Omega)} \le \left( \sum_{0\le|\alpha|\le k} \|D^\alpha u\|^r_{L^r(\Omega)} \right)^{\frac{s}{r}} = \|u\|^s_{W^{k,r}(\Omega)}.$$

Inserting this into (2.8) gives

$$\|u\|^s_{W^{k,s}(\Omega)} \le C_1 \sum_{0\le|\alpha|\le k} \|u\|^s_{W^{k,r}(\Omega)} = C_1 C_2 \|u\|^s_{W^{k,r}(\Omega)},$$

where $C_2 = |\{\alpha \in \mathbb{N}_0^n : 0 \le |\alpha| \le k\}|$. Taking the $s$th root from both sides proves the claim for $r < \infty$.

Assume then that $r = \infty$. By Theorem 2.5 and the definition of $\|u\|_{W^{k,\infty}(\Omega)}$, we get

$$\begin{aligned}
\|u\|^s_{W^{k,s}(\Omega)} &= \sum_{0\le|\alpha|\le k} \|D^\alpha u\|^s_{L^s(\Omega)} \\
&\le C_1 \sum_{0\le|\alpha|\le k} \|D^\alpha u\|^s_{L^\infty(\Omega)} \\
&\le C_1 \sum_{0\le|\alpha|\le k} \|u\|^s_{W^{k,\infty}(\Omega)} \\
&= C_1 C_2 \|u\|^s_{W^{k,\infty}(\Omega)}.
\end{aligned}$$

Taking the $s$th root from both sides proves the claim for $r = \infty$. □

Functions in Sobolev spaces resemble classically differentiable functions, although they need not be continuous or bounded in the interior of their domain. This resemblance is reflected by the properties of weak derivatives but also by the fact that the space $C^\infty(\Omega) \cap W^{k,s}(\Omega)$ is dense in $W^{k,s}(\Omega)$ for all positive integers $k$ and real numbers $1 \le s < \infty$. In other words, for every $u \in W^{k,s}(\Omega)$, there exists a sequence of infinitely differentiable functions which converges to $u$ with respect to the norm $\|\cdot\|_{W^{k,s}(\Omega)}$. For a proof, see [17, Theorem 3.17 on p. 67]. This is an immensely useful result that enables one to extend many existing results for smooth functions to the Sobolev spaces by passing to a limit. Note, however, that the density result does not hold when $s = \infty$.

It is also common to consider the closure of the space $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_{W^{k,s}(\Omega)}$. The resulting subspace of $W^{k,s}(\Omega)$ is denoted by $W_0^{k,s}(\Omega)$ for a general $s$ and by $H_0^k(\Omega)$ for $s = 2$. A function in this space vanishes on the boundary $\partial\Omega$ in the trace sense. Boundary traces will be discussed soon. These types of Sobolev spaces are important in the study of partial differential equations with Dirichlet boundary conditions.

For a function in $W_0^{1,p}(\Omega)$, there exists a useful inequality between the $L^s$-norms of the function and its gradient. This is commonly known as Poincaré's inequality.

**Theorem 2.10** (Poincaré's inequality)**.** *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain, and let $1 \leq s < \infty$. Then there exists a constant $C > 0$ such that*

$$\|u\|_{L^s(\Omega)} \leq C|u|_{W^{1,s}(\Omega)}$$

*for all $u \in W_0^{1,s}(\Omega)$.*

For a proof, see [17, Theorem 6.30 on p. 183]. Theorem 2.10 is sometimes also called Friedrichs' inequality or Poincaré-Friedrichs inequality. We will later prove a variant of Poincaré's inequality for which $u$ does not necessarily belong to the space $W_0^{1,p}(\Omega)$.

### 2.4.2 Sobolev Imbeddings

We previously saw two types of imbeddings between Sobolev spaces $W^{k,s}(\Omega)$ where either $k$ or $s$ was fixed: when $k$ and $m$ are such that $k \leq m$, then $W^{m,s}(\Omega) \subset W^{k,s}(\Omega)$ for all $1 \leq s \leq \infty$, and when $\Omega$ is bounded, then $W^{k,r}(\Omega) \subset W^{k,s}(\Omega)$ whenever $1 \leq s \leq r \leq \infty$. When the boundary of $\Omega$ satisfies some additional regularity assumptions, further imbeddings exist for which both $k$ and $s$ can vary. Moreover, in some cases for some positive integer $j$, the space $W^{k,s}(\Omega)$ can be considered as a subset of $j$ times continuously differentiable functions in the sense that each equivalence class $u \in W^{k,s}(\Omega)$ contains a function that is $j$ times continuously differentiable. Such an imbedding is denoted like the other imbeddings, i.e. $W^{k,s}(\Omega) \subset C^j(\Omega)$. These additional imbeddings are given by the Sobolev imbedding theorem [17].

The Sobolev imbedding theorem also states that the imbeddings are continuous. This means that when we consider an imbedding of the form $X \subset Y$ between two normed spaces $X$ and $Y$ via the identity mapping $I : X \longrightarrow Y$, $Ix = x$, then the operator $I$ is continuous. Since the operator $I$ is linear, this is equivalent to $I$ being bounded by Theorem 2.2. For example, the imbedding $W^{m,s}(\Omega) \subset W^{k,s}(\Omega)$ for $k \leq m$ is obviously continuous since $\|u\|_{W^{k,s}(\Omega)} \leq \|u\|_{W^{m,s}(\Omega)}$ for all $u \in W^{m,s}(\Omega)$. By Theorem 2.9, the imbedding $W^{k,r}(\Omega) \subset W^{k,s}(\Omega)$ for $1 \leq s \leq r \leq \infty$ is also continuous. Similarly, an imbedding is said to be compact if the operator $I$ is compact, see Definition 2.15. Most Sobolev imbeddings are compact, and this result is known as the Rellich-Kondrachov theorem [17].

We now present a simplified combined version of the Sobolev imbedding theorem and the Rellich-Kondrachov theorem which contains the relevant imbeddings for our purposes. In their most general forms, both theorems consider both bounded and unbounded domains, different types of boundary regularity, and the results depend on the dimension $n$. We are only interested in the case $n = 2$ and bounded domains with Lipschitz boundaries. Note that below $C^k(\overline{\Omega})$ denotes the normed space of functions $u \in C^k(\Omega)$ for which $D^\alpha u$ is bounded and uniformly continuous on $\Omega$ for all $0 \le |\alpha| \le k$, and its norm is given by

$$\|u\|_{C^k(\overline{\Omega})} = \max_{0\le|\alpha|\le k} \sup_{x\in\Omega} |D^\alpha u(x)|.$$

As usual, we use the alias $C(\overline{\Omega}) = C^0(\overline{\Omega})$. The notation $C^k(\overline{\Omega})$ signifies that the function $u$ along with its partial derivatives of order $|\alpha| \le k$ can be uniquely continuously extended to the boundary $\partial\Omega$.

**Theorem 2.11.** *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary. Let $k \ge 1$ and $j \ge 0$ be integers, and let $1 \le s < \infty$ be a real number. Then the following imbeddings are continuous and compact.*

*(i) If $ks < 2$, then*

$$W^{j+k,s}(\Omega) \subset W^{j,r}(\Omega) \quad \text{if } 1 \le r < 2s/(2 - ks).$$

*(ii) If $ks = 2$, then*

$$W^{j+k,s}(\Omega) \subset W^{j,r}(\Omega) \quad \text{if } 1 \le r < \infty.$$

*(iii) If $ks > 2$, then*

$$W^{j+k,s}(\Omega) \subset W^{j,r}(\Omega) \quad \text{if } 1 \le r < \infty,$$
$$W^{j+k,s}(\Omega) \subset C^j(\overline{\Omega}).$$

*Note that $W^{0,s}(\Omega) = L^s(\Omega)$.*

For the proofs and complete versions of the above imbeddings, see [17, Theorem 4.12 on p. 85] for the Sobolev imbedding theorem and [17, Theorem 6.3 on p. 168] for the Rellich-Kondrachov theorem.

Let us immediately note one important consequence of Theorem 2.11 regarding the Dirac delta functional $\delta_{x_0}$. In Section 2.2.3, we defined $\delta_{x_0}$ as an element of the dual space $C(\overline{\Omega})'$ such that for a given $x_0 \in \Omega$ we have $\delta_{x_0}(u) = u(x_0)$ for all $u \in C(\overline{\Omega})$. If we now try to define $\delta_{x_0}$ similarly as an element of the dual space of an arbitrary Sobolev space $W^{k,s}(\Omega)$, we run into the subtlety that $W^{k,s}(\Omega)$ does not really consist of functions but equivalence classes of functions. For every possible value $c \in \mathbb{R}$, each such equivalence class contains a function $u$ such that $u(x_0) = c$. The problem then becomes to choose one of these functions to evaluate $\delta_{x_0}(u)$. When $W^{k,s}(\Omega) \subset C(\overline{\Omega})$, we can simply choose the continuous function and $\delta_{x_0}(u)$ becomes well-defined. Theorem 2.11 states when this is possible. For example, the imbedding $W^{1,s}(\Omega) \subset C(\overline{\Omega})$ is true when $s > 2$ and, thus, $\delta_{x_0} \in W^{1,s}(\Omega)'$. Note that the continuity of $\delta_{x_0}$ follows from the continuity of the imbedding:

$$|\delta_{x_0}(u)| = |u(x_0)| \le \|u\|_{C(\overline{\Omega})} \le C\|u\|_{W^{1,s}(\Omega)}$$

for all $u \in W^{1,s}(\Omega)$.

### 2.4.3 Boundary Traces

By definition, a boundary value problem requires the ability to consider, in some sense, the restriction of a function on the boundary of its domain. This is obviously possible in the usual pointwise sense for functions in $C(\overline{\Omega})$ and, thus by Theorem 2.11, for functions in e.g. the Sobolev space $W^{1,s}(\Omega)$ when $\Omega$ is a two-dimensional bounded domain with Lipschitz boundary and $s > 2$. But what about the case $s \leq 2$ for which pointwise evaluation does not necessarily make sense? Fortunately, for all $1 \leq s < \infty$, the restriction of a function $u \in W^{1,s}(\Omega)$ on the boundary $\partial\Omega$ can always be considered as a function in $L^r(\partial\Omega)$ for some $r$ in a way that is consistent with the usual pointwise restriction for functions in $C(\overline{\Omega})$. This is the message of the following trace theorem.

**Theorem 2.12** (Trace theorem)**.** *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary, and let $1 \leq s < \infty$. Then there exists a unique bounded linear operator*

$$T : W^{1,s}(\Omega) \to L^s(\partial\Omega)$$

*such that $Tu = u|_{\partial\Omega}$ for all $u \in C(\overline{\Omega}) \cap W^{1,s}(\Omega)$.*

For a proof, see [16, Theorem 4.2 on p. 79 and Theorem 4.6 on p. 81]. The operator $T$ is called the trace operator. The notation $u|_{\partial\Omega}$ naturally extends to traces as well. Similarly, $\|u\|_{L^s(\partial\Omega)}$ is shorthand for $\|Tu\|_{L^s(\partial\Omega)}$.

The trace operator is not surjective. Thus, if we wish to pick a function $g \in L^s(\partial\Omega)$ such that there exists a $u \in W^{1,s}(\Omega)$ for which $Tu = g$, then we need to require that $g$ belongs to the range of $T$, i.e. $g \in T(W^{1,s}(\Omega))$. The range of the trace operator can be characterized as a fractional-order Sobolev space on $\partial\Omega$, which we shall not consider any further. For more information, see [17] and [18].

With the trace concept, integration by parts can be extended to the Sobolev spaces.

**Theorem 2.13** (Integration by parts)**.** *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary. Let $1 < s < \infty$ and $1 < s' < \infty$ be such that $1/s + 1/s' \leq 3/2$, and let $u \in W^{1,s}(\Omega)$ and $v \in W^{1,s'}(\Omega)$. Then for $i = 1$ and $i = 2$, it holds that*

$$\int_\Omega u \frac{\partial v}{\partial x_i}\, dx = -\int_\Omega \frac{\partial u}{\partial x_i} v\, dx + \int_{\partial\Omega} u v n_i \, dS,$$

*where $n = (n_1, n_2)$ is the exterior unit normal to the boundary $\partial\Omega$.*

For a proof, see [16, Theorem 1.1 on p. 117]. Note that the exterior unit normal exists almost everywhere on $\partial\Omega$ [16, Lemma 4.2 on p. 83]. As a corollary to Theorem 2.13, let us prove the following Green's formula that will be useful later.

**Theorem 2.14** (Green's formula)**.** *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary. Let $1 < s < \infty$ and $1 < s' < \infty$ be such that $1/s + 1/s' \leq 3/2$, and let $u \in W^{2,s}(\Omega)$ and $v \in W^{1,s'}(\Omega)$. Then*

$$\int_\Omega \nabla u \cdot \nabla v \, dx = -\int_\Omega \Delta u v \, dx + \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, dS,$$

*where $\partial u/\partial n = \nabla u \cdot n$ is the directional derivative of $u$ in the direction of the exterior unit normal on $\partial\Omega$.*

*Proof.* Since $u \in W^{2,s}(\Omega)$, it holds that $\nabla u \in W^{1,s}(\Omega) \times W^{1,s}(\Omega)$. Now by the integration by parts formula in Theorem 2.13, we get

$$\begin{aligned}\int_\Omega \nabla u \cdot \nabla v \, dx &= \int_\Omega \frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1}\, dx + \int_\Omega \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2}\, dx \\ &= -\int_\Omega \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right) v \, dx + \int_{\partial\Omega} \left(\frac{\partial u}{\partial x_1} n_1 + \frac{\partial u}{\partial x_2} n_2\right) v \, dS \\ &= -\int_\Omega \Delta u v \, dx + \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, dS.\end{aligned}$$

□

# 3 Poisson's Equation in a Polygon

Assume that $\Omega \subset \mathbb{R}^2$ is a bounded polygonal domain. Let $\Gamma_j$ denote the $j$th linear boundary segment on $\partial\Omega$ for $j = 1, 2, \ldots, J$, where $J$ is the total number of boundary segments. That is,

$$\partial\Omega = \bigcup_{1 \le j \le J} \overline{\Gamma_j}.$$

The linear boundary segments are assumed to be analogous to one-dimensional open intervals, which is why the union is taken over the closures of the boundary segments. It is also assumed that $\Omega$ does not contain any slits. In other words, the angle between two boundary segments with a common vertex is never 0 nor $2\pi$. This implies that $\partial\Omega$ is Lipschitz according to Definition 2.2 [11].

Poisson's equation in $\Omega$ with prescribed boundary conditions is classically formulated as finding a function $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ such that

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ \quad u = g_j & \text{on } \Gamma_j, \quad j \in D \\ \dfrac{\partial u}{\partial n} = g_j & \text{on } \Gamma_j, \quad j \in N, \end{cases} \tag{3.1}$$

where the functions $f$ and $g_j$ for $j = 1, 2, \ldots, J$ constitute the data of the problem. On each linear boundary segment, either a Dirichlet or a Neumann boundary condition is imposed according to the index sets $D$ and $N$, respectively, that are disjoint subsets of the full index set $\{j \in \mathbb{N} : 1 \le j \le J\}$.

Of course, $\Omega$ need not be polygonal for the boundary value problem (3.1) to make sense. One could consider a general open set $\Omega \subset \mathbb{R}^n$, but, for the purposes of this thesis, the assumption that $\Omega$ is a polygon has important theoretical and practical implications. This is especially true in the light of the finite element method.

The unknown function $u$ in the problem (3.1) can be used to model quantities, such as chemical concentration, temperature or electric potential, over a physical domain $\Omega$ when the quantity is in equilibrium, that is, the quantity does not change over time [1]. The function $f$ is typically called the load term or the source term. For example, in the context of heat diffusion, it can correspond to a source of heat, and its unit is energy per unit volume and time. In the context of electrostatics, the function $f$ can correspond to the charge density over $\Omega$. A Dirichlet boundary condition means that the quantity is held fixed on that boundary segment, and a Neumann boundary condition corresponds to the flux on that boundary segment. For example, in the context of heat diffusion, setting $g_j = 0$ for some $j \in N$ means that the boundary segment $\Gamma_j$ is perfectly insulated.

A solution to the classically formulated problem (3.1) is accordingly said to be classical. However, proving the existence of classical solutions is typically difficult and may turn out to be impossible if the data are not at least continuous and the boundary $\partial\Omega$ smooth enough. Even when a classical solution does exist, approximating it via e.g. finite differences can become unwieldy in complex domains. This makes the classical problem unfit for many practical applications. Instead, the problem is typically formulated in a weak form, for which the existence and uniqueness of a solution, even for irregular data, follows rather easily based on abstract results from functional analysis. Then the weak solution can be approximated with the finite element method over virtually any polygonal domain $\Omega$.

Let us formulate the boundary value problem (3.1) in weak form. Let $f \in L^2(\Omega)$, $g_j \in T(H^2(\Omega))$ for $j \in D$ and $g_j \in T(H^1(\Omega))$ for $j \in N$, where $T$ is the trace operator in Theorem 2.12. Assume that $u \in H^2(\Omega)$ is a classical solution to the problem (3.1). Let $v \in H^1(\Omega)$ be such that $v|_{\Gamma_j} = 0$ for all $j \in D$. Multiplying both sides of Poisson's equation by $v$, integrating both sides over $\Omega$ and applying Green's formula from Theorem 2.14 gives

$$\int_\Omega \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, dS = \int_\Omega f v \, dx. \tag{3.2}$$

Using the fact that $v|_{\Gamma_j} = 0$ for all $j \in D$ and substituting in the Neumann boundary conditions, the boundary integral in (3.2) becomes

$$\begin{aligned} \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, dS &= \sum_{j\in D} \int_{\Gamma_j} \frac{\partial u}{\partial n} v \, dS + \sum_{j \in N} \int_{\Gamma_j} \frac{\partial u}{\partial n} v \, dS \\ &= \sum_{j\in N} \int_{\Gamma_j} g_j v \, dS. \end{aligned} \tag{3.3}$$

Combining (3.2) and (3.3) gives

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx + \sum_{j\in N} \int_{\Gamma_j} g_j v \, dS. \tag{3.4}$$

Note that the expression (3.4) also makes sense when only $u \in H^1(\Omega)$. This motivates the definition of a weak solution to the problem (3.1).

**Definition 3.1.** A function $u \in H^1(\Omega)$ is said to be a weak solution to the boundary value problem (3.1) if

(i) it satisfies the identity (3.4) for all $v \in H^1(\Omega)$ with $v|_{\Gamma_j} = 0$ for all $j \in D$,

(ii) $u|_{\Gamma_j} = g_j$ for all $j \in D$.

It is difficult to deduce the existence of a weak solution to the problem (3.1) when it is formulated in its most general form. Instead, it is more convenient to separately consider the cases $D \neq \emptyset$, i.e. the pure Dirichlet and mixed Dirichlet-Neumann problems, and $D = \emptyset$, i.e. the pure Neumann problem. For both cases, we consider a special instance of the problem whose solution will enable us to deduce the existence of a weak solution to the general problem as per Definition 3.1. Let us begin with the case $D \neq \emptyset$.

It will be easier to work with a problem that has homogeneous Dirichlet boundary conditions $g_j = 0$ for all $j \in D$. Thus, we would like to transform the problem (3.1) to an equivalent problem with such boundary conditions. To make this simple, we always assume that there exists a function $u_D \in H^2(\Omega)$ such that $u_D|_{\Gamma_j} = g_j$ for all $j \in D$. This implies that when $g_i$ and $g_j$ approach a common vertex, they approach the same value because $H^2(\Omega) \subset C(\overline{\Omega})$ by the Sobolev imbedding theorem (Theorem 2.11).

Let us then consider finding a weak solution $w \in H^1(\Omega)$ to the modified problem

$$\begin{cases} -\Delta w = f + \Delta u_D & \text{in } \Omega \\ w = 0 & \text{on } \Gamma_j, \quad j \in D \\ \dfrac{\partial w}{\partial n} = g_j - \dfrac{\partial u_D}{\partial n} & \text{on } \Gamma_j, \quad j \in N. \end{cases} \tag{3.5}$$

Note that since $u_D \in H^2(\Omega)$, it holds that $f + \Delta u_D \in L^2(\Omega)$, and since the normal vector $n$ is constant on each linear boundary segment, it clearly holds that $g_j - \partial u_D/\partial n \in T(H^1(\Omega))$ for all $j \in N$.

If $w \in H^1(\Omega)$ is a weak solution to the problem (3.5), then $u = w + u_D \in H^1(\Omega)$ is a weak solution to the original non-homogeneous problem (3.1). Namely, we easily see that $u|_{\Gamma_j} = g_j$ for all $j \in D$, and, by the definition of a weak solution and by Green's formula, we have

$$\begin{aligned} \int_\Omega \nabla u \cdot \nabla v \, dx &= \int_\Omega \nabla w \cdot \nabla v \, dx + \int_\Omega \nabla u_D \cdot \nabla v \, dx \\ &= \int_\Omega (f + \Delta u_D) v \, dx + \sum_{j \in N} \int_{\Gamma_j} \left( g_j - \frac{\partial u_D}{\partial n} \right) v \, dS \\ &\quad - \int_\Omega \Delta u_D v \, dx + \sum_{j \in N} \int_{\Gamma_j} \frac{\partial u_D}{\partial n} v \, dS \\ &= \int_\Omega f v \, dx + \sum_{j \in N} \int_{\Gamma_j} g_j v \, dS \end{aligned}$$

for all $v \in H^1(\Omega)$ with $v|_{\Gamma_j} = 0$ for all $j \in D$. This means that to solve the non-homogeneous problem (3.1), it suffices to only consider the solvability of the homogeneous problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma_j, \quad j \in D \\ \dfrac{\partial u}{\partial n} = g_j & \text{on } \Gamma_j, \quad j \in N. \end{cases} \tag{3.6}$$

Homogeneity brings symmetricity to the weak formulation. Namely, define a subspace $V_D$ of $H^1(\Omega)$ as

$$V_D = \{ v \in H^1(\Omega) : v|_{\Gamma_j} = 0 \text{ for all } j \in D \}.$$

Then the weak formulation of the problem (3.6) can be succinctly written as: Find a $u \in V_D$ such that it satisfies (3.4) for all $v \in V_D$. We consider the solvability of this problem shortly.

Let us then consider how to apply a similar transformation to the pure Neumann problem, i.e. $D = \emptyset$. By Definition 3.1, the weak formulation is already quite simple and, most importantly, symmetric: Find a $u \in H^1(\Omega)$ such that it satisfies (3.4) for all $v \in H^1(\Omega)$. However, there are a few subtleties. First, the data $f$ and $g_j$ must satisfy the following compatibility condition. Define $v \equiv 1$ in $\Omega$. Since $\Omega$ is bounded, clearly $v \in H^1(\Omega)$. Substituting $v$ into (3.4) gives

$$\int_\Omega f\,dx + \sum_{j\in N}\int_{\Gamma_j} g_j\,dS = 0. \tag{3.7}$$

Thus, $f$ and $g_j$ for all $j \in N$ must satisfy (3.7).

Second, if $u \in H^1(\Omega)$ is a weak solution to the pure Neumann problem, then $u + C \in H^1(\Omega)$ is also a weak solution for any constant $C \in \mathbb{R}$. In other words, a weak solution is never unique as per Definition 3.1. To fix a solution, it is typically searched from the subspace of $H^1(\Omega)$ whose elements have zero mean value over $\Omega$:

$$V_N = \left\{ v \in H^1(\Omega) : \int_\Omega v\,dx = 0 \right\}.$$

The weak formulation then becomes: Find a $u \in V_N$ such that it satisfies (3.4) for all $v \in H^1(\Omega)$. The symmetry is now lost, but, fortunately, if the compatibility condition (3.7) is satisfied and (3.4) holds for all $v \in V_N$, then (3.4) also holds for all $v \in H^1(\Omega)$. Let us prove this. Let $v \in H^1(\Omega)$. Then $v - \overline{v} \in V_N$, where

$$\overline{v} = \frac{1}{|\Omega|}\int_\Omega v\,dx.$$

Now

$$\begin{aligned}
\int_\Omega \nabla u\cdot\nabla v\,dx &= \int_\Omega \nabla u\cdot\nabla(v-\overline{v})\,dx \\
&= \int_\Omega f(v-\overline{v})\,dx + \sum_{j\in N}\int_{\Gamma_j} g_j(v-\overline{v})\,dS \\
&= \int_\Omega fv\,dx + \sum_{j\in N}\int_{\Gamma_j} g_j v\,dS - \overline{v}\left(\int_\Omega f\,dx + \sum_{j\in N}\int_{\Gamma_j} g_j\,dS\right) \\
&= \int_\Omega fv\,dx + \sum_{j\in N}\int_{\Gamma_j} g_j v\,dS,
\end{aligned}$$

where the last equality follows from the compatibility condition (3.7). Thus, in the case of a pure Neumann problem, we rather consider the weak formulation: Find a $u \in V_N$ such that it satisfies (3.4) for all $v \in V_N$.

The above weak forms of the boundary value problem (3.1) are the most commonly studied formulations, see e.g. [1], [11], [19] and [20]. Due to the symmetricity and the fact that $H^1(\Omega)$ is a Hilbert space, the existence and uniqueness of solutions can be shown relatively easily, as we will see. Moreover, in many cases the weak solution can be shown to belong to the space $H^2(\Omega)$. We refer to the weak formulations presented above as the classical weak formulations.

Let us now discuss how to formulate the problem (3.1) when the load is the Dirac delta, i.e. $f = \delta_{x_0}$ for some $x_0 \in \Omega$. We use the definition of $\delta_{x_0}$ provided in Section 2.4.2, that is, $\delta_{x_0}(v) = v(x_0)$ for all $v \in W^{1,s'}(\Omega)$ for any $s' > 2$. By the Sobolev imbedding theorem, $W^{1,s'}(\Omega) \subset C(\overline{\Omega})$ so the pointwise evaluation makes sense. The definition of $\delta_{x_0}$ already suggests that the problem (3.1) can be understood in some weak sense only. We can characterize it as a limit of the classical weak problems where $f \in L^2(\Omega)$. Let us do this in the context of electrostatics.

Assume that we would like to model the electric potential in $\Omega$ caused by a single point charge located at the point $x_0 \in \Omega$. The load $f$ corresponds to the charge density of the point charge which is infinite at $x_0$ and zero everywhere else in $\Omega$. Such a function is zero almost everywhere, which means that it would vanish in (3.4). This would then incorrectly correspond to the scenario where there are no charges in $\Omega$. Let us instead consider the point charge as the limit of small charged disks. Let $B(x_0, \varepsilon) \subset \Omega$ be a disk centered at the point $x_0$ with a small radius $\varepsilon > 0$. For simplicity, assume that the total charge over the disk is one so that the charge density over $\Omega$ is given by

$$f_\varepsilon = \frac{1}{|B(x_0,\varepsilon)|}\mathbb{1}_{B(x_0,\varepsilon)}.$$

Clearly, $f_\varepsilon \in L^2(\Omega)$. Assume that $u_\varepsilon \in H^1(\Omega)$ is a weak solution to the corresponding weak problem with the load $f_\varepsilon$. Choose now a $v \in W^{1,s'}(\Omega) \subset H^1(\Omega)$, $s' > 2$, according to the type of the boundary value problem, i.e. either $v \in V_D$ or $v \in V_N$, and substitute it into (3.4):

$$\begin{aligned}
\int_\Omega \nabla u_\varepsilon \cdot \nabla v \, dx &= \int_\Omega f_\varepsilon v \, dx + \sum_{j\in N} \int_{\Gamma_j} g_j v \, dS \\
&= \frac{1}{|B(x_0,\varepsilon)|} \int_{B(x_0,\varepsilon)} v \, dx + \sum_{j\in N} \int_{\Gamma_j} g_j v \, dS \\
&\xrightarrow{\varepsilon\to 0} \delta_{x_0}(v) + \sum_{j\in N} \int_{\Gamma_j} g_j v \, dS.
\end{aligned}$$

The limit follows from Theorem 2.7 and the definition of $\delta_{x_0}$. Thus, the Dirac delta can be thought to model the effect of a point charge as a limit.

Let us now define what we mean by a weak solution to the problem

$$\begin{cases}
-\Delta u = \delta_{x_0} & \text{in } \Omega \\
u = g_j & \text{on } \Gamma_j, \quad j \in D \\
\dfrac{\partial u}{\partial n} = g_j & \text{on } \Gamma_j, \quad j \in N,
\end{cases} \tag{3.8}$$

where $x_0 \in \Omega$ and the functions $g_j \in T(H^2(\Omega))$ for $j \in D$ and $g_j \in T(H^1(\Omega))$ for $j \in N$ are the same as before.

**Definition 3.2.** Let $1 < s < 2$ and $2 < s' < \infty$ be conjugate exponents, i.e. $1/s + 1/s' = 1$. A function $u \in W^{1,s}(\Omega)$ is said to be a weak solution to the boundary value problem (3.8) if

(i) it satisfies

$$\int_\Omega \nabla u \cdot \nabla v \, dx = v(x_0) + \sum_{j \in N} \int_{\Gamma_j} g_j v \, dS \tag{3.9}$$

for all $v \in W^{1,s'}(\Omega)$ with $v|_{\Gamma_j} = 0$ for all $j \in D$,

(ii) $u|_{\Gamma_j} = g_j$ for all $j \in D$.

An identical definition for the homogeneous Dirichlet problem can be found in [6] and [10]. The conjugate exponents are obviously needed for the expression (3.9) to make sense. It will not be necessary to consider the cases $D \neq \emptyset$ and $D = \emptyset$ separately, other than that the boundary data must again satisfy the compatibility condition when $D = \emptyset$. We will need to consider the solvability of the classical weak formulation first before we are able to find a weak solution to the general problem (3.8).

## 3.1 Solvability of the Classical Weak Formulation

As a reminder, the classical weak formulation of the boundary value problem (3.1) is the following. Let $f \in L^2(\Omega)$ and $g_j \in T(H^1(\Omega))$ for all $j \in N$. It suffices to only consider homogeneous Dirichlet boundary conditions $g_j = 0$ for all $j \in D$. When $D \neq \emptyset$, let

$$V_D = \{v \in H^1(\Omega) : v|_{\Gamma_j} = 0 \text{ for all } j \in D\}. \tag{3.10}$$

When $D = \emptyset$, let

$$V_N = \left\{ v \in H^1(\Omega) : \int_\Omega v \, dx = 0 \right\}. \tag{3.11}$$

We wish to find a unique function $u \in V_D$ (resp. $u \in V_N$) such that

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx + \sum_{j \in N} \int_{\Gamma_j} g_j v \, dS \tag{3.12}$$

for all $v \in V_D$ (resp. $v \in V_N$).

### 3.1.1 Existence and Uniqueness of Solutions

The identity (3.12) can be written in the form

$$a(u, v) = \varphi(v), \tag{3.13}$$

where $a : V \times V \to \mathbb{R}$ and $\varphi : V \to \mathbb{R}$ are defined by

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx \tag{3.14}$$

and

$$\varphi(v) = \int_\Omega f v \, dx + \sum_{j \in N} \int_{\Gamma_j} g_j v \, dS \tag{3.15}$$

for all $u, v \in V$, where either $V = V_D$ or $V = V_N$.

For an arbitrary vector space $V$, a mapping of the form $a : V \times V \to \mathbb{R}$ is said to be bilinear if both univariate mappings $u \mapsto a(u, v)$ and $v \mapsto a(u, v)$ are linear when the other argument is kept fixed. The mapping $a$ is said to be symmetric if $a(u, v) = a(v, u)$ for all $u, v \in V$. Clearly, (3.14) defines a symmetric bilinear mapping. The functional in (3.15) is also obviously linear.

When $V$ is a Hilbert space and the mappings $a$ and $\varphi$ are bilinear and linear with some additional assumptions, there exists an abstract result that asserts the existence and uniqueness of a vector $u \in V$ such that (3.13) holds for all $v \in V$. This is the message of the well-known Lax-Milgram theorem.

**Theorem 3.1** (Lax-Milgram Theorem)**.** *Let $V$ be a Hilbert space with the inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|\cdot\|$. Let $a : V \times V \to \mathbb{R}$ be a symmetric bilinear mapping that satisfies the following two properties.*

*(i) Boundedness: There exists a constant $C > 0$ such that*

$$|a(u, v)| \leq C\|u\|\|v\|$$

*for all $u, v \in V$.*

*(ii) Ellipticity: There exists a constant $\alpha > 0$ such that*

$$a(u, u) \geq \alpha\|u\|^2$$

*for all $u \in V$.*

*Then for any $\varphi \in V'$, there exists a unique $u \in V$ such that*

$$a(u, v) = \varphi(v)$$

*for all $v \in V$.*

*Proof.* From the boundedness and ellipticity it follows that $a(u, u) \geq 0$ for all $u \in V$ and $a(u, u) = 0$ if and only if $u = 0$. Since the mapping $a$ is also bilinear and symmetric, it is an inner product on the space $V$.

Let us consider the inner product space $(V, a(\cdot, \cdot))$ with the induced norm $\|\cdot\|_a$. By the boundedness of $a$, we have that

$$\|u\|_a^2 = a(u, u) \leq C\|u\|^2,$$

and by the ellipticity of $a$, we have that

$$\|u\|_a^2 = a(u, u) \geq \alpha\|u\|^2.$$

Combining these, we have that

$$\sqrt{\alpha}\|u\| \leq \|u\|_a \leq \sqrt{C}\|u\|$$

for all $u \in V$, which means that the norms $\|\cdot\|$ and $\|\cdot\|_a$ are equivalent. It easily follows from this that $(V, a(\cdot,\cdot))$ is also a Hilbert space, and the dual spaces of $(V, \langle\cdot,\cdot\rangle)$ and $(V, a(\cdot,\cdot))$ are the same.

Finally, let $\varphi$ be a functional that belongs to the dual space of $(V, \langle\cdot,\cdot\rangle)$. Then $\varphi$ also belongs to the dual space of $(V, a(\cdot,\cdot))$, and, by the Riesz representation theorem (Theorem 2.3), there exists a unique $u \in V$ such that $\varphi(v) = a(u, v)$ for all $v \in V$, which is what we wanted to show. □

The assumption that the mapping $a$ is symmetric is not necessary, but it simplifies the proof quite a lot, and we will not be dealing with non-symmetric mappings anyway. For a proof without the symmetricity assumption, see [1, Theorem 1 on p. 297].

The Lax-Milgram theorem now provides a means to prove the existence of a unique solution to the classical weak formulation of Poisson's equation for which either $V = V_D$ or $V = V_N$ and the mappings $a$ and $\varphi$ are given by (3.14) and (3.15), respectively. To fulfil the assumptions of the Lax-Milgram theorem, we need to show the following.

(i) The subspaces $V_D$ and $V_N$ of $H^1(\Omega)$ are Hilbert spaces.

(ii) The symmetric bilinear mapping (3.14) is bounded and elliptic.

(iii) The linear functional (3.15) is continuous, i.e. bounded, so it belongs to $V'$.

To prove the first assertion, it is enough to show that the subspaces $V_D$ and $V_N$ are closed.

**Theorem 3.2.** *Let the subspaces $V_D \subset H^1(\Omega)$ and $V_N \subset H^1(\Omega)$ be as in (3.10) and (3.11). Then both $V_D$ and $V_N$ are Hilbert spaces.*

*Proof.* It suffices to show that $V_D$ and $V_N$ are closed subspaces. Consider first the subspace $V_D$. Let $u \in H^1(\Omega)$ be a closure point of $V_D$, that is, there exists a sequence $(u_i)_{i=1}^{\infty}$ in $V_D$ that converges to $u$ in the $H^1(\Omega)$-norm. Now by the trace theorem (Theorem 2.12) and the fact that $u_i|_{\Gamma_j} = 0$ for all $j \in D$, we get for all $j \in D$ that

$$\begin{aligned}\|u\|_{L^2(\Gamma_j)} &= \|u - u_i\|_{L^2(\Gamma_j)} \\ &\leq \|u - u_i\|_{L^2(\partial\Omega)} \\ &\leq C\|u - u_i\|_{H^1(\Omega)} \\ &\xrightarrow{i\to\infty} 0.\end{aligned}$$

Thus, $\|u\|_{L^2(\Gamma_j)} = 0$, which implies that $u|_{\Gamma_j} = 0$ and $u \in V_D$. We conclude that $V_D$ is closed.

Let us then consider the subspace $V_N$. Let $u \in H^1(\Omega)$ be a closure point of $V_N$, that is, there exists a sequence $(u_i)_{i=1}^\infty$ in $V_N$ that converges to $u$ in the $H^1(\Omega)$-norm. We may estimate the integral of $u$ by

$$\begin{aligned}\left|\int_\Omega u \, dx\right| &= \left|\int_\Omega u - u_i \, dx\right| \\ &\leq \|u - u_i\|_{L^1(\Omega)} \\ &\leq C\|u - u_i\|_{L^2(\Omega)} \\ &\leq C\|u - u_i\|_{H^1(\Omega)} \\ &\xrightarrow{i\to\infty} 0,\end{aligned}$$

where the second inequality follows from Theorem 2.5. This means that the integral of $u$ is zero, i.e. $u \in V_N$, which implies that $V_N$ is closed. □

Proving the ellipticity of the bilinear mapping $a$ is by far the most demanding step. For that, we need the following result which is a variant of Poincaré's inequality in Theorem 2.10.

**Theorem 3.3.** *Let $\Omega \subset \mathbb{R}^n$ be a bounded polygonal domain. Then there exists a constant $C > 0$ such that*

$$\|u\|_{L^2(\Omega)} \leq C|u|_{H^1(\Omega)}$$

*for all $u \in V$, where either $V = V_D$ or $V = V_N$.*

*Proof.* We proceed via proof by contradiction. Assume that the claim is not true for any constant $C > 0$. Then there exists a sequence $(u_i)_{i=1}^\infty$ in $V$ such that

$$\|u_i\|_{L^2(\Omega)} > i|u_i|_{H^1(\Omega)}, \quad i = 1, 2, \ldots .$$

Dividing both sides by $\|u_i\|_{L^2(\Omega)}$ and then by $i$ yields

$$|v_i|_{H^1(\Omega)} < \frac{1}{i}, \quad i = 1, 2, \ldots,$$

where $v_i = u_i/\|u_i\|_{L^2(\Omega)}$. Clearly, $\|v_i\|_{L^2(\Omega)} = 1$ for all $i$. Note that the sequence $(v_i)_{i=1}^\infty$ is bounded in $V \subset H^1(\Omega)$:

$$\begin{aligned}\|v_i\|_{H^1(\Omega)}^2 &= \|v_i\|_{L^2(\Omega)}^2 + |v_i|_{H^1(\Omega)}^2 \\ &\leq 1 + \frac{1}{i^2} \\ &\leq 2, \qquad i = 1, 2, \ldots .\end{aligned}$$

By Theorem 2.11, the imbedding $H^1(\Omega) \subset L^2(\Omega)$ is compact. Thus, there exists a subsequence of the sequence $(v_i)_{i=1}^\infty$ that converges to some $v \in L^2(\Omega)$ with respect to the norm $\|\cdot\|_{L^2(\Omega)}$. Without loss of generality, denote this subsequence by $(v_i)_{i=1}^\infty$ as well.

Let us then show that $v \in H^1(\Omega)$ with $\nabla v = 0$. Let $\phi \in C_0^\infty(\Omega)$. First note that

$$\begin{aligned}\left|\int_\Omega v D_k \phi \, dx - \int_\Omega v_i D_k \phi \, dx\right| &\leq \int_\Omega |(v - v_i) D_k \phi| \, dx \\ &\leq \|v - v_i\|_{L^2(\Omega)} \|D_k \phi\|_{L^2(\Omega)} \\ &\xrightarrow{i \to \infty} 0\end{aligned}$$

for all $k = 1, 2, \ldots, n$, where the second inequality follows from Hölder's inequality. Using the above limit and the definition of a weak partial derivative, we get

$$\begin{aligned}\left|\int_\Omega v D_k \phi \, dx\right| &= \left|\lim_{i \to \infty} \int_\Omega v_i D_k \phi \, dx\right| \\ &= \lim_{i \to \infty} \left|\int_\Omega D_k v_i \phi \, dx\right| \\ &\leq \limsup_{i \to \infty} \int_\Omega |D_k v_i \phi| \, dx \\ &\leq \limsup_{i \to \infty} \|D_k v_i\|_{L^2(\Omega)} \|\phi\|_{L^2(\Omega)} \\ &\leq \limsup_{i \to \infty} |v_i|_{H^1(\Omega)} \|\phi\|_{L^2(\Omega)} \\ &\leq \limsup_{i \to \infty} \frac{1}{i} \|\phi\|_{L^2(\Omega)} \\ &= 0.\end{aligned}$$

Note above also the use of Hölder's inequality and the bound $|v_i|_{H^1(\Omega)} < 1/i$ for all $i$. The above implies that

$$\int_\Omega v D_k \phi \, dx = 0$$

for all $k = 1, 2, \ldots, n$, which then implies that $D_k v = 0$ for all $k = 1, 2, \ldots, n$. In other words, $\nabla u = 0$.

Since $\Omega$ is connected, the fact that $\nabla v = 0$ implies that $v$ is constant almost everywhere in $\Omega$. Say $v = c$. We show next that $c$ must be equal to zero. Assume first that $V = V_D$. Clearly, $v|_{\partial\Omega} = c$. Recall that $v_i \in V_D$ and, thus, $v_i|_{\Gamma_j} = 0$ for all $j \in D$. Now by the trace theorem, we get for any $j \in D$ that

$$\begin{aligned}\|v\|_{L^2(\Gamma_j)}^2 &= \|v - v_i\|_{L^2(\Gamma_j)}^2 \\ &\leq \|v - v_i\|_{L^2(\partial\Omega)}^2 \\ &\leq C\|v - v_i\|_{H^1(\Omega)}^2 \\ &= C\left(\|v - v_i\|_{L^2(\Omega)}^2 + |v_i|_{H^1(\Omega)}^2\right) \\ &\leq C\left(\|v - v_i\|_{L^2(\Omega)}^2 + \frac{1}{i^2}\right) \\ &\xrightarrow{i \to \infty} 0.\end{aligned}$$

Thus,

$$0 = \|v\|_{L^2(\Gamma_j)} = \|c\|_{L^2(\Gamma_j)} = |c||\Gamma_j|,$$

where $|\Gamma_j|$ is the $(n-1)$-dimensional measure of $\Gamma_j$. Since this measure is non-zero, the constant $c$ must be zero when $V = V_D$.

Assume then that $V = V_N$. By Theorem 2.5, we get

$$\begin{aligned} \left|\int_\Omega v\,dx\right| &= \left|\int_\Omega v - v_i\,dx\right| \\ &\leq \|v - v_i\|_{L^1(\Omega)} \\ &\leq C\|v - v_i\|_{L^2(\Omega)} \\ &\xrightarrow{i\to\infty} 0. \end{aligned}$$

Thus,

$$0 = \int_\Omega v\,dx = c|\Omega|,$$

and since $|\Omega| > 0$, the constant $c$ must be zero.

However, the fact that $\|v_i\|_{L^2(\Omega)} = 1$ for all $i$ implies that $\|v\|_{L^2(\Omega)} = 1$ as follows. By the inverse triangle inequality, we get

$$\left|\|v\|_{L^2(\Omega)} - \|v_i\|_{L^2(\Omega)}\right| \leq \|v - v_i\|_{L^2(\Omega)} \xrightarrow{i\to\infty} 0,$$

from which we get that

$$\|v\|_{L^2(\Omega)} = \lim_{i\to\infty} \|v_i\|_{L^2(\Omega)} = 1.$$

We have now arrived at a contradiction between the results $v = 0$ almost everywhere in $\Omega$ and $\|v\|_{L^2(\Omega)} = 1$. The initial claim must thus hold. □

We may now prove that the mappings $a$ and $\varphi$ in the weak formulation of Poisson's equation satisfy the assumptions of the Lax-Milgram theorem.

**Theorem 3.4.** *Let either $V = V_D$ or $V = V_N$. The bilinear mapping $a : V \times V \to \mathbb{R}$ in (3.14) is bounded and elliptic, and the linear functional $\varphi : V \to \mathbb{R}$ in (3.15) is bounded.*

*Proof.* By the Cauchy-Schwarz and Hölder's inequalities, we have

$$\begin{aligned} |a(u,v)| &\leq \int_\Omega |\nabla u \cdot \nabla v|\,dx \\ &\leq \int_\Omega |\nabla u||\nabla v|\,dx \\ &\leq \|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} \\ &\leq \|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} \end{aligned}$$

for all $u, v \in V$. Thus, $a$ is bounded.

By Theorem 3.3, we have

$$\begin{aligned}
a(u,u) &= \int_\Omega \nabla u \cdot \nabla u \, dx \\
&= |u|^2_{H^1(\Omega)} \\
&= \frac{1}{2}|u|^2_{H^1(\Omega)} + \frac{1}{2}|u|^2_{H^1(\Omega)} \\
&\geq \frac{1}{2C}\|u\|^2_{L^2(\Omega)} + \frac{1}{2}|u|^2_{H^1(\Omega)} \\
&\geq \min\left\{\frac{1}{2C}, \frac{1}{2}\right\}\left(\|u\|^2_{L^2(\Omega)} + |u|^2_{H^1(\Omega)}\right) \\
&= \min\left\{\frac{1}{2C}, \frac{1}{2}\right\}\|u\|^2_{H^1(\Omega)} \\
&= \alpha\|u\|^2_{H^1(\Omega)}
\end{aligned}$$

for some constant $\alpha > 0$ and for all $u \in V$. Thus, $a$ is elliptic.

Finally, by Hölder's inequality and the trace theorem, we have

$$\begin{aligned}
|\varphi(v)| &\leq \int_\Omega |fv| \, dx + \sum_{j \in N} \int_{\Gamma_j} |g_j v| \, dS \\
&\leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \sum_{j \in N} \|g_j\|_{L^2(\partial\Omega)}\|v\|_{L^2(\partial\Omega)} \\
&\leq \|f\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)} + \sum_{j \in N} \|g_j\|_{L^2(\partial\Omega)} C\|v\|_{H^1(\Omega)} \\
&= \left(\|f\|_{L^2(\Omega)} + C\sum_{j \in N} \|g_j\|_{L^2(\partial\Omega)}\right)\|v\|_{H^1(\Omega)}
\end{aligned}$$

for all $v \in V$. Thus, $\varphi$ is bounded and $\varphi \in V'$. □

The existence of a unique solution to the classical weak formulation of Poisson's equation with homogeneous Dirichlet boundary conditions or with the zero mean value requirement follows now directly from the Lax-Milgram theorem. For completeness, let us also consider the existence of weak solutions as per Definition 3.1.

**Theorem 3.5.** *Let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain. Let $f \in L^2(\Omega)$, $g_j \in T(H^2(\Omega))$ for all $j \in D$ and $g_j \in T(H^1(\Omega))$ for all $j \in N$. Assume that there exists a function $g_D \in H^2(\Omega)$ such that $g_D|_{\Gamma_j} = g_j$ for all $j \in D$, and if $D = \emptyset$, assume that $f$ and $g_j$ satisfy the compatibility condition. Then the boundary value problem*

$$\begin{cases}
-\Delta u = f & \text{in } \Omega \\
u = g_j & \text{on } \Gamma_j, \quad j \in D \\
\dfrac{\partial u}{\partial n} = g_j & \text{on } \Gamma_j, \quad j \in N
\end{cases}$$

*has a weak solution $u \in H^1(\Omega)$. When $D \neq \emptyset$, $u$ is unique. When $D = \emptyset$, $u$ is unique up to an additive constant.*

*Proof.* When $D \neq \emptyset$, the existence of a weak solution follows directly from the Lax-Milgram theorem, i.e. Theorem 3.1, and the discussion at the beginning of this section. If $u$ and $v$ are weak solutions, then $u - v$ is clearly a weak solution to the problem

$$\begin{cases} -\Delta w = 0 & \text{in } \Omega \\ u = 0 & \text{on } \Gamma_j, \quad j \in D \\ \dfrac{\partial u}{\partial n} = 0 & \text{on } \Gamma_j, \quad j \in N. \end{cases} \tag{3.16}$$

Clearly, $w = 0$ is a weak solution to the problem (3.16), and it is unique by the Lax-Milgram theorem. Thus, $u - v = 0$, which implies that the solution is unique.

When $D = \emptyset$, the Lax-Milgram theorem implies the existence of a unique weak solution in the space $V_N$. If now $u$ and $v$ are two weak solutions, not necessarily in the space $V_N$, then $u - v$ is again a weak solution to the homogeneous problem (3.16). So is the function $u - v - (\overline{u} - \overline{v}) \in V_N$, where

$$\overline{u} = \frac{1}{|\Omega|} \int_\Omega u \, dx.$$

By the Lax-Milgram theorem, $u - v - (\overline{u} - \overline{v}) = 0$, i.e. $u = v + C$ with the constant $C = \overline{u} - \overline{v}$, which implies that a weak solution is unique up to an additive constant. □

The Lax-Milgram theorem is a very general existence result. It could be used to show the existence and uniqueness of solutions to other second-order elliptic boundary value problems as well and in higher dimensions than $n = 2$. One could also consider more general domains with Lipschitz boundaries with little modifications. For additional information, see [1, Chapter 6].

### 3.1.2 Regularity of the Weak Solutions

Poisson's equation contains the second-order differential operator $\Delta$, which motivates the question whether a weak solution $u \in H^1(\Omega)$ actually belongs to the space $H^2(\Omega)$. This is of particular importance in finite element analysis as typical convergence results rely on the answer being yes. Whether the answer is indeed yes depends on the domain $\Omega$. For a polygonal domain, we need the following additional assumptions.

**Assumption 3.1.** Let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain. The linear boundary segments $\Gamma_j$, $j = 1, 2, \ldots, J$, that constitute the boundary $\partial\Omega$ are arranged so that $\Gamma_j$ is followed by $\Gamma_{j+1}$. For $j = J$, set $\Gamma_{J+1} = \Gamma_1$. Denote the angle between the segments $\Gamma_j$ and $\Gamma_{j+1}$ by $\theta_j$. The angle is the one inside $\Omega$. Assume that the angles $\theta_j$ satisfy the following assumptions.

(i) $0 < \theta_j < \pi$ for all $j = 1, 2, \ldots, J$, i.e. $\Omega$ is convex.

(i) For all $j = 1, 2, \ldots, J$, if $\Gamma_j$ has a prescribed Dirichlet boundary condition and $\Gamma_{j+1}$ has a prescribed Neumann boundary condition, or the other way around, then $0 < \theta_j < \pi/2$.

Now we have the following regularity result from [11, Theorem 4.4.4.13 on p. 245] and [21, Theorem 1].

**Theorem 3.6.** *Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain that satisfies Assumption 3.1. Then the weak solutions in Theorem 3.5 belong to the space $H^2(\Omega)$.*

When $u \in H^2(\Omega)$, then $\Delta u \in L^2(\Omega)$ and $\partial u/\partial n$ on $\Gamma_j$ exists in the sense of traces for all $j \in N$. It turns out that when the weak solution belongs to the space $H^2(\Omega)$, Poisson's equation holds in the sense of $L^2$ and the Neumann boundary conditions hold in the sense of traces. Note that the Dirichlet boundary conditions already hold in the sense of traces because it is embedded into the solution space.

**Theorem 3.7.** *Let $u \in H^1(\Omega)$ be a weak solution provided by Theorem 3.5. Assume that $u \in H^2(\Omega)$. Then*

*(i)* $-\Delta u = f$,

*(ii)* $\frac{\partial u}{\partial n}|_{\Gamma_j} = g_j$ *for all* $j \in N$.

For a proof, see for example [22, Proposition 5.1.9 on p. 131]. Let us finish off the discussion on the classical weak formulation of Poisson's equation by proving an a priori $H^2$-norm estimate for the weak solution with homogeneous boundary values $g_j = 0$ for all $j = 1, 2, \ldots, J$. This estimate will be useful later.

**Theorem 3.8.** *By Theorem 3.5, let $u \in H^1(\Omega)$ be the unique weak solution to the boundary value problem*

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma_j, \quad j \in D \\ \dfrac{\partial u}{\partial n} = 0 & \text{on } \Gamma_j, \quad j \in N. \end{cases}$$

*When $D = \emptyset$, the uniqueness is enforced by requiring that $\int_\Omega u\,dx = 0$. Assume that $u \in H^2(\Omega)$. Then there exists a constant $C > 0$ independent of $u$ and $f$ such that*

$$\|u\|_{H^2(\Omega)} \le C\|f\|_{L^2(\Omega)}.$$

*Proof.* Note that $u \in V_D$ or $u \in V_N$ depending on the type of the boundary value problem. Theorem 3.3 now implies that

$$\begin{aligned} \|u\|_{H^2(\Omega)}^2 &= \|u\|_{L^2(\Omega)}^2 + |u|_{H^1(\Omega)}^2 + |u|_{H^2(\Omega)}^2 \\ &\le C^2|u|_{H^1(\Omega)}^2 + |u|_{H^1(\Omega)}^2 + |u|_{H^2(\Omega)}^2 \\ &= (C^2+1)|u|_{H^1(\Omega)}^2 + |u|_{H^2(\Omega)}^2, \end{aligned} \tag{3.17}$$

where $C > 0$ is some constant independent of $u$ and $f$.

Consider next the seminorm $|u|_{H^1(\Omega)}$. By the definition of a weak solution, Hölder's inequality and Theorem 3.3, we get

$$|u|_{H^1(\Omega)}^2 = \int_\Omega \nabla u \cdot \nabla u\,dx$$

$$\begin{aligned}&= \int_\Omega f u \, dx \\ &\leq \|f\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)} \\ &\leq \|f\|_{L^2(\Omega)} C |u|_{H^1(\Omega)},\end{aligned}$$

where $C$ is the same constant as in (3.17). Dividing both sides by $|u|_{H^1(\Omega)}$ gives

$$|u|_{H^1(\Omega)} \leq C \|f\|_{L^2(\Omega)}. \tag{3.18}$$

If $|u|_{H^1(\Omega)} = 0$, this inequality is also obviously true.

By [11, Proof of Theorem 4.3.1.4 on p. 199], it holds that

$$|u|_{H^2(\Omega)} \leq \|\Delta u\|_{L^2(\Omega)},$$

which combined with the result $-\Delta u = f$ from Theorem 3.7 yields

$$|u|_{H^2(\Omega)} \leq \|f\|_{L^2(\Omega)}. \tag{3.19}$$

Substituting now (3.18) and (3.19) into (3.17) gives

$$\begin{aligned}\|u\|^2_{H^2(\Omega)} &\leq (C^2 + 1) C^2 \|f\|^2_{L^2(\Omega)} + \|f\|^2_{L^2(\Omega)} \\ &= (C^4 + C^2 + 1) \|f\|^2_{L^2(\Omega)}\end{aligned}$$

Taking the square root from both sides finishes the proof. □

## 3.2 Solvability of Poisson's Equation with a Dirac Delta Load

Let us recall what we mean by a weak solution to the boundary value problem

$$\begin{cases} -\Delta u = \delta_{x_0} & \text{in } \Omega \\ u = g_j & \text{on } \Gamma_j, \quad j \in D \\ \dfrac{\partial u}{\partial n} = g_j & \text{on } \Gamma_j, \quad j \in N, \end{cases} \tag{3.20}$$

where the boundary data $g_j$ for $j = 1, 2, \ldots, J$ are the same as before. Let $1 < s < 2$. A function $u \in W^{1,s}(\Omega)$ is a weak solution if $u|_{\Gamma_j} = g_j$ for all $j \in D$, and it satisfies

$$\int_\Omega \nabla u \cdot \nabla v \, dx = v(x_0) + \sum_{j \in N} \int_{\Gamma_j} g_j v \, dS$$

for all $v \in W^{1,s'}(\Omega)$ with $v|_{\Gamma_j} = 0$ for all $j \in D$, where $2 < s' < \infty$ is the conjugate exponent of $s$.

The Lax-Milgram theorem cannot now be used to deduce the existence of a unique solution because the problem is not formulated in the Hilbert space $H^1(\Omega)$. In particular, $\delta_{x_0} \notin H^1(\Omega)'$ as there is no obvious well-defined meaning to the expression $\delta_{x_0}(v) = v(x_0)$ for an arbitrary $v \in H^1(\Omega)$. Fortunately, the problem does have a solution, and when $\Omega$ satisfies Assumption 3.1, it is possible to show that a solution is also unique. Casas [6] proves the existence and uniqueness of a solution to the pure Dirichlet problem with homogeneous boundary values in convex polygonal domains. The proof by Casas is an abstract existence proof that uses results from functional analysis. An outline for a somewhat more constructive proof is given by Araya et al. [10] to the same problem that Casas considers. Namely, there exists a fundamental solution to a specific instance of the problem (3.20) that arises from the classical theory of partial differential equations. This fundamental solution is commonly called Green's function, and it can be used to deduce the existence of a weak solution to the general problem (3.20).

### 3.2.1 Green's Function

Consider Poisson's equation in free space:

$$-\Delta u = f \quad \text{in } \mathbb{R}^2. \tag{3.21}$$

When $f \in C_0^2(\mathbb{R}^2)$, the problem (3.21) has a classical solution $u \in C^2(\mathbb{R}^2)$ with the explicit formula

$$u(x) = \int_{\mathbb{R}^2} \Phi(x-y) f(y)\, dy,$$

where $\Phi$ is the fundamental solution of Laplace's equation, that is, the equation $\Delta u = 0$ [1, Theorem 1 on p. 23]. The fundamental solution $\Phi$ is given by

$$\Phi(x) = -\frac{1}{2\pi} \ln|x|$$

for $x \in \mathbb{R}^2 \setminus \{0\}$. By a direct calculation,

$$\begin{aligned}\Delta\Phi(x) &= \frac{\partial^2\Phi}{\partial x_1^2}(x) + \frac{\partial^2\Phi}{\partial x_2^2}(x)\\ &= -\frac{1}{2\pi}\frac{x_2^2 - x_1^2}{|x|^4} - \frac{1}{2\pi}\frac{x_1^2 - x_2^2}{|x|^4}\\ &= 0\end{aligned}$$

for all $x \in \mathbb{R}^2 \setminus \{0\}$.

We may shift the singularity at the origin to an arbitrary point $x_0 \in \mathbb{R}^2$ and define

$$\Phi_{x_0}(x) = -\frac{1}{2\pi} \ln|x - x_0| \tag{3.22}$$

for $x \in \mathbb{R}^2 \setminus \{x_0\}$. Clearly, $\Delta\Phi_{x_0}(x) = 0$ for all $x \in \mathbb{R}^2 \setminus \{x_0\}$. Set $\Phi_{x_0}(x_0) = 0$. The function $\Phi_{x_0}$ turns out to solve the problem (3.20) when the boundary values are set accordingly. This is illustrated by the following theorem.

**Theorem 3.9.** *Let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain. Let $x_0 \in \Omega$, and define $\Phi_{x_0}$ by (3.22). Let $1 < s < 2$. Then $\Phi_{x_0} \in W^{1,s}(\Omega)$, and $\Phi_{x_0}$ is a weak solution to the boundary value problem*

$$\begin{cases} -\Delta u = \delta_{x_0} & \text{in } \Omega \\ \quad u = \Phi_{x_0} & \text{on } \Gamma_j, \quad j \in D \\ \dfrac{\partial u}{\partial n} = \dfrac{\partial \Phi_{x_0}}{\partial n} & \text{on } \Gamma_j, \quad j \in N. \end{cases} \tag{3.23}$$

*Proof.* The gradient of $\Phi_{x_0}$ is given by

$$\nabla \Phi_{x_0}(x) = -\frac{1}{2\pi} \frac{x - x_0}{|x - x_0|^2}$$

for $x \in \mathbb{R}^2 \setminus \{x_0\}$. Set $\nabla \Phi_{x_0}(x_0) = 0$. Let us show that $\Phi_{x_0} \in L^s(\Omega)$ and $\nabla \Phi_{x_0} \in L^s(\Omega) \times L^s(\Omega)$. Since $\Omega$ is bounded, let $B(x_0, R) \subset \mathbb{R}^2$ be a ball that is centered at $x_0$ with radius $R > 1$ such that $\Omega \subset B(x_0, R)$. Now

$$\begin{aligned} \|\Phi_{x_0}\|^s_{L^s(\Omega)} &\leq \|\Phi_{x_0}\|^s_{L^s(B(x_0,R))} \\ &= \frac{1}{(2\pi)^s} \int_{B(x_0,R)} |\ln|x - x_0||^s \, dx \\ &= \frac{1}{(2\pi)^s} \int_0^R \int_{\partial B(x_0,r)} |\ln r|^s \, dS \, dr \\ &= \frac{1}{(2\pi)^{s-1}} \int_0^R r|\ln r|^s \, dr \\ &= \frac{1}{(2\pi)^{s-1}} \left( \int_0^1 r(-\ln r)^s \, dr + \int_1^R r \ln^s r \, dr \right). \end{aligned} \tag{3.24}$$

The second integral in (3.24) is obviously finite. For the first integral, the loose estimate $x \leq e^x$ for all $x \in \mathbb{R}$ gives

$$\begin{aligned} \int_0^1 r(-\ln r)^s \, dr &\leq \int_0^1 r \left(e^{-\ln r}\right)^s dr \\ &= \int_0^1 r^{1-s} \, dr \\ &= \lim_{\varepsilon \to 0} \int_\varepsilon^1 r^{1-s} \, dr \\ &= \lim_{\varepsilon \to 0} \left( \frac{1}{2-s} \left(1 - \varepsilon^{2-s}\right) \right) \\ &= \frac{1}{2-s}. \end{aligned}$$

The limit is finite since $2 - s > 0$. Thus, $\|\Phi_{x_0}\|_{L^s(\Omega)} < \infty$ and $\Phi_{x_0} \in L^s(\Omega)$.

Consider then the integrability of the derivatives:

$$\|D_i \Phi_{x_0}\|^s_{L^s(\Omega)} \leq \|D_i \Phi_{x_0}\|^s_{L^s(B(x_0,R))}$$

$$
\begin{aligned}
&= \frac{1}{(2\pi)^s} \int_{B(x_0,R)} \frac{(x_i - x_{0i})^s}{|x - x_0|^{2s}}\, dx \\
&\leq \frac{1}{(2\pi)^s} \int_{B(x_0,R)} \frac{|x - x_0|^s}{|x - x_0|^{2s}}\, dx \\
&= \frac{1}{(2\pi)^s} \int_{B(x_0,R)} \frac{1}{|x - x_0|^s}\, dx \\
&= \frac{1}{(2\pi)^s} \int_0^R \int_{\partial B(x_0,r)} \frac{1}{r^s}\, dS\, dr \\
&= \frac{1}{(2\pi)^{s-1}} \int_0^R r^{1-s}\, dr, \quad i = 1, 2.
\end{aligned}
$$

This is the same integral as above which is finite. Thus, $\nabla\Phi_{x_0} \in L^s(\Omega) \times L^s(\Omega)$.

To conclude that $\Phi_{x_0} \in W^{1,s}(\Omega)$, we need to show that $\nabla\Phi_{x_0}$ is the weak gradient of $\Phi_{x_0}$. This is not obvious due to the singularity at the point $x_0$. Let $\varepsilon > 0$ be small so that $B(x_0, \varepsilon) \subset \Omega$. We need to consider $\Phi_{x_0}$ inside and outside this ball separately. Outside the ball, $\Phi_{x_0}$ is smooth and bounded so we may use classical results from calculus.

Let $\phi \in C_0^\infty(\Omega)$. Then for $i = 1$ and $i = 2$, we have

$$
\int_\Omega \Phi_{x_0} D_i\phi\, dx = \int_{B(x_0,\varepsilon)} \Phi_{x_0} D_i\phi\, dx + \int_{\Omega\setminus\overline{B(x_0,\varepsilon)}} \Phi_{x_0} D_i\phi\, dx. \tag{3.25}
$$

Bringing $\varepsilon$ to zero, the first integral in (3.25) becomes

$$
\lim_{\varepsilon\to 0} \int_{B(x_0,\varepsilon)} \Phi_{x_0} D_i\phi\, dx = \int_\Omega \lim_{\varepsilon\to 0} \mathbb{1}_{B(x_0,\varepsilon)} \Phi_{x_0} D_i\phi\, dx = 0, \tag{3.26}
$$

where we used the dominated convergence theorem with the estimate $|\Phi_{x_0} D_i\phi| \leq |\Phi_{x_0}| \|D_i\phi\|_{L^\infty(\Omega)} \in L^1(\Omega)$. By integration by parts, the second integral in (3.25) becomes

$$
\int_{\Omega\setminus\overline{B(x_0,\varepsilon)}} \Phi_{x_0} D_i\phi\, dx = -\int_{\Omega\setminus\overline{B(x_0,\varepsilon)}} D_i\Phi_{x_0}\phi\, dx - \int_{\partial B(x_0,\varepsilon)} \Phi_{x_0}\phi n_i\, dS, \tag{3.27}
$$

where $n = (n_1, n_2)$ is the exterior unit normal to $\partial B(x_0, \varepsilon)$. By the dominated convergence theorem, the first integral in (3.27) has the limit

$$
\begin{aligned}
\lim_{\varepsilon\to 0} \int_{\Omega\setminus\overline{B(x_0,\varepsilon)}} D_i\Phi_{x_0}\phi\, dx &= \int_\Omega \lim_{\varepsilon\to 0} \mathbb{1}_{\Omega\setminus\overline{B(x_0,\varepsilon)}} D_i\Phi_{x_0}\phi\, dx \\
&= \int_\Omega D_i\Phi_{x_0}\phi\, dx,
\end{aligned}
$$

where we used the estimate $|D_i\Phi_{x_0}\phi| \leq |D_i\Phi_{x_0}| \|\phi\|_{L^\infty(\Omega)} \in L^1(\Omega)$. The boundary term in (3.27) can be estimated by

$$
\left| \int_{\partial B(x_0,\varepsilon)} \Phi_{x_0}\phi n_i\, dS \right| \leq \int_{\partial B(x_0,\varepsilon)} |\Phi_{x_0}\phi n_i|\, dS
$$

$$\begin{aligned}&\leq \|\phi\|_{L^\infty(\Omega)} \int_{\partial B(x_0,\varepsilon)} |\ln \varepsilon| \, dS \\ &= \|\phi\|_{L^\infty(\Omega)} 2\pi\varepsilon |\ln \varepsilon| \\ &\xrightarrow{\varepsilon \to 0} 0.\end{aligned}$$

The limit follows from a simple application of L'Hôpital's rule. Thus, the boundary term vanishes as $\varepsilon \to 0$. Bringing now $\varepsilon$ to zero in (3.25), we get

$$\int_\Omega \Phi_{x_0} D_i \phi \, dx = -\int_\Omega D_i \Phi_{x_0} \phi \, dx.$$

That is, the weak derivatives exist and $\Phi_{x_0} \in W^{1,s}(\Omega)$.

It remains to show that $\Phi_{x_0}$ solves the problem (3.23) in the weak sense. The boundary values obviously hold. Let $v \in W^{1,s'}(\Omega)$ be such that $v|_{\Gamma_j} = 0$ for all $j \in D$ and $s'$ is the conjugate exponent of $s$. Now

$$\int_\Omega \nabla \Phi_{x_0} \cdot \nabla v \, dx = \int_{B(x_0,\varepsilon)} \nabla \Phi_{x_0} \cdot \nabla v \, dx + \int_{\Omega \setminus \overline{B(x_0,\varepsilon)}} \nabla \Phi_{x_0} \cdot \nabla v \, dx. \tag{3.28}$$

By the Cauchy-Schwarz and Hölder's inequalities, we have

$$|\nabla \Phi_{x_0} \cdot \nabla v| \leq |\nabla \Phi_{x_0}| |\nabla v| \in L^1(\Omega).$$

Thus, we may apply the dominated convergence theorem to the limit of the first integral in (3.28) as $\varepsilon$ tends to zero:

$$\lim_{\varepsilon \to 0} \int_{B(x_0,\varepsilon)} \nabla \Phi_{x_0} \cdot \nabla v \, dx = \int_\Omega \lim_{\varepsilon \to 0} \mathbb{1}_{B(x_0,\varepsilon)} \nabla \Phi_{x_0} \cdot \nabla v \, dx = 0.$$

By Green's formula, the second integral in (3.28) becomes

$$\begin{aligned}\int_{\Omega \setminus \overline{B(x_0,\varepsilon)}} \nabla \Phi_{x_0} \cdot \nabla v \, dx = -\int_{\Omega \setminus \overline{B(x_0,\varepsilon)}} \Delta \Phi_{x_0} v \, dx + \sum_{j \in N} \int_{\Gamma_j} \frac{\partial \Phi_{x_0}}{\partial n} v \, dS \\ - \int_{\partial B(x_0,\varepsilon)} \frac{\partial \Phi_{x_0}}{\partial n} v \, dS.\end{aligned}$$

The first integral vanishes because $\Delta \Phi_{x_0} = 0$ everywhere except at the point $x_0$. The last boundary integral can be simplified as follows. Note that $n = (x - x_0)/\varepsilon$. Now

$$\begin{aligned}-\int_{\partial B(x_0,\varepsilon)} \frac{\partial \Phi_{x_0}}{\partial n} v \, dS &= \int_{\partial B(x_0,\varepsilon)} \left( \frac{1}{2\pi} \frac{x - x_0}{|x - x_0|^2} \cdot \frac{x - x_0}{\varepsilon} \right) v \, dS \\ &= \int_{\partial B(x_0,\varepsilon)} \frac{1}{2\pi\varepsilon} \frac{|x - x_0|^2}{|x - x_0|^2} v \, dS \\ &= \frac{1}{2\pi\varepsilon} \int_{\partial B(x_0,\varepsilon)} v \, dS \\ &\xrightarrow{\varepsilon \to 0} v(x_0).\end{aligned}$$

The limit follows from Theorem 2.7. Thus, bringing $\varepsilon$ to zero in (3.28), we get

$$\int_\Omega \nabla\Phi_{x_0} \cdot \nabla v \, dx = v(x_0) + \sum_{j \in N} \int_{\Gamma_j} \frac{\partial \Phi_{x_0}}{\partial n} v \, dS,$$

which is what we wanted to show. □

The Green's function is commonly used as an auxiliary function to find a general classical solution to Poisson's equation. In the same spirit, we consider next how it can be used to prove a result similar to Theorem 3.9 but with general boundary data.

### 3.2.2 Existence and Uniqueness of Solutions with General Boundary Data

Theorem 3.9 and the solvability of the classical weak formulation of Poisson's equation imply that the problem (3.20) has a weak solution.

**Theorem 3.10.** *Let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain, and let $x_0 \in \Omega$. Let $g_j \in T(H^2(\Omega))$ for all $j \in D$ and $g_j \in T(H^1(\Omega))$ for all $j \in N$. Assume that there exists a function $g_D \in H^2(\Omega)$ such that $g_D|_{\Gamma_j} = g_j$ for all $j \in D$, and if $D = \emptyset$, assume that $g_j$ satisfy the compatibility condition. Let $1 < s < 2$. Then the boundary value problem*

$$\begin{cases} -\Delta u = \delta_{x_0} & \text{in } \Omega \\ u = g_j & \text{on } \Gamma_j, \quad j \in D \\ \dfrac{\partial u}{\partial n} = g_j & \text{on } \Gamma_j, \quad j \in N \end{cases} \tag{3.29}$$

*has a weak solution $u \in W^{1,s}(\Omega)$.*

*Proof.* Define $\Phi_{x_0}$ like before by (3.22). The strategy is to consider the problem

$$\begin{cases} -\Delta w = 0 & \text{in } \Omega \\ w = g_j - \Phi_{x_0} & \text{on } \Gamma_j, \quad j \in D \\ \dfrac{\partial w}{\partial n} = g_j - \dfrac{\partial \Phi_{x_0}}{\partial n} & \text{on } \Gamma_j, \quad j \in N \end{cases} \tag{3.30}$$

and use Theorem 3.5 and Theorem 3.9.

Assume that Theorem 3.5 implies the existence of a weak solution $w \in H^1(\Omega)$ to the problem (3.30). Then $u = \Phi_{x_0} + w$ is a weak solution to the problem (3.29) as follows. First, $u \in W^{1,s}(\Omega)$ because $\Phi_{x_0} \in W^{1,s}(\Omega)$ and the imbedding $H^1(\Omega) \subset W^{1,s}(\Omega)$ implies that $w \in W^{1,s}(\Omega)$. Second, it clearly holds that $u|_{\Gamma_j} = g_j$ for all $j \in D$. Finally, let $v \in W^{1,s'}(\Omega)$ be such that $v|_{\Gamma_j} = 0$ for all $j \in D$ and $s' \in (2, \infty)$ is the conjugate exponent of $s$. By the definition of a weak solution, we then have

$$\begin{aligned} \int_\Omega \nabla u \cdot \nabla v \, dx &= \int_\Omega \nabla\Phi_{x_0} \cdot \nabla v \, dx + \int_\Omega \nabla w \cdot \nabla v \, dx \\ &= v(x_0) + \sum_{j \in N} \int_{\Gamma_j} \frac{\partial \Phi_{x_0}}{\partial n} v \, dS + \sum_{j \in N} \int_{\Gamma_j} \left( g_j - \frac{\partial \Phi_{x_0}}{\partial n} \right) v \, dS \end{aligned}$$

$$= v(x_0) + \sum_{j \in N} \int_{\Gamma_j} g_j v \, dS.$$

It remains to show that Theorem 3.5 is applicable to the problem (3.30). It is not clear whether the Dirichlet and Neumann boundary functions belong to the spaces $T(H^2(\Omega))$ and $T(H^1(\Omega))$, respectively. If $\Phi_{x_0} \in H^2(\Omega)$ were true, then this would obviously be true, but it is easy to show that $\Phi_{x_0}$ does not even belong to the space $H^1(\Omega)$ because its gradient does not satisfy the integrability condition.

The only issue is the singularity of $\Phi_{x_0}$ at the point $x_0$. The function $\Phi_{x_0}$ is smooth everywhere else including the boundary. Since $\Phi_{x_0}$ is radial around the point $x_0$, let us remove the singularity by truncating $\Phi_{x_0}$ over a ball $B(x_0, \varepsilon) \subset \Omega$ and denote the truncated function by $\Phi_{x_0,\varepsilon}$ which is defined by

$$\Phi_{x_0,\varepsilon}(x) = (1 - \mathbb{1}_{B(x_0,\varepsilon)}(x))\Phi_{x_0}(x) + \mathbb{1}_{B(x_0,\varepsilon)}(x)\Phi_{x_0}(x_0 + \varepsilon r)$$

for all $x \in \overline{\Omega}$ and where $r \in \mathbb{R}^2$ is an arbitrary fixed unit vector. Clearly, $\Phi_{x_0,\varepsilon}|_{\partial\Omega} = \Phi_{x_0}|_{\partial\Omega}$.

Let us now smooth out $\Phi_{x_0,\varepsilon}$ over the edge $\partial B(x_0, \varepsilon)$. Assume that $\varepsilon$ is chosen so that the distance between the ball $B(x_0, \varepsilon)$ and the boundary $\partial\Omega$ is at least $\varepsilon$. By a standard mollification argument, there exists a function $\eta \in C_0^\infty(\mathbb{R})$ such that $\eta(t) = 1$ whenever $|t| < \varepsilon/4$ and $\eta(t) = 0$ whenever $|t| > \varepsilon/2$. Consider then the function

$$g_\Phi(x) = (1 - \eta(|x - x_0| - \varepsilon))\Phi_{x_0,\varepsilon}(x) + \eta(|x - x_0| - \varepsilon)\Phi_{x_0,\varepsilon}(x_0)$$

for all $x \in \overline{\Omega}$. The first term vanishes near the edge $\partial B(x_0, \varepsilon)$ so that only the smooth second term remains. Moreover, $g_\Phi$ is constant over the ball $B(x_0, \varepsilon/2)$, which implies that $g_\Phi$ is also smooth at the point $x_0$. In particular, $g_\Phi \in C^\infty(\overline{\Omega}) \subset H^2(\Omega)$.

The fact that $g_\Phi = \Phi_{x_0}$ near and on the boundary implies that $\Phi_{x_0}|_{\partial\Omega} \in T(H^2(\Omega))$ and $\partial\Phi_{x_0}/\partial n \in T(H^1(\Omega))$ on each boundary segment $\Gamma_j$ for $j \in N$. Moreover,

$$T(g_D - g_\Phi) = Tg_D - Tg_\Phi = g_j - \Phi_{x_0}$$

for all $j \in D$ and $g_D - g_\Phi \in H^2(\Omega)$. We may thus apply Theorem 3.5 to the problem (3.30), which concludes the proof. □

Theorem 3.10 does not say anything about the uniqueness of the weak solutions. When the domain $\Omega$ satisfies Assumption 3.1, uniqueness follows as well.

**Theorem 3.11.** *Let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain that satisfies Assumption 3.1. Let $u \in W^{1,s}(\Omega)$ be a weak solution to the problem (3.29). When $D \neq \emptyset$, $u$ is unique. When $D = \emptyset$, $u$ is unique up to an additive constant.*

*Proof.* Let $v \in W^{1,s}(\Omega)$ be another weak solution to the problem (3.29). Then $u - v \in W^{1,s}(\Omega)$ satisfies

$$\int_\Omega \nabla(u - v) \cdot \nabla w \, dx = 0 \tag{3.31}$$

for all $w \in W^{1,s'}(\Omega)$, where $s' \in (2, \infty)$ is the conjugate exponent of $s$. Moreover, it holds that $(u - v)|_{\Gamma_j} = 0$ for all $j \in D$.

Consider then the boundary value problem

$$\begin{cases} -\Delta h = u - v & \text{in } \Omega \\ \quad h = 0 & \text{on } \Gamma_j, \quad j \in D \\ \dfrac{\partial h}{\partial n} = 0 & \text{on } \Gamma_j, \quad j \in N, \end{cases} \tag{3.32}$$

where $u - v \in L^2(\Omega)$ by the Sobolev imbedding theorem. When $D = \emptyset$, the load term needs to satisfy the compatibility condition

$$\int_\Omega u - v \, dx = 0. \tag{3.33}$$

The general case, for which (3.33) does not necessarily hold, is considered at the end.

By Theorem 3.5 and Theorem 3.6, there exists a weak solution $h \in H^2(\Omega)$ to the problem (3.32). Moreover, by Theorem 3.7, it holds that $-\Delta h = u - v$ and $\partial h / \partial n = 0$ on each $\Gamma_j$ for $j \in N$. Now by these facts and Green's formula, we get

$$\begin{aligned} \|u - v\|^2_{L^2(\Omega)} &= \int_\Omega (u - v)^2 \, dx \\ &= -\int_\Omega (u - v)\Delta h \, dx \\ &= \int_\Omega \nabla(u - v) \cdot \nabla h \, dx - \sum_{j \in D} \int_{\Gamma_j} \frac{\partial h}{\partial n}(u - v) \, dS - \sum_{j \in N} \int_{\Gamma_j} \frac{\partial h}{\partial n}(u - v) \, dS \\ &= 0. \end{aligned}$$

Note that the first term is zero because the Sobolev imbedding theorem implies that $H^2(\Omega) \subset W^{1,s'}(\Omega)$ and we may thus apply the identity (3.31). The integrals over the Dirichlet boundary segments vanish because $(u - v)|_{\Gamma_j} = 0$ for all $j \in D$.

The result $\|u - v\|^2_{L^2(\Omega)} = 0$ implies that $u = v$ almost everywhere in $\Omega$. This concludes the proof for the case $D \neq \emptyset$. This also concludes the proof for the case $D = \emptyset$ with the condition (3.33), i.e. $u - v \in V_N$. If the condition (3.33) does not hold, then as usual we consider the normalized function $u - v - (\overline{u} - \overline{v}) \in V_N$, where

$$\overline{u} = \frac{1}{|\Omega|} \int_\Omega u \, dx.$$

Using the normalized function as the load term for the problem (3.32) then implies that $u - v - (\overline{u} - \overline{v}) = 0$, i.e. $u$ and $v$ differ by the constant $\overline{u} - \overline{v}$. This concludes the proof. □

An easy corollary to Theorem 3.11 is that the solution is the same for all $s \in (1, 2)$. We skip the proof, however.

# 4 The Finite Element Method

This section introduces the finite element method as a general procedure for obtaining approximate solutions to weakly formulated elliptic boundary value problems. After the introduction, we consider the error of the approximations for the $p$-version of the finite element method. This is done through the approximation properties of high-order polynomials in one and two dimensions.

## 4.1 Definition of a Finite Element Solution

Consider the task of finding a solution to the weakly formulated boundary value problem

$$\text{Find } u \in U \text{ s.t. } a(u, v) = \varphi(v) \text{ for all } v \in V, \tag{4.1}$$

where $U$ and $V$ are Sobolev subspaces, the mapping $a : U \times V \longrightarrow \mathbb{R}$ is bilinear and $\varphi \in V'$. As an example, consider the classical weak formulation of Poisson's equation with homogeneous Dirichlet boundary conditions for which

$$U = V = \left\{v \in H^1(\Omega) : v|_{\Gamma_j} = 0 \text{ for all } j \in D\right\}$$

or, if $D = \varnothing$,

$$U = V = \left\{v \in H^1(\Omega) : \int_\Omega v \, dx = 0\right\}.$$

With a Dirac delta load, the spaces $U$ and $V$ are otherwise defined identically but $U \subset W^{1,s}(\Omega)$ and $V \subset W^{1,s'}(\Omega)$, where $s \in (1, 2)$ and $s' \in (2, \infty)$ are conjugate exponents.

Assume that the problem (4.1) has a unique solution. Based on the previous section, the validity of this assumption typically relies on the applicability of the Lax-Milgram theorem. The Lax-Milgram theorem does not, however, provide a method for constructing the solution for computational purposes. In fact, it is rare to find an explicit formula for the solution, but we may still try to approximate it.

Let $S_U$ and $S_V$ be finite-dimensional subspaces of $U$ and $V$, and assume that $\dim S_U = \dim S_V$. Denote the dimension of the subspaces by $m \in \mathbb{N}$, and let $\{u_i\}_{i=1}^m$ and $\{v_i\}_{i=1}^m$ be bases for $S_U$ and $S_V$, respectively. Consider then the discretized problem

$$\text{Find } u_S \in S_U \text{ s.t. } a(u_S, v) = \varphi(v) \text{ for all } v \in S_V. \tag{4.2}$$

Note that a possible solution can be written as

$$u_S = \sum_{i=1}^m b_i u_i \tag{4.3}$$

for some coefficient vector $b = (b_1, \ldots, b_m) \in \mathbb{R}^m$. It turns out that the problem (4.2) is equivalent to solving a linear system of equations.

**Theorem 4.1.** *Let $S_U \subset U$ and $S_V \subset V$ as above with the bases $\{u_i\}_{i=1}^m$ and $\{v_i\}_{i=1}^m$, and consider the discretized problem (4.2). Define a matrix $K \in \mathbb{R}^{m\times m}$ such that $K_{ij} = a(u_j, v_i)$, and define a vector $r \in \mathbb{R}^m$ such that $r_i = \varphi(v_i)$. Then $u_S$ is a solution to the discretized problem if and only if the coefficient vector $b \in \mathbb{R}^m$ of $u_S$ solves the system of equations $Kb = r$.*

*Proof.* Assume first that $u_S$ is a solution to (4.2). Let $v \in S_V$ which can be written as

$$v = \sum_{i=1}^m c_i v_i \tag{4.4}$$

for some $c = (c_1, \dots, c_m) \in \mathbb{R}^m$. Substituting (4.3) and (4.4) into (4.2) and using the linearity of $a$ and $\varphi$ gives

$$\sum_{i=1}^m c_i \sum_{j=1}^m b_j a(u_j, v_i) = \sum_{i=1}^m c_i \varphi(v_i). \tag{4.5}$$

Note that (4.5) can be written as

$$c^T K b = c^T r. \tag{4.6}$$

The vector $c$ was chosen to be arbitrary, which means that by setting $c = Kb - r$ in (4.6), we deduce that the coefficient vector $b$ must solve the system $Kb = r$.

Assume then that $b$ solves the system $Kb = r$. Let $v \in S_V$ as in (4.4). Backtracking the steps above, we conclude that

$$a(u_S, v) = \varphi(v),$$

which means that $u_S$ solves the problem (4.2). □

For Poisson's equation, the matrix $K$ is called the stiffness matrix, and the vector $r$ is called the load vector. The assumption that $\dim S_U = \dim S_V = m$ is now useful because $K$ is a square matrix in that case. In particular, the discretized problem (4.2) has a unique solution if and only if $K$ is non-singular. It turns out that the matrix $K$ is non-singular precisely when the mapping $a$ and the subspaces $S_U$ and $S_V$ satisfy a certain regularity condition, which is presented in the following theorem. For reference, see also [12].

**Theorem 4.2.** *The matrix $K$ in Theorem 4.1 is non-singular if and only if for every $0 \neq u \in S_U$ there exists a $v \in S_V$ such that $a(u, v) \neq 0$.*

*Proof.* Assume first that $K$ is non-singular. Let $0 \neq u \in S_U$ and write

$$u = \sum_{i=1}^m b_i u_i \tag{4.7}$$

for some $b = (b_1, \dots, b_m) \in \mathbb{R}^m$. Since $u \neq 0$, also $b \neq 0$. Because $K$ is non-singular, $Kb \neq 0$. Let $v \in S_V$ be such that

$$v = \sum_{i=1}^{m} c_i v_i,$$

where $c = Kb \in \mathbb{R}^m$. Now

$$a(u, v) = c^T K b = (Kb)^T K b = \|Kb\|^2 > 0.$$

That is, $a(u, v) \neq 0$.

Assume then the other direction that for every $0 \neq u \in S_U$ there exists a $v \in S_V$ such that $a(u, v) \neq 0$. Aiming for a contradiction, assume that $K$ is singular. The singularity of $K$ implies that there exists a $0 \neq b \in \mathbb{R}^m$ such that $Kb = 0$. Defining $0 \neq u \in S_U$ as in (4.7), we now get that

$$a(u, v) = c^T K b = 0$$

for all $v \in S_V$. This contradicts the initial assumption, which means that $K$ must be non-singular. □

Summarizing the above discussion, the idea is that the solution of the discretized problem (4.2) approximates the solution of the initial problem (4.1), and the solution of the discretized problem can be computed by solving a linear system of equations that is guaranteed to have a unique solution as long as the subspaces $S_U$ and $S_V$ are chosen suitably according to Theorem 4.2. We generally assume that the mapping $a$ corresponds to an elliptic second-order differential operator which is typically formulated in the Hilbert space $H^1(\Omega)$. In this case, by setting $S_U = S_V = S \subset H^1(\Omega)$, the ellipticity of $a$ implies that $a(u, u) > 0$ for all $0 \neq u \in S$, which then implies via Theorem 4.2 that the discretized problem has a unique solution. Moreover, if the mapping $a$ is also symmetric, as it typically is, the matrix $K$ is symmetric and positive definite, which means that the system $Kb = r$ can be solved efficiently on a computer by using e.g. the Cholesky decomposition.

The subspace $S$ should naturally be chosen so that the matrix $K$ and the vector $r$ can be computed in practice. Recall from the previous section that the entries of $K$ and $r$ typically correspond to integrals over the domain $\Omega$. In the finite element method, the subspace $S$ is chosen as the space of certain piecewise polynomials of a given degree. Polynomials are particularly suitable for integration and differentiation. For a polygonal domain $\Omega \subset \mathbb{R}^2$, the finite element space $S$ is more specifically obtained by modeling $\Omega$ as a mesh of triangles, quadrilaterals or both, which are called elements, and then defining the functions in $S$ to be continuous over the domain $\Omega$ so that the piecewise parts correspond to the elements of the mesh.

We denote a mesh on the domain $\Omega$ by $\mathcal{M}$ which is a finite set of triangles or quadrilaterals whose closures' union is exactly the closure of $\Omega$. That is, $\mathcal{M} = \{E_i\}_{i=1}^{N}$, where $N$ is the total number of elements and $E_i$ is a non-degenerate triangle or quadrilateral for all $i = 1, 2, \dots, N$, and

$$\overline{\Omega} = \bigcup_{i=1}^{N} \overline{E_i}.$$

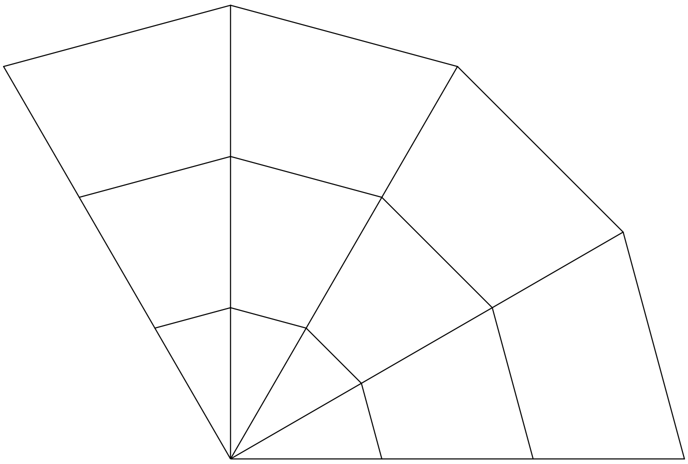


**Figure 2:** A finite element mesh.

For simplicity, it is also assumed that the quadrilaterals are convex and the intersection $\overline{E_i} \cap \overline{E_j}$ for all $i \neq j$ either is empty, consists of a common vertex or consists of an entire common side. Figure 2 contains an example of such a mesh.

Denote arbitrary triangular and quadrilateral domains by $T$ and $Q$, respectively. Define also a reference triangle by

$$\widehat{T} = \{(x_1, x_2) \in \mathbb{R}^2 : 0 < x_1 < 1,\ 0 < x_2 < 1 - x_1\}$$

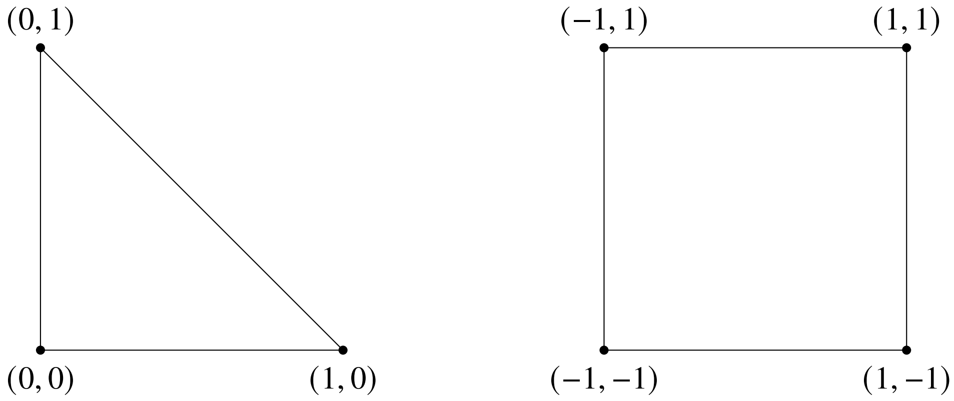
and a reference quadrilateral by $\widehat{Q} = (-1, 1) \times (-1, 1)$, see Figure 3. The symbol $\widehat{E}$ is used to mean the corresponding reference element of the element $E$.

For each $i = 1, 2, \ldots, N$, let $F_i : \widehat{E_i} \to E_i$ be a bijective mapping between the element $E_i$ and its corresponding reference element $\widehat{E_i}$. The finite element space $S$ is now defined by

$$\begin{aligned} S &= S(\Omega, \mathcal{M}, p) \\ &= \{u \in C(\overline{\Omega}) : u \circ F_i \in \mathcal{P}_p(\widehat{E_i}),\ i = 1, 2, \ldots, N(\mathcal{M}),\ Bu = 0\}, \end{aligned} \tag{4.8}$$

where $\mathcal{P}_p(\widehat{E})$ is the space of bivariate polynomials of degree $p \in \mathbb{N}$ and $Bu = 0$ corresponds to the homogeneous Dirichlet boundary conditions or, in the case of a pure Neumann problem, to the zero mean value requirement in order to fix the solution. The degree $p$ of the polynomials can be interpreted as the maximal sum of the powers of each monomial, so that for a monomial $x^i y^j$ it holds that $i + j \leq p$, or as the degree of each independent variable, i.e. $i \leq p$ and $j \leq p$. These two polynomial spaces are called the trunk space and the product space, respectively, although their definitions may slightly differ depending on whether the element is a triangle or a quadrilateral [3]. We consider these polynomial spaces for both triangles and quadrilaterals later.

The bijective element mapping $F_i : \widehat{E_i} \to E_i$ is commonly defined as the linear combination of the vertices of the element $E_i$, which is the definition we will always use as well. When the element $E_i$ is a triangle or a parallelogram, such a mapping is affine. It is easy to show that an affine mapping preserves polynomials, which means



**(a)** The reference triangle $\widehat{T}$.

**(b)** The reference quadrilateral $\widehat{Q}$.

**Figure 3:** The reference elements.

that when the mesh consists of triangles or parallelograms, a function $u \in S$ is truly a piecewise polynomial. However, a similarly defined mapping for a quadrilateral that is not a parallelogram does not preserve the polynomials. For a general quadrilateral, the convexity assumption is also necessary for the bijectivity [23].

Other choices are also possible for the element mappings [3]. A reasonable requirement for the element mappings is that if $u \in W^{k,s}(E_i)$, then $u \circ F_i \in W^{k,s}(\widehat{E}_i)$, and if $\hat{u} \in W^{k,s}(\widehat{E}_i)$, then $\hat{u} \circ F_i^{-1} \in W^{k,s}(E_i)$ for all $k \in \mathbb{N}$ and $s \in [1, \infty)$. This is true for the mappings that linearly combine the vertices of each element [24, Theorem 1 on p. 13]. With this requirement, it is easy to show that the finite element space $S$ is a subspace of $W^{1,s}(\Omega)$ for all $s \in [1, \infty)$ since polynomials on the reference elements obviously belong to the Sobolev spaces $W^{1,s}(\widehat{E}_i)$ and the functions in $S$ are continuous. We skip the proof of this result, but see [20, Theorem 5.2 on p. 62] for reference.

Using the weak formulation of Poisson's equation as an example, the computation of the entries of the stiffness matrix $K$ can be reduced to element-wise computations:

$$\begin{aligned}
K_{ij} &= a(u_j, u_i) \\
&= \int_\Omega \nabla u_j \cdot \nabla u_i \, dx \\
&= \sum_{k=1}^{N} \int_{E_k} \nabla u_j \cdot \nabla u_i \, dx \\
&= \sum_{k=1}^{N} \int_{\widehat{E}_k} \nabla u_j(F_k(\hat{x})) \cdot \nabla u_i(F_k(\hat{x})) |\det J_{F_k}(\hat{x})| \, d\hat{x} \\
&= \sum_{k=1}^{N} \int_{\widehat{E}_k} \hat{\nabla} \hat{u}_j(\hat{x})^T J_{F_k}^{-1}(\hat{x}) J_{F_k}^{-T}(\hat{x}) \hat{\nabla} \hat{u}_i(\hat{x}) |\det J_{F_k}(\hat{x})| \, d\hat{x}, \qquad (4.9)
\end{aligned}$$

where we applied a change of variables to the reference elements via the element mappings. Note that $\hat{\nabla}$ is the gradient with respect to the variable $\hat{x}$. Note also that $\hat{u}_i = u_i \circ F_k \in \mathcal{P}_p(\widehat{E}_k)$. The identity (4.9) implies that all computations can be done in the reference elements regardless of the domain $\Omega$. The same also applies to the assembly of the load vector.

When the element mappings are affine, i.e. $F_k(\hat{x}) = A_k\hat{x} + b_k$ for some matrix $A_k \in \mathbb{R}^{2\times 2}$ and some translation vector $b_k \in \mathbb{R}^2$, then the Jacobian matrix $J_{F_k}$ is equal to $A_k$ and (4.9) can be written as

$$K_{ij} = \sum_{k=1}^{N} \int_{\widehat{E}_k} \hat{\nabla}\hat{u}_j(\hat{x})^T A_k^{-1} A_k^{-T} \hat{\nabla}\hat{u}_i(\hat{x}) |\det A_k| \, d\hat{x}. \tag{4.10}$$

Note that the integrands in (4.10) are polynomials so the integrals can be evaluated exactly. In practice, numerical integration is used, and considering that the integrals consist of polynomials, a Gauss-Legendre quadrature rule is an apt choice, see for example [25] and [26].

The difference between the exact solution $u \in U$ and the finite element solution $u_S \in S$ is often measured in a suitable $W^{1,s}(\Omega)$-norm, typically the $H^1(\Omega)$-norm, or in a suitable $L^s(\Omega)$-norm, typically the $L^2(\Omega)$-norm or the $L^\infty(\Omega)$-norm. There are three commonly used strategies for controlling the error of the approximate solutions: making the mesh more refined, increasing the degree $p$ of the polynomials or using a combination of both of these strategies. In the FEM nomenclature, these three strategies are called the $h$-, $p$- and $hp$-version of the finite element method, respectively.

The parameter $h$ is usually defined by

$$h = \max_{E \in \mathcal{M}} h_E,$$

where $h_E$ is the diameter of the smallest sphere that contains the element $E$. By making the mesh more refined, the largest diameter $h$ becomes smaller and the quality of the finite element solution should hopefully improve. Convergence analysis of the $h$-version attempts to describe how the error behaves as $h \to 0$. To make this analysis feasible, the refinements of the mesh are typically assumed to satisfy some regularity condition. For example, letting $\rho_E$ denote the diameter of the largest sphere contained inside the element $E$, a collection of meshes $\{\mathcal{M}_h\}_{h>0}$ is said to be shape-regular if there exists a constant $\tau > 0$ independent of $h$ such that

$$\frac{h_E}{\rho_E} \leq \tau$$

for every element $E \in \mathcal{M}_h$ and for every mesh $\mathcal{M}_h \in \{\mathcal{M}_h\}_{h>0}$. It is common to use polynomials that have a low degree for the $h$-version, e.g. $p = 1$ or $p = 2$. The $h$-version is the classical version of the finite element method, and it is the main topic of several standard text books, see e.g. [19], [20] and [22].

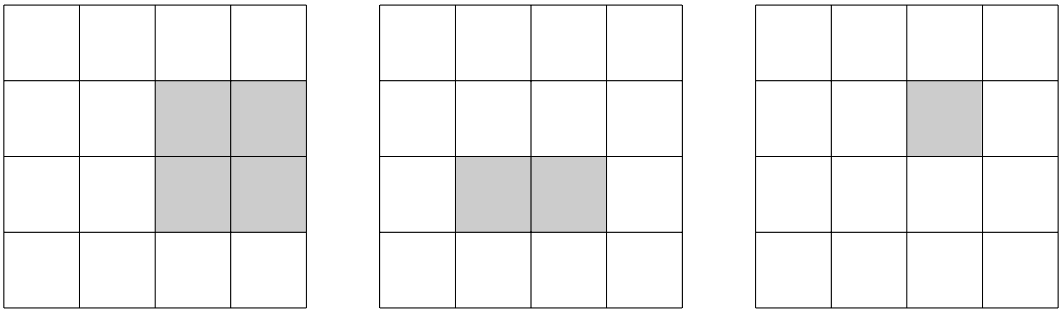
In the $p$-version, the quality of the approximation is improved by increasing the degree $p$ of the polynomials while the mesh is kept fixed. In the $hp$-version, the mesh is refined and the degree of the polynomials is increased at the same time, which can boost the rate of convergence. The mathematical foundations of the $p$-version and the $hp$-version were developed after the $h$-version. For references on the $p$-version, see e.g. [27], [28], [29] and [30]. For references on the $hp$-version, see e.g. [31], [32], [33] and [34]. For a text book on both the $p$-version and the $hp$-version, see [12].

There exist several extensions to the finite element method. Instead of elliptic second-order boundary value problems, one could consider other types of problems. Instead of two-dimensional polygonal domains, one could consider, for example, three-dimensional curved domains with curvilinear elements. Instead of using the same polynomial degree $p$ in every element, one could choose them on an element-by-element basis if the approximation needs to be more accurate in some specific subdomain only. Some of these generalizations are discussed in [3], but we shall not consider them any further. The definition of a finite element solution remains more or less the same, nevertheless.

## 4.2 Basis Functions

To compute the matrix $K$ and the vector $r$, a basis is needed for the finite element space $S$. The basis functions can be divided into nodal basis functions, side basis functions and internal basis functions. A nodal basis function is associated with a vertex of the mesh so that the value of the function is one at the given vertex and the function is supported on the elements sharing the given vertex. There exists one nodal basis function for each vertex. A side basis function is associated with a side of the mesh so that the function is typically a polynomial on the given side and the function is supported on the elements sharing the given side. For each side, there exist $p - 1$ side basis functions. An internal basis function is supported on a given element, and the number of internal basis functions depends on the degree $p$ and whether the polynomial space in the reference elements is the product space or the trunk space. Figure 4 illustrates the supports of the different types of basis functions. A notable implication of such a basis is that the matrix $K$ becomes relatively sparse.

By the definition of the finite element space, the restriction of a basis function on a given element is a polynomial after the change of variables to the corresponding reference element. Thus, the basis functions can be defined by defining a polynomial basis on the reference elements. The polynomials in the reference element basis are called shape functions, and they are also divided into nodal, side and internal shape



**(a)** Nodal basis function. **(b)** Side basis function. **(c)** Internal basis function.

**Figure 4:** The supports of the three types of basis functions.

functions according to the corresponding basis function. There exist several options for how to construct the shape functions. We present here hierarchic shape functions that can be found in [3]. Hierarchic means that the set of shape functions for any degree $p$ is a subset of the set of shape functions for the degree $p + 1$. The basis for the finite element space is then hierarchic as well.

We consider the shape functions for both quadrilaterals and triangles and for both the product space and the trunk space. The only difference between the product and trunk spaces is the number of internal shape functions.

### 4.2.1 Hierarchic Shape Functions for Quadrilaterals

The nodal shape functions are given by

$$\begin{aligned} N_1(\hat{x}_1, \hat{x}_2) &= \frac{1}{4}(1 - \hat{x}_1)(1 - \hat{x}_2) \\ N_2(\hat{x}_1, \hat{x}_2) &= \frac{1}{4}(1 + \hat{x}_1)(1 - \hat{x}_2) \\ N_3(\hat{x}_1, \hat{x}_2) &= \frac{1}{4}(1 + \hat{x}_1)(1 + \hat{x}_2) \\ N_4(\hat{x}_1, \hat{x}_2) &= \frac{1}{4}(1 - \hat{x}_1)(1 + \hat{x}_2). \end{aligned}$$

For the side and internal shape functions, define the auxiliary functions

$$\phi_k(s) = (k-1)k \int_{-1}^{s} P_{k-1}(t)\, dt \tag{4.11}$$

for $k = 2, 3, \ldots, p$, where $P_{k-1}$ is the Legendre polynomial of degree $k - 1$. Note that $\phi_k$ is a polynomial of degree $k$. The side shape functions are now given by

$$\begin{aligned} N_k^{(1)}(\hat{x}_1, \hat{x}_2) &= \frac{1}{2}(1 - \hat{x}_2)\phi_k(\hat{x}_1) \\ N_k^{(2)}(\hat{x}_1, \hat{x}_2) &= \frac{1}{2}(1 + \hat{x}_1)\phi_k(\hat{x}_2) \\ N_k^{(3)}(\hat{x}_1, \hat{x}_2) &= \frac{1}{2}(1 + \hat{x}_2)\phi_k(-\hat{x}_1) \\ N_k^{(4)}(\hat{x}_1, \hat{x}_2) &= \frac{1}{2}(1 - \hat{x}_1)\phi_k(-\hat{x}_2) \end{aligned}$$

for $k = 2, 3, \ldots, p$. The argument of $\phi_k$ is negative for the sides 3 and 4 so that the orientation of the polynomials is the same on all the sides. This way a side basis function can always be constructed by multiplying one of the corresponding side shape functions by $-1$, assuming that the element mappings preserve the orientation, i.e. their Jacobian determinants are positive.

The internal shape functions for the product space are given by

$$N_p^{(k,l)}(\hat{x}_1, \hat{x}_2) = \phi_k(\hat{x}_1)\phi_l(\hat{x}_2) \tag{4.12}$$

for $k = 2, 3, \dots, p$ and $l = 2, 3, \dots, p$. The product space is denoted by $\mathcal{P}_p^{(pr)}(\widehat{Q})$.

The internal shape functions for the trunk space are also given by (4.12), but the indices $k = 2, 3, \dots, p$ and $l = 2, 3, \dots, p$ satisfy $k + l \leq p$. This implies that there are no internal shape functions until $p = 4$. Note also that the trunk space for quadrilaterals contains the monomials $\hat{x}_1^p \hat{x}_2$ and $\hat{x}_1 \hat{x}_2^p$ for all $p$. The trunk space is denoted by $\mathcal{P}_p^{(tr)}(\widehat{Q})$.

### 4.2.2 Hierarchic Shape Functions for Triangles

The nodal shape functions are given by

$$
\begin{aligned}
L_1(\hat{x}_1, \hat{x}_2) &= 1 - \hat{x}_1 - \hat{x}_2 \\
L_2(\hat{x}_1, \hat{x}_2) &= \hat{x}_1 \\
L_3(\hat{x}_1, \hat{x}_2) &= \hat{x}_2.
\end{aligned}
$$

For the side shape functions, define the auxiliary functions

$$
\widetilde{\phi}_k(s) = 4\frac{\phi_k(s)}{1 - s^2} = 4P'_{k-1}(s) \tag{4.13}
$$

for $k = 2, 3, \dots, p$, where $\phi_k$ is given by (4.11) and the second equality follows from the properties of the Legendre polynomials [35]. The side shape functions are now given by

$$
\begin{aligned}
N_k^{(1)}(\hat{x}_1, \hat{x}_2) &= L_1 L_2 \widetilde{\phi}_k(L_2 - L_1) \\
N_k^{(2)}(\hat{x}_1, \hat{x}_2) &= L_2 L_3 \widetilde{\phi}_k(L_3 - L_2) \\
N_k^{(3)}(\hat{x}_1, \hat{x}_2) &= L_3 L_1 \widetilde{\phi}_k(L_1 - L_3)
\end{aligned}
$$

for $k = 2, 3, \dots, p$. The constant 4 in (4.13) is needed to make the side shape functions for triangles compatible with the side shape functions for quadrilaterals. The orientation is already the same for both.

The internal shape functions for the trunk space are given by

$$
N_p^{(k,l)}(\hat{x}_1, \hat{x}_2) = L_1 L_2 L_3 P_k(2\hat{x}_1 - 1) P_l(2\hat{x}_2 - 1) \tag{4.14}
$$

for $k = 0, 1, \dots, p - 3$ and $l = 0, 1, \dots, p - 3$ with $k + l \leq p - 3$. The trunk space is denoted by $\mathcal{P}_p^{(tr)}(\widehat{T})$.

The internal shape functions for the product space are also given by (4.14) but with the indices $k = 0, 1, \dots, p - 2$ and $l = 0, 1, \dots, p - 2$. However, note that the resulting space does not contain the monomial $\hat{x}_1^p \hat{x}_2^p$, for example. We still refer to the space as the product space and denote it by $\mathcal{P}_p^{(pr)}(\widehat{T})$.

Let us finally discuss how the basis functions are related to the boundary conditions of the boundary value problems, which is relevant for both quadrilateral and triangular elements. For homogeneous Dirichlet boundary conditions, consider the nodal and side basis functions that determine the value of a function $u \in S$ on the corresponding boundary segments. To then satisfy the Dirichlet boundary conditions, one only needs to set the coefficients of those basis functions to zero. To achieve this in partice, it is easier to first assemble the system $Kb = r$ without considering any boundary conditions and then eliminate the rows and columns that correspond to the aforementioned nodal and side basis functions, which essentially zeroes the corresponding coefficients.

For a Neumann boundary value problem that has a unique solution, the basis functions do not need any special handling since the boundary conditions are embedded into the load vector instead of the solution space. However, a pure Neumann problem for Poisson's equation does not have a unique solution, and we enforced uniqueness by requiring that the solution has zero mean value over the domain. The above basis does not satisfy the zero mean value requirement, and it would be cumbersome to try to construct such a basis. The system $Kb = r$ can still be assembled with the basis above, but it does not have a unique solution because the stiffness matrix $K$ has a one-dimensional kernel. Fortunately, it is possible to fix the solution by fixing its value at one of the vertices to zero. This is equivalent to setting the coefficient of the corresponding nodal basis function to zero in the coefficient vector $b$ and then eliminating the corresponding row and column from the system $Kb = r$. The resulting system has a unique solution, and the dropped equation holds as a consequence of the Neumann compatibility condition [20]. Finally, the solution can be normalized to have zero mean value.

## 4.3 Convergence of the $p$-Version of the Finite Element Method

One of the main objectives of this thesis is to study the convergence of the $p$-version of the finite element method when applied to Poisson's equation with a Dirac delta load. As in Section 3 regarding the solvability of the weak formulations, we carry out the convergence analysis in two phases. In the first phase, we again consider the classical elliptic weak formulation in the Hilbert space setting. In the second phase, the convergence results of the first phase are applied to the Dirac delta problem via a duality argument by Casas [6]. The rest of this section is concerned with the first phase, and the second phase is covered in the next section.

### 4.3.1 Céa's Lemma

We have seen that the finite element method is a rather complicated process with a lot of varying factors such as the mesh and the degree $p$ of the polynomials. Trying to measure the error between the exact solution and the finite element solution directly is unwieldy. It, however, turns out that the error is directly comparable to how well the finite element space $S$ approximates the ambient space where the exact solution resides. This result is known as Céa's lemma by Jean Céa [36].

**Theorem 4.3** (Céa's lemma)**.** *Let $V$ be a Hilbert space, $a : V \times V \to \mathbb{R}$ an elliptic bounded bilinear mapping and $\varphi \in V'$. Let $u \in V$ be the unique vector that satisfies*

$$a(u, v) = \varphi(v)$$

*for all $v \in V$. Let $S$ be a non-empty closed subspace of $V$, and let $u_S \in S$ be the unique vector that satisfies*

$$a(u_S, v) = \varphi(v)$$

*for all $v \in S$. Then there exists a constant $C > 0$ dependent only on the bilinear mapping such that*

$$\|u - u_S\|_V \leq C \inf_{v \in S} \|u - v\|_V.$$

*Proof.* We begin by observing that for all $v \in S$ it holds that

$$a(u - u_S, v) = a(u, v) - a(u_S, v) = \varphi(v) - \varphi(v) = 0.$$

This property is commonly called Galerkin orthogonality.

Let $v \in S$. By the Galerkin orthogonality and the ellipticity and boundedness of $a$, we get that

$$\begin{aligned} \|u - u_S\|_V^2 &\leq \frac{1}{\alpha} a(u - u_S, u - u_S) \\ &= \frac{1}{\alpha} a(u - u_S, u - v) \\ &\leq \frac{C}{\alpha} \|u - u_S\|_V \|u - v\|_V. \end{aligned}$$

Dividing both sides by $\|u - u_S\|_V$ and taking the infimum over all $v \in S$, the claim follows. □

For our purposes, the Hilbert space $V$ is a subspace of the Sobolev space $H^1(\Omega)$, and the approximation properties of the finite element space $S$ in the $H^1(\Omega)$-norm correspond to the general approximation properties of polynomials with respect to the degree $p$.

### 4.3.2 Approximation Properties of Polynomials

Let us begin with the approximation properties of polynomials in one dimension, after which we consider the two-dimensional setting. The one-dimensional analysis is based on the Legendre series which is discussed in great detail in e.g. [12] and [35].

Let $I = (-1, 1) \subset \mathbb{R}$ and $u \in L^2(I)$. The Legendre series of $u$ is given by

$$u(x) = \sum_{i=0}^{\infty} a_i P_i(x), \tag{4.15}$$

where $P_i$ is the Legendre polynomial of degree $i$. The coefficients of the series are given by

$$a_i = \frac{2i + 1}{2} \int_{-1}^{1} u(x) P_i(x)\, dx$$

for all $i = 0, 1, \ldots$, which follows from the orthogonality of the Legendre polynomials:

$$\int_{-1}^{1} P_i(x) P_j(x)\, dx = \begin{cases} \frac{2}{2i+1}, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

Convergence of the series (4.15) can always be understood in the sense of $L^2$. That is, it holds that

$$\lim_{p\to\infty} \|u - \sum_{i=0}^{p} a_i P_i\|_{L^2(I)} = 0.$$

This implies that the series converges pointwise almost everywhere in $I$. If $u$ has a continuous derivative, then the pointwise convergence holds everywhere in $I$.

When $u \in H^2(I)$, the Legendre series can be used to prove the following polynomial approximation theorem.

**Theorem 4.4.** *Let $u \in H^2(I)$ and $p \in \mathbb{N}$. Then there exists a polynomial $u_p \in \mathcal{P}_p(I)$ and a constant $C > 0$ independent of $u$ and $p$ such that*

$$u(\pm 1) = u_p(\pm 1) \tag{4.16}$$

*and*

$$\|u - u_p\|_{L^2(I)} \le Cp^{-2} A(u), \tag{4.17}$$

$$\|u - u_p\|_{H^1(I)} \le Cp^{-1} A(u), \tag{4.18}$$

$$\|u - u_p\|_{L^\infty(I)} \le Cp^{-1} A(u), \tag{4.19}$$

*where*

$$A(u) = \left(\int_{-1}^{1} |u''(x)|^2 (1 - x^2)\, dx\right)^{\frac{1}{2}}.$$

*Proof.* For the claims (4.16), (4.17) and (4.18), see [29, Lemma 3.3] or [37]. The polynomial $u_p$ is constructed by letting its derivative be the $(p-1)$th order truncated Legendre series of the derivative of $u$, that is,

$$u_p'(x) = \sum_{i=0}^{p-1} b_i P_i(x),$$

where

$$b_i = \frac{2i+1}{2} \int_{-1}^{1} u'(x) P_i(x)\, dx.$$

See also [12, Theorem 3.14 on p. 73] for the same construction. Moreover, by [12, Theorem 3.10 on p. 71], the $A(u)$ term and the coefficients $b_i$ are related by

$$A(u)^2 = \int_{-1}^{1} |u''(x)|^2 (1 - x^2)\, dx = \sum_{i=1}^{\infty} \frac{2}{2i+1} \frac{(i+1)!}{(i-1)!} |b_i|^2. \tag{4.20}$$

Let us use the above construction to prove (4.19) as well which is not included in the results of the aforementioned references.

Let $x \in I$. By the fundamental theorem of calculus and the definition of $u_p'$, we get that

$$u(x) - u_p(x) = \int_{-1}^{x} u'(y) - u_p'(y)\, dy$$

$$
\begin{aligned}
&= \int_{-1}^{x} \sum_{i=p}^{\infty} b_i P_i(y)\, dy \\
&= \sum_{i=p}^{\infty} b_i \int_{-1}^{x} P_i(y)\, dy. \qquad (4.21)
\end{aligned}
$$

By [35, p. 151], the Legendre polynomials satisfy

$$P_i(y) = \frac{1}{2i+1}(P'_{i+1}(y) - P'_{i-1}(y))$$

for all $i = 1, 2, \ldots$. We may thus compute the integral in (4.21) as

$$
\begin{aligned}
\int_{-1}^{x} P_i(y)\, dy &= \frac{1}{2i+1} \int_{-1}^{x} P'_{i+1}(y) - P'_{i-1}(y)\, dy \\
&= \frac{1}{2i+1}(P_{i+1}(x) - P_{i-1}(x)), \qquad (4.22)
\end{aligned}
$$

where we also used the property of the Legendre polynomials that $P_i(-1) = (-1)^i$ for all $i = 0, 1, \ldots$. The Legendre polynomials satisfy $|P_i(x)| \leq 1$ for all $x \in I$ and $i = 0, 1, \ldots$, which now implies together with (4.21) and (4.22) the pointwise estimate

$$
\begin{aligned}
|u(x) - u_p(x)| &\leq \sum_{i=p}^{\infty} |b_i| \frac{1}{2i+1}(|P_{i+1}(x)| + |P_{i-1}(x)|) \\
&\leq \sum_{i=p}^{\infty} \frac{2}{2i+1}|b_i| \\
&= \sum_{i=p}^{\infty} \left( \left(\frac{2}{2i+1}\right)^{\frac{1}{2}} \left(\frac{(i-1)!}{(i+1)!}\right)^{\frac{1}{2}} \right) \left( \left(\frac{2}{2i+1}\right)^{\frac{1}{2}} \left(\frac{(i+1)!}{(i-1)!}\right)^{\frac{1}{2}} |b_i| \right).
\end{aligned}
$$

Hölder's inequality and (4.20) imply that

$$
\begin{aligned}
|u(x) - u_p(x)|^2 &\leq \left( \sum_{i=p}^{\infty} \frac{2}{2i+1} \frac{(i-1)!}{(i+1)!} \right) \left( \sum_{i=p}^{\infty} \frac{2}{2i+1} \frac{(i+1)!}{(i-1)!} |b_i|^2 \right) \\
&\leq \left( \sum_{i=p}^{\infty} \frac{2}{2i+1} \frac{(i-1)!}{(i+1)!} \right) \left( \sum_{i=1}^{\infty} \frac{2}{2i+1} \frac{(i+1)!}{(i-1)!} |b_i|^2 \right) \\
&= \left( \sum_{i=p}^{\infty} \frac{2}{2i+1} \frac{(i-1)!}{(i+1)!} \right) A(u)^2. \qquad (4.23)
\end{aligned}
$$

The remaining series in (4.23) can be estimated by

$$\sum_{i=p}^{\infty} \frac{2}{2i+1} \frac{(i-1)!}{(i+1)!} = \sum_{i=p}^{\infty} \frac{2}{(2i+1)(i+1)i}$$

$$
\begin{aligned}
&\leq \sum_{i=p}^{\infty} i^{-3} \\
&\leq p^{-3} + \int_p^{\infty} t^{-3}\, dt \\
&= p^{-3} + \frac{1}{2} p^{-2} \\
&\leq \frac{3}{2} p^{-2}.
\end{aligned}
$$

This implies with (4.23) that

$$|u(x) - u_p(x)|^2 \leq \frac{3}{2} p^{-2} A(u)^2$$

for all $x \in I$, which then finally implies (4.19). □

Theorem 4.4 has a two-dimensional counterpart for the reference quadrilateral and the reference triangle. The polynomial space is assumed to be the product space for the quadrilateral and the trunk space for the triangle. This assumption will be relaxed later.

**Theorem 4.5.** *Let $\widehat{E} = \widehat{Q}$ (resp. $\widehat{E} = \widehat{T}$). Let $u \in H^2(\widehat{E})$ and $p \in \mathbb{N}$. Then there exists a $u_p \in \mathcal{P}_p^{(pr)}(\widehat{Q})$ (resp. $u_p \in \mathcal{P}_p^{(tr)}(\widehat{T})$) and a constant $C > 0$ independent of $u$ and $p$ such that*

$$\|u - u_p\|_{H^1(\widehat{E})} \leq C p^{-1} \|u\|_{H^2(\widehat{E})}, \tag{4.24}$$

$$\|u - u_p\|_{L^\infty(\widehat{E})} \leq C p^{-1} \|u\|_{H^2(\widehat{E})}, \tag{4.25}$$

$$\|u - u_p\|_{L^2(\gamma)} \leq C p^{-3/2} \|u\|_{H^2(\widehat{E})}, \tag{4.26}$$

$$\|u - u_p\|_{H^1(\gamma)} \leq C p^{-1/2} \|u\|_{H^2(\widehat{E})}, \tag{4.27}$$

*where $\gamma$ is any side of $\widehat{E}$.*

*Proof.* The proof for the case $\widehat{E} = \widehat{Q}$ can be found in [29, Lemma 3.1], where the result follows from the approximation properties of truncated Fourier series.

Let us use the result for the quadrilateral and extend it to the case $\widehat{E} = \widehat{T}$. We do this by dividing the quadrilateral $\widehat{Q}$ into two triangles along the diagonal line $x_2 = x_1$. Let $\widetilde{T}$ denote the resulting bottom triangle, and let $F : \widehat{T} \to \widetilde{T}$ be a bijective affine mapping between the reference triangle and the bottom triangle.

Define $\tilde{u} = u \circ F^{-1} \in H^2(\widetilde{T})$. By [38, Theorem 5 on p. 181], there exists an extension $\tilde{U} \in H^2(\widehat{Q})$ of $\tilde{u}$ such that $\tilde{U}|_{\widetilde{T}} = \tilde{u}$ and

$$\|\tilde{U}\|_{H^2(\widehat{Q})} \leq C_1 \|\tilde{u}\|_{H^2(\widetilde{T})}, \tag{4.28}$$

where the constant $C_1 > 0$ is independent of $\tilde{u}$.

Assume for now that $p \geq 2$. If $p$ is even, let $q = p/2 \in \mathbb{N}$, and if $p$ is odd, let $q = (p-1)/2 \in \mathbb{N}$. Now there exists a $\tilde{U}_q \in \mathcal{P}_q^{(pr)}(\widehat{Q}) \subset \mathcal{P}_p^{(tr)}(\widehat{Q})$ such that $\tilde{U}$ and $\tilde{U}_q$ satisfy (4.24)-(4.27) with $q$ instead of $p$.

Let $\tilde{u}_p \in \mathcal{P}_p^{(tr)}(\widetilde{T})$ be the restriction of $\tilde{U}_q$ on $\widetilde{T}$. Using (4.28), we may now prove (4.24)-(4.27) for $\tilde{u}$ and $\tilde{u}_p$ in the triangle $\widetilde{T}$. For example,

$$\begin{aligned}\|\tilde{u} - \tilde{u}_p\|_{H^1(\widetilde{T})} &\leq \|\tilde{U} - \tilde{U}_q\|_{H^1(\widehat{Q})} \\ &\leq Cq^{-1}\|\tilde{U}\|_{H^2(\widehat{Q})} \\ &\leq 4Cp^{-1}\|\tilde{U}\|_{H^2(\widehat{Q})} \\ &\leq 4CC_1p^{-1}\|\tilde{u}\|_{H^2(\widetilde{T})}.\end{aligned}$$

The other estimates follow more or less analogously. Regarding the side of $\widetilde{T}$ that corresponds to the diagonal line $x_2 = x_1$, it is shown in [29, Lemma 3.1] that (4.26) and (4.27) hold for it as well when $\widehat{E} = \widehat{Q}$.

The desired estimates for the case $\widehat{E} = \widehat{T}$ now follow by applying the affine coordinate transformation $F$ to $\tilde{u}$ and $\tilde{u}_p$. The resulting approximation $u_p = \tilde{u}_p \circ F$ belongs to the trunk polynomial space $\mathcal{P}_p^{(tr)}(\widehat{T})$ because an affine mapping preserves polynomials.

It was assumed above that $p \geq 2$. If $p = 1$, we can simply choose $u_p = 0 \in \mathcal{P}_1^{(tr)}(\widehat{T})$, and the estimates (4.24)-(4.27) follow from the Sobolev imbedding theorem and the trace theorem. □

Theorem 4.5 does not provide precise error bounds since the value of the constant $C$ is not known, but it enables us to consider whether the approximations converge as $p \to \infty$.

### 4.3.3 Approximation Properties of the Finite Element Space

We defined the finite element space by

$$\begin{aligned}S &= S(\Omega, \mathcal{M}, p) \\ &= \{u \in C(\overline{\Omega}) : u \circ F_i \in \mathcal{P}_p(\widehat{E}_i),\ i = 1, 2, \ldots, N(\mathcal{M}),\ Bu = 0\},\end{aligned}$$

where $\mathcal{P}_p(\widehat{E}_i)$ is either the product space or the trunk space.

Let $u \in H^2(\Omega)$ be such that it satisfies the condition $Bu = 0$. By e.g. [29] and [33], there exists a $u_p \in S$ and a constant $C > 0$ independent of $u$ and $p$ such that

$$\|u - u_p\|_{H^1(\Omega)} \leq Cp^{-1}\|u\|_{H^2(\Omega)}. \tag{4.29}$$

Céa's lemma then immediately implies the same estimate for the $H^1(\Omega)$ error between the exact solution and the finite element solution when the exact solution belongs to the Sobolev space $H^2(\Omega)$.

To be able to prove an $L^2(\Omega)$ error estimate for the Dirac delta problem, we will also need the pointwise estimate

$$\|u - u_p\|_{L^\infty(\Omega)} \leq Cp^{-1}\|u\|_{H^2(\Omega)}, \tag{4.30}$$

where the functions $u$ and $u_p$ are as above. The estimate (4.30) does not seem to be covered as such by the existing standard literature so we will prove it in full detail.

The estimate (4.30) can be proven with the same approach that is used in [29] to prove the estimate (4.29). The idea is to use Theorem 4.5 element-wise, but note that the resulting approximation is not necessarily continuous yet, i.e. it does not belong to the space $S$. The element-wise approximations need to be stitched together over the sides of the elements. This is achieved by adding suitable auxiliary functions to each element-wise approximation. By suitable it is meant that the estimates of Theorem 4.5 continue to hold. Homogeneous Dirichlet boundary conditions can be fixed with the same approach as well. We consider these auxiliary functions first.

There are two types of auxiliary functions. The first type is used to fix the values at the vertices of the mesh, and the second type is used to fix the values over the sides after the vertices have been fixed. Both types can be constructed with the help of the following theorem. More or less the same constructions and their relevant norm estimates that are needed to prove (4.29) can be found in [29]. In particular, see the proof of Theorem 4.1 in [29]. We supplement the estimates with the necessary $L^\infty(\Omega)$-norm estimates that are needed to prove (4.30).

**Theorem 4.6.** *Let $I = (-1, 1)$ and $p \in \mathbb{N}$. Then there exists a polynomial $\psi_p \in \mathcal{P}_p(I)$ and a constant $C > 0$ independent of $p$ such that*

$$\psi_p(-1) = 1 \quad \textit{and} \quad \psi_p(1) = 0 \tag{4.31}$$

*and*

$$\|\psi_p\|_{L^2(I)} \leq Cp^{-1/2}, \tag{4.32}$$

$$\|\psi_p\|_{H^1(I)} \leq Cp^{1/2}, \tag{4.33}$$

$$\|\psi_p\|_{L^\infty(I)} \leq C. \tag{4.34}$$

*Proof.* Like in [29], define

$$\phi_p(x) = \frac{e^{-p(x+1)} - e^{-2p}}{1 - e^{-2p}}.$$

Clearly, $\phi_p(-1) = 1$ and $\phi_p(1) = 0$. Applying now Theorem 4.4 to the function $\phi_p$ gives the desired polynomial $\psi_p \in \mathcal{P}_p(I)$ that satisfies (4.31), (4.32) and (4.33), see [29]. It remains to prove the claim (4.34).

The function $\phi_p$ is strictly decreasing in $I$, which implies that

$$\|\phi_p\|_{L^\infty(I)} \leq 1. \tag{4.35}$$

By [29], the term $A(\phi_p)$ in Theorem 4.4 has the upper bound

$$A(\phi_p) \leq Cp$$

for some constant $C > 0$ independent of $p$. Thus,

$$\|\phi_p - \psi_p\|_{L^\infty(I)} \leq C. \tag{4.36}$$

Combining (4.35) and (4.36) gives the desired estimate (4.34):

$$\|\psi_p\|_{L^\infty(I)} \leq \|\psi_p - \phi_p\|_{L^\infty(I)} + \|\phi_p\|_{L^\infty(I)} \leq C + 1.$$

□

The next theorem is used to fix the values at the vertices of the elements.

**Theorem 4.7.** *Let $\widehat{E} = \widehat{Q}$ (resp. $\widehat{E} = \widehat{T}$). Let $V_i$ be a vertex of $\widehat{E}$ for some $i \in \{1,2,3,4\}$ (resp. $i \in \{1,2,3\}$). Let $p \in \mathbb{N}$. Then there exists a polynomial $v_p \in \mathcal{P}_p^{(pr)}(\widehat{Q})$ (resp. $v_p \in \mathcal{P}_p^{(tr)}(\widehat{T})$) and a constant $C > 0$ independent of $p$ such that*

$$v_p(V_i) = 1 \quad \text{and} \quad v_p(V_j) = 0 \text{ for all } j \neq i \tag{4.37}$$

*and*

$$\|v_p\|_{H^1(\widehat{E})} \leq C, \tag{4.38}$$

$$\|v_p\|_{L^\infty(\widehat{E})} \leq C, \tag{4.39}$$

$$\|v_p\|_{L^2(\gamma)} \leq Cp^{-1/2}, \tag{4.40}$$

$$\|v_p\|_{H^1(\gamma)} \leq Cp^{1/2}, \tag{4.41}$$

*where $\gamma$ is any edge of $\widehat{E}$.*

*Proof.* Let first $\widehat{E} = \widehat{Q}$. Without loss of generality, assume that $V_i = (-1,-1)$. Define the polynomial $v_p$ by

$$v_p(x) = \psi_p(x_1)\psi_p(x_2) \in \mathcal{P}_p^{(pr)}(\widehat{Q}),$$

where $\psi_p \in \mathcal{P}_p(I)$ is given by Theorem 4.6. The claim for the case $\widehat{E} = \widehat{Q}$ now follows from the properties of $\psi_p$ in Theorem 4.6, see [29]. Let us only consider the estimate (4.39):

$$\|v_p\|_{L^\infty(\widehat{Q})} \leq \|\psi_p\|^2_{L^\infty(I)} \leq C^2.$$

Consider then the case $\widehat{E} = \widehat{T}$. Without loss of generality, assume that $V_i = (0,0)$. If $p$ is even, let $q = p/2 \in \mathbb{N}$ and define

$$v_p(x) = \psi_q(2x_1 - 1)\psi_q(2x_2 - 1) \in \mathcal{P}_p^{(tr)}(\widehat{T}).$$

If $p > 1$ and $p$ is odd, let $q = (p-1)/2 \in \mathbb{N}$ and define

$$v_p(x) = \psi_q(2x_1 - 1)\psi_{q+1}(2x_2 - 1) \in \mathcal{P}_p^{(tr)}(\widehat{T}).$$

The polynomial $\psi_q \in \mathcal{P}_q(I)$ is again given by Theorem 4.6. The claim now follows with similar arguments as above. If $p = 1$, then $v_p$ can be chosen as the nodal shape function

$$v_p(x) = 1 - x_1 - x_2 \in \mathcal{P}_1^{(tr)}(\widehat{T}).$$

□

After Theorem 4.7 has been applied, the next theorem is used to fix the values across the sides.

**Theorem 4.8.** *Let $\widehat{E} = \widehat{Q}$ (resp. $\widehat{E} = \widehat{T}$). Let $\gamma$ be a side of $\widehat{E}$ with the vertices $V_1$ and $V_2$ at the endpoints. Let $w_p \in \mathcal{P}_p(\gamma) \cong \mathcal{P}_p(I)$, $p \in \mathbb{N}$, be a polynomial on the side such that $w_p(V_1) = w_p(V_2) = 0$. Then there exists a polynomial $\xi_p \in \mathcal{P}_p^{(pr)}(\widehat{Q})$ (resp. $\xi_p \in \mathcal{P}_{2p}^{(tr)}(\widehat{T})$) and a constant $C > 0$ independent of $w_p$ and $p$ such that*

$$\xi_p = w_p \ \text{ on } \gamma \quad \text{and} \quad \xi_p = 0 \ \text{ on } \partial\widehat{E} \setminus \gamma \tag{4.42}$$

*and*

$$\|\xi_p\|_{H^1(\widehat{E})} \leq Cp^{-1/2}\|w_p\|_{H^1(\gamma)} + Cp^{1/2}\|w_p\|_{L^2(\gamma)}, \tag{4.43}$$

$$\|\xi_p\|_{L^\infty(\widehat{E})} \leq C\|w_p\|_{L^\infty(\gamma)}. \tag{4.44}$$

*Proof.* Let first $\widehat{E} = \widehat{Q}$. Without loss of generality, assume that $\gamma$ is the side corresponding to the horizontal line $x_2 = -1$. Define the polynomial $\xi_p$ by

$$\xi_p(x) = w_p(x_1)\psi_p(x_2) \in \mathcal{P}_p^{(pr)}(\widehat{Q}),$$

where $\psi_p \in \mathcal{P}_p(I)$ is given by Theorem 4.6. The claim for the case $\widehat{E} = \widehat{Q}$ now follows from the properties of $\psi_p$ in Theorem 4.6 and the assumption that $w_p(V_1) = w_p(V_2) = 0$, see [29]. We only consider the estimate (4.44) which follows easily:

$$\|\xi_p\|_{L^\infty(\widehat{Q})} \leq \|\psi_p\|_{L^\infty(I)}\|w_p\|_{L^\infty(\gamma)} \leq C\|w_p\|_{L^\infty(\gamma)}.$$

Assume then that $\widehat{E} = \widehat{T}$. Without loss of generality, let $\gamma$ be the side corresponding to the line $x_2 = 1 - x_1$. The side has the parametrization $\gamma(t) = [t, 1-t]^T$ for $t \in [0, 1]$, and we write $w_p(t)$ to mean $w_p(\gamma(t))$. For $p \geq 2$, define $\xi_p$ now by

$$\xi_p(x) = \psi_{p-1}(2(1 - x_1 - x_2) - 1)[x_2 w_p(x_1) + x_1 w_p(1 - x_2)] \in \mathcal{P}_{2p}^{(tr)}(\widehat{T}), \tag{4.45}$$

where $\psi_{p-1} \in \mathcal{P}_{p-1}(I)$ is again given by Theorem 4.6. The claim follows with similar arguments as above. In particular, the estimate (4.44) is again easy to prove:

$$\begin{aligned}\|\xi_p\|_{L^\infty(\widehat{T})} &\leq \|\psi_{p-1}\|_{L^\infty(I)}(\|w_p\|_{L^\infty(\gamma)} + \|w_p\|_{L^\infty(\gamma)}) \\ &\leq C\|w_p\|_{L^\infty(\gamma)}.\end{aligned}$$

Note that we have not defined $\psi_{p-1}$ for $p = 1$ in (4.45), but if $p = 1$, then $w_p(V_1) = w_p(V_2) = 0$ implies that $w_p = 0$ on $\gamma$ and then $\xi_p = 0$ obviously has the desired properties. □

We are now ready to take the above auxiliary theorems into use and consider the estimates (4.29) and (4.30), although we provide the full proof only for the second one.

**Theorem 4.9.** *Let $u \in H^2(\Omega)$ be such that it satisfies the condition $Bu = 0$. Then there exists a $u_p \in S(\Omega, \mathcal{M}, p)$ and a constant $C > 0$ independent of $u$ and $p$ such that*

$$\|u - u_p\|_{H^1(\Omega)} \leq Cp^{-1}\|u\|_{H^2(\Omega)}, \tag{4.46}$$

$$\|u - u_p\|_{L^\infty(\Omega)} \leq Cp^{-1}\|u\|_{H^2(\Omega)}. \tag{4.47}$$

*Proof.* Babuška and Suri [29, Theorem 4.1] prove the existence of an approximation $u_{p+1} \in S(\Omega, \mathcal{M}, p+1)$ such that (4.46) holds when the polynomial space in the definition of $S$ is the product space. The steps of their proof can essentially be mapped to applications of the above theorems 4.5, 4.6, 4.7 and 4.8. Let us prove the estimate (4.47) with the same approach but assume that the polynomial space is the trunk space. This is enough to cover the product space as well since $\mathcal{P}_p^{(tr)}(\widehat{E}) \subset \mathcal{P}_p^{(pr)}(\widehat{E})$. The arguments below could also easily be substituted into the proof of the estimate (4.46) in [29] to show that it holds for the trunk space as well and the $(p+1)$th order approximation $u_{p+1}$ can be replaced with a $p$th order approximation $u_p \in S(\Omega, \mathcal{M}, p)$.

Let $E_i \in \mathcal{M}$ for any $i \in \{1, 2, \ldots, N(\mathcal{M})\}$. Define $\hat{u}^{(i)} = u \circ F_i \in H^2(\widehat{E}_i)$. If $p$ is even, let $q = p/2 \in \mathbb{N}$. If $p \geq 2$ is odd, let $q = (p-1)/2 \in \mathbb{N}$. If $p = 1$, let $q = 1$. Note that $q^{-1} \leq 4p^{-1}$. Now by Theorem 4.5, depending on whether $\widehat{E}_i = \widehat{Q}$ or $\widehat{E}_i = \widehat{T}$, there exists a $\hat{u}_p^{(i)} \in \mathcal{P}_q^{(pr)}(\widehat{Q}) \subset \mathcal{P}_p^{(tr)}(\widehat{Q})$ or a $\hat{u}_p^{(i)} \in \mathcal{P}_q^{(tr)}(\widehat{T}) \subset \mathcal{P}_p^{(tr)}(\widehat{T})$ such that

$$\|\hat{u}^{(i)} - \hat{u}_p^{(i)}\|_{L^\infty(\widehat{E}_i)} \leq Cq^{-1}\|\hat{u}^{(i)}\|_{H^2(\widehat{E}_i)} \leq 4Cp^{-1}\|\hat{u}^{(i)}\|_{H^2(\widehat{E}_i)}, \tag{4.48}$$

where the constant $C > 0$ is independent of $u$ and $p$.

Define $u_p^{(i)} = \hat{u}_p^{(i)} \circ F_i^{-1}$ for all $i = 1, 2, \ldots, N$. We could try to define the desired approximation $u_p$ such that $u_p|_{E_i} = u_p^{(i)}$ for all $i = 1, 2, \ldots, N$, but the resulting function does not necessarily satisfy the continuity requirement or the condition that $Bu_p = 0$. In other words, $u_p \notin S$. However, if $u_p \in S$ were true, then (4.48) would imply (4.47) as we will see. Thus, the next step is to modify each $\hat{u}_p^{(i)}$ so that $u_p \in S$ while preserving the estimate (4.48). This can be achieved with the help of Theorem 4.7 and Theorem 4.8.

Let us begin with the continuity requirement. We first modify each $\hat{u}_p^{(i)}$ so that $u_p^{(i)}(V_j) = u(V_j)$ for each vertex $V_j$ of the element $E_i$. Let $\widehat{V}_j = F_i^{-1}(V_j)$ denote the corresponding vertex of the reference element $\widehat{E}_i$. By Theorem 4.7, there exists a polynomial $\hat{v}_q \in \mathcal{P}_q^{(pr)}(\widehat{Q})$ or $\hat{v}_q \in \mathcal{P}_q^{(tr)}(\widehat{T})$ depending on the type of $E_i$ such that $\hat{v}_q(\widehat{V}_j) = 1$ and $\hat{v}_q(\widehat{V}_k) = 0$ for all $k \neq j$, and it satisfies the pointwise estimate

$$\|\hat{v}_q\|_{L^\infty(\widehat{E}_i)} \leq C \tag{4.49}$$

for some constant $C > 0$ that is independent of $p$. Redefining now $\hat{u}_p^{(i)}$ to be the polynomial $\hat{u}_p^{(i)} + (u - u_p^{(i)})(V_j)\hat{v}_q \in \mathcal{P}_p^{(tr)}(\widehat{E}_i)$ satisfies the condition $u_p^{(i)}(V_j) = u(V_j)$. Moreover, the updated polynomial still satisfies the estimate (4.48):

$$\|\hat{u}^{(i)} - [\hat{u}_p^{(i)} + (u - u_p^{(i)})(V_j)\hat{v}_q]\|_{L^\infty(\widehat{E}_i)}$$

$$
\begin{aligned}
&\le \|\hat{u}^{(i)} - \hat{u}_p^{(i)}\|_{L^\infty(\widehat{E}_i)} + |(u - u_p^{(i)})(V_j)| \|\hat{v}_q\|_{L^\infty(\widehat{E}_i)} \\
&= \|\hat{u}^{(i)} - \hat{u}_p^{(i)}\|_{L^\infty(\widehat{E}_i)} + |(\hat{u}^{(i)} - \hat{u}_p^{(i)})(\widehat{V}_j)| \|\hat{v}_q\|_{L^\infty(\widehat{E}_i)} \\
&\le \|\hat{u}^{(i)} - \hat{u}_p^{(i)}\|_{L^\infty(\widehat{E}_i)} + \|\hat{u}^{(i)} - \hat{u}_p^{(i)}\|_{L^\infty(\widehat{E}_i)} \|\hat{v}_q\|_{L^\infty(\widehat{E}_i)} \\
&\le Cp^{-1} \|\hat{u}^{(i)}\|_{H^2(\widehat{E}_i)},
\end{aligned}
$$

where the last inequality follows from (4.48) and (4.49). This step is repeated for every vertex of every element, which results in the equality $u_p = u$ at every vertex of the mesh.

The next step is to achieve continuity across the sides of the elements. Let $E_i$ and $E_j$ be two elements with a common side $\gamma_{ij}$. Define the function $w_q^{(ij)} = u_p^{(j)} - u_p^{(i)}$ over the side $\gamma_{ij}$. Notice that $w_q^{(ij)}(V_1) = w_q^{(ij)}(V_2) = 0$, where $V_1$ and $V_2$ are the endpoints of $\gamma_{ij}$. Changing the variables via $F_i$ yields the polynomial $\hat{w}_q^{(ij)} = w_q^{(ij)} \circ F_i \in \mathcal{P}_q(\hat{\gamma}_i)$, where $\hat{\gamma}_i = F_i^{-1}(\gamma_{ij})$, which satisfies $\hat{w}_q^{(ij)}(\widehat{V}_1) = \hat{w}_q^{(ij)}(\widehat{V}_2) = 0$. Now by Theorem 4.8, there exists an extension $\hat{\xi}_q \in \mathcal{P}_q^{(pr)}(\widehat{Q})$ or $\hat{\xi}_q \in \mathcal{P}_{2q}^{(tr)}(\widehat{T}) \subset \mathcal{P}_p^{(tr)}(\widehat{T})$ of $\hat{w}_q^{(ij)}$ such that $\hat{\xi}_q = \hat{w}_q^{(ij)}$ on $\hat{\gamma}_i$ and $\hat{\xi}_q = 0$ on $\partial\widehat{E}_i \setminus \hat{\gamma}_i$, and it satisfies the pointwise estimate

$$
\begin{aligned}
\|\hat{\xi}_q\|_{L^\infty(\widehat{E}_i)} &\le C\|\hat{w}_q^{(ij)}\|_{L^\infty(\hat{\gamma}_i)} \\
&= C\|w_q^{(ij)}\|_{L^\infty(\gamma_{ij})} \\
&\le C\left(\|u - u_p^{(i)}\|_{L^\infty(\gamma_{ij})} + \|u - u_p^{(j)}\|_{L^\infty(\gamma_{ij})}\right) \\
&= C\left(\|\hat{u}^{(i)} - \hat{u}_p^{(i)}\|_{L^\infty(\hat{\gamma}_i)} + \|\hat{u}^{(j)} - \hat{u}_p^{(j)}\|_{L^\infty(\hat{\gamma}_j)}\right) \\
&\le C\left(\|\hat{u}^{(i)} - \hat{u}_p^{(i)}\|_{L^\infty(\widehat{E}_i)} + \|\hat{u}^{(j)} - \hat{u}_p^{(j)}\|_{L^\infty(\widehat{E}_j)}\right) \\
&\le Cp^{-1}\left(\|\hat{u}^{(i)}\|_{H^2(\widehat{E}_i)} + \|\hat{u}^{(j)}\|_{H^2(\widehat{E}_j)}\right).
\end{aligned} \tag{4.50}
$$

By redefining now $\hat{u}_p^{(i)}$ to be $\hat{u}_p^{(i)} + \hat{\xi}_q \in \mathcal{P}_p^{(tr)}(\widehat{E}_i)$, we achieve continuity over the side $\gamma_{ij}$ while keeping the other sides unmodified. Moreover, by the estimates (4.48) and (4.50), the new polynomial satisfies a slightly different form of the estimate (4.48):

$$
\begin{aligned}
\|\hat{u}^{(i)} - [\hat{u}_p^{(i)} + \hat{\xi}_q]\|_{L^\infty(\widehat{E}_i)} &\le \|\hat{u}^{(i)} - \hat{u}_p^{(i)}\|_{L^\infty(\widehat{E}_i)} + \|\hat{\xi}_q\|_{L^\infty(\widehat{E}_i)} \\
&\le Cp^{-1}\left(\|\hat{u}^{(i)}\|_{H^2(\widehat{E}_i)} + \|\hat{u}^{(j)}\|_{H^2(\widehat{E}_j)}\right).
\end{aligned} \tag{4.51}
$$

After repeating the above procedure for every side of every element, the approximation $u_p$ is continuous. Moreover, by (4.51), the approximation satisfies

$$
\|\hat{u}^{(i)} - \hat{u}_p^{(i)}\|_{L^\infty(\widehat{E}_i)} \le Cp^{-1}\left(\|\hat{u}^{(i)}\|_{H^2(\widehat{E}_i)} + \sum_{\overline{E}_i \cap \overline{E}_j = \gamma_{ij}} \|\hat{u}^{(j)}\|_{H^2(\widehat{E}_j)}\right) \tag{4.52}
$$

for all $i = 1, 2, \ldots, N$, where the summation iterates over the neighboring elements of $E_i$ with a common side $\gamma_{ij}$.

If the condition $Bu = 0$ corresponds to homogeneous Dirichlet boundary conditions, then the boundary values can be fixed with the exact same strategy that was used above to fix the values over the sides. After the boundary values have been fixed, the approximation $u_p$ belongs to the space $S$. We consider the pure Neumann case, i.e. the zero mean value requirement, after proving the estimate (4.47) for the approximation that we have now obtained.

The desired estimate (4.47) follows from the estimate (4.52):

$$\begin{aligned}\|u - u_p\|_{L^\infty(\Omega)} &= \max_{i=1,2,\dots,N} \|u - u_p\|_{L^\infty(E_i)} \\ &= \max_{i=1,2,\dots,N} \|\hat{u}^{(i)} - \hat{u}_p^{(i)}\|_{L^\infty(\widehat{E}_i)} \\ &\le \max_{i=1,2,\dots,N} Cp^{-1}\left(\|\hat{u}^{(i)}\|_{H^2(\widehat{E}_i)} + \sum_{\overline{E}_i \cap \overline{E}_j = \gamma_{ij}} \|\hat{u}^{(j)}\|_{H^2(\widehat{E}_j)}\right) \\ &\le \max_{i=1,2,\dots,N} Cp^{-1}\left(\|u\|_{H^2(E_i)} + \sum_{\overline{E}_i \cap \overline{E}_j = \gamma_{ij}} \|u\|_{H^2(E_j)}\right) \\ &\le Cp^{-1}\|u\|_{H^2(\Omega)}. \end{aligned} \tag{4.53}$$

The second to last inequality follows from the equivalence of the norms $\|\cdot\|_{H^2(\widehat{E}_i)}$ and $\|\cdot\|_{H^2(E_i)}$ under the change of variables $F_i : \widehat{E}_i \longrightarrow E_i$ [24, Theorem 1 on p. 13]. The same approximation also satisfies (4.46) [29]. This concludes the proof for the Dirichlet case.

When $Bu = 0$ corresponds to the zero mean value requirement, the desired approximation is given by $u_p - \overline{u}_p \in S$, where $u_p$ is the unnormalized approximation obtained above that already satisfies the desired estimates and

$$\overline{u}_p = \frac{1}{|\Omega|}\int_\Omega u_p \, dx.$$

This follows from the following result that the constant $\overline{u}_p$ is bounded by the same estimate:

$$\begin{aligned}|\overline{u}_p| &= |\overline{u} - \overline{u}_p| \\ &\le C\int_\Omega |u - u_p| \, dx \\ &\le C\|u - u_p\|_{L^2(\Omega)} \\ &\le C\|u - u_p\|_{H^1(\Omega)} \\ &\le Cp^{-1}\|u\|_{H^2(\Omega)}.\end{aligned}$$

We used above the assumption that $\overline{u} = 0$, Hölder's inequality and the estimate (4.46). Thus,

$$\|u - (u_p - \overline{u}_p)\|_{L^\infty(\Omega)} \le \|u - u_p\|_{L^\infty(\Omega)} + \|\overline{u}_p\|_{L^\infty(\Omega)} \le Cp^{-1}\|u\|_{H^2(\Omega)}.$$

A similar argument can be used to prove the estimate (4.46) as well. □

Consider then the usual weakly formulated boundary value problem

$$\text{Find } u \in V \text{ s.t. } a(u, v) = \varphi(v) \text{ for all } v \in V, \tag{4.54}$$

where the space $V \subset H^1(\Omega)$ is defined according to the type of the boundary value problem, $a : V \times V \to \mathbb{R}$ is an elliptic bounded bilinear mapping and $\varphi \in V'$. When the solution $u$ also belongs to the Sobolev space $H^2(\Omega)$, then Céa's lemma and Theorem 4.9 immediately imply that the corresponding finite element solutions converge towards the exact solution $u$ as the degree $p$ is increased.

**Theorem 4.10.** *Let $u \in V$ be the unique solution to the problem (4.54). Assume that $u \in H^2(\Omega)$. Let $u_S \in S(\Omega, \mathcal{M}, p)$ be the corresponding finite element solution. Then there exists a constant $C > 0$ independent of $u$ and $p$ such that*

$$\|u - u_S\|_{H^1(\Omega)} \leq Cp^{-1}\|u\|_{H^2(\Omega)}.$$

*Proof.* Let $u_p \in S$ be the approximation of $u$ provided by Theorem 4.9. Now by Céa's lemma, i.e. Theorem 4.3, we get that

$$\begin{aligned}
\|u - u_S\|_{H^1(\Omega)} &\leq C \inf_{v \in S}\|u - v\|_{H^1(\Omega)} \\
&\leq C\|u - u_p\|_{H^1(\Omega)} \\
&\leq Cp^{-1}\|u\|_{H^2(\Omega)},
\end{aligned}$$

where the constant $C > 0$ is independent of $u$ and $p$. □

Theorem 4.10 is the standard well-known result regarding the convergence of the $p$-version of the finite element method. In [29], it is shown that the rate of convergence $p^{-1}$ is optimal in the general case. The optimal convergence estimate for the $h$-version with first-order polynomials is analogous to Theorem 4.10 but the $p^{-1}$ is replaced with $h$, see e.g. [20].

Continuing with the $h$-version, there also exist error estimates in the $L^\infty(\Omega)$-norm. For example, in [20], it is shown that

$$\|u - u_S\|_{L^\infty(\Omega)} \leq Ch\|u\|_{H^2(\Omega)},$$

and [19] contains an even stronger estimate. For the $p$-version, however, similar pointwise estimates do not seem to be covered by the standard literature. Proving convergence estimates in the $L^\infty(\Omega)$-norm is in general more difficult than in the $H^1(\Omega)$-norm because Céa's lemma is not directly applicable. Nevertheless, we are still able to prove a pointwise estimate for the $p$-version by using a strategy identical to [20, p. 93]. The main enablers for this application are the pointwise estimate in Theorem 4.9 and a certain inverse estimate in [12]. We will need this result in the next section for the main theorem of this thesis.

**Theorem 4.11.** *Let $u \in V$ be the unique solution to the problem (4.54). Assume that $u \in H^2(\Omega)$. Let $u_S \in S(\Omega, \mathcal{M}, p)$ be the corresponding finite element solution. Then there exists a constant $C > 0$ independent of $u$ and $p$ such that*

$$\|u - u_S\|_{L^\infty(\Omega)} \leq Cp^{-1}\left(1 + \sqrt{\ln(p+1)}\right)\|u\|_{H^2(\Omega)}.$$

*Proof.* Let $u_p \in S$ be the approximation of $u$ provided by Theorem 4.9. The estimation is divided into two parts via the triangle inequality:

$$\|u - u_S\|_{L^\infty(\Omega)} \le \|u - u_p\|_{L^\infty(\Omega)} + \|u_p - u_S\|_{L^\infty(\Omega)}. \tag{4.55}$$

The first term can be estimated by (4.47) in Theorem 4.9. For the second term, notice that $u_p - u_S \in S$ and

$$\|u_p - u_S\|_{L^\infty(\Omega)} = \max_{i=1,\dots,N} \|u_p - u_S\|_{L^\infty(E_i)} = \max_{i=1,\dots,N} \|\hat{u}_p^{(i)} - \hat{u}_S^{(i)}\|_{L^\infty(\widehat{E}_i)}, \tag{4.56}$$

where $\hat{u}_p^{(i)} = u_p \circ F_i \in \mathcal{P}_p(\widehat{E}_i)$ and $\hat{u}_S^{(i)} = u_S \circ F_i \in \mathcal{P}_p(\widehat{E}_i)$.

By [12, Theorem 4.76 on p. 208] (covers both quadrilaterals and triangles) or [39, Proposition 3.1] (covers only triangles), the polynomials $\hat{u}_p^{(i)} - \hat{u}_S^{(i)} \in \mathcal{P}_p(\widehat{E}_i)$ satisfy the inverse inequality

$$\|\hat{u}_p^{(i)} - \hat{u}_S^{(i)}\|_{L^\infty(\widehat{E}_i)} \le C\sqrt{\ln(p+1)}\|\hat{u}_p^{(i)} - \hat{u}_S^{(i)}\|_{H^1(\widehat{E}_i)}, \tag{4.57}$$

where the constant $C > 0$ is independent of $u$ and $p$.

Define $\hat{u}^{(i)} = u \circ F_i \in H^2(\widehat{E}_i)$ for all $i = 1, 2, \dots, N$. The $H^1$-norm on the right-hand side of (4.57) can be estimated by

$$\begin{aligned}\|\hat{u}_p^{(i)} - \hat{u}_S^{(i)}\|_{H^1(\widehat{E}_i)} &\le \|\hat{u}_p^{(i)} - \hat{u}^{(i)}\|_{H^1(\widehat{E}_i)} + \|\hat{u}^{(i)} - \hat{u}_S^{(i)}\|_{H^1(\widehat{E}_i)} \\ &\le C\|u_p - u\|_{H^1(E_i)} + C\|u - u_S\|_{H^1(E_i)} \\ &\le C\|u_p - u\|_{H^1(\Omega)} + C\|u - u_S\|_{H^1(\Omega)} \\ &\le Cp^{-1}\|u\|_{H^2(\Omega)}, \end{aligned} \tag{4.58}$$

where we used the equivalence of the norms $\|\cdot\|_{H^1(\widehat{E}_i)}$ and $\|\cdot\|_{H^1(E_i)}$ under the change of variables $F_i : \widehat{E}_i \to E_i$ and then Theorems 4.9 and 4.10.

Substituting (4.58) into (4.57) and then (4.57) into (4.56), the second term in (4.55) has the bound

$$\|u_p - u_S\|_{L^\infty(\Omega)} \le Cp^{-1}\sqrt{\ln(p+1)}\|u\|_{H^2(\Omega)}.$$

We now arrive at the desired estimate:

$$\begin{aligned}\|u - u_S\|_{L^\infty(\Omega)} &\le \|u - u_p\|_{L^\infty(\Omega)} + \|u_p - u_S\|_{L^\infty(\Omega)} \\ &\le Cp^{-1}\|u\|_{H^2(\Omega)} + Cp^{-1}\sqrt{\ln(p+1)}\|u\|_{H^2(\Omega)} \\ &\le Cp^{-1}\left(1 + \sqrt{\ln(p+1)}\right)\|u\|_{H^2(\Omega)}.\end{aligned}$$

□

# 5 $p$- Finite Element Method for Poisson's Equation with a Dirac Delta Load

We are now ready to consider the $L^2$ convergence of the $p$- finite element method for the problem

$$\begin{cases} -\Delta u = \delta_{x_0} & \text{in } \Omega \\ u = g_j & \text{on } \Gamma_j, \quad j \in D \\ \dfrac{\partial u}{\partial n} = g_j & \text{on } \Gamma_j, \quad j \in N, \end{cases} \tag{5.1}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded polygonal domain, $\delta_{x_0}$ is the Dirac delta functional for an arbitrary point $x_0 \in \Omega$ and the boundary data satisfy $g_j \in T(H^2(\Omega))$ for all $j \in D$ and $g_j \in T(H^1(\Omega))$ for all $j \in N$. If $D = \emptyset$, the Neumann boundary data are assumed to satisfy the usual compatibility condition. It is also assumed that the domain $\Omega$ satisfies Assumption 3.1. Uniqueness of the solution for the pure Neumann problem is again enforced with the zero mean value requirement. We consider the $L^2$ convergence both analytically and numerically.

## 5.1 An $L^2$ Error Estimate

Casas [6] provides an $L^2$ error estimate for the $h$-version when applied to the pure Dirichlet variant of the problem (5.1) with homogeneous boundary conditions. The same proof strategy can be used to consider the $p$-version as well, and for that we will need the error estimates in the previous section. We consider both Dirichlet and Neumann boundary conditions, and we also assume that the Dirichlet conditions are homogeneous to simplify the proof. As in Section 3, a problem with non-homogeneous Dirichlet boundary conditions could be transformed into a new problem with homogeneous Dirichlet boundary conditions to which the finite element method could then be applied.

**Theorem 5.1.** *Consider the problem (5.1) with homogeneous Dirichlet boundary conditions. Assume that the domain $\Omega$ satisfies Assumption 3.1. Let $u \in W^{1,s}(\Omega)$, where $s \in (1,2)$, denote the exact solution, and let $u_S \in S(\Omega, \mathcal{M}, p)$ denote the finite element solution. Then there exists a constant $C > 0$ independent of p such that*

$$\|u - u_S\|_{L^2(\Omega)} \leq C p^{-1} \left(1 + \sqrt{\ln(p+1)}\right).$$

*Proof.* By the Sobolev imbedding theorem, $u \in L^2(\Omega)$. Thus, $u - u_S \in L^2(\Omega)$. Consider now the weak formulation of the problem

$$\begin{cases} -\Delta v = u - u_S & \text{in } \Omega \\ v = 0 & \text{on } \Gamma_j, \quad j \in D \\ \dfrac{\partial v}{\partial n} = 0 & \text{on } \Gamma_j, \quad j \in N. \end{cases} \tag{5.2}$$

By Theorem 3.5, the problem has a unique solution $v$, and by Theorem 3.6 and Theorem 3.7, it holds that $v \in H^2(\Omega)$ and $-\Delta v = u - u_S$ almost everywhere in $\Omega$. Moreover, by Theorem 3.8, the solution $v$ satisfies the estimate

$$\|v\|_{H^2(\Omega)} \leq C\|u - u_S\|_{L^2(\Omega)}, \tag{5.3}$$

where the constant $C > 0$ is independent of $u$ and $p$.

We may use the function $v$ to rewrite $\|u - u_S\|^2_{L^2(\Omega)}$. Substituting $u - u_S = -\Delta v$ and integrating by parts gives

$$\begin{aligned}
\|u - u_S\|^2_{L^2(\Omega)} &= \int_\Omega |u - u_S|^2 \, dx \\
&= -\int_\Omega (u - u_S)\Delta v \, dx \\
&= \int_\Omega \nabla(u - u_S) \cdot \nabla v \, dx - \int_{\partial\Omega} \frac{\partial v}{\partial n}(u - u_S) \, dS \\
&= \int_\Omega \nabla u \cdot \nabla v \, dx - \int_\Omega \nabla u_S \cdot \nabla v \, dx.
\end{aligned} \tag{5.4}$$

For the last equality, note that $u - u_S = 0$ on $\Gamma_j$ for all $j \in D$ and $\partial v/\partial n = 0$ on $\Gamma_j$ for all $j \in N$.

Since $v \in H^2(\Omega)$, the Sobolev imbedding theorem implies that $v \in W^{1,s'}(\Omega)$, where $s' \in (2, \infty)$ is the conjugate exponent of $s$. Moreover, $v = 0$ on $\Gamma_j$ for all $j \in D$. Hence, by the definition of a weak solution to the problem (5.1), the first integral in (5.4) can be written as

$$\int_\Omega \nabla u \cdot \nabla v \, dx = v(x_0) + \sum_{j \in N} \int_{\Gamma_j} g_j v \, dS. \tag{5.5}$$

Let $v_S \in S$ denote the finite element solution to the problem (5.2). The second integral in (5.4) can now be written as

$$\int_\Omega \nabla u_S \cdot \nabla v \, dx = \int_\Omega \nabla u_S \cdot \nabla v_S \, dx = v_S(x_0) + \sum_{j \in N} \int_{\Gamma_j} g_j v_S \, dS. \tag{5.6}$$

The first equality follows the Galerkin orthogonality between $v$ and $v_S$. The second equality follows from the definition of $u_S$ as the finite element solution to the problem (5.1).

Substituting (5.5) and (5.6) into (5.4) and applying Hölder's inequality and the trace theorem yield

$$\begin{aligned}
\|u - u_S\|^2_{L^2(\Omega)} &= v(x_0) - v_S(x_0) + \sum_{j \in N} \int_{\Gamma_j} g_j (v - v_S) \, dS \\
&\leq \|v - v_S\|_{L^\infty(\Omega)} + \sum_{j \in N} \|g_j\|_{L^2(\Gamma_j)} \|v - v_S\|_{L^2(\Gamma_j)} \\
&\leq \|v - v_S\|_{L^\infty(\Omega)} + C\|v - v_S\|_{H^1(\Omega)}.
\end{aligned}$$

The errors $\|v - v_S\|_{L^\infty(\Omega)}$ and $\|v - v_S\|_{H^1(\Omega)}$ can be further estimated by Theorem 4.11, Theorem 4.10 and the estimate (5.3):

$$\begin{aligned}\|u - u_S\|^2_{L^2(\Omega)} &\leq Cp^{-1}\left(1 + \sqrt{\ln(p+1)}\right)\|v\|_{H^2(\Omega)} + Cp^{-1}\|v\|_{H^2(\Omega)} \\ &\leq Cp^{-1}\left(1 + \sqrt{\ln(p+1)}\right)\|u - u_S\|_{L^2(\Omega)}.\end{aligned}$$

Dividing both sides by $\|u - u_S\|_{L^2(\Omega)}$ completes the proof. □

## 5.2 Numerical Results

Let us investigate the accuracy of the error estimate in Theorem 5.1 through numerical experiments. We consider the pure Neumann variant of the problem (5.1) in the domain $\Omega = (-1, 1)^2$ with several locations for the load $x_0$. The boundary conditions are chosen such that an exact solution is given by the Green's function

$$u(x) = -\frac{1}{2\pi}\ln|x - x_0|.$$

The exact solution and the finite element solution are defined up to an additive constant. Thus, they are normalized to have zero mean value over the domain $\Omega$ to fix the solutions.

We consider two meshes: one that consists of only quadrilaterals and one that consists of only triangles. For both meshes, we consider three locations for the Dirac delta load: load at a vertex, on a side and in the interior of an element. The different configurations are illustrated in Figure 5. The exact locations of the load are $(0, 0)$, $(1/4, 0)$ and $(1/4, 1/4)$ for the quadrilateral mesh and $(0, 0)$, $(1/2, 0)$ and $(1/3, 1/3)$ for the triangle mesh. We consider the product and trunk polynomial spaces for both meshes such that a given configuration uses the same polynomial space for all elements of the mesh.

The numerical error caused by finite-precision floating-point arithmetic has been tried to be kept close to a minimum by performing most of the computations exactly

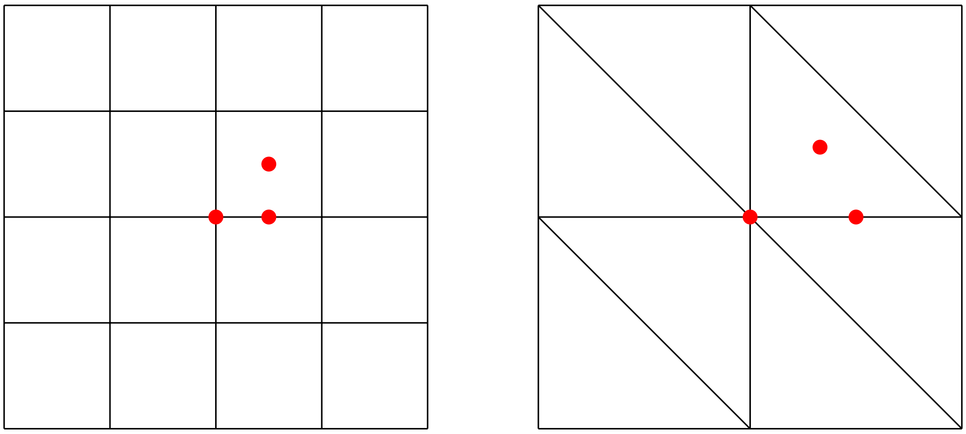
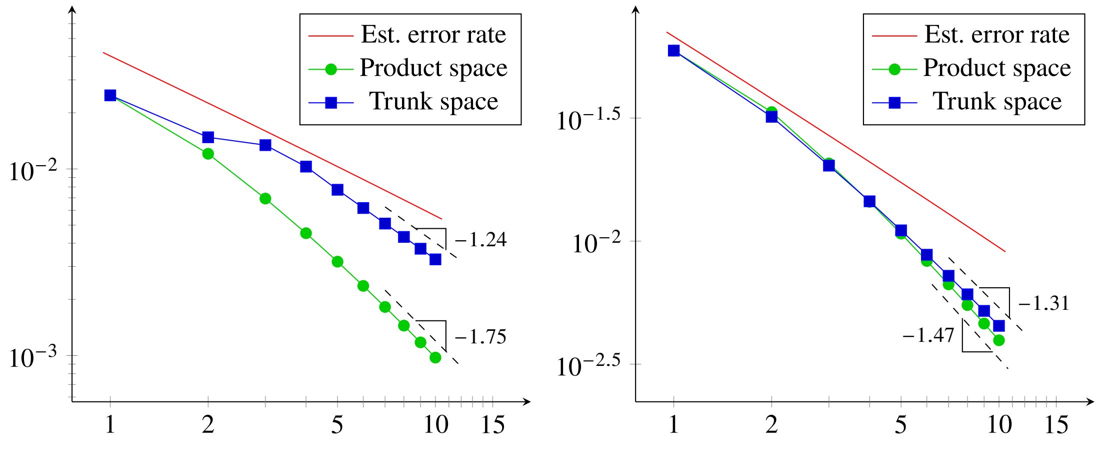


**Figure 5:** Meshes and locations of the Dirac delta load for the domain $\Omega = (-1, 1)^2$.

in rational number arithmetic with the GNU Multiple Precision Arithmetic Library (GMP) [40]. For example, the stiffness matrix has been assembled exactly with no numerical error by using the shape polynomials presented in Section 4.2. After the elimination of one row and column from the linear system, as described at the end of Section 4.2 for solving the Neumann problem, the resulting linear system is also solved exactly by using the Eigen software library [41] in combination with the GMP library. The only sources of numerical error are the computation of the load vector due to the Neumann boundary integrals and the computation of the error $\|u - u_S\|_{L^2(\Omega)}$. Both of those have been approximated with the Gauss-Legendre quadrature rule for which we have used precomputed weights and abscissae with double precision provided by the GNU Scientific Library [42]. Quadrature over the reference triangle has been implemented by mapping the quadrature points for the reference quadrilateral to the reference triangle and then mitigating the crowding of points caused by such a mapping, see [25] for more details.

Figures 6, 7 and 8 below show the error $\|u - u_S\|_{L^2(\Omega)}$ as a function of $p \in \{1, 2, \ldots, 10\}$ in log-log scale when the Dirac delta load is located at a vertex, on a side and in the interior of an element, respectively. Each figure shows the errors for both meshes and for both product and trunk spaces. The estimated error rate in Theorem 5.1 has also been plotted for comparison. The constant $C$ in Theorem 5.1 is not known, but note that changing it only shifts the error curve vertically when plotted in log-log scale. The constant is simply chosen so that the estimated error curve is close to the observed errors.

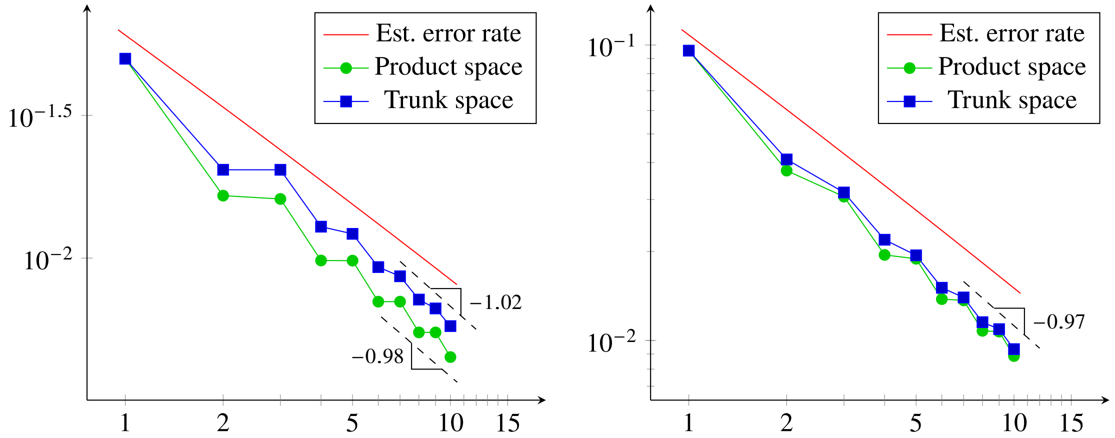
To quantify the exact observed error rates, there are regression lines and their slopes next to each observed error curve. The lines have been obtained via the method of least squares from some of the last observed data points, and they have been shifted slightly for clarity. Note that a linear error curve in log-log scale corresponds to the



**(a)** Quadrilateral mesh; $x_0 = (0, 0)$.

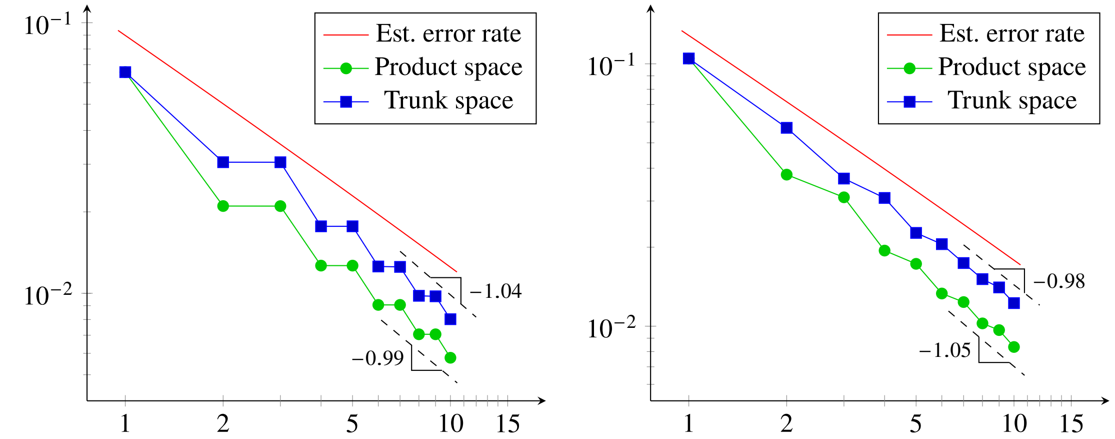
**(b)** Triangle mesh; $x_0 = (0, 0)$.

**Figure 6:** Log-log plots of $\|u - u_S\|_{L^2(\Omega)}$ vs. $p$ for the quadrilateral and triangle meshes when the load is at a vertex. The slope triangles are used to estimate the exact convergence rates, and they have been computed via linear regression.



**(a)** Quadrilateral mesh; $x_0 = (1/4, 0)$. **(b)** Triangle mesh; $x_0 = (1/2, 0)$.

**Figure 7:** Same as Figure 6 but when the load is on a side.



**(a)** Quadrilateral mesh; $x_0 = (1/4, 1/4)$. **(b)** Triangle mesh; $x_0 = (1/3, 1/3)$.

**Figure 8:** Same as Figure 6 but when the load is in the interior of an element.

error bound

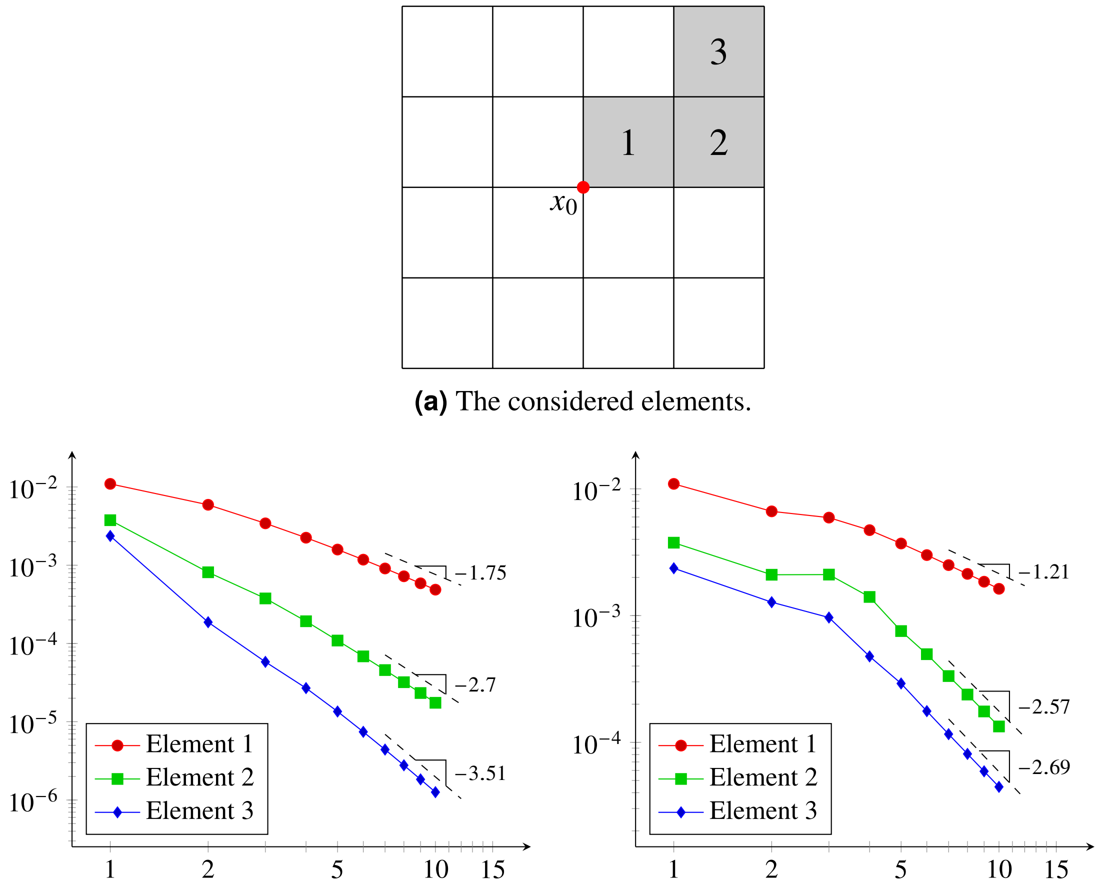
$$\|u - u_S\|_{L^2(\Omega)} \leq Cp^{-k},$$

where $C > 0$ is some constant independent of $p$ and $k > 0$ is the absolute value of the slope.

We may conclude from Figure 6 that when the load is located at a vertex, the $L^2$ error decays at a clearly higher rate than what Theorem 5.1 predicts for both quadrilateral and triangular meshes and for both product and trunk spaces. For the quadrilateral elements, the product space is a better choice than the trunk space when measured in terms of convergence rate, which is not surprising. The internal shape functions for quadrilaterals seem to be important for improving the quality of the approximation as can be seen when comparing, for example, the transitions from $p = 2$ to $p = 3$ and from $p = 3$ to $p = 4$ for the trunk space. Recall that the trunk space for quadrilaterals does not contain any internal shape functions until $p = 4$. The difference between the product and trunk spaces for triangular elements is not as striking. The asymptotic rate of convergence for the product space seems to be again better, however.

Based on Figure 7, when the Dirac delta load is on a side, Theorem 5.1 only slightly overestimates the true rate of convergence. The error is in general larger for the same values of $p$ when compared to the load being at a vertex in Figure 6. The error also decreases more irregularly. The decrease in error is larger for even values of $p$ than for odd values of $p$. The error may even remain the same for odd values of $p$ as can be seen for the product space especially. The load is set to be in the middle of the side for both meshes, but a similar phenomenon can be observed for other locations on the side as well, where the decrease in error fluctuates periodically. The above discussion also applies to the case where the load is in the interior of an element in Figure 8. For both cases, choosing between the product space and the trunk space does not seem to have a large effect on the asymptotic convergence rate of the error.

All in all, the results above suggest that Theorem 5.1 is more applicable to the cases

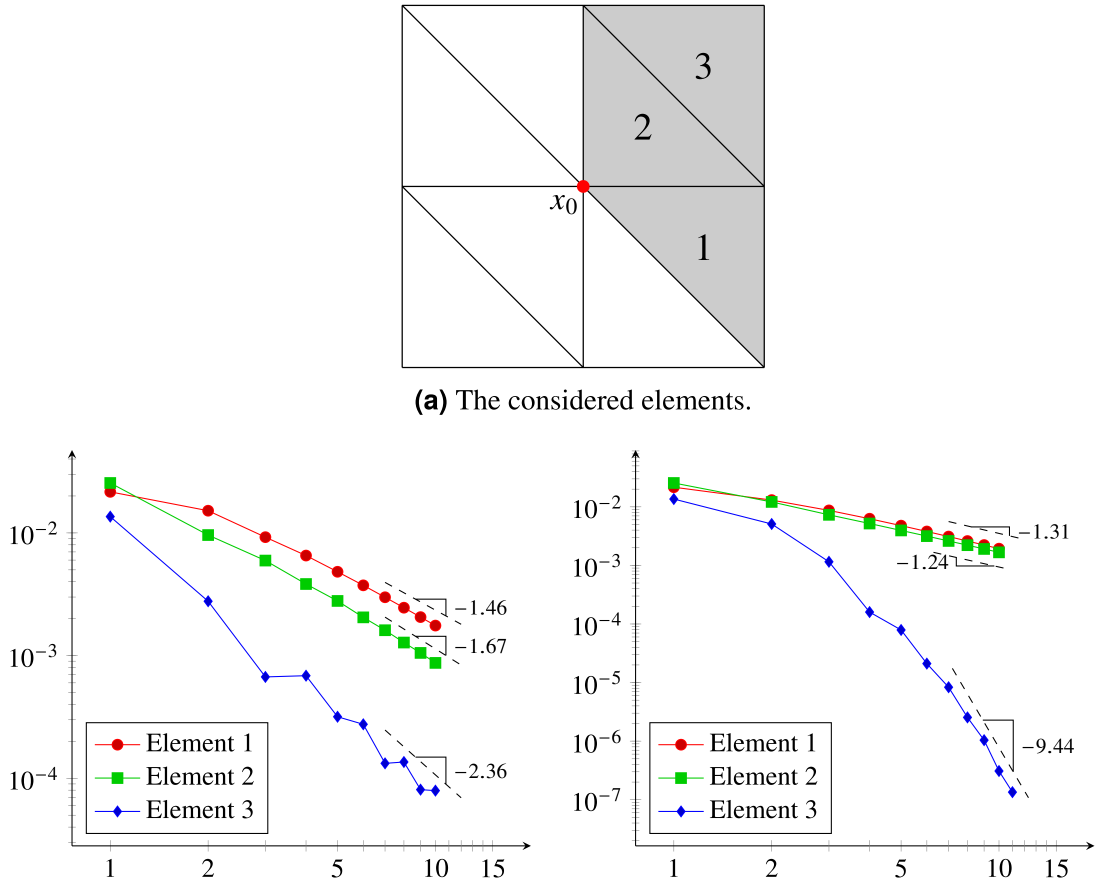


**(a)** The considered elements.

**(b)** Product space.

**(c)** Trunk space.

**Figure 9:** Log-log plots of $\|u - u_S\|_{L^2(E_i)}$ vs. $p$ over individual elements of the quadrilateral mesh. The Dirac delta load is at $x_0 = (0, 0)$.



**(a)** The considered elements.

**(b)** Product space.

**(c)** Trunk space.

**Figure 10:** Log-log plots of $\|u - u_S\|_{L^2(E_i)}$ vs. $p$ over individual elements of the triangle mesh. The Dirac delta load is at $x_0 = (0, 0)$.

where the Dirac delta load is located either on a side or in the interior of an element. It seems more beneficial to construct the mesh so that the load is at a vertex since, based on the examples above, this choice results in much better convergence rate and more robust error behavior. Let us now briefly consider this case further by investigating how the $L^2$ error is distributed between the individual elements. Figures 9 and 10 show the element-wise errors for the quadrilateral mesh and the triangle mesh, respectively. Due to symmetry, it suffices to only consider the designated elements to cover the other elements as well.

Unsurprisingly, the element-wise error is the largest and the rate of convergence is the slowest in the elements that overlap with the Dirac delta load. The rate of convergence in those elements also determines the rate of convergence in the whole domain as can be seen when comparing the results to Figure 6. The error converges significantly faster when the element does not overlap with the load. Thus, it would make sense to refine the mesh around the load to improve the error over the whole domain. In general, the further away the element is from the load, the faster the convergence seems to be. The convergence can even be exponential as can be seen for the trunk space for triangular elements. Interestingly, the convergence is not exponential when the trunk space is replaced with the product space or when triangles are replaced with quadrilaterals. Although we have only provided the computations for the case where the load is at a vertex, similar observations apply to the other locations of the load as well.

The computations presented above were performed using computer resources within the Aalto University School of Science “Science-IT” project.

# 6 Summary

In Section 1, we set out to study Poisson's equation with a Dirac delta load term in the context of the $p$-version of the finite element method for which we defined two objectives. The first objective was to consider the unique solvability of the problem with Dirichlet and Neumann boundary values in two-dimensional bounded polygonal domains. The second objective was to study the convergence of the $p$-version of the finite element method when applied to the same problem, which does not seem to be covered by the existing literature.

For the first objective, the standard theory for solving elliptic boundary value problems in a Hilbert space is not alone applicable to assert the existence of a solution to the Dirac delta problem simply because the problem cannot be formulated in the usual Hilbert space setting. It is, however, a well-known fact from the classical theory of partial differential equations that the Green's function for the Laplacian is a solution to the Dirac delta problem when the boundary values are set accordingly. We used the Green's function in combination with the standard Hilbert space PDE theory to show that the Dirac delta problem with general boundary data has a solution. Moreover, when the domain satisfies certain convexity assumptions, we were able to show that the solution is unique. This is in line with the existence and uniqueness result by Casas [6] who considers the homogeneous Dirichlet problem.

For the second objective, we proved the $L^2$ convergence of the $p$-version of the finite element method for the Dirac delta problem under the same convexity assumptions that were used to prove the uniqueness of the solution. We used the same proof strategy that Casas [6] used to prove $L^2$ convergence for the $h$-version. The proof relies on standard error estimates for the $p$-version and on some additional pointwise estimates that do not seem to be covered by the existing standard literature. The proofs for the additional estimates are largely based on the results of Babuška and Suri [29] and Schwab [12].

Finally, the accuracy of the obtained $L^2$ error bound was assessed by applying the $p$- finite element method to the Neumann problem whose exact solution is given by the Green's function mentioned before. We considered the error for several different configurations: quadrilateral and triangular elements, different locations for the Dirac delta load and different polynomial space types. The proven $L^2$ error bound applies to all of these configurations. When the Dirac delta load is on a side or in the interior of an element, the error bound overestimates the exact error rate only slightly. When the load is at a vertex of the mesh, the predicted error convergence rate is overly pessimistic, and the element-wise convergence rate can even be exponential for a triangular element that does not overlap with the load.

Nothing has been said about the optimality of the obtained $L^2$ error bound, which could be the topic for a future study. The numerical results suggest that the bound could be improved. Moreover, the behavior of the error heavily depends on the location of the Dirac delta load, which is not accounted for by the obtained error bound. Two additional open questions are whether the Dirac delta problem still admits a unique solution when the convexity assumption is dropped and, if so, whether Theorem 5.1 is still true.

# References

[1] L. C. Evans, *Partial Differential Equations*, vol. 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2. ed., 2010.

[2] J. Tinsley Oden, “Finite elements: An introduction,” in *Finite Element Methods (Part 1)*, vol. 2 of *Handbook of Numerical Analysis*, pp. 3–15, Elsevier, 1991.

[3] B. Szabó and I. Babuška, *Introduction to Finite Element Analysis: Formulation, Verification and Validation*. John Wiley & Sons, 1. ed., 2011.

[4] I. Babuška, A. M. Soane, and M. Suri, “The computational modeling of problems on domains with small holes,” *Computer Methods in Applied Mechanics and Engineering*, vol. 322, pp. 563–589, 2017.

[5] S. Wang, A. Lee, E. Alexov, and S. Zhao, “A regularization approach for solving Poisson’s equation with singular charge sources and diffuse interfaces,” *Applied Mathematics Letters*, vol. 102, p. 106144, 2020.

[6] E. Casas, “$L^2$ Estimates for the Finite Element Method for the Dirichlet Problem with Singular Data,” *Numerische Mathematik*, vol. 47, pp. 627–632, 1985.

[7] R. Scott, “Finite Element Convergence for Singular Data,” *Numerische Mathematik*, vol. 21, p. 317–327, 1973.

[8] A. H. Schatz and L. B. Wahlbin, “Interior Maximum Norm Estimates for Finite Element Methods,” *Mathematics of Computation*, vol. 31, no. 138, pp. 414–442, 1977.

[9] F. Millar, I. Muga, S. Rojas, and K. G. van der Zee, “Projection in negative norms and the regularization of rough linear functionals,” *Numerische Mathematik*, vol. 150, pp. 1087–1121, 2021.

[10] R. Araya, E. Behrens, and R. Rodríguez, “A posteriori error estimates for elliptic problems with Dirac delta source terms,” *Numerische Mathematik*, vol. 105, pp. 193–216, 2006.

[11] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*. Society for Industrial and Applied Mathematics, 2011.

[12] C. Schwab, *p- and hp- Finite Element Methods: Theory and Applications in Solid and Fluid Mechanics*. Numerical Mathematics and Scientific Computation, Clarendon Press, 1998.

[13] W. Rudin, *Functional Analysis*. International Series in Pure and Applied Mathematics, McGraw Hill, 2. ed., 1991.

[14] W. Rudin, *Real and Complex Analysis*. McGraw-Hill Series in Higher Mathematics, McGraw Hill, 3. ed., 1986.

[15] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts, John Wiley & Sons, 2. ed., 1999.

[16] J. Nečas, *Direct Methods in the Theory of Elliptic Equations*. Springer Monographs in Mathematics, Springer Berlin Heidelberg, 1. ed., 2011. Originally published in French “Les méthodes directes en théorie des équations elliptiques” by Academia, Praha, and Masson et Cie, Editeurs, Paris, 1967.

[17] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces*, vol. 140 of *Pure and Applied Mathematics Series*. Academic Press, 2. ed., 2003.

[18] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, vol. 1 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 1. ed., 1972.

[19] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, vol. 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, 2002.

[20] D. Braess, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, 3. ed., 2007.

[21] P. Grisvard, “Behavior of the Solutions of an Elliptic Boundary Value Problem in a Polygonal or Polyhedral Domain,” in *Numerical Solution of Partial Differential Equations–III* (B. Hubbard, ed.), pp. 207–274, Academic Press, 1976.

[22] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*. Texts in Applied Mathematics, Springer New York, 3. ed., 2007.

[23] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*. Prentice-Hall Series in Automatic Computation, Prentice-Hall, 1973.

[24] V. Maz’ya, *Sobolev Spaces: with Applications to Elliptic Partial Differential Equations*, vol. 342 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2. ed., 2011.

[25] F. Hussain, M. S. Karim, and R. Ahamad, “Appropriate Gaussian quadrature formulae for triangles,” *International Journal of Applied Mathematics and Computation*, vol. 4, pp. 23–38, 2012.

[26] M. Islam and M. A. Hossain, “Numerical integrations over an arbitrary quadrilateral region,” *Applied Mathematics and Computation*, vol. 210, pp. 515–524, April 2009.

[27] I. Babuška, B. A. Szabo, and I. N. Katz, “The $p$-Version of the Finite Element Method,” *SIAM Journal on Numerical Analysis*, vol. 18, no. 3, pp. 515–545, 1981.

[28] M. R. Dorr, “The Approximation Theory for the $p$-Version of the Finite Element Method,” *SIAM Journal on Numerical Analysis*, vol. 21, no. 6, pp. 1180–1207, 1984.

[29] I. Babuška and M. Suri, “The Optimal Convergence Rate of the $p$-Version of the Finite Element Method,” *SIAM Journal on Numerical Analysis*, vol. 24, no. 4, pp. 750–776, 1987.

[30] B. Szabó, A. Düster, and E. Rank, “The $p$-version of the Finite Element Method,” in *Encyclopedia of Computational Mechanics*, ch. 5, John Wiley & Sons, Ltd, 2004.

[31] I. Babuška and M. R. Dorr, “Error Estimates for the Combined $h$ and $p$ Versions of the Finite Element Method,” *Numerische Mathematik*, vol. 37, pp. 257–278, 1981.

[32] B. Guo and I. Babuška, “The $h$-$p$ version of the finite element method,” *Computational Mechanics*, vol. 1, pp. 21–41, 1986.

[33] I. Babuška and M. Suri, “The $h - p$ version of the finite element method with quasiuniform meshes,” *M2AN - Modélisation mathématique et analyse numérique*, vol. 21, no. 2, pp. 199–238, 1987.

[34] I. Babuška and M. Suri, “The P and H-P Versions of the Finite Element Method, Basic Principles and Properties,” *SIAM Review*, vol. 36, no. 4, pp. 578–632, 1994.

[35] L. C. Andrews, *Special Functions of Mathematics for Engineers*. SPIE Press, 2. ed., 1998.

[36] J. Céa, “Approximation variationnelle des problèmes aux limites,” *Annales de l’Institut Fourier*, vol. 14, no. 2, pp. 345–444, 1964.

[37] W. Gui and I. Babuška, “The $h$, $p$ and $hp$ Versions of the Finite Element Method in One Dimension. Part 1: The Error Analysis of the $p$-Version. Part 2: The Error Analysis of the $h$- and $hp$-Versions. Part 3: The Adaptive $hp$-Versions,” *Numerische Mathematik*, vol. 49, pp. 577–683, 1986.

[38] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press, 1970.

[39] E. Boillat, “On a Right-Inverse for the Divergence Operator in Spaces of Continuous Piecewise Polynomials,” *Mathematical Models and Methods in Applied Sciences*, vol. 7, no. 4, pp. 487–505, 1997.

[40] “The GNU MP Bignum Library.” [Online]. `https://gmplib.org/` (accessed 19 March 2024).

[41] "Eigen." [Online]. https://eigen.tuxfamily.org/ (accessed 19 March 2024).

[42] "GSL - GNU Scientific Library." [Online]. https://www.gnu.org/software/gsl/ (accessed 19 March 2024).