Freie Universität Berlin
Department of Mathematics and Computer Science
Institute of Mathematics

Bachelor Thesis

Bernstein Polynomials and their Application in the Finite Element Method

Student: Tianyi Hu
Matriculation Number: 5262896
Supervisor: Prof. Dr. Volker John
Second Reviewer: Dr. Alfonso Caiazzo

Berlin, 10 September 2021

Contents

Chapter 1

Introduction

Remark 1.1. *Contents of the bachelor thesis.* To solve differential equations numerically, the finite element method (FEM) is widely employed. Choosing an appropriate set of basis functions is important for the finite element method. Normally one takes so-called Lagrange polynomials, such as continuous and piecewise linear functions. However, Bernstein basis polynomials can also be used as a set of basis functions. For higher degree, Bernstein basis polynomials take only non-negative values, in contrast to Lagrange polynomials.

Chapter 2 first gives a brief introduction of the biography of the Russian mathematician Sergei Bernstein. Then we will introduce the Bernstein basis polynomials, the Bernstein polynomials and their properties. The Bernstein polynomials were used in a constructive proof of the Weierstrass approximation theorem, which will be shown at the end of this chapter.

In Chapter 3 we will introduce the model problem and investigate the solution of the model problem by analyzing linear homogeneous and inhomogeneous differential equations respectively. Main parts of this chapter follow the lecture notes of Prof. Dr. Volker John, see [4].

Chapter 4 presents the weak formulation and the minimization problem of a somewhat simplified model problem, namely the Poisson equation with homogeneous Dirichlet boundary conditions. The weak formulation and the minimization problem are more general than the strong formulation in Chapter 3.

Based on the weak formulation, the finite element method is used to solve the simplified model problem. Chapter 5 gives derivations and examples of the matrix-vector form for hat functions, piecewise quadratic basis functions and Bernstein basis polynomials. The bachelor thesis ends with the Mathematica programming of the Ritz method for a one-dimensional example. □

Chapter 2

Bernstein Polynomials

The idea of Bernstein polynomials is named after the Russian mathematician Sergei Bernstein. Let us first take a look at the biography of this reowned figure (see also [11] for full details).

2.1 The Biography of Sergei Bernstein

Sergei Natanowitsch Bernstein (Russian: Сергей Натанович Бернштейн) was born into a family of doctors in Odessa, nowadays a port city in southern Ukraine, in March 1880. His family was of Jewish origin. After graduating from high school education in 1898, he moved to Paris and studied mathematics at the Sorbonne. During his studies, he spent three semesters at the University of Göttingen, where he conducted his studies under the supervision of David Hilbert[1], who was one of the most influential German mathematicians of the nineteenth century. In 1904, he submitted and defended his doctoral dissertation about Hilbert's 19^{th} Problem, which considers the analytic solutions of elliptic differential equations.

However, since Russia did not recognise European academic qualifications, Bernstein went to St. Petersburg and started his second mathematical doctoral program. Accordingly, in 1906 he earned his first Russian master's degree there. Afterwards, he moved to Kharkov in 1908 and obtained another master's degree for his thesis *Investigation and Solution of Elliptic Partial Differential Equations of Second Degree*. Following his studies, he became a lecturer at the Kharkov University and eventually received his doctorate from the institution for his thesis titled *About the Best Approximation of Continuous Functions by Polynomials of Given Degree* in 1913.

For about 25 years Bernstein taught at the Kharkov University. He became a professor at the Kahrkov University in 1920. In 1933, he left to work at the Mathematical Institute of the USSR Academy of Sciences in Leningrad. Unfortunately, life was not smooth sailing for him. In the summer of 1941,

[1] David Hilbert (1862 - 1943)

Figure 2.1: Sergei Natanovich Bernstein (1880–1968)[3]

the rapid encroachment of German armies upon Leningrad caused Bernstein to flee his home and escape to Moscow. Then he taught at the University of Moscow, and in the years that followed, he devoted much time to the editing and publication of P.L. Chebyshev's[2] remaining works, see [6].

During his life, he made significant contributions to partial differential equations, differential geometry, probability theory and approximation theory.

2.2 Definition and Properties

As far back as 1911, Bernstein had already introduced the concept of now called the Bernstein polynomials in his paper *Démonstration du théoréme de Weierstrass fondée sur le calcul des probabilités*[1].

Definition 2.1. Bernstein basis polynomials. For $0 \leq k \leq n$, the $n+1$ Bernstein basis polynomials of degree n on $x \in [0,1]$ are defined as

$$b_k^n(x) := \binom{n}{k}(1-x)^{n-k}x^k, \tag{2.1}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ and $n, k \in \mathbb{N}_0$.
For $k < 0$ and $k > n$, it is defined that $b_k^n(x) \equiv 0$. □

Remark 2.2. *Domain of the Bernstein basis polynomials.* The Bernstein basis polynomials can be defined on any interval $[a,b]$. One replaces x by $\frac{t-a}{b-a}$, which

[2] Pafnuty Lvovich Chebyshev (1821 - 1894)
[3] Portrait collected from the Russian Academy of Sciences (RAS) - see `http://www.ras.ru`

maps $t \in [a, b]$ to $x \in [0, 1]$. Then

$$\begin{aligned} b_k^n\left(\frac{t-a}{b-a}\right) &= \binom{n}{k}\left(\frac{t-a}{b-a}\right)^k\left(1-\frac{t-a}{b-a}\right)^{n-k} \\ &= \binom{n}{k}\left(\frac{t-a}{b-a}\right)^k\left(\frac{b-t}{b-a}\right)^{n-k} \\ &= \binom{n}{k}\left(\frac{1}{b-a}\right)^k\left(\frac{1}{b-a}\right)^{n-k}(t-a)^k(b-t)^{n-k} \\ &= \left(\frac{1}{b-a}\right)^n\binom{n}{k}(t-a)^k(b-t)^{n-k}. \end{aligned}$$

Hence, one can consider the Bernstein basis polynomials on $x \in [0, 1]$ in this thesis without loss of generality. □

Example 2.3. *Bernstein basis polynomials.* Consider the Bernstein basis polynomials on $x \in [0, 1]$.

- $n = 0$. The Bernstein basis polynomial of degree 0 is given by

$$b_0^0(x) = 1.$$

- $n = 1$. The two Bernstein basis polynomials of degree 1 are given by

$$b_0^1(x) = 1 - x,\ b_1^1(x) = x.$$

- $n = 2$. The three Bernstein basis polynomials of degree 2 are given by

$$b_0^2(x) = (1-x)^2,\ b_1^2(x) = 2(1-x)x,\ b_2^2(x) = x^2.$$

- $n = 3$. The four Bernstein basis polynomials of degree 3 (see Figure 2.2) are given by

$$\begin{gathered} b_0^3(x) = (1-x)^3,\ b_1^3(x) = 3(1-x)^2x, \\ b_2^3(x) = 3(1-x)x^2,\ b_3^3(x) = x^3. \end{gathered}$$

- $n = 4$. The five Bernstein basis polynomials of degree 4 (see Figure 2.2) are given by

$$\begin{gathered} b_0^4(x) = (1-x)^4,\ b_1^4(x) = 4(1-x)^3x, \\ b_2^4(x) = 6(1-x)^2x^2,\ b_3^4(x) = 4(1-x)x^3,\ b_4^4(x) = x^4. \end{gathered}$$

□

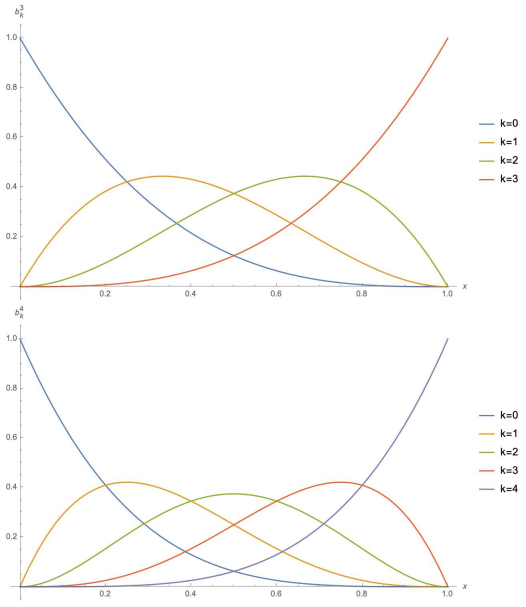

Figure 2.2: The four Bernstein basis polynomials of degree 3 (up) and the five Bernstein basis polynomials of degree 4 (down).

Remark 2.4. *A recursive definition of the Bernstein basis polynomials.* A Bernstein basis polynomial of degree n can be defined by blending together two Bernstein basis polynomials of degree $n-1$, i.e.,

$$b_k^n(x) := (1-x)b_k^{n-1}(x) + xb_{k-1}^{n-1}(x). \tag{2.2}$$

Using (2.1) and (2.2) yields

$$\begin{aligned}
(1-x)b_k^{n-1}(x) + xb_{k-1}^{n-1}(x) &= (1-x)\binom{n-1}{k}(1-x)^{(n-1)-k}x^k \\
&\quad + x\binom{n-1}{k-1}(1-x)^{(n-1)-(k-1)}x^{k-1} \\
&= \binom{n-1}{k}(1-x)^{n-k}x^k + \\
&\quad + \binom{n-1}{k-1}(1-x)^{n-k}x^k \\
&= \binom{n}{k}(1-x)^{n-k}x^k =: b_k^n(x).
\end{aligned}$$

□

Lemma 2.5. Properties of the Bernstein basis polynomials. *The Bernstein basis polynomials have the following properties:*

i) *The $n+1$ Bernstein basis polynomials of degree n are non-negative on $x \in [0,1]$.*

ii) *The Bernstein basis polynomials $b_k^n(x)$ and $b_{n-k}^n(x)$ of degree n are symmetrical about $x = \frac{1}{2}$, i.e.,*

$$b_k^n(x) = b_{n-k}^n(1-x).$$

iii) *The $n+1$ Bernstein basis polynomials of degree n form a partition of unity. The sum of them of degree n at x is 1, i.e.,*

$$\sum_{k=0}^{n} b_k^n(x) = 1.$$

iv) *The Bernstein basis polynomials of degree n satisfy*

$$b_k^n(0) = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{otherwise} \end{cases} \quad \textit{and} \quad b_k^n(1) = \begin{cases} 1 & \text{if } k = n \\ 0 & \text{otherwise.} \end{cases}$$

v) *The Bernstein basis polynomials of degree n can be written as a linear combination of the polynomials of degree $n+1$, i.e.,*

$$b_k^n(x) = \frac{n+1-k}{n+1} b_k^{n+1}(x) + \frac{k+1}{n+1} b_{k+1}^{n+1}(x).$$

vi) *Derivatives of the Bernstein basis polynomials of degree n can be written as a linear combination of the polynomials of degree $n-1$, i.e.,*

$$\frac{d}{dx} b_k^n(x) = n\left(b_{k-1}^{n-1}(x) - b_k^{n-1}(x)\right).$$

vii) *The maximum of the Bernstein basis polynomials of degree n occurs when $x = \frac{k}{n}$.*

viii) *Definite integrals of the Bernstein basis polynomials of degree n are $\frac{1}{n+1}$, i.e.,[3]*

$$\int_0^1 b_k^n(x)\, dx = \frac{1}{n+1}.$$

ix) *The $n+1$ Bernstein basis polynomials of degree n are linearly independent, i.e., if*

$$\lambda_0 b_0^n(x) + \lambda_1 b_1^n(x) + \cdots + \lambda_n b_k^n(x) = 0,$$

where the coefficients $\lambda_0, \lambda_1, \ldots, \lambda_n \in \mathbb{R}$, then $\lambda_0 = \lambda_1 = \cdots = \lambda_n = 0$.

Proof. i) It is trivial that $b_0^0(x) = 1, b_0^1(x) = 1 - x$ and $b_1^1(x) = x$ are non-negative for $0 \leq x \leq 1$. Using (2.2), one obtains that $b_k^n(x) := (1 - x)b_k^{n-1}(x) + xb_{k-1}^{n-1}(x)$. Then, it follows by induction that all the Bernstein basis polynomials of degree n are non-negative.

ii) Using the symmetric identity of binomial coefficients and (2.1), it follows that

$$\begin{aligned} b_{n-k}^n(1-x) &= \binom{n}{n-k}(1-(1-x))^{n-(n-k)}(1-x)^{n-k} \\ &= \binom{n}{k}x^k(1-x)^{n-k} = b_k^n(x). \end{aligned}$$

iii) It follows directly from the fact that the $n+1$ Bernstein basis polynomials are the $n+1$ terms in the binomial expansion of $((1-x)+x)^n$. Hence,

$$\sum_{k=0}^{n} b_k^n(x) = ((1-x)+x)^n = 1^n = 1.$$

iv) It is easily obtained by $b_k^n(0) = \binom{n}{k}1^{n-k}0^k$, $b_k^n(1) = \binom{n}{k}0^{n-k}1^k$ and $\binom{n}{0} = \binom{n}{n} = 1$ for all integers $n \geq 0$.

v) It holds that

$$\begin{aligned} b_k^n(x) &= ((1-x)+x)\binom{n}{k}(1-x)^{n-k}x^k \\ &= (1-x)\binom{n}{k}(1-x)^{n-k}x^k + x\binom{n}{k}(1-x)^{n-k}x^k \\ &= \binom{n}{k}(1-x)^{(n+1)-k}x^k + \binom{n}{k}(1-x)^{n-k}x^{k+1} \\ &= \frac{n+1-k}{n+1}\frac{n!(n+1)}{k!(n+1-k)(n-k)!}(1-x)^{(n+1)-k}x^k \\ &\quad + \frac{k+1}{n+1}\frac{n!(n+1)}{k!(k+1)((n+1)-(k+1))!}(1-x)^{(n+1)-(k+1)}x^{k+1} \\ &= \frac{n+1-k}{n+1}\binom{n+1}{k}(1-x)^{(n+1)-k}x^k \\ &\quad + \frac{k+1}{n+1}\binom{n+1}{k+1}(1-x)^{(n+1)-(k+1)}x^k \\ &= \frac{n+1-k}{n+1}b_k^{n+1}(x) + \frac{k+1}{n+1}b_{k+1}^{n+1}(x). \end{aligned}$$

vi) Using the chain rule and the product rule, one obtains that

$$\begin{aligned} \frac{d}{dx}b_k^n(x) &= \binom{n}{k}\left((n-k)(1-x)^{n-k-1}(-1)x^k + (1-x)^{n-k}kx^{k-1}\right) \\ &= k\binom{n}{k}(1-x)^{n-k}x^{k-1} - (n-k)\binom{n}{k}(1-x)^{(n-1)-k}x^k. \end{aligned}$$

Since

$$k\binom{n}{k} = \frac{n!}{(k-1)!(n-k)!} = n\frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} = n\binom{n-1}{k-1}$$

and

$$(n-k)\binom{n}{k} = \frac{n!}{k!(n-k-1)!} = n\frac{(n-1)!}{k!((n-1)-k)!} = n\binom{n-1}{k},$$

it follows that

$$\begin{aligned}\frac{d}{dx}b_k^n(x) &= n\binom{n-1}{k-1}(1-x)^{(n-1)-(k-1)}x^{k-1} \\ &\quad - n\binom{n-1}{k}(1-x)^{(n-1)-k}x^k \\ &= n\left(b_{k-1}^{n-1}(x) - b_k^{n-1}(x)\right).\end{aligned}$$

vii) By setting the derivative from vi) to zero, one gets

$$\begin{aligned}\frac{d}{dx}b_k^n(x) &= k\binom{n}{k}(1-x)^{n-k}x^{k-1} - (n-k)\binom{n}{k}(1-x)^{(n-1)-k}x^k \\ &= \binom{n}{k}\left(k(1-x) - (n-k)x\right)(1-x)^{n-k-1}x^{k-1} \\ &= \binom{n}{k}(k-nx)(1-x)^{n-k-1}x^{k-1} \\ &= 0.\end{aligned}$$

Hence, one obtains with i) and iv) that the maximum of the Bernstein basis polynomials occurs when $x = \frac{k}{n}$.

viii) From vi) one knows that $\frac{d}{dx}b_{k+1}^{n+1}(x) = (n+1)(b_k^n(x) - b_{k+1}^n(x))$. If $k \neq n$,

$$\begin{aligned}\int_0^1 \left(\frac{d}{dx}b_{k+1}^{n+1}\right)dx &= b_{k+1}^{n+1}(x)\big|_0^1 \\ &= \binom{n+1}{k+1}\left(0^{n-k}1^{k+1} - 1^{n-k}0^{k+1}\right) \\ &= 0 \\ &= (n+1)\int_0^1 \left(b_k^n(x) - b_{k+1}^n(x)\right)dx \\ &= (n+1)\left(\int_0^1 b_k^n(x)\,dx - \int_0^1 b_{k+1}^n(x)\,dx\right).\end{aligned}$$

Hence,

$$\int_0^1 b_0^n(x)\,dx = \int_0^1 b_1^n(x)\,dx = \cdots = \int_0^1 b_{(n-1)+1}^n(x)\,dx = \int_0^1 b_n^n(x)\,dx.$$

Combining with iii) and the linearity of the integral yields

$$\int_0^1 b_k^n(x)\,dx = \frac{1}{n+1}.$$

ix) Using (2.1) and applying the binomial theorem to expand the term $(1-x)^{n-k}$, one can write any Bernstein basis polynomial of degree n in terms of the monomial basis $\{1, x, x^2, \ldots, x^n\}$, which spans the space of polynomials of degree $\leq n$, such that,

$$\begin{aligned}
b_k^n(x) &= \binom{n}{k}(1-x)^{n-k}x^k \\
&= \binom{n}{k}x^k \sum_{i=0}^{n-k}\binom{n-k}{i}1^{n-k-i}(-x)^i \\
&= \sum_{i=0}^{n-k}\binom{n}{k}\binom{n-k}{i}(-1)^i x^{i+k} \\
&= \sum_{i=k}^{n}(-1)^{i-k}\binom{n}{k}\binom{n-k}{i-k}x^i \\
&= \sum_{i=k}^{n}(-1)^{i-k}\binom{n}{i}\binom{i}{k}x^i.
\end{aligned}$$

It follows that

$$\begin{aligned}
0 &= \lambda_0 b_0^n(x) + \lambda_1 b_1^n(x) + \cdots + \lambda_n b_n^n(x) \\
&= \lambda_0 \sum_{i=0}^{n}(-1)^i\binom{n}{i}\binom{i}{0}x^i + \lambda_1 \sum_{i=1}^{n}(-1)^{i-1}\binom{n}{i}\binom{i}{1}x^i + \ldots \\
&\quad + \lambda_n \sum_{i=n}^{n}(-1)^{i-n}\binom{n}{i}\binom{i}{n}x^i \\
&= \lambda_0 + \left(\sum_{i=0}^{1}(-1)^{1-i}\lambda_i\binom{n}{1}\binom{1}{i}\right)x^1 + \ldots \\
&\quad + \left(\sum_{i=0}^{n}(-1)^{n-i}\lambda_i\binom{n}{n}\binom{n}{i}\right)x^n.
\end{aligned}$$

One gets with the linear independence of the monomial basis

$$\lambda_0 = 0,\ \sum_{i=0}^{1}(-1)^{1-i}\lambda_i\binom{n}{1}\binom{1}{i} = 0,\ \ldots,\ \sum_{i=0}^{n}(-1)^{n-i}\lambda_i\binom{n}{n}\binom{n}{i} = 0.$$

By inserting $\lambda_0 = 0$ into the second equation and so on, one can easily obtain that $\lambda_0 = \lambda_1 = \cdots = \lambda_n = 0$.

□

2.3 Weierstrass Approximation Theorem

In 1911 Bernstein gave a constructive proof of the Weierstrass[4] approximation theorem in his paper *Démonstration du théoréme de Weierstrass fondée sur le calcul des probabilités*[1], which states that every continuous function defined on a compact interval $[a, b]$ can be uniformly approximated as closely as desired by a polynomial function. He sent his simple proof to the Belgium Academy of Science and was awarded a prize.

Definition 2.6. Bernstein polynomials. Consider a continuous function f on $[0, 1]$ without loss of generality (see Remark 2.2), the Bernstein polynomial of degree n associated with f is defined by

$$B_n(f)(x) := \sum_{k=0}^{n} f\left(\frac{k}{n}\right) b_k^n(x). \tag{2.3}$$

□

Lemma 2.7. Properties of the Bernstein polynomials. *The Bernstein polynomials have the following properties, also see [9]:*

i) If $f(x) = x$ and $x \in [0, 1]$, it holds for all $n \geq 1$ that $B_n(f)(x) = x$.

ii) If $f(x) = x(1 - x)$ and $x \in [0, 1]$, then it holds

$$B_n(f)(x) = \left(1 - \frac{1}{n}\right) x(1 - x).$$

Hence, for all $n \geq 2$,

$$0 \leq \sum_{k=0}^{n} \left(x - \frac{k}{n}\right)^2 b_k^n(x) = \frac{x(1-x)}{n} \leq \frac{1}{4n}.$$

Proof. i) Using Lemma 2.5 iii), one obtains that for all $n \geq 1$

$$\begin{aligned} B_n(f)(x) &= \sum_{k=0}^{n} \frac{k}{n} \binom{n}{k} (1-x)^{n-k} x^k \\ &= \sum_{k=1}^{n} \binom{n-1}{k-1} (1-x)^{(n-1)-(k-1)} x^{(k-1)} x \\ &= x \sum_{k=0}^{n-1} \binom{n-1}{k} (1-x)^{(n-1)-k} x^k \\ &= x \sum_{k=0}^{n-1} b_k^{n-1}(x) = x. \end{aligned}$$

[4]Karl Weierstrass (1815 - 1897)

ii) Similarly, using Lemma 2.5 iii) again, one obtains that for all $n \geq 2$

$$\begin{aligned} B_n(f)(x) &= \sum_{k=0}^{n} \frac{k}{n}\left(1-\frac{k}{n}\right)\binom{n}{k}(1-x)^{n-k}x^k \\ &= \frac{n-1}{n}\sum_{k=1}^{n-1}\binom{n-2}{k-1}(1-x)^{n-k}x^k \\ &= \frac{n-1}{n}\sum_{k=0}^{n-2}\binom{n-2}{k}(1-x)^{n-(k+1)}x^{k+1} \\ &= \left(1-\frac{1}{n}\right)x(1-x)\sum_{k=0}^{n-2}\binom{n-2}{k}(1-x)^{(n-2)-k}x^k \\ &= \left(1-\frac{1}{n}\right)x(1-x). \end{aligned}$$

Hence, if follows with Lemma 2.7 i) that for all $n \geq 2$ and $x \in [0,1]$

$$\begin{aligned} \sum_{k=0}^{n}\left(x-\frac{k}{n}\right)^2 b_k^n(x) &= \sum_{k=0}^{n}\left(x^2 - 2x\frac{k}{n} + \frac{k^2}{n^2}\right)b_k^n(x) \\ &= \sum_{k=0}^{n}\left(x^2 + (1-2x)\frac{k}{n} - \frac{k(n-k)}{n^2}\right)b_k^n(x) \\ &= x^2\sum_{k=0}^{n} b_k^n(x) + (1-2x)\sum_{k=0}^{n}\frac{k}{n}b_k^n(x) \\ &\quad - \sum_{k=0}^{n}\frac{k(n-k)}{n^2}b_k^n(x) \\ &= x^2 + (1-2x)x - \left(1-\frac{1}{n}\right)x(1-x) \\ &= \frac{x(1-x)}{n} \leq \frac{1}{4n}. \end{aligned}$$

□

Lemma 2.8. Uniform continuity on a compact set. *A continuous function f on a compact set K is uniformly continuous, i.e., if any $\epsilon > 0$ is given, there exists a $\delta > 0$ such that for all $x, y \in K$ with $|x-y| < \delta$*

$$|f(x) - f(y)| < \epsilon.$$

Proof. This lemma was proved in the basic module *Analysis II*, see [10]. □

Theorem 2.9. Weierstrass Approximation Theorem. *If f is a continuous function on the compact interval $[0,1]$ and if any $\epsilon > 0$ is given, then there exists a polynomial $p_n(x)$ of sufficiently high degree n such that for all $x \in [0,1]$*

$$|f(x) - p_n(x)| < \epsilon.$$

Proof. Consider (2.3) as the polynomial to approximate a given continuous function $f(x)$, i.e.,

$$p_n(x) = B_n(f)(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) b_k^n(x). \tag{2.4}$$

Since $[0,1]$ is a closed and bounded interval, one gets by Lemma 2.8 that $f(x)$ is uniformly continuous, more precisely, for a given $\epsilon > 0$, there exists a $\delta > 0$ such that for all $x, y \in [0,1]$ with $|x-y| < \delta$

$$|f(x) - f(y)| < \epsilon.$$

For the space $C([0,1])$, one denotes that the function $\|\cdot\|_{max} : C([0,1]) \to \mathbb{R}$, which is defined through

$$\|f\|_{max} := \max_{x\in[0,1]} |f(x)|,$$

maximum norm. The maximum norm is well defined on the basis of the extreme value theorem, which ensures the existence of the maximum, see [17].

Using the triangle inequality and the uniform continuity of $f(x)$, with Lemma 2.5 i) and Lemma 2.7, one obtains for a fix x that

$$\begin{aligned}
|f(x) - B_n(f)(x)| &= \left| \sum_{k=0}^{n} \left(f(x) - f\left(\frac{k}{n}\right)\right) b_k^n(x) \right| \\
&\leq \sum_{k=0}^{n} \left| f(x) - f\left(\frac{k}{n}\right)\right| b_k^n(x) \\
&= \sum_{|x-\frac{k}{n}|<\delta} \left| f(x) - f\left(\frac{k}{n}\right)\right| b_k^n(x) \\
&\quad + \sum_{|x-\frac{k}{n}|\geq\delta} \left| f(x) - f\left(\frac{k}{n}\right)\right| b_k^n(x) \\
&\leq \frac{\epsilon}{2} \sum_{|x-\frac{k}{n}|<\delta} b_k^n(x) + 2\|f\|_{max} \sum_{|x-\frac{k}{n}|\geq\delta} b_k^n(x) \\
&\leq \frac{\epsilon}{2} \sum_{|x-\frac{k}{n}|<\delta} b_k^n(x) + 2\frac{\|f\|_{max}}{\delta^2} \sum_{|x-\frac{k}{n}|\geq\delta} \left(x - \frac{k}{n}\right)^2 b_k^n(x) \\
&\leq \frac{\epsilon}{2} \sum_{k=0}^{n} b_k^n(x) + 2\frac{\|f\|_{max}}{\delta^2} \sum_{k=0}^{n} \left(x - \frac{k}{n}\right)^2 b_k^n(x) \\
&\leq \frac{\epsilon}{2} + \frac{\|f\|_{max}}{2n\delta^2} \\
&< \epsilon \text{ for all } n > \frac{\|f\|_{max}}{\epsilon\delta^2}.
\end{aligned}$$

□

Corollary 2.10. Bounds of a one-dimensional polynomial. *Let $p_n(x)$ be a one-dimensional polynomial of degree n on $[0,1]$, let f be a continuous function on $[0,1]$, then $p_n(x)$ is uniformly bounded by the largest and smallest Bernstein coefficients.*

Proof. Using (2.4) and Lemma 2.5 iii), also see [7], one obtains

$$\begin{aligned} p_n(x) &= B_n(f)(x) \\ &= \sum_{k=0}^{n} f\left(\frac{k}{n}\right) b_k^n(x) \\ &\leq \max_{0\leq k\leq n} f\left(\frac{k}{n}\right) \sum_{k=0}^{n} b_k^n(x) \\ &\leq \max_{0\leq k\leq n} f\left(\frac{k}{n}\right), \end{aligned}$$

where $f\left(\frac{k}{n}\right)$ are Bernstein coefficients. Similarly, one gets

$$p_n(x) = B_n(f)(x) \geq \min_{0\leq k\leq n} f\left(\frac{k}{n}\right).$$

□

Chapter 3

Linear Two-point Boundary Value Problems

Remark 3.1. *Motivation.* In this chapter we will discuss linear two-point boundary value problems and introduce the model problem. After that, we will investigate the solution of the model problem by analyzing linear homogeneous and inhomogeneous differential equations separately. Then the boundary conditions will be taken into account.

For more detailed derivations, it is referred to [4]. □

3.1 The Model Problem

Definition 3.2. Linear two-point boundary value problems. A linear two-point second order boundary value problem has the form

$$-\varepsilon u'' + b(x)u' + c(x)u = f(x) \quad \text{for } x \in (d, e), \tag{3.1}$$

with the boundary conditions

$$\begin{aligned} \alpha_d u(d) - \beta_d u'(d) &= \gamma_d, \\ \alpha_e u(e) - \beta_e u'(e) &= \gamma_e, \end{aligned} \tag{3.2}$$

where $\alpha_d, \alpha_e, \beta_d, \beta_e, \gamma_d, \gamma_e$ are given constants, and $b, c, f \in C([e, d])$, $0 < \varepsilon \in \mathbb{R}$. □

Definition 3.3. Boundary conditions. Let $\gamma_d, \gamma_e \in \mathbb{R}$, $\alpha_d, \alpha_e \in \mathbb{R} \setminus \{0\}$.

1. The most common boundary condition is to specify the value of the function on the boundary; this type of constraint is called Dirichlet[1] boundary conditions, i.e.,
$$u(d) = \gamma_d, \quad u(e) = \gamma_e.$$

[1] Johann Peter Gustav Lejeune Dirichlet (1805 - 1859)

2. A second type of boundary condition is to specify the derivative of the unknown function on the boundary; this type of constraint is called Neumann[2] boundary conditions, i.e.,

$$u'(d) = \gamma_d, \quad u'(e) = \gamma_e.$$

3. A third type of boundary condition is to specify a weighted combination of the function value and its derivative at the boundary; this is called Robin[3] or mixed boundary conditions, i.e.,

$$\alpha_d u(d) + u'(d) = \gamma_d, \quad \alpha_e u(e) + u'(e) = \gamma_e.$$

□

Remark 3.4. *Normalization of a linear two-point boundary value problem.*

- One can consider a linear two-point boundary value problem on $x \in [0, 1]$ without loss of generality. Similarly, see Remark 2.2, one replaces x by $\frac{x-d}{e-d}$.
- By subtracting from $u(x)$ a smooth function $\psi(x)$, which also satisfies the original boundary conditions, one can consider a linear two-point second order boundary value problem with homogeneous boundary conditions $\gamma_d = \gamma_e = 0$ without loss of generality.
  Let the Dirichlet boundary conditions be

  $$u(d) = \gamma_d, \quad u(e) = \gamma_e,$$

  and one sets

  $$\psi(x) = \gamma_d \frac{x-e}{d-e} + \gamma_e \frac{x-d}{e-d},$$

  and

  $$v(x) = u(x) - \psi(x).$$

  It can be easily verified that $v(d) = v(e) = 0$. Then $v(x)$ is the solution of a linear two-point second order boundary value problem with homogeneous boundary conditions.
- For linear two-point boundary value problems, a well-known approach to find the solution depends on whether it is possible to evaluate integrals analytically. Generally, it is not possible to find the analytical solution of a linear second order elliptic boundary value problem in higher dimensions. In this thesis we will restrict our investigations to one dimension for reducing the complexity. □

[2] Carl Gottfried Neumann (1832 - 1925)
[3] Victor Gustave Robin (1855 - 1897)

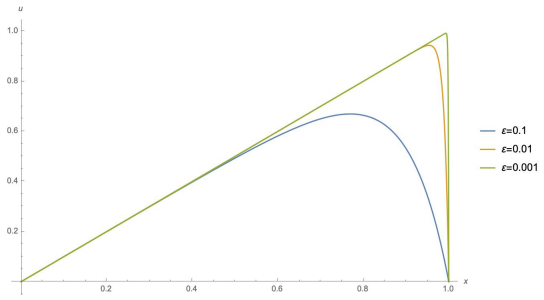

Figure 3.1: The solutions for $\varepsilon = 0.1, 0.01$ and 0.001.

Definition 3.5. The model problem. The model problem has the form

$$\mathcal{L}u := -\varepsilon u'' + b(x)u' + c(x)u = f(x) \quad \textit{for } x \in (0,1), \tag{3.3}$$

with the boundary conditions

$$u(0) = u(1) = 0, \tag{3.4}$$

where $b, c, f \in C([0,1])$, $0 < \varepsilon \in \mathbb{R}$. □

Definition 3.6. Differential operator. A operator is a map between two function spaces. A linear operator is a linear map A on a linear space X, so that

$$A(\alpha u + \beta v) = \alpha Au + \beta Av$$

for all scalars $\alpha, \beta \in \mathbb{R}$ and all $u, v \in X$. A differential operator is an operator defined as a function of the differentiation operator. For example, in (3.3), $\mathcal{L}$ is a linear differential operator. □

Example 3.7. Consider the boundary value problem

$$-\varepsilon u'' + u' = 1 \quad \textit{for } x \in (0,1),$$

with

$$u(0) = u(1) = 0.$$

Then the solution is

$$u(x) = x - \frac{\exp\left(-\frac{1-x}{\epsilon}\right) - \exp\left(-\frac{1}{\epsilon}\right)}{1 - \exp\left(-\frac{1}{\epsilon}\right)}.$$

The smaller the parameter ε is, the steeper the solution will be near the right-hand edge, see Figure 3.1. This steep part of the solution is called boundary layer. The dramatic change in a very small area will lead to difficulties for the numerical approximation of the solution. □

3.2 Solutions to the Model Problem

Remark 3.8. *Solvability.* In order to investigate if the model problem (3.3), (3.4) is solvable, one can consider the problem (3.5), (3.6) under the condition of $0 < \varepsilon \in \mathbb{R}$. By dividing (3.3) by ε, one obtains the problem

$$\mathcal{L}u := -u'' + b(x)u' + c(x)u = f(x) \quad \textit{for } x \in (0,1), \tag{3.5}$$

with the boundary conditions

$$u(0) = u(1) = 0, \tag{3.6}$$

where $b, c, f \in C([0,1])$. First we will discuss the existence and uniqueness of the solution of the model problem. □

Definition 3.9. Well-posedness. A boundary value problem is called well-posed, if

i) a solution exists,

ii) the solution is unique, and

iii) the solution changes continuously with changes in the data.

The problems that are not well-posed are termed ill-posed. □

Definition 3.10. The trivial solution. If a boundary value problem has only constant zero solution, then it is called a trivial solution, i.e., $u(x) \equiv 0$ is a trivial solution. □

Definition 3.11. The general solution. The general solution includes all possible solutions and typically includes arbitrary constants. □

Definition 3.12. The classical solution. A function $u(x)$ is called a classical solution of (3.5), (3.6), if

i) $u \in C^2(0,1) \cap C([0,1])$,

ii) $u(x)$ satisfies the equation (3.5) and

iii) $u(x)$ satisfies the boundary conditions (3.6). □

3.2.1 Linear homogeneous differential equations

Remark 3.13. *Homogeneous equation.* To study the solution of the problem (3.5), (3.6), we will analyze the homogeneous problem (right-hand side of (3.5) vanishes) and the inhomogeneous problem separately. We will start with linear homogeneous second order differential equations, since homogeneous equations are much easier to solve compared to their inhomogeneous counterparts. Consider first only the equation (3.5) with homogeneous right-hand side,

$$\mathcal{L}u := -u'' + b(x)u' + c(x)u = 0 \quad \textit{for } x \in (0,1), \tag{3.7}$$

where $b, c \in C([0,1])$. □

Theorem 3.14. Principle of superposition. *Consider the linear homogeneous second order differential equation (3.7). Let $u_1(x), u_2(x) \in C^2([0,1])$ be two linear independent solutions to the linear homogeneous differential equation (3.7), then the general solution can be expressed as a linear combination of the solutions $u_1(x), u_2(x)$.*

Proof. Suppose that $u_1(x), u_2(x)$ are both solutions to (3.7). Then it holds that

$$-u_1'' + b(x)u_1' + c(x)u_1 = 0 \quad \textit{and} \quad -u_2'' + b(x)u_2' + c(x)u_2 = 0 \quad \textit{for } x \in (0,1).$$

Let $\alpha, \beta \in \mathbb{R}$. Since $\mathcal{L}$ is a linear differential operator, one obtains

$$\begin{aligned}
\mathcal{L}u &= 0 \\
&= \alpha(-u_1'' + b(x)u_1' + c(x)u_1) + \beta(-u_2'' + b(x)u_2' + c(x)u_2) \\
&= -\alpha u_1'' - \beta u_2'' + b(x)\alpha u_1\prime + b(x)\beta u_2' + c(x)\alpha u_1 + c(x)\beta u_2 \\
&= -(\alpha u_1 + \beta u_2)'' + b(x)(\alpha u_1 + \beta u_2)' + c(x)(\alpha u_1 + \beta u_2).
\end{aligned}$$

Thus, the general solution can be expressed as $u(x) = \alpha u_1(x) + \beta u_2(x)$, $\alpha, \beta \in \mathbb{R}$, which satisfies (3.7). □

Example 3.15. *Non-existence or non-uniqueness of the solution of a linear homogeneous differential equation.* So far we have only considered linear homogeneous differential equations without the boundary conditions. However, a classical solution also needs to satisfy the boundary conditions. Consider the linear homogeneous differential equation

$$-u''(x) - u(x) = 0 \quad \textit{for } x \in (0, \pi),$$

with the boundary conditions

$$u(0) = u(\pi) = 1.$$

Following Theorem 3.14, the general solution of this linear homogeneous differential equation is

$$u(x) = \alpha \cos x + \beta \sin x, \ \ \alpha, \beta \in \mathbb{R}.$$

Then it has no solution, because α cannot be equal to 1 and -1 simultaneously. In another case, if the boundary conditions are

$$u(0) = 1, \quad u(\pi) = -1.$$

It follows that $\alpha = 1$, and β can be any real number, i.e., the equation has infinitely many solutions, see Figure 3.2. So this problem is ill-posed. □

Theorem 3.16. Existence and uniqueness of the solution of the model problem with homogeneous right-hand side. *Consider the linear homogeneous second order differential equation (3.7) with the homogeneous Dirichlet boundary conditions (3.6), where $b \in C^1([0,1])$, $c \in C([0,1])$. If it holds for all $x \in (0,1)$ that*

$$\tilde{c}(x) := \frac{1}{4}b^2(x) - \frac{1}{2}b'(x) + c(x) \geq 0,$$

then it has only the trivial solution.

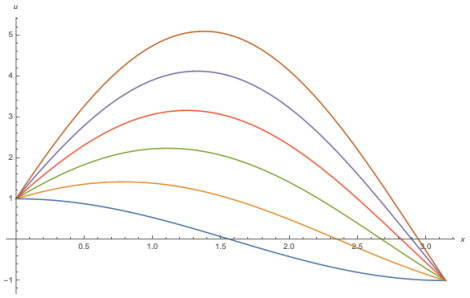

Figure 3.2: Different solutions to Example 3.15, which satisfy the boundary conditions $u(0) = 1$, $u(\pi) = -1$.

Proof. Obviously, $u(x) \equiv 0$ is a solution of (3.7), (3.6).
Assume that $v(x) \neq 0$ is another solution. Define

$$\tilde{u}(x) := v(x) \exp\left(-\frac{1}{2}\int_0^x b(\xi)\, d\xi\right), \ x \in [0,1],$$

so that (3.7), (3.6) can be transformed into a symmetric problem

$$-\tilde{u}''(x) + \tilde{c}(x)\tilde{u}(x) = 0 \quad \textit{for } x \in (0,1),$$

with the boundary conditions

$$\tilde{u}(0) = \tilde{u}(1) = 0,$$

whose solution is $\tilde{u}(x) \equiv 0$. Assume that $\tilde{u}(x) \neq 0$ is another solution of this symmetric problem. By multiplying the equation by $\tilde{u}(x)$ and then integrating it by parts, one gets

$$\begin{aligned} 0 &= \int_0^1 \left(-\tilde{u}''(x)\tilde{u}(x) + \tilde{c}(x)\tilde{u}^2(x)\right)\, dx \\ &= -\tilde{u}''(x)\tilde{u}(x)\Big|_0^1 + \int_0^1 \left(\tilde{u}'(x)\right)^2\, dx + \int_0^1 \tilde{c}(x)\tilde{u}^2(x)\, dx \\ &= \int_0^1 \left(\left(\tilde{u}'(x)\right)^2 + \tilde{c}(x)\tilde{u}^2(x)\right)\, dx. \end{aligned}$$

Since all terms in the integral are non-negative, it follows that $\left(\tilde{u}'(x)\right)^2 = 0$, i.e., $\tilde{u}'(x)^2 = 0$. According to the boundary conditions and the continuity of $\tilde{u}(x)$, one gets $\tilde{u}(x) \equiv 0$. Consequently, it follows also that

$$v(x) = \tilde{u}(x) \exp\left(\frac{1}{2}\int_0^x b(\xi)\, d\xi\right) \equiv 0,$$

which is a contradiction to the assumption. Hence, $u(x) \equiv 0$. □

Corollary 3.17. Another criterion for the uniqueness of the model problem with homogeneous right-hand side. *Consider the linear homogeneous second order differential equation (3.7) with homogeneous Dirichlet boundary conditions (3.6), where $b \in C^1([0,1])$, $c \in C([0,1])$. Suppose that $u_1(x), u_2(x)$ are two linear independent solutions to (3.7), (3.6), denote*

$$R := det \begin{pmatrix} u_1(0) & u_2(0) \\ u_1(1) & u_2(1) \end{pmatrix},$$

then the solution is unique if and only if $R \neq 0$.

Proof. Following Theorem 3.14, the general solution of linear homogeneous second order differential equations is

$$u(x) = \alpha u_1(x) + \beta u_2(x), \quad \alpha, \beta \in \mathbb{R}.$$

Since the solution has to satisfy the boundary conditions (3.6), one has

$$u(0) = \alpha u_1(0) + \beta u_2(0) = 0, \quad u(1) = \alpha u_1(1) + \beta u_2(1) = 0,$$

which can be transformed into a system of linear equations

$$\begin{pmatrix} u_1(0) & u_2(0) \\ u_1(1) & u_2(1) \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

From *Linear Algebra*, one knows that the system has a unique solution (the trivial solution) if and only if the determinant of its coefficient matrix is non-zero, i.e., $R \neq 0$. □

3.2.2 Linear inhomogeneous differential equations

Remark 3.18. *Inhomogeneous equation.* Consider the linear equation (3.5) with inhomogeneous right-hand side. Now we will examine the conditions under which a solution of this linear inhomogeneous differential equation (3.5) with the boundary conditions (3.6) exists and is unique. □

Definition 3.19. The particular solution. A particular solution $u_p(x)$ of a differential equation is a solution that contains no arbitrary constants. □

Theorem 3.20. The general solution of a linear inhomogeneous differential equations. *Consider the linear inhomogeneous second order differential equation (3.5) with $b, c, f \in C([0,1])$. Let $u_p(x)$ be a particular solution of the linear inhomogeneous equation (3.5), and let $u_1(x), u_2(x)$ be two linear independent solutions to the corresponding linear homogeneous equation (3.7), then the general solution can be expressed as*

$$u(x) = \alpha u_1(x) + \beta u_2(x) + u_p(x), \ \alpha, \beta \in \mathbb{R}.$$

Proof. To prove that $u(x)$ is the general solution, one must first show that it solves the linear inhomogeneous equation (3.5). Substituting the expression of $u(x)$ into (3.5), one obtains

$$\begin{aligned}-u'' + b(x)u' + c(x)u &= -(\alpha u_1 + \beta u_2 + u_p)'' + b(x)(\alpha u_1 + \beta u_2 + u_p)' \\ &\quad + c(x)(\alpha u_1 + \beta u_2 + u_p) \\ &= \alpha(-u_1'' + b(x)u_1' + c(x)u_1) + \beta(-u_2'' + b(x)u_2' + c(x)u_2) \\ &\quad + (-u_p'' + b(x)u_p' + c(x)u_p) \\ &= f(x).\end{aligned}$$

Secondly, let $u(x)$ be an arbitrary solution and $u_p(x)$ be a particular solution. Now one has to show that any solution of (3.5) can be written in that form. It holds that

$$\begin{aligned}-(u-u_p)'' + b(x)(u-u_p)' + c(x)(u-u_p) &= (-u'' + b(x)u' + c(x)u) \\ &\quad - (-u_p'' + b(x)u_p' + c(x)u_p) \\ &= f(x) - f(x) \\ &= 0.\end{aligned}$$

Thus, $u(x) - u_p(x)$ is a solution of the corresponding linear homogeneous equation (3.7). It follows from Theorem 3.14 that

$$u(x) - u_p(x) = \alpha u_1(x) + \beta u_2(x),\ \alpha, \beta \in \mathbb{R}.$$

Hence, the general solution can be expressed as

$$u(x) = \alpha u_1(x) + \beta u_2(x) + u_p(x),\ \alpha, \beta \in \mathbb{R}.$$

□

Example 3.21. *Non-existence or non-uniqueness of the solution of a linear inhomogeneous differential equation.* At this point, there are still the boundary conditions that we have to take into account. Consider the linear inhomogeneous differential equation

$$-u''(x) - u(x) = -x \quad \textit{for } x \in (0, \pi),$$

with the boundary conditions

$$u(0) = u(\pi) = 0.$$

Following Example 3.15 and Theorem 3.20, a particular solution of this linear homogeneous differential equation is $u_p(x) = x$, and the general solution is

$$u(x) = \alpha \cos x + \beta \sin x + x,\ \alpha, \beta \in \mathbb{R}.$$

Then it has no solution, although one has $\alpha = 0$, $u(\pi) = -\pi \neq 0$.
In another case, if the boundary conditions are

$$u(0) = 0, \quad u(\pi) = \pi.$$

It follows that $\alpha = 0$, and β can be any real number, i.e., the equation has infinitely many solutions, see Figure 3.3. So this problem is ill-posed. □

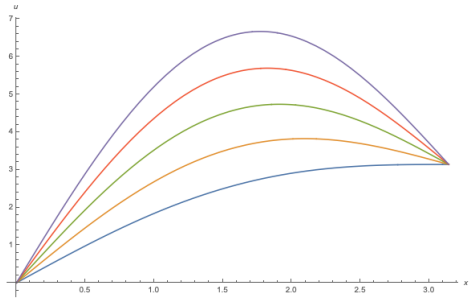

Figure 3.3: Different solutions to Example 3.21, which satisfy the boundary conditions $u(0) = 0$, $u(\pi) = \pi$.

To examine the conditions under which a solution of this linear inhomogeneous differential equation (3.5) with the boundary conditions (3.6) exists and is unique, it is necessary to introduce the Wronskian[4] determinant and the Green[5]'s function.

Definition 3.22. The Wronskian determinant. Consider the linear inhomogeneous differential equation (3.5) with the boundary conditions (3.6). Let $u_1(x), u_2(x)$ be two linearly independent solutions to the corresponding linear homogeneous equation (3.7). The Wronskian determinant (or simply the Wronskian) is the determinant of the square matrix

$$W(x) := det \begin{pmatrix} u_1(x) & u_2(x) \\ u_1'(x) & u_2'(x) \end{pmatrix}.$$

Since $u_1(x), u_2(x)$ are linear independent, the Wronskian determinant is not equal to zero for $x \in [0, 1]$, i.e., $W(x) \neq 0$. □

Definition 3.23. Green's function. The function $G(x, \xi)$ is called a Green's function for the linear homogeneous second order differential equation (3.7) with homogeneous Dirichlet boundary conditions (3.6), where $b \in C^1([0, 1])$, $c \in C([0, 1])$, if

i) $G(x, \xi)$ is continuous on the square $Q := \{(x, \xi)|\ x, \xi \in [0, 1]\}$,

ii) there are continuous partial derivatives $G_x(x, \xi)$ and $G_{xx}(x, \xi)$ in both of the domains

$$Q_1 := \{(x, \xi)|\ 0 < \xi < x < 1\}, \quad Q_1 := \{(x, \xi)|\ 0 < x < \xi < 1\},$$

iii) for a fixed $\xi \in (0, 1)$, $G(x, \xi)$ as a function of x is a solution of $\mathcal{L}G = 0$ for $x \neq \xi,\ x \in (0, 1)$,

[4] Józef Maria Hoëné-Wroński (1776 - 1853)
[5] George Green (1793 - 1841)

iv) the first partial derivative of $G(x,\xi)$ has a "jump" of the form for $x \in (0,1)$

$$G_x(x+0,x) - G_x(x-0,x) = -1,$$

v) and $G(0,\xi) = G(1,\xi) = 0$ for all $\xi \in (0,1)$. □

Theorem 3.24. Existence and uniqueness of the solution of the model problem with inhomogeneous right-hand side. *Consider the linear inhomogeneous second order differential equation (3.5) with the homogeneous Dirichlet boundary conditions (3.6), where $b, c, f \in C([0,1])$. If the corresponding linear homogeneous equation (3.7) has only the trivial solution, then (3.5), (3.6) has exactly one classical solution, which has the form*

$$u(x) = \int_0^1 G(0,\xi) f(\xi)\, d\xi$$

with the Green's function

$$G(x,\xi) = \frac{1}{RW(\xi)} \begin{cases} A(\xi)B(x), & \text{if } x \in \overline{Q_1}, \\ A(x)B(\xi), & \text{if } x \in \overline{Q_2}, \end{cases}$$

where

$$A(x) := det \begin{pmatrix} u_1(0) & u_2(0) \\ u_1(x) & u_2(x) \end{pmatrix}, \; B(x) := det \begin{pmatrix} u_1(x) & u_2(x) \\ u_1(1) & u_2(1) \end{pmatrix}.$$

Proof. Only a sketch proof of this theorem will be given here, since the idea of this proof is similar to those of Theorem 3.14, Theorem 3.16, Theorem 3.20. The According to Definition 3.23, one can first show that $G(x,\xi)$ is a Green's function. To show the existence of the solution, one substitutes $u(x)$ into (3.5), then it shows that $u(x)$ solves (3.5). One proves the uniqueness of the solution by contradiction. Assume that there is another solution $v(x)$, then one finds out that the difference of two solutions $u(x) - v(x)$ solves (3.7) as well. Since (3.7) has only the trivial solution, both solutions are actually the same, which is a contradiction to the assumption. Hence, the classical solution is unique. □

Corollary 3.25. The converse of Theorem 3.24. *Under the assumptions of Theorem 3.24, if (3.5), (3.6) has exactly one classical solution, then the corresponding linear homogeneous equation (3.7) has only the trivial solution.*

Proof. One proves the corollary by contradiction. Suppose that $u_{inh}(x)$ is the unique classical solution of (3.5), (3.6) and assume that $u_{hom}(x)$ is a nontrivial solution of the corresponding linear homogeneous equation (3.7). It follows from Theorem 3.20 that $u_{inh}(x) + u_{hom}(x)$ is another solution of (3.5), which is a contradiction to the assumption. Hence, (3.7) has only the trivial solution. □

Corollary 3.26. Existence and uniqueness of the solution of the model problem with arbitrary Dirichlet boundary conditions. *Consider the linear inhomogeneous second order differential equation (3.5) with arbitrary Dirichlet boundary conditions $u(d) = \gamma_d, \quad u(e) = \gamma_e, \; \gamma_d, \gamma_e \in \mathbb{R}$, where $b \in C^1([0,1])$,*

$c \in C([0,1])$. *Similarly, see Theorem 3.16, if it holds for all* $x \in (0,1)$ *that*

$$\tilde{c}(x) := \frac{1}{4}b^2(x) - \frac{1}{2}b'(x) + c(x) \geq 0,$$

then the model problem with arbitrary Dirichlet boundary conditions has exactly one classical solution.

Proof. It follows from Remark 3.4, Theorem 3.16, Theorem 3.24. □

Remark 3.27. *Well-posedness of the linear two-point second order boundary value problems.* Until now, we have shown that under the certain conditions, a linear two-point second order boundary value problem has exactly one classical solution, which also means that, the linear two-point second order boundary value problem is well-posed, see Definition 3.9. It tells us that the unique solution of the model problem varies continuously with the boundary data. □

Chapter 4

Weak Formulation of the Model Problem

Remark 4.1. *Motivation.* In fact, there does not always exist a classical solution of the model problem (3.5), (3.6), because many solutions are not sufficiently smooth, which does not satisfy the first condition of the classical solution, see Definition 3.12. The so-called weak solutions are much easier to describe in terms of linear algebra and its infinite-dimensional analogues than classical solutions are, see [13].

In this chapter we will show the existence and uniqueness of the weak solution. Several functional analysis tools might be used here. Lebesgue[1] spaces are introduced in the basic module *Analysis III*, see [5]. Some inequalities, e.g. Young[2]'s inequality, Cauchy[3]-Schwarz[4] inequality, Hölder[5]'s inequality and Poincaré[6]-Friedrichs[7] inequality, can be found in the literature, see [16]. Besides, Sobolev[8] spaces are required in the theory of weak formulations. A brief introduction to Sobolev spaces will be given here and we will introduce the weak derivative, which is also required for weak formulations. □

[1] Henri Léon Lebesgue (1875 - 1941)
[2] William Henry Young (1863 - 1942)
[3] Baron Augustin-Louis Cauchy (1789 - 1857)
[4] Hermann Schwarz (1843 - 1921)
[5] Otto Hölder (1859 - 1937)
[6] Jules Henri Poincaré (1854 - 1912)
[7] Kurt Otto Friedrichs (1901 - 1982)
[8] Sergei Sobolev (1908 - 1989)

4.1 Weak Derivative and Sobolev Spaces

Definition 4.2. Weak derivative. Let $u \in L^1_{loc}(a,b)$. A function $v \in L^1_{loc}(a,b)$ is a weak derivative of u if it holds that

$$\int_a^b u(x)\varphi'(x)\,dx = -\int_a^b v(x)\varphi(x)\,dx$$

for all $\varphi \in C_0^\infty(a,b) \subset C^\infty(a,b)$, i.e., for all infinitely differentiable functions with compact support in (a,b). We also write $u'(x) = v(x)$. □

Lemma 4.3. Fundamental lemma of calculus of variations. *Let $u \in L^1_{loc}(a,b)$ and suppose that*

$$\int_a^b u(x)\varphi'(x)\,dx = 0$$

for all $\varphi \in C_0^\infty(a,b)$. Then $u(x) = 0$ almost everywhere.

Proof. For the proof, it is referred to the literature, see [15]. □

Lemma 4.4. Uniqueness of the weak derivative. *Let $u \in L^1_{loc}(a,b)$. Then the weak derivative $u'(x)$ of $u(x)$ is uniquely determined.*

Proof. Assume that $\tilde{u}(x) \neq 0$ is another weak derivative of $u(x)$. It holds for all $\varphi \in C_0^\infty(a,b)$ that

$$\int_a^b \left(\tilde{u}'(x) - u'(x)\right)\varphi(x)\,dx = \int_a^b \left(-u(x) + u(x)\right)\varphi'(x)\,dx = 0.$$

Then it follows from the fundamental lemma of calculus of variations that $\tilde{u}'(x) = u'(x)$, which is a contradiction to the assumption. □

Lemma 4.5. Linearity of the weak derivative. *Let $u_1, u_2 \in L^1_{loc}(a,b)$ and suppose that there exist weak derivatives $v_1 = u_1'$, $v_2 = u_2'$. Then there exists $(\alpha u1 + \beta u2)'$, $\alpha, \beta \in \mathbb{R}$, and*

$$(\alpha u_1 + \beta u_2)' = \alpha v_1 + \beta v_2.$$

Proof. Using Definition 4.2, one obtains

$$\begin{aligned}\int_a^b (\alpha u_1(x) + \beta u_2(x))\varphi'(x)\,dx &= \alpha\int_a^b u_1(x)\varphi'(x)\,dx + \beta\int_a^b u_2(x)\varphi'(x)\,dx \\ &= -\alpha\int_a^b v_1(x)\varphi(x)\,dx - \beta\int_a^b v_2(x)\varphi(x)\,dx \\ &= -\int_a^b (\alpha v_1(x) + \beta v_2(x))\varphi(x)\,dx\end{aligned}$$

for all $\varphi \in C_0^\infty(a,b)$. Hence, $(\alpha u_1 + \beta u_2)' = \alpha v_1 + \beta v_2$. □

Remark 4.6. *Weak derivative and classical derivative.* Each classical derivative is also a weak derivative. But a derivative can exist in the weak sense without existing in the classical sense. □

Example 4.7. *The absolute value function.* Consider the absolute value function $u(x) = |x|$, which is not differentiable at $x = 0$. But it has a weak derivative

$$v(x) = \begin{cases} -1, & \text{if } x < 0, \\ 1, & \text{if } x > 0, \end{cases}$$

since it holds for all $\varphi \in C_0^\infty(a, b)$ that

$$\begin{aligned} \int_{\mathbb{R}} u(x)\varphi'(x)\,dx &= \int_{-\infty}^{0} -x\varphi'(x)\,dx + \int_0^\infty x\varphi'(x)\,dx \\ &= -x\varphi(x)\Big|_{-\infty}^{0} - \int_{-\infty}^{0} -\varphi(x)\,dx + x\varphi(x)\Big|_0^\infty + - \int_0^\infty \varphi(x)\,dx \\ &= \int_{-\infty}^{0} \varphi(x)\,dx - \int_0^\infty \varphi(x)\,dx \\ &= -\int_{\mathbb{R}} v(x)\varphi(x)\,dx. \end{aligned}$$

□

Definition 4.8. Sobolev spaces $H^k(a, b)$. Let $k \in \mathbb{N}_0$. The Sobolev space $H^k(a, b)$ consists of all $u \in L^2(a, b)$ such that weak derivatives $u^{(i)} \in L^2(a, b)$ for $i \in \mathbb{N}$, $1 \le i \le k$, i.e.,

$$H^k(a, b) = W^{k,2}(a, b) := \big\{u \in L^2(a, b)\big|\ u^{(i)} \in L^2(a, b),\ i = 1, \dots, k\big\}.$$

Obviously, it holds $H^0(a, b) = L^2(a, b)$, which is also a Lebesgue space.
The scalar product of $H^1(a, b)$ is defined by

$$(u_1, u_2)_{H^1} := \int_a^b (u_1(x)u_2(x) + u_1'(x)u_2'(x))\,dx,$$

and the norm is then given by

$$||u||_{H^1} := (u, u)_{H^1}^{1/2}.$$

□

Definition 4.9. The Sobolev space $H_0^1(a, b)$. The Sobolev space $H_0^1(a, b)$ is defined by

$$H_0^1(a, b) := \big\{u \in H^1(a, b)\big|\ u(a) = u(b) = 0\big\}.$$

It is equipped with the scalar product

$$(u_1, u_2)_{H_0^1} := \int_a^b u_1'(x)u_2'(x)\,dx,$$

and with the norm

$$\|u\|_{H_0^1} := (u,u)_{H_0^1}^{1/2}.$$

The boundary values are defined in the sense of traces here. Using Poincaré's inequality, one can show that $(u_1,u_2)_{H_0^1}$ is a scalar product in $H_0^1(a,b)$. □

Definition 4.10. The dual space $H^{-1}(a,b)$ of $H_0^1(a,b)$. The dual space $H^{-1}(a,b)$ of $H_0^1(a,b)$ consists of all continuous linear functionals on $H_0^1(a,b)$. □

Remark 4.11. On the different spaces.

- The Lebesgue spaces $L^p(a,b)$, $p \in [1,\infty)$ are complete normed spaces, i.e., Banach[9] spaces. Hilbert spaces are a subset of Banach spaces. For example, the Lebesgue space $L^2(a,b)$, which is equipped with the scalar product

  $$(u_1,u_2)_{L^2} = \int_a^b u_1(x)u_2(x)\,dx,\ u_1,u_2 \in L^2(a,b),$$

  and with the norm

  $$\|u\|_{L^2} = \|u\|_2 = (u,u)_{L^2}^{1/2},$$

  is a Hilbert space.
- The Sobolev spaces $H^k(a,b)$, $k \in \mathbb{N}_0$ are Hilbert spaces.
- Generally, $H^{-k}(a,b)$, $k \in \mathbb{N}$ is the dual space of $H_0^k(a,b)$. It is $H^{-k}(a,b) = \big(H_0^k(a,b)\big)'$. In particular, it is $H^{-1}(a,b) = \big(H_0^1(a,b)\big)'$.
- $H_0^1(a,b) \subset L^2(a,b) \subset H^{-1}(a,b)$ is a so-called Gelfand[10] triple. □

Before we get into the weak formulation, we will introduce the bilinear form in this section, which will be used later.

Definition 4.12. Bilinear form. A bilinear form on a Banach space V is a mapping $B(\cdot,\cdot) : V \times V \to \mathbb{R}$, which is linear in each argument:

$$B(u, v_1 + \lambda v_2) = B(u,v_1) + \lambda B(u,v_2),\ u, v_1, v_2 \in V, \lambda \in \mathbb{R},$$
$$B(u_1 + \lambda u_2, v) = B(u_1,v) + \lambda B(u_2,v),\ u_1, u_2, v \in V, \lambda \in \mathbb{R}.$$

The bilinear form $B(\cdot,\cdot)$ is called

i) symmetric, if $B(u,v) = B(v,u)$, $u,v \in V$,

ii) positive, if $B(u,u) \geq 0$, $u \in V$,

iii) coercive, if there is a constant $\mu > 0$, such that for all $u \in V$,

$$B(u,u) \geq \mu \|u\|_V^2,$$

iv) bounded, if there is a constant $\delta > 0$, such that for all $u,v \in V$,

$$|B(u,v)| \leq \delta \|u\|_V \|v\|_V.$$

□

[9] Stefan Banach (1892 - 1945)
[10] Israel Moiseevich Gelfand (1913 - 2009)

4.2 Weak Formulation and Weak Solution

Remark 4.13. *The Poisson*[11] *equation with homogeneous Dirichlet boundary conditions.* Consider the simplified model problem (3.3), (3.4) with $\varepsilon = 1$, $b(x) = c(x) = 0$, i.e.,

$$-u'' = f(x) \quad \text{for } x \in (0,1), \tag{4.1}$$

with homogeneous Dirichlet boundary conditions

$$u(0) = u(1) = 0.$$

The differential equation (4.1) is a so-called Poisson equation. Next, we will concentrate on the Poisson equation with homogeneous Dirichlet boundary conditions for investigation of the weak solution's behaviour. □

Remark 4.14. *Derivation of the weak formulation.* Consider the Poisson equation (4.1) with homogeneous Dirichlet boundary conditions. By multiplying (4.1) with a so-called test function $v(x)$, $v(0) = v(1) = 0$, then integrating the equation by parts over $(0,1)$, one obtains

$$\begin{aligned}\int_0^1 -u''(x)v(x)\,dx &= u'(x)v(x)\Big|_0^1 - \int_0^1 -u'(x)v'(x)\,dx \\ &= \int_0^1 u'(x)v'(x)\,dx \\ &= \int_0^1 f(x)v(x)\,dx.\end{aligned}$$

This transformation only makes sense when a proper space is chosen, such that all integrals are well defined. The Sobolev space $H_0^1(0,1)$ and $L^2(0,1)$ will be considered in this chapter and we denote

$$a(u,v) := (u',v')_{L^2} = (u,v)_{H_0^1} = \int_0^1 u'(x)v'(x)\,dx.$$

Let $f \in H^{-1}(0,1)$. Then one gets

$$f(v) := (f,v)_{H^{-1},H_0^1} = \int_0^1 f(x)v(x)\,dx,$$

where $f(v)$ is a linear functional f acting on $v \in H_0^1(0,1)$, and $(\cdot,\cdot)_{H^{-1},H_0^1}$ is the dual pairing of the spaces $H^{-1}(0,1)$ and $H_0^1(0,1)$. □

Example 4.15. *A bilinear form* $a(\cdot,\cdot)$. Consider the Sobolev space $H_0^1(0,1)$ and let $u, v \in H_0^1(0,1)$. It follows directly from the linearity of differentiation and integration that $a(\cdot,\cdot)$ is a bilinear form on $H_0^1(0,1)$. The bilinear form $a(\cdot,\cdot)$ is

i) symmetric, since $a(u,v) = a(v,u)$,

[11] Siméon Denis Poisson (1781 - 1840)

ii) positive, since $a(u,u) = (u', u')_{L^2} = \|u\|_2^2 \geq 0$,

iii) coercive, since $a(u,u) = (u,u)_{H_0^1} = \|u\|_{H_0^1}^2$,

iv) bounded. Using the Cauchy-Schwarz inequality yields

$$|a(u,v)| = |(u,v)_{H_0^1(0,1)}| \leq \|u\|_{H_0^1(0,1)} \|v\|_{H_0^1(0,1)}.$$

□

Definition 4.16. Weak formulation and weak solution. Let $f \in L^2(0,1) \subset H^{-1}(0,1)$. The weak formulation of the Poisson equation (4.1) with homogeneous Dirichlet boundary conditions is to find $u \in H_0^1(0,1) \subset L^2(0,1)$, such that

$$a(u,v) = f(v) \quad \forall\ v \in H_0^1(0,1). \tag{4.2}$$

The functions $v(x)$ are called test functions. A solution of (4.2) is called a weak solution. □

Remark 4.17. *Weak solution and classical solution.* Each classical solution is also a weak solution. The corresponding space of the weak solution is called ansatz space or solution space. Compared to the classical solution in $C^2(0,1) \cap C([0,1])$, only the first weak derivative of the weak solution is required. □

Theorem 4.18. Riesz[12] representation theorem. *Let H be a Hilbert space and H' be the dual space of H. Then for each continuous linear functional $f \in H'$ there exists a unique $u \in H$ such that*

$$(u,v)_H = f(v) \quad \forall\ v \in H.$$

Proof. Let $ker(f) := \{v \in H |\ f(v) = 0\}$, which is a linear subspace of H. Let $\{v_k\}_{k\in\mathbb{N}}$ be a sequence in $ker(f)$ with $v_k \to v$ for $k \to \infty$. In fact, since f is continuous and linear, it is also bounded. Using the continuity, linearity and boundedness of f, it holds for all $k \in \mathbb{N}$ that

$$|f(v)| = |f(v) - f(v_k)| = |f(v - v_k)| \leq \|f\|_{H'} \|v - v_k\|_H.$$

Hence, $|f(v)| = 0$. If $f = 0$, then $H = ker(f)$. Since $(v,v)_H = \|v\|_H^2 > 0$ for all $v \neq 0$, it follows that $u = 0$ is the unique solution.

If $f \neq 0$, then $H \neq ker(f)$. Let $(ker(f))^\perp$ be the orthogonal complement of $ker(f)$, i.e., $H \cong ker(f) \oplus (ker(f))^\perp$. Consider $z \in (ker(f))^\perp$ with $f(z) = 1$. It follows from the linearity of f that

$$f(v - f(v)z) = f(v) - f(v)f(z) = 0 \quad \forall v \in H,$$

i.e., $(v - f(v)z) \in ker(f)$. So it holds for all $v \in H$ that

$$0 = (z, v - f(v)z)_H = (z,v)_H - f(v)(z,z)_H.$$

[12] Frigyes Riez (1880 - 1956)

Thus, $(u, v) = f(v)$ for $u = (\|z\|_H^2)^{-1} z$.

Until now we have shown the existence of u. At last, the uniqueness of u will be proved. Assume $\tilde{u}$ is another solution. It holds

$$(u - \tilde{u}, v)_H = (u, v)_H - (\tilde{u}, v)_H = f(v) - f(v) = 0 \quad \forall v \in H.$$

Hence, $u = \tilde{u}$, which is a contradiction to the assumption. □

Lemma 4.19. Lax[13]-Milgram[14] Theorem. *Let H be a Hilbert space and $B(\cdot, \cdot)$ be a bounded and coercive bilinear form on H. Then, for each bounded linear functional $f \in H'$ there exists a unique $u \in H$ such that*

$$B(u, v) = f(v) \quad \forall v \in H.$$

Proof. One can prove the Lax-Milgram theorem with the help of the Riesz representation theorem. For the proof, it is referred to the literature, see [4]. □

Corollary 4.20. Existence and Uniqueness of the weak solution of the weak formulation. *Let $f \in L^2(0, 1) \subset H^{-1}(0, 1)$. Then there exists a unique weak solution of (4.2).*

Proof. It follows from Remark 4.14, Example 4.15 and Theorem 4.18. □

Remark 4.21. *Well-posedness of the weak formulation (4.2).* According to Corollary 4.20, the weak formulation (4.2) admits a unique solution. Setting the test function $v = u$ and using the Cauchy-Schwarz inequality for the right-hand side of (4.2), one gets $\|u\|_{H_0^1} \leq \|f\|_{H^{-1}}$, which implies that u changes continuously with f. Thus, the weak formulation (4.2) is well-posed. □

Remark 4.22. *The Lax-Milgram theorem for non-symmetric bilinear forms.* The Lax-Milgram theorem is a generalization of the Riesz representation theorem. The main interest of the Lax-Milgram theorem is that it also works without the bilinear form $a(\cdot, \cdot)$ being symmetric. □

Example 4.23. Consider a Poisson equation (4.1) with homogeneous Dirichlet boundary conditions, where

$$f(x) = \begin{cases} 16, & \text{if } 0 < x < \frac{1}{2}, \\ -16, & \text{if } \frac{1}{2} \leq x < 1. \end{cases}$$

Since $f(x)$ is discontinuous at $x = \frac{1}{2}$, this problem does not have a classical solution. However, this problem has a weak solution

$$u(x) = \begin{cases} 8x^2 - 4x, & \text{if } 0 < x < \frac{1}{2}, \\ -8x^2 + 12x - 4, & \text{if } \frac{1}{2} \leq x < 1, \end{cases}$$

[13] Peter David Lax (1926 -)

[14] Arthur Norton Milgram (1912 - 1961)

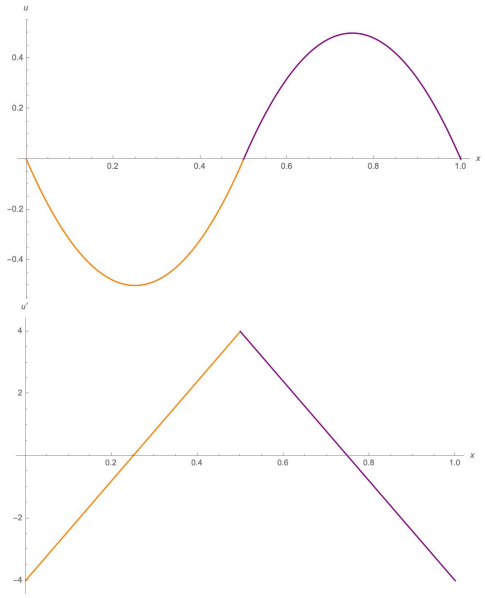

Figure 4.1: The weak solution of Example 4.21 and its first derivative.

and its first derivative is

$$u'(x) = \begin{cases} 16x - 4, & \text{if } 0 < x < \frac{1}{2}, \\ -16x + 12, & \text{if } \frac{1}{2} \leq x < 1, \end{cases}$$

see Figure 4.1.

The weak formulation of this problem is to find $u \in H_0^1(0,1)$, such that

$$a(u,v) = f(v) \quad \forall\ v \in H_0^1(0,1),$$

since it holds for all $v \in H_0^1(0,1)$ that

$$\begin{aligned} a(u,v) &= \int_0^1 u'(x)v'(x)\,dx \\ &= \int_0^{\frac{1}{2}} u'(x)v'(x)\,dx + \int_{\frac{1}{2}}^1 u'(x)v'(x)\,dx \\ &= u'(x)v(x)\big|_0^{\frac{1}{2}} - \int_0^{\frac{1}{2}} u''(x)v(x)\,dx + u'(x)v(x)\big|_{\frac{1}{2}}^1 - \int_{\frac{1}{2}}^1 u''(x)v(x)\,dx \\ &= \int_0^1 f(x)v(x)\,dx \\ &= f(v). \end{aligned}$$

□

4.3 The Minimization Problem

Remark 4.24. *Another weak formulation.* The Poisson equation with homogeneous Dirichlet boundary conditions can be interpreted as a so-called minimization problem, which is equivalent to the weak formulation (4.2). □

Definition 4.25. The minimization problem. Let $f \in L^2(0,1) \subset H^{-1}(0,1)$. The minimization problem of the Poisson equation (4.1) with homogeneous Dirichlet boundary conditions is to find $u \in H_0^1(0,1)$, which minimizes the so-called energy functional

$$F(v) := \frac{1}{2}a(v,v) - f(v) \quad \forall v \in H_0^1(0,1), \tag{4.3}$$

i.e.,

$$F(u) \leq F(v) \quad \forall\, v \in H_0^1(0,1).$$

□

Lemma 4.26. The solution of (4.3). *Under the assumptions of Definition 4.16, $u \in H_0^1(0,1)$ is the solution of (4.2) if and only if u minimizes (4.3).*

Proof. Suppose that $u \in H_0^1(0,1)$ is the solution of (4.2). Using the bilinearity and positiveness of $a(\cdot,\cdot)$ and the linearity of f, one obtains

$$\begin{aligned} F(v) &= F(u) + \frac{1}{2}a(v-u, v-u) + a(u, v-u) - f(v-u) \\ &= F(u) + \frac{1}{2}a(v-u, v-u) \geq F(u) \quad \forall v \in H_0^1(0,1). \end{aligned}$$

The bilinear form $a(v-u, v-u)$ is equal to zero here if and only if $u = v$.

Now suppose that u minimizes (3.3). Let $v \in H_0^1(0,1)$ be arbitrary, and $\lambda \in (0,1)$. Using the bilinearity of $a(\cdot,\cdot)$ and the linearity of f again, one obtains

$$\begin{aligned} F(u+\lambda v) - F(u) &= \left(\frac{1}{2}a(u+\lambda v, u+\lambda v) - f(u+\lambda v)\right) - \left(\frac{1}{2}a(u,u) - f(u)\right) \\ &= \frac{1}{2}\lambda^2 a(v,v) + \lambda(a(u,v) - f(v)) \geq 0 \quad \forall v \in H_0^1(0,1). \end{aligned}$$

Dividing the above mentioned inequality by λ and then setting $\lambda \to 0$ yields

$$a(u,v) \geq f(v) \quad \forall\, v \in H_0^1(0,1).$$

Similarly, let $\lambda \in (-1,0)$ and after a direct calculation, one gets the equality. □

Remark 4.27. *Well-posedness of the minimization problem (4.3).* Since the weak formulation (4.2) and the minimization problem (4.3) are equivalent, it follows from Corollary 4.20, Remark 4.21 and Lemma 4.26 that the minimization problem (4.3) is also well-posed. □

Chapter 5

Finite Element Method

Remark 5.1. *Idea.* It is widely known that the three classical methods to obtain the numerical solution of differential equations are the finite difference method, the finite element method and the finite volume method. The most frequently used finite element method is the so-called the Ritz[1] method, which is commonly called the Rayleigh[2]-Ritz method or the Ritz-Galerkin[3] method according to different assumptions.

The Ritz method is based on the weak formulation that we have constructed in Chapter 4. The idea is to reduce the problem to finite-dimensional subspaces and then numerically compute the solution as a finite linear combination of the basis vectors in the subspaces. □

5.1 The Ritz Method

Definition 5.2. The Ritz approximation. Let H be a separable Hilbert space, which has a countable orthonormal basis. There are finite-dimensional subspaces $H_k \subset H$, $k = 1, \ldots, n$, such that, for each $u \in H$ and $\varepsilon \geq 0$ there is a $N \in \mathbb{N}$ and a $u_k \in H_k$ with

$$\|u - u_k\|_H \leq \varepsilon \quad \forall\ k \geq N. \tag{5.1}$$

The Ritz approximation of (4.2) is to find $u_k \in H_k$ such that

$$a(u_k, v_k) = f(v_k) \quad \forall\ v_k \in H_k. \tag{5.2}$$

Note that the bilinear form $a(\cdot, \cdot)$ is symmetric here and the inclusion of the subspaces $H_k \subset H_{k+1}$ is not required. □

Lemma 5.3. Existence and uniqueness of the solution of (5.2). *Under the assumptions of Definition 5.2, there exists a unique solution of (5.2).*

[1] Walther Ritz (1878 - 1909)
[2] John William Strutt, 3^{rd} Baron Rayleigh (1842 - 1919)
[3] Boris Grigoryevich Galerkin (1871 - 1945)

Proof. Since the subspaces $H_k \subset H$ are still Hilbert spaces, one can apply the Riesz representation theorem to (5.2), which implies that there exists a unique solution of (5.2). □

Lemma 5.4. Consistency of the Ritz approximation, Galerkin orthogonality. *Let u and u_k be the solutions of (4.2) and (5.2) respectively. Then the error $\epsilon_k := u - u_k$ is orthogonal to the subspace H_k, i.e.,*

$$a(\epsilon_k, v_k) := a(u - u_k, v_k) = 0 \quad \forall\ v_k \in H_k.$$

Proof. Using the linearity of $a(\cdot, \cdot)$ and Definition 5.2, one obtains

$$a(u - u_k, v_k) = a(u, v_k) - a(u_k, v_k) = f(v_k) - f(v_k) = 0.$$

Hence, The error ϵ_k is orthogonal to H_k, i.e., $\epsilon_k \perp H_k$. □

Lemma 5.5. Best approximation property, convergence of the Ritz approximation. *Let u and u_k be the solutions of (4.2) and (5.2) respectively. Then u_k is the best approximation of u in H_k, i.e.,*

$$\|u - u_k\|_H \leq \|u - v_k\|_H \quad \forall\ v_k \in H_k.$$

Hence, the sequence of the Ritz approximation (u_k) converges to u with respect to the norm in H, i.e.,

$$\lim_{k \to \infty} \|u - u_k\|_H = 0.$$

Proof. Obviously, if $\|u - u_k\|_H = 0$, then u_k is the best approximation of u.

Suppose that $\|u - u_k\|_H \neq 0$. Let $v_k = u_k - w_k$ with an arbitrary $w_k \in H_k$. Using the linearity of $a(\cdot, \cdot)$, the Galerkin orthogonality and the Cauchy–Schwarz inequality yields

$$\begin{aligned} \|u - u_k\|_H^2 &= a(u - u_k, u - u_k) \\ &= a(u - u_k, u - (v_k + w_k)) \\ &= a(u - u_k, u - v_k) - a(u - u_k, w_k) \\ &= a(u - u_k, u - v_k) \leq \|u - u_k\|_H \|u - v_k\|_H. \end{aligned}$$

Dividing the above mentioned inequality by $\|u - u_k\|_H$ gives the statement of the best approximation.

Then it follows from the best approximation property and (5.1) that

$$\|u - u_k\|_H \leq \|u - v_k\|_H \leq \epsilon \quad \forall\ v_k \in H_k,$$

which implies that

$$\lim_{k \to \infty} \|u - u_k\|_H = 0.$$

□

Remark 5.6. *Well-posedness of the Ritz approximation (4.2), stability of the Ritz approximation.* Since the subspaces $H_k \subset H$ are still Hilbert spaces, the properties of the bilinear form $a(\cdot,\cdot)$ on H_k can be inherited from them on H. From Lemma 5.3 we know that there exists a unique solution of (5.2). In addition, the proof of stability on the data for the discrete problem is analog as for the continuous problem, see Remark 4.21. Hence, the Ritz approximation is well-posed. □

Remark 5.7. *Matrix-vector form of the Ritz method.* One can reformulate the Ritz method as a linear system of equations, which can be used to compute the numerical solution of the Ritz approximation (5.2) algorithmically. Let $\{\phi_i\}_{i=1}^k$ be a set of basis functions of H_k and assume that $f \in L^2(0,1)$. Then $v_k \in H_k$ can be represented as a linear combination of the basis functions $\{\phi_i\}_{i=1}^k$ in H_k, i.e., $v_k = \sum_{i=1}^k \lambda_i \phi_i$, $\lambda_i \in \mathbb{R}$. Since $a(\cdot,\cdot)$ is a bilinear form and f is a linear functional, it is sufficient to test (5.2) with these basis functions, i.e.,

$$\begin{aligned} a(u_k, v_k) &= a\left(u_k, \sum_{i=1}^k \lambda_i \phi_i\right) \\ &= \sum_{i=1}^k \lambda_i a(u_k, \phi_i) \\ &= \sum_{i=1}^k \lambda_i f(\phi_i) \\ &= f\left(\sum_{i=1}^k \lambda_i \phi_i\right) = f(v_k), \end{aligned}$$

where $a(u_k, \phi_i) = f(\phi_i)$ for $i = 1, \ldots, k$, see Definition 5.2. Expanding u_k with respect to these basis functions, i.e., $u_k = \sum_{j=1}^k \zeta_j \phi_j$, $\zeta_j \in \mathbb{R}$, and then inserting it into the equation above, one obtains

$$a(u_k, \phi_i) = a\left(\sum_{j=1}^k \zeta_j \phi_j, \phi_i\right) = \sum_{j=1}^k \zeta_j a(\phi_j, \phi_i) = f(\phi_i) = (f, \phi_i)_{L^2}$$

for $i = 1, \ldots, k$, where the coefficients $\zeta_j \in \mathbb{R}, j = 1, \ldots, k$ are to be determined for the numerical solution u_k. This previous equation gives the matrix-vector form $A\underline{u} = \underline{f}$, where

$$A_{ij} = a(\phi_j, \phi_i), \quad \underline{u} = (\zeta_1, \ldots, \zeta_k)^T, \quad \underline{f} = (f(\phi_1), \ldots, f(\phi_k))^T,$$

The matrix A is called stiffness matrix, and the right-hand side vector $\underline{f}$ is called load vector. The symmetry and positive definiteness of the stiffness matrix A follow directly from the symmetry and positiveness of the bilinear form $a(\cdot,\cdot)$. □

5.2 Mesh and Basis Functions

Remark 5.8. *Goal.* In the last section, we have introduced how the Ritz method solves the Poisson equation (4.1) with homogeneous Dirichlet boundary conditions in the weak formulation (4.2). Now we need to generate a mesh on which a set of basis functions based. □

Definition 5.9. Mesh, node, mesh cell, and step size. For the model problem, a mesh is a decomposition I_h of the interval $I = [0, 1]$

$$I_h = \{x_0 = 0, x_1, \ldots, x_N = 1\}$$

with $x_0 < x_1 < \cdots < x_N$. Here, $x_i, i = 0, \ldots, N$, is called a node, or nodal point, and $(x_0, x_1), \ldots, (x_{N-1}, x_N)$ are called mesh cells. The differences between neighboring nodes $h_i = x_{i+1} - x_i$ are called step sizes. For an equidistant mesh on $I = [0, 1]$, we denote

$$x_i = ih, \quad i = 0, \ldots, N, \quad h = h_i = \frac{1}{N}.$$

□

Remark 5.10. *Choices of the basis functions.* Let the solution be in the separable Hilbert space H, which is $H_0^1(0, 1)$ in the model problem. Then there are finite-dimensional spaces $H_k \subset H$, $k = 1, \ldots, n$ based on the grid, see Definition 5.2. It is worth mentioning that different finite-dimensional spaces generate different finite element solutions. Hence, it is important to choose appropriate linearly independent basis functions $\{\phi_i\}_{i=1}^k$, which can span these finite-dimensional spaces, i.e.,

$$H_k = \left\{ v_k \middle| v_k = \sum_{i=1}^{k} \lambda_i \phi_i, \lambda_i \in \mathbb{R} \right\}.$$

From the numerical point of view, the stiffness matrix A should be sparse. Otherwise, the computation of integrals in the matrix A would be very costly. So we should choose a set of basis functions, which

- is simple so that the integrals can be computed easily,
- is zero in most mesh cells so that the stiffness matrix A can be sparse,
- and is continuous and differentiable except at nodes.

Next, the piecewise linear basis functions, the piecewise quadratic basis functions and the Bernstein basis polynomials will be considered and discussed separately. □

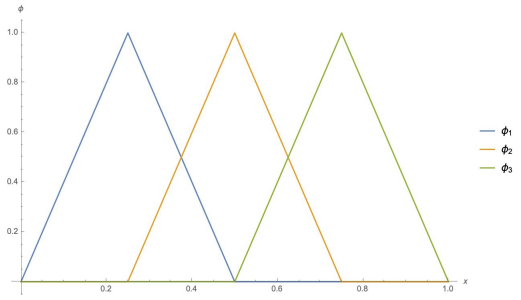

Figure 5.1: Hat functions for $N = 4$.

5.2.1 Piecewise linear basis functions

Definition 5.11. Hat functions. The simplest piecewise linear basis functions are so-called hat functions. They have the form for $i = 1, \ldots, N-1$

$$\phi_i(x) = \begin{cases} \frac{x-x_{i-1}}{x_i-x_{i-1}} = \frac{x-x_{i-1}}{h_i}, & \text{if } x \in [x_{i-1}, x_i], \\ \frac{x_{i+1}-x}{x_{i+1}-x_i} = \frac{x_{i+1}-x}{h_{i+1}}, & \text{if } x \in [x_i, x_{i+1}], \\ 0, & \text{otherwise.} \end{cases}$$

Note that

$$\phi_i(x_j) = \delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases} \quad i, j = 1, \ldots, N-1,$$

when x_j is a node in the mesh with global node number j. The spanned finite element space $\operatorname{span}\{\phi_i(x)\}_{i=1}^{N-1}$ has the finite dimension $N-1$. For an equidistant mesh on $I = [0, 1]$, see Figure 5.1 for examples for $N = 4$. □

Remark 5.12. *Derivation of the matrix-vector form for hat functions.* According to Definition 5.9, the interval $I = [0, 1]$ can be divided into non-overlapping mesh cells $(x_0, x_1), \ldots, (x_{N-1}, x_N)$. We denote $(x_0, x_1), \ldots, (x_{N-1}, x_N)$ as elements. For the model problem, the stiffness matrix entries $A_{ij} = a(\phi_j, \phi_i) = \int_0^1 \phi_j' \phi_i' \, dx$, $i, j = 1, \ldots, N-1$ can be split up element by element, i.e.,

$$A = \begin{pmatrix} \int_{x_0}^{x_1} (\phi_1')^2 \, dx & \int_{x_0}^{x_1} \phi_1' \phi_2' \, dx & \ldots & \int_{x_0}^{x_1} \phi_1' \phi_{N-1}' \, dx \\ \int_{x_0}^{x_1} \phi_2' \phi_1' \, dx & \int_{x_0}^{x_1} (\phi_2')^2 \, dx & \ldots & \int_{x_0}^{x_1} \phi_2' \phi_{N-1}' \, dx \\ \vdots & \vdots & \ddots & \vdots \\ \int_{x_0}^{x_1} \phi_2' \phi_1' \, dx & \int_{x_0}^{x_1} (\phi_2')^2 \, dx & \ldots & \int_{x_0}^{x_1} (\phi_{N-1}')^2 \, dx \end{pmatrix} + \ldots$$

$$+ \begin{pmatrix} \int_{x_{N-1}}^{x_N} (\phi_1')^2 \, dx & \int_{x_{N-1}}^{x_N} \phi_1' \phi_2' \, dx & \ldots & \int_{x_{N-1}}^{x_N} \phi_1' \phi_{N-1}' \, dx \\ \int_{x_{N-1}}^{x_N} \phi_2' \phi_1' \, dx & \int_{x_{N-1}}^{x_N} (\phi_2')^2 \, dx & \ldots & \int_{x_{N-1}}^{x_N} \phi_2' \phi_{N-1}' \, dx \\ \vdots & \vdots & \ddots & \vdots \\ \int_{x_{N-1}}^{x_N} \phi_2' \phi_1' \, dx & \int_{x_{N-1}}^{x_N} (\phi_2')^2 \, dx & \ldots & \int_{x_{N-1}}^{x_N} (\phi_{N-1}')^2 \, dx \end{pmatrix}.$$

In each element, since there are not more than two neighboring hat functions, then each row has at most two nonzero entries, i.e.,

$$A = \begin{pmatrix} \int_{x_0}^{x_1} (\phi_1')^2 \, dx & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} + \begin{pmatrix} \int_{x_1}^{x_2} (\phi_1')^2 \, dx & \int_{x_1}^{x_2} \phi_1' \phi_2' \, dx & \dots & 0 \\ \int_{x_1}^{x_2} \phi_2' \phi_1' \, dx & \int_{x_1}^{x_2} (\phi_2')^2 \, dx & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}$$
$$+ \cdots + \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \int_{x_{N-1}}^{x_N} (\phi_{N-1}')^2 \, dx \end{pmatrix}.$$

For $i = 1, \dots, N-2$, in the element (x_i, x_{i+1}), there are exactly two nonzero hat functions. A straightforward calculation gives

$$\int_{x_i}^{x_{i+1}} (\phi_i')^2 \, dx = \int_{x_i}^{x_{i+1}} \frac{1}{h_i^2} \, dx = \frac{1}{h_i}, \quad \int_{x_i}^{x_{i+1}} (\phi_{i+1}')^2 \, dx = \int_{x_i}^{x_{i+1}} \frac{1}{h_i^2} \, dx = \frac{1}{h_i},$$
$$\int_{x_i}^{x_{i+1}} \phi_i' \phi_{i+1}' \, dx = \int_{x_i}^{x_{i+1}} \phi_{i+1}' \phi_i' \, dx = \int_{x_i}^{x_{i+1}} -\frac{1}{h_i^2} \, dx = -\frac{1}{h_i}.$$

In the elements (x_0, x_1) and (x_{N-1}, x_N), there is only one nonzero hat function, then one has

$$\int_{x_0}^{x_1} (\phi_1')^2 \, dx = \frac{1}{h_0}, \quad \int_{x_{N-1}}^{x_N} (\phi_{N-1}')^2 \, dx = \frac{1}{h_{N-1}}.$$

Finally, we can add elementwise contributions together, which is called assemble finite elements. Then one gets a tridiagonal stiffness matrix A, i.e.,

$$A = \begin{pmatrix} \frac{1}{h_0} + \frac{1}{h_1} & -\frac{1}{h_1} & & & & \\ -\frac{1}{h_1} & \frac{1}{h_1} + \frac{1}{h_2} & -\frac{1}{h_2} & & & \\ & \ddots & \ddots & \ddots & & \\ & & -\frac{1}{h_{N-3}} & \frac{1}{h_{N-3}} + \frac{1}{h_{N-2}} & -\frac{1}{h_{N-2}} & \\ & & & -\frac{1}{h_{N-2}} & \frac{1}{h_{N-2}} + \frac{1}{h_{N-1}} \end{pmatrix}.$$

This process of adding elementwise contributions together is called finite element assembly or simply assembly. Notice that the stiffness matrix A is sparse.

Similarly, the load vector $\underline{f}$ can also be assembled element by element, i.e.,

$$\underline{f} = \begin{pmatrix} \int_{x_0}^{x_1} f\phi_1 \, dx \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \int_{x_1}^{x_2} f\phi_1 \, dx \\ \int_{x_1}^{x_2} f\phi_2 \, dx \\ \vdots \\ 0 \\ 0 \end{pmatrix} + \cdots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \int_{x_{N-1}}^{x_N} f\phi_{N-1} \, dx \end{pmatrix}.$$

□

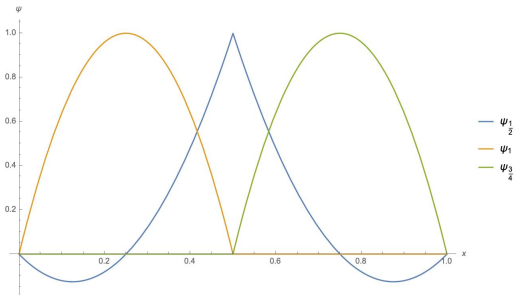

Figure 5.2: Piecewise quadratic basis functions for $N = 2$.

Example 5.13. *The stiffness matrix A for hat functions.* For an equidistant mesh on $I = [0, 1]$, one has $h = h_i = 1/N$. Then one obtains the stiffness matrix

$$A = \frac{1}{h} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{(N-1)\times(N-1)}.$$

□

5.2.2 Piecewise quadratic basis functions

Definition 5.14. Piecewise quadratic basis functions. The typical piecewise quadratic basis functions are Lagrange basis polynomials of second order. They have the form for $i = 1, \ldots, N-1$

$$\psi_i(x) = \begin{cases} \frac{2(x-x_{i-1})\left(x-x_{i-1/2}\right)}{(x_i-x_{i-1})^2} = \frac{2}{h_i^2}\left(x-x_{i-1}\right)\left(x-x_{i-1/2}\right), & \text{if } x \in [x_{i-1}, x_i], \\ \frac{2(x_{i+1}-x)\left(x_{i+1/2}-x\right)}{(x_{i+1}-x_i)^2} = \frac{2}{h_{i+1}^2}\left(x_{i+1}-x\right)\left(x_{i+1/2}-x\right), & \text{if } x \in [x_i, x_{i+1}], \\ 0, & \text{otherwise,} \end{cases}$$

together with the bubble functions for $i = 1, \ldots, N$

$$\psi_{i-1/2}(x) = \begin{cases} \frac{4(x-x_{i-1})(x_i-x)}{(x_i-x_{i-1})^2} = \frac{4}{h_i^2}\left(x-x_{i-1}\right)\left(x_i-x\right), & \text{if } x \in [x_{i-1}, x_i], \\ 0, & \text{otherwise,} \end{cases}$$

where $x_{i-1/2}$, $i = 1, \ldots, N$ is the midpoint of the two neighboring nodes x_{i-1} and x_i. Note that for $i, j = 1, \ldots, N$

$$\psi_i(x_j) = \psi_{i-1/2}(x_{j-1/2}) = \delta_{ij} \quad \text{and} \quad \psi_i(x_{j-1/2}) = \psi_{i-1/2}(x_j) = 0.$$

The spanned finite element space $\operatorname{span}\{\psi_{1/2}(x), \psi_1(x), \ldots, \psi_{N-1}(x), \psi_{N-1/2}(x)\}$ has the finite dimension $2N-1$. For an equidistant mesh on $I = [0, 1]$, see Figure 5.2 for examples for $N = 2$. □

Example 5.15. *The stiffness matrix A for piecewise quadratic basis functions.* The derivation of the matrix-vector form for piecewise quadratic basis functions is referred to the literature, e.g., see [14]. The Lagrange basis polynomials of second order as polynomials are still relatively simple so that the integrals can be computed easily. Another advantage is that they are zero in most mesh cells, and the corresponding stiffness matrix A is sparse. For example, when $N = 3$, one gets the stiffness matrix

$$A = \begin{pmatrix} 16 & -8 & 0 & 0 & 0 \\ -8 & 14 & -8 & 1 & 0 \\ 0 & -8 & 16 & -8 & 0 \\ 0 & 1 & -8 & 14 & -8 \\ 0 & 0 & 0 & -8 & 16 \end{pmatrix} \in \mathbb{R}^{5\times 5}.$$

There are more nonzero entries in the matrix, and the linear system of equations in the same grid is larger for the same dimension. Although the computation of the stiffness matrix A is more costly compared to the hat functions, one gets a more accurate numerical solution. □

5.2.3 Bernstein basis polynomials

Remark 5.16. *Bernstein basis polynomials as a set of basis functions.* Based on the weak formulation, theoretically a lot of different sets of basis functions can be chosen to solve the model problem. The $n + 1$ Bernstein basis polynomials of degree n are continuous and differentiable except at nodes, however, from Lemma 2.5 iv) we know that $b_0^n(0) = 1$ and $b_n^n(1) = 1$. These two Bernstein basis polynomials are nonzero at the boundaries, which are not in the Sobolev space $H_0^1(0, 1)$.

In the proof of Lemma 2.5 ix), we have shown that the $n + 1$ Bernstein basis polynomials $b_0^n(x), \ldots, b_n^n(x)$ of degree n are linearly independent. Because of the linearly independence of the $n + 1$ Bernstein basis polynomials, after dropping $b_0^n(x)$ and $b_n^n(x)$, the remaining $n - 1$ Bernstein basis polynomials $b_1^n(x), \ldots, b_{n-1}^n(x)$ of degree n are still linearly independent. These can be a appropriate basis for the finite-dimensional spaces. □

Example 5.17. *The stiffness matrix A for Bernstein basis polynomials.* The Bernstein basis polynomials are not zero in most mesh cells on $x \in (0, 1)$, see Lemma 2.5 i) and iv). Consequently, the stiffness matrix A is not sparse in this case. Hence, there is no need to assemble element by element to form the stiffness matrix A for Bernstein basis polynomials. For example, if the Bernstein basis polynomials $b_1^4(x), b_2^4(x), b_3^4(x)$ of degree 4 are chosen to be a set of basis functions, see Figure 5.3. A direct calculation gives the stiffness matrix

$$A = \begin{pmatrix} \frac{48}{35} & \frac{12}{35} & -\frac{8}{35} \\ \frac{12}{35} & \frac{24}{35} & \frac{12}{35} \\ -\frac{8}{35} & \frac{12}{35} & \frac{48}{35} \end{pmatrix} \in \mathbb{R}^{3\times 3}.$$

Generally, the spanned finite element space $\operatorname{span}\{b_i^n(x)\}_{i=1}^{n-1}$ has the finite dimension $n - 1$. □

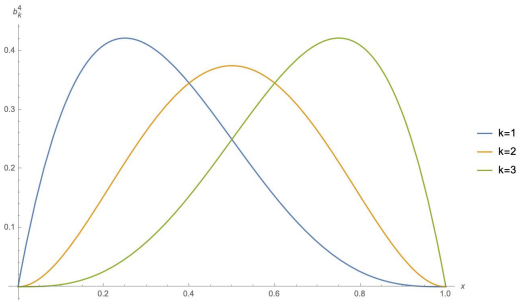

Figure 5.3: The Bernstein basis polynomials $b_1^4(x), b_2^4(x), b_3^4(x)$ of degree 4 as a set of basis functions.

5.3 A One-dimensional Example

Example 5.18. One-dimensional Poisson equation with homogeneous Dirichlet boundary conditions. Consider the following one-dimensional Poisson equation

$$-u'' = f(x) \quad \textit{for } x \in (0,1),$$

where $f(x) := 4e^{2x}$, with homogeneous Dirichlet boundary conditions

$$u(0) = u(1) = 0.$$

The analytical solution of this problem is

$$u(x) = -e^{2x} + e^2 x - x + 1.$$

The weak formulation of this problem is to find $u \in H_0^1(0,1)$, such that

$$a(u,v) = f(v) \quad \forall\ v \in H_0^1(0,1).$$

Let $H = H_0^1(0,1)$. Let $H_k \subset H$, $k = 1, \ldots, N-1$ be finite-dimensional subspaces. The Ritz approximation of this problem is to find $u_k \in H_k$ such that

$$a(u_k, v_k) = f(v_k) \quad \forall v_k \in H_k.$$

The Mathematica code is mainly referred to [8], which can be found in the appendix. The equidistant mesh will be used here. The code for this problem includes the following steps:

i) Set up the mesh and the basis functions;

ii) Compute the stiffness matrices A and load vectors $\underline{f}$ for different sets of basis functions;

iii) Define the right-hand side of the equation, then solve the problem analytically and numerically;

iv) Plot the solutions and generate a log plot of the errors respectively.

The solutions and logarithmic errors are shown on the next two pages, see Figure 5.4 and Figure 5.5. For better comparison, the results are presented in the finite element space with the same dimension respectively. uH is the numerical solution by using hat functions, uQ is the numerical solution by using piecewise quadratic basis functions, uB is the numerical solution by using Bernstein basis polynomials, and $uExact$ is the analytical solution of this problem. Moreover, $errH$, $errQ$ or $errB$ are the differences between uH, uQ or uB and $uExact$. □

Remark 5.19. *Comparison between hat functions, piecewise quadratic basis functions and Bernstein basis polynomials.* Figure 5.5 shows that, for the same finite dimension, the Ritz method by using Bernstein basis polynomials performs generally better than by using hat functions and piecewise quadratic basis functions. As the finite dimension increases, the errors by using the Bernstein basis polynomials gets smaller much faster than for the other two sets of basis functions.

However, the stiffness matrices for the Bernstein basis polynomials are not sparse, whereas those for the hat functions and piecewise quadratic basis functions are sparse. On the one hand, the sparse matrices are much cheaper to store. On the other hand, sometimes the integration by using the Bernstein basis polynomials is complicated and the resulting stiffness matrices are not sparse which makes the computation much more costly. □

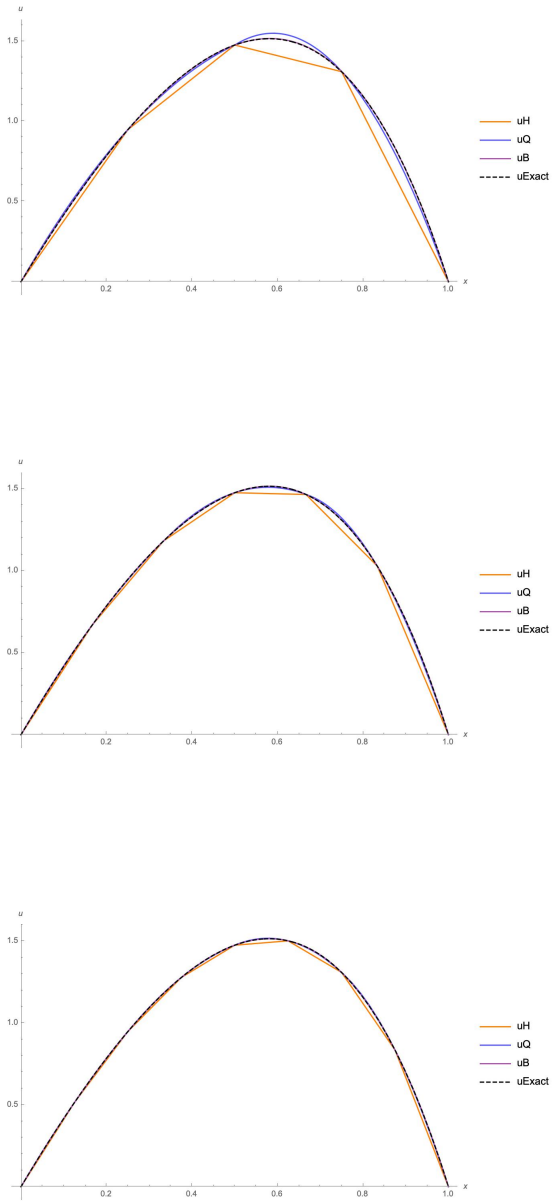

Figure 5.4: From top to bottom: the solutions in the three-dimensional, five-dimensional and seven-dimensional finite element space.

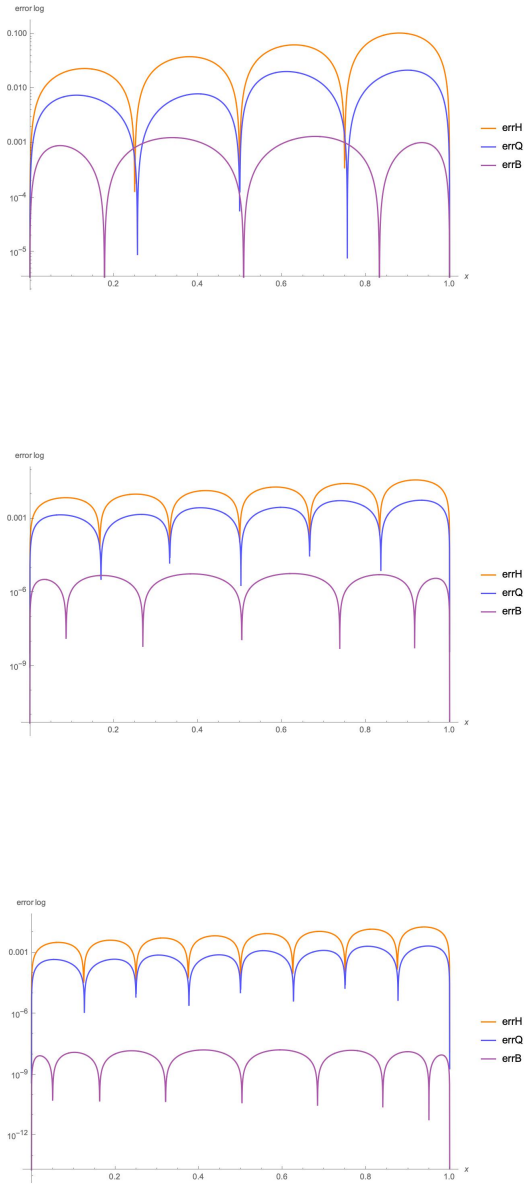

Figure 5.5: From top to bottom: the logarithmic errors in the three-dimensional, five-dimensional and seven-dimensional finite element space.

Chapter 6

Outlook

Remark 6.1. *hp−finite element method.* If the numerical solution is not accurate enough, then we can refine it by the $hp-$finite element method (hp-FEM). The $hp-$FEM is more general than the finite element method. $h-$refinements means dividing elements into smaller ones and $p-$refinements means increasing the polynomial degree. Since the pioneering works on the $hp-$finite element method by Gui, Babuška[1] in the 1980s, where it was shown that for one-dimensional problems $hp-$refinement leads to exponential convergence with respect to the number of degrees of freedom on a priori adapted meshes, there has been a great amount of work devoted to developing adaptive $hp-$refinement strategies based on a posteriori errors estimates, e.g., see [2]. □

Remark 6.2. *Other applications of the Bernstein polynomials.* The Bernstein polynomials have been implemented for solving differential, integro-differential and fractional differential equations. The Bernstein polynomials are also used for important applications in many branches of mathematics and the other sciences, for instance, approximation theory, probability theory, statistic theory, number theory, numerical analysis, constructing Bézier curves, q-calculus, operator theory and applications in computer graphics, see [12]. □

Remark 6.3. *Finite element method with local Bernstein basis polynomials.* One can apply the approach from Section 5.2.3 for each mesh cell $[x_i, x_{i+1}]$, $i = 1, \ldots, N-2$. For the first mesh cell $[x_0, x_1]$ one can drop b_0^n and then use the remaining n Bernstein basis polynomials $b_1^n, \ldots, b_n^n$ as local basis functions. Similarly, for the last mesh cell $[x_{N-1}, x_N]$, one can drop b_n^n and then use remaining n Bernstein basis polynomials $b_0^n, \ldots, b_{n-1}^n$ as local basis functions. For other mesh cells, one uses all the $n+1$ Bernstein basis polynomials of degree n. □

[1] Ivo M. Babuška (1926 -)

Appendix A

Mathematica Code of the Ritz Method

```
In[1]:= xNode[N_, k_] := k / N (* Equidistant Mesh *)

In[2]:= ϕ[N_, k_, x_] :=
         Piecewise[{{(x - xNode[N, k - 1]) / (1 / N), xNode[N, k - 1] ≤ x < xNode[N, k]},
           {(xNode[N, k + 1] - x) / (1 / N), xNode[N, k] ≤ x < xNode[N, k + 1]}}]
        (* Hat functions *)

In[3]:= ψ1[N_, k_, x_] :=
         Piecewise[
          {{2 × (x - xNode[N, k - 1]) × (x - 1/2 (xNode[N, k - 1] + xNode[N, k])) / (1 / N)^2,
            xNode[N, k - 1] ≤ x < xNode[N, k]},
           {2 × (xNode[N, k + 1] - x) × (1/2 (xNode[N, k + 1] + xNode[N, k]) - x) / (1 / N)^2,
            xNode[N, k] ≤ x < xNode[N, k + 1]}}]
        (* Piecewise quadratic basis functions - Part 1 *)

In[4]:= ψ2[N_, k_, x_] :=
         Piecewise[
          {{4 × (x - xNode[N, k - 1]) × (xNode[N, k] - x) / (1 / N)^2,
            xNode[N, k - 1] ≤ x < xNode[N, k]}}]
        (* Piecewise quadratic basis functions - Part 2 *)

In[5]:= ψ[N_, k_, x_] := Piecewise[{{ψ2[N, (k + 1) / 2, x], OddQ[k]}, {ψ1[N, k / 2, x], EvenQ[k]}}]
        (* Piecewise quadratic basis functions *)

In[6]:= φ[N_, k_, x_] := Binomial[N, k] × x^k × (1 - x)^(N - k)
        (* Bernstein basis polynomials *)

In[7]:= stiffnessMatrixH[N_] := N × ToeplitzMatrix[PadRight[{2, -1}, (N - 1)]]
        (* Stiffness matrix for hat functions *)
```

```
In[8]:= stiffnessMatrixQ[N_] := Table[Integrate[D[ψ[N, i, x], x] × D[ψ[N, j, x], x], {x, 0, 1}],
          {i, 1, 2 × N - 1}, {j, 1, 2 × N - 1}]
        (* Stiffness matrix for piecewise quadratic basis functions *)

In[9]:= stiffnessMatrixB[N_] := Table[Integrate[D[φ[N, j, x], x] × D[φ[N, i, x], x], {x, 0, 1}],
          {i, 1, N - 1}, {j, 1, N - 1}]
        (* Stiffness matrix for Bernstein basis polynomials *)

In[10]:= loadVectorH[N_] := Table[Integrate[ϕ[N, j, x] × f[x], {x, 0, 1}], {j, 1, N - 1}]
         (* Load vectors for hat functions *)

In[11]:= loadVectorQ[N_] := Table[Integrate[ψ[N, j, x] × f[x], {x, 0, 1}], {j, 1, 2 × N - 1}]
         (* Load vectors for piecewise quadratic basis functions *)

In[12]:= loadVectorB[N_] := Table[Integrate[φ[N, j, x] × f[x], {x, 0, 1}], {j, 1, N - 1}]
         (* Load vectors for Bernstein basis polynomials *)

In[13]:= f[x_] = 4 × Exp[2 x] (* Tfemhe right-hand side of the equation *)
Out[13]= 4 e^(2 x)

In[14]:= uExact[x_] = u[x] /. DSolve[{-u''[x] == f[x], u[0] == 0, u[1] == 0}, u[x], x][[1]]
         (* Find the exact solution of the problem *)
Out[14]= 1 - e^(2 x) - x + e^2 x

In[15]:= uCoeffH[N_] := Block[{A = stiffnessMatrixH[N], fϕ = loadVectorH[N]}, LinearSolve[A, fϕ]]
         (* Determine the coefficients by using hat functions as a set of basis
          functions *)

In[16]:= uCoeffQ[N_] := Block[{A = stiffnessMatrixQ[N], fψ = loadVectorQ[N]}, LinearSolve[A, fψ]]
         (* Determine the coefficients by using piecewise quadratic basis functions
          as a set of basis functions *)

In[17]:= uCoeffB[N_] := Block[{A = stiffnessMatrixB[N], fφ = loadVectorB[N]},
           LinearSolve[A, fφ]]
         (* Determine the coefficients by using Bernstein basis polynomials as a
          set of basis functions *)

In[18]:= uFemH[N_, x_] := Block[{u = uCoeffH[N]}, Sum[u[[i]] × ϕ[N, i, x], {i, 1, N - 1}]]
         (* Find the numerical solution by using hat functions as a set of basis
          functions *)

In[19]:= uFemQ[N_, x_] := Block[{u = uCoeffQ[N]}, Sum[u[[i]] × ψ[N, i, x], {i, 1, 2 × N - 1}]]
         (* Find the numerical solution by using piecewise quadratic basis functions
          as a set of basis functions *)

In[20]:= uFemB[N_, x_] := Block[{u = uCoeffB[N]}, Sum[u[[j]] × φ[N, j, x], {j, 1, N - 1}]]
         (* Find the numerical solution by using Bernstein basis polynomials as a
          set of basis functions *)

In[21]:= sol[N_] := Block[{}, uH[x_] = uFemH[N, x]; uQ[x_] = uFemQ[N / 2, x];
           uB[x_] = uFemB[N, x];
           Plot[{uH[x], uQ[x], uB[x], uExact[x]}, {x, 0, 1},
            PlotStyle → {Orange, Lighter[Blue], Lighter[Purple], {Dashed, Black}},
            PlotLegends → {uH, uQ, uB, uExact}, AxesLabel → {x, u}]]
         (* Plot the solutions *)

In[22]:= logerr[N_] := Block[{}, uH[x_] = uFemH[N, x]; uQ[x_] = uFemQ[N / 2, x];
           uB[x_] = uFemB[N, x]; errH[x_] = uH[x] - uExact[x]; errQ[x_] = uQ[x] - uExact[x];
           errB[x_] = uB[x] - uExact[x];
           LogPlot[{1/Sqrt[N - 1] × Norm[errH[x]], 1/Sqrt[N - 1] × Norm[errQ[x]], 1/Sqrt[N - 1] × Norm[errB[x]]},
            {x, 0, 1}, PlotLegends → {errH, errQ, errB},
            PlotStyle → {Orange, Lighter[Blue], Lighter[Purple]}, AxesLabel → {x, log error}]]
         (* Generate a log plot of the errors *)
```

References

[1] Sergei N. Bernstein. "Démonstration du théoréme de Weierstrass fondée sur le calcul des probabilités". In: *Communications de la Société Mathématique de Kharkov* (1911). English translation.

[2] P. Daniel and A. Ern. "An adaptive hp-refinement strategy with computable guaranteed bound on the error reduction factor". In: *Computers Mathematics with Applications, Elsevier* (2018), pp. 967–983.

[3] Johannes Erath. *Bernstein-Polynome, Schriftliche Ausarbeitung zum Vortrag*. Ruprecht-Karls-Universität Heidelberg, 2009.

[4] Prof. Dr. Volker John. *Numerik Partieller Differentialgleichungen - Eine elementare Einführung*. Universität des Saarlandes, 2009.

[5] Prof. Dr. Péter Koltai. *Mitschrift zur Analysis III*. Freie Universität Berlin, 2019.

[6] Samuel Kotz and Norman L. Johnson. *Leading personalities in statistical sciences: from the 17th century to the present*. Section 4: Probability Theory. Wiley-Blackwell, 2011.

[7] C. Lohmann and D. Kurmin. "Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements". In: *Journal of Computational Physics* (2017), pp. 151–186.

[8] Dr. Katharine Long. *Finite element solution of a 1D BVP*. Principles of Classical Applied Analysis. Texas Tech University, 2010.

[9] Malte Milatz. *Vortrag zum Proseminar zur Analysis*. RWTH Aachen, 2010.

[10] Prof. Dr. H. Reich. *Skript zur Analysis I*. Freie Universität Berlin, 2018.

[11] Edmund Robertson and John O'Connor. *MacTutor History of Mathematics Archive*. URL: `https://mathshistory.st-andrews.ac.uk/Biographies/Bernstein_Sergi`. (accessed: 10.06.2021).

[12] Yilmaz Simsek. "On Interpolation Function of the Bernstein Polynomials". In: *Digital Proceedings, International Conference on the Applications of Computer Science and Mathematics in Architecture and Civil Engineering* (2012). DOI: `https://doi.org/10.25643/bauhaus-universitaet.2786`.

[13] Prof. Dr. T. J. Sullivan. *A brief introduction to weak formulations of PDEs and the finite element method*. Mathematics Institute and School of Engineering, University of Warwick, 2020.

[14] Prof. Dr. Volker Ulbricht. *Numerische Methoden (FEM, REM)*. Technische Universität Dresden, 2012.

[15] Prof. Dr. Timo Weidl. *Skript zur Vorlesung Funktionalanalysis*. Chapter 1: Sobolev Spaces. Universität Stuttgart, 2004.

[16] Prof. Dr. Dirk Werner. *Funktionalanalysis*. Freie Universität Berlin, 1997.

[17] Wikipedia. *Maximumsnorm*. URL: `https://de.wikipedia.org/wiki/Maximumsnorm`. (accessed: 13.06.2021).