the PnPm Schemes for the one dimensional Hyperbolic Conservation Laws

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

von **Abdulatif Badenjki**
geb. am 01. April 1979 in Aleppo, Syrien

genehmigt durch die Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg

Gutachter:
Prof. Dr. Gerald Warnecke

Eingereicht am: August 9, 2018
Verteidigung am: June 18, 2018

Acknowledgements

In the Name of **ALLAH**, the Merciful, the Compassionate. Praise be to Allah, Lord of the worlds. I praise Him for His favors and ask Him to increase His grace.

((and say " My Lord! Increase me in knowledge ")) Quran [20/114]

My master **Muhammad**, the blessings and peace of **ALLAH** be upon him, upon the rest of the prophets and messengers, and upon all their families, their companions, and the rest of godly persons.

I am very grateful to my supervisor **Prof. Dr. Gerald Warnecke** for providing me an interesting subject of the research and for his remarkable suggestions. His useful guidance and scientific advices inspired me to develop my work throughout my thesis.

I deeply thanks my country Syria for the financial support. I pray that God saves the blood of the Syrian peoples and to have mercy on the martyrs.

I also thanks the German government for the financial support for the last years of my work.

I greet all members of the group of the Prof. Warnecke.

I would like to dedicate this thesis to the soul of my father. He was very interest of my study.

I would like to express my deep obligation to my mother, my brothers, my sisters. I am sure that they had prayed in order I to be able to complete my Ph.D. studies.

Finally, I am heartily thankful to my wife **Rahaf Badenjki**, to my sons. Their love, care, support and advice have encouraged me to work hard for this degree.

Abstract

The aim of this dissertation is to present the details of the $P_N P_M$ DG schemes in one space dimension for $N, M \in \mathbb{N}_0$ with $M \geqslant N$, to study their numerical properties, to apply them to scalar equations and systems of the hyperbolic conservation laws, and to make some basic comparisons between numerical schemes from this large class of schemes. The $P_N P_M$ DG schemes were originally introduced by Dumbser et al. [7].

The $P_N P_M$ DG schemes use reconstruction operators applied to the discontinuous Galerkin (DG) scheme. First, the given data are projected on elements of the numerical space domain, which is discretized by an appropriate partition. This projection gives in each element at the starting time an approximation which is written in each element as a sum of piecewise polynomials of a maximal degree N. We define this projection procedure and associate with it an appropriate basis and an appropriate space of piecewise polynomials. With these algebraic elements (basis and space), we prove the existence and the uniqueness of the approximation. Then we prove some of its properties, where we will consider this procedure as an operator.

After that, we prove estimates of this operator using the L^1 and L^2 norms. This proof is done using the Bramble-Hilbert lemma. A smooth order $N + 1$ is proven for the projection operator which gives polynomials of the degree N. As supplement, we give some examples of the projection operators using various continuous and discontinuous functions.

We give an extra work for the extension of the projection operator to the 2D case, with some properties and estimates.

The first step of the $P_N P_M$ DG schemes is to produce new piecewise polynomials of degree M. They are reconstructed in each element from these approximate polynomials of the degree N. We present this reconstruction and define the stencils which have a main role in this step. We prove the existence of the solution using these operators with conditions on choosing the size of the stencil with respect to the orders N and M. This proof is a first general proof of this fact which previously had been obtained for special cases, $M = 2N + 1$ only, see [14]. We consider two cases of the solution, either a unique exact solution or a unique solution obtained by using the least squares approach in the overdetermined case.

Then we prove some of the properties of the reconstruction operator, especially the identity property. We also prove that there are some relations between the two previous operators. Finally, we prove estimates of this operator using the L^2 norm. A smooth order $M+1$ is proven for the reconstruction operator which gives polynomials of degree M. We again demonstrate, as supplement, some examples of the reconstruction operators with several types and sizes of the stencils using various functions.

The second step of the $P_N P_M$ DG schemes is a time evolution that gives other polynomials of degree M in time and space. This improvement of the data depends on applying the local space time Galerkin scheme. We demonstrate the details of this step and explain how to insert the reconstructed polynomials inside the solutions of this step. After that, we provide some simple examples for our work in this step.

The third step of $P_N P_M$ DG schemes is to apply the DG scheme to the conservation laws. The numerical flux, taken to solve the Riemann problem at the interfaces of the elements, has to be computed by using the solutions of the local Galerkin step.

We consider the linear advection equation as a standard equation of the 1D hyperbolic equations. We show the general formula of the $P_N P_M$ DG schemes for the cases $a > 0$ and $a < 0$, discuss the boundary conditions, and view some special formulas for some choices of the orders N and M. Then we study the linear stability, obtaining tables of the maximal limits of the stability for the schemes for all orders till $M = 5$. We also study the efficiency of the schemes measuring the cost in the computational time and mesh discretization. We also study the influence of the size and the type of the stencils.

We also apply the $P_N P_M$ DG schemes to the Burgers equation and associate to them the Lax-Friedrichs and Godunov fluxes. We study two different cases. First, we apply the schemes for the Riemann problem and discuss the effect of the use of the slope limiter. We note that the Godunov flux gives better results than the Lax-Friedrichs flux. Second, we apply the schemes for smooth function and compare the solutions using the Lax-Friedrichs and the Godunov fluxes at different times using the TVDM and TVBM limiters.

In the last Chapter, we apply the schemes to the system of the shallow water equations.

Zusammenfassung

Das Ziel dieser Arbeit ist es, die Details der $P_N P_M$ DG Schemata für $N, M \in \mathbb{N}_0$ mit $M \geqslant N$ in einer Raumdimension zu präsentieren, ihre numerischen Eigenschaften zu studieren, sie auf skalare und Systeme von hyperbolischen Erhaltungssätzen anzuwenden, und einige grundlegende Vergleiche zwischen numerischen Schemata aus dieser großen Klasse von Schemata zu machen. Die $P_N P_M$ DG Schemata wurden ursprünglich von Dumbser et al. [7] eingeführt.

Die $P_N P_M$ DG Schemata wenden Rekonstruktionsoperatoren auf die diskontinuierliche Galerkin Schema an. Die angegebene Daten werden zunächst auf Elemente des numerischen Intervalls, welches durch eine entsprechende Aufteilung diskretisiert wird, projiziert. Diese Projektion liefert zur Startzeit in jedem Element eine Approximation, welche in jedem Element als eine Summe von stückweisen Polynomen vom maximalen Grad N geschrieben wird. Wir definieren dieses Projektionsverfahren und ordnen ihm eine geeignete Basis und einen geeigneten Raum stückweiser Polynome zu. Mit diesen algebraischen Elementen (bestehend aus Basis und Raum) beweisen wir die Existenz und die Eindeutigkeit der Approximation. Darüber hinaus beweisen wir einige seiner Eigenschaften, indem wir dieses Verfahren als Operator betrachten.

Im Anschluss beweisen wir einige Abschätzungen dieses Operators mit Hilfe der L^1- und L^2-Normen. Dieser Beweis wird mit dem Bramble-Hilbert-Lemma durchgeführt. Eine glatte Ordnung $N + 1$ wird für den Projektionsoperator, welcher Polynome vom Grad N liefert, bewiesen. Als Ergänzung geben wir einige Beispiele der Projektionsoperatoren an, bei denen sowohl kontinuierliche als auch diskontinuierliche Funktionen verwendet werden.

Der Projektionsoperator wird auf den 2D-Fall erweitert und wir geben einige Eigenschaften und Abschätzungen an.

Im ersten Schritt der $P_N P_M$ DG Schemata werden neue stückweise Polynome vom Grad M erzeugt. In jedem Element werden diese aus den annähernden Polynomen eines Grades N rekonstruiert. Wir präsentieren diese Rekonstruktion und definieren die Abhängigkeitsgebiete, die in diesem Schritt eine Hauptrolle spielen. Wir beweisen die Existenz der Lösung unter Verwendung dieser Operatoren mit Bedingungen zur Auswahl der Größe des Gebietes in Bezug auf die Ordnungen N und M. Dieser Beweis ist der erste allgemeine Beweis dieser Tatsache, welche bisher für Sonderfälle, nur $M = 2N + 1$ erhalten wurde, siehe [14]. Wir betrachten

zwei Fälle der Lösung, entweder eine eindeutige Lösung oder eine, die unter Verwendung des least-squares Ansatzes erhalten wird.

Dann beweisen wir einige Eigenschaften des Rekonstruktionsoperators, insbesondere die Identitätseigenschaft. Wir beweisen auch, dass zwischen den beiden vorherigen Operatoren einige Beziehungen bestehen. Schließlich beweisen wir Abschätzungen dieses Operators mit der L^2-Norm. Eine glatte Ordnung $M + 1$ wird für den Rekonstruktionsoperator bewiesen, welcher Polynome vom Grad M liefert. Als Ergänzung zeigen wir einige Beispiele der Rekonstruktionsoperatoren mit verschiedenen Arten und Größen der Abhängigkeitsgebieten unter Verwendung verschiedener Funktionen.

Der zweite Schritt der $P_N P_M$ DG Schemata ist eine Zeitentwicklung, welche andere Polynome vom Grad M in Zeit und Raum ergibt. Diese Verbesserung der Daten hängt von der Anwendung des lokalen Raum-Zeit Galerkin Schemas ab. Wir demonstrieren die Details dieses Schritts und erklären, wie die rekonstruierten Polynome in die Lösungen dieses Schrittes eingefügt werden. Danach geben wir einige einfache Beispiele für unsere Arbeit in diesem Schritt.

Im dritten Schritt eines $P_N P_M$ DG Schemas wird das DG-Schema auf die Erhaltungssätze angewendet. Der numerische Fluss, der zur Lösung des Riemann-Problems an den Grenzflächen der Elemente benötigt wird, wird unter Verwendung der Lösungen des lokalen Galerkin-Schritts berechnet.

Wir betrachten die lineare Advektionsgleichung als eine Standard Gleichung der 1D hyperbolischen Gleichungen. Wir zeigen die allgemeine Formel der $P_N P_M$ DG Schemata für die Fälle $a > 0$ und $a < 0$, diskutieren die Rand-Bedingungen, und zeigen einige spezielle Formeln für einige Auswahlmöglichkeiten der Ordnungen N und M. Im Anschluss studieren wir die lineare Stabilität und erhalten Tabellen der maximalen Stabilitätsgrenzen für die Schemata für alle Ordnungen bis $M = 5$. Wir studieren auch die Effizienz der Schemata durch Bestimmung der Kosten für die Rechenzeit und Gitterdiskretisierung. Wir studieren auch den Einfluss der Größe und der Art der Gebiete.

Wir wenden die $P_N P_M$ DG Schemata auch auf die Burgers Gleichung an unter Verwendung des Lax-Friedrichs oder Godunov Flusses. Wir studieren zwei verschiedene Fälle. Zuerst wenden wir die Schemata auf das Riemann-Problem an und diskutieren den Effekt der Verwendung des Slope-Limiters. Es stellt sich heraus, dass der Godunov-Fluss bessere Ergebnisse liefert als der Lax-Friedrichs-Fluss. Zweitens wenden wir die Schemata auf eine glatte Funktion an und vergleichen die Lösungen mit TVDM- und TVBM-Limitern unter Verwendung der Lax-Friedrichs und der Godunov Flüsse zu verschiedenen Zeiten.

Im letzten Kapitel wenden wir die Schemata auf das System der Shallow-Water Gleichungen an.

Contents

1 Introduction **10**
- 1.1 Overview of the $P_N P_M$ DG Schemes . . . 10
- 1.2 The Main Results . . . 11
- 1.3 Outline of the Structure . . . 12

2 The Projection onto Piecewise Polynomials **14**
- 2.1 Mathematical Preliminaries . . . 14
- 2.2 The Space of the Piecewise Polynomials . . . 16
- 2.3 Approximating Using Piecewise Polynomials . . . 18
- 2.4 Analytical Study . . . 20

3 Examples of the Projections **27**
- 3.1 Preface . . . 27
- 3.2 Polynomial of Degree One . . . 28
- 3.3 Polynomial of Degree Two . . . 29
- 3.4 Polynomial of Degree Three . . . 30
- 3.5 Polynomial of Degree Four . . . 30
- 3.6 A Trigonometric Function . . . 31
- 3.7 Summary . . . 31

4 The Projection onto Piecewise Polynomials: 2D Case **40**
- 4.1 2D Discretization . . . 40
- 4.2 2D Basis Functions . . . 41
- 4.3 The Projection . . . 43

5 The Reconstruction of Higher Order Polynomials **45**
- 5.1 The Idea of the Reconstruction . . . 45
- 5.2 Approximations by the Projection Polynomials . . . 46
- 5.3 Computing the Coefficients . . . 47
- 5.4 Motivating Examples . . . 52
- 5.5 Analytical Study . . . 55

6 Examples of the Reconstruction **61**
6.1 The Function $v(x) = x - 1$ . . . 61
6.2 The Function $v(x) = x^2 - 3x + 2$ . . . 61
6.3 The Function $v(x) = x^3 - x$ . . . 63
6.4 The Function $v(x) = \sin(x)$ . . . 64

7 The Local Space Time Galerkin Scheme **68**
7.1 The Local Space Time Basis Functions . . . 68
7.2 The Formulas of the Solutions . . . 71
7.3 The Matrix Form . . . 72
7.4 Inserting the Reconstructed Polynomials . . . 74
7.5 Reducing the Algebraic System . . . 75
7.6 Iterating the Reduced Algebraic System . . . 76
7.7 Example 1: the Linear Advection Equation . . . 77
7.8 Example 2: Nonlinear Burgers Equation . . . 80

8 The Discontinuous Galerkin Schemes **81**
8.1 Preface . . . 81
8.2 The $P_N P_M$ DG Schemes . . . 81

9 Linear Advection Equation and Numerical Studies **86**
9.1 The $P_N P_M$ Schemes . . . 86
9.2 Numerical studies . . . 90

10 The Burgers Equation **102**
10.1 The $P_N P_M$ Schemes . . . 102
10.2 The Slope Limiter . . . 103
10.3 The Riemann Problems . . . 104
10.4 The Burgers Equation with Smooth Initial Data . . . 106

11 The Shallow Water Equations **114**
11.1 The Mathematical Model . . . 114
11.2 Numerical Test . . . 115

Conclusion **116**

A 2D Hierarchical Orthogonal Basis on Rectangles **117**

Bibliography **118**

Ehrenerklärung **120**

Chapter 1

Introduction

1.1 Overview of the $P_N P_M$ DG Schemes

A conservation law [12] is a system of hyperbolic PDEs that states that the rate of change of a physical state or conserved quantity is governed by a flux function. We are interested in solving 1D hyperbolic systems of conservation laws

$$\mathbf{v}_t(t,x) + \mathbf{f}(\mathbf{v}(t,x))_x = 0,$$

where the dependent variable $\mathbf{v} = \mathbf{v}(t,x)$ is the vector of conserved quantities, the continuously differentiable function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$ is the flux function, $x \in I = [a,b] \subset \mathbb{R}$ is the space variable, and $t \geqslant 0$ is the time variable. The hyperbolicity means that the Jacobian matrix of $\mathbf{f}(\mathbf{v})$ with respect to $\mathbf{v}$ has real eigenvalues associated with a set of linearly independent eigenvectors which form a basis of $\mathbb{R}^d$, where $d \in \mathbb{N}$ is the dimension of the vector $\mathbf{v}$.

The $P_N P_M$ DG schemes, originally developed by Dumbser et al. [7], are able, in general, to treat these systems even with source term $\mathbf{s}(\mathbf{v})$ of the form $\mathbf{v}_t + \mathbf{f}(\mathbf{v})_x = \mathbf{s}(\mathbf{v})$.

The data are approximated at each time step t_n on the space interval I by a piecewise polynomial u^n of degree $N \in \mathbb{N}_0$ using some time independent piecewise basis polynomials $\Phi_{i,j}$ of degree $i \leqslant N$ and with time dependent coefficients $\widehat{u}^n_{i,j}(t)$. This approximation is written as

$$u^n(t,x) = \sum_{j=1}^{Z} \sum_{i=0}^{N} \widehat{u}^n_{i,j}(t) \Phi_{i,j}(x),$$

where Z is the number of cells in the discretization.

The $P_N P_M$ DG schemes start with using a linear reconstruction operator applied at the beginning of each time step of the discontinuous Galerkin schemes (DG). This operator is applied to the numerical data u^n. It increases the order of accuracy in space obtaining high order piecewise polynomials of an arbitrary degree $M \geqslant N$. These polynomials are linear combinations of the basis functions $\Phi_{i,j}$ of degree M with time dependent coefficients $\widehat{w}^n_{i,j}(t)$. In the special case when $M = N$ the reconstruction operator reduces to the identity. Dumbser and Munz [8] were the first to propose the application of the reconstruction operators to the DG schemes at the beginning of each time step.

The $P_N P_M$ DG schemes use also a local continuous space time Galerkin method. This second step evolves the numerical solution in time inside each element and provides a high order time discretization of the same order of accuracy as the space discretization. This gives, as a high order accurate predictor, space time polynomials for the functions $\mathbf{v}$, $\mathbf{f}$ and $\mathbf{s}$. Dumbser et al. [7] proposed to use this approach to obtain smaller algebraic systems and solved these systems efficiently by a simple iteration scheme. They also explained that at least for linear hyperbolic PDE this iteration scheme has a unique solution and the convergence to this solution is guaranteed, according to the Banach fixed point theorem. For linear homogeneous scalar equations the method always converges for any initial guess vector after at most M iterations. But for nonlinear systems only about M or $M+1$ iterations are taken as an approximation.

The third step is to apply the DG schemes. The result of the iteration method is used for the time integration of the flux and source evolutions. The ADER methods, for Arbitrary Accuracy DERivative Riemann problem, of Toro et al. [24], solve a high order Riemann problem approximately at the interface. They need the Gaussian quadrature in space and time in order to compute the fluxes across the interface. Dumbser et al. [9, 11] proposed a quadrature free version of the scheme of arbitrary accuracy in space and time on unstructured meshes in two and three space dimensions. This version of the ADER schemes is similar to the original ENO scheme proposed by Harten et al. [15], since it first evolves the data for each element via the Cauchy-Kovalewski procedure and then solves the interactions across the boundary. In the $P_N P_M$ DG schemes, to obtain a quadrature free version of the schemes, the space time informations has been used, which was neither done in the ENO scheme nor in previous ADER schemes.

The resulting $P_N P_M$ DG schemes are one step schemes, i.e. only two time levels are involved in one time step. They are quadrature free, fully discrete, and can be chosen of arbitrary order of accuracy in space and time. These schemes contain both the finite volume schemes when $N = 0$, and the discontinuous Galerkin schemes when $M = N$.

1.2 The Main Results

In Chapter 2, we represent how we define a piecewise approximation u using the operator $\Pi_{N,Z}$. We prove that this operator is stable, linear, conservative and as a best approximation. As well as, it is an orthogonal projection. We give estimates for this operator in the L^1 and L^2 norms for the smooth functions and for the discontinuous ones.

The idea of the projection is extended to the 2D case. We prove in an analogous way similar properties in Chapter 4.

The main part of the $P_N P_M$ DG schemes is to use the reconstruction operators, as we will show in Chapter 5. It is of interest to increase the order of the solution in space arbitrarily. The reconstruction operator uses a domain around an arbitrary element. This domain is called the reconstruction stencil. The reconstruction operators are given by equalities of the projections on the elements of the stencil.

The idea of using the reconstruction operators is inspired from the work of Dumbser et al. [7]. They applied these operators to conservation laws in two and three space dimensions. Our

work on the reconstruction operators is only to the one space dimension and this work has driven us to prove properties more precisely. A central result is a proof of the unique solvability of the reconstruction step. Our choices of the stencil sizes for the reconstruction operators always generate systems of equations with full column rank. Furthermore, the reconstruction operators give approximations of the data considered, but as a special case they recover the same data when these originally are polynomials of the same degree.

We prove that the reconstruction operator gives unique solutions and it is stable, linear, conservative, and consistent. We give some theorems to show the relation between the projection and the reconstruction operators. Moreover, we prove estimates for the reconstruction operator meaning that the reconstructed polynomials of degree M are accurate of order $M+1$ in the L^2 norm, provided that the stencil gives equations, whose number is greater than or equal to the number of the coefficients of the reconstructed polynomials.

The $P_N P_M$ DG schemes develop the data in time once in each time step. Furthermore, with the nonlinear equations we use a slope limiter. We apply this limiter once in each time step. On the other hand, with the RKDG schemes one needs to develop the data in time and to use the slope limiter several times in each time step, according to the Runge-Kutta stages. Thus, precisely due to this point, for high order schemes the $P_N P_M$ are faster than the RKDG schemes.

Courant numbers are important for the stability of explicit schemes for conservation laws. We computationally explore maximal limits of these numbers for the $P_N P_M$ DG schemes by applying the von-Neumann analysis and using an experimental procedure. We obtain a wide variety of stability limits, including some unstable cases for which we have only one value $\lambda = 1$ that gives a stable solution. Moreover, there are some semi-stable cases with a minimal bound on the time step and some cases with a larger stability interval that $]0, 1]$. This study of the stability is for the application of the $P_N P_M$ DG schemes to the linear advection equation.

1.3 Outline of the Structure

Now we give an overview of our thesis. The definition and properties of the projection procedure are presented in Chapter 2. This projection produces piecewise representations of the data. These representations are polynomials of degree $N \geqslant 0$. We prove the existence and the uniqueness of the approximation and prove some properties of the projection operator. At the end we give estimates of this operator using the L^1 and L^2 norms. A smooth order $N+1$ is proven for the projection operator which gives polynomials of the degree N.

Some examples of the projection operator using various functions are given in Chapter 3 with figures and tables of errors. The projection always has the order $N+1$ of accuracy using the L^2 norm.

The Chapter 4 includes an extension of the projection operator to the 2D case, with some properties and estimates.

The main part of the work is in Chapter 5 where we introduce the idea of the reconstruction operators. We define the stencils and their types and sizes. We prove the existence and uniqueness of the solution using these operators. This is a first general proof of this fact which

is previously had been obtained for special cases, $M = 2N + 1$ only, see [14]. Some examples of the reconstruction operators with several types and sizes of the reconstruction stencils view the formulas of these operators. We consider two cases of the solution, either unique exact solutions or unique solutions by using the least square approach in the overdetermined case. Furthermore, some theorems review the properties of these operators and review the relation between the two operators. Finally, we give estimates using the L^2 norm. A smooth order $M + 1$ is proven for the reconstruction operator which gives polynomials of degree M.

Some examples of the reconstruction operator using various functions are given in Chapter 6 with figures and tables of errors. The reconstruction always has the order $M + 1$ of accuracy using the L^2 norm.

In Chapter 7 we treat the idea of improvement the data in time. This step increases the degree of the time variable for the data by applying the space time continuous Galerkin method to the conservation law considered. We give examples of building the nodal bases, explain how to insert the reconstructed polynomials and how to reduce the linear algebraic system, and show the iterative method to solve the reduced linear algebraic system. After that, we view some formulas of the solutions for the advection equation and for the Burgers equation.

In Chapter 8 the DG schemes are applied to the conservation laws. The numerical flux, taken to solve the Riemann problem at the interfaces of the elements, will be applied to the space time solutions of the continuous Galerkin method of Chapter 7.

The Chapter 9 is specified to study the numerical properties of the $P_N P_M$ DG schemes. The linear advection equation is of interest to determine the efficiency and the ability of the numerical schemes to capture the best approximations. We view some examples of the $P_N P_M$ DG schemes applied to the advection equation, study the Fourier stability analysis, give the limits of the Courant numbers associated with these schemes. Also we study the numerical effect of the size and form of these stencils on the efficiency of the $P_N P_M$ DG schemes.

In Chapter 10 we apply the $P_N P_M$ DG schemes to the Burgers equation for the Riemann problem and for smooth functions. We discuss the influence of the slope limiter on the solutions and compare the solutions using the Lax-Friedrichs and the Godunov fluxes at different times and using the TVDM and TVBM limiters.

Finally, in Chapter 11, we apply the $P_N P_M$ DG schemes to the system the shallow water equations.

Chapter 2

The Projection onto Piecewise Polynomials

2.1 Mathematical Preliminaries

2.1.1 Function Classes

Given a closed finite interval $I = [a, b] \subset \mathbb{R}$, we define the space of continuous functions on I, $C(I) := \{v : v \text{ is continuous at each } x \in I\}$, see Adams [1]. This space is a normed linear space with the norm $||v||_{C(I)} = \max_{x \in I} |v(x)|$. We often deal with smoother functions which have derivatives. If r is a positive integer, we introduce the space of r-times continuously differentiable functions as $C^r(I) := \{v : v, D^{(1)}v, \ldots, D^{(r)}v \in C(I)\}$, where $D^{(i)}v$ is the derivative of order i of v. At the boundary points $a, b \in \mathbb{R}$ we assume one-sided continuous differentiability.

We introduce for $p \in [1, \infty]$ the classical Lebesgue spaces

$$L^p(I) := \{v : v \text{ is measurable on } I \text{ and } ||v||_{L^p(I)} < \infty\},$$

with the norms

$$\begin{aligned} ||v||_{L^p(I)} &:= \left(\int_I |v(x)|^p dx \right)^{\frac{1}{p}}, \quad 1 \leqslant p < \infty, \\ ||v||_{L^\infty(I)} &:= \operatorname{ess\,sup}_{x \in I} |v(x)|, \quad p = \infty, \end{aligned}$$

Especially, we use extensively the L^2 norm and the corresponding scalar product

$$\langle f, g \rangle := \int_I f(x) g(x) dx. \tag{2.1}$$

Let $p \in [1, \infty[$ and let r be a non negative integer, the Sobolev space is defined by

$$W^{r,p}(I) := \{v \in L^p(I) : D^{(r)}v \in L^p(I)\}, \tag{2.2}$$

where the derivatives are taken in the sense of distributions. This space is associated with the norm

$$||v||_{W^{r,p}(I)} := \left(\sum_{i=0}^{r} (||D^{(i)}v||_{L^p(I)})^p \right)^{1/p}, \tag{2.3}$$

and with the seminorm

$$|v|_{W^{r,p}(I)} := ||D^{(r)}v||_{L^p(I)}. \tag{2.4}$$

For the case $p = 2$, we have the spaces $W^{r,2}(I)$ with the seminorm $|v|_{W^{r,2}(I)}$. The function $|\cdot|_{W^{r,p}(I)}$ is a seminorm, since we may get $|v|_{W^{r,p}(I)} = 0$ even if $v \neq 0$, e.g. if $v \equiv 1$ and $r \geqslant 1$, therefore it is not a norm.

2.1.2 The Projection Operator

Let X be a normed linear vector space and $V \subset X$ be a linear subspace. Then a bounded idempotent operator $P : X \to V$ with $P = P^2$ is called a *projection* operator. Moreover, if X is a Hilbert space and the image $\mathrm{Im}(P)$ is orthogonal to the kernel $\ker(P)$, then P is called an *orthogonal* projection operator. We are interested only in the case where $V = \mathrm{Im}(P)$ is a finite dimensional subspace.

2.1.3 The Polynomials

Polynomials play a main role in approximation theory and numerical analysis. To indicate why this might be the case, let $N \in \mathbb{N}_0$ and $I \subseteq \mathbb{R}$ be an interval. We call

$$P_{N,I} := \left\{ p : p(x) = \sum_{i=0}^{N} a_i x^i, \quad a_0, \ldots, a_N \in \mathbb{R}, \quad x \in I \right\}, \tag{2.5}$$

the space of polynomials of a maximal degree N with support I. This space is a finite dimensional linear space with the monomials $1, x, \ldots, x^N$ as basis. One may also consider any other convenient basis. The polynomials from $P_{N,I}$ have many attractive features. (1) They are smooth functions. (2) The coefficients $a_0, \ldots, a_N$ can be stored and evaluated on a digital computer. (3) The derivative of a polynomial is again a polynomial. (4) The number of zeros of a polynomial of degree N cannot exceed N. (5) Given any continuous function on an interval $[a, b]$, there exists a polynomial which is uniformly close to it. (6) Precise rates of convergence can be given for the approximation of smooth functions by polynomials. These properties indicate, indeed, that polynomials should be ideal for approximation purposes, however, it has been observed that the polynomials possess one unfortunate feature. Many approximation processes involving polynomials tend to produce polynomial approximations that oscillate wildly. This main drawback of the space $P_{N,I}$ of polynomials is a kind of inflexibility. Polynomials seem to do all right on sufficiently small intervals, but when we go to larger intervals, severe oscillations often appear particularly if the degree $N \geqslant 3, 4$. This observation suggests that in order to achieve a greater flexibility, one should divide up the interval of interest into smaller sub intervals. We are motivated to define another space of polynomials.

2.2 The Space of the Piecewise Polynomials

Let $Z \in \mathbb{N}$, $a < b$ be real, $\Delta := \{x_{j+\frac{1}{2}}\}_0^Z$ be a grid of $Z+1$ equally distant points $a = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \ldots < x_{Z-\frac{1}{2}} < x_{Z+\frac{1}{2}} = b$ which are the called the nodes of the partition or grid. The set Δ partitions the interval $I = [a, b]$ into Z disjoint subintervals which we call elements $I_j := [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[$ for $j = 1, \ldots, Z$. We define the midpoints by $x_j := \frac{1}{2}(x_{j+\frac{1}{2}} + x_{j-\frac{1}{2}})$ and the constant element size $h := x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}} = \frac{b-a}{Z}$.

2.2.1 The Definition

Given a non negative integer $N \in \mathbb{N}_0$. We call

$$P_{N,I,Z} := \{p : \text{for all } \ j = 1, \ldots, Z, \ \ \exists\, p_j \in P_{N,I_j}; \ \ p(x) = p_j(x) \ \text{ for } \ x \in I_j\},$$

the space of piecewise polynomials of degree at most N with respect to the partition Δ, where P_{N,I_j} for $j = 1, \ldots, Z$ are the spaces of polynomials of the degree N given by (2.5). Any polynomial on $[a, b]$ is also a smooth piecewise polynomial with respect to any partition of this interval. This implies that the space $P_{N,I}$ is a subspace of $P_{N,I,Z}$.

By going over from polynomials to piecewise polynomials, we have gained flexibility, but at the same time we have lost an important property. The piecewise polynomial functions are not necessarily smooth and they can even be discontinuous. Each function $p \in P_{N,I,Z}$ consists of Z polynomial terms. This means that the two polynomial terms p_{j-1} and p_j associated with the intervals I_{j-1} and I_j respectively have one common node $x_{j-\frac{1}{2}}$ and are unrelated to each other. Thus there may be a jump discontinuity at $x_{j-\frac{1}{2}}$. Thus, the space $P_{N,I,Z}$ contains functions with possible jump discontinuities at the interior nodes $x_{j\mp\frac{1}{2}}$.

Remark 2.1. In order to maintain the flexibility of piecewise polynomials with achieving some degree of global smoothness, one can enforce smoothness conditions on the polynomial terms and their derivatives at the interior nodes. In other words, one can force the two polynomial terms to tie together smoothly in the sense that a piecewise polynomial $p \in P_{N,I,Z}$ and its first derivatives are all continuous across the node. For more details, one can see [19]. Here we are concerned only with the type of piecewise polynomials without such conditions on the terms.

2.2.2 The Legendre Polynomials

The mutually orthogonal Legendre polynomials of degree $i \in \mathbb{N}_0$ on the reference interval $J = [-1, 1]$ can be determined by the Rodrigues formula $\mathcal{L}_i(s) = \frac{(-1)^i}{2^i i!} \frac{d^i}{ds^i} \{(1 - s^2)^i\}$ for $s \in J$, see e.g. Stegun [21]. On J they satisfy the orthogonality condition

$$\int_{-1}^{1} \mathcal{L}_m(s)\mathcal{L}_n(s)ds = \frac{2}{2n+1}\delta_{mn},$$

where δ_{mn} is the Kronecker delta and satisfy also

$$\mathcal{L}_i(-s) = (-1)^i \mathcal{L}_i(s) \quad \text{ and } \quad \mathcal{L}_i(1) = 1, \tag{2.6}$$

see e.g. Koornwinder et al. [17, Table 18.6.1]. For example, the first four polynomials are

$$\left\{1, s, \frac{3s^2-1}{2}, \frac{5s^3-3s}{2}\right\}.$$

2.2.3 Finding a Basis for the Space $P_{N,I,Z}$

Let I_j for $j = 1, \ldots, Z$ be our discrete intervals with constant length h and midpoints x_j. We define linear reference transformations $\gamma_j : I_j \to J$ by

$$\gamma_j(x) := \frac{2}{h}(x - x_j), \qquad x \in I_j. \tag{2.7}$$

We find that $\gamma_j(x) \in [-1, 1]$ if $x \in I_j$ and 0 elsewhere. We can suppose that γ_1 is a variable for the Legendre polynomial associated with I_1, and γ_2 with I_2, etc. Thus using these transformations we obtain the transformed piecewise Legendre basis functions

$$\Phi_{i,j}(x) = \begin{cases} \mathcal{L}_i(\gamma_j(x)) & x \in I_j \\ 0 & x \in I \backslash I_j \end{cases} \qquad j = 1, \ldots, Z, \quad i = 0, \ldots, N. \tag{2.8}$$

The functions $\Phi_{i,j}$ are orthogonal and defined on I, not only on I_j. For example, in I_j

$$\Phi_{0,j}(x) = 1, \qquad \Phi_{1,j}(x) = \frac{2}{h}(x - x_j), \qquad \Phi_{2,j}(x) = \frac{3}{2}\left(\frac{2}{h}(x - x_j)\right)^2 - \frac{1}{2},$$

and in $I \backslash I_j$ we have $\Phi_{0,j}(x) = \Phi_{1,j}(x) = \Phi_{2,j}(x) = 0$.

We collect the first $N + 1$ such functions in the set $A_{N,j} := \{\Phi_{0,j}, \ldots, \Phi_{N,j}\}$. Let finally $\Omega_{N,j} := \text{span}\{A_{N,j}\}$ be the space spanned by $A_{N,j}$. The restriction $A_{N,j}|_{I_j}$ forms a basis for $\Omega_{N,j}|_{I_j}$, this follows directly from the linearly independence of the Legendre polynomials. Also, the space $\Omega_{N,j}|_{I_j}$ is a subspace of $L^2(I_j)$ and $L^1(I_j)$, therefore we can use the following scalar product

$$\langle v, w \rangle_j := \int_{I_j} v(x)w(x)dx, \qquad \text{for all} \quad v, w \in L^2(I_j), \tag{2.9}$$

and the norms

$$||v||_{L^2(I_j)} := \sqrt{\int_{I_j} |v(x)|^2 dx}, \qquad \text{for all} \quad v \in L^2(I_j),$$

$$||v||_{L^1(I_j)} := \int_{I_j} |v(x)|dx, \qquad \text{for all} \quad v \in L^1(I_j).$$

Each space $\Omega_{N,j}|_{I_j}$ has a basis which consists of $N + 1$ polynomials and is a subspace of P_{N,I_j}, thus $\Omega_{N,j}|_{I_j} = P_{N,I_j}$. We can now take a new definition of the space $P_{N,I,Z}$ as follows

$$P_{N,I,Z} := \{p : \text{for all} \quad j = 1, \ldots, Z, \quad \exists\, p_j \in \Omega_{N,j}|_{I_j}; \quad p(x) = p_j(x) \quad \text{for} \quad x \in I_j\}. \tag{2.10}$$

For all $i = 0, \ldots, N$ and $j = 1, \ldots, Z$, the function $\Phi_{i,j}$ given by (2.8) is a piecewise polynomial defined on I, thus $\Phi_{i,j} \in P_{N,I,Z}$. Then the set $B_{N,Z} := \bigcup_{j=1}^{Z} A_{N,j} = \{\Phi_{i,j}\}_{i=0,\ldots,N}^{j=1,\ldots,Z}$ satisfies

$B_{N,Z}|_{I_j} = A_{N,j}|_{I_j}$, for all $j = 1, \ldots, Z$. Thus, $B_{N,Z}|_{I_j}$ is linearly independent. Since $A_{N,k} \cap A_{N,i} = \{0\}$ when $k \neq i$, one may easily deduce that $B_{N,Z}$ itself is linearly independent. Then $B_{N,Z}$ is a basis of the direct sum of the spaces span$\{A_{N,j}|_{I_j}\}$, which is itself the sum $\sum_{j=1}^{Z} P_{N,I}|_{I_j}$. On the other hand, according to the definition (2.10), we have $P_{N,I,Z} = \sum_{j=1}^{Z} P_{N,I}|_{I_j}$. Then $B_{N,Z}$ is a basis of $P_{N,I,Z}$. Moreover, the basis $B_{N,Z}$ consists of orthogonal functions on I with

$$\int_{I_j} \Phi_{m,j}(x)\Phi_{n,j'}(x)dx = \begin{cases} \frac{h}{2m+1}\delta_{mn}, & j = j', \\ 0, & j \neq j', \end{cases} \qquad m, n = 0, \ldots, N. \tag{2.11}$$

Thus we have proved the following theorem

Theorem 2.2. $P_{N,I,Z}$ has the dimension $Z(N+1)$ and the set $B_{N,Z}$ is an orthogonal basis for it.

2.3 Approximating Using Piecewise Polynomials

In this chapter we will study various functions of different order of smoothness, e.g. starting from the space C^∞ of infinitely continuously differentiable functions, such as e.g. $\sin x$, up to the discontinuous functions, such as a jump function. Therefore, the function, which will be considered, must be at least bounded on the interval I. The Riemann integral of a bounded function on a closed interval I always exists, provided that the set of points in I, at which the function is not continuous, has Lebesgue measure 0, see also [13]. Therefore, we deal with integrable functions, for example, deal with $v \in L^2(I)$.

2.3.1 The Approximation

Given $v \in L^2(I)$. We want to approximate v using piecewise polynomials finding $u \in P_{N,I,Z}$ in such a way that the error in L^2 norm is minimal. Suppose that the piecewise polynomial u is written in the form

$$u(x) = \sum_{j=1}^{Z} \sum_{i=0}^{N} \widehat{u}_{i,j}\Phi_{i,j}(x), \qquad x \in I.$$

The coefficients $\widehat{u}_{i,j}$ are the unknowns and must be computed in such a way that the error of finding them in the L^2 norm is minimal.

The error is the difference $E(\widehat{u}_{i,j}, x) = v(x) - u(x)$. The error becomes minimal when the derivatives of its norm squared with respect to the unknowns $\widehat{u}_{k,j}$ vanish, i.e. $\frac{\partial}{\partial \widehat{u}_{k,j}}||E||^2_{L^2(I)} = 0$ for all $k = 0, \ldots, N$ and $j = 1, \ldots, Z$. The norm of the error is given by

$$||E||^2_{L^2(I)} = \int_a^b E^2 dx = \sum_{j=1}^{Z} \int_{I_j} \left(E|_{I_j}(x)\right)^2 dx = \sum_{j=1}^{Z} \int_{I_j} \left(v(x) - \sum_{i=0}^{N} \widehat{u}_{i,j}\Phi_{i,j}(x)\right)^2 dx.$$

Thus we have for all $k = 0, \ldots, N$ and $j = 1, \ldots, Z$

$$\begin{aligned} 0 &= \frac{\partial}{\partial \widehat{u}_{k,j}} ||E||^2_{L^2(I)} = \frac{\partial}{\partial \widehat{u}_{k,j}} \left(\sum_{j=1}^{Z} \int_{I_j} \left(v(x) - \sum_{i=0}^{N} \widehat{u}_{i,j} \Phi_{i,j}(x) \right)^2 dx \right) \\ &= \int_{I_j} \frac{\partial}{\partial \widehat{u}_{k,j}} \left(v(x) - \sum_{i=0}^{N} \widehat{u}_{i,j} \Phi_{i,j}(x) \right)^2 dx = 2 \int_{I_j} \left(v(x) - \sum_{i=0}^{N} \widehat{u}_{i,j} \Phi_{i,j}(x) \right) [-\Phi_{k,j}(x)] dx. \end{aligned}$$

Therefore, we must have $\int_{I_j} \left(\sum_{i=0}^{N} \widehat{u}_{i,j} \Phi_{i,j}(x) \right) \Phi_{k,j}(x) dx = \int_{I_j} v(x) \Phi_{k,j}(x) dx$, or in detail

$$\begin{aligned} \widehat{u}_{0,j} \int_{I_j} \Phi_{0,j}(x) \Phi_{0,j}(x) dx + \ldots + \widehat{u}_{N,j} \int_{I_j} \Phi_{N,j}(x) \Phi_{0,j}(x) dx &= \int_{I_j} v(x) \Phi_{0,j}(x) dx, \\ \widehat{u}_{0,j} \int_{I_j} \Phi_{0,j}(x) \Phi_{1,j}(x) dx + \ldots + \widehat{u}_{N,j} \int_{I_j} \Phi_{N,j}(x) \Phi_{1,j}(x) dx &= \int_{I_j} v(x) \Phi_{1,j}(x) dx, \\ &\vdots \\ \widehat{u}_{0,j} \int_{I_j} \Phi_{0,j}(x) \Phi_{N,j}(x) dx + \ldots + \widehat{u}_{N,j} \int_{I_j} \Phi_{N,j}(x) \Phi_{N,j}(x) dx &= \int_{I_j} v(x) \Phi_{N,j}(x) dx. \end{aligned}$$

According to the orthogonality (2.11), we obtain for $j = 1, \ldots, Z$, $h\widehat{u}_{0,j} = \int_{I_j} v(x) \Phi_{0,j}(x) dx$, $\frac{h}{3} \widehat{u}_{1,j} = \int_{I_j} v(x) \Phi_{1,j}(x) dx$, and $\frac{h}{2N+1} \widehat{u}_{N,j} = \int_{I_j} v(x) \Phi_{N,j}(x) dx$, or simply

$$\boxed{\widehat{u}_{i,j} = \frac{2i+1}{h} \langle v, \Phi_{i,j} \rangle_j, \quad i = 0, \ldots, N, \quad j = 1, \ldots, Z.}$$

2.3.2 The Definition

The solution of our problem of approximating a function $v \in L^2(I)$ using piecewise polynomials with minimal error using the L^2 norm is a piecewise polynomial $u \in P_{N,I,Z}$ which is given by

$$u(x) = \sum_{j=1}^{Z} \sum_{i=0}^{N} \widehat{u}_{i,j} \Phi_{i,j}(x), \quad x \in I.$$

The coefficients $\widehat{u}_{i,j}$ have to computed by the following $Z(N+1)$ equations

$$\widehat{u}_{i,j} = \frac{2i+1}{h} \int_{I_j} v(x) \Phi_{i,j}(x) dx. \tag{2.12}$$

We could say also that the solution consists of Z terms. Each term is denoted by

$$u_j(x) = \sum_{i=0}^{N} \widehat{u}_{i,j} \Phi_{i,j}(x), \quad x \in I. \tag{2.13}$$

It is a polynomial of degree N related to the element I_j and is 0 on $I \backslash I_j$.

2.4 Analytical Study

In fact, one can look at the procedure of finding the approximation as the application of an operator. We define this operator $\Pi_{N,Z} : L^2(I) \to P_{N,I,Z}$ by $\Pi_{N,Z}(v) := u$ to give

$$\Pi_{N,Z}(v)(x) = u(x) = \sum_{j=1}^{Z} \sum_{i=0}^{N} \widehat{u}_{i,j} \Phi_{i,j}(x), \quad \text{for } x \in I, \tag{2.14}$$

where the coefficients $\widehat{u}_{i,j}$ are given by (2.12) depending on v. We have for $j = 1, \ldots, Z$

$$\Pi_{N,Z}(v)|_{I_j}(x) = u|_{I_j}(x) = u_j(x) = \sum_{i=0}^{N} \widehat{u}_{i,j} \Phi_{i,j}(x), \quad \text{for } x \in I_j.$$

The operator $\Pi_{N,Z}$ is linear, that is obvious from the definition. Also the operator $\Pi_{N,Z}$ has the following properties.

2.4.1 Identity

The operator $\Pi_{N,Z}$ is a *projection*, i.e. it is idempotent meaning that if v itself is a polynomial of degree N defined on I, then $\Pi_{N,Z}(v) = v$. This implies that $\Pi_{N,Z}(\Pi_{N,Z}(v)) = \Pi_{N,Z}(v)$.

Proof. Let $u = \Pi_{N,Z}(v)$ and $u_j = u|_{I_j}$ for $j = 1, \ldots, Z$. Since v and u_j are polynomials of the degree N defined on I_j, then these both can be written in the form $v(x) = \sum_{i=0}^{N} \widehat{v}_{i,j} \Phi_{i,j}(x)$ and $u_j(x) = \sum_{i=0}^{N} \widehat{u}_{i,j} \Phi_{i,j}(x)$ for $x \in I_j$. We have to prove $\widehat{v}_{i,j} = \widehat{u}_{i,j}$ for all $i = 0, \ldots, N$. According to (2.12) we have for all $i = 0, \ldots, N$ and $j = 1, \ldots, Z$

$$\begin{aligned}
\widehat{u}_{i,j} &= \frac{2i+1}{h} \int_{I_1} v(x) \Phi_{i,j}(x) dx = \frac{2i+1}{h} \int_{I_1} \left(\sum_{k=0}^{N} \widehat{v}_{k,j} \Phi_{k,j}(x) \right) \Phi_{i,j}(x) dx \\
&= \frac{2i+1}{h} \sum_{k=0}^{N} \widehat{v}_{k,j} \left(\int_{I_1} \Phi_{k,j}(x) \Phi_{i,j}(x) dx \right) = \frac{2i+1}{h} \sum_{k=0}^{N} \widehat{v}_{k,j} \left(\frac{h}{2i+1} \delta_{i,k} \right) \\
&= \frac{2i+1}{h} \widehat{v}_{i,j} \frac{h}{2i+1} = \widehat{v}_{i,j}.
\end{aligned}$$

□

2.4.2 This Operator is an Orthogonal Projection

The solution $u = \Pi_{N,Z}(v)$ is an *orthogonal projection* of v on $P_{N,I,Z}$ with respect to the L^2 scalar product in space.

Proof. Let $u = \Pi_{N,Z}(v)$ and $\Phi_{i,j}$ one of the basis functions of the space $P_{N,I,Z}$. We have

$$\begin{aligned}
\langle v-u, \Phi_{i,j}\rangle &= 0+\ldots+0+\langle v-u_j, \Phi_{i,j}\rangle_j+0+\ldots+0 = \langle v, \Phi_{i,j}\rangle_j - \langle u_j, \Phi_{i,j}\rangle_j \\
&= \int_{I_j} v(x)\Phi_{i,j}(x)dx - \int_{I_j}\left(\sum_{k=0}^{N}\widehat{u}_{k,j}\Phi_{k,j}(x)\right)\Phi_{i,j}(x)dx \\
&= \int_{I_j} v(x)\Phi_{i,j}(x)dx - \sum_{k=0}^{N}\widehat{u}_{k,j}\left(\int_{I_j}\Phi_{k,j}(x)\Phi_{i,j}(x)dx\right) \\
&= \int_{I_j} v(x)\Phi_{i,j}(x)dx - \frac{h}{2i+1}\widehat{u}_{i,j},
\end{aligned}$$

and according to (2.12)

$$\langle v-u, \Phi_{i,j}\rangle = \int_{I_j} v(x)\Phi_{i,j}(x)dx - \frac{h}{2i+1}\frac{2i+1}{h}\int_{I_j} v(x)\Phi_{i,j}(x)dx = 0.$$

This means that

$$\langle v-u, \Phi_{i,j}\rangle = 0. \tag{2.15}$$

Thus $v - \Pi_{N,Z}(v)$ is orthogonal to all $\Phi_{i,j}$, i.e. orthogonal to the space $P_{N,I,Z}$ with respect to the L^2 scalar product. In other words $\Pi_{N,Z}(v)$ is the piecewise polynomial closest to v with respect to the L^2 norm. So we call $\Pi_{N,Z}(v)$ the L^2 projection, or projection of v. □

2.4.3 This Operator is the Best Approximation

Directly from the previous property we find $||v-u||_{L^2(I)} \leqslant ||v-p||_{L^2(I)}$ for all $p \in P_{N,I,Z}$. It means that the piecewise polynomial $u = \Pi_{N,Z}(v)$ is the *best approximation* using piecewise polynomials, in the sense that the error in the L^2 norm is as small as possible.

Proof. Let $p \in P_{N,I,Z}$ be arbitrary and set $q = u - p \in P_{N,I,Z}$. Using (2.15) with p replaced by q, we get, using the Cauchy's inequality also,

$$\begin{aligned}
||v-u||^2_{L^2(I)} &= \langle v-u, v-u\rangle = \langle v-u, v-u\rangle + \langle v-u, q\rangle = \langle v-u, v-u+q\rangle \\
&= \langle v-u, v-p\rangle \leqslant ||v-u||_{L^2(I)}||v-p||_{L^2(I)}.
\end{aligned}$$

Dividing by $||v-u||_{L^2(I)}$, if $||v-u||_{L^2(I)} \neq 0$, we get the result. □

2.4.4 Boundedness

Let $v \in L^2(I)$. The following estimate boundedness holds

$$||\Pi_{N,Z}(v)||_{L^2(I)} \leqslant ||v||_{L^2(I)}. \tag{2.16}$$

Proof. Let $u(x) = \Pi_{N,Z}(v)(x) = \sum_{j=1}^{Z}\sum_{i=0}^{N}\widehat{u}_{i,j}\Phi_{i,j}(x)$. We have

$$\begin{aligned}
||u_j||^2_{L^2(I_j)} &= \int_{I_j}|u_j(x)|^2dx = \int_{I_j}\left(\sum_{i=0}^{N}\widehat{u}_{i,j}\Phi_{i,j}(x)\right)^2 dx \\
&= \int_{I_j}\left(\sum_{i=0}^{N}\widehat{u}^2_{i,j}\Phi^2_{i,j}(x) + \sum_{i=0}^{N}\widehat{u}_{i,j}\Phi_{i,j}(x)\sum_{\substack{k=0\\k\neq i}}^{N}\widehat{u}_{k,j}\Phi_{k,j}(x)\right)dx \\
&= \sum_{i=0}^{N}\widehat{u}^2_{i,j}||\Phi_{i,j}||^2_{L^2(I_j)} + \sum_{i=0}^{N}\sum_{\substack{k=0\\k\neq i}}^{N}\widehat{u}_{i,j}\widehat{u}_{k,j}\langle\Phi_{i,j},\Phi_{k,j}\rangle_j.
\end{aligned}$$

According to (2.11), where $\Phi_{i,j}$ are orthogonal, we get

$$||u_j||^2_{L^2(I_j)} = \sum_{i=0}^{N}\widehat{u}^2_{i,j}||\Phi_{i,j}||^2_{L^2(I_j)} = \sum_{i=0}^{N}\widehat{u}_{i,j}\widehat{u}_{i,j}\frac{h}{2i+1}. \tag{2.17}$$

According to (2.12) we obtain

$$\begin{aligned}
||u_j||^2_{L^2(I_j)} &= \sum_{i=0}^{N}\widehat{u}_{i,j}\left[\frac{2i+1}{h}\int_{I_j}v(x)\Phi_{i,j}(x)dx\right]\frac{h}{2i+1} \\
&= \int_{I_j}v(x)\left(\sum_{i=0}^{N}\widehat{u}_{i,j}\Phi_{i,j}(x)\right)dx = \int_{I_j}v(x)u_j(x)dx,
\end{aligned}$$

and using the Cauchy Schwarz inequality we find $||u_j||^2_{L^2(I_j)} \leqslant ||v||_{L^2(I_j)}||u_j||_{L^2(I_j)}$ or $||u_j||_{L^2(I_j)} \leqslant ||v||_{L^2(I_j)}$. Squaring and taking the summation over all elements I_j we get

$$||u||^2_{L^2(I)} = \sum_{j=1}^{Z}||u_j||^2_{L^2(I_j)} \leqslant \sum_{j=1}^{Z}||v||^2_{L^2(I_j)} = ||v||^2_{L^2(I)}.$$

Finally, by taking the square root, we obtain the result. □

Corollary 2.3. From the equality (2.17), we obtain for all $j = 1, \ldots, Z$

$$\frac{h}{2N+1}\sum_{i=0}^{N}\widehat{u}^2_{i,j} \leqslant ||u_j||^2_{L^2(I_j)} \leqslant h\sum_{i=0}^{N}\widehat{u}^2_{i,j}.$$

Let $\widehat{\mathbf{u}}_j := (\widehat{u}_{0,j}, \ldots, \widehat{u}_{N,j})^T$. Using the Euclidean vector norm[1] we obtain the estimates

$$\frac{h}{2N+1}||\widehat{\mathbf{u}}_j||^2_e \leqslant ||u_j||^2_{L^2(I_j)} \leqslant h||\widehat{\mathbf{u}}_j||^2_e. \tag{2.19}$$

[1]The Euclidean norm is defined in $\mathbb{R}^{N+1}$ by

$$||p||_e = \sqrt{p_1^2 + \ldots + p_{N+1}^2}, \quad \text{for all } p = (p_1, \ldots, p_{N+1})^T \in \mathbb{R}^{N+1}. \tag{2.18}$$

Setting $\underline{m} := ||\Phi_{N,j}||^2_{L^2(I_j)} = \frac{h}{2N+1}$ and $\overline{m} := ||\Phi_{0,j}||^2_{L^2(I_j)} = h$ we obtain

$$\underline{m}||\widehat{\mathbf{u}}_j||^2_e \leqslant ||u_j||^2_{L^2(I_j)} \leqslant \overline{m}||\widehat{\mathbf{u}}_j||^2_e. \tag{2.20}$$

2.4.5 The Error Estimates of the Projection Operator

We prove error estimates with help of the following theorem, which is a version of the Bramble-Hilbert Lemma given by Watkins [25, Theorem 1].

Theorem 2.4. Let $\Upsilon \in \mathbb{R}^n$ be a domain such that the identity map $I : W^{N+1,2}(\Upsilon) \to W^{N,2}(\Upsilon)$ is a compact operator, i.e. we have a compact embedding. Let $W^{N+1,2}(J)$, with $J = [-1, 1]$, be the Sobolev space defined by (2.2), and let the seminorm $|\cdot|_{W^{N+1,2}(J)}$, which is defined by (2.4). Let $\mathcal{B} : W^{N+1,2}(J) \to Y$ be a bounded linear operator with domain $W^{N+1,2}(J)$ and range in a normed linear space Y, and let $||\cdot||_Y$ be its norm. Thus there exists a constant $||\mathcal{B}||$ such that $||\mathcal{B}(f)||_Y \leqslant ||\mathcal{B}|| \cdot ||f||_{W^{N+1,2}(J)}$ for all $f \in W^{N+1,2}(J)$. Suppose also that $\mathcal{B}(p) = 0$ for any $p \in P_{N,J}$ in the space of polynomials of degree N. Then there is a constant C_1 which depends on J and N, but not on $\mathcal{B}$, such that

$$||\mathcal{B}(f)||_Y \leqslant C_1 \cdot ||\mathcal{B}|| \cdot |f|_{W^{N+1,2}(J)}, \quad \text{for all} \quad f \in W^{N+1,2}(J).$$

Theorem 2.5. Suppose that the interval $I = [a, b]$ has a uniform partition of Z subintervals with constant mesh size $h = (b - a)/Z$. Then, for each $v \in W^{N+1,2}(I)$, the following error estimates hold

$$\begin{aligned} ||\Pi_{N,Z}(v) - v||_{L^2(I)} &\leqslant C_2 h^{N+1}|v|_{W^{N+1,2}(I)}, \\ ||\Pi_{N,Z}(v) - v||_{L^1(I)} &\leqslant C_3 h^{N+1}|v|_{W^{N+1,2}(I)}, \end{aligned}$$

where $||\cdot||_{W^{N+1,2}(I)}$ and $|\cdot|_{W^{N+1,2}(I)}$ the norm and the seminorm, which are defined in (2.3) and (2.4), respectively.

Proof. **The First Inequality.** Consider $\Upsilon = J = [-1, 1]$ and $f \in W^{N+1,2}(J)$ from Theorem 2.4. We set $Y = L^2(J)$. Let $\Pi_{N,1}$ be the projection operator in the special case where the interval $I = J$ and $Z = 1$. In this case the assumption of a compact embedding for Theorem 2.4 holds. Then, we can say that the operator $\mathcal{B}(f) = \Pi_{N,1}(f) - f$ is

1. linear, due to the linearity of the projection operator,
2. bounded, due to (2.16) for the unique element J, and using the triangle inequality, we get $||\mathcal{B}(f)||_{L^2(J)} = ||\Pi_{N,1}(f) - f||_{L^2(J)} \leqslant ||\Pi_{N,1}(f)||_{L^2(J)} + ||f||_{L^2(J)} \leqslant 2||f||_{L^2(J)}$,
3. due to the property 2.4.1, we have $\mathcal{B}(p) = \Pi_{N,1}(p) - p = 0$ for all $p \in P_{N,J}$.

Thus, by Theorem 2.4, there is a constant C_1 with $||\Pi_{N,1}(f) - f||_{L^2(J)} \leqslant C_1|f|_{W^{N+1,2}(J)}$. Then by squaring

$$\int_{-1}^{1} |\Pi_{N,1}(f)(\xi) - f(\xi)|^2 d\xi \leqslant C_1^2 \int_{-1}^{1} |D^{(N+1)}f(\xi)|^2 d\xi.$$

Let $j = 1, \ldots, Z$ be fixed and x_j be the midpoint of the element I_j. We use the inverses $\gamma_j^{-1} : J \to I_j$ of the linear transformations γ_j given in (2.7), and write $x = \gamma_j^{-1}(\xi) := \frac{h}{2}\xi + x_j$ for $\xi \in J$. We also suppose that $f = v \circ \gamma_j^{-1}$. Then for $x \in I_j$ we have $v(x) = v(\gamma_j^{-1}(\xi)) = (v \circ \gamma_j^{-1})(\xi) = f(\xi)$ and the chain rule gives

$$\frac{df(\xi)}{d\xi} = \frac{df(\xi)}{dx}\frac{dx}{d\xi} = \frac{dv(x)}{dx}\frac{dx}{d\xi} = \frac{h}{2}\frac{dv(x)}{dx} \quad \Rightarrow \qquad D^{(N+1)}f(\xi) = \left(\frac{h}{2}\right)^{N+1} D^{(N+1)}v(x).$$

Furthermore, by defining the operators $\mathcal{B}_j : W^{N+1,2}(I_j) \to L^2(I_j)$ by

$$\mathcal{B}_j(v) = \Pi_{N,Z}(v) - v = \Pi_{N,1}(f \circ \gamma_j) - (f \circ \gamma_j), \qquad \text{for } \; v \in W^{N+1,2}(I_j),$$

and noting that $d\xi = \frac{2}{h}dx$, we can now rewrite the last inequality with the variable x and with the main projection operator $\Pi_{N,Z}$ as follows

$$\frac{2}{h}\int_{I_j} |\Pi_{N,Z}(v)(x) - v(x)|^2 dx \leqslant \frac{2}{h}C_1^2 \int_{I_j} \left| \left(\frac{h}{2}\right)^{N+1} D^{(N+1)}v(x) \right|^2 dx,$$

or

$$\int_{I_j} |\Pi_{N,Z}(v)(x) - v(x)|^2 dx \leqslant \frac{C_1^2}{2^{2N+2}} h^{2N+2} \int_{I_j} \left|D^{(N+1)}v(x)\right|^2 dx.$$

Then, by taking $C_2 = C_1 2^{-N-1}$, we get

$$||\Pi_{N,Z}(v) - v||^2_{L^2(I_j)} \leqslant C_2^2 h^{2N+2} |v|^2_{W^{N+1,2}(I_j)}. \tag{2.21}$$

By summation over all j we get

$$||\Pi_{N,Z}(v) - v||^2_{L^2(I)} = \sum_{j=1}^{Z} ||\Pi_{N,Z}(v) - v||^2_{L^2(I_j)} \leqslant C_2^2 h^{2N+2} \sum_{j=1}^{Z} |v|^2_{W^{N+1,2}(I_j)} = C_2^2 h^{2N+2} |v|^2_{W^{N+1,2}(I)}.$$

By taking the square root follows the first inequality.

The Second Inequality. Using the Cauchy Schwarz inequality and (2.21), we have

$$\begin{aligned} ||\Pi_{N,Z}(v) - v||_{L^1(I_j)} &= \int_{I_j} |\Pi_{N,Z}(v) - v| dx \leqslant \left(\int_{I_j} dx\right)^{\frac{1}{2}} ||\Pi_{N,Z}(v) - v||_{L^2(I_j)} \\ &= h^{\frac{1}{2}} ||\Pi_{N,Z}(v) - v||_{L^2(I_j)} \leqslant C_2 h^{N+\frac{3}{2}} |v|_{W^{N+1,2}(I_j)}. \end{aligned}$$

By summation over all j, we get

$$||\Pi_{N,Z}(v) - v||_{L^1(I)} = \sum_{j=1}^{Z} ||\Pi_{N,Z}(v) - v||_{L^1(I_j)} \leqslant C_2 h^{N+\frac{3}{2}} \sum_{j=1}^{Z} |v|_{W^{N+1,2}(I_j)}. \tag{2.22}$$

On the other hand, according to the inequality $\epsilon\vartheta \leqslant \frac{1}{2}(\epsilon^2 + \vartheta^2)$, which always holds for all $\epsilon, \vartheta \in \mathbb{R}$, we have

$$\begin{aligned}
\left(\sum_{j=1}^{Z} |v|_{W^{N+1,2}(I_j)}\right)^2 &= \sum_{j=1}^{Z} |v|^2_{W^{N+1,2}(I_j)} + \sum_{j=1}^{Z}\sum_{\substack{i=1\\ i\neq j}}^{Z} |v|_{W^{N+1,2}(I_j)}|v|_{W^{N+1,2}(I_i)} \\
&\leqslant \sum_{j=1}^{Z} |v|^2_{W^{N+1,2}(I_j)} + \frac{1}{2}\sum_{j=1}^{Z}\sum_{\substack{i=1\\ i\neq j}}^{Z} \left(|v|^2_{W^{N+1,2}(I_j)} + |v|^2_{W^{N+1,2}(I_i)}\right) \\
&= \sum_{j=1}^{Z} |v|^2_{W^{N+1,2}(I_j)} + \frac{1}{2}\sum_{j=1}^{Z}\sum_{\substack{i=1\\ i\neq j}}^{Z} |v|^2_{W^{N+1,2}(I_j)} + \frac{1}{2}\sum_{j=1}^{Z}\sum_{\substack{i=1\\ i\neq j}}^{Z} |v|^2_{W^{N+1,2}(I_i)} \\
&= \sum_{j=1}^{Z} |v|^2_{W^{N+1,2}(I_j)} + \frac{1}{2}\sum_{\substack{i=1\\ i\neq j}}^{Z}\sum_{j=1}^{Z} |v|^2_{W^{N+1,2}(I_j)} + \frac{1}{2}\sum_{\substack{j=1\\ i\neq j}}^{Z}\sum_{i=1}^{Z} |v|^2_{W^{N+1,2}(I_i)}.
\end{aligned}$$

This implies that

$$\begin{aligned}
\left(\sum_{j=1}^{Z} |v|_{W^{N+1,2}(I_j)}\right)^2 &\leqslant \sum_{j=1}^{Z} |v|^2_{W^{N+1,2}(I_j)} + \frac{Z-1}{2}\sum_{j=1}^{Z} |v|^2_{W^{N+1,2}(I_j)} + \frac{Z-1}{2}\sum_{i=1}^{Z} |v|^2_{W^{N+1,2}(I_i)} \\
&= Z\sum_{j=1}^{Z} |v|^2_{W^{N+1,2}(I_j)}.
\end{aligned}$$

Since $Z = (b-a)/h$, then we have

$$\left(\sum_{j=1}^{Z} |v|_{W^{N+1,2}(I_j)}\right)^2 \leqslant \frac{b-a}{h}\sum_{j=1}^{Z} |v|^2_{W^{N+1,2}(I_j)} = \left(\sqrt{b-a}h^{-\frac{1}{2}}|v|_{W^{N+1,2}(I)}\right)^2.$$

Taking the square root and then substituting in (2.22) and taking $C_3 = C_2\sqrt{b-a}$, we get finally $||\Pi_{N,Z}(v) - v||_{L^1(I)} \leqslant C_3 h^{N+1}|v|_{W^{N+1,2}(I)}$, thus the second inequality holds. □

The above estimates are all related to smooth data. However, we sometimes face discontinuities which effect the accuracy and the order.

Theorem 2.6. Let $N \in \mathbb{N}_0$, $I \subset \mathbb{R}$, and $v \in B(I)$ be a bounded function which has a discontinuity in I. Suppose h is the length of I. Then, there is a constant C_4 which is only related to N, such that the following error estimate, for all $1 \leqslant p < \infty$, holds

$$||\Pi_{N,Z}(v) - v||_{L^p(I)} \leqslant C_4||v||_{L^\infty(I)}h^{1/p}.$$

Proof. Since $h = |I|$, we take a partition of one element $I_1 = I$ and assume that $u(x) = \Pi_{N,Z}(v)(x) = \sum_{i=0}^{N} \widehat{u}_{i,1}\Phi_{i,1}(x)$ with $\widehat{u}_{i,1} = \frac{2i+1}{h}\int_I v(x)\Phi_{i,1}(x)dx$ for $i = 0, \ldots, N$.
For all $i = 0, \ldots, N$ we have $||\Phi_{i,1}||_{L^\infty(I)} = 1$ and, for $x \in I$

$$\begin{aligned}
|u(x) - v(x)| &\leqslant |u(x)| + |v(x)| \leqslant \sum_{i=0}^{N} |\widehat{u}_{i,1}||\Phi_{i,1}(x)| + ||v||_{L^\infty(I)} \\
&\leqslant \sum_{i=0}^{N} \left| \frac{2i+1}{h} \int_I v(x)\Phi_{i,1}(x)dx \right| ||\Phi_{i,1}||_{L^\infty(I)} + ||v||_{L^\infty(I)} \\
&\leqslant \sum_{i=0}^{N} \left(\frac{2i+1}{h} h||v||_{L^\infty(I)} \underbrace{||\Phi_{i,1}||_{L^\infty(I)}}_{=1} \right) \underbrace{||\Phi_{i,1}||_{L^\infty(I)}}_{=1} + ||v||_{L^\infty(I)} \\
&= \sum_{i=0}^{N} (2i+1)||v||_{L^\infty(I)} + ||v||_{L^\infty(I)} \\
&= \underbrace{\left(\sum_{i=0}^{N} (2i+1) + 1 \right)}_{:= C_4(N)} ||v||_{L^\infty(I)} = C_4||v||_{L^\infty(I)}.
\end{aligned}$$

Then for all $1 \leqslant p < \infty$ we have

$$||u - v||^p_{L^p(I)} = \int_I |u(y) - v(y)|^p \, dy \leqslant \int_I C_4^p ||v||^p_{L^\infty(I)} dy = C_4^p ||v||^p_{L^\infty(I)} \int_I dy = C_4^p ||v||^p_{L^\infty(I)} h.$$

Thus the error in the L^p norm is $||u - v||_{L^p(I)} \leqslant C_4||v||_{L^\infty(I)} h^{1/p} = \mathcal{O}\left(h^{1/p}\right)$. □

Chapter 3

Examples of the Projections

3.1 Preface

Again let Z be a positive integer, N be a non negative integer, $\Delta = \{x_{j+\frac{1}{2}}\}_{j=0}^{Z}$ be a uniform partition of equally points of an interval $I \subset \mathbb{R}$ with a constant mesh size h. Let I_j; $j = 1, \ldots, Z$, be an element of the partition.

In this chapter we will study two types of functions from the space $L^2(I)$ with a support I which changes with respect to the examples. The types are polynomials of various degrees and one of the trigonometric functions. Then we will give an example of a function with less smoothness.

We will check that the Identity 2.4.1 holds for the examples of the polynomials. Furthermore, the error estimates will be shown, such that the smooth order $N + 1$ will appear when we use the projection $\Pi_{N,Z}$ of degree N.

We use the projections $\Pi_{N,Z}$ to polynomials of degree N, which we defined in the previous chapter. We usually view the formula of the projection of one element I_j and the formula of the complete projection on I will be the summation.

We will use the following notations. The projection of a function v to a polynomial of degree N is denoted by $u = \Pi_{N,Z}(v)$ and the restriction on I_j is $u_j = u|_{I_j}$.

After that we will compute the errors on the complete interval I by computing the L^2 norm numerically using some Gaussian quadrature rule.

3.1.1 Gaussian Quadrature Rule

Let $n \in \mathbb{N}$. The Gaussian quadrature rule of order n computes the integral of some function $v \in L^2(I)$ on the interval $I = [a, b]$, numerically, by the following formula $\int_a^b v(x)dx \equiv \sum_{i=1}^{n} \omega_i v(\xi_i)$, where $\xi_1, \ldots, \xi_n \in [a, b]$ are called nodes and $\omega_1, \ldots, \omega_n$ are called weights and chosen to minimize the expected error obtained in the approximation.

The roots of the Legendre polynomials give us the nodes and weights for the quadrature rule. In this way, the nodes $\xi_1, \ldots, \xi_n$ produce an integral approximation formula that gives exact results for any polynomial of degree less than $2n$ are the roots of the Legendre polynomial

of degree n. This is established by Theorem 4.7 in [4]. The Gaussian rule by using the roots of Legendre polynomials becomes on the interval $[-1, 1]$.

To do the integral over an arbitrary interval $[a, b]$, the integral can be transformed into an integral over $[-1, 1]$ by using the change of variables $\int_a^b v(x)dx = \int_{-1}^{1} v\left(\frac{(b-a)s+(b+a)}{2}\right)\frac{b-a}{2}ds$. This permits Gaussian quadrature to be applied to any interval $[a, b]$.

For our work, we choose the Gaussian rule of $\max\{\deg, N\}$ points, where deg is the degree of the initial function v, and $\deg = 1$ for the $v(x) = \sin(x)$.

3.1.2 Experimental Order of Convergence EOC

We also investigate the orders of the accuracy of the projections numerically by calculating the experimental order of the convergence (EOC).

Let generally X ba a linear space with some norm $||\cdot||_X$ and let $v_h \in X$ be a numerical approximation of a given function $v \in X$ which depends on a parameter h of the discretization. The convergence of v_h towards v as h tends to zero can be quantified by $||v_h - v||_X \leqslant Ch^\kappa$, with the order of convergence κ. This gives a possibility to quantify the quality of a numerical scheme. If we can compute two numerical solutions v_h and $v_{h'}$, then the order κ can be estimated experimentally by $\kappa \simeq EOC(h, h') = \frac{\log(||v_{h'}-v||_X/||v_h-v||_X)}{\log(h'/h)}$.

3.2 Polynomial of Degree One

We consider the polynomial $v(x) = x - 1$ defined on $I = [0, 2]$. We have according to (2.12)

$$\widehat{u}_{0,j} = \frac{1}{h}\int_{I_j}(x-1)dx = x_j - 1, \quad \widehat{u}_{1,j} = \frac{3}{h}\int_{I_j}(x-1)\frac{2}{h}(x-x_j)dx = \frac{h}{2}.$$

3.2.1 $\Pi_{0,Z}$

This projection is given according to (2.14) by $\Pi_{0,Z}(v)|_{I_j}(x) = u_j(x) = x_j - 1$. The error using the L^2 norm is equal to $||v-\Pi_{0,Z}(v)||^2_{L^2(I)} = \frac{h^2}{6}$. Then we have $||v-\Pi_{0,Z}(v)||_{L^2(I)} = \frac{h}{\sqrt{6}} = \mathcal{O}(h)$, i.e. we get a first order projection using the L^2 norm. Tables 3.1 and 3.2 show the errors in the L^1 and L^2 norms computed numerically and analytically, respectively, by using the Matlab, as well as the elapsed time, in seconds, during the computations. Note that we can obtain the same values of the error by using the exact error $\frac{h}{\sqrt{6}}$. We note that the numerical integral is faster than the analytical one and has almost similar values as the analytical. So, we always will do only the numerical computations.

Example. $Z = 5$, $h = 0.4$, $x \in \{0, 0.4, 0.8, 1.2, 1.6, 2\}$ and the projection is

$$\Pi_{0,5}(v)|_{I_j}(x) = \begin{cases} -0.8 & \text{for} \quad x \in [0\ \ , 0.4[= [0\ \ , 0.38], \\ -0.4 & \text{for} \quad x \in [0.4, 0.8[= [0.4, 0.78], \\ 0 & \text{for} \quad x \in [0.8, 1.2[= [0.8, 1.18], \\ 0.4 & \text{for} \quad x \in [1.2, 1.6[= [1.2, 1.58], \\ 0.8 & \text{for} \quad x \in [1.6, 2\ \ ]. \end{cases}$$

Figure 3.1 shows this projection.

3.2.2 $\Pi_{1,Z}$

This projection is given according to (2.14) by

$$\Pi_{1,Z}(v)|_{I_j}(x) = u_j(x) = \widehat{u}_{0,j} + \widehat{u}_{1,j}\Phi_{1,j}(x) = x_j - 1 + \frac{h}{2}\frac{2}{h}(x - x_j) = x - 1 = v(x).$$

Thus $\Pi_{1,Z}$ becomes the identity operator and this agrees with the Property 2.4.1.

3.3 Polynomial of Degree Two

We consider the polynomial $v(x) = x^2 - 3x + 2$ defined on $I = [0, 3]$. We have

$$\begin{aligned}
\widehat{u}_{0,j} &= \frac{1}{h}\int_{I_j}(x^2 - 3x + 2)dx = x_j^2 - 3x_j + \frac{h^2}{12} + 2, \\
\widehat{u}_{1,j} &= \frac{3}{h}\int_{I_j}(x^2 - 3x + 2)\frac{2}{h}(x - x_j)dx = x_j h - \frac{3h}{2}, \\
\widehat{u}_{2,j} &= \frac{5}{h}\int_{I_j}(x^2 - 3x + 2)\left(\frac{3}{2}\left(\frac{2}{h}(x - x_j)\right)^2 - \frac{1}{2}\right)dx = \frac{h^2}{6}.
\end{aligned}$$

One could easily find that the operator $\Pi_{2,Z}$ is the identity operator $\Pi_{2,Z}(v)(x) = v(x)$.

3.3.1 $\Pi_{0,Z}$

This projection is given by $\Pi_{0,Z}(v)|_{I_j}(x) = u_j(x) = x_j^2 - 3x_j + \frac{h^2}{12} + 2$. We present the way of computing the L^2 norm of the error in detail. We first take the integral related to I_j

$$\begin{aligned}
&\int_{I_j}(v(x) - u(x))^2\,dx = \int_{I_j}\left((x^2 - 3x + 2) - (x_j^2 - 3x_j + \frac{h^2}{12} + 2)\right)^2 dx \\
&= h^5\left(x_j^4 - 2x_j^3 + \frac{3}{2}x_j^2 - \frac{1}{2}x_j + \frac{49}{720}\right) + h^4\left(-6x_j^3 + 9x_j^2 - \frac{9}{2}x_j + \frac{3}{4}\right) \\
&+h^3\left(-2x_j^4 + 8x_j^3 + \frac{17}{6}x_j^2 - \frac{17}{2}x_j + 3\right) + h^2(6x_j^3 - 21x_j^2 + 9x_j) + h(x_j^4 - 6x_j^3 + 9x_j^2).
\end{aligned}$$

We have $x_j = (j - \frac{1}{2})h$ then the integral becomes

$$\int_{I_j}(v(x) - u(x))^2\,dx = h^5\left(\frac{j^2}{3} - \frac{j}{3} + \frac{4}{45}\right) - h^4\left(j - \frac{1}{2}\right) + \frac{3h^3}{4}.$$

By using the following rules of summation $\sum_{j=1}^{Z} j^2 = \frac{Z(Z+1)(2Z+1)}{6}$, $\sum_{j=1}^{Z} j = \frac{Z(Z+1)}{2}$, and $\sum_{j=1}^{Z} 1 = Z$, and noting the relation $h = \frac{3}{Z}$ we find that the summation over all I_j gives

$$||v-u||^2_{L^2(I)} = \sum_{j=1}^{Z} \int_{I_j} (v(x)-u(x))^2 \, dx = \frac{3h^2}{4} - \frac{h^4}{15}.$$

Finally we get $||v-u||_{L^2(I)} = \sqrt{\frac{3h^2}{4} - \frac{h^4}{15}} = \mathcal{O}(h)$, i.e. we get a first order projection using the L^2 norm. Table 3.3 shows the errors in the L^1 and L^2 norms computed by using the Matlab.
Example. $Z = 5$ we have $h = \frac{3}{5} = 0.6$ and the solution is given by

$$\Pi_{0,5}(v)|_{I_j}(x) = \frac{1}{50} \begin{cases} 61 & \text{for} \quad x \in [0 \ \ , 0.57], \\ 7 & \text{for} \quad x \in [0.6, 1.17], \\ -11 & \text{for} \quad x \in [1.2, 1.77], \\ 7 & \text{for} \quad x \in [1.8, 2.17], \\ 61 & \text{for} \quad x \in [2.4, 3]. \end{cases}$$

Figure 3.2 shows this projection.

3.3.2 $\Pi_{1,Z}$

It is given by $\Pi_{1,Z}(v)|_{I_j}(x) = u_j(x) = \widehat{u}_{0,j} + \widehat{u}_{1,j}\Phi_{1,j}(x) = \frac{h^2}{12} - x_j^2 + 2 + (2x_j - 3)x$. The error is equal to $||v - \Pi_{1,Z}(v)||_{L^2(I)} = \frac{h^2}{\sqrt{60}} = \mathcal{O}(h^2)$, i.e. we get a second order projection using the L^2 norm. We get by using Matlab the Table 3.4. With $Z = 5$ we have the following solution

$$\Pi_{1,5}(v)|_{I_j}(x) = \frac{1}{50} \begin{cases} 97 + 120x & \text{for} \quad x \in [0 \ \ , 0.57], \\ 61 + 60x & \text{for} \quad x \in [0.6, 1.17], \\ -11 & \text{for} \quad x \in [1.2, 1.77], \\ -119 + 60x & \text{for} \quad x \in [1.8, 2.17], \\ -263 + 120x & \text{for} \quad x \in [2.4, 3]. \end{cases}$$

Figure 3.3 shows this projection.

3.4 Polynomial of Degree Three

We consider the polynomial $v(x) = x^3 - x$ defined on $I = [-2, 2]$. We have $\widehat{u}_{0,j} = x_j^3 + \left(\frac{h^2}{4} - 1\right) x_j$, $\widehat{u}_{1,j} = \frac{3h}{2}x_j^2 + \frac{3h^3}{40} - \frac{h}{2}$, $\widehat{u}_{2,j} = \frac{h^2}{2}x_j$, and $\widehat{u}_{3,j} = \frac{h^3}{20}$. The operator $\Pi_{3,Z}$ is the identity operator and give the same function v. Table 3.5 presents the errors using Matlab. In Figures 3.4 we view the projections with $Z = 5$.

3.5 Polynomial of Degree Four

We consider the function $v(x) = x^4 - 2x^3 - x^2 + 2x$ defined on $I = [-1, 2]$. The operator $\Pi_{3,Z}$ is the identity operator. Table 3.6 shows the computations by using Matlab. In Figures 3.5 we view the four projections $\Pi_{N,5}$ with $N = 0, 1, 2$.

3.6 A Trigonometric Function

We first consider the trigonometric function $v(x) = \sin(x)$ defined on $I = [0, 2\pi]$. Table 3.7 gives the errors of computing the projections. Figures 3.6 show $\Pi_{N,5}$ with $N = 0, 1, 2$.

Now we give a test for the discontinuous case and view how the discontinuity effects on the orders. We consider the function

$$v(x) = \begin{cases} \sin(x) & \text{for} \quad 0 \leqslant x \leqslant \frac{\pi}{2}, \\ \sin(x + \pi) = -\sin(x) & \text{for} \quad \frac{\pi}{2} < x \leqslant 3. \end{cases} \tag{3.1}$$

Figures 3.7 show some projections with $Z = 15$. Table 3.8 shows the errors. Note that the order is lost and it is 1 with L^1 norm and $1/2$ with L^2 norm, but with large meshes.

Finally, we consider the same function but with a computational domain symmetric with respect to the location of the jump. We define the function

$$v(x) = \begin{cases} \sin(x) & \text{for} \quad 0 \leqslant x \leqslant \frac{\pi}{2}, \\ -\sin(x) & \text{for} \quad \frac{\pi}{2} < x \leqslant \pi. \end{cases}$$

We give, in Tables 3.9 and 3.10, the errors in two groups of meshes, odd and even. We note that with the odd meshes where the jump is included inside one element the order is lost. It is clear that we get solutions of order 1 with L^1 norm and of order $1/2$ with L^2 norm. However, we get solutions of smooth orders $N + 1$ with the even numbers of the meshes, since the jump point isn't an internal point of any element.

3.7 Summary

In this chapter we have applied the projection operators $\Pi_{N,Z}$ to various types of functions.

- We found that the Property 2.4.1 holds for the first four examples.
- The error estimates in Section 2.4.5 have clearly appeared. The smooth order $N + 1$ was achieved when we have used continuous functions. While the order $\mathcal{O}\left(h^{1/p}\right)$ for $p = 1, 2$ was obtained with discontinuous functions.
- We noted that the discontinuity affects the orders, according to Theorem 2.6. Table 3.8 showed that the order is lost and it was 1 for the L^1 norm and $1/2$ for the L^2 norm, for suitably fine meshes. On rather coarse meshes the solution error could become larger after a refinement.
- In the special case where the computational domain was symmetric with respect to the location of the jump and when the jump was included inside one element the order is lost. While when the jump point isn't an internal point of any element the solutions were of smooth orders $N + 1$.

Z	L^1 errors	EOC	L^2 errors	EOC	Elapsed time (seconds)
8	0.1075829		0.1020621		0.447907
16	0.0537914	1	0.0510310	1	0.990749
32	0.0268957	1	0.0255155	1	1.772117
64	0.0134479	1	0.0127578	1	3.546628
128	0.0067239	1	0.0063789	1	7.105968
256	0.0033620	1	0.0031894	1	13.777892
512	0.0016810	1	0.0015947	1	28.650745

Table 3.1: The errors of computing $\Pi_{0,Z}(v)$ for $v(x) = x - 1$. Numerical integration using a Gaussian quadrature rule.

Z	L^1 errors	EOC	L^2 errors	EOC	Elapsed time (seconds)
8	0.12500000		0.1020621		0.592521
16	0.06250000	1	0.0510310	1	1.117595
32	0.03125000	1	0.0255155	1	2.580826
64	0.01562500	1	0.0127578	1	5.009351
128	0.00781250	1	0.0063789	1	14.078991
256	0.00390630	1	0.0031894	1	28.575431
512	0.00195314	1	0.0015947	1	58.079580

Table 3.2: The errors of computing $\Pi_{0,Z}(v)$ for $v(x) = x - 1$. Analytical integration.

Z	L^1 errors	EOC	L^2 errors	EOC
8	0.4398194	1	0.3227234	0.97
16	0.2199097	1	0.1621258	0.99
32	0.1099548	1	0.0811582	1
64	0.0549774	1	0.0405910	1

Table 3.3: The errors of computing $\Pi_{0,Z}(v)$ for $v(x) = x^2 - 3x + 2$.

Z	L^1 errors	EOC	L^2 errors	EOC
8	0.0243801	2	0.0181546	2
16	0.0060950	2	0.0045387	2
32	0.0015238	2	0.0011347	2
64	0.0003809	2	0.0002837	2

Table 3.4: The errors of computing $\Pi_{1,Z}(v)$ for $v(x) = x^2 - 3x + 2$.

Z	$N=0$		$N=1$		$N=2$	
	L^1 errors	EOC	L^1 errors	EOC	L^1 errors	EOC
8	1.6586	0.95	0.1939931	2	0.0071762	3
16	0.8141305	1.03	0.0484983	2	0.0008970	3
32	0.4036951	1.01	0.0121246	2	0.0001121	3
64	0.2009659	1.01	0.0030311	2	0.0000140	3
	L^2 errors	EOC	L^2 errors	EOC	L^2 errors	EOC
8	1.3185	0.90	0.1281740	1.97	0.0047246	3
16	0.6702478	0.98	0.0322172	1.99	0.0005906	3
32	0.3365006	0.99	0.0080651	2	0.0000738	3
64	0.1684224	1	0.0020170	2	0.0000092	3

Table 3.5: The errors of computing $\Pi_{N,Z}(v)$ with $N = 0, 1, 2$ for $v(x) = x^3 - x$.

Z	$N=0$		$N=1$		$N=2$		$N=3$	
	L^1 errors	EOC	L^1 errors	EOC	L^1 errors	EOC	L^1 errors	EOC
8	08.3043	0.95	1.4139	1.86	0.0980560	3	0.0027948	4
16	04.2006	0.98	0.3549049	1.99	0.0122570	3	0.0001747	4
32	02.1141	0.99	0.0888158	2	0.0015321	3	0.0000109	4
64	01.0578	1	0.0222095	2	0.0001915	3	0.0000007	4
	L^2 errors	EOC	L^2 errors	EOC	L^2 errors	EOC	L^2 errors	EOC
8	6.6149	0.77	0.9523059	1.88	0.0591194	2.97	0.0016247	4
16	3.4364	0.94	0.2426263	1.97	0.0074317	2.99	0.1015467	4
32	1.7345	0.99	0.0609405	1.99	0.0009303	3	0.0000063	4
64	0.8693072	1	0.0152529	2	0.0001163	3	0.0000004	4

Table 3.6: The errors of computing $\Pi_{N,Z}(v)$ with $N = 0, 1, 2, 3$ for $v(x) = x^4 - 2x^3 - x^2 + 2x$.

Z	$N=0$		$N=1$		$N=2$		$N=3$	
	L^1 errors	EOC	L^1 errors	EOC	L^1 errors	EOC	L^1 errors	EOC
8	0.7295349	1.14	0.0889003	2.07	0.0053313	3.10	0.0002853	4.08
16	0.3503279	1.06	0.0219792	2.02	0.0006517	3.03	0.0000176	4.02
32	0.1719585	1.03	0.0054797	2	0.0000808	3.01	0.0000011	4
64	0.0852231	1.01	0.0013690	2	0.0000101	3	0.0000001	4
	L^2 errors	EOC	L^2 errors	EOC	L^2 errors	EOC	L^2 errors	EOC
8	0.3977513	0.96	0.0403887	1.96	0.0026849	2.97	0.0001330	3.97
16	0.2004140	0.99	0.0101642	1.99	0.0003375	2.99	0.0000083	3.99
32	0.1004003	1	0.0025452	2	0.0000422	3	0.0000005	4
64	0.0502244	1	0.0006366	2	0.0000053	3	0.0000001	4

Table 3.7: The errors of computing $\Pi_{N,Z}(v)$ with $N = 0, 1, 2, 3$ for $v(x) = \sin(x)$.

Z	$N=0$		$N=1$		$N=2$		$N=3$	
	L^1 errors	EOC	L^1 errors	EOC	L^1 errors	EOC	L^1 errors	EOC
60	0.0679552		0.0204950		0.0164041		0.0158297	
120	0.0214332	1.66	0.0103905	0.98	0.0080134	1.03	0.0043200	1.87
240	0.0169807	0.34	0.0050974	1.03	0.0041007	0.97	0.0039573	0.13
480	0.0084895	1	0.0025464	1	0.0020503	1	0.0019787	1
960	0.0042445	1	0.0012727	1	0.0010252	1	0.0009893	1
1920	0.0021222	1	0.0006362	1	0.0005126	1	0.0004947	1
	L^2 errors	EOC	L^2 errors	EOC	L^2 errors	EOC	L^2 errors	EOC
60	0.2150411		0.1117078		0.0861445		0.0715561	
120	0.1025510	1.07	0.0835836	0.42	0.0568214	0.60	0.0276167	1.37
240	0.1072711	-0.06	0.0558500	0.58	0.0430697	0.40	0.0357772	-0.37
480	0.0758219	0.50	0.0394917	0.50	0.0304548	0.50	0.0252982	0.50
960	0.0536034	0.50	0.0279248	0.50	0.0215348	0.50	0.0178885	0.50
1920	0.0378995	0.50	0.0197458	0.50	0.0152274	0.50	0.0126491	0.50

Table 3.8: The errors of computing $\Pi_{N,Z}(v)$ with $N = 0, 1, 2, 3$ for v given by (3.1).

Z	$N=0$		$N=1$		$N=2$		$N=3$	
	L^1 errors	EOC	L^1 errors	EOC	L^1 errors	EOC	L^1 errors	EOC
5	0.8582418		0.2727565		0.2084071		0.1989590	
15	0.2915306	0.98	0.0875612	1.03	0.0687936	1	0.0663061	1
45	0.0972266	1	0.0286908	1.02	0.0229056	1	0.0221018	1
	L^2 errors	EOC	L^2 errors	EOC	L^2 errors	EOC	L^2 errors	EOC
5	0.7797969		0.3979689		0.3076241		0.2536529	
15	0.4443682	0.51	0.2287304	0.5	0.1764430	0.5	0.1464466	0.5
45	0.2545078	0.51	0.1319953	0.5	0.1017944	0.5	0.0845510	0.5

Table 3.9: The errors with odd meshes.

Z	$N=0$		$N=1$		$N=2$		$N=3$	
	L^1 errors	EOC	L^1 errors	EOC	L^1 errors	EOC	L^1 errors	EOC
10	0.1510497		0.0057227		1.6627e-4		3.8365e-6	
20	0.0748294	1.01	0.0014271	2	2.0664e-5	3	2.3920e-7	4
40	0.0372549	1.01	0.0003566	2	2.5768e-6	3	1.4941e-8	4
	L^2 errors	EOC	L^2 errors	EOC	L^2 errors	EOC	L^2 errors	EOC
10	0.1134762		0.0046034		1.2226e-4		2.4201e-6	
20	0.0568081	1	0.0011521	2	1.5295e-5	3	1.5136e-7	4
40	0.0284128	1	0.0002881	2	1.9124e-6	3	9.4616e-9	4

Table 3.10: The errors with even meshes.

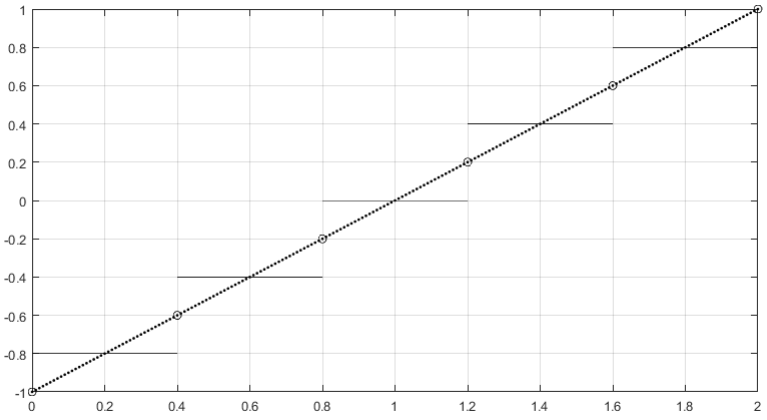


Figure 3.1: The projection $\Pi_{0,Z}(v)$ for $v(x) = x - 1$. The small circles are at the mesh points.

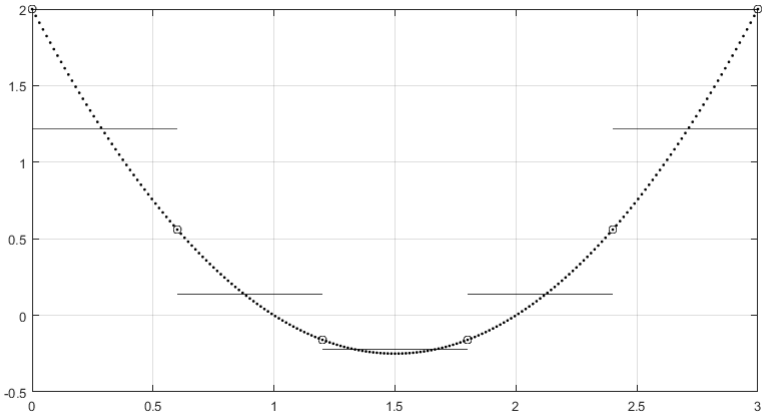


Figure 3.2: The projection $\Pi_{0,Z}$ for $v(x) = x^2 - 3x + 2$.

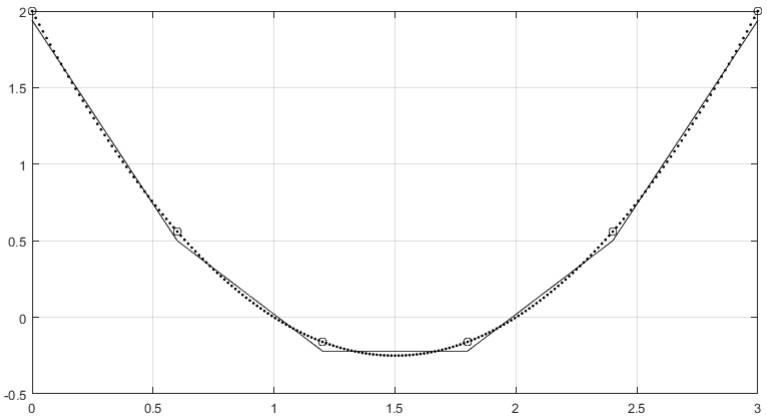


Figure 3.3: The projection $\Pi_{1,Z}$ for $v(x) = x^2 - 3x + 2$.

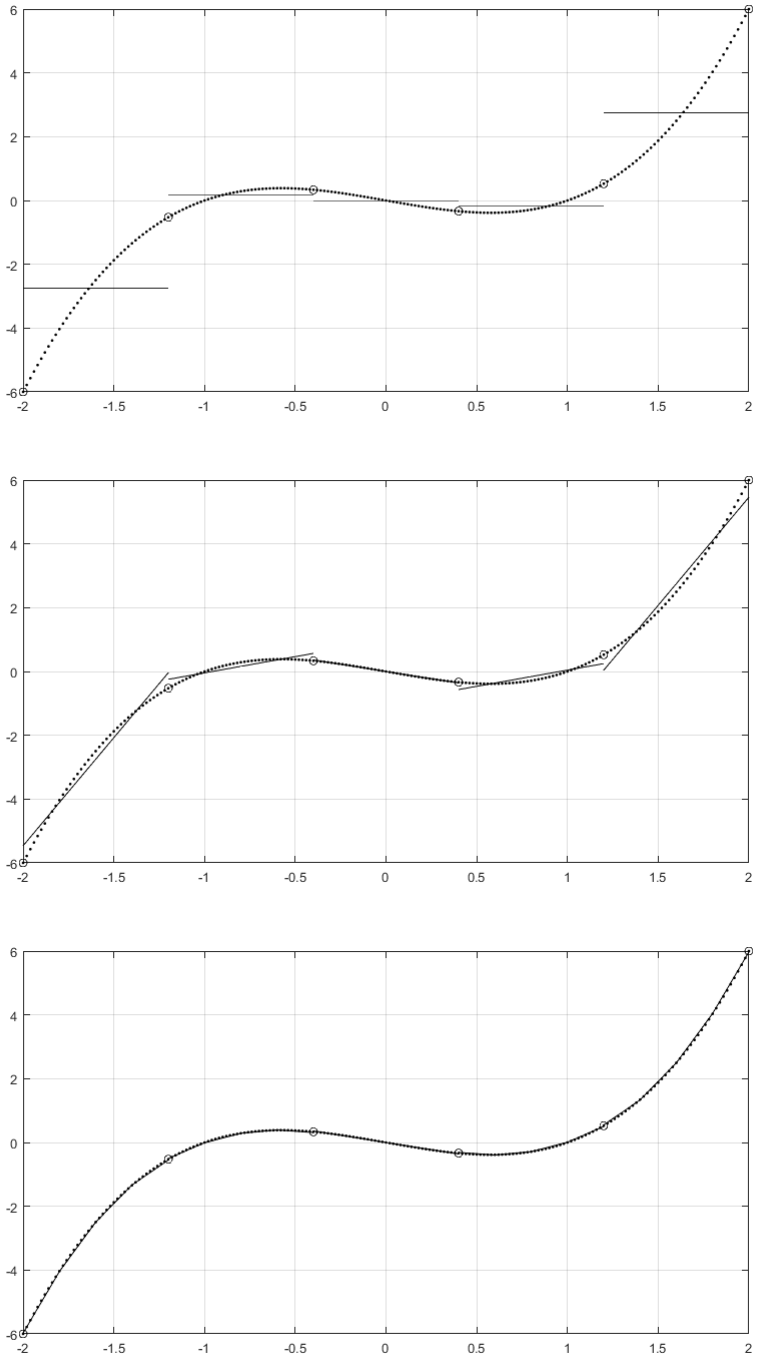


Figure 3.4: The projections $\Pi_{N,5}(v)$ with $N = 0, 1, 2$ for $v(x) = x^3 - x$.

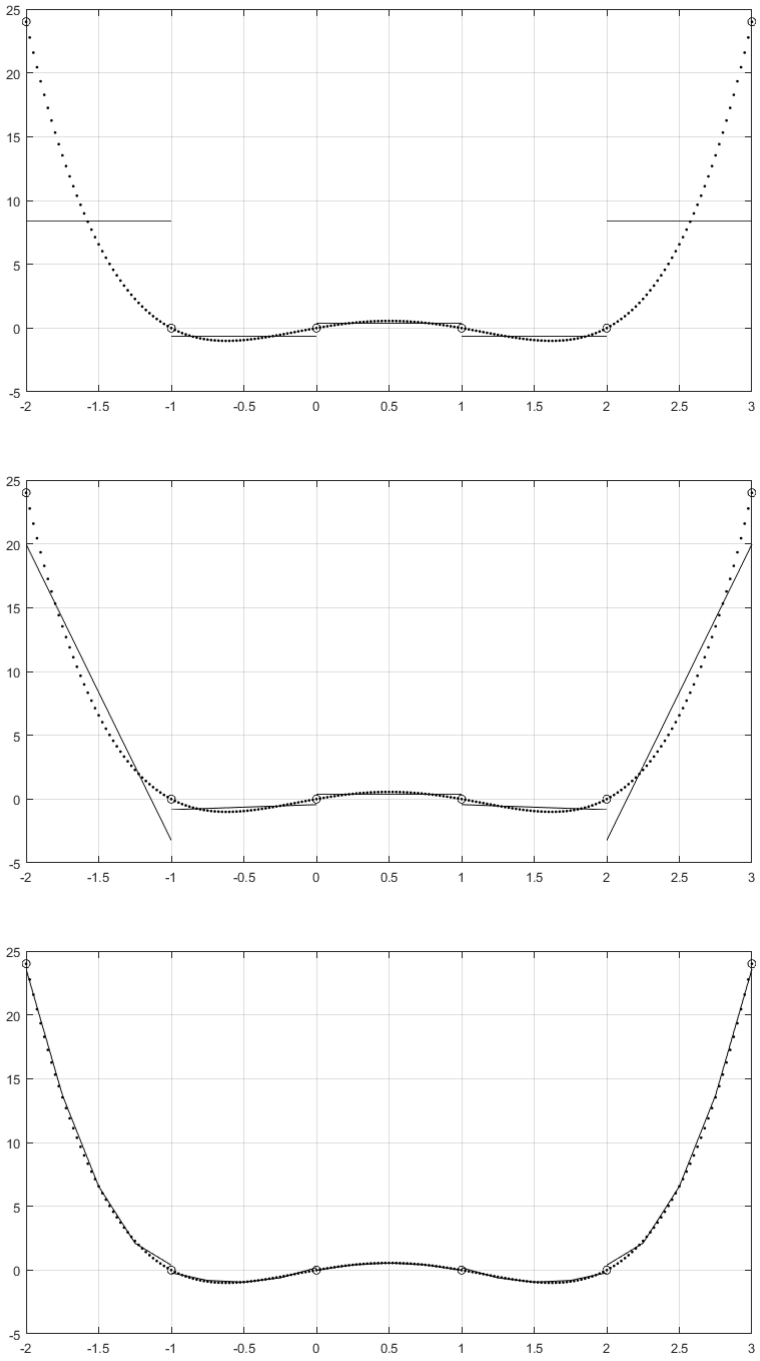


Figure 3.5: The projections $\Pi_{N,5}(v)$ with $N = 0, 1, 2$ for $v(x) = x^4 - 2x^3 - x^2 + 2x$.

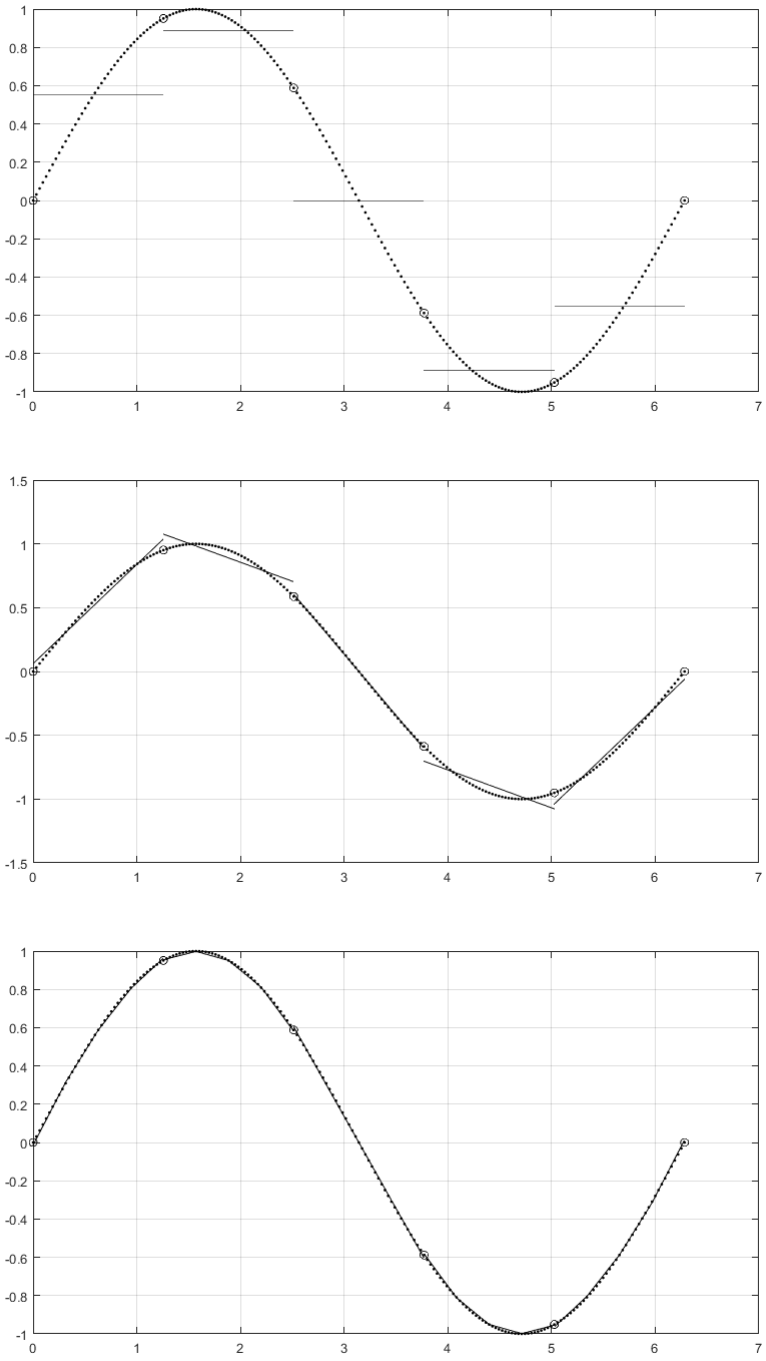


Figure 3.6: The projections $\Pi_{N,5}(v)$ with $N = 0, 1, 2$ for $v(x) = \sin(x)$.

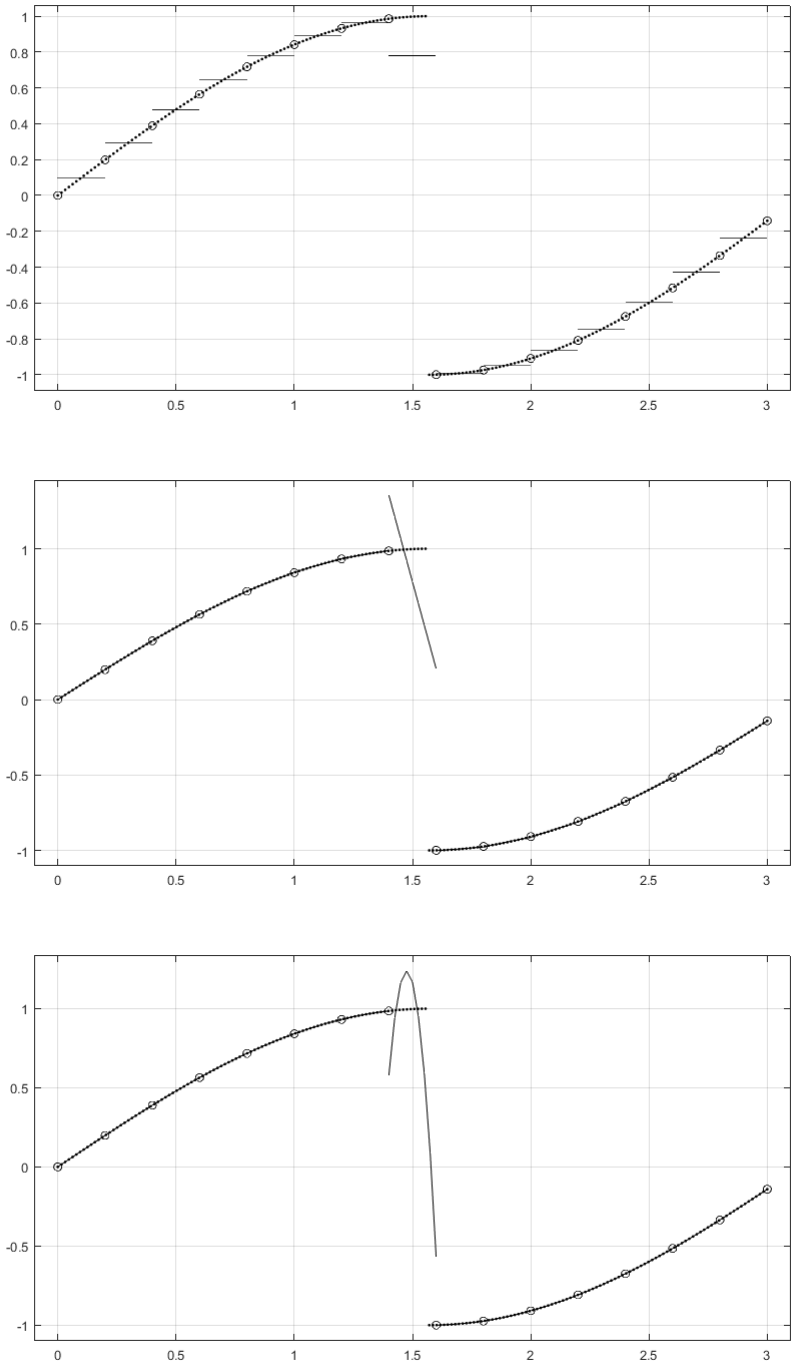


Figure 3.7: The projections $\Pi_{N,15}(v)$ with $N = 0, 1, 2$.

Chapter 4

The Projection onto Piecewise Polynomials: 2D Case

The concept of the projection treated for the approximation in the 1D case can be extended to the 2D case. This extension consists of defining basis functions over a domain $\Omega \subset \mathbb{R}^2$ and compute coefficients $\widehat{u}_{i,j}$ by integrating over Ω.

4.1 2D Discretization

We use a discretization based on rectangles. Let the rectangle $\Omega = [a, b] \times [a', b'] \subset \mathbb{R}^2$ be the computational domain. We assume that Ω can be partitioned using K elements $\Omega = \bigcup_{j=1}^{K} T_j$ where T_j is a rectangle and the partition $\Omega_K = \{T_1, \ldots, T_K\}$ is assumed to be geometrically conforming and non-overlapping. By conforming we mean that each corner point in the interior is shared by exactly four rectangles and each side in the interior by exactly two rectangles. We also assume that the rectangles have the same length h_1 and the same height h_2, where by choosing $Z_1, Z_2 \in \mathbb{N}$ we have $h_1 = \frac{b-a}{Z_1}$, $h_2 = \frac{b'-a'}{Z_2}$, and $K = Z_1 \cdot Z_2$.

We consider one rectangle element $T_j \in \Omega_K$, see Figure 4.1, with $j = 1, \ldots, K$ whose vertices are counter clockwise starting at the lower left corner $v_{j,1} = (x_{j-\frac{1}{2}}, y_{j-\frac{1}{2}})$, $v_{j,2} = (x_{j+\frac{1}{2}}, y_{j-\frac{1}{2}})$, $v_{j,3} = (x_{j+\frac{1}{2}}, y_{j+\frac{1}{2}})$, and $v_{j,4} = (x_{j-\frac{1}{2}}, y_{j+\frac{1}{2}})$ with length$(v_{j,1}v_{j,2})$ =length$(v_{j,3}v_{j,4}) = h_1$ and length$(v_{j,1}v_{j,4})$ =length$(v_{j,2}v_{j,3}) = h_2$.
Now we define the standard reference square $T_S = [-1, 1] \times [-1, 1]$ whose vertices are $v_{S,1} = (-1, -1)$, $v_{S,2} = (1, -1)$, $v_{S,3} = (1, 1)$, $v_{S,4} = (-1, 1)$. We also define piecewise linear reference transformations, see Figure 4.2, $R_j : T_j \to T_S$ from the physical coordinate system $(x, y) \in T_j$ into the reference coordinate system $(\xi, \eta) \in T_S$ by

$$(\xi, \eta) := R_j(x, y) = \left(\frac{2(x - x_j)}{h_1}, \frac{2(y - y_j)}{h_2} \right), \tag{4.1}$$

or in detail by the formulas $\xi(x) = \frac{2(x-x_j)}{h_1}$ and $\eta(y) = \frac{2(y-y_j)}{h_2}$ where (x_j, y_j) is the center of T_j.

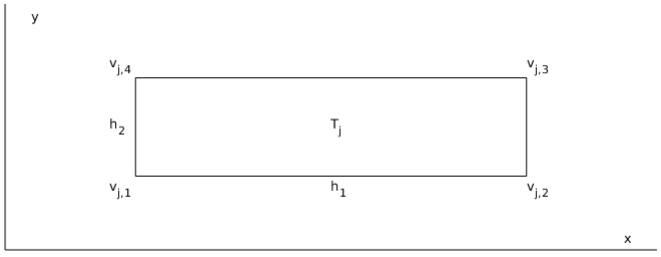


Figure 4.1: A rectangle of the 2D discretization.

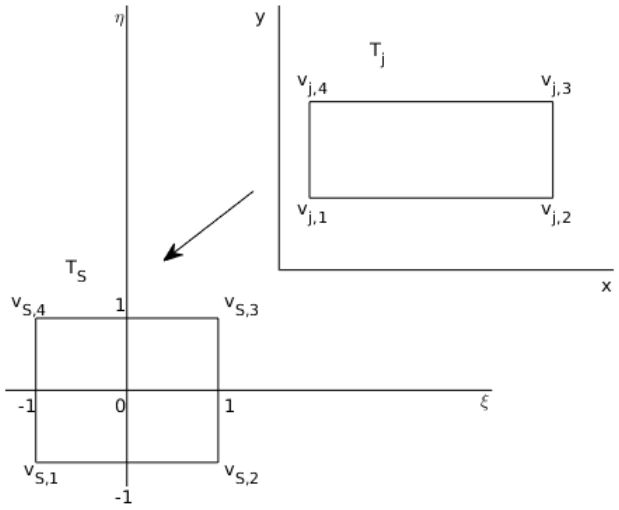


Figure 4.2: Transformation from T_j to T_S.

The differentiations are related by $\begin{pmatrix} dx \\ dy \end{pmatrix} = J_j \begin{pmatrix} d\xi \\ d\eta \end{pmatrix}$ with the Jacobian

$$|J_j| = \left|\frac{\partial(x,y)}{\partial(\xi,\eta)}\right| = \begin{vmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{vmatrix} = \begin{vmatrix} \frac{h_1}{2} & 0 \\ 0 & \frac{h_2}{2} \end{vmatrix} = \frac{1}{4} h_1 h_2.$$

4.2 2D Basis Functions

Let $N \in \mathbb{N}_0$ and

$$P_{N,T_j} = \operatorname{span}\left\{\xi^p \eta^q; 0 \leqslant p, q, p+q \leqslant N \text{ and } (\xi,\eta) = R_j(x,y) \text{ with } (x,y) \in T_j\right\},$$

be the space of the piecewise polynomials of degree N in two dimensions. The dimension of this space is given by

$$d_N := \dim(P_{N,T_j}) = \frac{(N+1)(N+2)}{2}.$$

For example we have with $(\xi,\eta) \in T_S$, $P_{2,T_j} = \text{span}\{1,\xi,\eta,\xi^2,\xi\eta,\eta^2\}$ with $d_2 = 6$.

Using the linear transformations (4.1) and recalling the basis functions Φ defined in (2.8), we define basis functions $\Psi_{i,j} \in P_{N,T_j}$ as

$$\Psi_{i,j}(x,y) := \Phi_{p,j}(x)\Phi_{q,j}(y) = \mathcal{L}_p(\xi(x))\mathcal{L}_q(\eta(y)), \tag{4.2}$$

where $\mathcal{L}_p$, $\mathcal{L}_q$ are the Legendre polynomials defined in Section 2.2.2, and

$$\xi \in [-1,1], \quad \eta \in [-1,1], \quad 0 \leqslant p \leqslant N, \quad 0 \leqslant q \leqslant N, \quad 0 \leqslant p+q \leqslant N,$$

$$i = \frac{(p+q)(p+q+1)}{2} + q, \quad 0 \leqslant i \leqslant d_N - 1.$$

Note that we can find p and q for a given i. We set $m := \max\left\{k \in \mathbb{N};\ \frac{k(k+1)}{2} \leqslant i\right\}$, then we have $q = i - m$ and $p = m - q$.

If we denote the mass matrix B_N then $B_N \in \mathbb{R}^{d_N \times d_N}$. In Appendix A we give examples of these bases and their mass matrices. These basis functions belong to the space P_{N,T_j} and we have $P_{N,T_j} \subseteq L^2(T_j)$ and they are orthogonal on T_j with respect to the scalar product

$$\langle f,g\rangle_{T_j} = \iint_{T_j} f(x,y)g(x,y)dydx.$$

We will use for the functions $v \in L^2(T_j)$ the following norms

$$||v||_{L^2(T_j)} = \sqrt{\langle v,v\rangle_{T_j}} = \left(\iint_{T_j} v^2(x,y)dydx\right)^{\frac{1}{2}}, \quad ||v||_{L^1(T_j)} = \iint_{T_j} |v(x,y)|dydx.$$

We will also use the notations $\langle f,g\rangle_\Omega$, $||v||_{L^2(\Omega)}$, and $|v|_{W^{N+1,2}(\Omega)}$, as in Chapter 2 for the one dimensional case. The multiple integrals over the rectangles T_j can be computed by using Gaussian quadrature rules over T_S. The integral of a function $f \in P_{N,T_j}$ over T_j is computed in the form

$$\iint_{T_j} f(x,y)dydx = |J_j| \int_{-1}^{1}\int_{-1}^{1} f(R_j^{-1}(\xi,\eta))d\eta d\xi = \frac{h_1h_2}{4}\int_{-1}^{1}\int_{-1}^{1} f(R_j^{-1}(\xi,\eta))d\eta d\xi.$$

The integral over T_S of any function $g \in P_{N,T_S}$ is evaluated numerically by the following Gaussian quadrature rule of order $N_G \in \mathbb{N}$

$$\int_{-1}^{1}\int_{-1}^{1} g(\xi,\eta)d\eta d\xi = \sum_{k=1}^{N_G}\sum_{\ell=1}^{N_G} \omega_k\omega_\ell g(\xi_k,\eta_\ell),$$

with the weights ω_k, ω_ℓ and the roots ξ_k, η_ℓ. For more details, see Szabo and Babuška [23].

4.3 The Projection

Given $v \in L^2(\Omega)$. We approximate v using piecewise polynomials finding $u \in P_{N,\Omega,K}$ in such a way that the error in L^2 norm is minimal. Let $u_j = u|_{T_j}$ with $j = 1, \ldots, K$. The piecewise polynomial u is given by the form

$$u(x,y) = \sum_{j=1}^{K} \sum_{i=0}^{d_N-1} \widehat{u}_{i,j} \Psi_{i,j}(x,y), \qquad (x,y) \in \Omega.$$

The coefficients $\widehat{u}_{i,j}$ must be computed in such a way that the error in the L^2 norm is minimal. The error is the difference $E(\widehat{u}_{i,j}, x, y) = v(x,y) - u(x,y)$. The error becomes minimal when the derivatives of its norm squared with respect to the unknowns $\widehat{u}_{k,j}$ vanish, i.e. $\frac{\partial}{\partial \widehat{u}_{k,j}} ||E||^2_{L^2(\Omega)} = 0$ for all $k = 0, \ldots, d_N - 1$ and $j = 1, \ldots, K$. We have

$$\begin{aligned} ||E||^2_{L^2(\Omega)} &= \iint_\Omega E^2 dydx = \sum_{j=1}^{K} \iint_{T_j} \left(E|_{T_j}(x,y)\right)^2 dydx \\ &= \sum_{j=1}^{K} \iint_{T_j} \left(v(x,y) - \sum_{i=0}^{d_N-1} \widehat{u}_{i,j} \Psi_{i,j}(x,y) \right)^2 dydx. \end{aligned}$$

Then we have for all $k = 0, \ldots, d_N - 1$ and $j = 1, \ldots, K$

$$\iint_{T_j} \left(\sum_{i=0}^{d_N-1} \widehat{u}_{i,j} \Psi_{i,j}(x,y) \right) \Psi_{k,j}(x,y) dydx = \iint_{T_j} v(x,y) \Psi_{k,j}(x,y) dydx.$$

This can be written as

$$\sum_{i=0}^{d_N-1} \widehat{u}_{i,j} \langle \Psi_{i,j}, \Psi_{k,j} \rangle_{T_j} = \langle v, \Psi_{k,j} \rangle_{T_j}.$$

According to (4.2) there are $p, p', q, q' \in \{0, \ldots, N\}$ where $\Psi_{k,j}(x,y) = \Phi_{p,j}(x)\Phi_{q,j}(y)$ and $\Psi_{i,j}(x,y) = \Phi_{p',j}(x)\Phi_{q',j}(y)$. Due to the orthogonality of the one dimensional basis functions we get

$$\begin{aligned} \langle \Psi_{i,j}, \Psi_{k,j} \rangle_{T_j} &= \iint_{T_j} \Psi_{i,j}(x,y) \Psi_{k,j}(x,y) dydx \\ &= \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \Phi_{p',j}(x)\Phi_{p,j}(x) dx \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \Phi_{q',j}(y)\Phi_{q,j}(y) dy = \frac{h_1}{2p+1} \frac{h_2}{2q+1} \delta_{pp'} \delta_{qq'}. \end{aligned}$$

Then the last sum in the left hand side reduces to a unique term with index $i = k$. Then, for all values $j = 1, \ldots, K$, we find the $K \times d_N$ solutions of our problem

$$\boxed{\widehat{u}_{k,j} = \frac{(2p+1)(2q+1)}{h_1 h_2} \langle v, \Psi_{k,j} \rangle_{T_j}, \qquad k = 0, \ldots, d_N - 1, \quad j = 1, \ldots, K.} \tag{4.3}$$

Properties

We define the operator $\Gamma_{N,K} : L^2(\Omega) \to P_{N,\Omega,K}$ by $u := \Gamma_{N,K}(v)$ to give

$$\Gamma_{N,K}(v)(x,y) = u(x,y) = \sum_{j=1}^{K} \sum_{i=0}^{d_N-1} \widehat{u}_{i,j} \Psi_{i,j}(x,y), \qquad \text{for } \ (x,y) \in \Omega,$$

where the coefficients $\widehat{u}_{i,j}$ are given by (4.3) depending on v. In the same way as the one dimensional case, we find that the operator $\Gamma_{N,K}$ is linear, orthogonal projection, best approximation, bounded. Also, for each $v \in W^{N+1,2}(\Omega)$, the following error estimates hold

$$\begin{aligned} ||\Gamma_{N,K}(v) - v||_{L^2(\Omega)} &\leqslant \frac{C_1}{2^{N+1}} \widehat{h}^{N+1} |v|_{W^{N+1,2}(\Omega)}, \\ ||\Gamma_{N,K}(v) - v||_{L^1(\Omega)} &\leqslant C_2 \sqrt{h_1 h_2}\, \widehat{h}^{N+1} |v|_{W^{N+1,2}(\Omega)}, \end{aligned}$$

where $\widehat{h} := \max\{h_1, h_2\}$.

Chapter 5

The Reconstruction of Higher Order Polynomials

5.1 The Idea of the Reconstruction

The idea of the projection, as we have found in Chapter 2, is to project a given function v onto a scalar product space spanned by basis functions with the condition that the error in the L^2 norm is minimal in order to get a piecewise polynomial u of degree N. The projection is then an orthogonal projection. The idea of the reconstruction in the current chapter is somewhat similar and is to map the projection u onto a set of basis functions of higher order polynomials with the condition that the error is minimal over a stencil of elements in order to get a piecewise polynomial of a larger degree $M \geqslant N$.

The reconstruction is, in fact, a non-local map, where the solution does not depend only on one element, as in the projection, but on a stencil of several adjoining elements.

In this chapter we describe how we reconstruct polynomials of degree M from other polynomials of lower degree $N \leqslant M$. The reconstruction procedure is the main step in the $P_N P_M$ DG schemes which will be presented later.

In the same way as in Chapter 2 we find that the reconstruction is an application of an operator. The difference here is that this operator is applied to the projections of degree N obtained in Chapter 2.

As in the ENO and WENO schemes of Harten et al. [15], the key idea of the reconstruction starts with choosing a stencil, i.e. domain, from where the reconstruction will be computed. For those schemes a process of choosing the locally smoothest stencil is automatically done, such that the scheme will avoid discontinuities during the reconstruction as much as possible. On the other hand, for the reconstruction during a computation in this thesis, we fix the size and the shape of the stencil used. Nevertheless, we have different choices we make at the beginning of a computation.

5.2 Approximations by the Projection Polynomials

To define the reconstructed polynomial we need some preliminaries. We start with discretizing the computational domain. Recall $Z \in \mathbb{N}$, $N, M \in \mathbb{N}_0$ satisfy $N \leqslant M$, $\Delta = \{x_{j+\frac{1}{2}}\}_0^Z$ to be a grid of $Z+1$ equally distant points in $I = [a, b]$. Also recall the subintervals $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[$ and their midpoints x_j for $j = 1, \ldots, Z$, and h to be the constant mesh size.

5.2.1 The Reconstruction Stencil

A reconstruction stencil in one dimensional problems, is the union of the interval I_j with a finite number of its adjoining intervals. We build the stencil of L elements directly to the left of I_j, and R elements directly to the right of I_j, and I_j itself. The resulting interval has no gaps. The size of the stencil is denoted by $n_e = 1 + L + R$ with $R, L \geqslant 0$. We denote the stencil by $S_{I_j,n_e,L}$. This notation points out to the shape of the stencil and the location of I_j in the stencil. In general, we have with $R = n_e - (L+1)$, $S_{I_j,n_e,L} = \bigcup_{c=-L}^{R} I_{j+c}$. For example, we have

$$\begin{array}{ll} S_{I_j,2,0} = I_j \cup I_{j+1} & \\ S_{I_j,2,1} = I_{j-1} \cup I_j & \end{array}, \quad \begin{array}{ll} S_{I_j,3,0} = I_j \cup I_{j+1} \cup I_{j+2} & \\ S_{I_j,3,1} = I_{j-1} \cup I_j \cup I_{j+1} & \\ S_{I_j,3,2} = I_{j-2} \cup I_{j-1} \cup I_j & \end{array}, \quad \text{etc.}$$

Furthermore, to build the reconstruction polynomial on the whole interval I which has associated with it a partition of Z elements, we need Z stencils of the same size n_e and the same shape, i.e. the index L is fixed for all stencils. If we change the size or the shape or both we will get another polynomial.

5.2.2 The Necessity of Extra Elements

The two stencils $S_{I_1,n_e,L}$ and $S_{I_Z,n_e,L}$ associated to the first and the last elements in the partition cover their elements and some adjoining elements. Therefore, we have to add extra elements which are located outside of I either at the left side or at the right side or on both sides. At these external ghost elements we need to compute the projections, since their values are needed for the reconstruction. For this purpose we define the extended interval $I_{ex} := \bigcup_{j=1}^{Z} S_{I_j,n_e,L}$. It has $Z_{ex} := Z + L + R$ elements and $I \subset I_{ex}$.

Remark 5.1. Another possibility, that is not assumed here, would be to change the stencils as the boundary is approched in order to take elements from I only. For example, as the left boundary is approached one could add elements to the right of I_j in order to avoid ghost elements and to maintain the number of elements n_e.

5.2.3 The Projection

Let $v \in L^2(I_{ex})$, $u \in P_{N,I_{ex},Z_{ex}}$ be the projection of degree N of v, and $S_{I_j,n_e,L}$, with $L \in \{0, \ldots, n_e - 1\}$, be a stencil related to some element I_j. To build one term of the reconstruction polynomial, using this stencil, we have to use n_e terms of u. According to (2.12) and (2.13)

the function $u \in P_{N,I_{ex},Z_{ex}}$ is given by $u(x) = \sum_{j=1}^{Z_{ex}} u_j(x)$ with $u_j(x) = \sum_{i=0}^{N} \widehat{u}_{i,j}\Phi_{i,j}(x)$ for $x \in I_j$ where for $i = 0, \ldots, N$ we have $\widehat{u}_{i,j} = \frac{2i+1}{h}\int_{I_j} v(x)\Phi_{i,j}(x)dx$. The functions $\Phi_{i,j}$ are the orthogonal basis functions given by (2.8).

5.2.4 The Reconstruction Polynomial

Suppose that $v \in L^2(I_{ex})$ is any given function and $u \in P_{N,I_{ex},Z_{ex}}$ is the projection piecewise polynomial of degree N of v. The reconstructed polynomial is a piecewise polynomial $w \in P_{M,I,Z}$ of degree $M \geqslant N$ built using the projection u and is written in the form $w = \sum_{j=1}^{Z} w_j$ and its terms are given by

$$w_j(x) = \sum_{i=0}^{M} \widehat{w}_{i,j}\Phi_{i,j}(x), \qquad x \in I. \tag{5.1}$$

5.3 Computing the Coefficients

5.3.1 The Conditions of Computing $\widehat{w}_{i,j}$

The reconstruction must satisfy the following three conditions.

(C1) Sufficient Number of Equations

The stencil must give a sufficient number of conditions. By finding the derivative of the norm of the error, we get for each element in the stencil $N+1$ equations. Then we get $n_e(N+1)$ equations and their number must satisfy $n_e(N+1) \geqslant M+1$. Then the following condition

$$n_e \geqslant \frac{M+1}{N+1}, \tag{5.2}$$

must be satisfied when choosing any stencil. We will consider the cases $n_e = \frac{M+1}{N+1}$ and $n_e > \frac{M+1}{N+1}$. The third case $n_e < \frac{M+1}{N+1}$, which generates an under-determined system, will be neglected.

(C2) The Extension of the Terms w_j

The terms w_j can be extended to the whole stencil associated to I_j as follows. The basis functions $\Phi_{i,j}$ are originally defined on the interval I_j and their value is set equal to zero on $I \backslash I_j$. Since they are polynomials on I_j, these polynomials are also defined everywhere on $\mathbb{R}$. In the following we say that we consider the extensions $\Phi^e_{i,j}$ of the $\Phi_{i,j}$ when they are considered to be the same polynomials outside of the interval I_j. For example

$$\Phi_{1,j}(x) = \begin{cases} \frac{2}{h}(x - x_j) & x \in I_j \\ 0 & x \in I \backslash I_j \end{cases}, \quad \Phi^e_{1,j}(x) = \begin{cases} \frac{2}{h}(x - x_j) & x \in S_{I_j,n_e,L} \\ 0 & x \in I_{ex} \backslash S_{I_j,n_e,L}. \end{cases}$$

Given a stencil $S_{I_j,n_e,L}$ associated to I_j with size n_e which consists of L left elements and R right elements and I_j itself, and we set $c = -L, \ldots, R$. Let $i = 0, \ldots, M$. The extension w^e_j

of the term w_j onto $I_{j+c} = [x_{j-\frac{1}{2}} + ch, x_{j+\frac{1}{2}} + ch[$ is defined by $w_j^e(y) = \sum_{i=0}^{M} \widehat{w}_{i,j}\Phi_{i,j}^e(y)$ for $y \in I_{j+c}$. Note that we keep the index j of $\widehat{w}_{i,j}$ without changing. For example, as shown in Figure 5.1, with the stencil $S_{I_j,3,1}$, we have

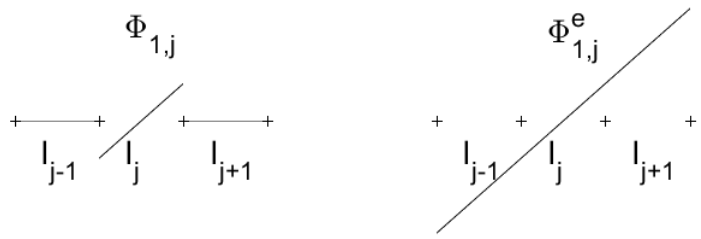


Figure 5.1: The basis function $\Phi_{1,j}$ (left) and its extension $\Phi_{1,j}^e$ (right) onto the stencil $S_{I_j,3,1}$.

(C3) Minimal Errors

The error in L^2 norm of computing each term must be minimal on the stencil. The error is the difference $E(\widehat{u}_{i,j}, y) = u(y) - w(y)$. Let $j = 1, \ldots, Z$ and $c = -L, \ldots, R$. We have for $y \in I_{j+c}$ and $E_{j+c} = E|_{I_{j+c}}$

$$E_{j+c}(\widehat{u}_{i,j+c}, y) = u_{j+c}(y) - w_j^e(y) = \sum_{i=0}^{N} \widehat{u}_{i,j+c}\Phi_{i,j+c}(y) - \sum_{\ell=0}^{M} \widehat{w}_{\ell,j}\Phi_{\ell,j}^e(y),$$

and then the least squares function is given by

$$||E_{j+c}||^2_{L^2(I_{j+c})} = \int_{I_{j+c}} \left(\sum_{i=0}^{N} \widehat{u}_{i,j+c}\Phi_{i,j+c}(y) - \sum_{\ell=0}^{M} \widehat{w}_{\ell,j}\Phi_{\ell,j}^e(y) \right)^2 dy.$$

The error becomes minimal when the derivatives of its norm squared with respect to the unknowns $\widehat{u}_{k,j+c}$ vanish, i.e. $\frac{\partial}{\partial \widehat{u}_{k,j+c}} ||E_{j+c}||^2_{L^2(I_{j+c})} = 0$ for all $k = 0, \ldots, N$, $c = -L, \ldots, R$, and $j = 1, \ldots, Z$. Since $(\int \cdot dx)' = \int (\cdot)' dx$ for a parameter then we get for $k = 0, \ldots, N$ and $c = -L, \ldots, R$ the following normal equations

$$\begin{aligned}
&\Rightarrow \quad 2\int_{I_{j+c}} \left(\sum_{i=0}^{N} \widehat{u}_{i,j+c}\Phi_{i,j+c}(y) - \sum_{\ell=0}^{M} \widehat{w}_{\ell,j}\Phi_{\ell,j}^e(y) \right) [-\Phi_{k,j+c}(y)] dy = 0, \\
&\Rightarrow \quad \int_{I_{j+c}} \left(\sum_{\ell=0}^{M} \widehat{w}_{\ell,j}\Phi_{\ell,j}^e(y) \right) \Phi_{k,j+c}(y) dy = \int_{I_{j+c}} \left(\sum_{i=0}^{N} \widehat{u}_{i,j+c}\Phi_{i,j+c}(y) \right) \Phi_{k,j+c}(y) dy, \\
&\Rightarrow \quad \sum_{\ell=0}^{M} \widehat{w}_{\ell,j} \langle \Phi_{\ell,j}^e, \Phi_{k,j+c} \rangle_{j+c} = \sum_{i=0}^{N} \widehat{u}_{i,j+c} \langle \Phi_{i,j+c}, \Phi_{k,j+c} \rangle_{j+c}.
\end{aligned}$$

The orthogonality holds in I_{j+c} for the functions $\Phi_{i,j+c}$, and according to (2.11) we have $\langle \Phi_{i,j+c}, \Phi_{k,j+c} \rangle_{j+c} = \frac{h}{2k+1}\delta_{ik}$. This leads to the system

$$\sum_{\ell=0}^{M} \widehat{w}_{\ell,j} \langle \Phi^e_{\ell,j}, \Phi_{k,j+c} \rangle_{j+c} = \frac{h}{2k+1} \widehat{u}_{k,j+c}. \tag{5.3}$$

This system consists of $n_e(N+1)$ equations and is obtained due to the minimal condition (C3) for the errors.

Remark 5.2. Due to the orthogonality of the Legendre basis functions we obtain the equalities

$$\widehat{w}_{i,j} = \widehat{u}_{i,j}, \quad i = 0, \dots, N, \quad j = 1, \dots, Z. \tag{5.4}$$

Consequently, the reconstructed polynomial w can be written as

$$w(x) = u(x) + \sum_{j=1}^{Z} \sum_{i=N+1}^{M} \widehat{w}_{i,j} \Phi_{i,j}(x), \quad x \in I.$$

Moreover, in the special case $M = N$, the equalities (5.4) cover all $\widehat{w}_{i,j}$ for $i = 0, \dots, M$ and we have $w = u|_I$. Then the $P_N P_N$ DG schemes are equivalent to the classical DG schemes, see [7].

5.3.2 The Matrix Form

We define the vectors $\widehat{\mathbf{w}}_j := (\widehat{w}_{0,j}, \dots, \widehat{w}_{M,j})^T \in \mathbb{R}^{M+1}$ and $\widehat{\mathbf{u}}_{j+c} := (\widehat{u}_{0,j+c}, \dots, \widehat{u}_{N,j+c})^T \in \mathbb{R}^{N+1}$ and the matrices

$$\mathbf{M}_{j,c} := \begin{pmatrix} \langle \Phi^e_{0,j}, \Phi_{0,j+c} \rangle_{j+c} & \dots & \langle \Phi^e_{M,j}, \Phi_{0,j+c} \rangle_{j+c} \\ \langle \Phi^e_{0,j}, \Phi_{1,j+c} \rangle_{j+c} & \dots & \langle \Phi^e_{M,j}, \Phi_{1,j+c} \rangle_{j+c} \\ \vdots & \vdots & \vdots \\ \langle \Phi^e_{0,j}, \Phi_{N,j+c} \rangle_{j+c} & \dots & \langle \Phi^e_{M,j}, \Phi_{N,j+c} \rangle_{j+c} \end{pmatrix} \in \mathbb{R}^{(N+1)\times(M+1)}.$$

$$\mathbf{A}_{j+c} = \begin{pmatrix} h & 0 & \dots & 0 \\ 0 & \frac{h}{3} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{h}{2N+1} \end{pmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}.$$

Then, for the system (5.3), we can write the following elemental matrix forms

$$\mathbf{M}_{j,c} \cdot \widehat{\mathbf{w}}_j = \mathbf{A}_{j+c} \cdot \widehat{\mathbf{u}}_{j+c} \quad \text{for} \quad c = -L, \dots, R. \tag{5.5}$$

Taking $\mathbf{y}_{j,c} := \mathbf{A}_{j+c} \cdot \widehat{\mathbf{u}}_{j+c}$ these forms become $\mathbf{M}_{j,c} \cdot \widehat{\mathbf{w}}_j = \mathbf{y}_{j,c}$ for $c = -L, \dots, R$. Also defining vectors $\mathbf{y}_j$ and matrices $\mathbf{M}_j$ by $\mathbf{y}_j := (\mathbf{y}_{j,-L}, \dots, \mathbf{y}_{j,R})^T \in \mathbb{R}^{n_e(N+1)}$ and $\mathbf{M}_j := (\mathbf{M}_{j,-L}, \dots, \mathbf{M}_{j,R})^T \in \mathbb{R}^{n_e(N+1)\times(M+1)}$ we can merge the elemental matrix forms into the following full matrix form

$$\mathbf{M}_j \cdot \widehat{\mathbf{w}}_j = \mathbf{y}_j, \tag{5.6}$$

which is related to the stencil $S_{I_j,n_e,L}$.

5.3.3 The Rank of the Reconstruction

We depend on the following Lemma to prove the existence of the solution of the system (5.6).

Lemma 5.3. Suppose that $(p_n)_{n\in\mathbb{N}_0}$ is a sequence of orthogonal polynomials on the interval $[a,b]$ with p_n of degree n. Then, for each $k \in \{0,\dots,n\}$, the polynomial p_k has k simple zeros that lie in $]a,b[$.

Proof. We consider p_n with the zeros $x_1^n,\dots,x_n^n \in \mathbb{C}$. We have $p_0 = c \neq 0$ and

$$0 = \langle p_0, p_n\rangle = c\int_a^b (x-x_1^n)\dots(x-x_n^n)dx.$$

This means that p_n must have at least one real zero, e.g. x_*, in $]a,b[$, at which the polynomial p_n changes its sign. This zero must have an odd multiplicity. Let $\Psi = \{x \in]a,b[: x \in \{x_1^n,\dots,x_n^n\}$ with odd multiplicities $\}$. Then we know that $\Psi \neq \emptyset$, since at least $x_* \in \Psi$. We set $\pi(x) := \prod_{t\in\Psi}(x-t)$. The function π has only simple zeros in $]a,b[$, since it is a product of different linear factors. Then the function $p_n\cdot\pi$ has in $]a,b[$ real zeros with even multiplicities only. This implies that $p_n\cdot\pi$ has no sign change on $]a,b[$. Thus we obtain $\langle p_n,\pi\rangle \neq 0$. Now we assume that $\pi \in P_\ell$ with $\ell < n$, i.e. $\pi = \sum_{j=0}^{\ell} a_j p_j$. Then $\langle p_n,\pi\rangle = \sum_{j=0}^{\ell} a_j\langle p_j,p_n\rangle = 0$. This is a contradiction to $\langle p_n,\pi\rangle \neq 0$. This means that $\pi \in \operatorname{span}(p_n)$, thus $\pi = \lambda p_n$, for some $\lambda \in \mathbb{R}$. Consequently, p_n has only simple zeros all of which lie in $]a,b[$. □

Corollary 5.4. Suppose that on the interval $I = [a,b]$ we have a set $\{p_0,\dots,p_N\}$ of $N+1$ orthogonal polynomials with $p_k \in P_{k,I}$ for $k = 0\dots,N$. Suppose that $p \in P_{M,I}$ is a polynomial of degree $M > N$ which is orthogonal to all p_k. Then it follows, by an analogous proof as for Lemma 5.3, that p has at least $N+1$ different zeros on $]a,b[$.

Theorem 5.5. The matrix $\mathbf{M}_j$ has a full column rank $M+1$.

Proof. We consider the homogeneous matrix form $\mathbf{M}_j\cdot\widehat{\mathbf{w}}_j = \mathbf{0}$ of (5.6). This system means that the coefficient vector of the reconstruction polynomial w_j, see (5.1), satisfies

$$\mathbf{M}_{j,c}\cdot\widehat{\mathbf{w}}_j = \begin{pmatrix} \langle w_j,\Phi_{0,j+c}\rangle_{j+c} \\ \vdots \\ \langle w_j,\Phi_{N,j+c}\rangle_{j+c}\end{pmatrix} = \mathbf{0}.$$

Therefore, the polynomial w_j of degree M is orthogonal to the $N+1$ basis functions $\Phi_{i,j+c}$ on all elements I_{j+c} of the stencil $S_{I_j,n_e,L}$, with $i = 0,\dots,N$ and $c = -L,\dots,R$. According to Corollary 5.4, there are at least $N+1$ different zeros of w on each element I_{j+c}. This gives $n_e(N+1)$ different zeros on the whole stencil. Also according to the condition (5.2) we find $n_e(N+1) \geqslant (M+1) > M$. Therefore, w is the zero polynomial. This proves the injectivity of the reconstruction and implies that the matrix $\mathbf{M}_j$ has the full column rank $M+1$. □

5.3.4 The Solution of the Reconstruction Problem

For $c = 0$, the equations (5.3) directly give the equalities (5.4), due to the orthogonality of the basis functions on I_j. Thus we can ignore the equations related to I_j in the system (5.3). We now consider the corresponding reduced system. Defining vectors $\widehat{\mathbf{y}}_j$ and matrices $\widetilde{\mathbf{M}}_j$ by

$$\widehat{\mathbf{y}}_j := (\mathbf{y}_{j,-L}, \ldots, \mathbf{y}_{j,-1}, \mathbf{y}_{j,1}, \ldots, \mathbf{y}_{j,R})^T \in \mathbb{R}^{(n_e-1)(N+1)},$$
$$\widetilde{\mathbf{M}}_j := (\mathbf{M}_{j,-L}, \ldots, \mathbf{M}_{j,-1}, \mathbf{M}_{j,1}, \ldots, \mathbf{M}_{j,R})^T \in \mathbb{R}^{(n_e-1)(N+1)\times(M+1)},$$

then we get the following reduced system $\widetilde{\mathbf{M}}_j \cdot \widehat{\mathbf{w}}_j = \widehat{\mathbf{y}}_j$.

The vector $\widehat{\mathbf{w}}_j$ can be divided into two vectors, $\widehat{\mathbf{u}}_j := (\widehat{u}_{0,j}, \ldots, \widehat{u}_{N,j})^T \in \mathbb{R}^{N+1}$ of the known coefficients and $\widehat{\mathbf{x}}_j := (\widehat{w}_{N+1,j}, \ldots, \widehat{w}_{M,j})^T \in \mathbb{R}^{M-N}$ of unknown coefficients. Moreover, the first $N+1$ columns in each matrix $\widetilde{\mathbf{M}}_j$ are related to the known coefficients. Thus the matrices $\widetilde{\mathbf{M}}_j$ can be divided into two parts, in the form $\widetilde{\mathbf{M}}_j = \left(\widetilde{\mathbf{M}}_{j,1}, \widetilde{\mathbf{M}}_{j,2}\right)$ where $\widetilde{\mathbf{M}}_{j,1} \in \mathbb{R}^{(n_e-1)(N+1)\times(N+1)}$ and $\widetilde{\mathbf{M}}_{j,2} \in \mathbb{R}^{(n_e-1)(N+1)\times(M-N)}$. We can rewrite the reduced system as follows $\left(\widetilde{\mathbf{M}}_{j,1}, \widetilde{\mathbf{M}}_{j,2}\right) \cdot \begin{pmatrix} \widehat{\mathbf{u}}_j \\ \widehat{\mathbf{x}}_j \end{pmatrix} = \widehat{\mathbf{y}}_j$, or

$$\widetilde{\mathbf{M}}_{j,2} \cdot \widehat{\mathbf{x}}_j = \widehat{\mathbf{y}}_j - \widetilde{\mathbf{M}}_{j,1} \cdot \widehat{\mathbf{u}}_j. \tag{5.7}$$

Since the matrix $\widetilde{\mathbf{M}}_{j,2}$ is a sub matrix from $\mathbf{M}_j$ and $\mathbf{M}_j$ has, according to Theorem 5.5, a full column rank, then $\widetilde{\mathbf{M}}_{j,2}$ has also a full column rank. Thus we conclude the following cases:

1. If $n_e = \frac{M+1}{N+1}$, then the matrix $\widetilde{\mathbf{M}}_{j,2}$ is square and invertible. We get the following unique solution
$$\widehat{\mathbf{x}}_j = \widetilde{\mathbf{M}}_{j,2}^{-1} \cdot \left(\widehat{\mathbf{y}}_j - \widetilde{\mathbf{M}}_{j,1} \cdot \widehat{\mathbf{u}}_j\right). \tag{5.8}$$

2. If $n_e > \frac{M+1}{N+1}$, then according to the least squares method[1], see e.g. Strang [22, p. 200], we consider $\widetilde{\mathbf{M}}_{j,2}^T \cdot \widetilde{\mathbf{M}}_{j,2}$ which is invertible. Moreover, the system (5.7) is over-determined. Now with $\mathbf{A} = \widetilde{\mathbf{M}}_{j,2}$, $\widetilde{\mathbf{x}} = \widehat{\mathbf{x}}_j$, and $\mathbf{b} = \widehat{\mathbf{y}}_j - \widetilde{\mathbf{M}}_{j,1} \cdot \widehat{\mathbf{u}}_j$, the normal equations are
$$\left(\widetilde{\mathbf{M}}_{j,2}^T \widetilde{\mathbf{M}}_{j,2}\right) \widehat{\mathbf{x}}_j = \widetilde{\mathbf{M}}_{j,2}^T \left(\widehat{\mathbf{y}}_j - \widetilde{\mathbf{M}}_{j,1} \cdot \widehat{\mathbf{u}}_j\right),$$
and the least squares solution is given by
$$\widehat{\mathbf{x}}_j = \left(\widetilde{\mathbf{M}}_{j,2}^T \cdot \widetilde{\mathbf{M}}_{j,2}\right)^{-1} \cdot \widetilde{\mathbf{M}}_{j,2}^T \cdot \left(\widehat{\mathbf{y}}_j - \widetilde{\mathbf{M}}_{j,1} \cdot \widehat{\mathbf{u}}_j\right). \tag{5.9}$$

3. If $n_e < \frac{M+1}{N+1}$, we ignore this case where the system (5.7) is under-determined.[2]

[1] For an overdetermined problem $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$, $m > n$, and rank$\mathbf{A} = n$, the quadratic minimization problem $\widetilde{\mathbf{x}} = \min_{\mathbf{x}} ||\mathbf{Ax} - \mathbf{b}||_e^2$ with the Euclidean norm (2.18) has a unique solution, provided that the n columns of $\mathbf{x}$ are linearly independent, given by solving the normal equations $\left(\mathbf{A}^T\mathbf{A}\right)\widetilde{\mathbf{x}} = \mathbf{A}^T\mathbf{b}$. Moreover, the least squares solution is given by $\widetilde{\mathbf{x}} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}$. For details see e.g. Strang [22].

[2] One could consider the solution of smallest Euclidean norm which is given by $\widehat{\mathbf{x}}_j = \widetilde{\mathbf{M}}_{j,2}^T \cdot \left(\widetilde{\mathbf{M}}_{j,2} \cdot \widetilde{\mathbf{M}}_{j,2}^T\right)^{-1} \cdot \left(\widehat{\mathbf{y}}_j - \widetilde{\mathbf{M}}_{j,1} \cdot \widehat{\mathbf{u}}_j\right)$. For details see Strang [22, p. 405].

5.3.5 The Solutions as Linear Combinations

We conclude from the formulas (5.8) and (5.9) that for one fixed element I_j each of the coefficients $\widehat{w}_{i,j}$ for $i = 0, \ldots, M$ of the term w_j can be written as a linear combination of all coefficients $\widehat{u}_{k,j+c}$ with $k = 0, \ldots, N$ and $c = -L, \ldots, R$. That means that there are constants $c_{i,k,j+c} \in \mathbb{R}$ such that

$$\widehat{w}_{i,j} = \sum_{c=-L}^{R} \sum_{k=0}^{N} c_{i,k,j+c} \widehat{u}_{k,j+c}. \tag{5.10}$$

We define, for $i = 0, \ldots, M$ and $j = 1, \ldots, Z$, the vectors

$$\mathbf{c}_i := (c_{i,0,j-L}, c_{i,1,j-L}, \ldots, c_{i,N,j-L}, c_{i,0,j-L+1}, \ldots, c_{i,N,j-L+1}, \ldots, c_{i,N,j+R}) \in \mathbb{R}^{n_e(N+1)},$$

$$\widehat{\mathbf{u}}_{j,s} := (\widehat{u}_{0,j-L}, \widehat{u}_{1,j-L}, \ldots, \widehat{u}_{N,j-L}, \widehat{u}_{0,j-L+1}, \ldots, \widehat{u}_{N,j-L+1}, \ldots, \widehat{u}_{N,j+R})^T \in \mathbb{R}^{n_e(N+1)}.$$

The vectors $\mathbf{c}_i$ are identical for different j. They only depend on the form of the stencil. With these vectors we can write $\widehat{w}_{i,j} = \mathbf{c}_i \cdot \widehat{\mathbf{u}}_{j,s}$. Also by defining the matrix $\mathbf{C} := (\mathbf{c}_0, \ldots, \mathbf{c}_M)^T \in \mathbb{R}^{(M+1)\times n_e(N+1)}$, we can write

$$\widehat{\mathbf{w}}_j = \mathbf{C} \cdot \widehat{\mathbf{u}}_{j,s}. \tag{5.11}$$

In the same way, as we studied the matrix $\widetilde{\mathbf{M}}_{j,2}$, we have the following cases:

1. If $n_e > \frac{M+1}{N+1}$, then the matrix $\mathbf{C}^T\mathbf{C}$ is invertible, it is positive definite.
2. If $n_e = \frac{M+1}{N+1}$, then $\mathbf{C}$ is invertible, and thus the product $\mathbf{C}^T\mathbf{C}$ is positive definite.

Whether the matrix $\mathbf{C}$ is square or rectangle, the product $\mathbf{C}^T\mathbf{C}$ will always be positive definite. Then, by using the Euclidean norm $||\cdot||_e$ defined in (2.18) and the spectral norm[3] $||\mathbf{C}||_2 = \sqrt{\beta_{\max}(\mathbf{C}^T\mathbf{C})}$ we have

$$||\widehat{\mathbf{w}}_j||_e^2 \leqslant ||\mathbf{C}||_2^2 ||\widehat{\mathbf{u}}_{j,s}||_e^2 = ||\mathbf{C}||_2^2 \left(||\widehat{\mathbf{u}}_{j-L}||_e^2 + \ldots + ||\widehat{\mathbf{u}}_{j+R}||_e^2 \right). \tag{5.12}$$

The last inequality is needed for proving the boundedness later.

5.4 Motivating Examples

In this section we present some examples of the reconstruction solutions with various stencils up to degree $M = 3$. We will consider one element I_j of the discretization and give formulas only of the term w_j which associated with I_j.

[3] Let $\mathbf{Q} \in \mathbb{R}^{n\times m}$. The spectral norm of $\mathbf{Q}$ is the largest singular value of $\mathbf{Q}$, i.e. the square root of the largest eigenvalue $\beta_{\max}(\mathbf{Q}^T\mathbf{Q})$ of the positive semidefinite matrix $\mathbf{Q}^T\mathbf{Q}$

$$||\mathbf{Q}||_2 = \sqrt{\beta_{\max}(\mathbf{Q}^T\mathbf{Q})}.$$

5.4.1 $N = 0,\ M = 1$

The projection on $I_{j+c} \in S_{I_j,n_e,L}$ is constant and is given by $u_{j+c}(x) = \widehat{u}_{0,j+c}\Phi_{0,j+c}(x) = \widehat{u}_{0,j+c}$ for $x \in I_{j+c}$ and $c = -L, \ldots, R$, and the reconstructed polynomial related to I_j is given by $w_j(x) = u_j(x) + \widehat{w}_{1,j}\Phi_{1,j}(x)$ for $x \in I_j$.

5.4.1.1 Stencils of 2-Elements

Let $S_{I_j,2,1} = I_{j-1} \cup I_j$, then for $c = -1$ the system (5.7) has only one equation and gives $(-2h)\widehat{w}_{1,j} = h(\widehat{u}_{0,j-1} - \widehat{u}_{0,j})$, then the unique solution is

$$\widehat{w}_{1,j} = \frac{1}{2}(\widehat{u}_{0,j} - \widehat{u}_{0,j-1}), \tag{5.13}$$

and thus for $x \in I$ we have

$$w_j(x) = \widehat{u}_{0,j} + \frac{1}{2}(\widehat{u}_{0,j} - \widehat{u}_{0,j-1})\frac{2(x - x_j)}{h} = \widehat{u}_{0,j} - \frac{x_j}{h}(\widehat{u}_{0,j} - \widehat{u}_{0,j-1}) + \frac{1}{h}(\widehat{u}_{0,j} - \widehat{u}_{0,j-1})x.$$

Similarly, for $S_{I_j,2,0} = I_j \cup I_{j+1}$, and, for $c = 1$ the system (5.7) has also only one equation and gives $(2h)\widehat{w}_{1,j} = h(\widehat{u}_{0,j+1} - \widehat{u}_{0,j})$, and the unique solution is

$$\widehat{w}_{1,j} = \frac{1}{2}(\widehat{u}_{0,j+1} - \widehat{u}_{0,j}), \tag{5.14}$$

and thus for $x \in I$ we have

$$w_j(x) = \widehat{u}_{0,j} + \frac{1}{2}(\widehat{u}_{0,j+1} - \widehat{u}_{0,j})\frac{2(x - x_j)}{h} = \widehat{u}_{0,j} - \frac{x_j}{h}(\widehat{u}_{0,j+1} - \widehat{u}_{0,j}) + \frac{1}{h}(\widehat{u}_{0,j+1} - \widehat{u}_{0,j})x.$$

5.4.1.2 Stencils of 3-Elements

Let $S_{I_j,3,2} = I_{j-2} \cup I_{j-1} \cup I_j$. Then for $c = -2, -1$, the system (5.7) has two equations and gives

$$\begin{pmatrix} -4h \\ -2h \end{pmatrix} \widehat{w}_{1,j} = \begin{pmatrix} h\widehat{u}_{0,j-2} \\ h\widehat{u}_{0,j-1} \end{pmatrix} - \begin{pmatrix} h \\ h \end{pmatrix} \widehat{u}_{0,j} \quad \Rightarrow \quad \begin{pmatrix} 4 \\ 2 \end{pmatrix} \widehat{w}_{1,j} = \begin{pmatrix} \widehat{u}_{0,j} - \widehat{u}_{0,j-2} \\ \widehat{u}_{0,j} - \widehat{u}_{0,j-1} \end{pmatrix}.$$

By applying the least squares approach, where $(4, 2)\begin{pmatrix} 4 \\ 2 \end{pmatrix} = 20$, we get

$$\widehat{w}_{1,j} = \frac{1}{20}(4, 2)\begin{pmatrix} \widehat{u}_{0,j} - \widehat{u}_{0,j-2} \\ \widehat{u}_{0,j} - \widehat{u}_{0,j-1} \end{pmatrix} = \frac{1}{10}(3\widehat{u}_{0,j} - \widehat{u}_{0,j-1} - 2\widehat{u}_{0,j-2}). \tag{5.15}$$

Similarly, we get

$$\text{for} \quad S_{I_j,3,1} = I_{j-1} \cup I_j \cup I_{j+1} \ : \quad \widehat{w}_{1,j} = \frac{1}{4}(\widehat{u}_{0,j+1} - \widehat{u}_{0,j-1}), \tag{5.16}$$

$$\text{for} \quad S_{I_j,3,0} = I_j \cup I_{j+1} \cup I_{j+2} \ : \quad \widehat{w}_{1,j} = \frac{1}{10}(2\widehat{u}_{0,j+2} + \widehat{u}_{0,j+1} - 3\widehat{u}_{0,j}). \tag{5.17}$$

5.4.2 $N = 0,\ M = 2$

The reconstructed polynomial related to I_j is given by $w_j(x) = u_j(x) + \widehat{w}_{1,j}\Phi_{1,j}(x) + \widehat{w}_{2,j}\Phi_{2,j}(x)$ for $x \in I_j$.

5.4.2.1 Stencils of 2-Elements, Neglected Case

Although this case does not verify the condition (5.2), but we present it for more explanation. Let $S_{I_j,2,1} = I_{j-1} \cup I_j$, then for $c = -1$ the system (5.7) has only one equation and gives

$$(-2h, 6h)\begin{pmatrix} \widehat{w}_{1,j} \\ \widehat{w}_{2,j} \end{pmatrix} = h(\widehat{u}_{0,j-1} - \widehat{u}_{0,j}) \quad \Rightarrow \quad -2\widehat{w}_{1,j} + 6\widehat{w}_{2,j} = \widehat{u}_{0,j-1} - \widehat{u}_{0,j}.$$

This gives an equation of two variable, thus we have infinite number of solutions.

5.4.2.2 Stencils of 3-Elements

$$S_{I_j,3,2} \ \Rightarrow \ \widehat{w}_{1,j} = \frac{1}{4}(\widehat{u}_{0,j-2} - 4\widehat{u}_{0,j-1} + 3\widehat{u}_{0,j}),\ \widehat{w}_{2,j} = \frac{1}{12}(\widehat{u}_{0,j-2} - 2\widehat{u}_{0,j-1} + \widehat{u}_{0,j}), \quad (5.18)$$

$$S_{I_j,3,1} \ \Rightarrow \ \widehat{w}_{1,j} = \frac{1}{4}(\widehat{u}_{0,j+1} - \widehat{u}_{0,j-1}), \qquad \widehat{w}_{2,j} = \frac{1}{12}(\widehat{u}_{0,j-1} - 2\widehat{u}_{0,j} + \widehat{u}_{0,j+1}), \quad (5.19)$$

$$S_{I_j,3,0} \ \Rightarrow \ \widehat{w}_{1,j} = \frac{1}{4}(4\widehat{u}_{0,j+1} - 3\widehat{u}_{0,j} - \widehat{u}_{0,j+2}),\ \widehat{w}_{2,j} = \frac{1}{12}(\widehat{u}_{0,j} - 2\widehat{u}_{0,j+1} + \widehat{u}_{0,j+2}).$$

5.4.3 $N = 0,\ M = 3$

The stencils of 2- or 3-elements do not verify the condition (5.2). We view stencils of 4-elements

$$S_{I_j,4,3} \quad \begin{cases} \widehat{w}_{1,j} = \frac{1}{120}(-19\widehat{u}_{0,j-3} + 87\widehat{u}_{0,j-2} - 177\widehat{u}_{0,j-1} + 109\widehat{u}_{0,j}), \\ \widehat{w}_{2,j} = \frac{1}{12}(-\widehat{u}_{0,j-3} + 4\widehat{u}_{0,j-2} - 5\widehat{u}_{0,j-1} + 2\widehat{u}_{0,j}), \\ \widehat{w}_{3,j} = \frac{1}{120}(-\widehat{u}_{0,j-3} + 3\widehat{u}_{0,j-2} - 3\widehat{u}_{0,j-1} + \widehat{u}_{0,j}). \end{cases}$$

$$S_{I_j,4,2} \quad \begin{cases} \widehat{w}_{1,j} = \frac{1}{120}(11\widehat{u}_{0,j-2} - 63\widehat{u}_{0,j-1} + 33\widehat{u}_{0,j} + 19\widehat{u}_{0,j+1}), \\ \widehat{w}_{2,j} = \frac{1}{12}(\widehat{u}_{0,j-1} - 2\widehat{u}_{0,j} + \widehat{u}_{0,j+1}), \\ \widehat{w}_{3,j} = \frac{1}{120}(-\widehat{u}_{0,j-2} + 3\widehat{u}_{0,j-1} - 3\widehat{u}_{0,j} + \widehat{u}_{0,j+1}). \end{cases}$$

$$S_{I_j,4,1} \quad \begin{cases} \widehat{w}_{1,j} = \frac{1}{120}(-19\widehat{u}_{0,j-1} - 33\widehat{u}_{0,j} + 63\widehat{u}_{0,j+1} - 11\widehat{u}_{0,j+2}), \\ \widehat{w}_{2,j} = \frac{1}{12}(\widehat{u}_{0,j-1} - 2\widehat{u}_{0,j} + \widehat{u}_{0,j+1}), \\ \widehat{w}_{3,j} = \frac{1}{120}(-\widehat{u}_{0,j-1} + 3\widehat{u}_{0,j} - 3\widehat{u}_{0,j+1} + \widehat{u}_{0,j+2}). \end{cases}$$

$$S_{I_j,4,0} \quad \begin{cases} \widehat{w}_{1,j} = \frac{1}{120}(-109\widehat{u}_{0,j} + 177\widehat{u}_{0,j+1} - 87\widehat{u}_{0,j+2} + 19\widehat{u}_{0,j+3}), \\ \widehat{w}_{2,j} = \frac{1}{12}(2\widehat{u}_{0,j} - 5\widehat{u}_{0,j+1} + 4\widehat{u}_{0,j+2} - \widehat{u}_{0,j+3}), \\ \widehat{w}_{3,j} = \frac{1}{120}(-\widehat{u}_{0,j} + 3\widehat{u}_{0,j+1} - 3\widehat{u}_{0,j+2} + \widehat{u}_{0,j+3}). \end{cases}$$

5.4.4 $N = 1,\ M = 2$

The projection on I_{j+c} is linear and is given by $u_{j+c}(x) = \widehat{u}_{0,j+c} + \widehat{u}_{1,j+c}\Phi_{1,j+c}(x)$ for $x \in I_{j+c}$ and $c = -L, \ldots, R$. The reconstructed polynomial related to I_j is given by $w_j(x) = u_j(x) + \widehat{w}_{2,j}\Phi_{2,j}(x)$ for $x \in I_j$. We consider stencils of 2-elements. For $S_{I_j,2,1}$ and $c = -1$ the system (5.7) has two equations and gives $\begin{pmatrix} 6h \\ -2h \end{pmatrix} \widehat{w}_{2,j} = \begin{pmatrix} h\widehat{u}_{0,j-1} \\ \frac{h}{3}\widehat{u}_{1,j-1} \end{pmatrix} - \begin{pmatrix} h & -2h \\ 0 & \frac{1}{3}h \end{pmatrix} \begin{pmatrix} \widehat{u}_{0,j} \\ \widehat{u}_{1,j} \end{pmatrix}$, then we have $\begin{pmatrix} 6 \\ -2 \end{pmatrix} \widehat{w}_{2,j} = \begin{pmatrix} \widehat{u}_{0,j-1} - \widehat{u}_{0,j} + 2\widehat{u}_{1,j} \\ \frac{1}{3}\widehat{u}_{1,j-1} - \frac{1}{3}\widehat{u}_{1,j} \end{pmatrix}$. By applying the least squares approach, where $(6, -2)\begin{pmatrix} 6 \\ -2 \end{pmatrix} = 40$, we get

$$\widehat{w}_{2,j} = \frac{1}{40}(6, -2) \begin{pmatrix} \widehat{u}_{0,j-1} - \widehat{u}_{0,j} + 2\widehat{u}_{1,j} \\ \frac{1}{3}\widehat{u}_{1,j-1} - \frac{1}{3}\widehat{u}_{1,j} \end{pmatrix} = \frac{1}{60}(9\widehat{u}_{0,j-1} - \widehat{u}_{1,j-1} - 9\widehat{u}_{0,j} + 19\widehat{u}_{1,j}). \tag{5.20}$$

Similarly, we find, for $S_{I_j,2,0}$, $\widehat{w}_{2,j} = \frac{1}{60}(9\widehat{u}_{0,j+1} + \widehat{u}_{1,j+1} - 9\widehat{u}_{0,j} - 19\widehat{u}_{1,j})$.

5.4.5 $N = 1,\ M = 3$

$$S_{I_j,2,1} \quad \Rightarrow \quad \begin{cases} \widehat{w}_{2,j} = \frac{1}{24}(15(\widehat{u}_{0,j-1} - \widehat{u}_{0,j}) + 11\widehat{u}_{1,j-1} + 19\widehat{u}_{1,j}), \\ \widehat{w}_{3,j} = \frac{1}{8}(\widehat{u}_{0,j-1} - \widehat{u}_{0,j} + \widehat{u}_{1,j-1} + \widehat{u}_{1,j}). \end{cases}$$

$$S_{I_j,2,0} \quad \Rightarrow \quad \begin{cases} \widehat{w}_{2,j} = \frac{1}{24}(15(\widehat{u}_{0,j+1} - \widehat{u}_{0,j}) - 11\widehat{u}_{1,j+1} - 19\widehat{u}_{1,j}), \\ \widehat{w}_{3,j} = \frac{1}{8}(\widehat{u}_{0,j} - \widehat{u}_{0,j+1} + \widehat{u}_{1,j} + \widehat{u}_{1,j+1}). \end{cases}$$

5.4.6 $N = 2,\ M = 3$

$$S_{I_j,2,1} \quad \Rightarrow \quad \widehat{w}_{3,j} = \frac{1}{4410}(165(\widehat{u}_{0,j} - \widehat{u}_{0,j-1}) + 25\widehat{u}_{1,j-1} - 355\widehat{u}_{1,j} - 3\widehat{u}_{2,j-1} + 1143\widehat{u}_{2,j}).$$

$$S_{I_j,2,0} \quad \Rightarrow \quad \widehat{w}_{3,j} = \frac{1}{4410}(165(\widehat{u}_{0,j+1} - \widehat{u}_{0,j}) + 25\widehat{u}_{1,j+1} - 355\widehat{u}_{1,j} + 3\widehat{u}_{2,j+1} - 1143\widehat{u}_{2,j}).$$

5.5 Analytical Study

Let $M, N \in \mathbb{N}_0$ and $S_{I_j,n_e,L}$ be a stencil whose size satisfies the condition $n_e \geqslant \frac{M+1}{N+1}$. We define an operator $\Re_{N,M,S,Z} : P_{N,I_{ex},Z_{ex}} \to P_{M,I,Z}$ by $\Re_{N,M,S,Z}(u) := w$ to give

$$\Re_{N,M,S,Z}(u)(x) = w(x) = u(x) + \sum_{j=1}^{Z} \sum_{i=N+1}^{M} \widehat{w}_{i,j}\Phi_{i,j}(x), \quad \text{for } x \in I. \tag{5.21}$$

We call this operator the *reconstruction* operator. We have for $j = 1, \ldots, Z$

$$\Re_{N,M,S,Z}(u)|_{I_j}(x) = w|_{I_j}(x) = w_j(x) = u_j(x) + \sum_{i=N+1}^{M} \widehat{w}_{i,j}\Phi_{i,j}(x), \quad \text{for } x \in I.$$

We can define this operator in another form

$$\Re_{N,M,S,Z} : L^2(I_{ex}) \to P_{M,I,Z}, \quad \text{with } \Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(v)) = w,$$

where $\Pi_{N,Z_{ex}}$ is the projection operator, which is defined in (2.14), of a function $v \in L^2(I_{ex})$.

Remark 5.6. The reconstruction operator $\Re_{N,M,S,Z}$ depends not only on the orders M and N and the mesh size h, but also on the stencil $S_{I_j,n_e,L}$ which is chosen. Thus, if we change one of these four factors we will get a new operator. For example, if we fix the mesh size h and take $M = 1$ and $N = 0$, then we still need to determine the stencil. There is a wide variety of choices of stencils which are available, but their sizes have to satisfy the condition (5.2), i.e. here $n_e \geqslant 2$. Two of these choices are shown as follows

- With $S = S_{I_j,2,1}$ we obtain the operator $\Re_{0,1,S,Z}$, whose solutions are (5.13).
- With $S = S_{I_j,2,0}$ we obtain the operator $\Re_{0,1,S,Z}$, whose solutions are (5.14).

The operator $\Re_{N,M,S,Z}$ has the following properties.

5.5.1 Linearity

The reconstruction operator is linear. This means that, for all $p, q \in P_{N,I_{ex},Z_{ex}}$ and $\varrho \in \mathbb{R}$, $\Re_{N,M,S,Z}(p+q) = \Re_{N,M,S,Z}(p) + \Re_{N,M,S,Z}(q)$ and $\Re_{N,M,S,Z}(\varrho p) = \varrho\Re_{N,M,S,Z}(p)$.

Proof. In (5.21) the reconstruction operator is written as the projection, which is linear, and a correction. Due to (5.10) the coefficients of the correction $\widehat{w}_{i,j}$ depend linearly on u. Therefore, the reconstruction operator is linear. □

5.5.2 Conservativity

The following conservation property $\langle \Re_{N,M,S,Z}(u), u|_I \rangle = \langle u|_I, u|_I \rangle$ holds for all $u \in P_{N,I_{ex},Z_{ex}}$ where $\langle \cdot, \cdot \rangle$ is the scalar product on $L^2(I)$ which was defined in (2.1).

Proof. Let $w = \Re_{N,M,S,Z}(u)$. We use the equalities (5.4) which hold for the terms w_j and u_j on the elements I_j. We have using the orthogonality of the basis functions $\langle w, u|_I \rangle = \sum_{j=1}^{Z} \langle w_j, u_j \rangle_j = \sum_{j=1}^{Z} \langle u_j, u_j \rangle_j = \langle u|_I, u|_I \rangle$ where $\langle \cdot, \cdot \rangle_j$ is the L^2 scalar product on the element I_j, defined in (2.9). □

5.5.3 Consistency for $p \in P_M$ and Identity

We prove that in the special case when the given function v is a polynomial $p \in P_M = P_{M,\mathbb{R}}$ of degree M then the system (5.6), $\mathbf{M}_j \cdot \widehat{\mathbf{w}}_j = \mathbf{y}_j$ has a consistent right hand side.

Theorem 5.7. Let $p \in P_M$, $u = \Pi_{N,Z_{ex}}(p)$ and $w = \Re_{M,M,S,Z}(u)$. The system (5.6) has a consistent right hand side. This means that for $p \in P_M$ the least squares solution is an exact solution of the linear system (5.6) in the overdetermined case.

Proof. Since $p \in P_M$ and the extended polynomials $\Phi^e_{0,j}, \ldots, \Phi^e_{M,j}$ are a basis of P_M, there exist constants $\widehat{p}_{0,j}, \cdots, \widehat{p}_{M,j} \in \mathbb{R}$ such that $p(x) = \sum_{m=0}^{M} \widehat{p}_{m,j} \Phi^e_{m,j}(x)$. For $k = 0, \ldots, N$ and $c = -L, \ldots, R$ let us consider the entries of the system $\mathbf{A}_{j+c} \cdot \widehat{\mathbf{u}}_{j+c} = \mathbf{M}_{j,c} \cdot \widehat{\mathbf{w}}_j$ defined in (5.5). Using (2.12) we obtain for a row of the system

$$\begin{aligned} \frac{h}{2k+1} \widehat{u}_{k,j+c} &= \int_{I_{j+c}} \Phi_{k,j+c}(x) p(x) dx = \sum_{m=0}^{M} \widehat{p}_{m,j} \int_{I_{j+c}} \Phi_{k,j+c}(x) \Phi^e_{m,j}(x) dx \\ &= \sum_{m=0}^{M} \widehat{p}_{m,j} \langle \Phi_{k,j+c}, \Phi^e_{m,j} \rangle_{j+c}. \end{aligned}$$

This is the k-th row of the equation $\mathbf{y}_{j,c} = \mathbf{A}_{j+c} \cdot \widehat{\mathbf{u}}_{j+c} = \mathbf{M}_{j,c} \cdot \widehat{\mathbf{p}}_j$ for $\widehat{\mathbf{p}}_j := (\widehat{p}_{0,j}, \ldots, \widehat{p}_{M,j})^T$. This means that $\mathbf{y}_{j,c}$ is in the range of $\mathbf{M}_{j,c}$ for all $c = -L, \ldots, R$. This holds also for the system (5.6), i.e. the vector $\mathbf{y}_j$ is in the range of $\mathbf{M}_j$. Thus the system (5.6) has a consistent right hand side. □

Lemma 5.8. Let $p \in P_M$. For all $N \in \{0, \ldots, M\}$ and by using any stencil $S = S_{I_j, n_e, L}$ with size n_e satisfying $n_e \geqslant \frac{M+1}{N+1}$ we have $\Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(p)) = p|_I$. So $\Re_{N,M,S,Z} \circ \Pi_{N,Z_{ex}}$ is a quasi identity map for polynomials in P_M. It uses the known values of $p \in P_M$ on the extended interval I_{ex}.

Proof. Let $u = \Pi_{N,Z_{ex}}(p) \in P_{N,I_{ex},Z_{ex}}$ and $w = \Re_{M,M,S,Z}(u) \in P_{M,I,Z}$. Since P_M may be identified one to one with $P_{M,I_{ex}}$ the polynomial p can be expanded on I_{ex} piecewise. Then we have

$$p = \sum_{j=1}^{Z_{ex}} \sum_{i=0}^{M} \widehat{p}_{i,j} \Phi_{i,j}, \qquad u = \sum_{j=1}^{Z_{ex}} \sum_{k=0}^{N} \widehat{u}_{k,j} \Phi_{k,j}, \qquad w = \sum_{j=1}^{Z} \sum_{\ell=0}^{M} \widehat{w}_{\ell,j} \Phi_{\ell,j},$$

where all $\widehat{u}_{k,j}$ and $\widehat{p}_{i,j}$ are given by (2.12) and the $\widehat{w}_{\ell,j}$ are the solutions of the reconstruction equations. We have on the extended interval $\widehat{p}_{i,j} = \widehat{u}_{i,j}$ for $i = 0, \ldots, N$ and $j = 1, \ldots, Z_{ex}$ and on the original interval the reconstruction coefficients satisfy $\widehat{w}_{\ell,j} = \widehat{p}_{\ell,j} = \widehat{u}_{\ell,j}$ for $\ell = 0, \ldots, N$ and $j = 1, \ldots, Z$. We want to prove that $\widehat{w}_{\ell,j} = \widehat{p}_{\ell,j}$ for $\ell = N+1, \ldots, M$ and $j = 1, \ldots, Z$. By Theorem 5.5 the solution to the reconstruction system (5.6) is unique since $\mathbf{M}_j$ has full column rank. This means that for $p \in P_M$ we have $\widehat{p}_{\ell,j} = \widehat{w}_{\ell,j}$ for $\ell = N+1, \ldots, M$ and $j = 1, \ldots, Z$. □

5.5.4 Further Relations between $\Re_{N,M,S,Z}$ and $\Pi_{N,Z_{ex}}$

Theorem 5.9. Let $p, q \in P_{N,I_{ex},Z_{ex}}$ such that $\Re_{N,M,S,Z}(p) = \Re_{N,M,S,Z}(q)$. Then, we have $p|_I = q|_I$.

Proof. Let $P = \Re_{N,M,S,Z}(p) = \Re_{N,M,S,Z}(q) = Q$. Then we may write

$$\sum_{j=1}^{Z}\sum_{i=0}^{M} \widehat{P}_{i,j}\Phi_{i,j}(x) = \sum_{j=1}^{Z}\sum_{i=0}^{M} \widehat{Q}_{i,j}\Phi_{i,j}(x),$$

or $\sum_{j=1}^{Z}\sum_{i=0}^{M}(\widehat{P}_{i,j} - \widehat{Q}_{i,j})\Phi_{i,j}(x) = 0$. Since the piecewise polynomials $\Phi_{i,j}$ are linearly independent basis functions in $P_{M,I,Z}$ then we have $\widehat{P}_{i,j} = \widehat{Q}_{i,j}$ for all $i = 0, \ldots, M$ and $j = 1, \ldots, Z$. On the other hand, the equalities (5.4) give $\widehat{p}_{i,j} = \widehat{P}_{i,j}$ and $\widehat{q}_{i,j} = \widehat{Q}_{i,j}$ for all $i = 0, \ldots, N$ and $j = 1, \ldots, Z$. Thus we find $\widehat{p}_{i,j} = \widehat{q}_{i,j}$ for $i = 0, \ldots, N$ and $j = 1, \ldots, Z$. Then $p|_I = q|_I$. □

Theorem 5.10. For any stencil $S_{I_j,ne,L}$ with $n_e \geqslant \frac{M+1}{N+1}$, we have for any $u \in P_{N,I_{ex},Z_{ex}}$

$$\Pi_{N,Z}(\Re_{N,M,S,Z}(u)) = u|_I.$$

Proof. Let $w = \Re_{N,M,S,Z}(u)$, and $q = \Pi_{N,Z}(w)$. We want to prove that $q = u|_I$. We have

$$u = \sum_{j=1}^{Z_{ex}}\sum_{i=0}^{N} \widehat{u}_{i,j}\Phi_{i,j}, \qquad w = \sum_{j=1}^{Z}\sum_{i=0}^{M} \widehat{w}_{i,j}\Phi_{i,j}, \quad \text{and} \quad q = \sum_{j=1}^{Z}\sum_{i=0}^{N} \widehat{q}_{i,j}\Phi_{i,j}.$$

For $i = 0, \ldots, N$ and $j = 1, \ldots, Z$ according to (2.12), we have

$$\widehat{q}_{i,j} = \frac{2i+1}{h}\langle w, \Phi_{i,j}\rangle_j = \frac{2i+1}{h}\sum_{j=1}^{Z}\sum_{k=0}^{M} \widehat{w}_{k,j}\langle \Phi_{k,j}, \Phi_{i,j}\rangle_j.$$

Since the basis functions satisfy (2.11), we have $\widehat{q}_{i,j} = \frac{2i+1}{h}\left(\widehat{w}_{i,j}\frac{h}{2i+1}\right) = \widehat{w}_{i,j}$. According to the conservation property (5.4), we have $\widehat{w}_{i,j} = \widehat{u}_{i,j}$, then $\widehat{q}_{i,j} = \widehat{u}_{i,j}$ for $i = 0, \ldots, N$ and $j = 1, \ldots, Z$. This implies finally that $q = u|_I$. □

Theorem 5.11. For all $w \in P_{M,I_{ex},Z_{ex}}$ and all $N \in \{0, \ldots, M\}$, the relation

$$\Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(w)) = w|_I$$

holds

Proof. Follows directly from the Lemma 5.8. □

5.5.5 Boundedness

Theorem 5.12. Let $w = \Re_{N,M,S,Z}(u)$ and let w_j and u_j be the restrictions to I_j of w and u, respectively. Then the following inequalities hold

$$||u_j||^2_{L^2(I_j)} \leqslant ||w_j||^2_{L^2(I_j)} \leqslant C_5 \sum_{c=-L}^{R} ||u_{j+c}||^2_{L^2(I_{j+c})}. \tag{5.22}$$

Proof. **1.** We start with the first part of the inequality (5.22). From the conservation property 5.5.2 and from the Cauchy Schwarz inequality, we have

$$||u_j||^2_{L^2(I_j)} = \langle u_j, u_j\rangle_j = \langle w_j, u_j\rangle_j \leqslant ||u_j||_{L^2(I_j)}||w_j||_{L^2(I_j)}.$$

If $||u_j||_{L^2(I_j)} = 0$ then trivially $||u_j||^2_{L^2(I_j)} \leqslant ||w_j||^2_{L^2(I_j)}$. If $||u_j||_{L^2(I_j)} > 0$, then we get $||u_j||_{L^2(I_j)} \leqslant ||w_j||_{L^2(I_j)}$ and hence $||u_j||^2_{L^2(I_j)} \leqslant ||w_j||^2_{L^2(I_j)}$, thus the first part of inequality (5.22) follows.
2. We apply (2.19) to the term w_j, then we have by taking $\widehat{\mathbf{w}}_j = (\widehat{w}_{0,j}, \ldots, \widehat{w}_{M,j})^T$

$$\frac{h}{2M+1}||\widehat{\mathbf{w}}_j||^2_e \leqslant ||w_j||^2_{L^2(I_j)} \leqslant h||\widehat{\mathbf{w}}_j||^2_e. \tag{5.23}$$

On the other hand, according to the formula (5.11), we have $\widehat{\mathbf{w}}_j = \mathbf{C}\widehat{\mathbf{u}}_{j,s}$. From (5.12), we get

$$||\widehat{\mathbf{w}}_j||^2_e \leqslant ||\mathbf{C}||^2_2||\widehat{\mathbf{u}}_{j,s}||^2_e = ||\mathbf{C}||^2_2\left(||\widehat{\mathbf{u}}_{j-L}||^2_e + \ldots + ||\widehat{\mathbf{u}}_{j+R}||^2_e\right).$$

According to (2.20), we have $||\widehat{\mathbf{u}}_{j+c}||^2_e \leqslant \frac{1}{\underline{m}}||u_{j+c}||^2_{L^2(I_{j+c})}$ for $c = -L, \ldots, R$, and then

$$||\widehat{\mathbf{w}}_j||^2_e \leqslant \frac{||\mathbf{C}||^2_2}{\underline{m}}\left(||u_{j-L}||^2_{L^2(I_{j-L})} + \ldots + ||u_{j+R}||^2_{L^2(I_{j+R})}\right).$$

Substituting into (5.23) we obtain

$$||w_j||^2_{L^2(I_j)} \leqslant h\frac{||\mathbf{C}||^2_2}{\underline{m}}\left(||u_{j-L}||^2_{L^2(I_{j-L})} + \ldots + ||u_{j+R}||^2_{L^2(I_{j+R})}\right).$$

With noting that $\underline{m} = \frac{h}{2N+1}$, we find

$$||w_j||^2_{L^2(I_j)} \leqslant (2N+1)||\mathbf{C}||^2_2\left(||u_{j-L}||^2_{L^2(I_{j-L})} + \ldots + ||u_{j+R}||^2_{L^2(I_{j+R})}\right).$$

Finally, noting that the coefficients in $\mathbf{C}$ only depend on the basis polynomials and not on the v or w_j, and by taking $C_5 := (2N+1)||\mathbf{C}||^2_2$ we get the right inequality. □

5.5.6 The Error Estimates of the Reconstruction Operator

Theorem 5.13. Suppose that the interval $I = [a, b]$ has a uniform partition of Z subintervals with constant mesh size $h = (b-a)/Z$. Let $N \leqslant M$, $S = S_{I_j,ne,L}$ be a stencil with $n_e \geqslant \frac{M+1}{N+1}$, and $I_{ex} = \bigcup_{j=1}^{Z} S_{I_j,n_e,L}$ be the extended interval. Then, for each $v \in W^{M+1,2}(I_{ex})$, the following error estimates hold

$$||\Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(v)) - v|_I||_{L^2(I)} \leqslant C_6 h^{M+1} |v|_{W^{M+1,2}(I)},$$

where $|\cdot|_{W^{M+1,2}(I)}$ is the seminorm on $W^{M+1,2}(I)$ given in (2.4).

Proof. Let I_j be an element with $j = 1, \ldots, Z$ fixed. Using the triangle inequality, we obtain

$$\begin{aligned} ||\Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(v)) - v|_I||^2_{L^2(I_j)} &\leqslant ||\Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(v)) - \Pi_{M,Z_{ex}}(v)|_I||^2_{L^2(I_j)} \\ &\quad + ||\Pi_{M,Z_{ex}}(v)|_I - v|_I||^2_{L^2(I_j)}. \end{aligned} \tag{5.24}$$

Due to the identity in Lemma 5.8 and the linearity of the reconstruction operator we have

$$\begin{aligned} ||\Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(v)) - \Pi_{M,Z_{ex}}(v)|_I||^2_{L^2(I_j)} &= ||\Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(v)) - \Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(\Pi_{M,Z_{ex}}(v)))||^2_{L^2(I_j)} \\ &= ||\Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(v) - \Pi_{N,Z_{ex}}(\Pi_{M,Z_{ex}}(v)))||^2_{L^2(I_j)}. \end{aligned}$$

By virtue of inequality (5.22) and linearity as well as boundedness of the projection operator we obtain

$$\begin{aligned} ||\Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(v)) - \Pi_{M,Z_{ex}}(v)|_I||^2_{L^2(I_j)} &\leqslant C_5 \sum_{c=-L}^{R} ||(\Pi_{N,Z_{ex}}(v) - \Pi_{N,Z_{ex}}(\Pi_{M,Z_{ex}}(v)))|_{I_{j+c}}||^2_{L^2(I_{j+c})} \\ &= C_5 \sum_{c=-L}^{R} ||(\Pi_{N,Z_{ex}}(v - \Pi_{M,Z_{ex}}(v))|_{I_{j+c}}||^2_{L^2(I_{j+c})} \\ &= C_5 \sum_{c=-L}^{R} ||(v - \Pi_{M,Z_{ex}}(v))|_{I_{j+c}}||^2_{L^2(I_{j+c})}. \end{aligned}$$

Substituting into (5.24) we get

$$||\Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(v)) - v|_I||^2_{L^2(I_j)} \leqslant C_5 \sum_{c=-L}^{R} ||(v - \Pi_{M,Z_{ex}}(v))|_{I_{j+c}}||^2_{L^2(I_{j+c})} + ||\Pi_{M,Z_{ex}}(v)|_I - v|_I||^2_{L^2(I_j)}.$$

Now the error estimate (2.21) obtained in the proof of Theorem 2.5 gives

$$\begin{aligned} ||\Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(v)) - v|_I||^2_{L^2(I_j)} &\leqslant C_5 \sum_{c=-L}^{R} \left(C_2^2 h^{2M+2} |v|^2_{W^{M+1,2}(I_{j+c})} \right) + C_2^2 h^{2M+2} |v|^2_{W^{M+1,2}(I_j)} \\ &= (1 + n_e C_5) C_2^2 h^{2M+2} |v|^2_{W^{M+1,2}(I_j)}. \end{aligned}$$

By taking $C_6 = C_2\sqrt{1 + n_e C_5}$ and by summation over all j, we get

$$||\Re_{N,M,S,Z}(\Pi_{N,Z_{ex}}(v)) - v|_I||^2_{L^2(I)} \leqslant C_6^2 h^{2M+2} |v|^2_{W^{M+1,2}(I)}.$$

Finally, taking the square root, we obtain the result. □

Chapter 6

Examples of the Reconstruction

The aim of this chapter is to view examples of reconstructed polynomials and to study the numerical effect of the reconstruction operators. We apply them and study their accuracy and how do they increase the order of the solutions. We use the same examples considered in Chapter 3 and the same solutions computed there for the coefficients $\widehat{u}_{i,j}$.

6.1 The Function $v(x) = x - 1$

We have $\widehat{u}_{0,j} = x_j - 1$ and $\widehat{u}_{1,j} = \frac{h}{2}$.

- When $M = N = 0$, we get the same operator $\Pi_{0,Z}$. It is of first order, see Table 3.1.
- When $M = N = 1$, we get the identity operator.

For the operator $\Re_{0,1,S,Z}$ we first choose the stencil $S_{I_j,2,1}$. Using $\widehat{u}_{0,j-1} = x_{j-1} - 1 = x_j - h - 1$ and according to (5.13) we get $\widehat{w}_{0,j} = \widehat{u}_{0,j} = x_j - 1$ and $\widehat{w}_{1,j} = \frac{1}{2}(\widehat{u}_{0,j} - \widehat{u}_{0,j-1}) = \frac{h}{2}$, and the reconstructed polynomial is $w_j(x) = x_j - 1 + \frac{h}{2}\frac{2}{h}(x - x_j) = x - 1 = v(x)$. Thus this operator is the identity.

If we choose another stencil we will get the same result. For example, for $S_{I_j,3,2}$ and according to (5.15) we have

$$\widehat{w}_{1,j} = \frac{3(x_j - 1) - (x_{j-1} - 1) - 2(x_{j-2} - 1)}{10} = \frac{1}{10}(3x_j - 3 - (x_j - h - 1) - 2(x_j - 2h - 1)) = \frac{h}{2}.$$

6.2 The Function $v(x) = x^2 - 3x + 2$

We have $\widehat{u}_{0,j} = x_j^2 - 3x_j + \frac{h^2}{12} + 2$, $\widehat{u}_{1,j} = x_j h - \frac{3h}{2}$, and $\widehat{u}_{2,j} = \frac{h^2}{6}$.

- When $M = N = 0$, we get the same operator $\Pi_{0,Z}$. It is of first order, see Table 3.3.
- When $M = N = 1$, we get the same operator $\Pi_{1,Z}$. It is of second order, see Table 3.4.
- When $M > 1$, we get the identity operator.

6.2.1 $N = 0$, $M = 1$, with $S_{I_j,2,1}$

Using $\widehat{u}_{0,j-1} = x_{j-1}^2 - 3x_{j-1} + \frac{h^2}{12} + 2 = x_j^2 - (2h+3)x_j + \frac{13h^2}{12} + 3h + 2$ and according to (5.13), we find $\widehat{w}_{0,j} = \widehat{u}_{0,j} = x_j^2 - 3x_j + \frac{h^2}{12} + 2$ and $\widehat{w}_{1,j} = \frac{1}{2}(\widehat{u}_{0,j} - \widehat{u}_{0,j-1}) = hx_j - \frac{h^2}{2} - \frac{3h}{2}$ and

$$\begin{aligned} w_j &= x_j^2 - 3x_j + \frac{h^2}{12} + 2 + \left(hx_j - \frac{h^2}{2} - \frac{3h}{2}\right)\frac{2}{h}(x - x_j) \\ &= -x_j^2 + hx_j + \frac{h^2}{12} + 2 + (2x_j - h - 3)x. \end{aligned}$$

These polynomials are of second order of accuracy and the L^2 errors are equal to $||v-w||_{L^2(I)} = h^2\sqrt{4/15} = \mathcal{O}(h^2)$. We get by using Matlab the Table 6.1.
Example. We take the interval $]0, 3[$ and discretize it into 5 elements (with $Z = 5$) then we have $h = \frac{3}{5} = 0.6$ and the solution is given by

$$w_j = \frac{1}{125}\begin{cases} 610 - 375x & \text{for} \quad x \in [0\ \ , 0.6[, \\ 49 - 225x & \text{for} \quad x \in [0.6, 1.2[, \\ -44 - 75x & \text{for} \quad x \in [1.2, 1.8[, \\ 7 + 75x & \text{for} \quad x \in [1.8, 2.4[, \\ -122 + 225x & \text{for} \quad x \in [2.4, 3\ \ ]. \end{cases}$$

Figure 6.1 shows this solution.

6.2.2 $N = 0$, $M = 1$, with $S_{I_j,3,2}$

According to (5.15), we have $\widehat{w}_{1,j} = \frac{1}{10}(3\widehat{u}_{0,j} - \widehat{u}_{0,j-1} - 2\widehat{u}_{0,j-2}) = hx_j - \frac{9h^2}{10} - \frac{3h}{2}$, and the reconstructed polynomials are

$$\begin{aligned} w_j = \ & \widehat{u}_{0,j} = x_j^2 - 3x_j + \tfrac{h^2}{12} + 2 + \left(hx_j - \tfrac{9h^2}{10} - \tfrac{3h}{2}\right)\tfrac{2(x-x_j)}{h} \\ = \ & -x_j^2 + \tfrac{9h}{5}x_j + \tfrac{h^2}{12} + 2 + \left(2x_j - \tfrac{9h}{5} - 3\right)x. \end{aligned}$$

The L^2 errors are equal to $||v - w||_{L^2(I)} = h^2\sqrt{62/75} = \mathcal{O}(h^2)$, i.e. second order, see Table 6.1. Figure 6.2 shows the solution with $Z = 5$.

6.2.3 $N = 0$, $M = 1$, with $S_{I_j,3,1}$

According to (5.16), we have $\widehat{w}_{1,j} = \frac{1}{4}(\widehat{u}_{0,j+1} - \widehat{u}_{0,j-1}) = x_j h - \frac{3h}{2}$, and the reconstructed polynomials are $w_j = x_j^2 + \frac{h^2}{12} + x_j h\frac{2}{h}(x - x_j) = \frac{h^2}{12} - x_j^2 + 2x_j x$. The L^2 errors are equal to $||v - w||_{L^2(I)} = h^2\sqrt{1/60} = \mathcal{O}(h^2)$, i.e. second order, see Table 6.1. Figure 6.3 shows the solution with $Z = 5$.

6.2.4 $N = 0$, $M = 1$, with $S_{I_j,3,0}$

According to (5.17), we have $\widehat{w}_{1,j} = \frac{1}{10}(\widehat{u}_{0,j+1} - 3\widehat{u}_{0,j} + 2\widehat{u}_{0,j+2}) = hx_j + \frac{9h^2}{10} - \frac{3h}{2}$ and the reconstructed polynomials are $w_j = -x_j^2 - \frac{9h}{5}x_j + \frac{h^2}{12} + 2 + \left(2x_j + \frac{9h}{5} - 3\right)x$. The L^2 errors are the same as of the case with $S_{I_j,3,2}$.

6.2.5 $M > 1$

With any stencil the reconstruction operator will be the identity. For example we choose $N = 1$ and $M = 2$ with $S_{I_j,2,1}$. Then according to (5.20) we have $\widehat{w}_{0,j} = x_j^2 - 3x_j + \frac{h^2}{12} + 2$, $\widehat{w}_{1,j} = \widehat{u}_{1,j} = hx_j - \frac{3h}{2}$, $\widehat{w}_{2,j} = \frac{1}{12}(\widehat{u}_{0,j-1} - \widehat{u}_{0,j} + 3\widehat{u}_{1,j} - \widehat{u}_{1,j-1}) = \frac{h^2}{6}$. The reconstructed polynomials are

$$w_j = x_j^2 - 3x_j + \frac{h^2}{12} + 2 + \left(hx_j - \frac{3h}{2}\right)\frac{2(x - x_j)}{h} + \frac{h^2}{6}\left(\frac{3}{2}\left(\frac{2(x - x_j)}{h}\right)^2 - \frac{1}{2}\right) = x^2 - 3x + 2.$$

6.3 The Function $v(x) = x^3 - x$

We have $\widehat{u}_{0,j} = x_j^3 + \left(\frac{h^2}{4} - 1\right)x_j$, $\widehat{u}_{1,j} = \frac{3h}{2}x_j^2 + \frac{3h^3}{40} - \frac{h}{2}$, $\widehat{u}_{2,j} = \frac{h^2}{2}x_j$, and $\widehat{u}_{3,j} = \frac{h^3}{20}$.

- When $M = N$ with $N = 0, 1, 2$, we get the same operators $\Pi_{N,Z}$. They are of order $N+1$, see Table 3.5.
- When $M > 2$, we get the identity operator.

For $N = 0$, $M = 1$, with $S_{I_j,3,1}$, we have $||v - w||_{L^2(I)} = \sqrt{\frac{163}{420}h^6 + \frac{16}{15}h^4}$ and

$$w_j = -2x_j^3 - h^2x_j + \left(3x_j^2 + \frac{5h^2}{4} - 1\right)x.$$

For $N = 0$, $M = 2$, with $S_{I_j,3,1}$, we have $||v - w||_{L^2(I)} = \sqrt{\frac{17}{42}}h^3$ and

$$w_j = x_j^3 - \frac{5h^2}{4}x_j + \left(-3x_j^2 + \frac{5h^2}{4} - 1\right)x + 3x_jx^2.$$

For $N = 1$, $M = 2$, with $S_{I_j,2,1}$, we have $||v - w||_{L^2(I)} = \sqrt{\frac{663}{21875}}h^3$ and

$$w_j = x_j^3 - \frac{57h}{50}x_j^2 - \frac{3h^2}{20}x_j + \frac{19h^3}{200} + \left(-3x_j^2 + \frac{57h}{25}x_j + \frac{3h^2}{20} - 1\right)x + \left(3x_j - \frac{57h}{50}\right)x^2.$$

For example, we view the operator when $N = 0$ and $M = 2$ with $S_{I_j,3,1}$ and take the interval $[-2, 2]$. With $Z = 5$ then we have $h = \frac{4}{5} = 0.8$ and the solution is given by

$$w_j = \begin{cases} -1426864/78125 - 11837x/625 - 24x^2/5 & \text{for} \quad x \in [-2, -1.2[, \\ -152/3125 + 311x/125 - 12x^2/5 & \text{for} \quad x \in [-1.2, -0.4[, \\ -x/5 & \text{for} \quad x \in [-0.4, 0.4[, \\ 152/3125 + 311x/125 + 12x^2/5 & \text{for} \quad x \in [0.4, 1.2[, \\ 1426864/78125 - 11837x/625 + 24x^2/5 & \text{for} \quad x \in [1.2, 2]. \end{cases}$$

Figure 6.4 shows this solution and Table 6.2 views the L^2 errors.

6.4 The Function $v(x) = \sin(x)$

We consider the function $v(x) = \sin(x)$ defined on $I = [0, 2\pi]$. Table 6.3 views the L^2 errors of some operators. Figures 6.5 and 6.6 view two solutions.

Z	L^2 errors	EOC
8	0.072618	2
16	0.018155	2
32	0.004539	2
64	0.001135	2
	$\Uparrow$	
	$h^2\sqrt{4/15}$	

Z	L^2 errors	EOC
8	0.127858	2
16	0.031964	2
32	0.007991	2
64	0.001998	2
	$\Uparrow$	
	$h^2\sqrt{62/75}$	

Z	L^2 errors	EOC
8	0.018155	2
16	0.004539	2
32	0.001135	2
64	0.000284	2
	$\Uparrow$	
	$h^2\sqrt{1/60}$	

Table 6.1: The errors of computing the reconstructed polynomials for $v(x) = x^2 - 3x + 2$ when $N = 0$ and $M = 1$ with the stencils $S_{I_j,2,1}$ (left), $S_{I_j,3,2}$ (middle), $S_{I_j,3,1}$ (right).

Z	L^2 errors	EOC
8	0.079526	3
16	0.009941	3
32	0.001243	3
64	0.000155	3
	$\Uparrow$	
	$\sqrt{\frac{17}{42}}h^3$	

Table 6.2: The errors of computing the reconstructed polynomials for $v(x) = x^3 - x$ when $N = 0$ and $M = 2$ with $S_{I_j,3,1}$

Z	L^2 errors	EOC	L^2 errors	EOC	L^2 errors	EOC
	P_0P_1					
8	5.908358e-2	2.50				
16	1.160803e-2	2.35				
32	2.641818e-3	2.14				
64	6.427222e-4	2.04				
	P_0P_2		P_1P_2			
8	4.328446e-2	2.78	3.843646e-3	3.53		
16	5.619406e-3	2.95	3.810827e-4	3.33		
32	7.091082e-4	2.99	4.368701e-5	3.12		
64	8.884859e-5	3	5.327866e-6	3.04		
	P_0P_3		P_1P_3		P_2P_3	
8	1.677594e-2	3.67	2.756360e-3	3.84	1.808543e-4	4.52
16	1.109078e-3	3.92	1.772770e-4	3.96	9.223925e-6	4.29
32	7.029687e-5	3.98	1.115948e-5	3.99	5.368237e-7	4.10
64	4.408990e-6	3.99	6.987184e-7	4	3.289320e-8	4.03

Table 6.3: The errors of computing some reconstructed polynomials for $v(x) = \sin(x)$.

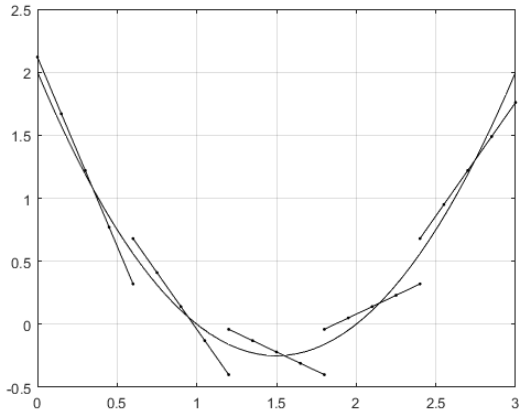


Figure 6.1: The operator $\Re_{0,1,S,5}$ for $v(x) = x^2 - 3x + 2$ with $S_{I_j,2,1}$.

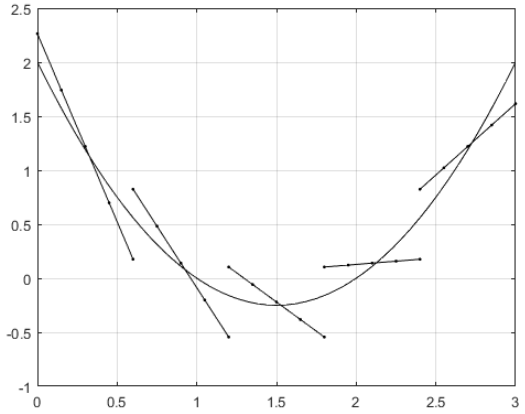


Figure 6.2: The operator $\Re_{0,1,S,5}$ for $v(x) = x^2 - 3x + 2$ with $S_{I_j,3,2}$.

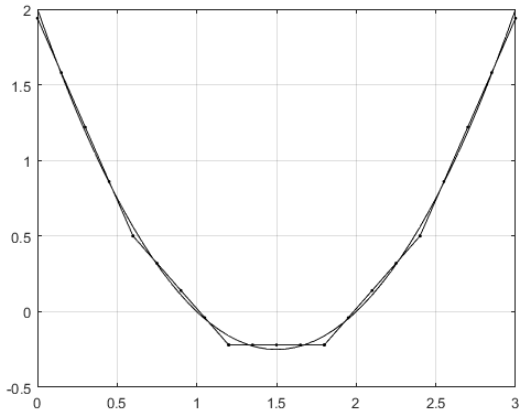


Figure 6.3: The operator $\Re_{0,1,S,5}$ for $v(x) = x^2 - 3x + 2$ with $S_{I_j,3,1}$.

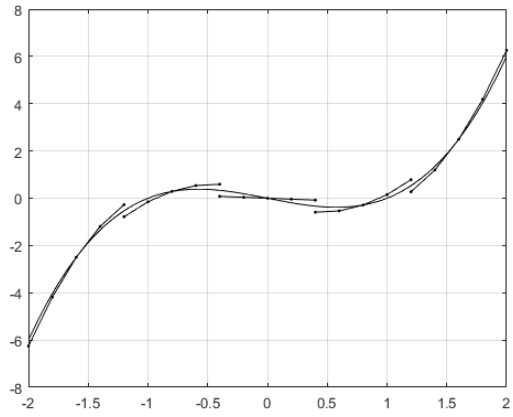


Figure 6.4: The operator $\Re_{0,2,S,5}$ for $v(x) = x^3 - x$ with $S_{I_j,3,1}$.

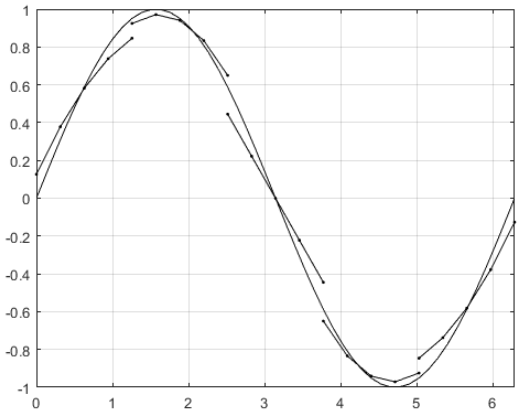


Figure 6.5: The operator $\Re_{0,2,S,5}$ for $v(x) = \sin(x)$ with $S_{I_j,3,1}$

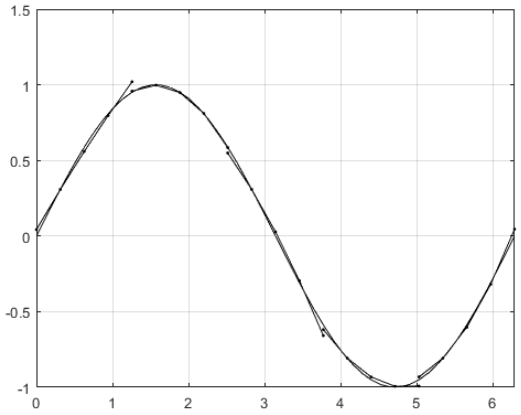


Figure 6.6: The operator $\Re_{1,2,S,5}$ for $v(x) = \sin(x)$ with $S_{I_j,2,1}$

Chapter 7

The Local Space Time Galerkin Scheme

The local space time Galerkin scheme is used as a predictor step for solving high order Riemann problems. This scheme evolves the reconstructed polynomials, produced by the reconstruction operators in Chapter 5, locally in time inside each element to the same order of accuracy as in space, by using the governing equations.

This time evolution is a part of the algorithm, proposed by Dumbser et al. [10], of the FV schemes for solving systems of hyperbolic balance laws with stiff source terms. In the same manner, another algorithm proposed by Dumbser et al. [7] uses this time evolution, namely, the $P_N P_M$ DG schemes. The suggestion of Dumbser of using this approach avoids the Cauchy-Kovalewski procedure used by Harten et al. [15] to evolve the data in time.

In this thesis we follow the procedure of Dumbser used in [7] for the $P_N P_M$ DG schemes. The process starts by choosing space time functions $\theta_{i,j}$ used as basis functions and as test functions where they are multiplied by the governing PDEs obtaining a weak form in space and time.

The solutions of the function, flux, and source terms are represented by linear combinations in terms of the basis functions $\theta_{i,j}$ weighted by coefficients. Some of these coefficients are known and computed with help of the basis functions $\Phi_{i,j}$ of degree M defined in (2.8). To compute the remaining coefficients we solve the weak form by replacing these formulas in the weak form obtaining a local system of linear algebraic equations. This local system is solved by an iteration scheme.

7.1 The Local Space Time Basis Functions

Let $M \in \mathbb{N}_0$, $T > 0$, $Z \in \mathbb{N}$, I_j be an element of a partition with constant mesh size h of the space interval $I = [a, b]$ with $j = 1, \ldots, Z$ fixed. Suppose that we have the times $0 = t_0 < t_1 < \ldots < t_{\max} = T$ with different time steps and let $\check{T}_n = [t_n, t_{n+1}[$, with $t_n < T$, be a time interval with time step $k_n = t_{n+1} - t_n$.

In the following we proceed according to Dumbser et al. [7]. We want to build basis functions $\theta_{i,j}$ according to the following properties:

1. They are from the space $P_{M,\breve{T}_n\times I_j}$, i.e. they are polynomials of the same degree M in the space and the time variables.

2. They take their value to be zero outside of $\breve{T}_n \times I_j$.

3. They are nodal functions. This means that we choose some nodes on $\breve{T}_n \times I_j$. Then we relate each node to a function such that this function equals to 1 at this node and equals to 0 at the others. The number of nodes should equal the number of degrees of freedom of these polynomials.

4. These functions will be used not only as basis functions to represent the solutions of the local Galerkin scheme, but also as test functions for finding the weak form of the local space time Galerkin scheme.

In the following, we present two examples of building these basis functions.

7.1.1 $M = 2$

The general form of a polynomial of degree 2 in space and time, which is zero out of I_j, is

$$\theta_{i,j}(t,x) = \begin{cases} a_1 + a_2x + a_3x^2 + a_4t + a_5xt + a_6t^2, & \text{if } x \in I_j, \\ 0 & \text{if } x \in I\backslash I_j, \end{cases} \qquad t \in \breve{T}_n,$$

with $a_1, \ldots, a_6 \in \mathbb{R}$. The number of coefficients is $(M+1)(M+2)/2 = 6$ therefore we need 6 conditions. We take the following 6 nodes

$$\begin{array}{lll} \beta_1 = (t_n, x_{j-\frac{1}{2}}) & \beta_2 = (t_n, x_j) & \beta_3 = (t_n, x_{j+\frac{1}{2}}) \\ \beta_4 = (t_{n+\frac{1}{2}}, x_{j-\frac{1}{2}}) & \beta_5 = (t_{n+\frac{1}{2}}, x_{j+\frac{1}{2}}) & \beta_6 = (t_{n+1}, x_j), \end{array}$$

see Figure 7.1-b. After that we relate to each node β_i one polynomial $\theta_{i,j}$ for $i = 1, \ldots, 6$. In total we have $6 \times 6 = 36$ conditions defined as follows

$$\theta_{i,j}(\beta_k) = \delta_{ik}, \quad \text{for } i,k = 1, \ldots, 6.$$

The node polynomial can be obtained as follows. Let $\theta_{1,j}$ be the polynomial of degree 2 related to the first node $\beta_1 = (t_n, x_{j-\frac{1}{2}})$. Then we have

$$\theta_{1,j}(t,x) = a_1 + a_2x + a_3x^2 + a_4t + a_5xt + a_6t^2.$$

The 6 conditions related to $\theta_{1,j}$ are listed as follows

$$\begin{aligned} \theta_{1,j}(t_n, x_{j-\frac{1}{2}}) &= a_1 + a_2x_{j-\frac{1}{2}} + a_3x_{j-\frac{1}{2}}^2 + a_4t_n + a_5x_{j-\frac{1}{2}}t_n + a_6t_n^2 = 1, \\ \theta_{1,j}(t_n, x_j) &= a_1 + a_2x_j + a_3x_j^2 + a_4t_n + a_5x_jt_n + a_6t_n^2 = 0, \\ \theta_{1,j}(t_n, x_{j+\frac{1}{2}}) &= a_1 + a_2x_{j+\frac{1}{2}} + a_3x_{j+\frac{1}{2}}^2 + a_4t_n + a_5x_{j+\frac{1}{2}}t_n + a_6t_n^2 = 0, \\ \theta_{1,j}(t_{n+\frac{1}{2}}, x_{j-\frac{1}{2}}) &= a_1 + a_2x_{j-\frac{1}{2}} + a_3x_{j-\frac{1}{2}}^2 + a_4t_{n+\frac{1}{2}} + a_5x_{j-\frac{1}{2}}t_{n+\frac{1}{2}} + a_6t_{n+\frac{1}{2}}^2 = 0, \\ \theta_{1,j}(t_{n+\frac{1}{2}}, x_{j+\frac{1}{2}}) &= a_1 + a_2x_{j+\frac{1}{2}} + a_3x_{j+\frac{1}{2}}^2 + a_4t_{n+\frac{1}{2}} + a_5x_{j+\frac{1}{2}}t_{n+\frac{1}{2}} + a_6t_{n+\frac{1}{2}}^2 = 0, \\ \theta_{1,j}(t_{n+1}, x_j) &= a_1 + a_2x_j + a_3x_j^2 + a_4t_{n+1} + a_5x_jt_{n+1} + a_6t_{n+1}^2 = 0. \end{aligned}$$

The matrix form is

$$\begin{pmatrix} 1 & x_{j-\frac{1}{2}} & x_{j-\frac{1}{2}}^2 & t_n & x_{j-\frac{1}{2}}t_n & t_n^2 \\ 1 & x_j & x_j^2 & t_n & x_j t_n & t_n^2 \\ 1 & x_{j+\frac{1}{2}} & x_{j-\frac{1}{2}}^2 & t_n & x_{j+\frac{1}{2}}t_n & t_n^2 \\ 1 & x_{j-\frac{1}{2}} & x_{j-\frac{1}{2}}^2 & t_{n+\frac{1}{2}} & x_{j-\frac{1}{2}}t_{n+\frac{1}{2}} & t_{n+\frac{1}{2}}^2 \\ 1 & x_{j+\frac{1}{2}} & x_{j+\frac{1}{2}}^2 & t_{n+\frac{1}{2}} & x_{j+\frac{1}{2}}t_{n+\frac{1}{2}} & t_{n+\frac{1}{2}}^2 \\ 1 & x_j & x_j^2 & t_{n+1} & x_j t_{n+1} & t_{n+1}^2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The coefficient matrix is invertible, since its determinant equals to $k_n^4 h^4/32$. Inverting this matrix and setting $\varsigma = \frac{2}{h}(x - x_j)$ and $\zeta = \frac{1}{k_n}(t - t_n)$, then we get the solution

$$\theta_{i,j}(\zeta(t), \varsigma(x)) = \begin{cases} -\frac{1}{2}\varsigma + \frac{1}{2}\varsigma^2 - 2\zeta + \varsigma\zeta + 2\zeta^2, & \text{if} \quad x \in I_j, \\ 0 & \text{if} \quad x \in I \backslash I_j, \end{cases} \qquad t \in \breve{T}_n.$$

In the same way we get for $x \in I_j$ and $t \in \breve{T}_n$

$$\begin{aligned} &\theta_{1,j}(\zeta(t), \varsigma(x)) = -\tfrac{1}{2}\varsigma + \tfrac{1}{2}\varsigma^2 - 2\zeta + \varsigma\zeta + 2\zeta^2, \quad \theta_{2,j}(\zeta(t), \varsigma(x)) = 1 - \varsigma^2 + \zeta - 2\zeta^2, \\ &\theta_{3,j}(\zeta(t), \varsigma(x)) = \tfrac{1}{2}\varsigma + \tfrac{1}{2}\varsigma^2 - 2\zeta - \zeta\varsigma + 2\zeta^2, \qquad \theta_{4,j}(\zeta(t), \varsigma(x)) = 2\zeta - \zeta\varsigma - 2\zeta^2, \\ &\theta_{5,j}(\zeta(t), \varsigma(x)) = 2\zeta + \zeta\varsigma - 2\zeta^2, \qquad\qquad\qquad \theta_{6,j}(\zeta(t), \varsigma(x)) = -\zeta + 2\zeta. \end{aligned} \tag{7.1}$$

We set the basis to be $\Theta_{2,j} = \{\theta_{1,j}, \theta_{2,j}, \theta_{3,j}, \theta_{4,j}, \theta_{5,j}, \theta_{6,j}\}$. Note that at $t = t_n$ we have

$$\theta_{4,j}(t_n, x) = \theta_{5,j}(t_n, x) = \theta_{6,j}(t_n, x) = 0, \quad \text{for all} \quad x \in I_j.$$

The other functions in $\Theta_{2,j}$ depend on the spatial points at $t = t_n$. The number of these other functions is $M + 1 = 3$, namely, $\theta_{1,j}, \theta_{2,j}$, and $\theta_{3,j}$.

7.1.2 $M = 1$

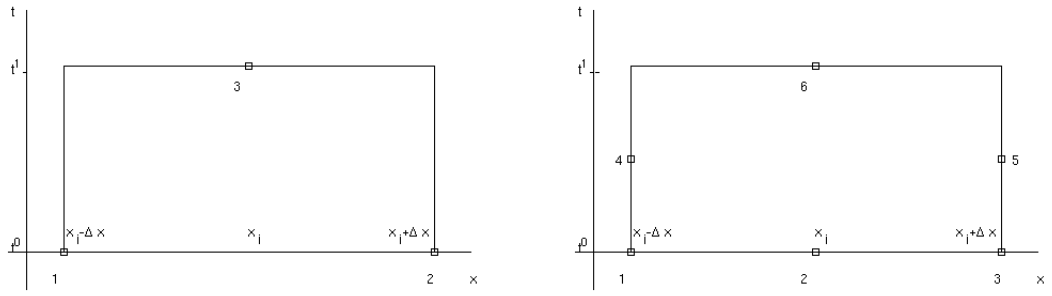
The general form of a polynomial of first degree in space and time, which is zero out of I_j, is

$$\theta_{i,j}(t, x) = \begin{cases} a_1 + a_2 x + a_3 t, & \text{if} \quad x \in I_j, \\ 0 & \text{if} \quad x \in I \backslash I_j, \end{cases} \qquad t \in \breve{T}_n.$$

The number of nodes is $(M + 1)(M + 2)/2 = 3$ which are $\beta_1 = (t_n, x_{j-\frac{1}{2}})$, $\beta_2 = (t_n, x_{j+\frac{1}{2}})$, and $\beta_3 = (t_{n+1}, x_j)$, see Figure 7.1-a. With $\varsigma = \frac{2(x-x_j)}{h}$ and $\zeta = \frac{t-t_n}{k_n}$ the functions are given in I_j by

$$\theta_{1,j}(\zeta(t), \varsigma(x)) = \frac{1}{2}(1 - \varsigma - \zeta), \quad \theta_{2,j}(\zeta(t), \varsigma(x)) = \frac{1}{2}(1 + \varsigma - \zeta), \quad \theta_{3,j}(\zeta(t), \varsigma(x)) = \zeta. \tag{7.2}$$

We set the basis to be $\Theta_{1,j} = \{\theta_{1,j}, \theta_{2,j}, \theta_{3,j}\}$. Note that at $t = t_n$ we have $\theta_{3,j}(t_n, x) = 0$ for $x \in I_j$. The other functions in $\Theta_{1,j}$ depend on the spatial points at $t = t_n$. The number of these other functions is $M + 1 = 2$, namely, $\theta_{1,j}$ and $\theta_{2,j}$.



(a) $M = 1, \mathcal{N} = 3$. (b) $M = 2, \mathcal{N} = 6$.

Figure 7.1: The nodes associated with the nodal functions of degrees $M = 1, 2$.

7.1.3 The General Case

For an arbitrary degree M, the number of nodes, which we denote by $\mathcal{N}$, is given by $\mathcal{N} := (M+1)(M+2)/2$. The nodes β_r with $r = 1, \ldots, \mathcal{N}$, except the last node, are chosen via the following distribution

$$\beta_r = (t_{i,k}, x_{i,k}) = \left(t_n + \frac{k}{M} k_n, x_{j-\frac{1}{2}} + \frac{i}{M-k} h \right), \qquad \begin{array}{l} k = 0, 1, \ldots, M-1, \\ i = 0, 1, \ldots, M-k, \\ r = 1, \ldots, \mathcal{N}, \quad \beta_r \neq (t_{n+1}, x_j). \end{array}$$

The last node is set to be $(t_{0,M}, x_{0,M}) = (t_{n+1}, x_j)$. According to these $\mathcal{N}$ nodes there will be $\mathcal{N}$ nodal space time functions $\theta_{i,j}$ for $i = 1, \ldots, \mathcal{N}$. They must be zero out of the element I_j. Moreover, each function must be associated with one node and is computed by solving the general formula of the space time polynomial of degree M at this node. Then we set the basis to be

$$\Theta_{M,j} = \{\theta_{1,j}, \ldots, \theta_{\mathcal{N},j}\}.$$

The first $M+1$ basis functions are taken to depend on the spatial variable x at $t = t_n$. They could be grouped together in a subbasis $\Theta^0_{M,j}$. All other basis functions vanish for $t = t_n$ and could be grouped together in a subbasis $\Theta^1_{M,j}$. This means $\Theta_{M,j} = \Theta^0_{M,j} \cup \Theta^1_{M,j}$.

7.2 The Formulas of the Solutions

The general form of the hyperbolic systems of balance laws in one dimension is given by

$$v_t(t,x) + f(v(t,x))_x = s(v(t,x)), \qquad \text{for} \quad x \in I, \quad t \in [0,T]. \tag{7.3}$$

We suppose that the solution v, the flux f, and the source term s using the local space time Galerkin scheme are approximated on $\breve{T}_n \times I$ by the formulas

$$\begin{aligned} v(t,x) &:= U^n(t,x) = \sum_{j=1}^{Z}\sum_{i=1}^{\mathcal{N}} \widehat{U}_{i,j}^n \theta_{i,j}(t,x), \\ f(v(t,x)) &:= F^n(t,x) = \sum_{j=1}^{Z}\sum_{i=1}^{\mathcal{N}} f(\widehat{U}_{i,j}^n)\theta_{i,j}(t,x), \quad \text{for } (t,x) \in \breve{T}_n \times I, \\ s(v(t,x)) &:= S^n(t,x) = \sum_{j=1}^{Z}\sum_{i=1}^{\mathcal{N}} s(\widehat{U}_{i,j}^n)\theta_{i,j}(t,x), \end{aligned} \tag{7.4}$$

where the coefficients $\widehat{U}_{i,j}^n \in \mathbb{R}$ are unknowns. According to the space discretization, these solutions have the following terms

$$U_j^n(t,x) = \sum_{i=1}^{\mathcal{N}} \widehat{U}_{i,j}^n \theta_{i,j}(t,x), \quad F_j^n(t,x) = \sum_{i=1}^{\mathcal{N}} f(\widehat{U}_{i,j}^n)\theta_{i,j}(t,x), \quad S_j^n(t,x) = \sum_{i=1}^{\mathcal{N}} s(\widehat{U}_{i,j}^n)\theta_{i,j}(t,x).$$

7.3 The Matrix Form

We multiply (7.3) by $\theta_{k,j}$ for $k = 1, \ldots, \mathcal{N}$ and integrate over $\breve{T}_n \times I$. Since the value of these test functions are taken to be zero outside $\breve{T}_n \times I_j$, thus we get for $k = 1, \ldots, \mathcal{N}$

$$\int_{\breve{T}_n}\int_{I_j} \theta_{k,j}(t,x) v_t(t,x) dxdt + \int_{\breve{T}_n}\int_{I_j} \theta_{k,j}(t,x) f(v(t,x))_x dxdt = \int_{\breve{T}_n}\int_{I_j} \theta_{k,j}(t,x) s(v(t,x)) dxdt.$$

Introducing the scalar product

$$\langle g, h\rangle_{tx} := \int_{\breve{T}_n}\int_{I_j} g(t,x)h(t,x)dxdt. \tag{7.5}$$

and using the previous definitions, the systems become for $k = 1, \ldots, \mathcal{N}$

$$\sum_{i=1}^{\mathcal{N}} \langle \theta_{k,j}, (\theta_{i,j})_t\rangle_{tx} \widehat{U}_{i,j}^n + \sum_{i=1}^{\mathcal{N}} \langle \theta_{k,j}, (\theta_{i,j})_x\rangle_{tx} f(\widehat{U}_{i,j}^n) = \sum_{i=1}^{\mathcal{N}} \langle \theta_{k,j}, \theta_{i,j}\rangle_{tx} s(\widehat{U}_{i,j}^n).$$

We introduce the following matrix entries

$$G_{ki} := \langle \theta_{k,j}, (\theta_{i,j})_t\rangle_{tx}, \quad H_{ki} := \langle \theta_{k,j}, (\theta_{i,j})_x\rangle_{tx}, \quad W_{ki} := \langle \theta_{k,j}, \theta_{i,j}\rangle_{tx}, \quad 1 \leqslant i,k \leqslant \mathcal{N}.$$

The values of these entries do not depend on j, since we use the same shifted basis for each j via a reference transformation. Introducing the vectors $\widehat{\mathbf{U}}_j^n := \left(\widehat{U}_{1,j}^n, \ldots, \widehat{U}_{\mathcal{N},j}^n\right)^T$, $\widehat{\mathbf{F}}_j^n := \left(f(\widehat{U}_{1,j}^n), \ldots, f(\widehat{U}_{\mathcal{N},j}^n)\right)^T$, and $\widehat{\mathbf{S}}_j^n := \left(s(\widehat{U}_{1,j}^n), \ldots, s(\widehat{U}_{\mathcal{N},j}^n)\right)^T$, we get the matrix forms

$$\mathbf{G}\widehat{\mathbf{U}}_j^n + \mathbf{H}\widehat{\mathbf{F}}_j^n = \mathbf{W}\widehat{\mathbf{S}}_j^n.$$

The first $M+1$ degrees of freedom are related to the functions $\Theta^0_{M,j}$. We group them together into the subvector $\widehat{\mathbf{U}}_j^{n,0} \in \mathbb{R}^{M+1}$. All other degrees of freedom are grouped together into the subvector $\widehat{\mathbf{U}}_j^{n,1} \in \mathbb{R}^{\mathcal{N}-M-1}$. Analogously, we define the subvectors $\widehat{\mathbf{F}}_j^{n,0}$, $\widehat{\mathbf{F}}_j^{n,1}$, $\widehat{\mathbf{S}}_j^{n,0}$, and $\widehat{\mathbf{S}}_j^{n,1}$. Then the matrix forms become

$$\mathbf{G}\begin{pmatrix} \widehat{\mathbf{U}}_j^{n,0} \\ \widehat{\mathbf{U}}_j^{n,1} \end{pmatrix} + \mathbf{H}\begin{pmatrix} \widehat{\mathbf{F}}_j^{n,0} \\ \widehat{\mathbf{F}}_j^{n,1} \end{pmatrix} = \mathbf{W}\begin{pmatrix} \widehat{\mathbf{S}}_j^{n,0} \\ \widehat{\mathbf{S}}_j^{n,1} \end{pmatrix},$$

We write the matrices $\mathbf{G}$, $\mathbf{H}$, and $\mathbf{W}$ as block matrices

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}^{00} & \mathbf{G}^{01} \\ \mathbf{G}^{10} & \mathbf{G}^{11} \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} \mathbf{H}^{00} & \mathbf{H}^{01} \\ \mathbf{H}^{10} & \mathbf{H}^{11} \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} \mathbf{W}^{00} & \mathbf{W}^{01} \\ \mathbf{W}^{10} & \mathbf{W}^{11} \end{pmatrix},$$

where

$$\begin{array}{ll} \mathbf{G}^{00}, \mathbf{H}^{00}, \mathbf{W}^{00} \in \mathbb{R}^{(M+1)\times(M+1)} & , \quad \mathbf{G}^{01}, \mathbf{H}^{01}, \mathbf{W}^{01} \in \mathbb{R}^{(M+1)\times(\mathcal{N}-M-1)} \\ \mathbf{G}^{10}, \mathbf{H}^{10}, \mathbf{W}^{10} \in \mathbb{R}^{(\mathcal{N}-M-1)\times(M+1)} & , \quad \mathbf{G}^{11}, \mathbf{H}^{11}, \mathbf{W}^{11} \in \mathbb{R}^{(\mathcal{N}-M-1)\times(\mathcal{N}-M-1)} \end{array}$$

Then we get

$$\begin{pmatrix} \mathbf{G}^{00} & \mathbf{G}^{01} \\ \mathbf{G}^{10} & \mathbf{G}^{11} \end{pmatrix}\begin{pmatrix} \widehat{\mathbf{U}}_j^{n,0} \\ \widehat{\mathbf{U}}_j^{n,1} \end{pmatrix} + \begin{pmatrix} \mathbf{H}^{00} & \mathbf{H}^{01} \\ \mathbf{H}^{10} & \mathbf{H}^{11} \end{pmatrix}\begin{pmatrix} \widehat{\mathbf{F}}_j^{n,0} \\ \widehat{\mathbf{F}}_j^{n,1} \end{pmatrix} = \begin{pmatrix} \mathbf{W}^{00} & \mathbf{W}^{01} \\ \mathbf{W}^{10} & \mathbf{W}^{11} \end{pmatrix}\begin{pmatrix} \widehat{\mathbf{S}}_j^{n,0} \\ \widehat{\mathbf{S}}_j^{n,1} \end{pmatrix}. \tag{7.6}$$

In the following we give two examples to the matrices arising in the matrix form.

7.3.1 Example $M = 2$

In this case the space time basis is given by (7.1) and we have the following matrices

$$\mathbf{G} = \frac{h}{6}\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & -1 \\ -2 & -1 & -2 & 2 & 2 & 1 \\ 0 & 1 & 0 & 0 & 0 & -1 \\ -1 & -2 & 1 & 1 & -1 & 2 \\ 1 & -2 & -1 & -1 & 1 & 2 \\ 2 & -3 & 2 & -2 & -2 & 3 \end{pmatrix}, \quad \mathbf{H} = \frac{k_n}{6}\begin{pmatrix} 0 & 0 & 0 & 1 & -1 & 0 \\ -1 & 0 & 1 & -2 & 2 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ -2 & 4 & -2 & -2 & 2 & 0 \\ 2 & -4 & 2 & -2 & 2 & 0 \\ 1 & 0 & -1 & -2 & 2 & 0 \end{pmatrix}, \tag{7.7}$$

$$\mathbf{W} = \frac{hk_n}{180}\begin{pmatrix} 18 & -27 & 8 & -19 & -9 & -1 \\ -27 & 80 & -27 & 34 & 34 & -4 \\ 8 & -27 & 18 & -9 & -19 & -1 \\ -19 & 34 & -9 & 44 & 4 & 6 \\ -9 & 34 & -19 & 4 & 44 & 6 \\ -1 & -4 & -1 & 6 & 6 & 24 \end{pmatrix}. \tag{7.8}$$

7.3.2 Example $M = 1$

In this case the space time basis is given by (7.2) and we have the following matrices

$$\mathbf{G} = \frac{h}{8}\begin{pmatrix} -1 & -1 & 2 \\ -1 & -1 & 2 \\ -2 & -2 & 4 \end{pmatrix}, \quad \mathbf{H} = \frac{k_n}{4}\begin{pmatrix} -1 & 1 & 0 \\ -1 & 1 & 0 \\ -2 & 2 & 0 \end{pmatrix}, \quad \mathbf{W} = \frac{hk_n}{12}\begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 4 \end{pmatrix}. \tag{7.9}$$

7.4 Inserting the Reconstructed Polynomials

In Chapter 5 we have defined the reconstructed polynomial w of degree M as a function only of the space variable, $w = w(x)$. Since we deal with the time variable t starting from this chapter, we denote the polynomial w as w^n with $w(t_n, x) = w^n(x)$ at time $t = t_n$. Furthermore, the coefficients $\widehat{w}_{i,j}$ will be denoted as $\widehat{w}^n_{i,j}$.

We determine the vector $\widehat{\mathbf{U}}^{n,0}_j$ of the first degrees of freedom related to $\Theta^0_{M,j}$ by projecting the reconstructed polynomial w^n at time $t = t_n$ onto the space spanned by the first nodal functions $\Theta^0_{M,j}$. This gives the following system of equations

$$\begin{aligned}
\int_{I_j} \theta_{1,j}(t_n,x)U^n(t_n,x)dx &= \int_{I_j} \theta_{1,j}(t_n,x)w(t_n,x)dx,\\
&\vdots\\
\int_{I_j} \theta_{M+1,j}(t_n,x)U^n(t_n,x)dx &= \int_{I_j} \theta_{M+1,j}(t_n,x)w(t_n,x)dx,
\end{aligned}$$

or, for $k = 1, \ldots, M+1$

$$\int_{I_j} \theta_{k,j}(t_n,x)\sum_{i=1}^{\mathcal{N}} \widehat{U}^n_{i,j}\theta_{i,j}(t_n,x)dx = \int_{I_j} \theta_{k,j}(t_n,x)\sum_{\ell=0}^{M} \widehat{w}^n_{\ell,j}\Phi_{\ell,j}(x)dx.$$

Since $\theta_{i,j}(t_n,x) = 0$ for $i = M+2, \ldots, \mathcal{N}$, then the sum in the left hand side reduces to the first $M+1$ terms. Thus we get

$$\sum_{i=1}^{M+1} \langle \theta_{k,j}(t_n,\cdot), \theta_{i,j}(t_n,\cdot)\rangle_j \widehat{U}^n_{i,j} = \sum_{\ell=0}^{M} \langle \theta_{k,j}(t_n,\cdot), \Phi_{\ell,j}(\cdot)\rangle_j \widehat{w}^n_{\ell,j}.$$

Now we define the vector $\widehat{\mathbf{w}}^n_j = \left(\widehat{w}^n_{0,j}, \ldots, \widehat{w}^n_{M,j}\right)^T$, and the matrices

$$\mathbf{J} := \begin{pmatrix}
\langle \theta_{1,j}(t_n,\cdot), \theta_{1,j}(t_n,\cdot)\rangle_j & \ldots & \langle \theta_{1,j}(t_n,\cdot), \theta_{M+1,j}(t_n,\cdot)\rangle_j \\
\langle \theta_{2,j}(t_n,\cdot), \theta_{1,j}(t_n,\cdot)\rangle_j & \ldots & \langle \theta_{2,j}(t_n,\cdot), \theta_{M+1,j}(t_n,\cdot)\rangle_j \\
\vdots & & \vdots \\
\langle \theta_{M+1,j}(t_n,\cdot), \theta_{1,j}(t_n,\cdot)\rangle_j & \ldots & \langle \theta_{M+1,j}(t_n,\cdot), \theta_{M+1,j}(t_n,\cdot)\rangle_j
\end{pmatrix},$$

$$\mathbf{K} := \begin{pmatrix}
\langle \theta_{1,j}(t_n,\cdot), \Phi_{0,j}(\cdot)\rangle_j & \ldots & \langle \theta_{1,j}(t_n,\cdot), \Phi_{M,j}(\cdot)\rangle_j \\
\langle \theta_{2,j}(t_n,\cdot), \Phi_{0,j}(\cdot)\rangle_j & \ldots & \langle \theta_{2,j}(t_n,\cdot), \Phi_{M,j}(\cdot)\rangle_j \\
\vdots & & \vdots \\
\langle \theta_{M+1,j}(t_n,\cdot), \Phi_{0,j}(\cdot)\rangle_j & \ldots & \langle \theta_{M+1,j}(t_n,\cdot), \Phi_{M,j}(\cdot)\rangle_j
\end{pmatrix},$$

then, making k takes the values $1, \ldots, M+1$, we get the following matrix forms $\mathbf{J}\widehat{\mathbf{U}}_j^{n,0} = \mathbf{K}\widehat{\mathbf{w}}_j^n$. Since the functions $\theta_{i,j}$ for $i = 1, M+1$, belong to the basis, they are linearly independent, thus the columns of $\mathbf{J}$ are linearly independent. This implies that $\mathbf{J}$ is invertible, for all orders $M \geqslant 0$. So by denoting the following projection Matrix $\mathbf{O}_M := \mathbf{J}^{-1}\mathbf{K}$ we can write $\widehat{\mathbf{U}}_j^{n,0} = \mathbf{O}_M\widehat{\mathbf{w}}_j^n$. The matrix $\mathbf{O}_M$ is square of size $M+1$. We give now four examples of these projection matrices

$$\mathbf{O}_0 = 1, \quad \mathbf{O}_1 = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{O}_2 = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -\frac{1}{2} \\ 1 & 1 & 1 \end{pmatrix}, \quad \mathbf{O}_3 = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & -\frac{1}{3} & -\frac{1}{3} & \frac{11}{27} \\ 1 & \frac{1}{3} & -\frac{1}{3} & -\frac{11}{27} \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

7.4.1 Example $M = 2$

We choose $N = 0$ and $S_{I_j,3,1}$. According to (5.19) we have

$$\widehat{\mathbf{U}}_j^{n,0} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -\frac{1}{2} \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \widehat{u}_{0,j}^n \\ \frac{\widehat{u}_{0,j+1}^n - \widehat{u}_{0,j-1}^n}{4} \\ \frac{\widehat{u}_{0,j-1}^n - 2\widehat{u}_{0,j}^n + \widehat{u}_{0,j+1}^n}{12} \end{pmatrix} = \frac{1}{24} \begin{pmatrix} 8\widehat{u}_{0,j-1}^n + 20\widehat{u}_{0,j}^n - 4\widehat{u}_{0,j+1}^n \\ -\widehat{u}_{0,j-1}^n + 26\widehat{u}_{0,j}^n - \widehat{u}_{0,j+1}^n \\ -4\widehat{u}_{0,j-1}^n + 20\widehat{u}_{0,j}^n + 8\widehat{u}_{0,j+1}^n \end{pmatrix}.$$

7.4.2 Example $M = 1$

We choose $N = 0$ and $S_{I_j,2,1}$. According to (5.13) we have

$$\begin{pmatrix} \widehat{U}_{1,j}^n \\ \widehat{U}_{2,j}^n \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \widehat{u}_{0,j}^n \\ \frac{1}{2}(\widehat{u}_{0,j}^n - \widehat{u}_{0,j-1}^n) \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \widehat{u}_{0,j}^n + \widehat{u}_{0,j-1}^n \\ 3\widehat{u}_{0,j}^n - \widehat{u}_{0,j-1}^n \end{pmatrix}.$$

7.5 Reducing the Algebraic System

Since we now have determined $M+1$ known coefficients, we no longer need the upper blocks in (7.6). Therefore, we cancel the first $M+1$ rows of these systems to obtain the smaller systems

$$\begin{pmatrix} \mathbf{G}^{10} & \mathbf{G}^{11} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{U}}_j^{n,0} \\ \widehat{\mathbf{U}}_j^{n,1} \end{pmatrix} + \begin{pmatrix} \mathbf{H}^{10} & \mathbf{H}^{11} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{F}}_j^{n,0} \\ \widehat{\mathbf{F}}_j^{n,1} \end{pmatrix} = \begin{pmatrix} \mathbf{W}^{10} & \mathbf{W}^{11} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{S}}_j^{n,0} \\ \widehat{\mathbf{S}}_j^{n,1} \end{pmatrix}.$$

In order to determine the vector $\widehat{\mathbf{U}}_j^{n,1}$ we have to solve the nonlinear equations

$$\mathbf{G}^{11}\widehat{\mathbf{U}}_j^{n,1} + \mathbf{H}^{11}\widehat{\mathbf{F}}_j^{n,1} - \mathbf{W}^{11}\widehat{\mathbf{S}}_j^{n,1} = -\mathbf{G}^{10}\widehat{\mathbf{U}}_j^{n,0} - \mathbf{H}^{10}\widehat{\mathbf{F}}_j^{n,0} + \mathbf{W}^{10}\widehat{\mathbf{S}}_j^{n,0}.$$

The quadratic matrix $\mathbf{G}^{11}$ depends on the mesh size h but not on the time step or the equations to be solved. For all orders of accuracy, the matrix $\mathbf{G}^{11}$ is invertible, since its columns are linearly independent. Therefore we obtain a fixed point problem for the unknowns $\widehat{\mathbf{U}}_j^{n,1}$

$$\widehat{\mathbf{U}}_j^{n,1} = \left(\mathbf{G}^{11}\right)^{-1} \left[\mathbf{W}^{11}\widehat{\mathbf{S}}_j^{n,1} - \mathbf{H}^{11}\widehat{\mathbf{F}}_j^{n,1} + \mathbf{W}^{10}\widehat{\mathbf{S}}_j^{n,0} - \mathbf{H}^{10}\widehat{\mathbf{F}}_j^{n,0} - \mathbf{G}^{10}\widehat{\mathbf{U}}_j^{n,0}\right].$$

7.5.1 Example $M = 2$

We depend on the matrices given by (7.7) and (7.8). Then we have

$$\mathbf{G}^{10} = \frac{h}{6}\begin{pmatrix} -1 & -2 & 1 \\ 1 & -2 & -1 \\ 2 & -3 & 2 \end{pmatrix}, \quad \mathbf{H}^{10} = \tfrac{k_n}{6}\begin{pmatrix} -2 & 4 & -2 \\ 2 & -4 & 2 \\ 1 & 0 & -1 \end{pmatrix}, \quad \mathbf{W}^{10} = \frac{hk_n}{180}\begin{pmatrix} -19 & 34 & -9 \\ -9 & 34 & -19 \\ -1 & -4 & -1 \end{pmatrix},$$

$$\mathbf{G}^{11} = \frac{h}{6}\begin{pmatrix} 1 & -1 & 2 \\ -1 & 1 & 2 \\ -2 & -2 & 3 \end{pmatrix}, \quad \mathbf{H}^{11} = \tfrac{k_n}{3}\begin{pmatrix} -1 & 1 & 0 \\ -1 & 1 & 0 \\ -1 & 1 & 0 \end{pmatrix}, \quad \mathbf{W}^{11} = \frac{hk_n}{90}\begin{pmatrix} 22 & 2 & 3 \\ 2 & 22 & 3 \\ 3 & 3 & 12 \end{pmatrix}.$$

The unknown degrees of freedom are then given by

$$\begin{aligned}
\begin{pmatrix} \widehat{U}^n_{4,j} \\ \widehat{U}^n_{5,j} \\ \widehat{U}^n_{6,j} \end{pmatrix} &= \frac{k_n}{120}\begin{pmatrix} 70 & -10 & -15 \\ -10 & 70 & -15 \\ 48 & 48 & 12 \end{pmatrix}\begin{pmatrix} s(\widehat{U}^n_{4,j}) \\ s(\widehat{U}^n_{5,j}) \\ s(\widehat{U}^n_{6,j}) \end{pmatrix} + \frac{k_n}{4h}\begin{pmatrix} 1 & -1 & 0 \\ 1 & -1 & 0 \\ 4 & -4 & 0 \end{pmatrix}\begin{pmatrix} f(\widehat{U}^n_{4,j}) \\ f(\widehat{U}^n_{5,j}) \\ f(\widehat{U}^n_{6,j}) \end{pmatrix} \\
&+ \frac{k_n}{120}\begin{pmatrix} -30 & 55 & -10 \\ -10 & 55 & -30 \\ -28 & 68 & -28 \end{pmatrix}\begin{pmatrix} s(\widehat{U}^n_{1,j}) \\ s(\widehat{U}^n_{2,j}) \\ s(\widehat{U}^n_{3,j}) \end{pmatrix} + \frac{k_n}{4h}\begin{pmatrix} 5 & -8 & 3 \\ -3 & 8 & -5 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} f(\widehat{U}^n_{1,j}) \\ f(\widehat{U}^n_{2,j}) \\ f(\widehat{U}^n_{3,j}) \end{pmatrix} \\
&+ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} \widehat{U}^n_{1,j} \\ \widehat{U}^n_{2,j} \\ \widehat{U}^n_{3,j} \end{pmatrix}.
\end{aligned}$$

7.5.2 Example $M = 1$

We depend on the matrices given by (7.9).

$$\mathbf{G}^{10} = \frac{h}{4}(-1,-1), \quad \mathbf{H}^{10} = \frac{k_n}{2}(-1,1), \quad \mathbf{W}^{10} = \frac{hk_n}{12}(1,1), \quad \mathbf{G}^{11} = \frac{h}{2}, \quad \mathbf{H}^{11} = 0, \quad \mathbf{W}^{11} = \frac{hk_n}{3}. \tag{7.10}$$

The unique unknown degree of freedom $\widehat{U}^n_{3,j}$ is given by

$$\widehat{U}^n_{3,j} = \tfrac{2k_n}{3} s\left(\widehat{U}^n_{3,j}\right) + \tfrac{k_n}{6}(1,1)\begin{pmatrix} s\left(\widehat{U}^n_{1,j}\right) \\ s\left(\widehat{U}^n_{2,j}\right) \end{pmatrix} + \tfrac{k_n}{h}(1,-1)\begin{pmatrix} f\left(\widehat{U}^n_{1,j}\right) \\ f\left(\widehat{U}^n_{2,j}\right) \end{pmatrix} + \left(\tfrac{1}{2},\tfrac{1}{2}\right)\begin{pmatrix} \widehat{U}^n_{1,j} \\ \widehat{U}^n_{2,j} \end{pmatrix}.$$

7.6 Iterating the Reduced Algebraic System

We solve the final system for $\widehat{\mathbf{U}}^{n,1}_j$ using the fixed point iteration

$$\widehat{\mathbf{U}}^{n,1,i+1}_j = \left(\mathbf{G}^{11}\right)^{-1}\left[\mathbf{W}^{11}\widehat{\mathbf{S}}^{n,1,i}_j - \mathbf{H}^{11}\widehat{\mathbf{F}}^{n,1,i}_j + \mathbf{W}^{10}\widehat{\mathbf{S}}^{n,0}_j - \mathbf{H}^{10}\widehat{\mathbf{F}}^{n,0}_j - \mathbf{G}^{10}\widehat{\mathbf{U}}^{n,0}_j\right]. \tag{7.11}$$

The superscript i denotes the iteration number. This approach works since $\left(\mathbf{G}^{11}\right)^{-1}\mathbf{W}^{11}$ and $\left(\mathbf{G}^{11}\right)^{-1}\mathbf{H}^{11}$ turn out to be contraction mappings, see Dumbser et al. [7, p.8218]. In our practical computations the fixed point was determined after at most $M+1$ iterations.

As suggested in [7] we begin iterating by using a stationary solution in time of (7.3) as an initial guess value for $\widehat{\mathbf{U}}_j^{n,1}$. The stationary equation is $v_t = 0$. The matrix form is $\mathbf{G}\widehat{\mathbf{U}}_j^n = 0$. Then we get the initial guess with $i = 0$

$$\widehat{\mathbf{U}}_j^{n,1,0} = -\left(\mathbf{G}^{11}\right)^{-1}\mathbf{G}^{10}\widehat{\mathbf{U}}_j^{n,0}. \tag{7.12}$$

7.7 Example 1: the Linear Advection Equation

Now we apply the local Galerkin scheme to the initial value problem of the linear advection equation $v_t(t,x) + av_x(t,x) = 0$ for $(t,x) \in \mathbb{R}_{\geqslant 0} \times I$ with $I \subset \mathbb{R}$, $a \in \mathbb{R}$, and with an initial function $v_0(x) = v(0,x)$ defined on I. We show formulas of the solutions of this approach on some time element $\breve{T}_n = [t_n, t_{n+1}[$ with time step $k_n > 0$. We have $f(\widehat{U}_{i,j}^n) = a\widehat{U}_{i,j}^n$ for all $i = 1, \ldots, \mathcal{N}$ and $j = 1, \ldots, Z$. The iterative scheme (7.11) becomes

$$\widehat{\mathbf{U}}_j^{n,1,i+1} = \left(\mathbf{G}^{11}\right)^{-1}\left[-a\mathbf{H}^{11}\widehat{\mathbf{U}}_j^{n,1,i} - a\mathbf{H}^{10}\widehat{\mathbf{U}}_j^{n,0} - \mathbf{G}^{10}\widehat{\mathbf{U}}_j^{n,0}\right]. \tag{7.13}$$

This system has to be solved iteratively for $\widehat{\mathbf{U}}_j^{n,1}$ with the initial guess given in (7.12).

7.7.1 $M = 0$

The solution is piecewise constant $\widehat{U}_{1,j}^n = \widehat{w}_{0,j}^n = \widehat{u}_{0,j}^n$ where no need to the iteration.

7.7.2 $M = 1$

We have $\mathcal{N} = 3$ and the known coefficients are given by

$$\begin{pmatrix} \widehat{U}_{1,j}^n \\ \widehat{U}_{2,j}^n \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \widehat{w}_{0,j}^n \\ \widehat{w}_{1,j}^n \end{pmatrix} = \begin{pmatrix} \widehat{w}_{0,j}^n - \widehat{w}_{1,j}^n \\ \widehat{w}_{0,j}^n + \widehat{w}_{1,j}^n \end{pmatrix}.$$

On the other hand, since, according to (7.10), $\mathbf{H}^{11} = 0$ and there is no source terms with our advection equation, then the local Galerkin scheme with $M = 1$ is fully explicit and hence it does not need any iteration. The iterative equation (7.13) becomes

$$\widehat{U}_{3,j}^{n,i+1} = \frac{2ak_n}{h}\left(\frac{1}{2}, \frac{-1}{2}\right)\begin{pmatrix} \widehat{U}_{1,j}^n \\ \widehat{U}_{2,j}^n \end{pmatrix} + \left(\frac{1}{2}, \frac{1}{2}\right)\begin{pmatrix} \widehat{U}_{1,j}^n \\ \widehat{U}_{2,j}^n \end{pmatrix} = \widehat{w}_{0,j}^n - \frac{2ak_n}{h}\widehat{w}_{1,j}^n.$$

7.7.2.1 $N = 0$ and $M = 1$

With $S_{I_j,2,1}$. According to (5.13) we have

$$\widehat{\mathbf{U}}_j^n = \frac{1}{2}\begin{pmatrix} \widehat{u}_{0,j-1}^n + \widehat{u}_{0,j}^n \\ -\widehat{u}_{0,j-1}^n + 3\widehat{u}_{0,j}^n \\ 2\widehat{u}_{0,j}^n + \frac{2ak_n}{h}(\widehat{u}_{0,j-1}^n - \widehat{u}_{0,j}^n) \end{pmatrix}.$$

With $S_{I_j,2,0}$. According to (5.14) we have

$$\widehat{\mathbf{U}}_j^n = \frac{1}{2}\begin{pmatrix} 3\widehat{u}_{0,j}^n - \widehat{u}_{0,j+1}^n \\ \widehat{u}_{0,j}^n + \widehat{u}_{0,j+1}^n \\ 2\widehat{u}_{0,j}^n + \frac{2ak_n}{h}(\widehat{u}_{0,j}^n - \widehat{u}_{0,j+1}^n) \end{pmatrix}.$$

With $S_{I_j,3,2}$. According to (5.15) we have

$$\widehat{\mathbf{U}}_j^n = \frac{1}{10}\begin{pmatrix} 2\widehat{u}_{0,j-2}^n + \widehat{u}_{0,j-1}^n + 7\widehat{u}_{0,j}^n \\ -2\widehat{u}_{0,j-2}^n - \widehat{u}_{0,j-1}^n + 13\widehat{u}_{0,j}^n \\ 10\widehat{u}_{0,j}^n + \frac{2ak_n}{h}(2\widehat{u}_{0,j-2}^n + \widehat{u}_{0,j-1}^n - 3\widehat{u}_{0,j+1}^n) \end{pmatrix}.$$

With $S_{I_j,3,1}$. According to (5.16) we have

$$\widehat{\mathbf{U}}_j^n = \frac{1}{4}\begin{pmatrix} \widehat{u}_{0,j-1}^n + 4\widehat{u}_{0,j}^n - \widehat{u}_{0,j+1}^n \\ -\widehat{u}_{0,j-1}^n + 4\widehat{u}_{0,j}^n + \widehat{u}_{0,j+1}^n \\ 4\widehat{u}_{0,j}^n + \frac{2ak_n}{h}(\widehat{u}_{0,j-1}^n - \widehat{u}_{0,j+1}^n) \end{pmatrix}.$$

With $S_{I_j,3,0}$ and the degrees of freedom (5.17)

$$\widehat{\mathbf{U}}_j^n = \frac{1}{10}\begin{pmatrix} 13\widehat{u}_{0,j}^n - \widehat{u}_{0,j+1}^n - 2\widehat{u}_{0,j+2}^n \\ 7\widehat{u}_{0,j}^n + \widehat{u}_{0,j+1}^n + 2\widehat{u}_{0,j+2}^n \\ 10\widehat{u}_{0,j}^n + \frac{2ak_n}{h}(3\widehat{u}_{0,j}^n - \widehat{u}_{0,j+1}^n - 2\widehat{u}_{0,j+2}^n) \end{pmatrix}.$$

7.7.2.2 $M = N = 1$

We have $\widehat{w}_{i,j}^n = \widehat{u}_{i,j}^n$ for $i = 0, 1$ and $\widehat{\mathbf{U}}_j^n = \begin{pmatrix} \widehat{u}_{0,j}^n - \widehat{u}_{1,j}^n \\ \widehat{u}_{0,j}^n + \widehat{u}_{1,j}^n \\ \widehat{u}_{0,j}^n - \frac{2ak_n}{h}\widehat{u}_{1,j}^n \end{pmatrix}$.

7.7.3 $M = 2$

We have 6 coefficients where $\mathcal{N} = 6$ and the known coefficients are computed by

$$\begin{pmatrix} \widehat{U}_{1,j}^n \\ \widehat{U}_{2,j}^n \\ \widehat{U}_{3,j}^n \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -\frac{1}{2} \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} \widehat{w}_{0,j}^n \\ \widehat{w}_{1,j}^n \\ \widehat{w}_{2,j}^n \end{pmatrix} = \begin{pmatrix} \widehat{w}_{0,j}^n - \widehat{w}_{1,j}^n + \widehat{w}_{2,j}^n \\ \widehat{w}_{0,j}^n - \frac{1}{2}\widehat{w}_{2,j}^n \\ \widehat{w}_{0,j}^n + \widehat{w}_{1,j}^n + \widehat{w}_{2,j}^n \end{pmatrix}.$$

The iterative equation (7.13) becomes

$$\begin{pmatrix} \widehat{U}_{4,j}^n \\ \widehat{U}_{5,j}^n \\ \widehat{U}_{6,j}^n \end{pmatrix}^{i+1} = \frac{ak_n}{4h}\begin{pmatrix} 1 & -1 & 0 \\ 1 & -1 & 0 \\ 4 & -4 & 0 \end{pmatrix}\begin{pmatrix} \widehat{U}_{4,j}^n \\ \widehat{U}_{5,j}^n \\ \widehat{U}_{6,j}^n \end{pmatrix}^i + \left[\frac{ak_n}{4h}\begin{pmatrix} 5 & -8 & 3 \\ -3 & 8 & -5 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}\right]\begin{pmatrix} \widehat{U}_{1,j}^n \\ \widehat{U}_{2,j}^n \\ \widehat{U}_{3,j}^n \end{pmatrix}.$$

7.7.3.1 $N = 0$ and $M = 2$

With $S_{I_j,3,2}$. According to (5.18), the known coefficients become

$$\begin{pmatrix} \widehat{U}^n_{1,j} \\ \widehat{U}^n_{2,j} \\ \widehat{U}^n_{3,j} \end{pmatrix} = \frac{1}{24} \begin{pmatrix} -4\widehat{u}^n_{0,j-2} + 20\widehat{u}^n_{0,j-1} + 8\widehat{u}^n_{0,j} \\ -\widehat{u}^n_{0,j-2} + 2\widehat{u}^n_{0,j-1} + 23\widehat{u}^n_{0,j} \\ 8\widehat{u}^n_{0,j-2} - 28\widehat{u}^n_{0,j-1} + 44\widehat{u}^n_{0,j} \end{pmatrix}.$$

We begin with the guess solution (7.12) corresponding to the iteration index $i = 0$ to get

$$\begin{pmatrix} \widehat{U}^n_{4,j} \\ \widehat{U}^n_{5,j} \\ \widehat{U}^n_{6,j} \end{pmatrix}^1 = \frac{1}{48} \begin{pmatrix} -8\widehat{u}^n_{0,j-2} + 40\widehat{u}^n_{0,j-1} + 16\widehat{u}^n_{0,j} + \frac{6ak_n}{h}(\widehat{u}^n_{0,j-2} - \widehat{u}^n_{0,j-2}) \\ 16\widehat{u}^n_{0,j-2} - 56\widehat{u}^n_{0,j-1} + 88\widehat{u}^n_{0,j} - \frac{6ak_n}{h}(3\widehat{u}^n_{0,j-2} - 8\widehat{u}^n_{0,j-1} + 5\widehat{u}^n_{0,j}) \\ -2\widehat{u}^n_{0,j-2} + 4\widehat{u}^n_{0,j-1} + 46\widehat{u}^n_{0,j} \end{pmatrix}.$$

We iterate the solution again to get

$$\begin{pmatrix} \widehat{U}^n_{4,j} \\ \widehat{U}^n_{5,j} \\ \widehat{U}^n_{6,j} \end{pmatrix}^2 = \frac{1}{48} \begin{pmatrix} -8\widehat{u}^n_{0,j-2}+40\widehat{u}^n_{0,j-1}+16\widehat{u}^n_{0,j}+\frac{24ak_n}{h}(\widehat{u}^n_{0,j-1}-\widehat{u}^n_{0,j})+\frac{6(ak_n)^2}{h^2}(\widehat{u}^n_{0,j-2}-2\widehat{u}^n_{0,j-1}+\widehat{u}^n_{0,j}) \\ +16\widehat{u}^n_{0,j-2}-56\widehat{u}^n_{0,j-1}+88\widehat{u}^n_{0,j}-\frac{24ak_n}{h}(\widehat{u}^n_{0,j-2}-3\widehat{u}^n_{0,j-1}+2\widehat{u}^n_{0,j})+\frac{6(ak_n)^2}{h^2}(\widehat{u}^n_{0,j-2}-2\widehat{u}^n_{0,j-1}+\widehat{u}^n_{0,j}) \\ -2(\widehat{u}^n_{0,j-2}-2\widehat{u}^n_{0,j-1}-23\widehat{u}^n_{0,j})-\frac{24ak_n}{h}(\widehat{u}^n_{0,j-2}-4\widehat{u}^n_{0,j-1}+3\widehat{u}^n_{0,j})+\frac{24(ak_n)^2}{h^2}(\widehat{u}^n_{0,j-2}-2\widehat{u}^n_{0,j-1}+\widehat{u}^n_{0,j}) \end{pmatrix}.$$

If we repeat the iteration we will get the same result $\widehat{\mathbf{U}}^{n,1,i}_j = \widehat{\mathbf{U}}^{n,1,2}_j$ for all $i \geqslant 3$.

7.7.3.2 $N = 1$ and $M = 2$

With $S_{I_j,2,1}$. According to (5.20) we have

$$\begin{pmatrix} \widehat{U}^n_{1,j} \\ \widehat{U}^n_{2,j} \\ \widehat{U}^n_{3,j} \end{pmatrix} = \frac{1}{120} \begin{pmatrix} 18\widehat{u}^n_{0,j-1} + 102\widehat{u}^n_{0,j} - 2\widehat{u}^n_{1,j-1} - 82\widehat{u}^n_{1,j} \\ -9\widehat{u}^n_{0,j-1} + 129\widehat{u}^n_{0,j} + \widehat{u}^n_{1,j-1} - 19\widehat{u}^n_{1,j} \\ 18\widehat{u}^n_{0,j-1} + 102\widehat{u}^n_{0,j} - 2\widehat{u}^n_{1,j-1} + 158\widehat{u}^n_{1,j} \end{pmatrix}.$$

In the same way and after two iterations we get the solution

$$\begin{aligned}
\widehat{U}^n_{1,j} = \frac{1}{60} \quad & \Big[9\widehat{u}^n_{0,j-1} + 51\widehat{u}^n_{0,j} - \widehat{u}^n_{1,j-1} - 41\widehat{u}^n_{1,j} + \frac{3ak_n}{h}(9(\widehat{u}^n_{0,j-1} - \widehat{u}^n_{0,j}) \\
& -\widehat{u}^n_{1,j-1} - \widehat{u}^n_{1,j}) + \frac{3(ak_n)^2}{2h^2}(9(\widehat{u}^n_{0,j-1} - \widehat{u}^n_{0,j}) - \widehat{u}^n_{1,j-1} + 19\widehat{u}^n_{1,j})\Big], \\
\widehat{U}^n_{2,j} = \frac{1}{60} \quad & \Big[9\widehat{u}^n_{0,j-1} + 51\widehat{u}^n_{0,j} - \widehat{u}^n_{1,j-1} + 79\widehat{u}^n_{1,j} - \frac{3ak_n}{h}(9(\widehat{u}^n_{0,j-1} - \widehat{u}^n_{0,j}) \\
& -\widehat{u}^n_{1,j-1} + 39\widehat{u}^n_{1,j}) + \frac{3(ak_n)^2}{2h^2}(9(\widehat{u}^n_{0,j-1} - \widehat{u}^n_{0,j}) - \widehat{u}^n_{1,j-1} + 19\widehat{u}^n_{1,j})\Big], \\
\widehat{U}^n_{3,j} = \frac{1}{60} \quad & \Big[\frac{-1}{2}(9\widehat{u}^n_{0,j-1} - 129\widehat{u}^n_{0,j} - \widehat{u}^n_{1,j-1} + 19\widehat{u}^n_{1,j}) - \frac{120ak_n}{h}\widehat{u}_{1,j} \\
& +\frac{6(ak_n)^2}{h^2}(9(\widehat{u}^n_{0,j-1} - \widehat{u}^n_{0,j}) - \widehat{u}^n_{1,j-1} + 19\widehat{u}^n_{1,j})\Big].
\end{aligned}$$

7.7.3.3 $M = N = 2$

We have $\widehat{w}^n_{i,j} = \widehat{u}^n_{i,j}$ for $i = 0, 1, 2$ and $\begin{pmatrix} \widehat{U}^n_{1,j} \\ \widehat{U}^n_{2,j} \\ \widehat{U}^n_{3,j} \end{pmatrix} = \begin{pmatrix} \widehat{u}^n_{0,j} - \widehat{u}^n_{1,j} + \widehat{u}^n_{2,j} \\ \widehat{u}^n_{0,j} - \frac{1}{2}\widehat{u}^n_{2,j} \\ \widehat{u}^n_{0,j} + \widehat{u}^n_{1,j} + \widehat{u}^n_{2,j} \end{pmatrix}$. After two iterations we get the following solution

$$\begin{pmatrix} \widehat{U}^n_{4,j} \\ \widehat{U}^n_{5,j} \\ \widehat{U}^n_{6,j} \end{pmatrix} = \begin{pmatrix} \widehat{u}^n_{0,j} - \widehat{u}^n_{1,j} + \widehat{u}^n_{2,j} - \frac{ak_n}{h}(\widehat{u}^n_{1,j} - 3\widehat{u}^n_{2,j}) + \frac{3(ak_n)^2}{2h^2}\widehat{u}^n_{2,j} \\ \widehat{u}^n_{0,j} + \widehat{u}^n_{1,j} + \widehat{u}^n_{2,j} - \frac{ak_n}{h}(\widehat{u}^n_{1,j} + 3\widehat{u}^n_{2,j}) + \frac{3(ak_n)^2}{2h^2}\widehat{u}^n_{2,j} \\ \widehat{u}^n_{0,j} - \frac{1}{2}\widehat{u}^n_{2,j} - \frac{2ak_n}{h}\widehat{u}^n_{1,j} + \frac{6(ak_n)^2}{h^2}\widehat{u}^n_{2,j} \end{pmatrix}.$$

7.8 Example 2: Nonlinear Burgers Equation

Now we apply the local Galerkin scheme to the initial value problem of the nonlinear Burgers equation $v_t(t,x) + (v^2/2)_x(t,x) = 0$ for $(t,x) \in \mathbb{R}_{\geqslant 0} \times I$ with $I \subset \mathbb{R}$ and with an initial function $v_0(x) = v(0,x)$ defined on I. We have $f(\widehat{U}^n_{i,j}) = \frac{1}{2}(\widehat{U}^n_{i,j})^2$ for all $i = 1, \ldots, \mathcal{N}$ and $j = 1, \ldots, Z$.

7.8.1 $M = 1$

The known coefficients are given by

$$\begin{pmatrix} \widehat{U}^n_{1,j} \\ \widehat{U}^n_{2,j} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \widehat{w}^n_{0,j} \\ \widehat{w}^n_{1,j} \end{pmatrix} = \begin{pmatrix} \widehat{w}^n_{0,j} - \widehat{w}^n_{1,j} \\ \widehat{w}^n_{0,j} + \widehat{w}^n_{1,j} \end{pmatrix}.$$

We found above that the local Galerkin scheme with $M = 1$ does not need any iteration. The iterative equation becomes

$$\widehat{U}^{n,i+1}_{3,j} = \frac{2k_n}{h}\left(\frac{1}{2}, \frac{-1}{2}\right)\begin{pmatrix} \frac{1}{2}\left(\widehat{U}^n_{1,j}\right)^2 \\ \frac{1}{2}\left(\widehat{U}^n_{2,j}\right)^2 \end{pmatrix} + \left(\frac{1}{2}, \frac{1}{2}\right)\begin{pmatrix} \widehat{U}^n_{1,j} \\ \widehat{U}^n_{2,j} \end{pmatrix} = \widehat{w}^n_{0,j} - \frac{2k_n}{h}\widehat{w}^n_{0,j}\widehat{w}^n_{1,j}.$$

7.8.2 $N = 0$, $M = 1$, with $S_{I_j,2,1}$

$$\widehat{\mathbf{U}}^n_j = \frac{1}{2}\begin{pmatrix} \widehat{u}^n_{0,j-1} + \widehat{u}^n_{0,j} \\ -\widehat{u}^n_{0,j-1} + 3\widehat{u}^n_{0,j} \\ 2\widehat{u}^n_{0,j} + \frac{2k_n}{h}\widehat{u}^n_{0,j}(\widehat{u}^n_{0,j-1} - \widehat{u}^n_{0,j}) \end{pmatrix}.$$

7.8.3 $M = N = 1$

We have $\widehat{w}^n_{i,j} = \widehat{u}^n_{i,j}$ for $i = 0, 1$ and

$$\widehat{\mathbf{U}}^n_j = \begin{pmatrix} \widehat{u}^n_{0,j} - \widehat{u}^n_{1,j} \\ \widehat{u}^n_{0,j} + \widehat{u}^n_{1,j} \\ \widehat{u}^n_{0,j} - \frac{2k_n}{h}\widehat{u}^n_{0,j}\widehat{u}^n_{1,j} \end{pmatrix}.$$

Chapter 8

The Discontinuous Galerkin Schemes

8.1 Preface

Here we apply the DG schemes [5, 6] and use numerical fluxes whose arguments are the solutions U^n and F^n of the previous step.

Let $T > 0$ and $I \subset \mathbb{R}$. We discretize the intervals I and $[0, T]$ and consider $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[$ and $\breve{T}_n = [t_n, t_{n+1}[$ with a time step $k_n > 0$ and a constant mesh size h for $j = 1, \ldots, Z$.

In this thesis we will study various functions of different order of smoothness, e.g. starting from the space $C^\infty(I)$ of infinitely continuously differentiable functions, such as e.g. $\sin x$, up to the discontinuous functions, such as a jump function. Therefore, at each time $t \in [0, T]$, the function which will be considered must be at least bounded on the interval I.

The integral of a bounded function on a closed interval I always exists, provided that the set of points in I, at which the function is not continuous, has Lebesgue measure 0, see also [13]. Then we can deal with integrable functions, for example, deal with $v(t, \cdot) \in L^2(I)$. The conserved functions will be from the following function space

$$L^\infty([0, T], L^2(I)) = \{v : [0, T] \times I \to \mathbb{R}, \text{ such that ess } \sup_{t \in [0,T]} ||v(t, \cdot)|| < \infty\}.$$

8.2 The $P_N P_M$ DG Schemes

Here we formulate the final fully discrete $P_N P_M$ DG schemes for the general nonlinear hyperbolic systems of balance laws in one space dimension. We have

$$\mathbf{v}_t(t, x) + \mathbf{f}(\mathbf{v}(t, x))_x = \mathbf{s}(\mathbf{v}(t, x)), \quad \text{for } (t, x) \in [0, T] \times I.$$

Let v_p, f_p, and s_p be arbitrary components of the vectors $\mathbf{v}$, $\mathbf{f}$, and $\mathbf{s}$, respectively. According to these components we have the following equation

$$(v_p(t, x))_t + (f_p(v_q(t, x))_x = s_p(v_q(t, x)).$$

The components f_p and s_p are based on the vector $\mathbf{v}$, so we write them in the form $f_p(v_q)$ and $s_p(v_q)$, respectively, where v_q of the subindex q indicates an arbitrary component of the

vector $\mathbf{v}$. We consider one space element I_j, with $j = 1, \ldots, Z$ fixed, and one time element $\breve{T}_n = [t_n, t_{n+1}[$.

8.2.1 Step 1.

Multiplying with an arbitrary smooth function $\chi \in L^2(I_j)$, integrating over $\breve{T}_n \times I_j$, and using integration by parts in space we get

$$\int_{\breve{T}_n}\int_{I_j} \chi(x)(v_p(t,x))_t dxdt + \int_{\breve{T}_n} \chi(x) f_p(v_q(t,x)) \Big|_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} dt$$
$$- \int_{\breve{T}_n}\int_{I_j} \chi_x(x) f_p\left(v_q(t,x)\right) dxdt = \int_{\breve{T}_n}\int_{I_j} \chi(x) s_p\left(v_q(t,x)\right) dxdt. \tag{8.1}$$

8.2.2 Step 2.

We assume that the numerical solution, defined on $\breve{T}_n \times I$, of the DG scheme, which we denote as u_p^n, is a piecewise polynomial of the degree N and is given by

$$u_p^n(t,x) = \sum_{j=1}^{Z}\sum_{\ell=0}^{N} \widehat{u}_{p,\ell,j}^n(t)\Phi_{\ell,j}(x), \quad \text{for } \; x \in I,$$

where $\widehat{u}_{p,\ell,j}^n \in \mathbb{R}$, for $j = 1, \ldots, Z$ and $\ell = 0, \ldots, N$, are unknowns, and $\Phi_{\ell,j}$ are the Legendre basis functions of degree $N \geqslant 0$, which are given by (2.8). This solution has Z terms, each term is written as

$$u_{p,j}^n(t,x) = \sum_{\ell=0}^{N} \widehat{u}_{p,\ell,j}^n(t)\Phi_{\ell,j}(x), \quad \text{for } \; j = 1, \ldots, Z.$$

8.2.3 Step 3.

Substituting the numerical solution u_p^n and replacing the test function χ by the basis functions $\Phi_{k,j}$ for $k = 0, \ldots, N$ in (8.1) leads to the following forms

$$\int_{\breve{T}_n}\int_{I_j} \Phi_{k,j}(x)(u_p^n(t,x))_t dxdt + \int_{\breve{T}_n} \Phi_{k,j}(x) f_p(u_q^n(t,x)) \Big|_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} dt$$
$$- \int_{\breve{T}_n}\int_{I_j} (\Phi_{k,j}(x))_x f_p(u_q^n(t,x)) dxdt = \int_{\breve{T}_n}\int_{I_j} \Phi_{k,j}(x) s_p(u_q^n(t,x)) dxdt.$$

The sum over the index j in the numerical solution u_p^n is reduced only to one term $u_{p,j}^n$ which has values on I_j and the other terms have value zero on I_j. By the index $k = 0, \ldots, N$ we have $N+1$ equations for the $N+1$ unknown coefficients $\widehat{u}_{k,j}^n(t)$. Now with $u_{p,j}^n(t,x) = \sum_{\ell=0}^{N} \widehat{u}_{p,\ell,j}^n(t)\Phi_{\ell,j}(x)$,

we have

$$\int_{\breve{T}_n}\int_{I_j}\Phi_{k,j}(x)\left(\sum_{\ell=0}^{N}\widehat{u}^n_{p,\ell,j}(t)\Phi_{\ell,j}(x)\right)_t dxdt+\int_{\breve{T}_n}\Phi_{k,j}(x)f_p(u^n_{q,j}(t,x))\Big|_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}}dt$$
$$-\int_{\breve{T}_n}\int_{I_j}(\Phi_{k,j}(x))_x f_p\left(u^n_{q,j}(t,x)\right)dxdt=\int_{\breve{T}_n}\int_{I_j}\Phi_{k,j}(x)s_p\left(u^n_{q,j}(t,x)\right)dxdt. \tag{8.2}$$

8.2.3.1 The First Term in (8.2)

The time derivative and the time integral in the first term are applied to $\widehat{u}^n_{p,\ell,j}$, since the functions $\Phi_{\ell,j}$ are independent of the time variable. This derivative can be replaced as follows for $x\in I_j$

$$\int_{\breve{T}_n}\left(\sum_{\ell=0}^{N}\widehat{u}^n_{p,\ell,j}(t)\Phi_{\ell,j}(x)\right)_t dt=\sum_{\ell=0}^{N}\left(\int_{\breve{T}_n}\left(\widehat{u}^n_{p,\ell,j}(t)\right)_t dt\right)\Phi_{\ell,j}(x)=\sum_{\ell=0}^{N}\left(\widehat{u}^{n+1}_{p,\ell,j}-\widehat{u}^n_{p,\ell,j}\right)\Phi_{\ell,j}(x).$$

Thus we have for the first term in (8.2)

$$\int_{I_j}\Phi_{k,j}(x)\sum_{\ell=0}^{N}\left(\widehat{u}^{n+1}_{p,\ell,j}-\widehat{u}^n_{p,\ell,j}\right)\Phi_{\ell,j}(x)dx.$$

Due to the orthogonality of the basis functions, the sum in the first term reduces to one term with index $k=\ell$ and the space integral has the value $\frac{h}{2k+1}$. Thus the first term becomes

$$\frac{h}{2k+1}\left(\widehat{u}^{n+1}_{p,k,j}-\widehat{u}^n_{p,k,j}\right).$$

8.2.3.2 The Second Term in (8.2)

According to the special values (2.6) of the Legendre polynomials, we have $\Phi_{k,j}(x_{j-\frac{1}{2}})=(-1)^k$ and $\Phi_{k,j}(x_{j+\frac{1}{2}})=1$ for $k=0,\ldots,N$. Then we have for the second term in (8.2)

$$\int_{\breve{T}_n}\left[f_p\left(u^n_{q,j}(t,x_{j+\frac{1}{2}})\right)-(-1)^k f_p\left(u^n_{q,j}(t,x_{j-\frac{1}{2}})\right)\right]dt.$$

Substituting these first and second terms in (8.2) we obtain the following system

$$\frac{h}{2k+1}\left(\widehat{u}^{n+1}_{p,k,j}-\widehat{u}^n_{p,k,j}\right)+\int_{\breve{T}_n}\left[f_p\left(u^n_{q,j}(t,x_{j+\frac{1}{2}})\right)-(-1)^k f_p\left(u^n_{q,j}(t,x_{j-\frac{1}{2}})\right)\right]dt$$
$$-\int_{\breve{T}_n}\int_{I_j}(\Phi_{k,j}(x))_x f_p\left(u^n_{q,j}(t,x)\right)dxdt=\int_{\breve{T}_n}\int_{I_j}\Phi_{k,j}(x)s_p\left(u^n_{q,j}(t,x)\right)dxdt. \tag{8.3}$$

8.2.4 Step 4.

Since the piecewise test functions $\Phi_{k,j}$ are discontinuous at the interfaces between the elements, then along an element boundary two distinct values of the solution can be obtained. To remove this non uniqueness in the second term the flux function f_p is replaced by a numerical flux function $\mathcal{F}_p$ which produces a single unique value. We can use any consistent monotone numerical flux that guarantees stable schemes and implies the convergence to the entropy solutions. For more details see [5].

For the values $f_p\left(u^n_{q,j}(t,x_{j+\frac{1}{2}})\right)$, we follow Dumbser et al. [7] and use numerical flux functions whose arguments are the solutions $U^n_{q,j}$, see (7.4), of the local Galerkin scheme, i.e.

$$f_p\left(u^n_{q,j}(t,x_{j+\frac{1}{2}})\right) \approx \mathcal{F}^n_{p,j+\frac{1}{2}}(t) := \mathcal{F}_p\left(U^n_{q,j}(t,x_{j+\frac{1}{2}}), U^n_{q,j+1}(t,x_{j+\frac{1}{2}})\right).$$

Similarly, we have

$$f_p\left(u^n_{q,j}(t,x_{j-\frac{1}{2}})\right) \approx \mathcal{F}^n_{p,j-\frac{1}{2}}(t) := \mathcal{F}_p\left(U^n_{q,j-1}(t,x_{j-\frac{1}{2}}), U^n_{q,j}(t,x_{j-\frac{1}{2}})\right).$$

We insert also these solutions $U^n_{q,j}$ in the formulas of the flux and source terms

$$\begin{aligned} f_p\left(u^n_{q,j}(t,x)\right) &\approx \sum_{i=1}^{\mathcal{N}} f_p\left(\widehat{U}^n_{q,i,j}\right)\theta_{i,j}(t,x), \\ s_p\left(u^n_{q,j}(t,x)\right) &\approx \sum_{i=1}^{\mathcal{N}} s_p\left(\widehat{U}^n_{q,i,j}\right)\theta_{i,j}(t,x), \quad \text{for } (t,x)\in \breve{T}_n\times I_j. \end{aligned}$$

Substituting into (8.3) we get using (7.4)

$$\begin{aligned} \frac{h}{2k+1}\left(\widehat{u}^{n+1}_{p,k,j}-\widehat{u}^n_{p,k,j}\right) &+ \int_{\breve{T}_n}\mathcal{F}^n_{p,j+\frac{1}{2}}(t)dt-(-1)^k\int_{\breve{T}_n}\mathcal{F}^n_{p,j-\frac{1}{2}}(t)dt \\ &- \int_{\breve{T}_n}\int_{I_j}(\Phi_{k,j}(x))_x\left[\sum_{i=1}^{\mathcal{N}} f_p\left(\widehat{U}^n_{q,i,j}\right)\theta_{i,j}(t,x)\right]dxdt \\ &= \int_{\breve{T}_n}\int_{I_j}\Phi_{k,j}(x)\left[\sum_{i=1}^{\mathcal{N}} s_p\left(\widehat{U}^n_{q,i,j}\right)\theta_{i,j}(t,x)\right]dxdt. \end{aligned}$$

The coefficients $f_p\left(\widehat{U}^n_{q,i,j}\right)$ and $s_p\left(\widehat{U}^n_{q,i,j}\right)$ are constants, thus, using the scalar product $\langle\cdot,\cdot\rangle_{tx}$, given by (7.5), we get

$$\begin{aligned} &\frac{h}{2k+1}\left(\widehat{u}^{n+1}_{p,k,j}-\widehat{u}^n_{p,k,j}\right)+\int_{\breve{T}_n}\mathcal{F}^n_{p,j+\frac{1}{2}}(t)dt-(-1)^k\int_{\breve{T}_n}\mathcal{F}^n_{p,j-\frac{1}{2}}(t)dt \\ &-\sum_{i=1}^{\mathcal{N}} f_p\left(\widehat{U}^n_{q,i,j}\right)\langle(\Phi_{k,j})_x,\theta_{i,j}\rangle_{tx}=\sum_{i=1}^{\mathcal{N}} s_p\left(\widehat{U}^n_{q,i,j}\right)\langle\Phi_{k,j},\theta_{i,j}\rangle_{tx}. \end{aligned}$$

8.2.5 Step 5.

Finally, by rearranging the terms, we get the fully discrete one step $P_N P_M$ DG scheme

$$\begin{aligned}\widehat{u}_{p,k,j}^{n+1} &= \widehat{u}_{p,k,j}^{n} - \frac{2k+1}{h}\int_{\breve{T}_n} \mathcal{F}_{p,j+\frac{1}{2}}^{n}(t)dt + (-1)^k \frac{2k+1}{h}\int_{\breve{T}_n} \mathcal{F}_{p,j-\frac{1}{2}}^{n}(t)dt \\ &+ \frac{2k+1}{h}\sum_{i=1}^{\mathcal{N}} f_p\left(\widehat{U}_{q,i,j}^{n}\right)\langle(\Phi_{k,j})_x, \theta_{i,j}\rangle_{tx} + \frac{2k+1}{h}\sum_{i=1}^{\mathcal{N}} s_p\left(\widehat{U}_{q,i,j}^{n}\right)\langle\Phi_{k,j}, \theta_{i,j}\rangle_{tx}.\end{aligned} \tag{8.4}$$

The equations (8.4) give the updates of $\widehat{u}_{p,k,j}^{n}$ from the time t_n to t_{n+1}. The numerical discrete solution updated at the new time t_{n+1} related to the element I_j is the term

$$u_{p,j}^{n+1}(x) = \sum_{k=0}^{N} \widehat{u}_{p,k,j}^{n+1}\Phi_{k,j}(x), \quad \text{for} \quad (t,x) \in \breve{T}_n \times I_j,$$

taking the summation over all elements of the discretization we get

$$u_{p}^{n+1}(x) = \sum_{j=1}^{Z}\sum_{k=0}^{N} \widehat{u}_{p,k,j}^{n+1}\Phi_{k,j}(x), \quad \text{for} \quad (t,x) \in \breve{T}_n \times I.$$

Chapter 9

Linear Advection Equation and Numerical Studies

Starting from this chapter for previty we say only the $P_N P_M$ schemes.

Let us consider the scalar linear advection equation $v_t(t,x) + av_x(t,x) = 0$ for $(t,x) \in [0,T] \times I$ with $T > 0$, $I = [\epsilon_1, \epsilon_2] \subset \mathbb{R}$, and $a \in \mathbb{R}$. Suppose that the initial solution is $v_0 = v(0,\cdot) \in L^2(I)$. The exact solution is given by $v_e(t,x) = v_0(x-at)$.

9.1 The $P_N P_M$ Schemes

We use the Lax-Friedrichs flux given by Cockburn and Shu [6]

$$\mathcal{F}^n_{LF,j+\frac{1}{2}}(t) = \frac{1}{2}\left[aU^n_{j+1}(t,x_{j+\frac{1}{2}}) + aU^n_j(t,x_{j+\frac{1}{2}}) - C\left(U^n_{j+1}(t,x_{j+\frac{1}{2}}) - U^n_j(t,x_{j+\frac{1}{2}})\right)\right],$$

$$\text{where} \quad C = \max_{\min_I(v_0)\leqslant s\leqslant \max_I(v_0)} |f'(s)| = |a|.$$

We have the following two cases. If $a > 0$ we have

$$\mathcal{F}^n_{LF,j+\frac{1}{2}}(t) = aU^n_j(t,x_{j+\frac{1}{2}}) = a\sum_{i=1}^{\mathcal{N}} \widehat{U}^n_{i,j}\theta_{i,j}(t,x_{j+\frac{1}{2}})$$

$$\mathcal{F}^n_{LF,j-\frac{1}{2}}(t) = aU^n_{j-1}(t,x_{j-\frac{1}{2}}) = a\sum_{i=1}^{\mathcal{N}} \widehat{U}^n_{i,j-1}\theta_{i,j-1}(t,x_{j-\frac{1}{2}})$$

and substituting into (8.4) we obtain

$$\begin{aligned}\widehat{u}^{n+1}_{k,j} &= \widehat{u}^n_{k,j} - \frac{(2k+1)a}{h}\sum_{i=1}^{\mathcal{N}}\left\{\widehat{U}^n_{i,j}\left(\int_{\breve{T}_n}\theta_{i,j}(t,x_{j+\frac{1}{2}})dt\right)\right. \\ &\left. -(-1)^k\widehat{U}^n_{i,j-1}\left(\int_{\breve{T}_n}\theta_{i,j-1}(t,x_{j-\frac{1}{2}})dt\right) - \widehat{U}^n_{i,j}\langle(\Phi_{k,j})_x,\theta_{i,j}\rangle_{tx}\right\}, \qquad (9.1)\end{aligned}$$

If $a < 0$ we have

$$\mathcal{F}^n_{LF,j+\frac{1}{2}}(t) = aU^n_{j+1}(t, x_{j+\frac{1}{2}}) = a\sum_{i=1}^{\mathcal{N}} \widehat{U}^n_{i,j+1}\theta_{i,j+1}(t, x_{j+\frac{1}{2}})$$

$$\mathcal{F}^n_{LF,j-\frac{1}{2}}(t) = aU^n_{j}(t, x_{j-\frac{1}{2}}) = a\sum_{i=1}^{\mathcal{N}} \widehat{U}^n_{i,j}\theta_{i,j}(t, x_{j-\frac{1}{2}})$$

and

$$\begin{aligned}\widehat{u}^{n+1}_{k,j} &= \widehat{u}^n_{k,j} - \frac{(2k+1)a}{h}\sum_{i=1}^{\mathcal{N}}\left\{\widehat{U}^n_{i,j+1}\left(\int_{\check{T}_n}\theta_{i,j+1}(t, x_{j+\frac{1}{2}})dt\right)\right. \\ &\left. -(-1)^k\widehat{U}^n_{i,j}\left(\int_{\check{T}_n}\theta_{i,j}(t, x_{j-\frac{1}{2}})dt\right) - \widehat{U}^n_{i,j}\langle(\Phi_{k,j})_x, \theta_{i,j}\rangle_{tx}\right\}. \qquad (9.2)\end{aligned}$$

The $N+1$ relations (9.1) or (9.2) give updates of the degrees of freedom.

9.1.1 The Extra Elements

The space interval is always discretized into Z elements. The relations (9.1) (for $a > 0$) and (9.2) (for $a < 0$) indicate that the new degrees of freedom $\widehat{u}^{n+1}_{k,j}$ depend on the values $\widehat{u}^n_{k,j}$ and $\widehat{U}^n_{i,j}$, as well as on $\widehat{U}^n_{i,j-1}$ (for $a > 0$) or $\widehat{U}^n_{i,j+1}$ (for $a < 0$), of the past time t_n. The values $\widehat{U}^n_{i,j-1}$ and $\widehat{U}^n_{i,j+1}$ locate at the left element I_{j-1} and the right element I_{j+1}, respectively. Therefore, we add two extra element. Thus we need $Z+2$ elements.

In additional, the degrees of freedom $\widehat{U}^n_{i,\complement}$ for $\complement = j-1, j, j+1$ are related to the reconstructed degrees of freedom $\widehat{w}^n_{k,\complement}$, respectively. The degrees of freedom $\widehat{w}^n_{k,\complement}$ depend on the elements within the stencil which contains n_e elements.

- If $M = N$, then $n_e = 1$ and $\widehat{w}^n_{k,\complement} = \widehat{u}^n_{k,\complement}$. We do not need any neighbors.
- If $M > N$, then $n_e > 1$ and $\widehat{w}^n_{k,\complement}$ need $n_e - 1$ new neighbors. Then we have to add $(2n_e - 2)$ elements in order to cover all shapes of the stencil.

Thus we have to add $2n_e$ extra elements. For example, the $P_N P_M$ schemes with $N = M$ have single stencils, i.e. $ne = 1$, and need only two extra elements, (one at left, one at right). The stencils of size 2 require 4 extra elements, (2 at left, 2 at right), and so on.

9.1.2 The Boundary Conditions

Let us discretize the space interval into Z elements and attach them with the extra elements. The relations (9.1) and (9.2) give updates only to the main Z elements. In order to continue with the time variable we have to update the solutions on the $2n_e$ extra elements. Therefore, we need to add $2n_e$ boundary conditions.

Example: Periodic Initial Function

Let $M, N \in \mathbb{N}$ with $M \geqslant N$. We discretize the space interval $[\epsilon_1, \epsilon_2]$ into Z elements and add $2n_e$ extra elements. The boundary conditions are given in the Table 9.1.

The size of the stencil, n_e	The number of all elements	The Boundary Conditions for all $i = 0, 1, \ldots, N$	
		Left	Right
1	$Z+2$	$\widehat{u}_{i,1}^{n+1} = \widehat{u}_{i,Z+1}^{n+1}$	$\widehat{u}_{i,Z+2}^{n+1} = \widehat{u}_{i,2}^{n+1}$
2	$Z+4$	$\widehat{u}_{i,1}^{n+1} = \widehat{u}_{i,Z+1}^{n+1}$ $\widehat{u}_{i,2}^{n+1} = \widehat{u}_{i,Z+2}^{n+1}$	$\widehat{u}_{i,Z+3}^{n+1} = \widehat{u}_{i,3}^{n+1}$ $\widehat{u}_{i,Z+4}^{n+1} = \widehat{u}_{i,4}^{n+1}$
3	$Z+6$	$\widehat{u}_{i,1}^{n+1} = \widehat{u}_{i,Z+1}^{n+1}$ $\widehat{u}_{i,2}^{n+1} = \widehat{u}_{i,Z+2}^{n+1}$ $\widehat{u}_{i,3}^{n+1} = \widehat{u}_{i,Z+3}^{n+1}$	$\widehat{u}_{i,Z+4}^{n+1} = \widehat{u}_{i,4}^{n+1}$ $\widehat{u}_{i,Z+5}^{n+1} = \widehat{u}_{i,5}^{n+1}$ $\widehat{u}_{i,Z+6}^{n+1} = \widehat{u}_{i,6}^{n+1}$

Table 9.1: The boundary conditions of the $P_N P_M$ schemes applied to the advection equation with a periodic initial function.

9.1.3 Some $P_N P_M$ Formulas for $a > 0$

The arguments of the discretization k_n and h are related, see [5], by the Courant number $\lambda = \frac{ak_n}{h}$. This number is important for the stability of the schemes.

The P_0P_0 Scheme

$$\widehat{u}_{0,j}^{n+1} = \widehat{u}_{0,j}^{n} + \lambda \left(\widehat{u}_{0,j-1}^{n} - \widehat{u}_{0,j}^{n} \right). \tag{9.3}$$

The P_0P_1 Scheme with $S_{I_j,2,1}$

$$\widehat{u}_{0,j}^{n+1} = \widehat{u}_{0,j}^{n} - \frac{\lambda}{2} \left(\widehat{u}_{0,j-2}^{n} - 4\widehat{u}_{0,j-1}^{n} + 3\widehat{u}_{0,j}^{n} \right) + \frac{\lambda^2}{2} \left(\widehat{u}_{0,j-2}^{n} - 2\widehat{u}_{0,j-1}^{n} + \widehat{u}_{0,j}^{n} \right).$$

The P_0P_1 Scheme with $S_{I_j,2,0}$

$$\widehat{u}_{0,j}^{n+1} = \widehat{u}_{0,j}^{n} + \frac{\lambda}{2} \left(\widehat{u}_{0,j-1}^{n} - \widehat{u}_{0,j+1}^{n} \right) + \frac{\lambda^2}{2} \left(\widehat{u}_{0,j-1}^{n} - 2\widehat{u}_{0,j}^{n} + \widehat{u}_{0,j+1}^{n} \right).$$

The P_1P_1 Scheme

$$\widehat{u}_{0,j}^{n+1} = \widehat{u}_{0,j}^{n} + \lambda\left(\widehat{u}_{0,j-1}^{n} - \widehat{u}_{0,j}^{n} + \widehat{u}_{1,j-1}^{n} - \widehat{u}_{1,j}^{n}\right) - \lambda^2\left(\widehat{u}_{1,j-1}^{n} - \widehat{u}_{1,j}^{n}\right),$$

$$\widehat{u}_{1,j}^{n+1} = \widehat{u}_{1,j}^{n} - 3\lambda\left(\widehat{u}_{0,j-1}^{n} - \widehat{u}_{0,j}^{n} + \widehat{u}_{1,j-1}^{n} + \widehat{u}_{1,j}^{n}\right) + 3\lambda^2\left(\widehat{u}_{1,j-1}^{n} - \widehat{u}_{1,j}^{n}\right).$$

The P_0P_2 Scheme with $S_{I_j,3,2}$

$$\begin{aligned}\widehat{u}_{0,j}^{n+1} &= \widehat{u}_{0,j}^{n} + \frac{\lambda}{6}\left(2\widehat{u}_{0,j-3}^{n} - 9\widehat{u}_{0,j-2}^{n} + 18\widehat{u}_{0,j-1}^{n} - 11\widehat{u}_{0,j}^{n}\right) - \frac{\lambda^2}{2}\left(\widehat{u}_{0,j-3}^{n} - 4\widehat{u}_{0,j-2}^{n}\right.\\ &\quad \left. + 5\widehat{u}_{0,j-1}^{n} - 2\widehat{u}_{0,j}^{n}\right) + \frac{\lambda^3}{6}\left(\widehat{u}_{0,j-3}^{n} - 3\widehat{u}_{0,j-2}^{n} + 3\widehat{u}_{0,j-1}^{n} - \widehat{u}_{0,j}^{n}\right).\end{aligned}$$

The P_0P_2 Scheme with $S_{I_j,3,1}$

$$\begin{aligned}\widehat{u}_{0,j}^{n+1} &= \widehat{u}_{0,j}^{n} - \frac{\lambda}{6}\left(\widehat{u}_{0,j-2}^{n} - 6\widehat{u}_{0,j-1}^{n} + 3\widehat{u}_{0,j}^{n} + 2\widehat{u}_{0,j+1}^{n}\right) + \frac{\lambda^2}{2}\left(\widehat{u}_{0,j-1}^{n} - 2\widehat{u}_{0,j}^{n} + \widehat{u}_{0,j+1}^{n}\right)\\ &\quad + \frac{\lambda^3}{6}\left(\widehat{u}_{0,j-2}^{n} - 3\widehat{u}_{0,j-1}^{n} + 3\widehat{u}_{0,j}^{n} - \widehat{u}_{0,j+1}^{n}\right).\end{aligned}$$

The P_0P_2 Scheme with $S_{I_j,3,0}$

$$\begin{aligned}\widehat{u}_{0,j}^{n+1} &= \widehat{u}_{0,j}^{n} + \frac{\lambda}{6}\left(2\widehat{u}_{0,j-1}^{n} + 3\widehat{u}_{0,j}^{n} - 6\widehat{u}_{0,j+1}^{n} + \widehat{u}_{0,j+2}^{n}\right) + \frac{\lambda^2}{2}\left(\widehat{u}_{0,j-1}^{n} - 2\widehat{u}_{0,j}^{n} + \widehat{u}_{0,j+1}^{n}\right)\\ &\quad + \frac{\lambda^3}{6}\left(\widehat{u}_{0,j-1}^{n} - 3\widehat{u}_{0,j}^{n} + 3\widehat{u}_{0,j+1}^{n} - \widehat{u}_{0,j+2}^{n}\right).\end{aligned}$$

The P_1P_2 Scheme with $S_{I_j,2,1}$

$$\begin{aligned}\widehat{u}_{0,j}^{n+1} &= \widehat{u}_{0,j}^{n} + \tfrac{\lambda}{60}\left(9\widehat{u}_{0,j-2}^{n} - \widehat{u}_{1,j-2}^{n} + 42\widehat{u}_{0,j-1}^{n} + 80\widehat{u}_{1,j-1}^{n} - 51\widehat{u}_{0,j}^{n} - 79\widehat{u}_{1,j}^{n}\right)\\ &\quad - \tfrac{\lambda^2}{20}\left(9\left(\widehat{u}_{0,j-2}^{n} - 2\widehat{u}_{0,j-1}^{n} + \widehat{u}_{0,j}^{n}\right) - \widehat{u}_{1,j-2}^{n} + 40\widehat{u}_{1,j-1}^{n} - 39\widehat{u}_{1,j}^{n}\right)\\ &\quad + \tfrac{\lambda^3}{30}\left(9\left(\widehat{u}_{0,j-2}^{n} - 2\widehat{u}_{0,j-1}^{n} + \widehat{u}_{0,j}^{n}\right) - \widehat{u}_{1,j-2}^{n} + 20\widehat{u}_{1,j-1}^{n} - 19\widehat{u}_{1,j}^{n}\right),\\ \widehat{u}_{1,j}^{n+1} &= \widehat{u}_{1,j}^{n} - \tfrac{\lambda}{20}\left(9\widehat{u}_{0,j-2}^{n} - \widehat{u}_{1,j-2}^{n} + 60\widehat{u}_{0,j-1}^{n} + 78\widehat{u}_{1,j-1}^{n} - 69\widehat{u}_{0,j}^{n} + 79\widehat{u}_{1,j}^{n}\right)\\ &\quad + 3\tfrac{\lambda^2}{20}\left(9\left(\widehat{u}_{0,j-2}^{n} - \widehat{u}_{0,j}^{n}\right) - \widehat{u}_{1,j-2}^{n} + 38\widehat{u}_{1,j-1}^{n} - \widehat{u}_{1,j}^{n}\right)\\ &\quad - \tfrac{\lambda^3}{10}\left(9\left(\widehat{u}_{0,j-2}^{n} - 2\widehat{u}_{0,j-1}^{n} + \widehat{u}_{0,j}^{n}\right) - \widehat{u}_{1,j-2}^{n} + 20\widehat{u}_{1,j-1}^{n} - 19\widehat{u}_{1,j}^{n}\right).\end{aligned}$$

The P_2P_2 Scheme

$$\begin{aligned}
\widehat{u}_{0,j}^{n+1} &= \widehat{u}_{0,j}^{n} + \lambda\left(\widehat{u}_{0,j-1}^{n} + \widehat{u}_{1,j-1}^{n} + \widehat{u}_{2,j-1}^{n} - \widehat{u}_{0,j}^{n} - \widehat{u}_{1,j}^{n} - \widehat{u}_{2,j}^{n}\right) \\
&\quad -\lambda^2\left(\widehat{u}_{1,j-1}^{n} + 3\widehat{u}_{2,j-1}^{n} - \widehat{u}_{1,j}^{n} - 3\widehat{u}_{2,j}^{n}\right) + 2\lambda^3\left(\widehat{u}_{2,j-1}^{n} - \widehat{u}_{2,j}^{n}\right), \\
\widehat{u}_{1,j}^{n+1} &= \widehat{u}_{1,j}^{n} - 3\lambda\left(\widehat{u}_{0,j-1}^{n} + \widehat{u}_{1,j-1}^{n} + \widehat{u}_{2,j-1}^{n} - \widehat{u}_{0,j}^{n} + \widehat{u}_{1,j}^{n} + \widehat{u}_{2,j}^{n}\right) \\
&\quad +3\lambda^2\left(\widehat{u}_{1,j-1}^{n} + 3\widehat{u}_{2,j-1}^{n} - \widehat{u}_{1,j}^{n} + 3\widehat{u}_{2,j}^{n}\right) - 6\lambda^3\left(\widehat{u}_{2,j-1}^{n} - \widehat{u}_{2,j}^{n}\right), \\
\widehat{u}_{2,j}^{n+1} &= \widehat{u}_{2,j}^{n} + 5\lambda\left(\widehat{u}_{0,j-1}^{n} + \widehat{u}_{1,j-1}^{n} + \widehat{u}_{2,j-1}^{n} - \widehat{u}_{0,j}^{n} + \widehat{u}_{1,j}^{n} - \widehat{u}_{2,j}^{n}\right) \\
&\quad -5\lambda^2\left(\widehat{u}_{1,j-1}^{n} + 3\widehat{u}_{2,j-1}^{n} - \widehat{u}_{1,j}^{n} + 3\widehat{u}_{2,j}^{n}\right) + 10\lambda^3\left(\widehat{u}_{2,j-1}^{n} - \widehat{u}_{2,j}^{n}\right).
\end{aligned}$$

9.2 Numerical studies

In fact, there are several parameters which control the P_NP_M schemes, namely: (1) The orders N and M. (2) The size n_e of any stencil $S_{I_j,n_e,L}$, that must satisfy the condition $n_e \geqslant \frac{M+1}{N+1}$. (3) The index L of the stencil $S_{I_j,n_e,L}$ that indicates the form of the stencil. (4) The mesh size Z which gives the length h of the elements. (5) The maximal time value T with the time step $k_n \leqslant T$. (6) The Courant number $\lambda = \frac{|a|k_n}{h}$ which relates the time step k_n to the mesh length h, see [5].

In the following we study the stability and efficiency of the P_NP_M schemes by studying three of these parameters, namely, the Courant number as well as the size and form of the stencils.

9.2.1 Stability Analysis

The Courant number is important for the stability of the schemes. We determine maximal Courant numbers which are limits of the stability. We study the P_NP_M schemes applying to the linear advection equation $v_t + v_x = 0$ with $a = 1$. At first, we apply the von Neumann stability analysis [16] with the special case $N = 0$. Then we follow an experimental procedure for higher order P_NP_M schemes with $N > 0$.

9.2.1.1 Von Neumann Analysis

The computational domain of the Fourier representations is the region $[-z, z]$ which is discretized into $2Z_f$ mesh elements with equidistant length element $h_f = z/Z_f$ and $z \in \mathbb{R}$ is a period of the initial data. We decompose the coefficients $\widehat{u}_{0,j}^n$ inside the element I_j, into a Fourier sum as

$$\widehat{u}_{0,j}^{n} = \sum_{\ell=-Z_f}^{Z_f} \mathcal{A}_\ell^n e^{ij\phi_\ell}, \tag{9.4}$$

where $\mathcal{A}_\ell^n$ is called the amplitude vector at time level t_n, $i = \sqrt{-1}$ is the imaginary unit, and ϕ_ℓ is the wave number and it is given by $\phi_\ell = \ell\pi/Z_f$ with $\ell = -Z_f, \ldots, Z_f$. This finite sum

splits the time dependence from the spatial one, where the time evolution is included in the time dependence of the amplitude $\mathcal{A}_\ell^n$.

Now we substitute this finite sum into the scheme considered. Then, dividing by $e^{ij\phi_\ell}$, we obtain a relation between the amplitude vectors $\mathcal{A}_\ell^n$ and $\mathcal{A}_\ell^{n+1}$ with some space shifts $e^{\mp i\phi_\ell}$. This relation can be written as $\mathcal{A}_\ell^{n+1} = D_\ell \mathcal{A}_\ell^n$, where D_ℓ is called the amplification factor for $\ell = -Z_f, \ldots, Z_f$.

The stability condition of the von Neumann analysis states that the Euclidean norm of the amplitude vector $\mathcal{A}_\ell^n$ for any wave number ϕ_ℓ does not grow in time. In other words, this condition can be written as $|D_\ell| \leqslant 1$ for all ϕ_ℓ.

The P_0P_0 Scheme

In this case we can obtain the Courant number $\lambda = \frac{k_n}{h}$ exactly. Substituting (9.4) in (9.3) we get, for $\ell = -Z_f, \ldots, Z_f$,

$$\mathcal{A}_\ell^{n+1} e^{ij\phi_\ell} = \mathcal{A}_\ell^n e^{ij\phi_\ell} + \lambda \left(\mathcal{A}_\ell^n e^{i(j-1)\phi_\ell} - \mathcal{A}_\ell^n e^{ij\phi_\ell} \right) = \mathcal{A}_\ell^n e^{ij\phi_\ell} \left(1 + \lambda e^{-i\phi_\ell} - \lambda \right).$$

Dividing by $e^{ij\phi_\ell}$ we get $\mathcal{A}_\ell^{n+1} = \underbrace{\left(1 - \lambda + \lambda e^{-i\phi_\ell}\right)}_{D_\ell} \mathcal{A}_\ell^n$. Then we have

$$\begin{aligned} |D_\ell|^2 &= D_\ell^T D_\ell = \left(1 - \lambda + \lambda e^{i\phi_\ell}\right)\left(1 - \lambda + \lambda e^{-i\phi_\ell}\right) \\ &= 1 + \lambda \left[-2 + e^{i\phi_\ell} + e^{-i\phi_\ell}\right] + \lambda^2 \left[1 - e^{i\phi_\ell} - e^{-i\phi_\ell} + e^{-i\phi_\ell} e^{i\phi_\ell}\right] \\ &= 1 + \lambda[-2 + 2\cos(\phi_\ell)] + \lambda^2[2 - 2\cos(\phi_\ell)] = 1 + 2\lambda(\lambda - 1)[1 - \cos(\phi_\ell)]. \end{aligned}$$

The condition $\max |D_\ell| \leqslant 1$ is equivalent to $\lambda(\lambda - 1)[1 - \cos(\phi_\ell)] \leqslant 0$, this indicates directly that the Courant number is bounded by $0 < \lambda \leqslant 1$.

For other P_NP_M Schemes with $N = 0$

We can determine the maximal Courant numbers numerically. The amplification factor D_ℓ is a function of two variables $D_\ell = D_\ell(\phi_\ell, \lambda)$. We take $Z_f = 3$ then $\ell = -3, \ldots, 3$ and $\phi_\ell \in \{-\pi, -2\pi/3, -\pi/3, 0, \pi/3, 2\pi/3, \pi\}$, and define the variable $\lambda_s = s/10$ with $1 \leqslant s \leqslant 30$ which covers the interval $[1/10, 3]$. Then we compute the modulus of D_ℓ at each value of ϕ_ℓ and of λ, then we get a 7×30 matrix of these values. Each column is related to one value of λ_s. If all entries of the column are less or equal to one then the value λ_s, to which this column associated, gives a stable solution of the scheme.

For example, the P_0P_1 scheme with stencil $S_{I_j,2,0}$, we obtain the following matrix

$$\begin{pmatrix} 0.98 & 0.92 & \dots & 0.62 & 1 & 1.42 & 1.88 & \dots \\ 0.98 & 0.95 & \dots & 0.80 & 1 & 1.25 & 1.55 & \dots \\ 0.99 & 0.99 & \dots & 0.98 & 1 & 1.03 & 1.07 & \dots \\ 1 & 1 & \dots & 1 & 1 & 1 & 1 & \dots \\ 0.99 & 0.99 & \dots & 0.98 & 1 & 1.03 & 1.07 & \dots \\ 0.98 & 0.95 & \dots & 0.80 & 1 & 1.25 & 1.55 & \dots \\ 0.98 & 0.92 & \dots & 0.62 & 1 & 1.42 & 1.88 & \dots \\ \Uparrow & \Uparrow & & \Uparrow & \Uparrow & \Uparrow & & \\ \lambda_1 = 0.1 & \lambda_2 = 0.2 & & \lambda_9 = 0.9 & \lambda_{10} = 1 & \lambda_{11} = 1.1 & & \end{pmatrix}$$

Note that, starting from the eleventh column, the entries are larger than one. This proves that the value $\lambda_{11} = 1.1$ gives an unstable solution. Thus the maximum value of λ_s which gives a stable solution is λ_{10}, thus $\lambda_{\max} \approx \lambda_{10} = 1$. On the other hand, all columns to the left of the eleventh have entries less or equal one, thus their λ_s give stable solutions. We obtain that the range of the Courant number for the P_0P_1 scheme using $S_{I_j,2,0}$ is the interval $\lambda \in (0, 1]$.

n_e	L	P_0P_1	P_0P_2	P_0P_3	P_0P_4	P_0P_5
2	0	$(0,1]$				
	1	$(\mathbf{0},\mathbf{2}]$				
3	0	$(0,1]$	$*$			
	1	$(0,1]$	$(0,1]$			
	2	$(0,1]$	$[\mathbf{1},\mathbf{2}]$			
4	0	$(0,1]$	$*$	$*$		
	1	$(0,1]$	$(0,1]$	$(0,1]$		
	2	$(0,1]$	$(0,1]$	$(\mathbf{0},\mathbf{2}]$		
	3	$(0,1]$	$*$	$[\mathbf{1},\mathbf{2}]$		
5	0	$(0,1]$	$(\mathbf{0.5},\mathbf{1}]$	$*$	$*$	
	1	$(0,1]$	$(0,1]$	$(0,1]$	$*$	
	2	$(0,1]$	$(0,1]$	$(0,1]$	$(0,1]$	
	3	$(0,1]$	$(0,1]$	$(0,1]$	$[\mathbf{1},\mathbf{2}]$	
	4	$(0,1]$	$*$	$*$	$*$	
6	0	$(0,1]$	$(0,1]$	$*$	$*$	$*$
	1	$(0,1]$	$(0,1]$	$(0,1]$	$*$	$*$
	2	$(0,1]$	$(0,1]$	$(0,1]$	$(0,1]$	$(0,1]$
	3	$(0,1]$	$(0,1]$	$(0,1]$	$(0,1]$	$(0,1]$
	4	$(0,1]$	$(0,1]$	$(0,1]$	$(0,1]$	$*$
	5	$(0,1]$	$(0,1]$	$*$	$[\mathbf{1},\mathbf{2}]$	$*$

Table 9.2: The maximal Courant numbers for some P_NP_M schemes, for $N = 0$ and $M = 1, 2, 3, 4, 5$.

Table 9.2 includes the maximal limits $\lambda_{\max}$, which are computed numerically in this way, of

the Courant numbers for the P_0P_M schemes with $N = 0$ and various orders $M = 1, \ldots, 5$ with all cases of the stencils $S_{I_j,n_e,L}$ with $n_e = \left\lceil \frac{M+1}{N+1} \right\rceil, \ldots, 6$ and $L = 0, \ldots, n_e - 1$. The symbol $*$ indicates unstable cases for which we have only one value $\lambda = 1$ that gives a stable solution. The fact that $\lambda = 1$ is stable is an artifact due to the equation $v_t + v_x = 0$, because for $\lambda = 1$ the numerical solution of these schemes is the exact solution. Moreover, the range $(0, 1]$ mostly appears, but there are some semi-stable cases which are written in boldface. Also, there are two cases with higher stability $\lambda \in (0, 2]$ that are also highlighted in boldface.

9.2.1.2 Another Experimental Procedure

The von Neumann analysis for higher order P_NP_M schemes with $N > 0$ is not possible without the use of computer algebra and numerical computation, see Dumbser [7, p. 8221]. Therefore, we consider another numerical procedure. We continue in the study of the advection equation $v_t + v_x = 0$ with the initial solution $v_0(x) = \sin(x)$ for $x \in [0, 2\pi]$ and periodicity as the boundary condition. So we have $v_0(x) = \sin(x)$ for all $x \in \mathbb{R}$. It is well known that the exact solution is $v_e(t, x) = v_0(x - t)$ on $[0, T] \times [0, 2\pi]$.

We found experimentally appropriate limits of the Courant numbers which guarantee the stability without resorting to von Neumann analysis. We checked the stability of the numerical solutions at the final time $T = 100\pi$ for a mesh with $Z = 50$. Let us set $\lambda_C := 1/(2N + 1)$. Cockburn [5] considered Runge-Kutta DG schemes and took for the linear equations λ somewhat smaller than λ_C as a limit in order to avoid unstable solutions. We start with this inequality and define variables λ_s in an interval around λ_C as follows $\lambda_s = \alpha_N\lambda_C + 0.001s$ for $s = 0, 1, 2, \ldots$, where $0 < \alpha_N < 1$ is constant associated to the order N and determines the starting point of a search algorithm. We use the values $\alpha_0 = 0.99$, $\alpha_1 = 0.9$, $\alpha_2 = 0.8$, $\alpha_3 = 0.7$, $\alpha_4 = 0.6$, and $\alpha_5 = 0.5$.

Increasing the index s, the variable λ_s comes closer to the ratio λ_C and then larger than λ_C. For each value λ_s, we associate the value $L_s^1 = \int_I |v_e(T, x) - w(T, x)|dx$, which is the L^1 error of the reconstructed polynomial (solution) w computed using the P_NP_M scheme at the last time T with time step $\Delta t = \lambda_s h$. We compute the errors L_s^1 numerically using Gaussian rules of enough large orders. As well, we compute the differences $d_{1,s} = |L_s^1 - L_{s-1}^1|$ for $s > 0$, defining $d_{1,0} = 0$. We also set a condition to stop this algorithm which is $d_{1,s} > TOL$, where TOL is a tolerance that we choose enough large, e.g. $TOL = 10$, to guarantee that the L^1 error is large, and this means that the solution is unstable. For example, we consider the P_2P_2 scheme. Then we have $\lambda_C = 0.2$, $\alpha_2 = 0.8$ and $\lambda_s = 0.16 + 0.001s$. We arrange the errors in the following table which leads to the conclusion that $\lambda_{\max} \approx 0.171$.

s	2	3	4	5	6	7	8	9	10	11	12
λ_s	0.162	0.163	0.164	0.165	0.166	0.167	0.168	0.169	0.170	**0.171**	**0.172**
L_s^1	0.074	0.047	0.009	0.040	0.064	0.079	0.005	0.009	0.011	**0.018**	$\mathbf{4 \times 10^{74}}$

Note that in solution plots the solution for $\lambda_s = 0.170$ looks smooth, whereas for $\lambda_s = 0.171$ small oscillations occur that become stronger for larger λ_s.

In the following, we give the approximate values of $\lambda_{\max}$ for all $P_N P_M$ schemes for

$$M = 0, \ldots, 5, \quad N = 0, \ldots, M, \quad n_e = \left\lceil \frac{M+1}{N+1} \right\rceil, \ldots, 6, \quad L = 0, \ldots, n_e - 1.$$

9.2.1.2.1 Case $M = 1$. For the P_1P_1 scheme we obtain $\lambda_C = 0.333$ and $\lambda_{\max} \approx 1/3$. For the P_0P_1, see Table 9.3.

$n_e \backslash L$	0	1	2	3	4	5
2	1.003	2.006				
3	1.005	1.006	1.008			
4	1.005	1.006	1.007	1.009		
5	1.005	1.005	1.006	1.008	1.007	
6	1.006	1.006	1.005	1.008	1.007	1.007

Table 9.3: The maximal Courant numbers for P_0P_1 scheme with $\lambda_C = 1$.

9.2.1.2.2 Case $M = 2$. For the P_2P_2 scheme we obtain $\lambda_C = 0.2$ and $\lambda_{\max} \approx 0.17$. Also, Tables 9.4 give the approximations of $\lambda_{\max}$ for the P_0P_2 and P_1P_2 schemes.

The P_0P_2 schemes with $\lambda_C = 1$						
$n_e \backslash L$	0	1	2	3	4	5
3	1.002	1.01	[1,2]			
4	1.013	1.012	1.005	1.011		
5	1.003	1.01	1.006	1.005	1.011	
6	1.004	1.007	1.006	1.006	1.006	1.01
The P_1P_2 schemes with $\lambda_C = 0.333$						
$n_e \backslash L$	0	1	2	3	4	5
2	1/3	*				
3	1/3	1/3	*			
4	1/3	1/3	*	*		
5	1/3	1/3	1/3	*	*	
6	1/3	1/3	1/3	*	*	0.305

Table 9.4: The maximal Courant numbers for P_0P_2 and P_1P_2 schemes.

9.2.1.2.3 Case $M = 3$. For the P_2P_3 schemes where $\lambda_C = 0.2$ and with all stencils considered above we obtain $\lambda_{\max} \approx 0.17$ and for the P_3P_3 scheme where $\lambda_C = 0.143$ we find $\lambda_{\max} \approx 0.103$. For the P_0P_3 and P_1P_3 schemes, see Table 9.5.

The P_0P_3 schemes with $\lambda_C = 1$						
$n_e \backslash L$	0	1	2	3	4	5
4	1.001	1.005	2.009	[1,2]		
5	1.002	1.01	1.006	1.005	1.02	
6	1.002	1.011	1.006	1.006	1.006	1.017
The P_1P_3 schemes with $\lambda_C = 0.333$						
$n_e \backslash L$	0	1	2	3	4	5
2	0.318	*				
3	0.328	0.34	*			
4	0.331	0.33	0.338	*		
5	0.332	1/3	0.332	*	*	
6	0.332	0.332	0.316	0.335	*	*

Table 9.5: The maximal Courant numbers for P_0P_3 and P_1P_3 schemes.

9.2.1.2.4 Case $M = 4$. For the P_3P_4 schemes with $\lambda_C = 0.143$ we obtain $\lambda_{\max} \approx 0.103$ and for the P_4P_4 scheme where $\lambda_C = 0.111$ we find $\lambda_{\max} \approx 0.069$. For the P_0P_4, P_1P_4, and P_2P_4 schemes, see Table 9.6.

The P_0P_4 schemes with $\lambda_C = 1$						
$n_e \backslash L$	0	1	2	3	4	5
5	1	1.002	1.012	2.02	[1,1.5]	
6	1	1.006	1.012	1	1	[1,2]
The P_1P_4 schemes with $\lambda_C = 0.333$						
3	0.316	0.346	*			
4	0.325	0.347	*	*		
5	0.328	0.338	0.337	*	*	
6	0.330	0.338	0.337	*	0.312	*
The P_2P_4 schemes with $\lambda_C = 0.2$						
$n_e \backslash L$	0	1	2	3	4	5
2	0.166	*				
3	0.169	0.176	*			
4	0.170	0.173	0.173	*		
5	0.170	0.170	0.172	0.170	0.170	
6	0.170	0.170	0.172	0.172	0.170	0.170

Table 9.6: The maximal Courant numbers for P_0P_4, P_1P_4, and P_2P_3 schemes.

9.2.1.2.5 Case $M = 5$. For the P_4P_5 schemes where $\lambda_C = 0.111$ we obtain $\lambda_{\max} \approx 0.069$ and for the P_5P_5 scheme where $\lambda_C = 0.091$ we find $\lambda_{\max} \approx 0.05$. For the P_0P_5, P_1P_5, P_2P_5, and

P_3P_4 schemes, see Table 9.7.

The P_0P_5 schemes with $\lambda_C = 1$						
$n_e \backslash L$	0	1	2	3	4	5
6	1	1.001	1.005	[1,2]	[1,3]	1
The P_2P_5 schemes with $\lambda_C = 0.2$						
2	*	*				
3	0.165	0.176	*			
4	0.167	0.176	0.175	*		
5	0.168	0.175	0.172	0.174	*	
6	0.169	0.172	0.172	0.172	0.172	*
The P_1P_5 schemes with $\lambda_C = 0.333$						
$n_e \backslash L$	0	1	2	3	4	5
3	*	0.402	*			
4	*	0.346	*	*		
5	*	0.345	0.335	*	*	
6	0.324	0.344	0.327	0.34	*	*
The P_3P_5 schemes with $\lambda_C = 0.143$						
2	0.1	*				
3	0.103	0.106	0.102			
4	0.103	0.105	0.104	0.103		
5	0.103	0.103	0.104	0.103	0.103	
6	0.103	0.103	0.104	0.104	0.103	0.103

Table 9.7: The maximal Courant numbers for P_0P_5, P_1P_5, P_2P_5, and P_3P_5 schemes.

9.2.2 Experimantal Order of Convergence (EOC)

We investigate the orders of the accuracy numerically by calculating the EOC. Let generally X ba a linear space with some norm $||\cdot||_X$ and let $v_h \in X$ be a numerical approximation of a given function $v \in X$ which depends on a parameter h of the discretization. The convergence of v_h towards v as h tends to zero can be quantified by $||v_h - v||_X \leqslant Ch^\kappa$, with the order of convergence κ. This gives a possibility to quantify the quality of a numerical scheme. If we can compute two numerical solutions v_h and $v_{h'}$, then the order κ can be estimated experimentally by $\kappa \simeq EOC(h, h') = \frac{\log(||v_{h'} - v||_X / ||v_h - v||_X)}{\log(h'/h)}$.

The maximum Courant numbers computed above are quite sharp since oscillations occur with slightly longer time steps. Therefore, for our further tests, we used the following restrictive bounds on the Courant number, see Table 9.8.

Now we consider the advection equation $v_t + v_x = 0$ with $v(0, x) = \sin(x)$ defined on $I = [0, 2\pi]$ and its solution at time $T = 2\pi$. We apply some P_NP_M schemes. The CFL numbers λ are taken from the Table 9.8. The L^1 errors are listed in the Table 9.9 where we always used the

The order N	0	1	2	3	4	5
The Courant number λ_{used}	1	0.25	0.16	0.08	0.05	0.05

Table 9.8: The Courant numbers λ_{used} for $N = 0, \ldots, 5$.

stencil $S_{I_j,5,2}$. The numbers for the EOC were truncated after the second decimal. Note that we always get the expected order of convergence close to $M + 1$. Some of the schemes produce a wrong experimental order on the coarsest meshes. This is not a problem, since the order is an asymptotic property for $h \to 0$.

Z	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC
	P_0P_0									
10	6.23e-1									
20	3.17e-1	0.97								
40	1.58e-1	1.00								
	P_0P_1		P_1P_1							
10	1.50e-1		1.67e-1							
20	2.34e-2	2.68	4.16e-2	2.01						
40	4.17e-3	2.49	1.03e-2	2.00						
	P_0P_2		P_1P_2		P_2P_2					
10	1.32e-1		2.61e-2		2.03e-1					
20	1.80e-2	2.87	2.14e-3	3.60	9.81e-4	7.69				
40	2.28e-3	2.98	2.17e-4	3.30	1.22e-4	3.00				
	P_0P_3		P_1P_3		P_2P_3		P_3P_3			
10	7.85e-3		2.10e-2		2.00e-1		1.96e-4			
20	4.56e-4	4.10	1.36e-3	3.94	2.68e-5	12.86	1.23e-5	3.99		
40	2.77e-5	4.04	8.61e-5	3.98	1.21e-6	4.46	7.80e-7	3.98		
	P_0P_4		P_1P_4		P_2P_4		P_3P_4		P_4P_4	
10	3.54e-3		1.28e-3		2.01e-1		1.27e-5		1.25e-1	
20	1.19e-4	4.89	3.57e-5	5.16	2.13e-5	13.20	2.39e-7	5.73	6.28e-2	1.00
40	3.75e-6	4.98	1.07e-6	5.04	6.72e-7	4.98	5.86e-9	5.34	5.60e-9	23.41
80									1.75e-10	4.99

Table 9.9: The L^1 errors and EOC of some P_NP_M schemes applied to the advection equation.

9.2.3 The Study of the Efficiency

We again consider the advection equation $v_t + v_x = 0$ with the initial function $v_0(x) = \sin(x)$ defined on $I = [0, 2\pi]$ and its solution at time $T = 2\pi$. The CFL numbers λ were taken from the Table 9.8. We study the efficiency of the P_NP_M schemes by setting the bound for the L^1 errors at time $T = 2\pi$ to be 0.01. We measure the speed of the schemes by the computational time

and number of time steps Z_1. Since we consider the linear advection equation then the time step Δt is constant and then it is equal to $\Delta t = T/Z_1 = 2\pi/Z_1$. Also the time step is computed using the Courant number λ by $\Delta t = \lambda h/a$. With our assumptions here we have $a = 1$ and $h = 2\pi/Z$ then $\Delta t = \lambda 2\pi/Z$. Thus we have $2\pi/Z_1 = \lambda 2\pi/Z$ which implies that $Z_1 = Z/\lambda$. A further indicator of and the cost of the discretization is the mesh size Z. To explain how do we perform these computations we take as example the P_0P_1 scheme using the stencil $S_{I_j,2,1}$ and take the mesh size Z changing from $Z = 2$ to $Z = 35$. We ended the computation when the L^1 error became lower than 0.01. For brevity, we give only some of these results for $Z = 28, \ldots, 35$. Table 9.10 shows that, when $Z = 33$, it is the first case where the L^1 error is less than 0.01. In this case we need 33 iterations and a computational time of 0.049 seconds.

L^1	0.0132681	0.0123711	0.0115621	0.0108299	0.0101650	0.0095595
time	0.0376	0.0364	0.0399	0.0460	0.0458	0.0490
Z_1	28	29	30	31	32	33
Z	28	29	30	31	32	33

Table 9.10: The computational time and the mesh size for the P_0P_1 scheme.

Now we will only give the data for the solution that satisfies the error bound on the coarsest mesh, which we obtain from a sequence of finer and finer meshes as explained. The errors will be rounded to 4 decimals.

9.2.3.1 The Influence of the Size n_e

We recall that the reconstruction stencil is given by $S_{I_j,n_e,L} = \bigcup_{c=-L}^{R} I_{j+c}$ and it consists of the interval I_j with L and R elements to the left and right of I_j, respectively, and its size is given by $n_e = 1 + L + R$ with $L \in \{0, \ldots, n_e - 1\}$ and $R \geqslant 0$. We used various stencils with different sizes n_e and fixed the index L at the values $L = 0$ and $L = n_e - 1$, see Table 9.11.

For $M = 1$, we have two schemes, the P_0P_1 scheme with various stencils and the P_1P_1 scheme with the unique stencil $S_{I_j,1,0} = I_j$. In all cases $N = M$ we have $n_e = 1$ since there is no reconstruction needed. Table 9.11 shows that the P_0P_1 scheme is faster than the P_1P_1 scheme. This is expected, since the piecewise constant solution P_0P_1 scheme has only one unknown degree of freedom. But the P_1P_1 scheme has a higher accuracy on the same mesh. Also, we find that the computational time grows when the size of stencil becomes larger, again this is expected, since the information comes from more cells. Thus the size of the stencil has negative influence on the efficiency of the scheme, as expected. An important point is that the larger stencils need more grid points to achieve the same accuracy.

For $M = 2$, we have the P_0P_2, P_1P_2, and P_2P_2 schemes. Table 9.11 shows that P_0P_2 scheme is faster than the others. Comparing tables we see that whereas in Table 9.9 on the same spatial mesh the error decreases from P_0P_2 to P_1P_2 to P_2P_2 schemes, on the other hand, in terms of the actual efficiency using the smallest possible stencil in Table 9.11 the order in terms of computational time is reversed. This is despite the fact that the other schemes need fewer mesh points to achieve the same accuracy. But they need more time steps due to their stability

$L=0$				
n_e	L^1	time	Z_1	Z
P_0P_1				
2	0.00955	0.00898	33	33
3	0.00904	0.01328	45	45
4	0.00961	0.01471	52	52
5	0.00903	0.01872	61	61
6	0.00974	0.01890	65	65
P_1P_1				
1	0.00879	0.01887	116	29
P_0P_2				
3	0.00626	0.00586	19	19
4	0.00991	0.00657	21	21
5	0.00802	0.00768	27	27
6	0.00982	0.00915	29	29
P_1P_2				
2	0.00851	0.01446	56	14
3	0.00537	0.01459	76	19
4	0.00721	0.01534	76	19
5	0.00895	0.01483	76	19
6	0.00914	0.01647	80	20
P_2P_2				
1	0.00203	0.01236	75	12

$L=n_e-1$				
n_e	L^1	time	Z_1	Z
P_0P_1				
2	0.00955	0.01030	33	33
3	0.00904	0.01288	45	45
4	0.00961	0.01511	52	52
5	0.00903	0.01741	61	61
6	0.00974	0.01905	65	65
P_1P_1				
1	0.00879	0.01991	116	29
P_0P_2				
3	0.00626	0.00630	19	19
4	0.00991	0.00638	21	21
5	0.00802	0.00826	27	27
6	0.00982	0.00919	29	29
P_1P_2				
2	unstable			
3				
4				
5				
6	0.00897	0.01508	76	19
P_2P_2				
1	0.00203	0.01313	75	12

Table 9.11: Numerical computations for some P_NP_M schemes with $M=1$ and $M=2$ for two values of L, $L=0$ and $L=n_e-1$.

restrictions. Note also that there is no real difference between choosing the larger stencils in an upwind $L=n_e-1$ or a downwind $L=0$ manner.

In Table 9.12 we now compare the case $L=0$ with the smallest stencil for different mesh sizes. The computational time of the P_0P_2 scheme is the smallest using the different meshes comparing with the P_1P_2 and P_2P_2 schemes. Again we see that the stability is crucial to the comparison since severer stability limits lead to a larger number of time steps.

9.2.3.2 The Influence of the Shifting L

We now use stencils of the same size n_e but with different type for the values $L=0,\ldots,n_e-1$. We choose $M=3$ and $n_e=5$.

We note in Table 9.13 that the symmetric stencil with $L=2$ is the best choice according to the computational time and the spatial discretization. On the other hand, the one side stencils with $L=0$ and $L=4$ require slightly longer computational time. The difference in choice of stencil is not very pronounced. Moreover, for the finite volume scheme P_0P_3, the number

$L=0$									
	P_0P_2 with $n_e=3$			P_1P_2 with $n_e=2$			P_2P_2		
Z	L^1	time	Z_1	L^1	time	Z_1	L^1	time	Z_1
10	0.04172	0.01471	10	0.02313	0.01881	40	0.08998	0.02117	63
11	0.99828	0.00517	12	0.25449	0.00995	45	0.04117	0.01186	69
12	0.02444	0.00462	12	0.23295	0.01008	49	0.00203	0.01282	75
13	0.84816	0.00503	14	0.01062	0.01122	52	0.10307	0.01438	82
14	0.01550	0.00806	14	0.00851	0.01421	56	0.06386	0.01830	88
15	0.73693	0.00530	16	0.18595	0.01169	61	0.02988	0.01558	94

Table 9.12: Numerical computations for some P_NP_M schemes with $M=2$ using the smallest possible stencil.

L	L^1	time	Z_1	Z
P_0P_3				
0	0.00676	0.00714	16	16
1	0.00784	0.00540	12	12
2	0.00940	0.00421	8	8
3	0.00784	0.00485	12	12
4	0.00676	0.00632	16	16
P_1P_3				
0	0.00772	0.01277	56	14
1	0.00661	0.01146	40	10
2	0.00931	0.01017	40	10
3	unstable			
4				

L	L^1	time	Z_1	Z
P_2P_3				
0	0.00491	0.01139	50	8
1	0.00405	0.01156	50	8
2	0.00097	0.01165	50	8
3	0.00350	0.01178	50	8
4	0.00440	0.01220	50	8
P_3P_3				
0	0.00009	0.02086	125	10

Table 9.13: The computational time and the mesh size of some P_NP_M schemes for $M=3$ with different types of the stencils of the size $n_e=5$.

of iterations relates to the type of the stencil, whereas with $N>0$ this number seems to be constant. This is seen also in Table 9.11.

Furthermore, when $N=0$, the number of iterations Z_1 is equal to the mesh size Z, whereas for $N>0$, this number is larger than Z by a factor due to the stability restriction. This indeed means that with larger N the cost of the computations is larger, but this improves the accuracy. This agrees with the results in the Table 9.9 where we find for example that for $Z=40$ the P_3P_3 scheme is more accurate than the P_2P_3 scheme which is in turn more accurate than the P_1P_3 and P_0P_3 schemes.

In Table 9.14 we again compare the computational time for some P_NP_M schemes with $M=3$ using the smallest stencil for different mesh sizes. The computational time of the P_0P_3 scheme is the smallest using the different meshes comparing with the P_1P_2 and P_2P_2 schemes.

Again we see that the stability is crucial to the comparison since severer stability limits lead to a larger number of time steps.

	P_0P_3 with $n_e = 4$ and $L = 1$		P_0P_3 with $n_e = 4$ and $L = 2$		P_1P_3 with $n_e = 2$ and $L = 0$		P_2P_3 with $n_e = 2$ and $L = 0$		P_3P_3	
Z	L^1	time	L^1	time	L^1	time	L^1	time	L^1	time
10	0.0070	0.0037	0.0070	0.0036	0.0040	0.0098	0.0884	0.0125	0.00009	0.0197
11	0.9979	0.0044	0.9994	0.0043	0.2501	0.0096	0.0400	0.0168	0.04045	0.0263
12	0.0034	0.0045	0.0034	0.0042	0.2299	0.0104	0.0003	0.0143	0.00004	0.0235
13	0.8480	0.0052	0.8486	0.0048	0.0014	0.0116	0.1025	0.0148	0.03424	0.0263
14	0.0018	0.0050	0.0018	0.0047	0.0010	0.0111	0.0634	0.0157	0.00002	0.0290
15	0.7368	0.0066	0.7371	0.0060	0.1847	0.0133	0.0295	0.0186	0.02968	0.0323

Table 9.14: Numerical computations for some P_NP_M schemes with $M = 3$ using the smallest possible stencil.

Chapter 10

The Burgers Equation

Here we apply the P_NP_M schemes to the initial value problem of the Burgers equation $v_t(t,x)+(v^2/2)_x(t,x)=0$ for $(t,x)\in\mathbb{R}_{\geqslant 0}\times[a,b]$ with an initial function $v_0(x)=v(0,x)$.

10.1 The P_NP_M Schemes

In this chapter we will use two numerical fluxes. The Lax-Friedrichs flux is given by

$$\mathcal{F}^n_{LF,j+\frac{1}{2}}(t)=\frac{1}{2}\left[\frac{\left(U^n_{j+1}(t,x_{j+\frac{1}{2}})\right)^2}{2}+\frac{\left(U^n_{j}(t,x_{j+\frac{1}{2}})\right)^2}{2}-C\left(U^n_{j+1}(t,x_{j+\frac{1}{2}})-U^n_{j}(t,x_{j+\frac{1}{2}})\right)\right]$$

$$\text{where}\quad C=\max_{\min_I(v_0)\leqslant s\leqslant\max_I(v_0)}|f'(s)|=\max_{\min_I(v_0)\leqslant s\leqslant\max_I(v_0)}|s|=\max_{x\in I}|v_0(x)|.$$

The Godunov flux is given by

$$\mathcal{F}^n_{G,j+\frac{1}{2}}(t)=\begin{cases}\max\{f(U^n_{j+1}(t,x_{j+\frac{1}{2}})),f(U^n_{j}(t,x_{j+\frac{1}{2}}))\} & \text{if } U^n_{j+1}(t,x_{j+\frac{1}{2}})<U^n_{j+1}(t,x_{j+\frac{1}{2}}),\\ \min\{f(U^n_{j+1}(t,x_{j+\frac{1}{2}})),f(U^n_{j}(t,x_{j+\frac{1}{2}}))\} & \text{if } U^n_{j+1}(t,x_{j+\frac{1}{2}})\geqslant U^n_{j+1}(t,x_{j+\frac{1}{2}}).\end{cases}$$

Then the scheme (8.4) becomes in the form

$$\begin{aligned}\widehat{u}^{n+1}_{k,j} &= \widehat{u}^n_{k,j}-\frac{2k+1}{h}\int_{\breve{T}_n}\mathcal{F}^n_{j+\frac{1}{2}}(t)dt+(-1)^k\frac{2k+1}{h}\int_{\breve{T}_n}\mathcal{F}^n_{j-\frac{1}{2}}(t)dt\\ &\quad+\frac{2k+1}{h}\sum_{i=1}^{\mathcal{N}}f\left(\widehat{U}^n_{i,j}\right)\langle(\Phi_{k,j})_x,\theta_{i,j}\rangle_{tx}.\end{aligned}$$

For example, we consider the P_0P_0 scheme. In this case we have $\widehat{U}^n_{1,j}=\widehat{w}^n_{0,j}=\widehat{u}^n_{0,j}$, and $f\left(\widehat{U}^n_{1,j}\right)=\frac{1}{2}(\widehat{u}^n_{0,j})^2$. We have $\Phi_{0,j}(x)=1$, $\theta_{1,j}(t,x)=1$, and $\langle(\Phi_{0,j})_x,\theta_{1,j}\rangle_{tx}=0$. The Lax-Friedrichs flux becomes

$$\mathcal{F}^n_{LF,j+\frac{1}{2}}(t)=\frac{\left(\widehat{u}^n_{0,j+1}\right)^2+\left(\widehat{u}^n_{0,j}\right)^2}{4}-\frac{C}{2}\left(\widehat{u}^n_{0,j+1}-\widehat{u}^n_{0,j}\right),$$

$$\mathcal{F}^n_{LF,j-\frac{1}{2}}(t) = \frac{\left(\widehat{u}^n_{0,j}\right)^2 + \left(\widehat{u}^n_{0,j-1}\right)^2}{4} - \frac{C}{2}\left(\widehat{u}^n_{0,j} - \widehat{u}^n_{0,j-1}\right).$$

Thus we get

$$\begin{aligned}
\widehat{u}^{n+1}_{0,j} &= \widehat{u}^n_{0,j} - \frac{1}{2h}\int_{\breve{T}_n}\left\{\frac{\left(\widehat{u}^n_{0,j+1}\right)^2 - \left(\widehat{u}^n_{0,j-1}\right)^2}{2} - C(\widehat{u}^n_{0,j-1} - 2\widehat{u}^n_{0,j} + \widehat{u}^n_{0,j+1})\right\}dt \\
&= \widehat{u}^n_{0,j} - \frac{k_n}{2h}\left\{\frac{\left(\widehat{u}^n_{0,j+1}\right)^2 - \left(\widehat{u}^n_{0,j-1}\right)^2}{2} - C(\widehat{u}^n_{0,j-1} - 2\widehat{u}^n_{0,j} + \widehat{u}^n_{0,j+1})\right\}.
\end{aligned}$$

10.2 The Slope Limiter

The $P_N P_M$ schemes for the Burgers equation generate oscillations near the discontinuities which in turn render the schemes unstable in the high order cases. So we use a slope limiter. We will use the limiter defined in Cockburn [5] for the RKDG. The author proved that the RKDG scheme obtained is total variation diminishing in the means (TVDM) under a CFL condition. At first we need the following notation.

The minmod function $\mu : \mathbb{R}^r \to \mathbb{R}$ with $r > 0$, is given, for $\eta_1, \ldots, \eta_r \in \mathbb{R}$ by

$$\mu(\eta_1, \ldots, \eta_r) = \begin{cases} \nu \min_{1\leqslant i\leqslant r} |\eta_i|, & \text{if } \nu = sign(\eta_1) = \ldots = sign(\eta_r) \\ 0, & \text{otherwise.} \end{cases}$$

We assume that the term u^n_j of the solution at time t_n is $u^n_j = \sum_{i=0}^N \widehat{u}^n_{i,j}\Phi_{i,j}$. We denote by $\tilde{u}^n_j$ to the slope limiter of u^n_j and calculate the two limits

$$\begin{aligned}
\left(\tilde{u}^n_{j+\frac{1}{2}}\right)^- &:= \widehat{u}^n_{0,j} + \mu\left(\left(u^n_{j+\frac{1}{2}}\right)^- - \widehat{u}^n_{0,j}, \widehat{u}^n_{0,j} - \widehat{u}^n_{0,j-1}, \widehat{u}^n_{0,j+1} - \widehat{u}^n_{0,j}\right), \\
\left(\tilde{u}^n_{j-\frac{1}{2}}\right)^+ &:= \widehat{u}^n_{0,j} - \mu\left(\widehat{u}^n_{0,j} - \left(u^n_{j-\frac{1}{2}}\right)^+, \widehat{u}^n_{0,j} - \widehat{u}^n_{0,j-1}, \widehat{u}^n_{0,j+1} - \widehat{u}^n_{0,j}\right).
\end{aligned}$$

We can finally define the slope limiter as follows.

1. If the conditions $\left(\tilde{u}^n_{j+\frac{1}{2}}\right)^- = \left(u^n_{j+\frac{1}{2}}\right)^-$ and $\left(\tilde{u}^n_{j-\frac{1}{2}}\right)^+ = \left(u^n_{j-\frac{1}{2}}\right)^+$ hold then we set $\tilde{u}^n_j = u^n_j$.
2. If not, then we set $\tilde{u}^n_j = \widehat{u}^n_{0,j} + \mu\left(\widehat{u}^n_{1,j}, \widehat{u}^n_{0,j} - \widehat{u}^n_{0,j-1}, \widehat{u}^n_{0,j+1} - \widehat{u}^n_{0,j}\right)\Phi_{1,j}$.

This slope limiter enforces the TVDM property, but it loses the high order accuracy. Therefore, we modify the slope limiter displayed above in such a way we preserve the high order accuracy even at local extrema. The resulting scheme will then be total variation bounded in the means (TVBM). We follow Shu [20] and modify the definition by using the TVB corrected minmod function $\widehat{\mu}$ defined as follows. We define

$$\widehat{\mu}(\eta_1, \ldots, \eta_r) = \begin{cases} \eta_1, & \text{if } |\eta_1| \leqslant \widehat{C}h^2, \\ \mu(\eta_1, \ldots, \eta_r), & \text{otherwise,} \end{cases}$$

where $\widehat{C}$ is an upper bound of $|(v_0)_{xx}|$ at the local extrema. In other words, if the initial function $v_0(x) \in C^2(I)$, then we can take

$$\widehat{C} = \sup\{|(v_0)_{xx}(x)|, \quad \text{for all} \quad x \in [a,b] \text{ such that } (v_0)_x(x) = 0\}.$$

10.3 The Riemann Problems

We first rewrite the Burgers equation in the form $v_t + (v^2/2)_x = v_t + vv_x = 0$. The method of characteristics enables us to determine some exact solutions. If we evaluate x as a function $x = x(t)$ of t, then the characteristics of the Burgers equation are curves in the x-t plane and given by $x'(t) = f'(v) = v(t, x(t))$. The function v is a constant along characteristics, since

$$\frac{d}{dt}v(t,x(t)) = v_t(t,x(t)) + v_x(t,x(t))x'(t) = v_t + vv_x = 0.$$

We suppose e.g. $v = v_0 = v(0, x(0))$, then the characteristics have constant slopes. This means that they are straight lines of the form $x(t) = v_0 t + x(0)$.

The conservation law together with piecewise constant data having a single discontinuity is known as the Riemann problem, see LeVeque [18]. An example of these initial data is given by

$$v_0(x) = \begin{cases} v_L, & \text{for} \quad x < x_*, \\ v_R, & \text{for} \quad x \geqslant x_*, \end{cases} \qquad \text{where} \quad x_* \in \mathbb{R}.$$

Since the initial solution is constant, except at $x = x_*$, the constant which is associated to the TVBM slope limiter is $\widehat{C} = 0$. So when we would use the slope limiter it will hold only the TVDM property and the high order accuracy will be lost. The Riemann problem can be solved depending on the relation between v_L and v_R.

When $v_L > v_R$. The problem has a unique weak solution which is the following shock wave

$$v(t,x) = \begin{cases} v_L, & \text{for} \quad x < x_* + \varphi t \\ v_R, & \text{for} \quad x \geqslant x_* + \varphi t. \end{cases}$$

where φ is the shock speed which is given by the following Rankine-Hugoniot jump condition

$$\varphi = \frac{f(v_L) - f(v_R)}{v_L - v_R} = \frac{v_L + v_R}{2}. \tag{10.1}$$

Since $v_L > \varphi > v_R$, which is known as the entropy condition, then the characteristics in each of the regions should go into the shock as time advances, see Figure 10.1.

When $v_L < v_R$. The solution is a rarefaction wave, see LeVeque [18],

$$v(t,x) = \begin{cases} v_L, & \text{for} \quad x < x_* + v_L t \\ (x(t) - x_*)/t, & \text{for} \quad x_* + v_L t \leqslant x < x_* + v_R t \\ v_R, & \text{for} \quad x_* + v_R t \leqslant x. \end{cases}$$

The rarefaction solution is shown in Figure 10.2.

In the following we apply some $P_N P_M$ schemes and study the influence of the slope limiter and the numerical fluxes on the solution.

10.3.1 Initial Data I

We first use the initial data

$$v_0(x) = \begin{cases} 2, & \text{for} \quad -1 \leqslant x < 0, \\ 0, & \text{for} \quad 0 \leqslant x \leqslant 1. \end{cases}$$

The shock speed is $\varphi = 1$. For $x < t$ the characteristics are the lines $x(t) = 2t + x(0)$, whilst for $x \geqslant t$ the characteristics are given by the vertical lines $x = x(0)$, until they cross, see Figure 10.3, then they go into the shock. So the solution is given by

$$v(t, x) = \begin{cases} 2, & \text{for} \quad -1 \leqslant x < t, \\ 0, & \text{for} \quad t \leqslant x \leqslant 1. \end{cases}$$

Now we apply the P_0P_0 scheme using the Lax-Friedrichs flux with final time $T = 0.1$. The shock appears at $x = T = 0.1$. Since $M = 0$ we do not need to apply the slope limiter. Table 10.1 shows that the solution is of a half order using the L^2 norm while it is of the first order with the L^1 norm, this agrees with the Theorem 2.6 about the discontinuous solutions. Figure 10.4 views this solution with $T = 0.1$ and $Z = 128$.

We apply the P_0P_1 scheme with the stencil $S_{I_j,3,1}$ and $T = 0.1$ for three cases, the first without using the slope limiter, the second and third with using a TVDM slope limiter but with Lax-Friedrichs and Godunov fluxes, respectively. The oscillations appear clearly in Figure 10.5. The use of the slope limiter, as shown in Figures 10.6 and 10.7, removes these oscillations at the jump, and by using the Godunov flux, the solution is slightly better.

We view in Table 10.2 the L^1 and L^2 errors of the P_0P_1 scheme using the TVDM slope limiter. We find that the Godunov flux gives better solutions.

10.3.2 Initial Data II

The initial data are

$$v(0, x) = \begin{cases} 2, & \text{for} \quad -1 \leqslant x \leqslant 0, \\ 1, & \text{for} \quad 0 < x \leqslant 2, \\ 0, & \text{for} \quad 2 < x \leqslant 4. \end{cases}$$

The solution consists of two moving shocks φ_1 and φ_2, the first shock at $x = 0$ moves with speed $\varphi_1 = 1.5$ and the second at $x = 2$ moves with speed $\varphi_2 = 0.5$. After a small time $t < 2$, the characteristics and shock lines are given by

$$\begin{array}{rcl} \varphi_1 & : & x(t) = \frac{3}{2}t, \\ \varphi_2 & : & x(t) = \frac{1}{2}t + 2, \end{array} \qquad x(t) = \begin{cases} 2t + x(0) & \text{for} \quad x \leqslant \frac{3}{2}t, \\ t + x(0) & \text{for} \quad \frac{3}{2}t < x \leqslant \frac{1}{2}t + 2, \\ x(0) & \text{for} \quad \frac{1}{2}t + 2 < x, \end{cases}$$

until the shocks would cross. To illustrate this meeting, the left part of the characteristics cross with the slope φ_1 when their times are equal, i.e.

$$t_{cha} = t_{sho} \quad \rightarrow \quad \frac{x - x(0)}{2} = \frac{2}{3}x \quad \rightarrow \quad x = -3x(0).$$

For example, the characteristic through $x(0) = -0.5$ is the line $x(t) = 2t - 0.5$. It goes into the shock at $x = -3(-0.5) = 1.5$ and so $1.5 = 2t - 0.5$ or $t = 1$. This is the same point if we substitute $t = 1$ in the equation of the φ_1. Hence for a small time ($t < 2$) the solution is given by

$$v(t,x) = \begin{cases} 2, & \text{for} \quad -1 \leqslant x \leqslant \frac{3}{2}t, \\ 1, & \text{for} \quad \frac{3}{2}t < x \leqslant \frac{1}{2}t + 2, \\ 0, & \text{for} \quad \frac{1}{2}t + 2 < x \leqslant 4. \end{cases}$$

With the time, the middle region, defined onto $1.5t < x \leqslant 0.5t + 2$, becomes smaller and smaller, and at $(x = 3, t = 2)$ this domain diminishes and the two shocks merge into one shock φ_3, see Figure 10.8. The new shock will connect between two values $v_L = 2$ to the left and $v_R = 0$ to the right. Using the Rankine-Hugoniot condition (10.1) the shock speed equals to $\varphi_3 = 1$ and the equation of the shock line is

$$1 = \frac{x(t) - x(2)}{t - 2} = \frac{x(t) - 3}{t - 2} \quad \rightarrow \quad x(t) = t + 1.$$

Therefore, we have

$$v(t,x) = \begin{cases} 2, & \text{for} \quad -1 \leqslant x \leqslant t + 1, \\ 0, & \text{for} \quad t + 1 < x \leqslant 4, \end{cases} \qquad \text{for} \quad t \geqslant 2.$$

Figures 10.9, 10.10, 10.11, and 10.12 views the P_1P_2 solution with the stencil $S_{I_j,3,1}$ using the TVDM limiter with $Z = 128$ at times $T = 1, 1.8, 2, 2.5$. Note how the two shocks merge when $T = 2$ and with $T > 2$ the soltion is shifting to right. The order of accuracy of the solutions are again a half with the L^2 norm and one with the L^1 norm.

Z	L^1	EOC	L^2	EOC
32	0.1259		0.3059	
64	0.0696	0.86	0.2140	0.52
128	0.0372	0.90	0.1604	0.42

Table 10.1: The L^1 and L^2 errors of the P_0P_0 scheme for the initial data I.

	Lax-Friedrichs				Godunov			
Z	L^1	EOC	L^2	EOC	L^1	EOC	L^2	EOC
32	0.1259		0.3097		0.0613		0.2552	
64	0.0696	0.86	0.2182	0.51	0.0239	1.36	0.1511	0.76
128	0.0372	0.90	0.1622	0.43	0.0154	0.64	0.1276	0.24

Table 10.2: The errors of the P_0P_1 scheme for the initial data I.

10.4 The Burgers Equation with Smooth Initial Data

We choose the initial data to be the periodic function $v_0(x) = \sin(x)$ for $x \in [0, 2\pi]$.

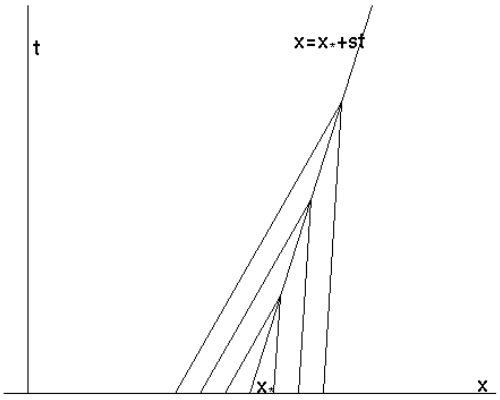


Figure 10.1: The shock solution corresponds to the case $v_L > v_R$.

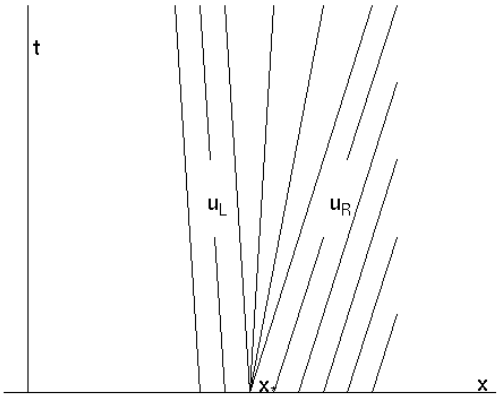


Figure 10.2: The rarefaction solution corresponds to the case $v_L < v_R$.

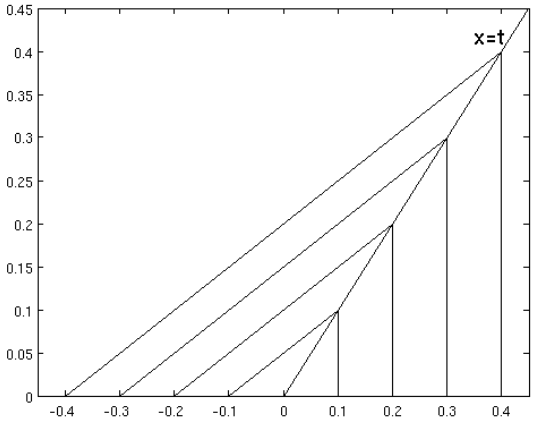


Figure 10.3: The characteristics of the initial data I.

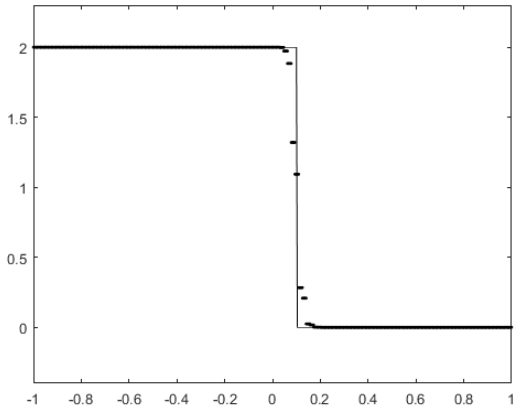


Figure 10.4: The P_0P_0 solution for data I.

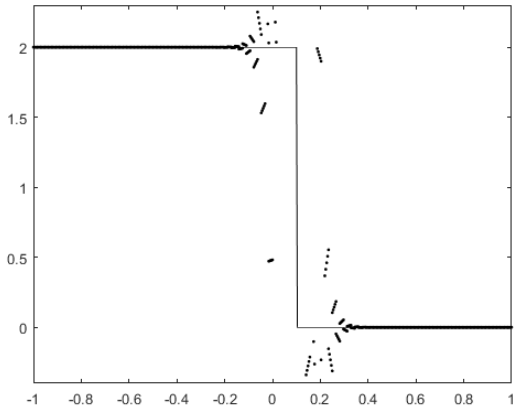


Figure 10.5: The P_0P_1 solution without using the slope limiter.

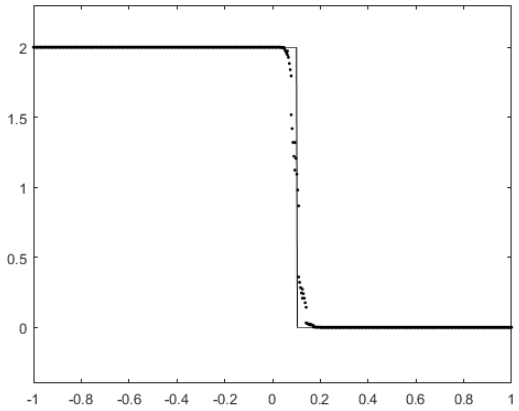


Figure 10.6: The P_0P_1 solution with $T = 0.1$ using slope limiter with the Lax-Friedrichs flux.

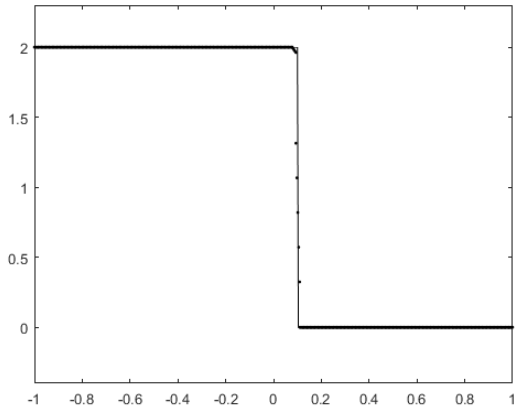


Figure 10.7: The P_0P_1 scheme with $T = 0.1$ using slope limiter with the Godunov flux.

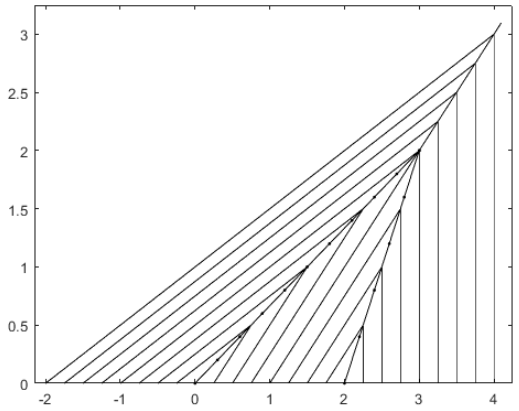


Figure 10.8: The characteristics and the shock solution corresponds to data II.

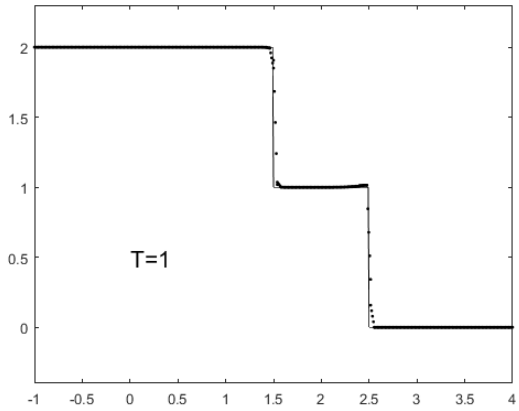


Figure 10.9: The P_1P_2 scheme for data II at $T = 1$.

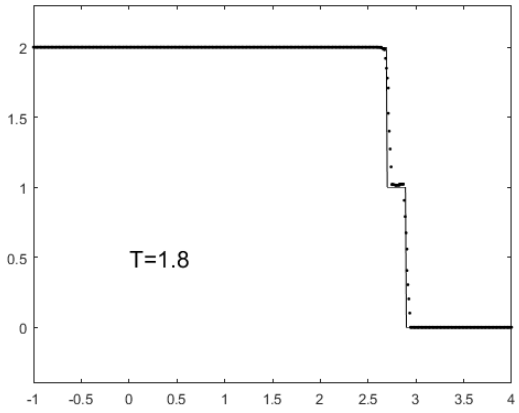


Figure 10.10: The P_1P_2 scheme for data II at $T = 1.8$.

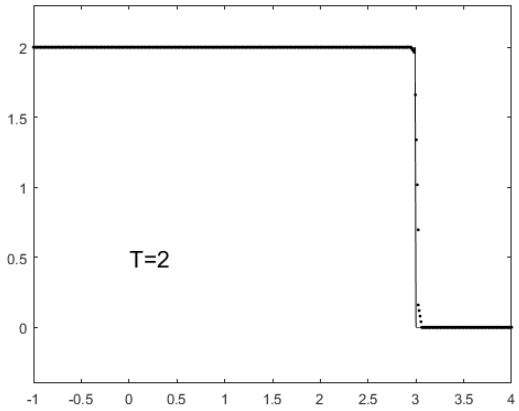


Figure 10.11: The P_1P_2 scheme for data II at $T = 2$.

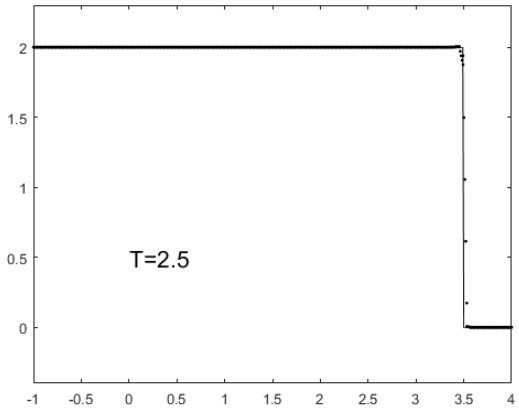


Figure 10.12: The P_1P_2 scheme for data II at T.

10.4.1 The Exact Solution

The exact solution after time $T > 0$, see LeVeque [18], is given implicitly by the function $v(T, x) = \sin(x - uT)$ for $x \in [0, 2\pi]$. The method of characteristics determines the exact solution also the characteristics are straight lines of the form $x(t) = v_0 t + x(0)$ with constant slopes $v_0 = v(0, x(0))$.

At the time $t = 0$ the initial function has the value zero at $x = \pi$. The characteristic at $x = \pi$ is vertical, it is so at $x = 0$ and $x = \pi$, and the solution on this line is constant and equals to zero, so a small shock appears immediately at the point of $x = \pi$. At $t = \pi/2$ the characteristics of the maximal value 1 of v_0 at $\pi/2$ and of the minimal value -1 of v_0 at $3\pi/2$ reach $x = \pi$. For later times the modulus of these values decreases, see Figure 10.13.

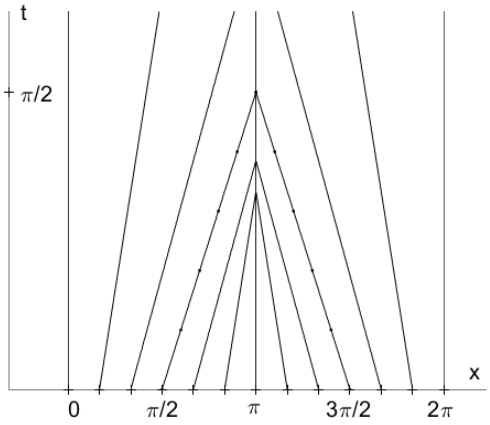


Figure 10.13: The exact solution of the Burgers equation for the initial data $v_0(x) = \sin(x)$ using the characteristic method.

10.4.2 The Numerical Solutions

We will compute the errors on the whole domain $x \in [0, 2\pi]$ and on the domain away from the shock which appears at $x = \pi$, i.e. the shock excluding domain will be chosen to be

$$x \in [0, \pi - 0.15[\ \cup \ ]\pi + 0.15, 2\pi].$$

The sine function has two local extrema in $[0, 2\pi]$, namely $y = \pi/2$ and $y = 3\pi/2$, and $|(v_0)_{xx}(y)| = 1$. Then by choosing $\widehat{C} = 1$ the slope limiter will verify the TVBM property.

In order to find the best solution of this example applying the $P_N P_M$ schemes, we consider all of the following points:

- The slope limiter will be not applied at first, then it will be applied with the TVDM property ($\widehat{C} = 0$) and with the TVBM property ($\widehat{C} = 1$).
- The Lax-Friedrichs and the Godunov fluxes will be considered.

- The L^1 errors will be computed on the whole and the shock excluding domains.

On the other hand, we fix the final time to be $T = 1$ and choose only the stencil $S_{I_j,3,1}$ which consists of three elements with one left neighbor of I_j. This stencil fits with all schemes we study in this section.

Tables 10.3 and 10.4 view the errors and orders of the P_0P_1 and P_1P_1 Schemes, respectively. We note, from these Tables that

- (Scheme) It clearly appears that the P_1P_1 is better than the P_0P_1.
- (Domain) The solution always is of second order when the errors are computed away from the shocks. While, we lose the orders, if we compute the errors onto the whole domain.
- (Limiter) The TVDM limiter increases the error with the coarse mesh, then with fine mesh works better. While the TVBM limiter always gives good results. Moreover, the slope limiters only remove the oscillations and do not improve the order.
- (Flux) The Lax-Friedrichs flux gives better results than these which the Godunov flux gives. This appears clearly, when we move from the P_0P_1 to P_1P_1, where we see that the errors using the Lax-Friedrichs flux reduce to about 50% or less, while using the Godunov flux, the errors reduce with a weaker rate.

From the remarks above we resume the progress onto the higher order schemes only by using the Lax-Friedrichs flux and using the TVBM-limiter.

Z	shock excluding domain				whole domain			
	Lax-Friedrichs		Godunov		Lax-Friedrichs		Godunov	
Without-limiter								
80	6.84e-3		3.92e-3		45.73e-3		16.44e-3	
160	1.02e-3	2.75	0.87e-3	2.17	19.22e-3	1.25	6.06e-3	1.44
320	0.22e-3	2.23	0.22e-3	1.98	8.40e-3	1.19	2.32e-3	1.39
With TVDM-limiter								
80	6.78e-3		4.09e-3		44.92e-3		11.87e-3	
160	1.04e-3	2.71	0.89e-3	2.19	19.24e-3	1.22	4.31e-3	1.46
320	0.22e-3	2.25	0.22e-3	2.01	8.40e-3	1.20	1.65e-3	1.38
With TVBM-limiter								
80	6.67e-3		3.92e-3		44.88e-3		11.72e-3	
160	1.02e-3	2.71	0.87e-3	2.17	19.22e-3	1.22	4.29e-3	1.45
320	0.22e-3	2.23	0.22e-3	1.98	8.40e-3	1.19	1.65e-3	1.38

Table 10.3: The L^1 errors of the P_0P_1 scheme.

Z	shock excluding domain				whole domain			
	Lax-Friedrichs		Godunov		Lax-Friedrichs		Godunov	
Without-limiter								
80	1.32e-3		3.20e-3		23.30e-3		8.87e-3	
160	0.22e-3	2.60	0.83e-3	1.95	9.86e-3	1.24	3.22e-3	1.46
320	0.06e-3	1.93	0.22e-3	1.93	4.17e-3	1.24	1.19e-3	1.44
With TVDM-limiter								
80	4.37e-3		5.55e-3		23.17e-3		11.22e-3	
160	0.91e-3	2.26	1.33e-3	2.06	9.51e-3	1.29	3.73e-3	1.59
320	0.21e-3	2.15	0.32e-3	2.06	3.98e-3	1.26	1.29e-3	1.53
With TVBM-limiter								
80	1.27e-3		2.41e-3		20.37e-3		8.09e-3	
160	0.22e-3	2.55	0.64e-3	1.92	8.88e-3	1.20	3.02e-3	1.42
320	0.06e-3	1.93	0.17e-3	1.90	3.84e-3	1.21	1.14e-3	1.41

Table 10.4: The L^1 errors of the P_1P_1 scheme.

Chapter 11

The Shallow Water Equations

11.1 The Mathematical Model

We are interested to the system of one dimensional shallow water equations with discontinuous bed topography $\Lambda(x)$, see [2], which can be written in the form

$$\mathbf{U}_t + \mathbf{F}(\mathbf{U})_x = \mathbf{S}(\mathbf{U}), \tag{11.1}$$

$$\text{where} \quad \mathbf{U} = \begin{bmatrix} \Lambda \\ \hbar \\ \hbar\tilde{v} \end{bmatrix}, \quad \mathbf{F}(\mathbf{U}) = \begin{bmatrix} 0 \\ \hbar\tilde{v} \\ \hbar\tilde{v}^2 + \frac{g}{2}\hbar^2 \end{bmatrix}, \quad \mathbf{S}(\mathbf{U}) = \begin{bmatrix} 0 \\ 0 \\ -g\hbar\Lambda_x \end{bmatrix},$$

and the variables Λ, $\hbar$, $\tilde{v}$ and g are respectively the bottom topography, the water height, the water velocity and the gravitational constant. We will study the shallow water equations on the flat bottom area, i.e. $\Lambda(x) = 0$, then the system becomes homogeneous

$$\mathbf{U}_t + \mathbf{F}(\mathbf{U})_x = 0. \tag{11.2}$$

The quasi linear form of the system (11.1) is written as $\mathbf{V}_t + \mathbf{A}(\mathbf{V})\mathbf{V}_x = 0$ where $\mathbf{V} = (\Lambda, \hbar, \hbar\tilde{v})^T$, the Jacobian matrix $\mathbf{A}(\mathbf{V})$ is given as

$$\mathbf{A}(\mathbf{V}) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ g\hbar & g\hbar - \tilde{v}^2 & 2\tilde{v} \end{bmatrix}.$$

The eigenvalues are $\lambda_0 = 0$, $\hbar_1 = \tilde{v} - \sigma$, and $\hbar_2 = \tilde{v} + \sigma$, where $\sigma = \sqrt{g\hbar}$ is the sound speed. The system (11.2) is not strictly hyperbolic due to that $\hbar_0$ can be equal to any of two other eigenvalues. The corresponding right eigenvectors are

$$\mathbf{R}_0 = \begin{bmatrix} \frac{\sigma^2 - \tilde{v}^2}{\sigma^2} \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{R}_1 = \begin{bmatrix} 0 \\ 1 \\ \tilde{v} - \sigma \end{bmatrix}, \quad \mathbf{R}_2 = \begin{bmatrix} 0 \\ 1 \\ \tilde{v} + \sigma \end{bmatrix}.$$

11.2 Numerical Test

The Lax-Friedrichs flux used here is given by

$$\mathcal{F}^n_{LF,j+\frac{1}{2}}(t) = \frac{1}{2}\left[f\left(U^n_{j+1}(t,x_{j+\frac{1}{2}})\right) + f\left(U^n_j(t,x_{j+\frac{1}{2}})\right) - C\left(U^n_{j+1}(t,x_{j+\frac{1}{2}}) - U^n_j(t,x_{j+\frac{1}{2}})\right)\right]$$

$$\text{where} \quad C = \max_I\left(|\tilde{v}(x)| + \sqrt{g\hbar(x)}\right).$$

This value of C is the spectral radius of the matrix of the quasi form, see [26]. We consider the following Riemann initial value problem

$$\hbar(x) = \begin{cases} \hbar_L = 4 & \text{for} \quad -3 \leqslant x < 0, \\ \hbar_R = 0.4 & \text{for} \quad 0 \leqslant x \leqslant 3, \end{cases} \qquad \tilde{v}(x) = 0.$$

The computational domain is $x \in [-3, 3]$. We will use the TVDM slope limiter defined in Section 10.2. Figure 11.1 views the P_1P_1 solution with $Z = 300$ at time $T = 0.2$.

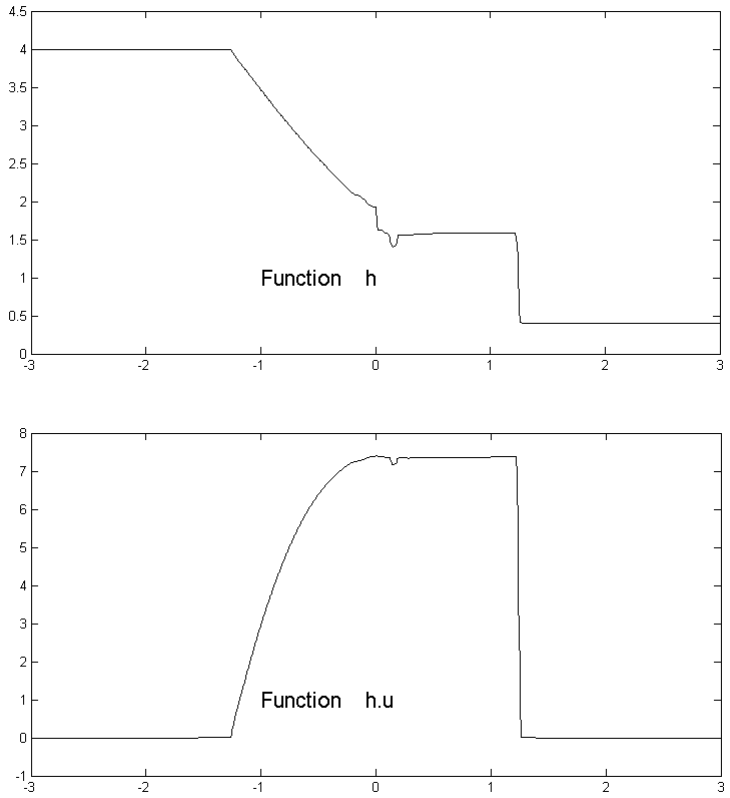


Figure 11.1: The P_1P_1 solution using TVDM limiter with $Z = 300$ at time $T = 0.2$.

Conclusion

We have considered the $P_N P_M$ DG schemes for the 1D problems of the hyperbolic conservation laws with $N \leqslant M$ introduced by Dumbser et al. [7].

We presented some properties of the projection operators. We proved that the piecewise polynomials of degree $N \geqslant 0$ are approximations of order $N+1$ for smooth initial data. The projection operators also recover the same initial data exactly if these initial data are polynomial of the same degree N.

We proved that the reconstruction operators give unique solutions and recover the same initial data exactly if these initial data are polynomial of the same degree M.

In Chapters 3 and 6 the solutions for some examples were approximations of the expected order. Also for those in Chapter 6 we found small differences by changing the stencils.

All allowed combinations $0 \leqslant N \leqslant M \leqslant 5$ had some stencils with stability for all CFL numbers between 0 and a maximal CFL number. We found a wide range of maximal stability limits being CFL numbers between 0.103 and 2. Some stencils have a strange semi-stability behaviour since they are stable for CFL numbers in an interval bounded away from 0. Also some stencils lead to unstable schemes.

Using the stability limits that we obtained, we checked the experimental order of convergence (EOC). We report only the cases $0 \leqslant N \leqslant M \leqslant 4$ for the stencil $S_{I_j,5,2}$. We always obtain an expected EOC close to $M+1$, also in other cases we did not put into the paper.

Based on the stability limits of the various schemes we also studied the efficiency of the schemes. We found that for given M the $P_0 P_M$ schemes are faster than the others with $M \geqslant N > 0$. Also, we found that the computational time grows when the size of stencil becomes larger and there was no real difference between choosing the larger stencils in an upwind $L = n_e - 1$ or a downwind $L = 0$ manner. We noted that the symmetric stencil, i.e. with $n_e > 1$ odd and $L = \frac{n_e}{2} - 1$, achieves a required accuracy on a coarser mesh leading to a faster computation in comparison to the asymmetric stencils of the same size.

Appendix A

2D Hierarchical Orthogonal Basis on Rectangles

Let T_j be a rectangle from the partition Ω_K and $(x, y) \in T_j$ and $(\xi, \eta) = R_j(x, y) \in T_S$.
For $N = 0$, the unique basis function is $\Psi_{0,j}(\xi, \eta) = 1$.
For $N = 1$, the basis has three functions $\Psi_{0,j}(\xi, \eta) = 1$, $\Psi_{1,j}(\xi, \eta) = \xi$, and $\Psi_{2,j}(\xi, \eta) = \eta$.
For $N = 2$, the basis has six functions

$$\Psi_{0,j} = 1, \quad \Psi_{1,j} = \xi, \quad \Psi_{2,j} = \frac{1}{2}(3\xi^2 - 1), \quad \Psi_{3,j} = \eta, \quad \Psi_{4,j} = \xi\eta, \quad \Psi_{5,j} = \frac{1}{2}(3\eta^2 - 1).$$

For $N = 3$, the basis has ten functions

$$\Psi_{0,j} = 1 \qquad \Psi_{1,j} = \xi \qquad \Psi_{2,j} = \tfrac{1}{2}(3\xi^2 - 1) \qquad \Psi_{3,j} = \tfrac{1}{2}(5\xi^3 - 3\xi)$$

$$\Psi_{4,j} = \eta \qquad \Psi_{5,j} = \eta\xi \qquad \Psi_{6,j} = \tfrac{1}{2}(3\xi^2 - 1)\eta \quad \Psi_{7,j} = \tfrac{1}{2}(3\eta^2 - 1)$$

$$\Psi_{8,j} = \tfrac{1}{2}\xi(3\eta^2 - 1) \quad \Psi_{9,j} = \tfrac{1}{2}(5\eta^3 - 3\eta),$$

The mass matrices are $B_0 = h_1 h_2$, $B_1 = h_1 h_2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix}$, and

$$B_2 = h_1 h_2 \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{5} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{9} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{5} \end{pmatrix}, \qquad B_3 = h_1 h_2 \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{9} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{15} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{15} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{7} \end{pmatrix}$$

Bibliography

[1] R. A. Adams, Sobolev Spaces, Academic Press, Inc. 1975.

[2] N. Andrianov, Performance of numerical methods on the non unique solution to the Riemann problem for the shallow water equations, International Journal for Numerical Methods in Fluids 47: 8-9, 825-831.

[3] G. E. L. Andrews, R. Askey, R. Roy, Special functions, Cambridge University Press, 1999.

[4] R. L. Burden, J. D. Faires, Numerical Analysis, ninth edition, 2011, Brooks/Cole, Cengage Learning.

[5] B. Cockburn, An introduction to the discontinuous Galerkin method for convection-dominated Problems, Lecture Notes in Mathematics vol. 1697, 1998, pp 151-268.

[6] B. Cockburn, C.W. Shu, TVB Runge Kutta local projection discontinuous Galerkin finite element method for conservation laws II, general framework, Mathematics of computation, 52, 1989, 411-435.

[7] M. Dumbser, D. Balsara, E.F. Toro, C.D. Munz, A unified framework for the construction of one-step finite volume and discontinuous Galerkin schemes on unstructured meshes, Journal of Computational Physics 227, 2008, 8209-8253.

[8] M. Dumbser, C.D. Munz, Arbitrary high order discontinuous Galerkin schemes, in: S. Cordier, T. Goudon, M. Gutnic, E. Sonnendrucker (Eds.), Numerical Methods for Hyperbolic and Kinetic Problems, IRMA Series in Mathematics and Theoretical Physics, EMS Publishing House, 2005, 295-333.

[9] M. Dumbser, M. Käser, Arbitrary high order non oscillatory finite volume schemes on unstructured meshes for linear hyperbolic systems, Journal of Computational Physics 221, 2007, 693-723.

[10] M. Dumbser, C. Enaux, E.F. Toro, Finite volume schemes of very high order of accuracy for stiff hyperbolic balance laws, Journal of Computational Physics 227, 2008, 3971-4001.

[11] M. Dumbser, M. Käser, V.A. Titarev, E.F. Toro, Quadrature-free non oscillatory finite volume schemes on unstructured meshes for non linear hyperbolic systems, Journal of Computational Physics 226, 2007, 204-243.

[12] L. C. Evans, Partial Differential Equations. Graduate Studies in Mathematics, Vol. 19, AMS, Providence, 1991.

[13] S. R. Ghorpade, B. V. Limaye, A Course in Calculus and Real Analysis, Springer, 2006.

[14] C. Goetz, M. Dumbser, A square entropy stable flux limiter for $P_N P_M$ DG schemes, submitted on 14 Dec 2016 (arXiv:1612.04793).

[15] A. Harten, B. Engquist, S. Osher, S. Chakravarthy, Uniformly high order essentially non oscillatory schemes, III, Journal of Computational Physics, 71, 1987, 231-303.

[16] C. Hirsch, Numerical Computation of Internal and External Flows vol I: Fundamentals of Numerical Discretization, Wiley, 1988.

[17] T. H. Koornwinder, R. Wong, R. Koekoek, R. F. Swarttouw, Orthogonal Polynomials, Chapter 18 in F. W. J. Olver, D. W. Lozier, R. F. Boisvert, C. W. Clark, NIST handbook of mathematical functions, Cambridge University Press, 2010.

[18] R. J. LeVeque, Numerical Methods for Conservation Laws. Birkhäuser, Basel, 3-7643-2723-5, 1992.

[19] L. L. Schumaker, Spline Functions: Basic Theory, third edition, Cambridge University Press, 2007.

[20] C.W. Shu. TVB uniformly high order schemes for conservation laws. Math. Comp., 49:105-121, 1987.

[21] I. A. Stegun, Legendre Functions, Chapter 8 in M. Abramowitz, I. A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9 ed, 1970.

[22] G. Strang, Introduction to Linear Algebra 3ed, Wellesley Cambridge Press, 2003.

[23] B. Szabo, I. Babuška, Finite Element Analysis, Wiley, New York, 1991.

[24] E.F. Toro, R.C. Millington, L.A.M. Nejad, Towards very high order Godunov schemes, in: E.F. Toro (Ed.), Godunov Methods. Theory and Applications, Kluwer/Plenum Academic Publishers, 2001, 905-938.

[25] D. S. Watkins, A generalization of the Bramble-Hilbert Lemma and applications to multivariate interpolation, Journal of approximation theory 26, 219-231, 1979.

[26] Y. Xing, X Zhang, C. W. Shu, Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow-water equations, Adv Water Resour (2010), doi:10.1016/j.advwatres.2010.08.005.

Ehrenerklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; verwendete fremde und eigene Quellen sind als solche kenntlich gemacht.

Ich habe insbesondere nicht wissentlich:

- Ergebnisse erfunden oder widersprüchliche Ergebnisse verschwiegen,
- statistische Verfahren absichtlich missbraucht, um Daten in ungerechtfertigter Weise zu interpretieren,
- fremde Ergebnisse oder Veröffentlichungen plagiiert oder verzerrt wiedergegeben.

Mir ist bekannt, dass Verstösse gegen das Urheberrecht Unterlassungs- und Schadenersatzansprüche des Urhebers sowie eine strafrechtliche Ahndung durch die Strafverfolgungsbehörden begründen kann.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

Magdeburg, August 9, 2018

Abdulatif Badenjki