UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

VAN NIEKERK A

THE DEVELOPMENT OF RATIONAL BASIS FUNCTIONS FOR THE FINITE ELEMENT METHOD

PhD UP 1989

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The development of rational basis functions for the finite element method

by

André van Niekerk

Submitted in partial fulfilment of the requirements for the degree

Ph.D.

in the

Faculty of Science
University of Pretoria
PRETORIA

November 1989

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

ACKNOWLEDGEMENTS

I wish to express my sincere thanks to Prof. F.D. van Niekerk for his support and encouragement. I am very grateful to him for his active interest in this work and for his personal interest in me. I am also indebted to Mrs A.E. van Rensburg who typed the manuscript so carefully and willingly.

Finally, I would like to thank my father, mother and Karen for their moral support and understanding during the research and writing of the manuscript.

I dedicate this thesis to my parents and my wife, Karen.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

INDEX

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

CHAPTER 1

MOTIVATION AND PURPOSE

1.1 MOTIVATION

The numerical solution of partial differential equations plays an important role in engineering and scientific applications [2,7,14,27]. Owing to improvements in numerical techniques and computer technology, the solution of partial differential equations which were previously difficult to solve is now possible. The finite element method, which is one of the general techniques for constructing approximate solutions to partial differential equations, was also used to achieve this. The method divides the domain of solution into a finite number of subdomains (the finite elements) and uses variational concepts to construct an approximation over the collection of finite elements [2,14,28,34]. It is because the method is of such a general nature that it has been applied with success to a wide range of problems in all areas of engineering and mathematical physics [2,33].

The basic idea of the finite element method can be explained briefly by means of the Galerkin method. Assuming that a solution to the operator equation, $L(u) = f$, has to be found, then a suitable finite dimensional space M (the trial space) and a basis $(w_1,\ldots,w_N)$ of M have to be selected. The Galerkin method consists in finding an approximation to u of the form

$$U = \sum_{j=1}^{N} \alpha_j w_j ,$$

where the unknown coefficients α_j, $j=1,\ldots,N$ have to be determined so that

$$(L(U) - f,V) = 0$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

for all $V \in M$, where $(\cdot,\cdot)$ denotes the L_2 inner product [14]. The quality of the approximation depends entirely on the choice of the space M and the construction of appropriate basis functions [2,14]. A basis for M can be selected in such a manner that each basis function is a piecewise polynomial with support on a small region of the spatial domain.

A serious shortcoming of this method, however, is that the approximate solution will depend to a large extent on the properties of the basis functions. Up till now, mainly piecewise polynomials have been considered as basis functions, and a number of polynomial basis functions were constructed to solve linear and nonlinear partial differential equations.

Although the class of finite element methods predominates, it has had less influence on nonlinear parabolic, hyperbolic and other time-dependent nonlinear problems such as convection-dominated phenomena. Usually, applications of Galerkin finite element methods and finite difference methods to these problems result in spurious oscillations [13,27]. The oscillations seem to occur mainly where convection dominates diffusion [5,7,27]. Although the oscillations can be removed by a stringent mesh refinement, this undermines the practical utility of the methods. This shortcoming prompted the development of numerical schemes which do not exhibit spurious oscillations. This in turn led to the classical upwind difference schemes [13,27,36,45] and Petrov-Galerkin methods [5,7,26,29] that use trial functions biased in the upstream direction, i.e. the direction from which the convection emanates. Although these methods usually suppress numerical oscil-

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

lations, they are at times inaccurate [17,27]. In addition, parabolic and hyperbolic flows have solutions that develop steep gradients and discontinuities, which are the mathematical equivalent of shock waves. Usually, it is only possible to model this kind of behaviour satisfactorily by using a very small mesh in the relevant numerical scheme. It would therefore be advantageous to develop a basis function that has the ability to approximate steep gradients and is also biased in the upstream direction. By constructing such a basis function an important property of fluid flow would be better modelled.

Since polynomials are the basic building blocks of most approximating functions, the approximation abilities of polynomials and rational functions are here compared. According to the theorem of Weierstrass [34] a polynomial p exists for every $f \in C[X]$ and every $\epsilon > 0$, such that

$$|f(x) - p(x)| < \epsilon \quad \text{for all} \quad x \in X,$$

where X is an interval of the real line. To determine the degree of approximation of f by polynomials define the nth degree of approximation $E_n(f)$ by

$$E_n(f) = \inf_{p_n} \{\sup_{x \in X} |f(x) - p_n(x)|\} ,$$

where the infimum is taken for all polynomials p_n of degree n on X. The question that then arises is whether one can estimate how fast $E_n(f)$ approaches zero. Powell [32] and Lorentz [24] found that the least maximum error of best approximation converges like $\frac{1}{n}$, i.e. $E_n(|x|) = 0(\frac{1}{n})$.

A rational function, $r_{m,n}$, on an interval X is defined as

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$r_{m,n}(x) = \frac{p_m(x)}{q_n(x)}, \quad x \in X, \tag{1.1}$$

where $p_m(x)$ and $q_n(x)$ are polynomials of degree m and n respectively [4,24,32]. According to a corollary of the theorem of Weierstrass, each function $f \in C[X]$ is approximable by $r_{n,n}$. Furthermore, it is a known fact that a best rational approximation exists [4,8,32]. Let $R_n(f)$ now denote the nth degree of approximation of f by rational functions, i.e. $r_{n,n}(x)$, then the infimum should be taken for all $r_{n,n}(x)$ that are finite on X; in other words $q_n(x) \neq 0$ for $x \in X$. Since the set of polynomials is contained in the set of rational functions, it is clear that $R_n(f) \leq E_n(f)$. To establish how much smaller $R_n(f)$ can be than $E_n(f)$, the degree of approximation of $f(x) = |x|$ on $X = [-1,1]$ by rational functions can be considered. Newman [24] has already demonstrated the existence of an upper bound, $R_n(|x|) \leq 3e^{-\sqrt{n}}$, $n \geq 5$. But, an optimal error bound for $|x|$ by rational functions has not yet been established, whereas the error bound by polynomials is optimal. By this example it can be seen that the error of approximation by rational functions is therefore of a different order of magnitude than that of polynomial approximation.

The approximation ability of a rational function is even better illustrated if its effect on the exponential function $\{e^x, -1 \leq x \leq 1\}$ is taken into consideration. Since $r_{m,n}(x)$ in equation (1.1) remains unchanged if $p_m(x)$ and $q_n(x)$ are replaced by $cp_m(x)$ and $cq_n(x)$, where c is any non-zero constant, the parameters provide $(m + n + 1)$ degrees of freedom. For this reason it is appropriate to compare the rational approximant $r_{m,n}$ with a polynomial approximant $p_{m+n}(x)$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

[24]. Then, according to [24], the least maximum error of a rational approximation to e^x, when $m = n = 2$, would be a mere 0.000087, while that of a polynomial approximation p_4 would be 0.000547. The gain in accuracy is remarkable and shows the ability of the rational function to approximate functions with a steep gradient or a suddenly changing solution without diverging rapidly to an unbounded value [24]. Rational approximations are therefore preferred to polynomials, and are used in computer subroutines almost exclusively for the calculation of transcendental functions such as sines, exponentials, etc. [32].

Recently, rational functions have been developed to construct numerical algorithms to compute the numerical solutions of linear and non-linear ordinary differential equations [11]. The power and efficiency of such algorithms were demonstrated by Luke et al [25] and Van Niekerk [49] in the numerical solution of ordinary differential equations.

It was the approximation abilities of rational functions and the success achieved by rational functions in the numerical solution of ordinary differential equations that prompted the introduction of rational functions as basis functions for the finite element method.

1.2 PURPOSE OF THIS STUDY

As stated in the previous paragraph, the approximation ability of rational functions has led to the introduction of rational functions, instead of polynomials, as basis functions for the finite element method. To avoid any poles in the rational function, the denominator

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

should be chosen in such a manner that it does not include any zeros in the range of the variable x, $x \in X$. Furthermore, owing to the flexibility of the rational function, an attempt should be made to construct a rational basis function that is biased in the upstream direction. By doing so, it is hoped to achieve, via the finite element method, a more efficient rational numerical scheme with improved approximation abilities than that of a numerical scheme utilising polynomials on the same mesh. Moreover, the effect of convection in parabolic and hyperbolic problems would then be physically modelled.

The main objectives of this thesis are to develop rational basis functions for the finite element method and to evaluate these rational schemes by applying it to time-dependent convection-diffusion, nonlinear parabolic and hyperbolic equations. Properties of the rational schemes, for example convergence, consistency and stability, have been restricted to the linearised versions of these equations.

1.3 OUTLINE OF THESIS

In Chapter 2 the concept of a rational basis function is established and the properties of such a function developed for use in later chapters. To approximate higher-order partial derivatives, Hermite rational basis functions are also constructed. The efficiency of the rational approximation method is then illustrated on a stiff ordinary differential equation.

In Chapter 3, a convection-diffusion equation

$$u_t + \delta u_x = \epsilon u_{xx}, \quad \epsilon > 0, \quad \delta > 0$$

is examined, where ϵ is the diffusion coefficient and δ the con-

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

vection coefficient. Oscillatory solutions appear if the cell Peclet number $\frac{\delta h}{\epsilon} > 2$ [7]. The rational basis functions that are constructed, however, are biased in the upstream direction as shown in Figure 1.1 and thus satisfy the convection property of the flow.

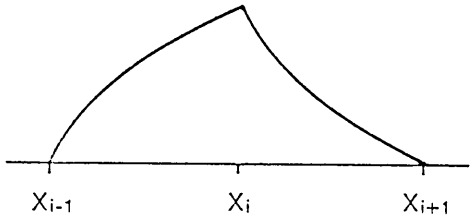

FIGURE 1.1 : Rational basis function

These rational basis functions are used both as trial and test functions in the Galerkin method and produce a numerical scheme where both test and trial functions are equally upwinded. This differs from the standard Petrov-Galerkin method where only the test function was biased in the upstream direction and where one had to change the magnitude of upwinding by means of a free parameter. Next the numerical scheme is employed with different boundary conditions and compared with the exact solutions available. The properties of the numerical scheme are also examined for consistency, stability and convergence. Then the rational basis functions are compared with the linear shaped basis functions with and without upwinding. From this comparison it follows that the rational basis function is found to give rise to a numerical scheme in which eigenvalues have reduced imaginary components, thereby increasing dissipation and reducing oscillations of the solution.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

In Chapter 4 the nonlinear Burgers equation

$$u_t + uu_x = \epsilon u_{xx}. \quad \epsilon > 0,$$

where ϵ is the coefficient of kinematic viscosity, is examined. For small values of ϵ the solution develops steep fronts, and numerical methods are unable to describe the sudden change in gradient, resulting in numerical oscillations [9]. It is the approximation abilities of a rational basis function that prompted this approach when it was found that the change in gradient had been satisfactorily approximated. The method was extended to include higher-order rational basis functions whereby the accuracy of the method was also improved. This extension is unique in the sense that higher-order rational basis functions did not involve the coupling of more node points in a mesh, but still only couple three nodes, leaving the tridiagonal structure of the matrices intact.

Next, the nonlinear Korteweg-de Vries equation

$$u_t + uu_x + \epsilon u_{xxx} = 0, \quad \epsilon > 0,$$

where ϵ is the coefficient of dispersion, is examined. The presence of the third-order partial derivative calls for rational basis functions that have higher-order continuous derivatives. This leads to the construction of the Hermite rational basis functions which are used as test functions in conjunction with the linear hat functions as trial functions in a Petrov-Galerkin method. In this way, a consistent numerical scheme is achieved that couples five nodes and is able to model a solitary wave without any significant phase errors.

In Chapter 5, the numerical solution of hyperbolic equations of the form

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$u_t + f(u)_x = 0$$

are examined. A rational numerical scheme is proposed that is accurate in smooth regions of the solution and produces a sharp resolution at a point of discontinuity or shock. The convergence of the numerical scheme is then proved.

In Chapter 6, a conclusion is drawn based on all the results found in the previous chapters, and finally, the numerical scheme is compared with other existing schemes.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

CHAPTER 2

RATIONAL APPROXIMATION

2.1 INTRODUCTION

The numerical approximation of a differential equation can be done by means of a weak formulation of the equation. In the variational statement, however, the test and trial functions have to be selected in order to obtain a numerical approximation to the exact solution. The Petrov-Galerkin method is obtained by choosing different test and trial functions; the Galerkin method by choosing identical test and trial functions. These functions are defined in terms of piecewise continuous functions, namely the finite element basis functions in the finite element method. The quality of approximation in the finite element method depends entirely on the choice of a test and trial function.

A rational function has better approximation abilities than ordinary polynomials and this property is used to construct C^0 continuous rational basis functions. Thus, by means of the finite element method a rational approximation method is derived.

The above-mentioned topics are discussed in this chapter and the effectiveness of the rational approximation method is demonstrated on a stiff ordinary differential equation. Further, C^1 continuous Hermite rational basis functions are also constructed.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

2.2 DEFINITIONS

For the sake of convenience, the notation and definitions of the terminology are introduced:

2.2.1 $C^0, C^1, \ldots, C^n$ spaces

The space that has k th-order continuous derivatives on $[a,b]$, $k \leq n$, is denoted by $C^k[a,b]$.

2.2.2 L_p norm

Define L_p norm as

$$\|\phi\|_p = \left(\int_a^b |\phi(x)|^p \, dx \right)^{1/p}$$

for $x \in [a,b]$.

2.2.3 Basis functions

A set $\{\phi_1, \phi_2, \ldots, \phi_n\}$ is a set of basis functions for the space X if

(i) all the members of the set are linearly independent, and

(ii) each element in X can be represented by a linear combination of ϕ_k, $1 \leq k \leq n$, i.e. $X = \text{span}\{\phi_1, \phi_2, \ldots, \phi_n\}$.

2.2.4 Support

The set on which the function ϕ is nonzero is called the support of the function.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

2.2.5 Inner product

Let X be a real linear space. Define for each $\phi_1, \phi_2 \in X$ a real number (ϕ_1, ϕ_2), such that

(i) $(\phi_1, \phi_1) \geq 0$; the equality holds if and only if $\phi_1 = 0$.

(ii) $(\phi_1, \phi_2) = (\phi_2, \phi_1)$.

(iii) $(a\phi_1, \phi_2) = a(\phi_1, \phi_2)$, a is any scalar.

(iv) $(\phi_1 + \phi_2, \phi_3) = (\phi_1, \phi_3) + (\phi_2, \phi_3)$.

If $(\phi_1, \phi_2) = 0$, then ϕ_1 and ϕ_2 are orthogonal to each other. The inner product

$$(\phi_1, \phi_2) = \int_X \phi_1 \phi_2 \, dx$$

is used in this thesis. In an inner product space the L_2 norm is defined by $\|\phi\|_2 = \{(\phi, \phi)\}^{1/2}$.

2.2.6 Sobolev space

The Sobolev space $W_p^m[a,b]$ of order (m,p) [30] is the linear space of functions in $L_p[a,b]$ whose distributional derivatives $D^\alpha u$ of all orders $|\alpha|$, $0 \leq |\alpha| \leq m$, are in $L_p[a,b]$, i.e.

$$W_p^m[a,b] = \{\phi | D^\alpha \phi \in L_p[a,b] \text{ for } 0 \leq |\alpha| \leq m\}$$

where $$L_p[a,b] = \{w | \int_a^b w^p \, dx < \infty\}.$$

Only $p = 2$ and $m = 1$ will be considered and a norm of the Sobolev space $W_2^1[a,b]$ is defined as

$$\|\phi\|_{W_2^1[a,b]} = \{\int_a^b [\phi^2 + (\phi')^2] dx\}^{1/2}.$$

Henceforth, the set of elements of $W_2^1[a,b]$ that are zero at end points a and b is indicated by H_0^1.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

2.2.7 Rational function

Let $R_{ST}(x)$ denote the rational functions of order (S,T), which is a function of the form

$$R_{S,T}(x) = \frac{a_0 x^S + a_1 x^{S-1} + \ldots + a_S}{b_0 x^T + b_1 x^{T-1} + \ldots + b_T}, \quad \sum_{i=0}^{T} |b_i| \neq 0.$$

2.2.8 Bounded variation

A function f defined on $[a,b]$ is of bounded variation [23] on $[a,b]$ if its variation $V_a^b f$ on $[a,b]$ is finite, where

$$V_a^b f = \sup \sum_{j=1}^{n} |f(t_j) - f(t_{j-1})|,$$

and the supremum is being taken over all arbitrary partitions,

$$a = t_0 < t_1 < \ldots < t_n = b, \quad \text{where } n \in \mathbb{N} \text{ is arbitrary.}$$

2.2.9 Existence of best approximation

Let X be a normed linear space with norm $\|\cdot\|$. Let X_N be a finite dimensional subspace of X. Then each point $x \in X$ has a best approximation $x_N \in X_N$; that is

$$\|x - x_N\| = \min_{y \in X_N} \|x - y\|.$$

This result is proven by Prenter [34] and assures the existence of a best approximation. The problem now remains to compute the approximation numerically. For this reason the variational method is discussed in the next section.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

2.3 VARIATIONAL STATEMENT OF A PROBLEM

Consider the problem of finding a function $u(x)$, $0 \leq x \leq 1$, which satisfies

$$\begin{aligned} Lu &= -u_{xx} + u = f(x) = x \ , \ 0 < x < 1 \\ u(0) &= 0, \ u(1) = 0. \end{aligned} \tag{2.1}$$

The data of the problem consists of all the information given in advance: the domain of the solution $(0 \leq x \leq 1)$, the nonhomogenous part given by the function $f(x) = x$, the coefficients of various derivatives of u and the values the solution attains at the boundary. Since the data in (2.1) is smooth, there exists a unique solution, $u(x) = x - \sinh x/\sinh 1$. In most applications, however, there is no solution to the classical statement of the problem because some of the data is not smooth.

To overcome this difficulty, the boundary-value problem is reformulated to admit weaker conditions on the solution and its derivatives, by means of a variational or weak formulation of the problem. The variational statement is given as follows: Find $u \in \tilde{H}$ such that

$$\int_0^1 (-u_{xx} + u - x)v \, dx = 0$$

or

$$\begin{aligned} &(-u_{xx} + u - x, v) = 0 \quad \text{for all} \quad v \in H \\ &u(0) = 0 \\ &u(1) = 0 \end{aligned} \tag{2.2}$$

where H is the set of all weight or test functions v that have zero values at $x = 0$ and $x = 1$.

It should be appreciated that the classical solution of (2.1), if it exists, is the only solution of (2.2) and that the specification of

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

the set H of test functions is an important aspect of an acceptable weak formulation. The set $\tilde{H}$ to which the solution u belongs is called the class of trial functions.

If u and v are sufficiently smooth, then

$$\int_0^1 -u_{xx}v\,dx = \int_0^1 u_x v_x\,dx - u_x v\Big|_0^1$$
$$= \int_0^1 u_x v_x\,dx\ .$$

Since the same order of derivatives of both trial and test functions appears one may choose $H = \tilde{H}$. Therefore, (2.2) can be replaced by the following variational problem: Find $u \in H_0^1$ such that

$$(u_x,v_x) + (u,v) - (x,v) = 0 \quad \text{for all} \quad v \in H_0^1 \qquad (2.3)$$

where H_0^1 is the set of all functions that vanish at the end points and whose first derivatives are square-integrable. Thus, $w \in H_0^1$ if

$$\int_0^1 (w_x)^2\,dx < \infty \quad \text{and} \quad w(0) = w(1) = 0. \qquad (2.4)$$

Since (2.3) has only first derivatives the smoothness requirements are weakened, thereby enlarging the class of data for which (2.3) makes sense.

2.4 GALERKIN METHOD

The Galerkin method [2] is explained by using the model problem in Section 2.3. Consider the problem in the following variational form: Find $u \in H_0^1$ such that

$$(u_x,v_x) + (u,v) = (x,v) \quad \text{for all} \quad v \in H_0^1. \qquad (2.5)$$

The set H_0^1 is a linear space of functions and secondly, H_0^1 is

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

infinite-dimensional. Suppose that an infinite set of functions $\{\psi_1(x), \psi_2(x), \ldots\}$ in H_0^1 has the property that each test function $v \in H_0^1$ can be represented as

$$v(x) = \sum_{i=1}^{\infty} \beta_i \psi_i(x) \tag{2.6}$$

where the β_i are constants and the series converges in a sense appropriate for the space H_0^1. It is clear that a finite number of terms in (2.6), say N, will yield an approximation v_N of v, that is

$$v_N(x) = \sum_{i=1}^{N} \beta_i \psi_i(x). \tag{2.7}$$

An approximation v_N converges to a function v in the H_0^1-norm [2,14] if

$$\lim_{N\to\infty} \int_0^1 [(v - v_N)^2 + (v_x - (v_N)_x)^2]dx = 0. \tag{2.8}$$

The N basis functions $\{\psi_1, \psi_2, \ldots, \psi_N\}$ define an N-dimensional subspace $H_0^{(N)}$ of H_0^1.

Galerkin's method consists of seeking an approximate solution to (2.5) in a finite-dimensional subspace $H_0^{(N)}$ of the space H_0^1. Thus, the variational statement of the approximate problem is: Find $u_N \in H_0^{(N)}$ where

$$u_N = \sum_{i=1}^{N} \alpha_i \psi_i(x) \tag{2.9}$$

such that

$$(u_N', v_N') + (u_N, v_N) = (x, v_N) \tag{2.10}$$

for all $v_N \in H_0^{(N)}$, where $u_N' = \frac{du_N}{dx}$.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

If the N coefficients α_i in (2.9) are determined, then the approximate solution u_N will be determined because ψ_i are known. To determine α_i substitute (2.7) and (2.9) into (2.10) to obtain

$$\int_0^1 \{\frac{d}{dx}[\sum_{i=1}^{N} \beta_i\psi_i(x)]\frac{d}{dx}[\sum_{j=1}^{N} \alpha_j\psi_j(x)] + \sum_{i=1}^{N} \beta_i\psi_i(x) \sum_{j=1}^{N} \alpha_j\psi_j(x) - x\sum_{i=1}^{N} \beta_i\psi_i(x)\}dx = 0 \qquad (2.11)$$

for all β_i, $i=1,2,\ldots,N$.

Because β_i are arbitrary, (2.11) represents N equations to be satisfied by the α_j, $j=1,\ldots,N$.

Consider the following natural choices,

$$\beta_1 = 1,\ \beta_i = 0,\ i \neq 1$$

for the parameters in (2.11). This yields

$$\sum_{j=1}^{N} K_{1j}\alpha_j = F_1,$$

where

$$K_{1j} = (\psi_1',\psi_j') + (\psi_1,\psi_j),$$
$$F_1 = (x,\psi_1)$$

and

$$\psi_i' = \frac{d\psi_i}{dx}.$$

Next, choose $\beta_2 = 1$, $\beta_i = 0$, $i \neq 2$, so that (2.11) yields

$$\sum_{j=1}^{N} K_{2j}\alpha_j = F_2.$$

Continuing in this way, a system of N equations in N unknown coefficients,

$$\sum_{j=1}^{N} K_{ij}\alpha_j = F_i,\ i=1,2,\ldots,N \qquad (2.12)$$

where

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

$$K_{ij} = (\psi_i', \psi_j') + (\psi_i, \psi_j)$$

and $$F_i = (x, \psi_i),$$

is obtained.

Since the trial functions and test functions coincide, the formulation has led to a symmetrical stiffness matrix $K = [K_{ij}]$, reducing the computational effort in obtaining an approximate solution. In addition the stiffness matrix is positive definite, for if $\beta \in \mathbb{R}^N$ is a nonzero vector and $w = \sum_{i=1}^{N} \beta_i \psi_i$, then

$$\begin{aligned} \beta^T K \beta &= \sum_{i,j=1}^{N} \{(\psi_i', \psi_j')\beta_i\beta_j + (\psi_i, \psi_j)\beta_i\beta_j\} \\ &= (w', w') + (w, w) \\ &\geq \|w'\|_2^2 . \end{aligned}$$

Since $\{\psi_1, \ldots, \psi_N\}$ is linearly independent and $\beta \neq 0$, it follows that $\|w\|_\infty = \max_{0 \leq x \leq 1} |w(x)| \neq 0$. The Sobolov inequality, [14], gives that

$$\|w\|_\infty \leq \frac{1}{2}\|w'\|_2$$

so that

$$\beta^T K \beta > 0, \quad \beta \neq 0.$$

Hence, K is positive definite [6] and the solution α_i in (2.12) is unique. Therefore, using (2.9) a unique Galerkin approximation u_N exists.

Further, the exact solution $u(x)$ satisfies

$$(u', v_N') + (u, v_N) = (x, v_N) \tag{2.13}$$

for all $v_N \in H_0^{(N)}$.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Subtract (2.10) from (2.13) and let $e(x) = u(x) - u_N(x)$, which indicates the error in the Galerkin approximation; then one obtains the following orthogonality condition:

$$(e',v_N') + (e,v_N) = 0 \quad \text{for all} \quad v_N \in H_0^{(N)}. \tag{2.14}$$

Consider, for any arbitrary test function, $v_N \in H_0^{(N)}$,

$$\begin{aligned} \|u - v_N\|_{H_0^1}^2 &= (u' - v_N', u' - v_N') + (u - v_N, u - v_N) \\ &= (u' - u_N' + u_N' - v_N', u' - u_N' + u_N' - v_N') \\ &\quad + (u - u_N + u_N - v_N, u - u_N + u_N - v_N) \end{aligned} \tag{2.15}$$

Using the orthogonality condition (2.14) the right-hand side of (2.15) reduces to

$$\begin{aligned} &(u' - u_N', u' - u_N') + (u - u_N, u - u_N) \\ &+ (u_N' - v_N', u_N' - v_N') + (u_N - v_N, u_N - v_N). \end{aligned} \tag{2.16}$$

The left-hand side of (2.15) is the error between the exact solution u and an arbitrary element $v_N \in H_0^{(N)}$. Since each term in (2.16) is positive the error is minimised by choosing $v_N = u_N$. This shows that the Galerkin method provides the best possible approximation of u in $H_0^{(N)}$.

However, it is important to realize that the quality of the Galerkin approximation is completely determined by the choice of basis functions ψ_i.

2.5 FINITE ELEMENT BASIS FUNCTIONS

The basis function ψ_i is defined piecewise over subregions of the domain called finite elements. Partition the domain (i.e., the interval $0 \le x \le 1$) of the model problem into a finite number of elements, say Ω_i, i=1,2,...,N. See Figure 2.1.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

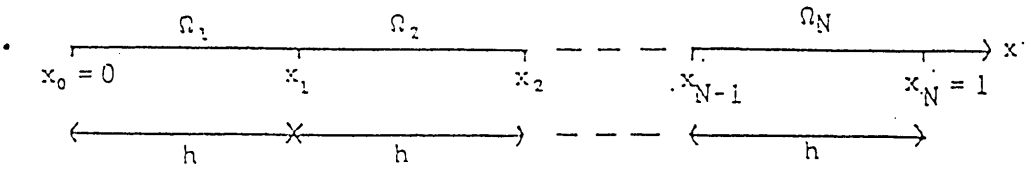

FIGURE 2.1: Elements Ω_i

The set of basis functions is now constructed along the following criteria:

1. The basis functions are generated by simple functions defined piecewise over each element Ω_i.
2. The basis functions are smooth enough to be members of H_0^1.
3. The basis functions are defined in such a way that the parameters α_i are precisely the values of $u_N(x)$ at the nodal points x_i.

Consider the following basis functions,

$$\psi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h} & , \quad x_{i-1} \leq x \leq x_i \\ \dfrac{x_{i+1} - x}{h} & , \quad x_i \leq x \leq x_{i+1} \\ 0 & , \quad \text{elsewhere,} \end{cases}$$

with

$$\psi_i'(x) = \begin{cases} \dfrac{1}{h} & , \quad x_{i-1} \leq x \leq x_i \\ -\dfrac{1}{h} & , \quad x_i \leq x \leq x_{i+1} \\ 0 & , \quad \text{elsewhere} \end{cases}$$

for $i=1,2,\ldots,N$, where $h = x_i - x_{i-1}$ is the length of element Ω_i. It is clear that the hat-shaped function ψ_i associated with node i is obtained by combining linear functions defined on elements Ω_i and Ω_{i+1} respectively as shown in Figure 2.2.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

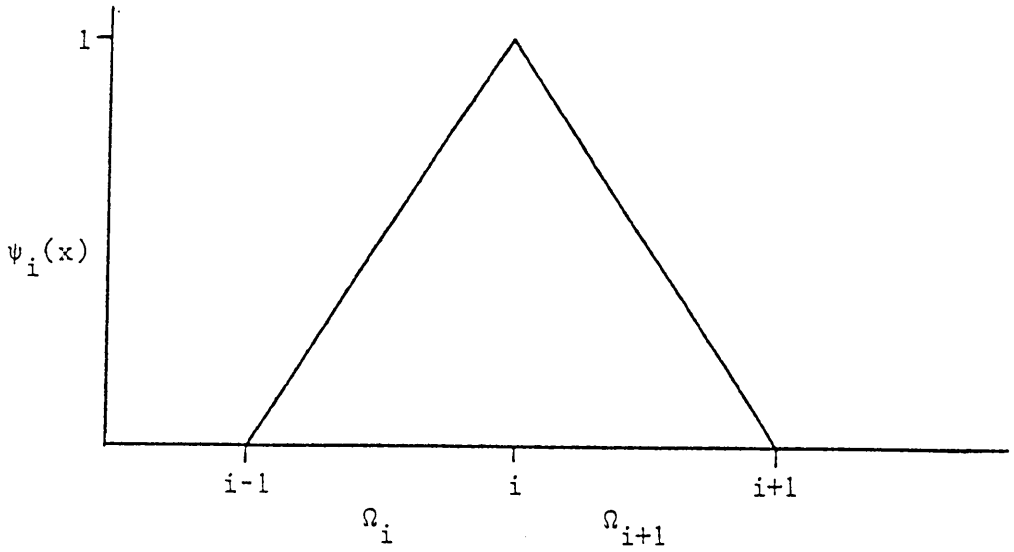

FIGURE 2.2 : Basis function ψ_i

The construction of the basis function is such that its value is unity at one node and zero at all other nodes. More specifically, if x_j is the x coordinate of node j, then

$$\psi_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

This implies that

$$U_N(x_j) = \sum_{i=1}^{N-1} \alpha_i \psi_i(x_j)$$
$$= \alpha_j \quad , \quad \text{for} \quad j=1,\ldots,N-1.$$

Note that $i = 0,N$ are not included because the basis functions satisfy the homogeneous conditions. Since

$$\int_0^1 [\psi_i'(x)]^2 dx = \frac{2}{h} < \infty$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

the basis functions have square-integrable first derivatives, i.e. $\psi_i \in H_0^1$.

Thus, it is clear that the linear hat functions satisfy the criteria 1, 2 and 3. Since the basis function ψ_i couples three nodes x_{i-1}, x_i and x_{i+1}, this results in a stiffness matrix of bandwidth three. This stiffness matrix is a symmetrical positive-definite matrix and possesses a unique inverse.

2.6 ERROR BOUNDS FOR THE GALERKIN METHOD

In Section 2.4 it was shown that the error in the Galerkin approximation satisfies

$$(e',v_N') + (e,v_N) = 0 \quad \text{for all} \quad v_N \in H_0^{(N)}, \tag{2.17}$$

where $e(x) = u(x) - u_N(x)$. Let $\tilde{u}(x)$ be the interpolation function that coincides with the exact solution $u(x)$ pointwise. Then, since $\tilde{u} - u_N \in H_0^{(N)}$ and using the Cauchy-Schwarz inequality [23], the error becomes

$$\begin{aligned} \|e\|_{H_0^1}^2 &= (e',u' - u_N') + (e,u-u_N) \\ &= (e',u'-\tilde{u}') + (e,u-\tilde{u}) \\ &\leq (e',e')^{1/2}(u'-\tilde{u}',u'-\tilde{u}')^{1/2} + (e,e)^{1/2}(u-\tilde{u},u-\tilde{u})^{1/2}. \end{aligned}$$

By definition of the Sobolev norm $\|e\|_{H_0^1}$ [14],

$$(e',e')^{1/2} \leq \|e\|_{H_0^1}$$

and

$$(e,e)^{1/2} \leq \|e\|_{H_0^1}$$

so that above equation becomes

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$\|e\|^2_{H^1_0} \leq \|e\|_{H^1_0} [(u'-\tilde{u}',u'-\tilde{u}')^{1/2} + (u-\tilde{u},u-\tilde{u})^{1/2}]. \qquad (2.18)$$

Since the inequality,

$$a_1^{1/2} + a_2^{1/2} \leq \sqrt{2}(a_1 + a_2)^{1/2}$$

is true for $a_1, a_2 \geq 0$, (2.18) finally becomes

$$\|e\|_{H^1_0} \leq C\|u - \tilde{u}\|_{H^1_0} \qquad (2.19)$$

where C is a constant. Thus, the error norm of the Galerkin solution is always smaller than the error of an interpolation function multiplied by a constant. For a linear hat function [13,14],

$$\|e\|_{H^1_0} \leq Ch.$$

As $h \to 0$, the Sobolev norm approaches zero and the Galerkin solution u_N converges to the exact solution u.

In general the approximation error satisfies [3,14,34]

$$\|u-u_N\|_{H^1_0} \leq Ch^{\mu}$$

where C is a constant independent of h and $\mu = \min(k,s)$, with s the order of the derivatives of u_N that are square-integrable on $[a,b]$ and k the degree of the basis functions on a uniform mesh h.

2.7 CONSTRUCTION OF RATIONAL BASIS FUNCTIONS

It is evident that sudden changes in a function and its derivatives could only be satisfactorily approximated by linear hat-shaped functions if a sufficiently small mesh interval h is used. This entails that the order of the stiffness matrix K would increase, thus introducing computational difficulties. A better approach would be to introduce basis functions that are able to cope with sudden changes in

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

a function and its derivatives, without substantially decreasing the mesh interval h.

Rational functions have been known to possess this ability to approximate a steep function [4,8,32]. The criteria developed in Section 2.4 will be used to construct a rational basis function.

For this purpose, consider a rational function

$$R(x) = a + \frac{b}{1 + cx}, \quad x \in [0,h] \tag{2.20}$$

where a, b and c are constants. Let $c = \frac{1}{h}$ in order to normalise x and define the functions $\phi_1(x)$ and $\phi_0(x)$ on $[0,h]$ by

$$\phi_1(x) = a_1 + \frac{b_1}{1 + \frac{x}{h}}$$

and

$$\phi_0(x) = a_0 + \frac{b_0}{1 + \frac{x}{h}}.$$

By enforcing the interpolation constraints, $\phi_1(0) = 0$ and $\phi_1(h) = 1$, two equations are obtained to solve a_1 and b_1, namely

$$a_1 + b_1 = 0$$

and

$$a_1 + \frac{1}{2} b_1 = 1$$

giving $a_1 = -b_1 = 2$. Thus,

$$\phi_1(x) = \frac{2x}{h + x}, \quad 0 \leq x \leq h.$$

Similarly, from the constraints, $\phi_0(0) = 1$ and $\phi_0(h) = 0$, one obtains $b_0 = 2$ and $a_0 = -1$ so that

$$\phi_0(x) = \frac{h - x}{h + x}, \quad 0 \leq x \leq h.$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The graphs of the functions are shown in Figure 2.3.

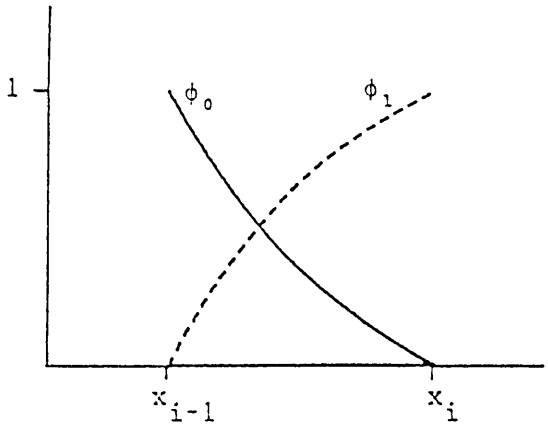

FIGURE 2.3 : The functions ϕ_0 and ϕ_1

From these piecewise functions a rational basis function at an interior node x_i with local support on $[x_{i-1}, x_{i-1}+2h]$ is defined by

$$\psi_i(x) = \begin{cases} \phi_1(x-x_{i-1}), & x \in [x_{i-1}, x_i] \\ \phi_0(x-x_i), & x \in [x_i, x_{i+1}]. \end{cases} \qquad (2.21)$$

The graph of the basis function is shown in Figure 2.4. The rational basis function is called a (1,1) rational basis function in agreement with the degree of the numerator and the degree of the denominator of (2.20).

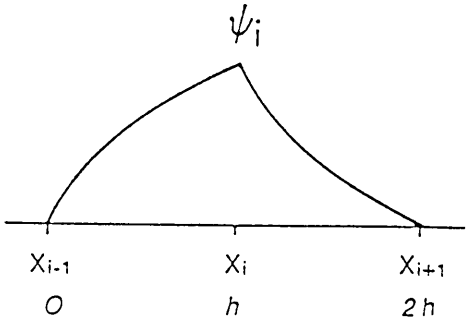

FIGURE 2.4: Rational basis function ψ_i

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Consider a (2,2) rational function on the interval [0,h], i.e.

$$R_{0,2}(x) = a + \frac{b}{1 + \frac{x}{h} + (\frac{x}{h})^2}, \quad x \in [0,h].$$

Define $\phi_1(x)$ and $\phi_0(x)$ by

$$\phi_1(x) = a_1 + \frac{b_1}{1 + \frac{x}{h} + (\frac{x}{h})^2}$$

and

$$\phi_0(x) = a_0 + \frac{b_0}{1 + \frac{x}{h} + (\frac{x}{h})^2},$$

where $x \in [0,h]$. From the interpolation constraints, $\phi_1(0) = 0$, $\phi_1(h) = 1$, $\phi_0(0) = 1$ and $\phi_0(h) = 0$, it follows that

$$\phi_1(x) = \frac{3}{2} - \frac{\frac{3}{2}}{1 + \frac{x}{h} + (\frac{x}{h})^2}$$

and $$\phi_0(x) = -\frac{1}{2} + \frac{\frac{3}{2}}{1 + \frac{x}{h} + (\frac{x}{h})^2},$$

where $x \in [0,h]$. The (2,2) rational basis function ψ_i at node x_i is then defined by (2.21).

In a similar manner the (3,3) rational basis function is defined by (2.21), where

$$\phi_1(x) = \frac{4}{3} - \frac{\frac{4}{3}}{1 + \frac{x}{h} + (\frac{x}{h})^2 + (\frac{x}{h})^3}$$

and

$$\phi_0(x) = -\frac{1}{3} + \frac{\frac{4}{3}}{1 + \frac{x}{h} + (\frac{x}{h})^2 + (\frac{x}{h})^3}.$$

Consequently, the (T,T) rational approximations on the interval [0,h] are defined by

$$\phi_1(x) = a_1 + \frac{b_1}{1 + \frac{x}{h} + \ldots + (\frac{x}{h})^T}$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

and

$$\phi_0(x) = a_0 + \frac{b_0}{1 + \frac{x}{h} + \ldots + (\frac{x}{h})^T}, \quad x \in [0,h].$$

Again, the interpolation constraints give $a_0 = -\frac{1}{T}$, $b_0 = 1 + \frac{1}{T}$, $a_1 = 1 + \frac{1}{T}$ and $b_1 = -1 - \frac{1}{T}$.

From these rational interpolatory functions a (T,T) rational basis function at x_i with support on $[x_{i-1},x_{i+1}]$ is defined by

$$\psi_i(x) = \begin{cases} \phi_1(x-x_{i-1}), & x \in [x_{i-1},x_i] \\ \phi_0(x-x_i), & x \in [x_i,x_{i+1}]. \end{cases} \qquad (2.22)$$

Note that all cases satisfy the relation

$$\phi_0(x) + \phi_1(x) = 1, \quad x \in [0,h].$$

This method to construct higher-order basis functions is unique in the sense that it is not extended to include more node points. This means that the danger of introducing real singularities is avoided and that all (T,T) rational basis functions couple only three adjoining nodes. Thus, the banded structure of the stiffness matrix in the Galerkin method will stay intact. It should be observed that these basis functions reduce to the piecewise linear basis functions in the limit as $T \longrightarrow \infty$.

The rational basis functions constructed are C^0 continuous and in order to approximate higher-order partial derivatives it is necessary to have C^1-continuous functions, namely Hermite rational basis functions.

Hermite rational basis functions can be constructed by considering

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

piecewise functions ψ_1^0, ψ_2^0, ψ_1^1, ψ_2^1 which satisfy the following interpolation properties on the interval $0 \leq \xi \leq 1$, namely

$$\psi_1^0(0) = 0, \quad \psi_1^0(1) = 1, \quad \frac{d\psi^0}{d\xi^1}(0) = \frac{d\psi^0}{d\xi^1}(1) = 0 ,$$

$$\psi_2^0(0) = 1, \quad \psi_2^0(1) = 0, \quad \frac{d\psi^0}{d\xi^2}(0) = \frac{d\psi^0}{d\xi^2}(1) = 0 ,$$

$$\psi_1^1(0) = \psi_1^1(1) = 0, \quad \frac{d\psi^1}{d\xi^1}(0) = 0, \quad \frac{d\psi^1}{d\xi^1}(1) = 1 ,$$

$$\psi_2^1(0) = \psi_2^1(1) = 0, \quad \frac{d\psi^1}{d\xi^2}(0) = 1, \quad \frac{d\psi^1}{d\xi^2}(1) = 0.$$

Let $\psi_1^0(\xi)$ be a (3,1) rational function of the form

$$\psi_1^0(\xi) = \frac{\xi^2(\alpha\xi + \beta)}{1 + \xi} , \quad 0 \leq \xi \leq 1.$$

Enforcing the interpolation constraints the parameters α and β are

$$\alpha = -3$$

and $$\beta = 5.$$

Similarly,

$$\psi_2^0(\xi) = \frac{(1 - \xi)^2(3\xi + 1)}{1 + \xi} .$$

Note that,

$$\psi_1^0(\xi) + \psi_2^0(\xi) = 1 , \quad 0 \leq \xi \leq 1.$$

Again, using the interpolation constraints, it follows that

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$\psi_1^1(\xi) = \frac{2\xi^2(\xi - 1)}{1 + \xi} \quad , \quad 0 \leq \xi \leq 1,$$

and

$$\psi_2^1(\xi) = \frac{\xi(\xi - 1)^2}{1 + \xi} \quad , \quad 0 \leq \xi \leq 1.$$

By analogous arguments, (3,2) Hermite rational basis functions are obtained, namely

$$\psi_1^0(\xi) = \frac{3\xi^2(-\xi + 2)}{1 + \xi + \xi^2} \quad , \quad 0 \leq \xi \leq 1,$$

$$\psi_2^0(\xi) = \frac{(1 - \xi)^2(3\xi + 1)}{1 + \xi + \xi^2} \quad , \quad 0 \leq \xi \leq 1,$$

$$\psi_1^1(\xi) = \frac{2\xi^2(\xi - 1)}{1 + \xi + \xi^2} \quad , \quad 0 \leq \xi \leq 1$$

and

$$\psi_2^1(\xi) = \frac{\xi(\xi - 1)^2}{1 + \xi + \xi^2} \quad , \quad 0 \leq \xi \leq 1.$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The graphs of the (3,1) Hermite rational basis functions are shown in Figure 2.5.

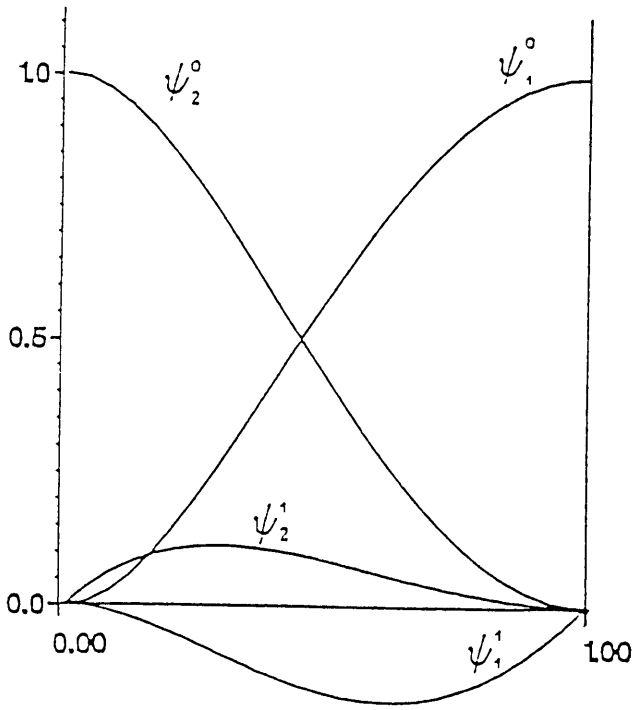

FIGURE 2.5 : (3,1) Hermite rational basis functions

This method is easily extendable to higher-order Hermite rational basis functions.

2.8 PROPERTIES OF INNER PRODUCTS OF RATIONAL BASIS FUNCTIONS

The properties of the rational basis functions are established by the following theorems:

Theorem 2.8.1

The (T,T) rational basis functions ψ_{j-1}, ψ_j and ψ_{j+1} satisfy the following relations:

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

(i) $(\psi_{j-1},\psi_j) = (\psi_{j+1},\psi_j)$

(ii) $(\psi_{j-1},\psi_j) + (\psi_j,\psi_j) + (\psi_{j+1},\psi_j) = h$

(iii) $(\psi'_{j-1},\psi_j) + (\psi'_{j+1},\psi_j) = 0$

(iv) $(\psi'_j,\psi_j) = 0$

(v) $(\psi'_{j-1},\psi_j) = -\frac{1}{2}$

(vi) $(\psi'_{j-1},\psi'_j) = (\psi'_{j+1},\psi'_j)$

(vii) $(\psi'_{j-1},\psi'_j) + (\psi'_j,\psi'_j) + (\psi'_{j+1},\psi'_j) = 0.$

Proof

Construct the (T,T) rational basis functions ψ_{j-1}, ψ_j and ψ_{j+1} defined by (2.22) in Figure 2.6.

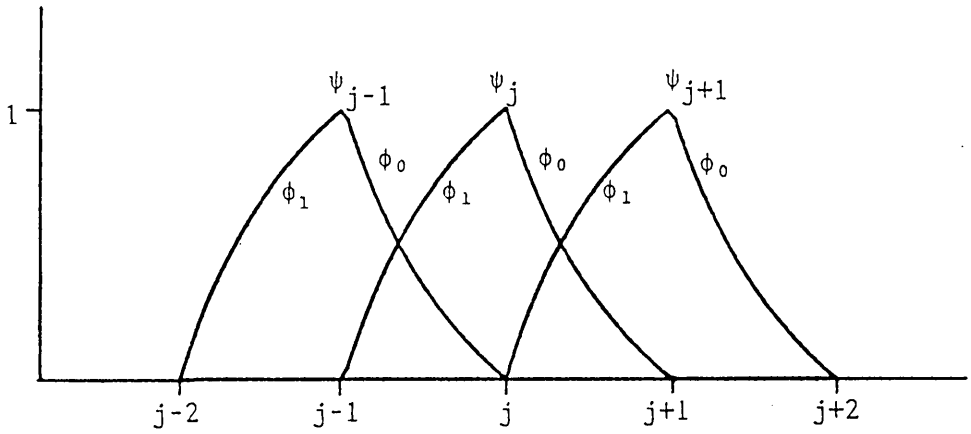

FIGURE 2.6 : Basis functions ψ_{j-1}, ψ_j and ψ_{j+1}

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

(i) From the construction of the basis functions, it follows that

$$(\psi_{j-1},\psi_j) = \int_0^h \phi_0\phi_1 \, dx$$

and

$$(\psi_{j+1},\psi_j) = \int_0^h \phi_1\phi_0 \, dx \ ,$$

therefore

$$(\psi_{j-1},\psi_j) = (\psi_{j+1},\psi_j).$$

(ii) Since $\phi_0 + \phi_1 = 1$, it follows that

$$\begin{aligned}
&(\psi_{j-1},\psi_j) + (\psi_j,\psi_j) + (\psi_{j+1},\psi_j) \\
&= \int_0^h \phi_0\phi_1 \, dx + \int_0^h \phi_1\phi_1 \, dx + \int_0^h \phi_0\phi_0 \, dx + \int_0^h \phi_0\phi_1 \, dx \\
&= \int_0^h \phi_1(\phi_0 + \phi_1)dx + \int_0^h (\phi_0 + \phi_1)\phi_0 \, dx \\
&= \int_0^h \phi_1 \, dx + \int_0^h \phi_0 \, dx \\
&= \int_0^h dx = h.
\end{aligned}$$

(iii) Now,

$$(\psi'_{j-1},\psi_j) = \int_0^h \phi'_0\phi_1 \, dx$$

and

$$(\psi'_{j+1},\psi_j) = \int_0^h \phi'_1\phi_0 \, dx$$

so that

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$
\begin{aligned}
&(\psi'_{j-1}, \psi_j) + (\psi'_{j+1}, \psi_j) \\
&= \int_0^h \phi'_0 \phi_1 \, dx + \int_0^h \phi'_1 \phi_0 \, dx \\
&= \phi_0 \phi_1 \Big|_0^h - \int_0^h \phi_0 \phi'_1 \, dx + \int_0^h \phi'_1 \phi_0 \, dx \\
&= 0.
\end{aligned}
$$

(iv) Consider,

$$
\phi'_1 = -b_1 \frac{\left(\frac{1}{h} + \ldots + T \frac{x^{T-1}}{h^T}\right)}{\left(1 + \frac{x}{h} + \ldots + \left(\frac{x}{h}\right)^T\right)^2}
$$

and

$$
\phi'_0 = -b_0 \frac{\left(\frac{1}{h} + \ldots + T \frac{x^{T-1}}{h^T}\right)}{\left(1 + \frac{x}{h} + \ldots + \left(\frac{x}{h}\right)^T\right)^2}
$$

Since $b_1 = -b_0$, it follows that

$$
\phi'_1 = -\phi'_0 .
$$

Now,

$$
(\psi'_j, \psi_j) = \int_0^h (\phi'_1 \phi_1 + \phi'_0 \phi_0) dx
$$

$$
= -\int_0^h \phi'_0 \phi_1 \, dx + \int_0^h \phi'_0 \phi_0 \, dx
$$

$$
= -\phi_0 \phi_1 \Big|_0^h + \int_0^h \phi_0 \phi'_1 \, dx + \int_0^h \phi'_0 \phi_0 \, dx
$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$= \int_0^h \phi_0 \phi_1' \, dx + \int_0^h \phi_0' \phi_0 \, dx$$

$$= 0 \, .$$

(v) $$(\psi_{j-1}', \psi_j) = \int_0^h \phi_0' \phi_1 \, dx$$

$$= \phi_0 \phi_1 \Big|_0^h - \int_0^h \phi_0 \phi_1' \, dx$$

$$= - \int_0^h \phi_0 \phi_1' \, dx$$

$$= \int_0^h \phi_0 \phi_0' \, dx$$

$$= \phi_0 \phi_0 \Big|_0^h - \int_0^h \phi_0' \phi_0 \, dx.$$

From this it follows that

$$2 \int_0^h \phi_0' \phi_0 \, dx = -1$$

and therefore

$$(\psi_{j-1}', \psi_j) = - \frac{1}{2} \, .$$

(vi) $$(\psi_{j-1}', \psi_j') = \int_0^h \phi_0' \phi_1' \, dx$$

$$= (\psi_j', \psi_{j-1}') .$$

(vii) It follows that

$$(\psi_{j-1}', \psi_j') = \int_0^h \phi_0' \phi_1' \, dx,$$

$$(\psi_{j+1}', \psi_j') = \int_0^h \phi_0' \phi_1' \, dx$$

and

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$(\psi_j', \psi_j') = \int_0^h \phi_0' \phi_0' \, dx + \int_0^h \phi_1' \phi_1' \, dx.$$

The above relations and the fact that $\phi_0' = -\phi_1'$ yield the relation

$$(\psi_{j-1}', \psi_j') + (\psi_{j+1}', \psi_j') + (\psi_j', \psi_j') = 0.$$

Theorem 2.8.2

For the (1,1) rational basis functions ψ_{j-1}, ψ_j and ψ_{j+1} the following relations hold:

(i) $(\psi_j, \psi_j) = 9h - 12h\ell n\ 2$

(ii) $(\psi_{j-1}, \psi_j) = (\psi_j, \psi_{j+1}) = 6h\ell n\ 2 - 4h$

(iii) $(\psi_j', \psi_j') = \frac{7}{3h}$

(iv) $(\psi_{j-1}', \psi_j') = (\psi_j', \psi_{j+1}') = \frac{-7}{6h}$.

Proof

(i) $(\psi_j, \psi_j) = (\phi_1, \phi_1) + (\phi_0, \phi_0)$

$$= \int_0^h \left(\frac{2x}{h+x}\right)^2 dx + \int_0^h \left(\frac{h-x}{h+x}\right)^2 dx$$

$$= 6h - 8h\ \ell n\ 2 + 3h - 4h\ \ell n\ 2$$

$$= 9h - 12h\ \ell n\ 2.$$

(ii) $(\psi_{j-1}, \psi_j) = (\phi_0, \phi_1)$

$$= \int_0^h \frac{h-x}{h+x} \cdot \frac{2x}{h+x} \, dx$$

$$= 6h\ \ell n\ 2 - 4h$$

and $(\psi_j, \psi_{j+1}) = (\phi_0, \phi_1) = 6h\ \ell n\ 2 - 4h.$

(iii) $(\psi_j', \psi_j') = (\phi_1', \phi_1') + (\phi_0', \phi_0')$

$$= \int_0^h \left[\frac{2h}{(h+x)^2}\right]^2 dx + \int_0^h \left[\frac{-2h}{(h+x)^2}\right]^2 dx$$

$$= \frac{7}{6h} + \frac{7}{6h}$$

$$= \frac{7}{3h} .$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

(iv) $$(\psi'_{j-1},\psi'_j) = (\psi'_j,\psi'_{j+1})$$
$$= (\phi'_0,\phi'_1)$$
$$= \int_0^h \frac{2h}{(h+x)^2}\,\frac{-2h}{(h+x)^2}\,dx$$
$$= -\frac{7}{6h}\,.$$

For the higher-order rational basis functions the integration becomes unwieldy. Numerical integration by a Gaussian Legendre quadrature with 20 points was used to obtain the following numerical values for the (2,2) rational basis function:

(i) $(\psi_j,\psi_j) = 0{,}6862h$

(ii) $(\psi_{j-1},\psi_j) = (\psi_j,\psi_{j+1}) = 0{,}1569h$

(iii) $(\psi'_j,\psi'_j) = 2{,}2092/h$

(iv) $(\psi'_{j-1},\psi'_j) = (\psi'_j,\psi'_{j+1}) = -1{,}1046/h.$

Likewise, for the (3,3) rational basis function:

(i) $(\psi_j,\psi_j) = 0{,}6877h$

(ii) $(\psi_{j-1},\psi_j) = (\psi_j,\psi_{j+1}) = 0{,}1562h$

(iii) $(\psi'_j,\psi'_j) = 2{,}1528/h$

(iv) $(\psi'_{j-1},\psi'_j) = (\psi'_j,\psi'_{j+1}) = -1{,}0764/h.$

Theorem 2.8.3

For the (T,T) rational basis functions, ψ_{j-1} and ψ_j, the inner product

$$-h(\psi'_{j-1},\psi'_j) \longrightarrow 1$$

if the order T tends to infinity.

Proof

The rational basis functions $\psi_j(x)$ and $\psi_{j-1}(x)$ on the interval [0,h] are defined by

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$\psi_{j-1}(x) = -\frac{1}{T} + \frac{1+\frac{1}{T}}{1+\frac{x}{h}+\ldots+(\frac{x}{h})^T} \quad , \quad x \in [0,h] \ ,$$

and

$$\psi_j(x) = 1 + \frac{1}{T} - \frac{1+\frac{1}{T}}{1+\frac{x}{h}+\ldots+(\frac{x}{h})^T} \quad , \quad x \in [0,h] \ ,$$

respectively. The derivatives of the basis functions are

$$\psi'_{j-1}(x) = \frac{-(1+\frac{1}{T})(1+\ldots+T(\frac{x}{h})^{T-1})}{h\{1+\frac{x}{h}+\ldots+(\frac{x}{h})^T\}^2}$$

and

$$\psi'_j(x) = \frac{(1+\frac{1}{T})(1+\ldots+T(\frac{x}{h})^{T-1})}{h\{1+\frac{x}{h}+\ldots+(\frac{x}{h})^T\}^2} \ .$$

Now,

$$-h(\psi'_{j-1},\psi'_j) = \int_0^1 \frac{(1+\frac{1}{T})^2(1+2x+\ldots+Tx^{T-1})^2}{(1+x+x^2+\ldots+x^T)^4}\,dx.$$

Define the function f_T by

$$f_T(x) = 1 - (1+\frac{1}{T})^2\frac{(1+2x+\ldots+Tx^{T-1})^2}{(1+x+x^2+\ldots+x^T)^4} \quad , \quad 0 \le x \le 1.$$

Now, for $0 \le x < 1$

$$\lim_{T\to\infty}(1+\frac{1}{T})^2\frac{(1+2x+\ldots+Tx^{T-1})^2}{(1+x+x^2+\ldots+x^T)^4}$$

$$= \lim_{T\to\infty}\frac{(1+\frac{1}{T})^2(\frac{1}{1-x}(\frac{1-x^T}{1-x}-Tx^T))^2}{(\frac{1-x^T}{1-x})^4}$$

$$= \lim_{T\to\infty}(1+\frac{1}{T})^2\frac{(1-(T+1)x^T+Tx^{T+1})^2}{(1-x^{T+1})^4}$$

$$= 1$$

and for $x = 1$,

$$\frac{(1+\frac{1}{T})^2[(T+1)\frac{T}{2}]^2}{[T+1]^4}$$

$$= \frac{1}{4} \ .$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Hence, the sequence of continuous functions $f_T(x)$ satisfies

$$\lim_{T\to\infty} f_T(x) = f(x),$$

where

$$f(x) = \begin{cases} 0 \ , & 0 \le x < 1 \\ \frac{3}{4} \ , & x = 1 \end{cases} .$$

Since, for all $0 \le x \le 1$, $T > 0$

$$(1 + 2x + \ldots + Tx^{T-1})$$
$$\le (1 + x + \ldots + x^T)^2$$
$$= 1 + 2x + \ldots + Tx^{T-1} + (T + 1)x^T + \ldots + x^{2T}$$

it follows that

$$\frac{(1 + 2x + \ldots + Tx^{T-1})^2}{(1 + x + \ldots + x^T)^4} \le 1.$$

Therefore,

$$|f_T(x)| \le |(1 + \tfrac{1}{T})^2 - 1| \le 3$$

for all $0 \le x \le 1$ and $T \ge 1$.

Hence, a function exists, say $g(x) = 3$, on $0 \le x \le 1$ such that

$$|f_T(x)| \le g(x) \ , \quad T = 1,2,\ldots$$

Then, according to Lebesque's dominated convergence theorem [38,39]

$$\lim_{T\to\infty} \int_0^1 f_T \, dx = \int_0^1 f \, dx = 0.$$

Thus,

$$\lim_{T\to\infty} -h(\psi'_{j-1}, \psi'_j) = \lim_{T\to\infty} \int_0^1 (1 + \tfrac{1}{T})^2 \frac{(1 + 2x + \ldots + Tx^{T-1})^2}{(1 + x + \ldots x^T)^4} \, dx$$
$$= 1 \ ,$$

which proves the theorem.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

2.9 ERROR ESTIMATES FOR RATIONAL APPROXIMATION

In this section the approximation abilities of the rational basis functions are investigated. However, the subject remains a study in itself, because the rational basis function represents a nonlinear approximation. The objective here is only to obtain an error bound for the approximation and not to establish an optimal error bound.

Let R_T^* denote the set of all rational functions of order T, i.e., $r_T \in R_T^*$ if

$$r_T(x) = \frac{a_m x^m + \ldots + a_0}{b_k x^k + \ldots + b_0}, \quad m \leq T, \quad k \leq T.$$

From this definition it is clear that in particular the $(1,1)$ rational basis function belongs to R_1^*. Denote by $\|u\|$ the supremum norm of the function u on the interval $[a,b]$, i.e.

$$\|u\| = \sup_{x \in [a,b]} |u(x)|.$$

Let $R_T(u;[a,b])$ be the error corresponding to the best uniform approximation of the function u in the interval $[a,b]$ by means of rational functions of order T, that is

$$R_T(u;[a,b]) = \inf\{\|u - r\|: r \in R_T^*\}.$$

Popov [31] has shown that if the k-th derivative of the function u, namely $u^{(k)}$, is of bounded variation in the interval $[a,b]$, then

$$R_T(u;[a,b]) \leq C(k)(b - a)^k V_a^b u^{(k)} / T^{k+1}$$

where $V_a^b \phi$ denotes the variation of the function ϕ on the interval $[a,b]$ and $C(k)$ is a constant, depending only on k.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

A rough estimate of the rational approximation to a function u is now obtained, namely:

Theorem 2.9.1

If $u \in C^2[a,b]$, and u_N is the $(1,1)$ rational approximation to u, and assume that u' and u'' are bounded, then

$$\|u - u_N\| < Ch \tag{2.23}$$

where $\|\cdot\|$ is the supremum norm on the subinterval $[x_i,x_{i+1}]$ of length h and C a constant.

Proof

Divide the interval $[a,b]$ into N subintervals of length h. Let u_N approximate u over the interval $[x_i,x_{i+1}]$, with $x_{i+1} = x_i + h$. Since u_N is a rational approximation of order 1,

$$u_N(x) = u(x_i)\phi_0(x - x_i) + u(x_{i+1})\phi_1(x - x_i) \tag{2.24}$$

where ϕ_0 and ϕ_1 are $(1,1)$ rational basis functions and $x \in [x_i,x_{i+1}]$.

The $(1,1)$ rational approximation over a fixed interval h is

$$u_N(x) = \frac{(2u(x_{i+1}) - u(x_i))x + hu(x_i)}{h + x}$$

$$= \frac{P(x)}{Q(x)}, \quad x \in [0,h].$$

Hence, the error in the approximation is

$$e(x) = u(x) - u_N(x)$$
$$= \frac{Q(x)u(x) - P(x)}{Q(x)}.$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Define

$$g(x) = (x + h)u(x) - (2u(x_{i+1}) - u(x_i))x - hu(x_i)$$
$$- \frac{x(x - h)}{\bar{x}(\bar{x} - h)}[(\bar{x} + h)u(\bar{x}) - (2u(x_{i+1}) - u(x_i))\bar{x} - hu(x_i)]$$

where $\bar{x} \in (0,h)$.

By definition, $g(0) = 0$, $g(h) = 0$ and $g(\bar{x}) = 0$.

Since $g(x)$ has three zeros, a point $\xi \in (0,h)$ exists such that $g''(\xi) = 0$. Therefore,

$$Q(\bar{x})e(\bar{x}) = \frac{\bar{x}(\bar{x} - h)}{2}[(\xi + h)u''(\xi) + 2u'(\xi)]$$

for $\bar{x},\xi \in (0,h)$.

Thus,

$$\|Qe\| \leq \frac{h^2}{4}[h\|u''\| + \|u'\|]$$

so that

$$\|e\| = \|\frac{Qe}{Q}\| \leq \frac{h}{4}[h\|u''\| + \|u'\|].$$

Assuming the derivatives to be bounded, it follows that $\|e\| \leq Ch$, where C is a constant. This gives another bound for the interpolation error in terms of the step size, which gives the assurance that the interpolation error would decrease as h tends to zero. It would be very difficult, however, to establish an optimal error bound. Therefore, in order to prove convergence of a numerical scheme the well-known Lax equivalence theorem [1,35], which unfortunately holds only for linear equations, will be reverted to.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

2.10 LAX EQUIVALENCE THEOREM

Consider a linear initial value problem

$$u_t = Lu, \quad t > 0 \tag{2.25}$$

and

$$u(x,0) = f(x).$$

The initial value problem can be approximated by a numerical scheme of the form

$$U^{n+1} = QU^n , \quad n > 0 , \tag{2.26}$$

where $U_i^n \simeq u(ih,nk)$ and Q is a difference operator. The difference scheme (2.26) is stable if constants K and β and some norm $\|\cdot\|$ exist such that,

$$\|U^n\| \leq Ke^{\beta t}\|f\| ,$$

where $t = nk$ and K and β are independent of h and k.

A difference scheme (2.26) is consistent up to time t in a norm $\|\cdot\|$ with (2.25) if the exact solution u to the initial value problem (2.25) "almost" satisfies the difference scheme, that is

$$u(ih,(n+1)k) = Qu(ih,nk) + \tau(h,k) ,$$

where $\tau(h,k) \rightarrow 0$ as $h,k \rightarrow 0$. The term, $\tau(h,k)$ is called the local truncation error at ih and nk.

Finally, the Lax equivalence theorem [1,35] is as follows:
Given a properly posed linear initial-value problem and a linear numerical approximation to it that satisfies the consistency condition, stability is the necessary and sufficient condition for convergence.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

2.11 GALERKIN'S METHOD APPLIED TO A STIFF ORDINARY DIFFERENTIAL EQUATION

The new method consisting of rational basis functions is applied to the test problem [16],

$$u' = \lambda(u - g(x)) + g'(x)$$

with $u(0) = 3$, $g(x) = \sin(0,1x) + 2$ and analytical solution

$$u(x) = g(x) + (u(0) - g(0))e^{\lambda x},$$

where λ is a stiffness ratio.

Let

$$u \simeq u_N = \sum_{i=0}^{N} U_i \psi_i(x).$$

with ψ_i a $(1,1)$ rational basis function.

The method of Galerkin gives

$$(u_N' - \lambda u_N, \psi_j) = (-\lambda g(x) + g'(x), \psi_j), \quad j=0,1,\ldots,N$$

or

$$\sum_{i=0}^{N} (\psi_i', \psi_j)U_i - \lambda \sum_{i=0}^{N} (\psi_i, \psi_j)U_i$$
$$= -\lambda(\sin(0,1x) + 2, \psi_j) + 0,1(\cos(0,1x), \psi_j), \quad j=0,\ldots,N$$

where $(u,v) = \int_0^1 uv \, dx$. The initial condition was imposed by replacing the first equation by $U_0 = 3$.

In matrix notation write the above system as

$$A\bar{U} = B$$

where A is a $(N + 1) \times (N + 1)$ matrix, B a $(N + 1) \times 1$ matrix and $\bar{U} = (U_0,\ldots,U_N)^T$. This problem has been solved on the interval $[0,1]$ for $h = 0,01$ and $\lambda = -10$. The numerical results of the method are compared with the analytical solution and the numerical

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

solution obtained by using linear hat functions. The results are given in Table 2.1. The errors are the relative errors.

TABLE 2.1

x	ANALYTICAL	LINEAR	RATIONAL	ERROR LINEAR	ERROR RATIONAL
0,1	2,377788	2,373429	2,377052	1,9E - 3	3,5E - 4
0,2	2,155334	2,142928	2,147861	5,8E - 3	3,5E - 3
0,3	2,079783	2,058134	2,063536	1,0E - 2	7,8E - 3
0,4	2,058305	2,026942	2,032506	1,5E - 2	1,3E - 2
0,5	2,056717	2,015471	2,021081	2,0E - 2	1,7E - 2
0,6	2,062443	2,011256	2,016866	2,5E - 2	2,2E - 2
0,7	2,070855	2,009713	2,015297	3,0E - 2	2,7E - 2
0,8	2,080250	2,009156	2,014695	3,4E - 2	3,2E - 2
0,9	2,090002	2,008965	2,014436	3,9E - 2	3,6E - 2
1,0	2,099879	2,008914	2,014288	4,3E - 1	4,1E - 2

In the interval [0;0,3] the solution has a steep gradient since it changes from 3 to 2,079. In the vicinity of this steep gradient the rational approximation is very effective as shown by the results in the table. This indicates that the rational approximation compares very favourably to the linear approximation for linear problems.

2.12 CONCLUSIONS

In this chapter the Galerkin method was introduced and the concept of a rational basis function was established. C^0-continuous rational basis functions and higher-order C^0 rational basis functions were constructed without coupling more than three adjoining nodes. This means that only tridiagonal matrices are obtained in the Galerkin method. C^1 Hermite rational basis functions were also constructed to

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

achieve continuity of the function and its first derivative at a node in the computational mesh.

Finally, the Galerkin method with (1,1) rational basis functions was applied to a stiff ordinary differential equation and it was found to perform excellently in the vicinity of a steep gradient.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

CHAPTER 3

CONVECTION-DIFFUSION EQUATIONS

3.1 INTRODUCTION

In transport equations which describe physical processes such as fluid flow, heat and mass transfer in chemical and nuclear engineering, etc., diffusion and convection may be involved. The dominance of one effect over the other will determine whether the transport is due to diffusion or to convection.

Numerical approximations of transport equations introduce errors, and thereby influencing the relative role between the diffusion and convection of a fluid. The consequent effects of these approximations on the numerical solution are observed either as numerical dissipation, in which sharp fronts or gradients are smeared, or as numerical oscillations, in which different components of the initial solution propagate with phase and amplitude errors. The last phenomenon is attributed to the presence of convection. Oscillations in a piecewise linear element solution of the convection-diffusion equation appear if the cell Peclet number, $\frac{\delta h}{\epsilon}$, exceeds two [7]. The disturbance in the solution is rectified by using the technique of upwinding [5,7]. In the Petrov-Galerkin method test functions which are biased in the upstream direction are implemented. However, the rational basis functions in Section 2.7, see Figure 2.3, are naturally biased in the upstream direction and therefore simulate the effect of upwinding in a natural manner. In this chapter a Galerkin method is formulated by using rational basis functions. The characteristics of the rational difference scheme are investigated with regard to consistency,

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

stability and numerical convergence of the method. A comparison is drawn with linear shaped basis functions and upwinding. Numerical results are also presented.

3.2 CONVECTION-DIFFUSION EQUATION

Consider

$$u_t = \epsilon u_{xx} - \delta u_x \ , \ \epsilon > 0, \ \delta > 0, \ x \in (0,1), \ t = (0,T] \qquad (3.1)$$

with initial condition

$$u(x,0) = u_0(x), \quad x \in (0,1).$$

Three different boundary values will be employed, namely

(a) Homogeneous Dirichlet

(b) Periodic

(c) Neumann.

The diffusion coefficient, ϵ, describes, for example, the spreading of ink in stationary water due to the Brownian motion of the water molecules. If, however, the water is in motion, then in addition to diffusion, the ink is also convected by the moving fluid. The convection is indicated by δ.

3.3 DISCRETISING THE CONVECTION-DIFFUSION EQUATION

3.3.1 Homogeneous Dirichlet boundary conditions

Consider the convection-diffusion equation

$$u_t = \epsilon u_{xx} - \delta u_x \ , \quad \epsilon > 0, \ \delta > 0, \ x \in (0,1), \ t \in (0,T] \qquad (3.2)$$

with the initial condition

$$u(x,0) = u_0(x), \quad x \in (0,1) \qquad (3.3)$$

and boundary conditions

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$u(0,t) = 0, \quad t \geq 0 \tag{3.4a}$$

$$u(1,t) = 0, \quad t \geq 0. \tag{3.4b}$$

Divide the interval $[0,1]$ into N subintervals of length h and the time interval $[0,T]$ into J subintervals of length k. Introduce rational basis functions

$$\psi_i(x) \quad , \quad i=1,\ldots,N-1,$$

which are independent of time at the nodes. Since the boundary values are zero, the associated system will only be solved in the $(N-1)$ interior nodes (see Figure 3.1).

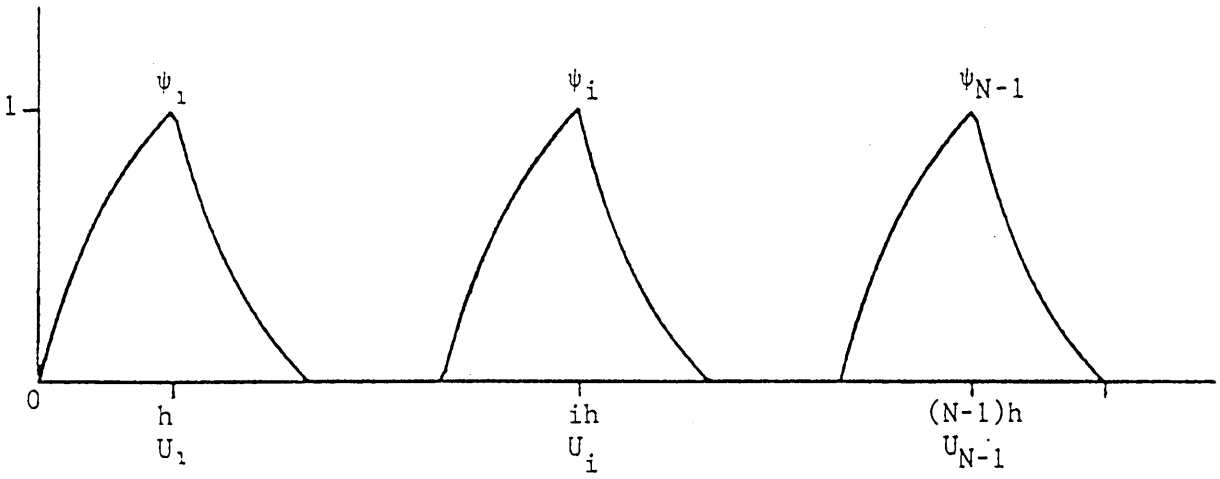

FIGURE 3.1 : Basis functions for space discretisation

Galerkin's method seeks an approximate solution to (3.2) in the form

$$u(x,t) \simeq \sum_{i=1}^{N-1} U_i(t)\psi_i(x) \tag{3.5}$$

which satisfies the following system

$$(u_t + \delta u_x - \epsilon u_{xx}, \psi_j) = 0 \quad j=1,\ldots,N-1, \tag{3.6}$$

where $(u,v) = \int_0^1 uv\, dx$. The substitution of (3.5) into (3.6) and the fact that $(u_{xx}, \psi_j) = -(u_x \psi_j')$ result in

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$(\psi_{j-1},\psi_j)\dot{U}_{j-1} + (\psi_j,\psi_j)\dot{U}_j + (\psi_{j+1},\psi_j)\dot{U}_{j+1}$$
$$+ \delta(\psi'_{j-1},\psi_j)U_{j-1} + \delta(\psi'_j,\psi_j)U_j + \delta(\psi'_{j+1},\psi_j)U_{j+1}$$
$$+ \epsilon(\psi'_{j-1},\psi'_j)U_{j-1} + \epsilon(\psi'_j,\psi'_j)U_j + \epsilon(\psi'_{j+1},\psi'_j)U_{j+1} = 0$$
$$\text{for } j=1,\ldots,N-1, \text{ where } \dot{U}(t) = \frac{dU(t)}{dt} . \qquad (3.7)$$

Discretize in time the approximating system (3.7) of first-order ordinary differential equations. For this purpose, let

$$U_i^n \simeq u(ih,nk)$$

and

$$T_\theta(U_i^n) = (1 - \theta)U_i^n + \theta U_i^{n+1}$$

where $\theta \in [0,1]$ is a parameter that determines the difference approximation in time. For example, $\theta = 0$ and 1 will correspond to forward and backward differencing respectively. The nodal values at a specific time level $(n + 1)k$ are determined from the following scheme:

$$(\psi_{j-1},\psi_j)(U_{j-1}^{n+1} - U_{j-1}^n) + (\psi_j,\psi_j)(U_j^{n+1} - U_j^n) + (\psi_{j+1},\psi_j)(U_{j+1}^{n+1} - U_{j+1}^n)$$
$$+ \delta k(\psi'_{j-1},\psi_j)T_\theta(U_{j-1}^n) + \delta k(\psi'_j,\psi_j)T_\theta(U_j^n) + \delta k(\psi'_{j+1},\psi_j)T_\theta(U_{j+1}^n)$$
$$+ \epsilon k(\psi'_{j-1},\psi'_j)T_\theta(U_{j-1}^n) + \epsilon k(\psi'_j,\psi'_j)T_\theta(U_j^n) + \epsilon k(\psi'_{j+1},\psi'_j)T_\theta(U_{j+1}^n)$$
$$= 0 \qquad (3.8)$$

for $j=1,\ldots,N-1$.

If $\theta = 1$ an implicit backward-difference scheme results, which in matrix notation is

$$A\bar{U}^{n+1} = B\bar{U}^n \qquad (3.9)$$

where $A = (a_{ij})$, $B = (b_{ij})$, $1 \le i, j \le (N-1)$ and
$\bar{U}^n = (U_1^n, \ldots, U_{N-1}^n)^T$.

The elements of A and B for the (1,1) rational basis function are:

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$
\begin{aligned}
a_{11} &= 9h - 12h \ln 2 + 7k\epsilon/3h \\
a_{12} &= 6h\ln 2 - 4h - 7k\epsilon/6h + k\delta/2 \\
a_{i,i-1} &= 6h \ln 2 - 4h - 7k\epsilon/6h - k\delta/2 \quad , \ i=2,\ldots,N-2 \\
a_{i,i} &= 9h - 12h \ln 2 + 7\, k\epsilon/3h \quad , \ i=2,\ldots,N-2 \\
a_{i,i+1} &= 6h \ln 2 - 4h - 7k\epsilon/6h + k\delta/2 \quad , \ i=2,\ldots,N-2 \\
a_{N-1,N-2} &= 6h \ln 2 - 4h - 7k\epsilon/6h - k\delta/2 \\
a_{N-1,N-1} &= 9h - 12h\ln 2 + 7k\epsilon/3h \\
b_{11} &= 9h - 12h\ln 2 \\
b_{12} &= 6h\ln 2 - 4h \\
b_{i,i-1} &= 6h\ln 2 - 4h \quad , \ i=2,\ldots,N-2 \\
b_{i,i} &= 9h - 12h\ln 2 \quad , \ i=2,\ldots,N-2 \\
b_{i,i+1} &= 6h \ln 2 - 4h \quad , \ i=2,\ldots,N-2 \\
b_{N-1,N-2} &= 6h\ln 2 - 4h \\
b_{N-1,N-1} &= 9h - 12h\ln 2.
\end{aligned}
$$

Elements not mentioned are zero.

If, however, $\theta = 0$ then a numerical scheme exists which is forward in time. From (3.8) with $\theta = 0$ the j-th equation for the (1,1) rational basis function is typically

$$
\begin{aligned}
&(6h\ln 2 - 4h)U_{j-1}^{n+1} + (9h - 12h\ln 2)U_j^{n+1} + (6h\ln 2 - 4h)U_{j+1}^{n+1} \\
&= (6h\ln 2 - 4h + \frac{7\epsilon k}{6h} + \frac{\delta k}{2})U_{j-1}^{n} \\
&\quad + (9h - 12h\ln 2 - \frac{7\epsilon k}{3h})U_j^n \\
&\quad + (6h\ln 2 - 4h + \frac{7\epsilon k}{6h} - \frac{\delta k}{2})U_{j+1}^n. \qquad (3.10)
\end{aligned}
$$

In matrix notation this forward-difference scheme may be written in the form

$$A(U^{n+1} - U^n) + kBU^n = 0$$

where $U^n = (U_1^n,\ldots,U_{N-1}^n)^T$ and A and B are tridiagonal matrices.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

This provides an explicit scheme which gives nodal values at time level $(n+1)k$.

3.3.2 Periodic boundary conditions

The boundary conditions (3.4) are replaced by

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

To find the approximate solution at the end point, a basis function is added at $x = 1$. Schematically the basis functions are represented as follows.

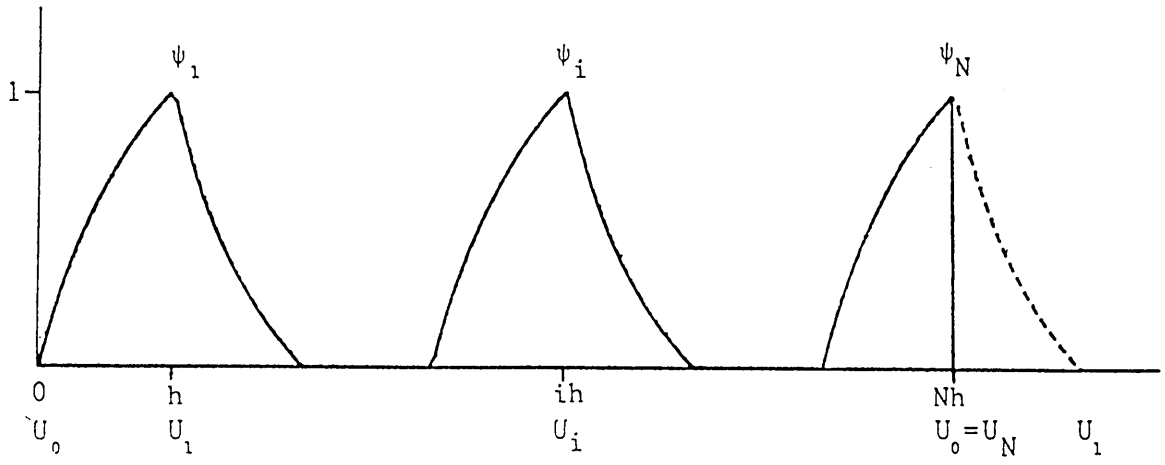

FIGURE 3.2: Basis functions for periodic boundary conditions

The exact solution $u(x,t)$ is approximated by

$$u(x,t) \simeq u_N(x,t) = \sum_{i=1}^{N} U_i(t)\psi_i(x).$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The Galerkin approximant results in the system

$$\sum_{i=1}^{N} (\psi_i,\psi_j)\dot{U}_i(t) + \epsilon \sum_{i=1}^{N} (\psi_i',\psi_j')U_i(t)$$

$$+ \delta \sum_{i=1}^{N} (\psi_i',\psi_j)U_i(t) = 0, \quad j=1,\ldots,N.$$

By using the implicit backward-difference scheme in time, a similar system to (3.9) is obtained.

3.3.3 Neumann boundary conditions

Consider the boundary conditions

$$u(0,t) = 0 \quad , \quad t > 0$$

$$\frac{\partial u}{\partial x}(1,t) = 0 \quad , \quad t > 0.$$

The Galerkin approximant $u_N(x,t)$ is defined by

$$(\frac{\partial u_N}{\partial t}, v) + \epsilon(\frac{\partial u_N}{\partial x}, \frac{\partial v}{\partial x}) + \delta(\frac{\partial u_N}{\partial x}, v) + <\frac{\partial u_N}{\partial x}, v> = 0$$

where

$$<\frac{\partial u_N}{\partial x}, v> = (\frac{\partial u_N}{\partial x} - 0)v\Big|_{x=1}$$

is added to include the Neumann boundary condition and

$v \in \text{span}\{\psi_1,\psi_2,\ldots,\psi_N\}$. This results in the system

$$\sum_{i=1}^{N} (\psi_i,\psi_j)\dot{U}_i(t) + \epsilon \sum_{i=1}^{N} (\psi_i',\psi_j')U_i(t)$$

$$+ \delta \sum_{i=1}^{N} (\psi_i',\psi_j)U_i(t) + (\sum_{i=1}^{N} \psi_i' U_i \psi_j)\Big|_{x=1} = 0, \quad j=1,\ldots,N.$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The basis functions are shown in Figure 3.3.

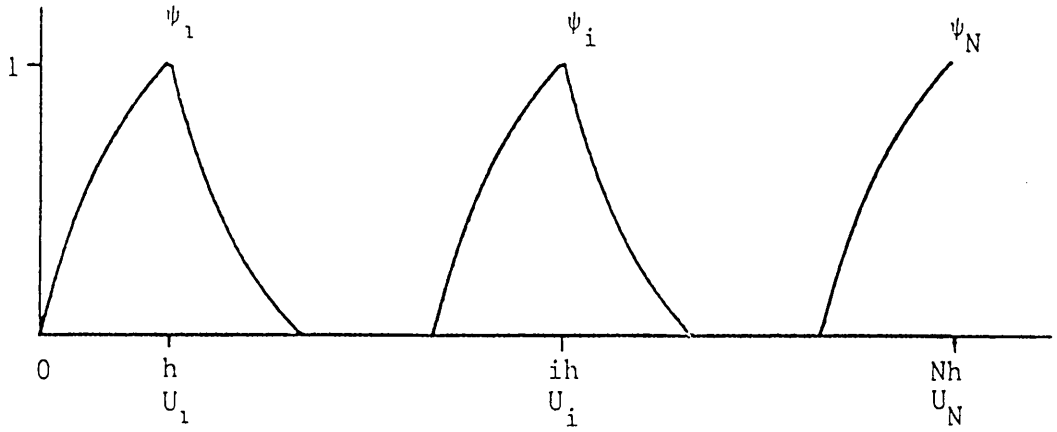

FIGURE 3.3 : Basis functions for Neumann boundary conditions

By using a backward-difference scheme in time, an implicit scheme similar to (3.9) is obtained.

3.4 CONSISTENCY

Consider the discrete equation (3.8) with $\theta = 0$:

$$(\psi_{j-1},\psi_j)U_{j-1}^{n+1} + (\psi_j,\psi_j)U_j^{n+1} + (\psi_{j+1},\psi_j)U_{j+1}^{n+1}$$

$$- \{(\psi_{j-1},\psi_j) - \epsilon k\,(\psi'_{j-1},\psi'_j) - \delta k\,(\psi'_{j-1},\psi_j)\}U_{j-1}^{n}$$

$$- \{(\psi_j,\psi_j) - \epsilon k\,(\psi'_j,\psi'_j) - \delta k\,(\psi'_j,\psi_j)\}U_j^{n}$$

$$- \{(\psi_{j+1},\psi_j) - \epsilon k\,(\psi'_{j+1},\psi'_j) - \delta k\,(\psi'_{j+1},\psi_j)\}U_{j+1}^{n} = 0. \qquad (3.11)$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Define the local truncation error $T_{j,n}$ at (jh, nk) by

$$T_{j,n} = \frac{1}{hk}\{(\psi_{j-1},\psi_j)u((j-1)h,(n+1)k) + (\psi_j,\psi_j)u(jh,(n+1)k)$$

$$+ (\psi_{j+1},\psi_j)u((j+1)h,(n+1)k) - \{(\psi_{j-1},\psi_j) - \epsilon k\,(\psi'_{j-1},\psi'_j)$$

$$- \delta k(\psi'_{j-1},\psi_j)\}u((j-1)h,nk) - \{(\psi_j,\psi_j) - \epsilon k(\psi'_j,\psi'_j)$$

$$- \delta k(\psi'_j,\psi_j)\}u(jh,nk) - \{(\psi_{j+1},\psi_j) - \epsilon k(\psi'_{j+1},\psi'_j)$$

$$- \delta k(\psi'_{j+1},\psi_j)\}u((j+1)h,nk)\}.$$

The application of the Taylor series

$$u((j\pm1)h,nk) = u \pm hu_x + \frac{h^2}{2}u_{xx} \pm \frac{h^3}{6}u_{xxx} + \frac{h^4}{24}u_{xxxx} + O(h^5)$$

and $u((j\pm1)h,(n+1)k) = u \pm hu_x + ku_t + \frac{1}{2}(h^2u_{xx} \pm 2hk\,u_{xt} + k^2u_{tt})$

$$+ \frac{1}{6}(\pm h^3u_{xxx} \pm 3hk^2u_{xtt} + 3h^2ku_{txx} + k^3u_{ttt})$$

$$+ O(h^4,h^3k,h^2k^2,hk^3,k^4)$$

yields

$$T_{j,n} = \frac{1}{hk}[(\psi_{j-1},\psi_j)\{u - hu_x + ku_t + \frac{1}{2}(h^2u_{xx} - 2hku_{xt} + k^2u_{tt})$$

$$+ \frac{1}{6}(-h^3u_{xxx} - 3hk^2u_{xtt} + 3h^2ku_{txx} + k^3u_{ttt})\}$$

$$+ (\psi_j,\psi_j)\{u + ku_t + \frac{1}{2}k^2u_{tt} + \frac{k^3}{6}u_{ttt}\}$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$+ (\psi_{j+1},\psi_j)\{u + hu_x + ku_t + \frac{1}{2}(h^2u_{xx} + 2hku_{xt}+k^2u_{tt})$$

$$+ \frac{1}{6}(h^3u_{xxx} + 3hk^2u_{xtt} + 3h^2ku_{txx} + k^3u_{ttt})\}$$

$$- \{(\psi_{j-1},\psi_j) - \epsilon k(\psi'_{j-1},\psi'_j) - \delta k(\psi'_{j-1},\psi_j)\}\{u - hu_x + \frac{h^2}{2}u_{xx} - \frac{h^3}{6}u_{xxx}\}$$

$$- \{(\psi_j,\psi_j) - \epsilon k(\psi'_j,\psi'_j) - \delta k(\psi'_j,\psi_j)\}u$$

$$- \{(\psi_{j+1},\psi_j) - \epsilon k(\psi'_{j+1},\psi'_j) - \delta k(\psi'_{j+1},\psi_j)\}\{u + hu_x + \frac{h^2}{2}u_{xx}+ \frac{h^3}{6}u_{xxx}\}]$$

$$+ 0(\frac{h^4}{k},\ h^3,\ h^2k,\ hk^2,\ k^3)$$

which by using the relations for the (T,T) rational basis functions, can be simplified to

$$T_{j,n} = \frac{1}{hk}\Big[hku_t + \frac{hk^2}{2}u_{tt} + \frac{hk^3}{6}u_{ttt}$$

$$+ \{(\psi_{j-1},\psi_j) + (\psi_{j+1},\psi_j)\}\{\frac{h^2}{2}u_{xx} + \frac{h^2k}{2}u_{txx}\}$$

$$- \{- \delta k(\psi'_{j+1},\psi_j) + \delta k(\psi'_{j-1},\psi_j)\}\{hu_x + \frac{h^3}{6}u_{xxx}\}$$

$$- \frac{h^2}{2}u_{xx}\{(\psi_{j-1},\psi_j) + (\psi_{j+1},\psi_j)$$

$$- \epsilon k(\psi'_{j+1},\psi'_j) - \epsilon k(\psi'_{j-1},\psi'_j) - \delta k(\psi'_{j-1},\psi_j) - \delta k(\psi'_{j+1},\psi_j)\}\Big]$$

$$+ 0(\frac{h^4}{k},\ h^3,h^2k,hk^2,k^3)$$

$$= u_t + \frac{k}{2}u_{tt} + \frac{k^2}{6}u_{ttt} + \{(\psi_{j-1},\psi_j) + (\psi_{j+1},\psi_j)\}\frac{h}{2}u_{txx}$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$+ \frac{h^2}{2}u_{xx}\epsilon\, k\{(\psi'_{j+1},\psi'_j) + (\psi'_{j-1},\psi'_j)\}\frac{1}{hk}$$

$$- (hu_x + \frac{h^3}{6}u_{xxx})\frac{2\delta}{h}(\psi'_{j-1},\psi_j) + 0(\frac{h^4}{k}, h^3, h^2k, hk^2, k^3)$$

$$= u_t + \epsilon h(\psi'_{j-1},\psi'_j)u_{xx} - 2\delta u_x(\psi'_{j-1},\psi_j) + 0(h^2,k). \qquad (3.12)$$

Since the (1,1) rational basis function satisfies

$$(\psi'_{j-1},\psi'_j) = -\frac{7}{6h}$$

and

$$(\psi'_{j-1},\psi_j) = -\frac{1}{2},$$

it follows from (3.12) that

$$\lim_{h,k\to 0} T_{j,n} = u_t - \frac{7}{6}\epsilon u_{xx} + \delta u_x.$$

The equation differs from the original equation (3.1) due to the added diffusion coefficient $\frac{\epsilon}{6}$.

The (2,2) and (3,3) rational basis functions yield

$$\lim_{h,k\to 0} T_{j,n} = u_t - 1{,}1046\epsilon u_{xx} + \delta u_x$$

and

$$\lim_{h,k\to 0} T_{j,n} = u_t - 1{,}0764\epsilon u_{xx} + \delta u_x$$

respectively.

The added diffusion coefficient in (3.12) is $-\epsilon\{1 + h(\psi'_{j-1},\psi'_j)\}$.

According to theorem (2.8.3), however, $h(\psi'_{j-1},\psi'_j) \to -1$ as $T \to \infty$. This means that the added diffusion coefficient tends to zero with increasing T, and the discrete equation almost coincides with equation (3.1). Therefore, by increasing the order of the rational basis

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

function, the consistency property of the discrete equation is improved. It is important to note that the added diffusion coefficient is independent of h and k and is negligible for small values of ϵ. From this argument it is evident that the numerical scheme is naturally dissipative and will tend to damp numerical oscillations arising from the convective term.

3.5 STABILITY

The stability of a numerical scheme is influenced by the algorithm, the choice of basis functions and the integration by time. Consider the forward-difference scheme (3.10)

$$U^{n+1} = (I - kA^{-1}B)U^n. \tag{3.13}$$

A numerical scheme (3.13) is stable if a small perturbation, $\bar{\epsilon}_0$, of the initial data remains bounded as t increases. The initial error vector, $\bar{\epsilon}_0$, satisfies the recurrence relation

$$\bar{\epsilon}_{n+1} = (I - kA^{-1}B)\bar{\epsilon}_n$$

and

$$\bar{\epsilon}_{n+1} = (I - kA^{-1}B)^{n+1}\bar{\epsilon}_0. \tag{3.14}$$

The von Neumann or Fourier stability analysis [36] studies the evolution of individual components in an eigenfunction expansion of the initial vector.

For this purpose, let $\{\lambda_i\}$ and $\{\bar{v}_i\}$ be the eigenvalues and eigenvectors respectively of the eigenvalue problem

$$B\bar{v} = \lambda A\bar{v}.$$

The eigenvalues, $\{\lambda_i\}$, may be complex, since the convection term yields a skew-symmetrical matrix B [7]. Expand the initial error vector, $\bar{\epsilon}_0$, in eigenvectors, that is

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$\bar{\epsilon}_0 = \sum_{i=1}^{N} \alpha_i \bar{v}_i \qquad (3.15)$$

where α_i are constants.

Using (3.15) in (3.14),

$$\begin{aligned}\bar{\epsilon}_{n+1} &= (I - kA^{-1}B)^n \left(\sum_{i=1}^{N} \alpha_i \bar{v}_i - kA^{-1}B \sum_{i=1}^{N} \alpha_i \bar{v}_i \right) \\ &= (I - kA^{-1}B)^n \sum_{i=1}^{N} (1 - \lambda_i k) \alpha_i \bar{v}_i \\ &= \sum_{i=1}^{N} (1 - \lambda_i k)^{n+1} \alpha_i \bar{v}_i .\end{aligned}$$

For each eigenvector, $\bar{v}_i$, the term $(1 - \lambda_i k)^{n+1}$ is the amplification factor of the component α_i in the initial error.

Therefore, no component of the error will grow if

$$|1 - \lambda_i k| < 1 \qquad (3.16)$$

i.e. the numerical scheme is stable.

With only diffusion present, the eigenvalues are real and positive and (3.16) yields the stability condition

$$k \leq 2/\lambda_{max}.$$

If convection is present, the eigenvalues may be complex. Then

$$\begin{aligned}|1 - \lambda_i k|^2 &= (1 - \mathrm{Re}(\lambda_i)k)^2 + (\mathrm{Im}(\lambda_i)k)^2 \\ &< 1\end{aligned}$$

gives the stability constraint

$$k < \frac{2\,\mathrm{Re}(\lambda_i)}{(\mathrm{Re}(\lambda_i))^2 + (\mathrm{Im}(\lambda_i))^2} .$$

Implicit backward integration schemes are unconditionally stable and there is no restriction on the time step necessary for stability.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The stability of the general forward-difference equation (3.11) is examined by the standard Fourier stability analysis [1,7,36]. This is in fact the same as above matrix eigenvalue analysis except that the boundary conditions are not taken into consideration. The substitution of

$$U_j^n = \xi^n e^{\hat{i}j\gamma h}$$

(where $\hat{i}^2 = -1$ and $\gamma \geq 0$) into (3.11) yields

$$\begin{aligned}\xi = [\{&(\psi_{j-1},\psi_j) - \delta k(\psi'_{j-1},\psi_j) - \epsilon k(\psi'_{j-1},\psi'_j) \\ &+ (\psi_{j+1},\psi_j) - \delta k\,(\psi'_{j+1},\psi_j) - \epsilon k(\psi'_{j+1},\psi'_j)\}\cos\gamma h \\ &+ \hat{i}\{(\psi_{j+1},\psi_j) - (\psi_{j-1},\psi_j) - \delta k(\psi'_{j+1},\psi_j) + \delta k(\psi'_{j-1},\psi_j) \\ &- \epsilon k(\psi'_{j+1},\psi'_j) + \epsilon k(\psi'_{j-1},\psi'_j)\}\sin\gamma h \\ &+ \{(\psi_j,\psi_j) - \delta k(\psi'_j,\psi_j) - \epsilon k\,(\psi'_j,\psi'_j)\}]/ \\ &[(\psi_j,\psi_j) + \{(\psi_{j-1},\psi_j) + (\psi_{j+1},\psi_j)\}\cos\gamma h].\end{aligned}$$

Using the relations for the (T,T) rational basis functions, the above equation reduces to

$$\begin{aligned}\xi = [&h - \{2(\psi_{j-1},\psi_j) + \epsilon k(\psi'_j,\psi'_j)\}(1 - \cos\gamma h) \\ &- 2\hat{i}\delta k(\psi'_{j+1},\psi_j)\sin\gamma h]/[h - 2(\psi_{j-1},\psi_j)(1 - \cos\gamma h)].\end{aligned}$$

A necessary condition for stability is that the amplification factor ξ satisfies $|\xi|^2 < 1$. This condition is equivalent to

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$k \leq \frac{2\epsilon(\psi'_j,\psi'_j)[h - 2(\psi_{j-1},\psi_j)(1 - \cos \gamma h)]}{\{\epsilon(\psi'_j,\psi'_j)\}^2(1 - \cos \gamma h) + \delta^2(1 + \cos \gamma h)} .$$

Denote the right-hand side as a function of $\cos \gamma h$, that is $f(\cos \gamma h)$. Then,

$$f'(\cos \gamma h) = 2\epsilon(\psi'_j,\psi'_j)h[-\delta^2 + \epsilon^2(\psi'_j,\psi'_j)^2 + 4\delta^2(\psi_{j-1},\psi_j)/h] / [\epsilon^2(\psi'_j,\psi'_j)^2(1 - \cos \gamma h) + \delta^2(1 + \cos \gamma h)]^2,$$

where $f' = \frac{df}{d \cos \gamma h}$. Since (ψ'_j,ψ'_j) and (ψ_{j-1},ψ_j) are positive, it follows that the function f is monotone increasing over $[-1,1]$ if

$$-\delta^2 + \epsilon^2(\psi'_j,\psi'_j)^2 + 4\delta^2(\psi_{j-1},\psi_j)/h \geq 0 \qquad (3.17)$$

holds. A minimum value of the function is then attained at the left-hand side of the interval, i.e. $\cos \gamma h = -1$, and it yields the condition

$$k \leq \frac{h - 4(\psi_{j-1},\psi_j)}{\epsilon(\psi'_j,\psi'_j)}. \qquad (3.18)$$

If, however,

$$-\delta^2 + \epsilon^2(\psi'_j,\psi'_j)^2 + 4\delta^2(\psi_{j-1},\psi_j)/h < 0$$

the function f is montone decreasing over $[-1,1]$ and a minimum value is attained at $\cos \gamma h = 1$, which yields the condition

$$k \leq \frac{\epsilon(\psi'_j,\psi'_j)h}{\delta^2} .$$

This last condition is less strict than (3.18), because from (3.17)

$$h - 4(\psi_{j-1},\psi_j) \leq \epsilon^2(\psi'_j,\psi'_j)^2 h/\delta^2$$

and by substituting this into (3.18) it yields

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$k \leq \frac{h - 4(\psi_{j-1}, \psi_j)}{\epsilon(\psi'_j, \psi'_j)}$$

$$\leq \frac{\epsilon^2 (\psi'_j, \psi'_j)^2 h}{\delta^2 \epsilon(\psi'_j, \psi'_j)}$$

$$= \frac{\epsilon(\psi'_j, \psi'_j) h}{\delta^2}.$$

Thus, (3.18) is the stability condition for the explicit scheme.

In particular the (1,1), (2,2) and (3,3) rational basis functions give

$$k \leq 0.156\ h^2/\epsilon\ ,$$

$$k \leq 0.168\ h^2/\epsilon$$

and

$$k \leq 0.174\ h^2/\epsilon$$

respectively. From conditions (vi) and (vii) of theorem (2.8.1) and theorem (2.8.3) it can be seen that

$$h(\psi'_j, \psi'_j) \rightarrow 2$$

as $T \rightarrow \infty$. Therefore, it is clear that the stability condition weakens with the increasing order of the rational basis functions.

3.6 CONVERGENCE

From a previous analysis in Section 3.4 it was seen that the discrete equation is not consistent with the original partial differential equation due to an added diffusion coefficient. This coefficient, however, depends only on the order of the rational basis function and is therefore not influenced by any choice of the mesh parameters. In addition, it becomes negligible with increasing T as shown by the

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

consistency analysis with (1,1), (2,2) and (3,3) rational basis functions.

If the time step is chosen according to criterion (3.18), the numerical scheme is stable and it converges to a solution of the convection-diffusion equation with a slightly different diffusion coefficient, ϵ. With the increasing order of rational basis functions, however, this added diffusion coefficient becomes negligible so that in the limit the Lax equivalence theorem assures the convergence of the numerical scheme to the actual solution of the convection-diffusion equation.

3.7 OSCILLATIONS DUE TO PHASE ERRORS

The propagation of oscillatory numerical errors inhibits the solution of a discrete equation. The eigenvalues λ_i and eigenvectors of $A^{-1}B$ may be complex due to the domination of the convection term [7]. The imaginary parts of the eigenvectors generate the oscillatory ripples in the numerical solution, while the real part determines the amplitude of the solution. This implies that each component in an eigenvector expansion of the initial vector has an associated amplitude and phase and these may differ for different components. The inability to propagate individual frequencies or wave numbers without phase and amplitude errors leads to oscillations.

Consider the convection-diffusion equation

$$u_t + \delta u_x - \epsilon u_{xx} = 0$$

with the initial condition $u(x,0) = u_0(x)$. Assume the initial data consists of the single component

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$u_0(x) = \gamma_0 \exp(\hat{i}\sigma_\ell x)$$

where σ_ℓ is the wave number of component ℓ and $\hat{i}^2 = -1$. The analytical solution to the above is

$$u(x,t) = \gamma_0 e^{-\epsilon\sigma_\ell^2 t}\, e^{\hat{i}\sigma_\ell(x-\delta t)}.$$

Thus, the solution may be viewed as a wave propagated along the constant lines $x - \delta t$, but with exponential decay of the amplitude in time as a result of diffusion. The wave number σ_ℓ also plays a role in the amplitude decay and translation of the initial disturbance. To compare the amplitude decay and phase change at a time $\tau = mk$ for the analytical solution and the forward-difference scheme (3.11), substitute

$$U_j^n = \gamma_n \exp(\hat{i}\sigma_\ell jh), \quad \text{amplitude } \gamma_n$$

into (3.11) to obtain after simplification

$$\begin{aligned}\frac{\gamma_{n+1}}{\gamma_n} &= A_\ell \\ &= [h - \{2(\psi_{j-1},\psi_j) + \epsilon k(\psi'_j,\psi'_j)\}(1 - \cos\sigma_\ell h) \\ &\quad - 2\hat{i}\delta k(\psi'_{j+1},\psi_j)\sin\sigma_\ell h]/ \\ &\quad [h - 2(\psi_{j-1},\psi_j)(1 - \cos\sigma_\ell h)] \qquad (3.19)\end{aligned}$$

or

$$\gamma_{n+1} = (A_\ell)^{n+1}\gamma_0.$$

The amplification factor, A_ℓ, ensures the stability of the forward-difference scheme if $|A_\ell| < 1$. This has been done in Section 3.5. After m time steps the solution is

$$U_j^m = (A_\ell)^m \gamma_0 \exp(\hat{i}\sigma_\ell jh)$$

which is to be compared with the analytical solution. The substitution of $A_\ell = \rho_\ell e^{\hat{i}\theta_\ell}$ into above yields

$$U_j^m = \gamma_0 \rho_\ell^m \exp(\hat{i}(m\theta_\ell + \sigma_\ell jh))$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

and the phase shift for component ℓ after m time steps is $m\theta_\ell$. Then, from (3.19) and using the relations for (T,T) rational basis functions,

$$\tan\,\theta_\ell = \frac{-2\delta k(\psi'_{j+1},\psi_j)\sin\,\sigma_\ell h}{h + \{2(\psi_{j-1},\psi_j) + \epsilon k(\psi'_j,\psi'_j)\}(\cos\,\sigma_\ell h - 1)}\ .$$

But, $k \le \dfrac{h - 4(\psi_{j-1},\psi_j)}{\epsilon(\psi'_j,\psi'_j)}$, then choosing an optimal time step, $k = \dfrac{h - 4(\psi_{j-1},\psi_j)}{\epsilon(\psi'_j,\psi'_j)}$, and using the (T,T) relations results in

$$\theta_\ell = \tan^{-1}[\delta k \sin\,\sigma_\ell h\ /[h + (\psi_j,\psi_j)(\cos\,\sigma_\ell h - 1)]].$$

From the numerical calculation of (ψ_j,ψ_j) in Section 2.8 it can be seen that (ψ_j,ψ_j) increases with increasing order of the rational basis functions. Since $(\cos\,\sigma_\ell h - 1)$ is negative, the argument of $\tan^{-1}$ increases with increasing order. Because, $\tan^{-1}$ is an increasing function on $(-1,1)$ it follows that the phase error also increases with higher-order rational basis functions for the forward-difference scheme.

3.8 COMPARISON WITH LINEAR SHAPED BASIS FUNCTIONS AND UPWINDING

In finite-difference analysis of convection-dominated problems, backward or upwind differencing of the convection term has been used to damp numerical oscillations. A Taylor series expansion of this backward-difference approximation gives a leading term $0(h)$ in the truncation error [7]. This term can be interpreted as an added numerical or artificial diffusion. However, in finite element methods upwinding can be interpreted as weighted-residual methods in which the

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

test functions differ from the trial functions in the sense that they are biased in the upwind direction.

If in the convection-diffusion problem $\delta > 0$, then the flow is from left to right. Thus, the upwind direction is to the left and test functions $\hat{\psi}_j$ can be constructed that are weighted in this direction. A simple approach is to add and subtract a quadratic function on the left and right sides respectively of the linear hat basis function ψ_j in Figure 3.4.

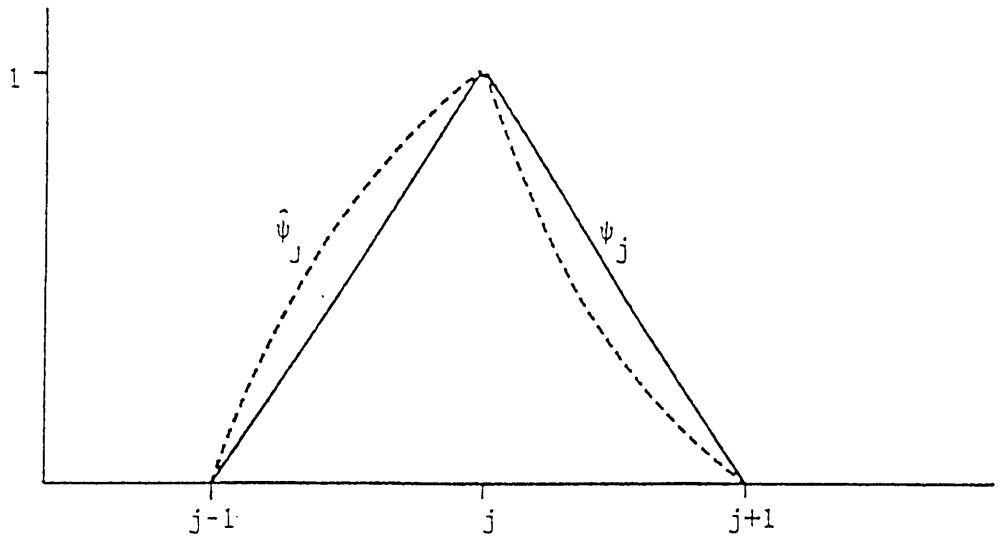

FIGURE 3.4 : Quadratically biased test functions $\hat{\psi}_j$

This is done by introducing a scaling factor ω that specifies the amount of upwind desired:

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$\hat{\psi}_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{h} + \dfrac{\omega(x - x_{j-1})(x_j - x)}{h^2}, & x_{j-1} \le x \le x_j \\ \dfrac{x_{j+1} - x}{h} - \dfrac{\omega(x - x_j)(x_{j+1} - x)}{h^2}, & x_j \le x \le x_{j+1} \\ 0, & x \le x_{j-1} \text{ and } x \ge x_{j+1}. \end{cases}$$

Using $\{\hat{\psi}_j(x)\}$, $j=1,\ldots,N$, as test functions and $\{\psi_j(x)\}$, $j=1,2,\ldots,N$, as trial functions in a variational formulation, yields [7]

$$\frac{h}{6}(\dot{U}_{j-1} + 4\dot{U}_j + \dot{U}_{j+1}) + \frac{\delta}{2}(U_{j+1} - U_{j-1}) - \frac{\epsilon}{h}(U_{j+1} - 2U_j + U_{j-1}) + \omega[\frac{h}{12}(-\dot{U}_{j-1} + \dot{U}_{j+1}) - \frac{\delta}{6}(U_{j+1} - 2U_j + U_{j-1})] = 0.$$

By lumping the time-dependent terms and using a Taylor series to expand U_{j+1} and U_{j-1} the expression

$$u_t + \delta u_x - (\epsilon + \frac{\delta\omega h}{6})u_{xx} + 0(h^2) = 0$$

is obtained at point $x = x_j$. Note that as ω increases, the numerical diffusion increases proportionally and will tend to damp numerical oscillations arising from the convection term. The term $\frac{\delta\omega h}{6}$ represents a numerical or artificial diffusion that is $0(h)$.

Discretising the lumped system forward in time yields

$$U_j^{n+1} = U_j^n + k\{(\frac{\epsilon}{h^2} + \frac{\delta\omega}{6h} + \frac{\delta}{2h})U_{j-1}^n - (\frac{2\epsilon}{h^2} + \frac{\omega\delta}{3h})U_j^n + (\frac{\epsilon}{h^2} + \frac{\delta\omega}{6h} - \frac{\delta}{2h})U_{j+1}^n\} .$$

The eigenvalues are

$$\lambda_i = (\frac{2\epsilon}{h^2} + \frac{\omega\delta}{3h}) - 2[(\frac{\epsilon}{h^2} + \frac{\omega\delta}{6h})^2 - (\frac{\delta}{2h})^2]^{1/2}\cos\frac{i\pi}{N+1}, \quad i=1,2,\ldots,N. \qquad (3.20) .$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Hence, the upwinding ω reduces the imaginary components and a larger convective term is now needed to dominate diffusion, thereby reducing oscillations. Furthermore, the real part becomes larger, thus increasing the dissipative decay.

Now, a lumped forward scheme with (T,T) rational basis functions is

$$U_j^{n+1} = U_j^n + \frac{k}{h}\{(-\epsilon(\psi'_{j-1},\psi'_j) + \frac{\delta}{2})U_{j-1}^n - \epsilon(\psi'_j,\psi'_j)U_j^n + (-\epsilon(\psi'_{j-1},\psi'_j) - \frac{\delta}{2})U_{j+1}^n\} \tag{3.21}$$

with eigenvalues

$$\lambda_i = \frac{\epsilon(\psi'_j,\psi'_j)}{h} - 2\left[\left\{\frac{\psi'_{j-1},\psi'_j)\epsilon}{h}\right\}^2 - (\frac{\delta}{2h})^2\right]^{1/2} \cos\frac{i\pi}{N+1}\ ,\ i=1,\ldots,N. \tag{3.22}$$

Since,

$$\lim_{T\to\infty} - h(\psi'_{j-1},\psi'_j) = 1$$

and

$$\lim_{T\to\infty} h(\psi'_j,\psi'_j) = 2$$

(3.22) coincides with (3.20) if upwinding is neglected, i.e. $\omega = 0$.

The eigenvalues of the lumped scheme of the (1,1) rational approximation and the linear hat-shaped approximation yield

$$\lambda_i^R = \frac{7\epsilon}{3h^2} - 2\left[(\frac{7}{6}\frac{\epsilon}{h^2})^2 - (\frac{\delta}{2h})^2\right]^{1/2} \cos\frac{i\pi}{N+1}$$

and

$$\lambda_i^L = \frac{2\epsilon}{h^2} - 2\left[(\frac{\epsilon}{h^2})^2 - (\frac{\delta}{2h})^2\right]^{1/2} \cos\frac{i\pi}{N+1}$$

respectively.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

If oscillations are present, that is when the cell Peclet number satisfies $\frac{\delta h}{\epsilon} > 2$, then the linear eigenvalues have real and imaginary components

$$\lambda_i^L = \frac{2\epsilon}{h^2} - \hat{i}2\left[\left(\frac{\delta}{2h}\right)^2 - \left(\frac{\epsilon}{h^2}\right)^2\right]^{1/2} \cos\frac{i\pi}{N+1}$$

where $\hat{i}^2 = -1$.

Similarly, if $\frac{\delta}{2} > \frac{7}{6}\frac{\epsilon}{h}$, then

$$\lambda_i^R = \frac{7\epsilon}{3h^2} - \hat{i}2\left[\left(\frac{\delta}{2h}\right)^2 - \left(\frac{7}{6}\frac{\epsilon}{h^2}\right)^2\right]^{1/2} \cos\frac{i\pi}{N+1}$$

and it is clear that the imaginary part of the eigenvalue is smaller than that of the linear eigenvalue, while the real part of it is bigger than that of the linear eigenvalue.

This suggests that the rational basis functions are usually better than linear hat-shaped functions, and a consistency analysis has shown that they naturally introduce numerical dissipation that prohibits numerical oscillations.

It can easily be shown that for the lumped difference schemes the following stability criteria prevail:

Rational; if $h \leq \frac{7}{3}\frac{\epsilon}{\delta}$ then $k < \frac{3h^2}{7\epsilon}$

and if $h > \frac{7}{3}\frac{\epsilon}{\delta}$ then $k < \frac{14}{3}\frac{\epsilon}{\delta^2}$

Linear; if $h \leq \frac{2\epsilon}{\delta}$ then $k < \frac{h^2}{2\epsilon}$

and if $h > \frac{2\epsilon}{\delta}$ then $k < \frac{4\epsilon}{\delta^2}$.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

It is clear that the rational approximation enables one to use larger time steps k and space steps h than its linear counterpart. Thus the same results could be obtained by using a standard piecewise linear scheme and, merely change ϵ to $7\epsilon/6$.

In conclusion, the shape of the rational basis functions is biased in the upstream direction, introducing natural numerical dissipation. Therefore, the eigenvalues of the discretised schemes have reduced imaginary components, thereby increasing dissipative decay and reducing oscillations of the solution.

3.9 NUMERICAL RESULTS

3.9.1 Homogeneous Dirichlet boundary conditions

In order to obtain a solution, the space interval $(0,1)$ was divided into 40 subintervals, i.e. $h = 0{,}025$, the time step $k = 0{,}01$, the diffusion coefficient $\epsilon = 0{,}01$ and the convection coefficient $\delta = 1$ were used. For the initial value u_0 an impulse of unity height and centralised at $x = 0{,}3$ was considered, namely

$$u_0(x) = \begin{cases} x/h_1 - 2, & 2h_1 \leq x \leq 3h_1 \\ -x/h_1 + 4, & 3h_1 \leq x \leq 4h_1 \\ 0, & \text{elsewhere} \end{cases}$$

and $h_1 = 0{,}1$.

The analytical solution [48] is given by

$$u(x,t) = \sum_{n=1}^{\infty} e^{\nu x+\lambda t} b_n \sin n\pi x$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

with $\lambda = -\delta^2/4\epsilon - \epsilon n^2\pi^2$, $\nu = \delta/2\epsilon$ and

$$b_n = \frac{2e^{-2h\nu}}{h_1(\nu^2 + n^2\pi^2)^2}\{(\nu^2 - n^2\pi^2)\sin 2h_1 n\pi + 2\nu n\pi \cos 2h_1 n\pi\}$$

$$- \frac{4e^{-3h\nu}}{h_1(\nu^2 + n^2\pi^2)^2}\{(\nu^2 - n^2\pi^2)\sin 3h_1 n\pi + 2\nu n\pi \cos 3h_1 n\pi\}$$

$$+ \frac{2e^{-4h\nu}}{h_1(\nu^2 + n^2\pi^2)^2}\{(\nu^2 - n^2\pi^2)\sin 4h_1 n\pi + 2\nu n\pi \cos 4h_1 n\pi\}.$$

Figure 3.5 shows the numerical and analytical solution at different time levels for the backward-difference scheme (3.9) with (1,1) rational basis functions.

The results obtained compare favourably to those obtained by Mitchell and Griffiths [26], who used full upwinding with quadratic trial functions.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

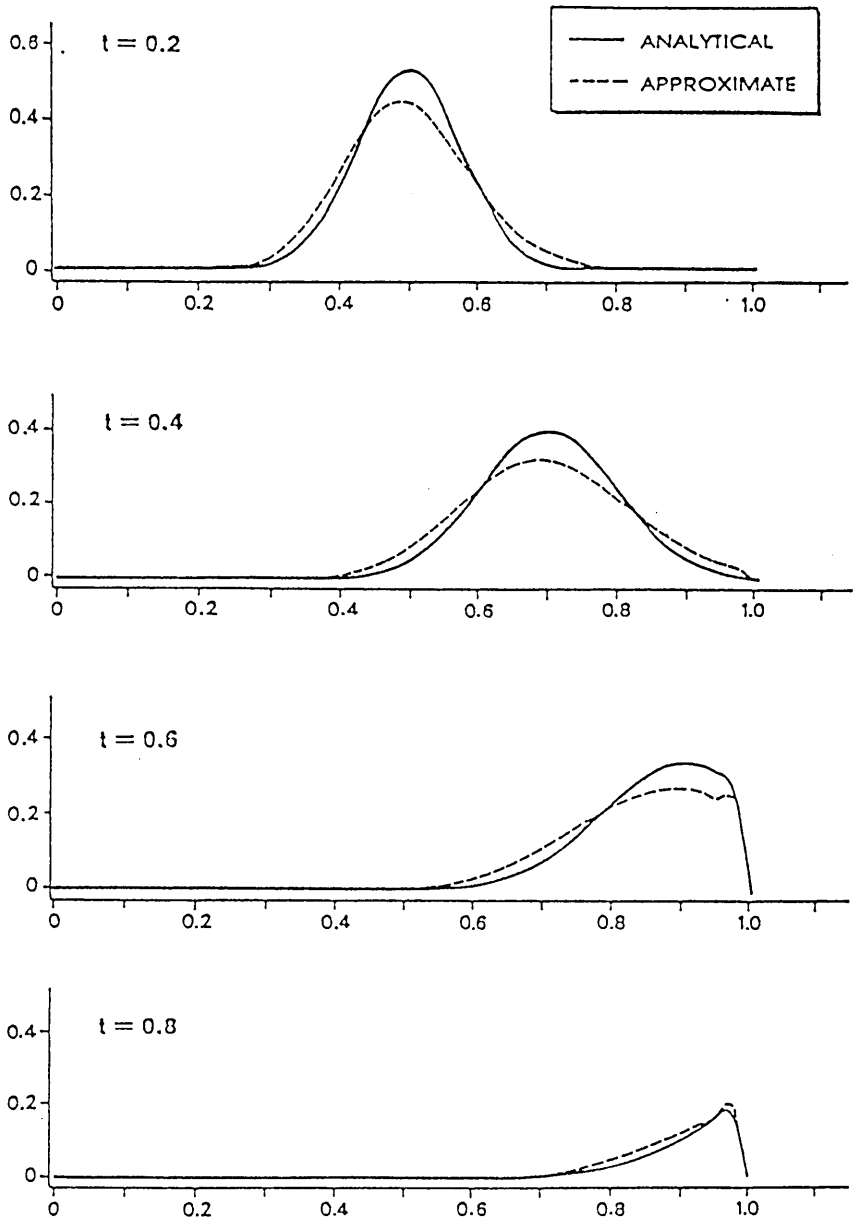

FIGURE 3.5 : Dirichlet boundary conditions solved with backward-difference scheme

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

A comparison of the forward- and backward-difference schemes is also presented. Both schemes have been solved with (1,1) rational basis functions for different values of h, namely h = 0,0333, 0,02, 0,0167 and 0,0125. The parameters are $\epsilon = 0{,}01$ and $\delta = 1$ with time step k = 0,001. The results are compared by means of the relative L_2 norm, i.e.

$$\|E\|_2 = \frac{\|V - U\|_2}{\|V\|_2}$$

where U is the approximant and V the analytical solution (3.23). The results at time T = 0,6 for the (0,1) rational basis function are presented in Table 3.1.

TABLE 3.1 : Relative L_2 norm for (1,1) rational basis functions

Difference scheme	h=0,033	h=0,025	h=0,020	h=0,0167	h=0,0125
Forward	6,93E-2	4,80E-2	4,23E-2	4,15E-2	4,25E-2
Backward	8,30E-2	7,21E-2	7,09E-2	7,17E-2	7,35E-2

The forward-difference scheme yields better results than the backward-difference scheme, where a skew-symmetrical matrix had to be inverted in comparison to the forward-difference scheme, where a symmetrical matrix had to be inverted. This is where the main difference exists between the two numerical schemes. Numerical results for the forward-difference scheme at time T = 0,6 for (1,1), (2,2) and (3,3) rational basis functions are shown in Table 3.2.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

TABLE 3.2: Relative L_2 norm for different rational basis functions using a forward-difference scheme

	h=0,033	h=0,025	h=0,020	h=0,0167	h=0,0125
(1,1)	6,93E-2	4,80E-2	4,23E-2	4,15E-2	4,25 E-2
(2,2)	7,80E-2	4,11E-2	2,80E-2	2,28E-2	2,07E-2
(3,3)	7,45E-2	4,22E-2	2,62E-2	1,80E-2	1,21E-2

From the table it is clear that higher-order methods improve the results, and that the accuracy increases with smaller h. In Figures 3.6, 3.7, and 3.8 the numerical and analytical solutions at T = 0,6 with h = 0,0125, k = 0,001 and different rational basis functions are compared. The graphs clearly demonstrate the superiority of the higher-order rational basis functions for the forward-difference schemes. The better performances of the higher- order basis functions correspond with the consistency analysis.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

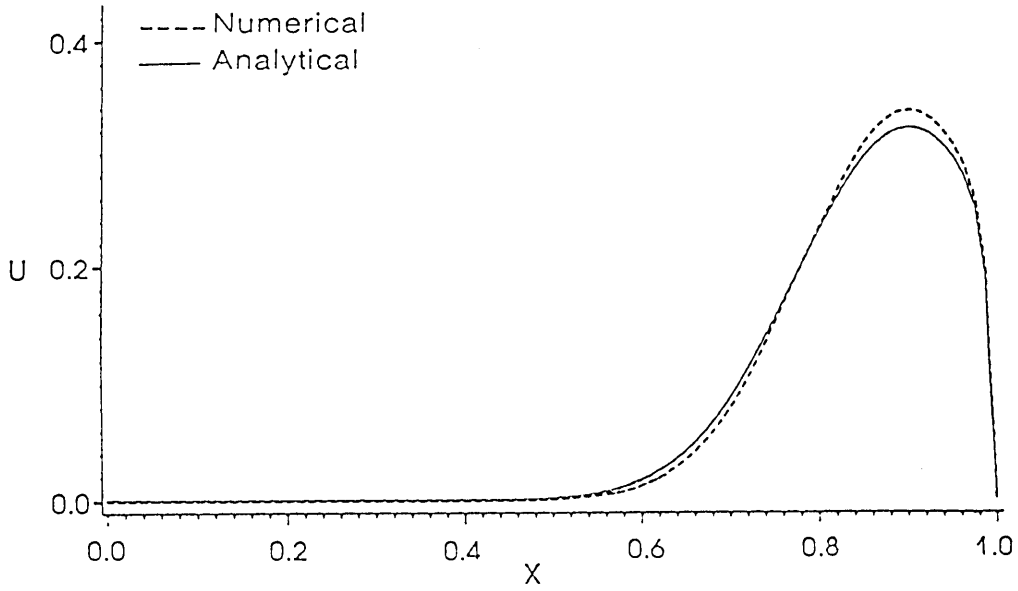

FIGURE 3.6 : (1,1) Rational basis function

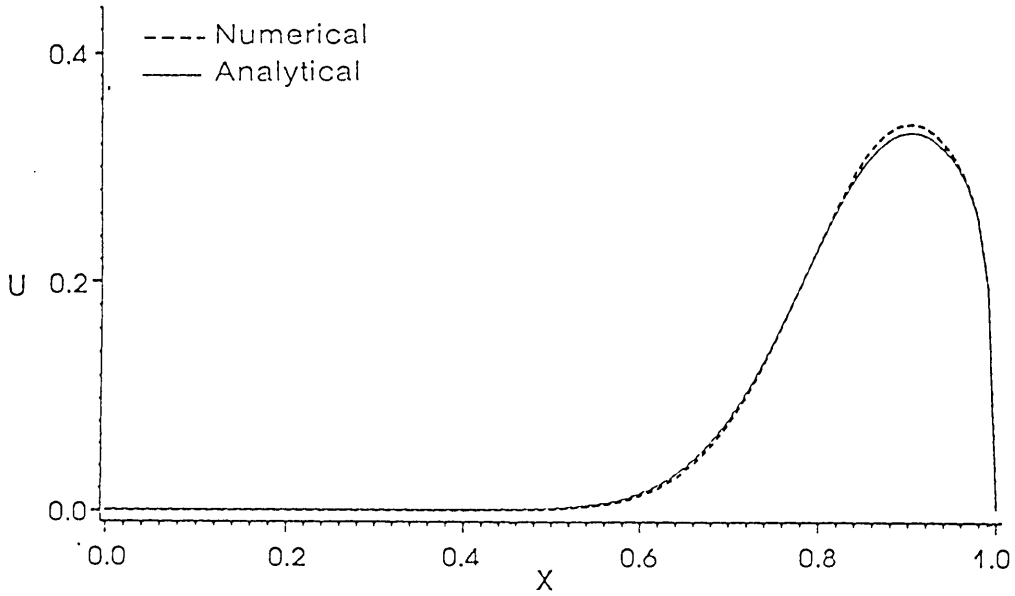

FIGURE 3.7 : (2,2) Rational basis function

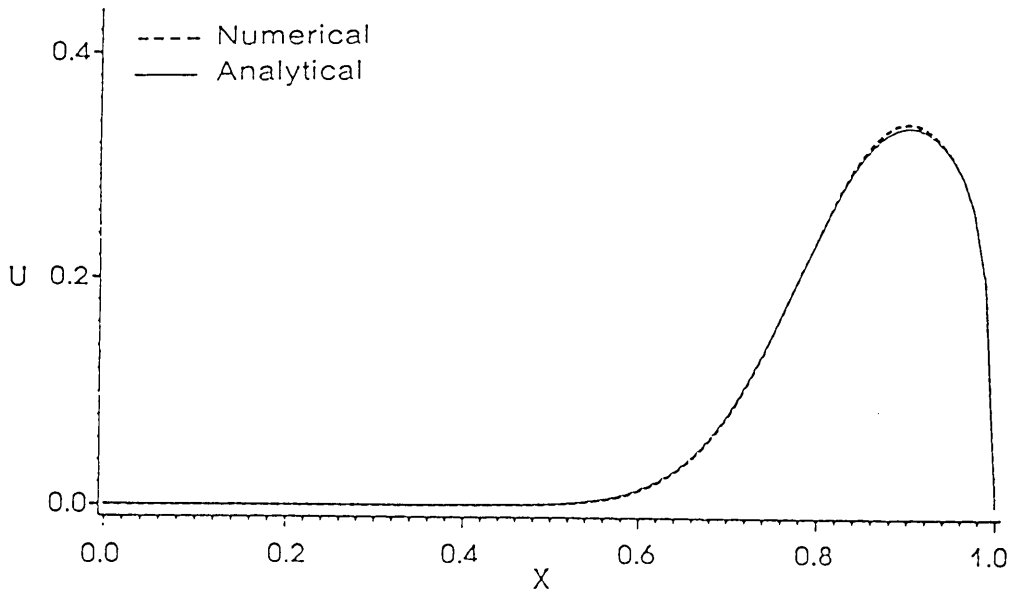

FIGURE 3.8 : (3,3) Rational basis function

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The forward-difference scheme is also implemented at different time steps to verify the stability condition. In Table 3.3 the discrete L_2 norm,

$$\|E\|_2^2 = h \sum_{i=1}^{N} (u_i - v_i)^2$$

at time $T = 1,0$ is tabulated for different rational basis functions. A dash in the table indicates that the stability condition is being violated.

TABLE 3.3 : L_2 norm for different space and time steps.

	(1,1) Rational basis function			
h/k	0,01	0,005	0,0025	0,001
0,0333	1,80E-3	0,91E-3	1,74E-3	2,37E-3
0,0200	-	0,29E-3	0,76E-3	1,40E-3
0,0167	-	-	0,65E-3	1,28E-3
0,0125	-	-	52,95E-3	1,18E-3
	(2,2) Rational basis function			
0,0333	2,31E-3	0,99E-3	1,40E-3	1,97E-3
0,0200	-	0,29E-3	0,31E-3	0,88E-3
0,0167	-	-	0,14E-3	0,74E-3
0,0125	-	-	0,05E-3	0,63E-3
	(3,3) Rational basis function			
0,0333	2,53E-3	1,13E-3	1,31E-3	1,81E-3
0,0200	-	1,09E-3	0,29E-3	0,67E-3
0,0167	-	1,15E-3	0,21E-3	0,57E-3
0,0125	-	-	0,24E-3	0,38E-3

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The numerical results correspond extremely well with the stability condition and also indicate that the higher-order rational basis functions have a less rigid restriction on the time step, which is in accordance with the theoretical analysis. The last column, $k = 0{,}001$, suggests that the error improves with $0(h^{1/2})$ when the order of the basis function increases. From this observation it is evident that higher-order basis functions improve the numerical convergence of the scheme.

3.9.2 Periodic boundary conditions

Again, for the initial value u_0 an impulse of unity height at $x = 0{,}3$ is considered. The parameters are $h = 0{,}025$, $k = 0{,}01$, $\epsilon = 0{,}01$ and $\delta = 1$. The exact solution of [48] is given by

$$u(x,t) = h + \sum_{n=1}^{\infty} \left[\frac{-e^{-(2n\pi)^2 \epsilon t}}{2(n\pi)^2 h}\right] \cdot \{\cos 2n\pi(x - \delta t - 2h) - 2\cos 2n\pi(x - \delta t - 3h) + \cos 2n\pi(x - \delta t - 4h)\}.$$

Using a backward-difference scheme with (1,1) rational basis functions, the numerical results are indicated by Figure 3.9.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

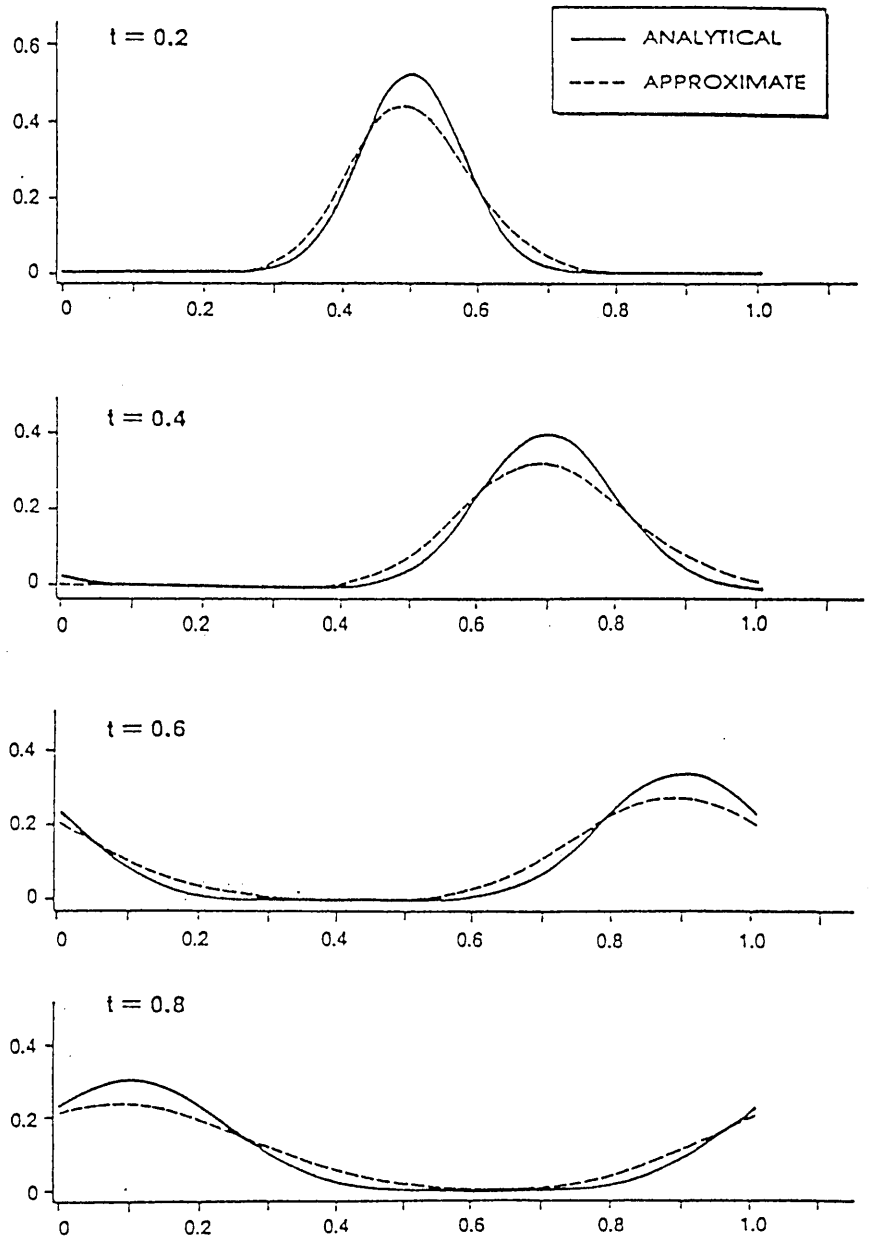

FIGURE 3.9 : Periodic boundary conditions

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

3.9.3 Neumann boundary conditions

The same parameters as in the previous section are used in a backward-difference scheme with (1,1) rational basis functions to obtain the numerical results indicated in Figure 3.10.

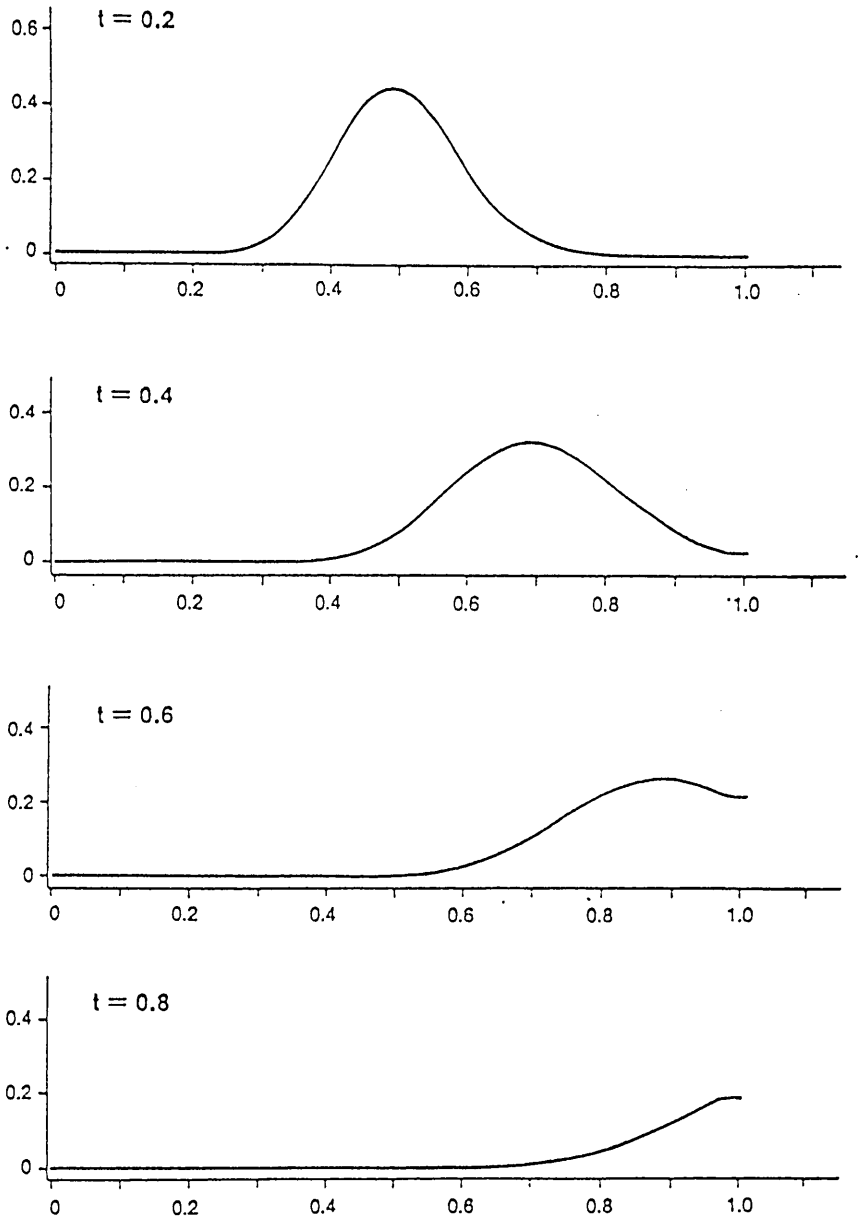

FIGURE 3.10 : Neumann boundary conditions

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

In this case, no analytical solution exists, but the solution seems to be fairly smooth.

3.10 CONCLUSIONS

The numerical scheme proposed in this chapter consists of a standard Galerkin method with identical upwinded test and trial functions. Since the rational basis function is biased in the upstream direction, the effect of convection-dominated flow is taken care of in a natural way. The numerical scheme is naturally dissipative and tends to damp numerical oscillations arising from the convective term.

In addition, the method is easily extended to include higher-order rational basis functions without coupling more than three node points. The advantage of this method is that the tridiagonal structure of the difference scheme stays intact while improved accuracy is achieved by the higher-order rational functions.

Finally, the numerical results validate the theoretical analysis that higher-order basis functions yield a numerical scheme with improved consistency properties and higher accuracies. The rational scheme is stable while convergence to the exact solution is assured by increasing the order of the rational basis functions. Moreover, the effect of upstream differencing is simulated in a natural way without adjusting any parameter to change the amount of upwinding as found in standard upwinded schemes [7,26,27,29].

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

CHAPTER 4

NONLINEAR DISSIPATIVE AND DISPERSIVE EQUATIONS

4.1 INTRODUCTION

In applications the numerical solution of nonlinear problems are most important. As examples the time-dependent Burgers equation and the Korteweg-de Vries (KdV) equation are studied using a rational approximation method.

Consider the time-dependent Burgers equation

$$Lu = u_t + uu_x - \epsilon u_{xx} = 0, \quad \epsilon > 0, \; x \in [0,1] \quad \text{and} \quad t \in [0,T],$$

where ϵ is the coefficient of kinematic viscosity. The equation arises in the studies of turbulence and shock wave theory. The mathematical properties of Burgers equation have been studied by Cole [10] and the equation provides a model where the transport of momentum of the fluid, $u_t + uu_x$, is dissipated by the viscous term ϵu_{xx}. For small values of ϵ the solution develops steep fronts and, depending upon the spatial mesh, numerical methods are unable to describe the sudden change in gradient, resulting in non-physical oscillations in the approximate solution. To suppress these unwanted non-physical disturbances various techniques have been developed for application in numerical methods. In finite-difference schemes the first derivative of the convective term is replaced by the standard backward (upwinded) and forward (downwinded) difference replacements, while in finite elements a Petrov-Galerkin method with different test and trial functions is introduced.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The Petrov-Galerkin approximation is the element $u \in \phi^{N+1}$ which satisfies

$$(Lu, \psi_j) = 0 \quad \text{for all} \quad \psi_j \in \psi^{N+1}.$$

The trial and test spaces ϕ^{N+1} and ψ^{N+1} respectively are different (N+1) dimensional subspaces of H_0^1. Typically, the trial functions ϕ_i are the linear hat functions and the test functions ψ_j are obtained by a linear combination of ϕ_i and a quadratic polynomial. For example Mitchell, et al [27], uses the following trial and test functions:

$$\phi_i(x) = \phi(x/h - i)$$

and $$\psi_i(x) = \psi(x/h - i), \quad i=0,1,\ldots,N,$$

where

$$\psi(s) = \phi(s) + \alpha\sigma(s),$$

$$\phi(s) = \begin{cases} 0 & |s| \geq 1 \\ 1 & |s| < 1 \end{cases},$$

$$\sigma(s) = \begin{cases} 0 & |s| > 1 \\ -3s(1 - s) & 0 \leq s \leq 1 \\ -\sigma(-s) & -1 \leq s \leq 0 \end{cases}$$

and α is an arbitrary parameter. This numerical scheme has a free parameter that has to be selected to vary the amount of upwinding. In fact, $\alpha = 1$ produces the upstream (backward difference) for the first-order space derivative. However, the rational basis functions developed have the same characteristic shape as the above test functions that are biased in the upstream direction. Thus, by using rational basis functions the upwinding parameter α is eliminated and

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

this justifies the application of rational basis functions for the Galerkin method.

The Korteweg-de Vries equation is a nonlinear wave equation which describes nonlinear dispersive wave phenomena. The equation will be considered in the form

$$Lu = u_t + uu_x + \epsilon u_{xxx} = 0,$$

where $u(x,t)$ is a real valued function for $-\infty < x < \infty$ and ϵ a real constant. The existence and uniqueness of a solution of the KdV equation for periodic initial data have been examined by Jeffrey and Kakutani [21]. The KdV equation on the real line has numerous conservation laws that have to be satisfied. The most important of these conservation laws is that any numerical scheme should aim to conserve the L_2 energy of the solution u, namely

$$\frac{dL_2}{dt} = \frac{d}{dt}\int_{-\infty}^{\infty} u^2 dx = 0.$$

In a Petrov-Galerkin method the third-order partial derivative demands a test function ψ_j which is at least C^1-continuous. Sanz-Serna et al [42] proposed a Petrov-Galerkin scheme,

$$(Lu,\psi_j) = 0 \quad \text{for all} \quad \psi_j \in \psi^{N+1}, \quad u \in \phi^{N+1}$$

where the trial functions ϕ_i are the linear hat functions and the test functions ψ_j are defined as a linear combination of the Hermite spline functions over an interval $[x_{j-2},x_{j+2}]$. In so doing a five-node replacement for the third-order partial derivative is obtained, assuring a better approximation. The unknown coefficients in the linear combination are chosen to satisfy consistency and to introduce upwinding in the polynomial scheme.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

To investigate the ability of a rational function to model a dispersive wave, continuous Hermite rational basis functions possessing continuous first derivatives at a node are constructed, as discussed in Section 2.7. These Hermite rational basis functions are used as test functions in conjunction with linear hat functions as trial functions in the Petrov-Galerkin scheme to solve the Korteweg-de Vries equation numerically.

In this chapter the rational difference schemes for both equations are developed. Some numerical results are also presented and comparisons are made with other numerical schemes to demonstrate the efficiency of the rational approximation method.

4.2 DISCRETISING THE BURGERS EQUATION

Consider the initial-boundary value problem

$$u_t + uu_x - \epsilon u_{xx} = 0, \quad \epsilon > 0, \quad x \in (0,1), \; t \in (0,T]$$

with

$$\begin{aligned} u(x,0) &= u_0(x) \;, \quad 0 < x < 1 \;, \\ u(0,t) &= q_0(t) \;, \; t \geq 0, \\ u(1,t) &= q_1(t) \;, \; t \geq 0. \end{aligned} \tag{4.1}$$

Divide the interval $[0,1]$ into N subintervals of length h and the interval $[0,T]$ in J subintervals of length k. Introduce rational basis functions

$$\psi_i(x), \quad i=1,\ldots,N-1$$

which are independent of time at the nodes. Galerkin's method consists of seeking an approximate solution to (4.1) in the form of

$$u(x,t) \simeq \sum_{i=1}^{N-1} U_i(t)\psi_i(x), \tag{4.2}$$

which satisfies the following system:

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$(u_t + (\frac{u^2}{2})_x - \epsilon u_{xx}, \psi_j) = 0 \quad , \; j=1,\ldots,N-1 \; , \tag{4.3}$$

where $(u,v) = \int_0^1 uv\, dx$.

Hence, the substitution of (4.2) into (4.3) and the application of product approximation [9] results in

$$\begin{aligned} &(\psi_{j-1},\psi_j)\dot{U}_{j-1} + (\psi_j,\psi_j)\dot{U}_j + (\psi_{j+1},\psi_j)\dot{U}_{j+1} \\ &- \tfrac{1}{2}\{(\psi_{j-1},\psi_j')U_{j-1}^2 + (\psi_j,\psi_j')U_j^2 + (\psi_{j+1},\psi_j')U_{j+1}^2\} \\ &+ \epsilon\{(\psi_{j-1}',\psi_j')U_{j-1} + (\psi_j',\psi_j')U_j + (\psi_{j+1}',\psi_j')U_{j+1}\} = 0, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j=1,\ldots,N-1. \end{aligned} \tag{4.4}$$

Discretise the approximating system (4.4) of first-order ordinary differential equations in time. For this purpose, let

$$U_i^n \simeq u(ih,nk)$$

and

$$T_\theta(U_i^n) = (1 - \theta)U_i^n + \theta U_i^{n+1},$$

where $\theta \in [0,1]$ is a parameter that determines the difference approximation in time. For example, $\theta = 0$ and 1 will correspond to forward and backward differencing respectively. The nodal values at a specific time level $(n+1)k$ are determined from the following scheme:

$$\begin{aligned} &(\psi_{j-1},\psi_j)(U_{j-1}^{n+1} - U_{j-1}^n) + (\psi_j,\psi_j)(U_j^{n+1} - U_j^n) + (\psi_{j+1},\psi_j)(U_{j+1}^{n+1} - U_{j+1}^n) \\ &- \tfrac{k}{2}\{(\psi_{j-1},\psi_j')T_\theta[(U_{j-1}^n)^2] + (\psi_j,\psi_j')T_\theta[(U_j^n)^2] + (\psi_{j+1},\psi_j')T_\theta[(U_{j+1}^n)^2]\} \\ &+ k\epsilon\{(\psi_{j-1}',\psi_j')T_\theta(U_{j-1}^n) + (\psi_j',\psi_j')T_\theta(U_j^n) + (\psi_{j+1}',\psi_j')T_\theta(U_{j+1}^n)\} = 0, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j=1,\ldots,N-1. \end{aligned}$$

In matrix notation this difference scheme may be written in the form

$$\begin{aligned} &A(U^{n+1} - U^n) + (1 - \theta)kB\overline{U^2}^n + (1 - \theta)kCU^n \\ &\qquad + \theta kB\overline{U^2}^{n+1} + \theta kCU^{n+1} = 0, \end{aligned} \tag{4.5}$$

where $\overline{U^2} = (U_1^2,\ldots,U_{N-1}^2)^T$.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

For the (1,1) rational basis functions the elements of the tridiagonal matrices are

$$
\begin{aligned}
a_{i,i-1} &= 6h\,\ell n\,2 - 4h\,, && i=2,\ldots,N-1,\\
a_{i,i+1} &= 6h\,\ell n\,2 - 4h\,, && i=1,\ldots,N-2,\\
a_{i,i} &= -a_{i,i-1} - a_{i,i+1} + h\,, && i=2,\ldots,N-2,\\
a_{1,1} &= a_{N-1,N-1} = a_{i,i}\,, && i=2,\ldots,N-2,\\
b_{i,i-1} &= -\frac{1}{4}\,, && i=2,\ldots,N-1,\\
b_{i,i+1} &= \frac{1}{4}\,, && i=1,\ldots,N-2,\\
b_{i,i} &= 0\,, && i=1,\ldots,N-1,\\
c_{i,i-1} &= -\frac{7\epsilon}{6h}\,, && i=2,\ldots,N-1,\\
c_{i,i+1} &= -\frac{7\epsilon}{6h}\,, && i=1,\ldots,N-2,\\
c_{i,i} &= -c_{i,i-1} - c_{i,i+1}\,, && i=2,\ldots,N-2,\\
c_{1,1} &= c_{N-1,N-1} = c_{i,i}\,, && i=2,\ldots,N-2.
\end{aligned}
$$

The implementation of higher-order rational basis functions leads to similar tridiagonal matrices. If $\theta = 0$ the difference scheme (4.5) reduces to

$$A\bar{U}^{n+1} = kA\bar{U}^{n} - kB\bar{U}^{2n} - C\bar{U}^{n}. \qquad (4.6)$$

Since A is strictly diagonally dominant, it is nonsingular [6], and an explicit forward-difference scheme is obtained.

However, if $\theta = 1$, an implicit backward-difference system of non-linear equations arises, which has to be solved iteratively by means of the Newton-Raphson method.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

4.3 CONVERGENCE, CONSISTENCY AND STABILITY FOR THE BURGERS EQUATION

The consistency analysis of the nonlinear scheme (4.5), with $\theta = 0$, by using u^2 instead of u in a Taylor series expansion at (jh,nk) yields

$$\begin{aligned} T_{j,n} &= u_y + \frac{k}{2} u_{tt} + \frac{k^2}{6} u_{tt} + (\psi_{j-1},\psi_j)hu_{txx} \\ &\quad + \frac{h}{2}\epsilon\{(\psi'_{j-1},\psi'_j) + (\psi'_{j+1},\psi'_j)\}u_{xx} + \frac{1}{2}(u^2)_x \\ &\quad + \frac{h^2}{12} u_{xx} + O(\frac{h^4}{k},h^3,h^2k,hk^2,k^3) \\ &= u_t + \frac{1}{2}(u^2)_x + \epsilon h(\psi'_{j-1},\psi')u_x + O(h^2,k). \end{aligned}$$

For the (1,1) rational basis function

$$(\psi'_{j-1},\psi'_j) = -\frac{7}{6h}$$

so that

$$\lim_{h,k\to 0} T_{j,n} = u_t + \frac{1}{2}(u^2)_x - \frac{7}{6}\epsilon u_{xx}.$$

This corresponds to the consistency analysis of the convection-diffusion equation.

The nonlinear difference scheme cannot be analysed, however, by means of the Lax equivalence theorem owing to the existence of the nonlinear convection term. By linearising the term uu_x to δu_x, however, where δ indicates the convection coefficient, the convection-diffusion equation is obtained with known consistency, stability and convergence properties as discussed in Chapter 3. Note that the behaviour of the linear equation need not be similar to that of the nonlinear equation.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

4.4 DISCRETISING THE KORTEWEG-DE VRIES EQUATION

Assume that the KdV equation

$$u_t + uu_x + \epsilon u_{xxx} = 0, \quad \epsilon > 0, \quad -\infty < x < \infty , \tag{4.7}$$

together with the initial condition

$$u(x,0) = u_0(x), \quad -\infty < x < \infty , \tag{4.8}$$

has a unique solution such that, for a fixed t, $u(x,t)$ and all its x derivatives tend to zero as $|x| \longrightarrow \infty$. Rewrite (4.7) in the conservation form

$$u_t + (\frac{1}{2} u^2)_x + \epsilon u_{xxx} = 0. \tag{4.9}$$

A subinterval $[a,b]$ is chosen on which the boundary conditions are approximately satisfied.

Divide the fixed interval $[a,b]$ into a uniform grid $a = x_0 < x_1 < \ldots < x_N = b$, with $h = x_j - x_{j-1}$, $j=1,\ldots,N$. Let $\psi_j \in C^1$ be a piecewise test function defined at each node x_j, $j=0,1,\ldots,N$.

The weak solution of (4.9) is given by

$$(u_t,\psi_j) + \frac{1}{2}((u^2)_x,\psi_j) + \epsilon(u_x,(\psi_j)_{xx}) = 0 \tag{4.10}$$

where $(f,g) = \int_a^b f(x)g(x)dx$.

Define by $\phi_i(x) = \phi((x - x_0)/h - i)$ the piecewise linear hat function associated with the node x_i, $i=0,1,\ldots,N$, where

$$\phi(\xi) = \begin{cases} \phi_1^0(\xi + 1) & , \ -1 \le \xi \le 0, \\ \phi_2^0(\xi) & , \ 0 \le \xi \le 1, \\ 0 & , \ \text{elsewhere} \end{cases}$$

with

$$\phi_1^0(\xi) = \xi$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

and

$$\phi_2^0(\xi) = 1 - \xi.$$

Since the trial function $\phi_i(x)$ has compact support, the support of the continuous differentiable test function $\psi_j(x)$ should be extended to obtain a better node replacement in the numerical scheme.

By using the same technique as Sanz-Serna, et al [42], define $\psi_j(x)$, $x \in [x_{j-2}, x_{j+2}]$ as

$$\psi_j(x) = \psi((x - x_0)/h - j), \quad j=0,1,\ldots,N, \tag{4.11}$$

where $\psi(\xi) \in C^1[-2,2]$. This function is defined as a linear combination of the Hermite rational basis functions, i.e.

$$\begin{aligned}\psi(\xi) = {} & \alpha_{-1}\psi^0(\xi + 1) + \alpha_0\psi^0(\xi) + \alpha_1\psi^0(\xi - 1) \\ & + \beta_{-1}\psi^1(\xi + 1) + \beta_0\psi^1(\xi) + \beta_1\psi^1(\xi - 1),\end{aligned}$$

where

$$\psi^0(\xi) = \begin{cases} \psi_1^0(\xi + 1) & , \ -1 \le \xi \le 0 , \\ \psi_2^0(\xi) & , \ 0 \le \xi \le 1 , \\ 0 & , \ \text{elsewhere} , \end{cases}$$

and

$$\psi^1(\xi) = \begin{cases} \psi_1^1(\xi + 1) & , \ -1 \le \xi \le 0 , \\ \psi_2^1(\xi) & , \ 0 \le \xi \le 1 , \\ 0 & , \ \text{elsewhere.} \end{cases}$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Approximate the exact solution $u(x,t)$ by

$$u_N(x,t) = \sum_{i=0}^{N} U_i(t)\phi_i(x), \tag{4.12}$$

where $U_i(t)$ denotes the approximant at node x_i, and use product approximation on the nonlinear term, i.e.

$$u_N^2(x,t) = \sum_{i=0}^{N} U_i^2(t)\phi_i(x). \tag{4.13}$$

The substitution of (4.11), (4.12) and (4.13) into (4.10) and the evaluation of the inner products yield

$$M\dot{U} + SU^2 + GU = 0 \tag{4.14}$$

where

$$U = [U_0(t),\ldots,U_N(t)]^T,$$

$$U^2 = [U_0^2(t),\ldots,U_N^2(t)]^T$$

and M, S and G are five-banded $(N+1) \times (N+1)$ matrices which are functions of the parameters α_i, β_i, $i=-1,0,1$.

The elements of the j-th row of the matrices $M = (m_{ij})$, $S = (s_{ij})$ and $G = (g_{ij})$ are

$$m_{j,j-2} = \alpha_{-1}(\phi_2^0,\psi_1^0) + \beta_{-1}(\phi_2^0,\psi_1^1),$$

$$m_{j,j-1} = \alpha_{-1}F_1 + \beta_{-1}F_2 + \alpha_0(\phi_2^0,\psi_1^0) + \beta_0(\phi_2^0,\psi_1^1),$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$m_{j,j} = \alpha_{-1}(\phi_1^0, \psi_2^0) + \beta_{-1}(\phi_1^0, \psi_2^1) + \alpha_0 F_1 + \beta_0 F_2$$

$$+ \alpha_1(\phi_2^0, \psi_1^0) + \beta_1(\phi_2^0, \psi_1^1) \ ,$$

$$m_{j,j+1} = \alpha_0(\phi_1^0, \psi_2^0) + \beta_0(\phi_1^0, \psi_2^1) + \alpha_1 F_1 + \beta_1 F_2 \ ,$$

$$m_{j,j+2} = \alpha_1(\phi_1^0, \psi_2^0) + \beta_1(\phi_1^0, \psi_2^1) \ ,$$

$$s_{j,j-2} = \frac{1}{2}[\alpha_{-1}(\frac{d\phi_2^0}{dx}, \psi_1^0) + \beta_{-1}(\frac{d\phi_2^0}{dx}, \psi_1^1)] \ ,$$

$$s_{j,j-1} = \frac{1}{2}[\alpha_{-1} G_1 + \beta_{-1} G_2 + \alpha_0(\frac{d\phi_2^0}{dx}, \psi_1^0) + \beta_0(\frac{d\phi_2^0}{dx}, \psi_1^1)] \ ,$$

$$s_{j,j} = \frac{1}{2}[-\alpha_{-1}(\frac{d\phi_2^0}{dx}, \psi_2^0) - \beta_{-1}(\frac{d\phi_2^0}{dx}, \psi_2^1) + \alpha_0 G_1 + \beta_0 G_2$$

$$+ \alpha_1(\frac{d\phi_2^0}{dx}, \psi_1^0) + \beta_1(\frac{d\phi_2^0}{dx}, \psi_1^1)] \ ,$$

$$s_{j,j+1} = \frac{1}{2}[-\alpha_0(\frac{d\phi_2^0}{dx}, \psi_2^0) - \beta_0(\frac{d\phi_2^0}{dx}, \psi_2^1) + \alpha_1 G_1 + \beta_1 G_2] \ ,$$

$$s_{j,j+2} = \frac{1}{2}[-\alpha_1(\frac{d\phi_2^0}{dx}, \psi_2^0) - \beta_1(\frac{d\phi_2^0}{dx}, \psi_2^1)] \ ,$$

$$g_{j,j-2} = -\frac{\epsilon}{h^2}\beta_{-1} \ ,$$

$$g_{j,j-1} = \frac{\epsilon}{h^2}(2\beta_{-1} - \beta_0) \ ,$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$g_{j,j} = \frac{\epsilon}{h^2}(2\beta_0 - \beta_1 - \beta_{-1}),$$

$$g_{j,j+1} = \frac{\epsilon}{h^2}(2\beta_1 - \beta_0),$$

$$g_{j,j+2} = -\frac{\epsilon}{h^2}\beta_1,$$

where

$$F_1 = (\phi_1^0, \psi_1^0) + (\phi_2^0, \psi_2^0),$$

$$F_2 = (\phi_1^0, \psi_1^1) + (\phi_2^0, \psi_2^1),$$

$$G_1 = -(\frac{d\phi_2^0}{dx}, \psi_1^0) + (\frac{d\phi_2^0}{dx}, \psi_2^0)$$

and

$$G_2 = -(\frac{d\phi_2^0}{dx}, \psi_1^1) + (\frac{d\phi_2^0}{dx}, \psi_2^1).$$

Note that $j=0,1,\ldots, N$ with $U_{-2} = U_{-1} = U_{N+1} = U_{N+2} = 0$.

The parameters α_i, β_i, $i=-1,0,1$, remain to be selected to give an approximation for (4.14) that is consistent with (4.7). Therefore, the application of a Taylor expansion to (4.14) eventually yields

$$\{h(\alpha_{-1} + \alpha_0 + \alpha_1) + (\beta_{-1} + \beta_0 + \beta_1)(1, \psi_1^1 + \psi_2^1)\}\dot{u}$$
$$+ \{h(\alpha_{-1} + \alpha_0 + \alpha_1) + (\beta_{-1} + \beta_0 + \beta_1)h(\frac{-d\phi_2^0}{dx}, \psi_1^1 + \psi_2^1)\}\frac{1}{2}(u^2)_x$$
$$- \epsilon(\beta_{-1} + \beta_0 + \beta_1)u_{xx} + \epsilon h(\beta_{-1} - \beta_1)u_{xx} + O(h^2) = 0.$$

By enforcing

$$\alpha_{-1} + \alpha_0 + \alpha_1 = 1$$
$$\beta_{-1} + \beta_0 + \beta_1 = 0$$
$$\beta_{-1} - \beta_1 = 1.$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

a local truncation error of $0(h)$ is obtained.

However, by selecting

$$\alpha = \alpha_{-1} = \alpha_1$$

and

$$\beta_0 = 0$$

additional constraints are obtained which yield $\beta_{-1} = \frac{1}{2}$ and $\beta_1 = -\frac{1}{2}$. This leaves only α as a free parameter to introduce upwinding in the numerical scheme. Discretise in time the approximating system (4.14) of first-order ordinary differential equations. For this purpose, let

$$U_j^\ell \simeq u(jh, \ell k),$$

where k is the time step, i.e. $0 \leq \ell k \leq T$, $0 \leq \ell \leq J$ and

$$T_\theta(U_j^\ell) = (1 - \theta)U_j^\ell + \theta U_j^{\ell+1}$$

where $\theta \in [0,1]$.

For example, $\theta = 0$ and $\frac{1}{2}$ will correspond to the forward differencing and Crank-Nicholson scheme respectively.

In matrix notation the difference scheme may be written in the form

$$M[U^{\ell+1} - U^\ell]/k + ST_\theta((U^2)^\ell) + GT_\theta(U^\ell) = \bar{0}. \qquad (4.15)$$

If $\theta = \frac{1}{2}$ an implicit scheme arises and the resulting system of non-linear equations has to be solved by Newton iteration. However, if $\theta = 1$, the inverse of the matrix M has to be computed and $U^{\ell+1}$ is computed explicitly in terms of U^ℓ.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

4.5 NUMERICAL RESULTS

4.5.1 The Burgers equation

Consider the Burgers equation

$$u_t + uu_x - \epsilon u_{xx} = 0, \quad x \in (0,1), \ t \in (0,T]$$

with initial and boundary values

$$u_0(x) = \sin \pi x , \qquad 0 < x < 1$$

$$q_0(t) = q_1(t) = 0, \quad t \geq 0$$

respectively. The theoretical solution to this problem is given by Cole [10], namely

$$u(x,t) = 2\epsilon\pi \frac{\sum_{n=1}^{\infty} \exp[-\epsilon n^2\pi^2 t/\ell^2] n A_n \sin(n\pi x)}{A_0 + \sum_{n=1}^{\infty} \exp[-\epsilon n^2\pi^2 t/\ell^2] A_n \cos(n\pi x)},$$

where

$$A_0 = \int_0^1 \exp\left[-\frac{1}{2\epsilon}\int_0^x f(\xi)d\xi\right]dx$$

and

$$A_n = 2\int_0^1 \exp\left[-\frac{1}{2\epsilon}\int_0^x f(\xi)d\xi\right]\cos(n\pi x)dx.$$

The solution develops a steep front near $x = 1$ which broadens and dies out with time.

The numerical results obtained from (4.5) are compared with the exact solution and the compact differencing technique of Hirsh, which is probably the most successful of all finite difference techniques for solving the Burgers equation [9,27]. The (1,1) rational scheme is solved implicitly with $\theta = 1$, $k = 0{,}01$ and $N = 19$, i.e. $h = 0{,}055$, while the compact differencing results have been obtained by means of

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

a time step $k = 0,001$ in a Crank-Nicholson scheme. The numerical results and absolute errors are shown in Table 4.1.

TABLE 4.1 :Sinc initial condition ($\epsilon = 0,01$, $t = 0,5$)

x	Exact	Compact differencing	Rational (1,1)	Error Compact	Error Rational
0,5	0,589	0,589	0,592	0,0	3,0E-3
0,56	0,649	0,648	0,651	1,0E-3	2,0E-3
0,62	0,707	0,709	0,708	2,0E-3	1,0E-3
0,67	0,762	0,760	0,762	2,0E-3	4,0E-4
0,72	0,814	0,820	0,812	6,0E-3	2,0E-3
0,78	0,861	0,852	0,856	9,0E-3	5,0E-3
0,83	0,902	0,917	0,894	1,5E-2	8,0E-3
0,89	0,934	0,911	0,917	2,3E-2	1,7E-2
0,94	0,937	0,964	0,936	2,7E-2	1,0E-3
1,0	0,0	0,0	0,0	-	-

From the results it is clear that the (1,1) rational approximation is performing excellently in comparison with the scheme of Hirsh.

Numerical results for this problem are also obtained by Christie, et al [9], using quadratic trial functions $\tilde{\psi}_j$ and quadratic test functions $\hat{\psi}_j$ with full upwinding in a standard Petrov-Galerkin technique where the nonlinear convective term is approximated by

$$(u\frac{\partial u}{\partial x}, \hat{\psi}_j) \simeq (\sum_{i=1}^{N-1} U_i\tilde{\psi}_i \sum_{i=1}^{N-1} U_i\tilde{\psi}'_i, \hat{\psi}_j).$$

Numerical results are also obtained for the above Petrov-Galerkin

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

scheme using product approximation on the nonlinear term, that is

$$(u\frac{\partial u}{\partial x}, \hat{\psi}_j) \simeq \frac{1}{2}(\sum_{i=1}^{N-1} U_i^2 \, \tilde{\psi}_i', \hat{\psi}_j).$$

The Galerkin method with (1,1) rational basis functions is compared to these Petrov-Galerkin schemes in the range where the solution changes its gradient using $h = 0{,}055$. The other parts of the domain are neglected in Table 4.2, since the solution is linear over these parts.

TABLE 4.2 :Comparison with Petrov-Galerkin techniques at at $t = 0{,}5$ $(\epsilon = 0{,}01)$

x	Exact	Standard Petrov-Galerkin	Product approximation	Rational (1,1)
0,83	0,902	0,895	0,907	0,894
0,89	0,934	0,911	0,952	0,917
0,94	0,937	0,764	0,774	0,936
1,0	0,0	0,0	0,0	0,0

Integration in time for the Petrov-Galerkin techniques in Table 4.2 is carried out using the Crank-Nicholson scheme with time step $k = 0{,}001$, while an implicit scheme with time step $k = 0{,}01$ is used for the (1,1) rational approximation. The results again show that the rational approximation is accurate in the vicinity of the wave front. This can only be attributed to the better approximation ability of a rational function when the solution suddenly changes.

Secondly, consider Burgers equation,

$$u_t + uu_x - \frac{1}{R} u_{xx} = 0, \quad x \in \mathbb{R}, \; t \in (0,T]$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

with initial condition

$$u(x,0) = \frac{2a}{R(1 + \exp(ax + b))} \quad , \quad x \in \mathbb{R}$$

where R is the Reynolds number and the constants a and b are chosen arbitrarily. The solution is given by the solitary wave

$$u(x,t) = \frac{2a}{R(1 + \exp(a(x - at/R) + b))} \quad .$$

Rosinger [37] solved this problem by means of a stable and convergent explicit nonlinear difference scheme. For purposes of comparison with his results, an explicit scheme, $\theta = 0$, with (1,1) rational basis functions is implemented with the following parameters: R = 10, a = 20, b = 1, h = 0,04, k = 0,001 and time T = 0,14. The results are shown in Table 4.3, where V and U denote the numerical and exact solutions while E denotes the relative error (U - V)/V. Despite the relative large h the (1,1) rational approximation performs satisfactorily and is obtained very easily by computation as a result of the explicit scheme. The maximum slope is attained at

$$x = \frac{aT}{R} - \frac{b}{a} = 0,23.$$

At this point the (1,1) rational approximation compares favourably with the exact solution as shown in Figure 4.1. At the extreme turning points the rational approximation tends to undershoot/overshoot the exact values.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

TABLE 4.3 : Solitary wave initial condition

		(1,1) Rational h=0,04, k=0,001 T = 0,14		Rosinger h=0,004, k=0,02 T = 0,14	
x	U	V	E	V	E
0,012	3,949531	3,925373	6,1E-3	3,968594	-4,8E-3
0,052	3,889390	3,844439	1,1E-2	3,905095	-4,0E-3
0,092	3,761903	3,685953	2,0E-2	3,830391	-1,8E-2
0,132	3,506132	3,398354	3,1E-2	3,598506	-2,6E-2
0,172	3,045331	2,936078	3,6E-2	3,046040	-2,3E-4
0,212	2,356164	2,309563	1,9E-2	2,243422	4,8E-2
0,252	1,566964	1,619406	-3,3E-2	1,444392	7,8E-2
0,292	0,897744	1,008460	-1,2E-1	0,837771	6,7E-2
0,332	0,460267	0,566468	-2,3E-1	0,453413	1,5E-2
0,372	0,220802	0,294721	-3,3E-1	0,235205	-6,5E-2
0,412	0,102323	0,145769	-4,2E-1	0,118771	-1,6E-1
0,452	0,046634	0,069848	-4,9E-1	0,058783	-2,6E-1

The relative L_2 error norm is 1,19E-2 for the rational approximation.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

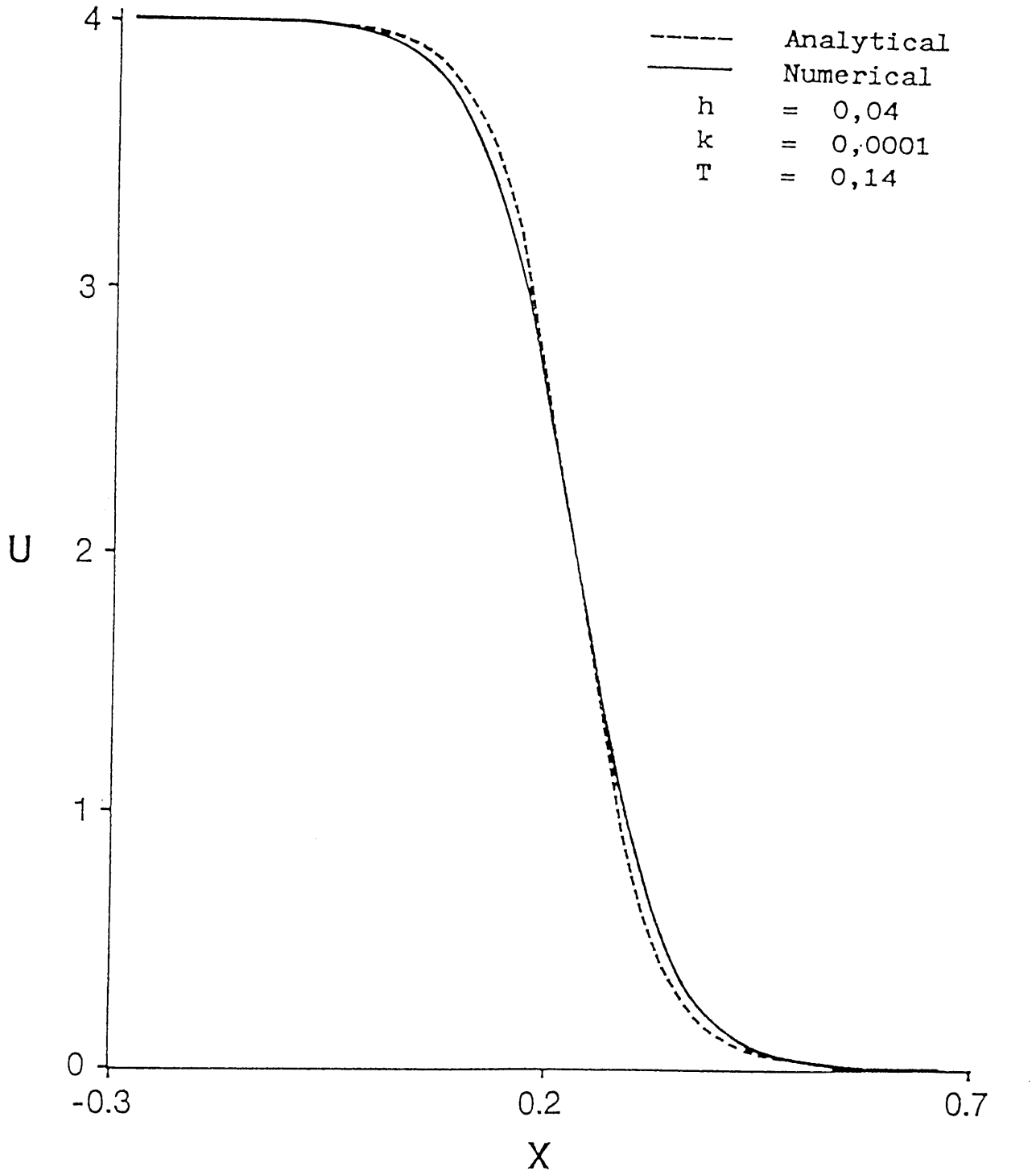

FIGURE 4.1 : Comparison between (1,1) rational approximation and analytical solution

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Higher-order rational basis functions improve the previous results considerably as shown in Table 4.4.

TABLE 4.4 : Higher-order rational approximations

	(2,2) Rational L_2 norm: 7,4E-3		(3,3) Rational L_2 norm: 5,3E-3	
x	V	E	V	E
0,012	3,938371	2,8E-3	3,943696	1,5E-3
0,052	3,866204	5,9E-3	3,875893	3,5E-3
0,092	3,718861	1,1E-2	3,734006	7,4E-3
0,132	3,439305	1,9E-2	3,458755	1,4E-2
0,172	2,972186	2,4E-2	2,989769	1,8E-2
0,212	2,322565	1,4E-2	2,329009	1,2E-2
0,252	1,601427	-2,2E-2	1,592654	-1,6E-2
0,292	0,970357	-8,1E-2	0,951974	-6,0E-2
0,332	0,526358	-1,4E-1	0,507473	-1,0E-1
0,372	0,263673	-1,9E-1	0,249492	-1,3E-1
0,412	0,125669	-2,3E-1	0,116779	-1,4E-1
0,452	0,058195	-2,5E-1	0,053199	-1,4E-1

Next, let R = 10, a = 7,5, b = 0,0 and apply the rational approximation method to the soliton solution. The relative L_2 error norm for Rosinger's method with h = 0,00681 and k = 0,025 is 0,13, while the (1,1), (2,2) and (3,3) rational approximation method with h = 0,05 and k = 0,001 yields 0,012, 0,008 and 0,0059 respectively. Different Reynolds numbers were also used to investigate the sensitivity of the method. Excellent results were obtained at high

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Reynolds numbers, e.g. when $R = 1000$ the relative L_2 error norm is 2,8E-4.

4.5.2 The Korteweg-de Vries equation

The first example considered is

$$u_t + uu_x + \epsilon u_{xxx} = 0, \quad x \in [0,2]$$

with initial condition

$$u(x,0) = 3c\ \mathrm{sech}^2(ax + d),$$

where

$$c = 0,3 \quad ,$$

$$\epsilon = 0,000484 \ ,$$

$$a = (c/4\epsilon)^{1/2}$$

and

$$d = -a.$$

The theoretical solution

$$u(x,t) = 3c\ \mathrm{sech}^2(ax - act + d)$$

represents a single soliton with amplitude 0,9 which moves across the interval in the time period, $0 \leq t \leq 1$. Outside the interval $[0,2]$ the solution is considered to be negligible.

Sanz-Serna, et al [42], considered a Petrov-Galerkin scheme with Hermite cubic test functions and subsequently modified the scheme to include product approximation. The amount of upwinding needed in the rational scheme was determined by numerical experimentation and $\alpha = 1/6$ yielded the best results. In this case the test function $\psi_j(x)$ is slightly biased in the upwind direction and assumes the form as shown in Figure 4.2.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

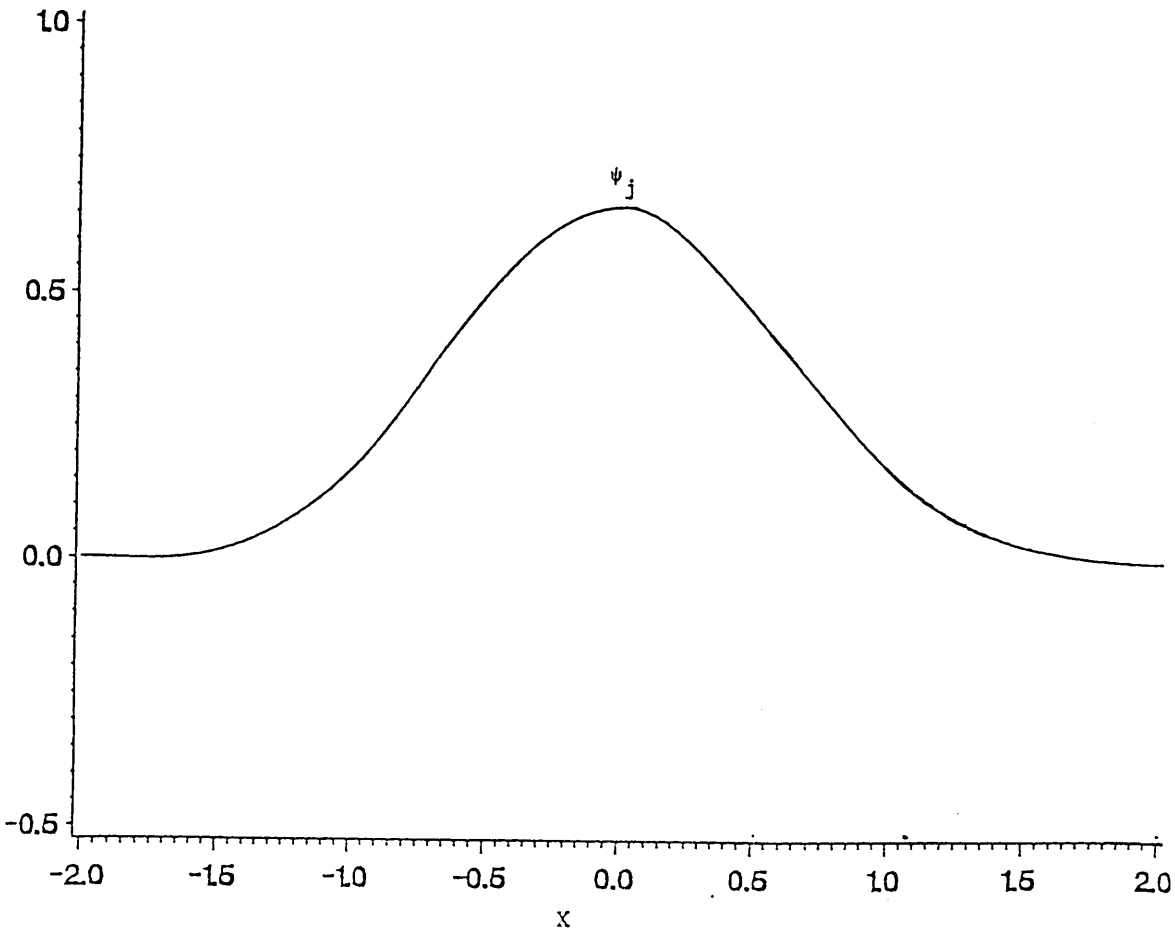

FIGURE 4.2 : Test function ψ_j with $\alpha = 1/6$

The numerical results of the (3,1) Hermite rational basis functions with $\alpha = \frac{1}{6}$ are compared to the results of Zabusky-Kruskal, Petrov-Galerkin and modified Petrov-Galerkin [42] in Table 4.5. The accuracy was measured by the norms

$$\|u - U\|_\infty = \max_{0 \le i \le N} \{u(x_i, t_\ell) - U_i^\ell\}$$

and

$$\|u - U\|_2 = \left[\sum_{i=0}^{N} \left\{ u(x_i, t_\ell) - U_i^\ell \right\}^2 h \right]^{1/2}$$

for a fixed $t_\ell = \ell k, \quad 1 \le \ell \le J.$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The rational scheme (4.15) was solved implicitly ($\theta = \frac{1}{2}$) and explicitly ($\theta = 0$) and the accuracy exceeded the first-order accuracy predicted by the linearised truncation error analysis. The rational method gives very good results for moderate values of h, but the other methods soon catch up for smaller values of h. This shows a slower rate of convergence for these methods.

TABLE 4.5 : L_2 and L_∞ errors (× 1000)

		Zabusky-Kruskal		Petrov-Galerkin		Modified Petrov-Galerkin		(3,1) Rational implicit		(3,1) Rational explicit	
		L_2	L_∞	L_2	L_∞	L_2	L_∞	L_2	L_∞	L_2	L_∞
h=0,05	t	k=0,025		k=0,025		k=0,025		k=0,025		k=0,001	
	0,25	34,64	19,4	81,39	42,18	52,15	30,22	7,77	17,75	8,50	20,28
	0,50	122,68	63,5	102,54	51,85	64,90	22,85	10,17	12,49	11,08	13,84
	0,75	210,44	122,4	125,84	87,60	89,01	35,86	11,25	16,46	12,80	23,10
	1,00	298,19	161,4	150,57	100,41	107,20	39,39	13,16	26,99	15,36	23,85
h=0,033				k=0,01		k=0,01		k=0,01		k=0,001	
	0,25			31,18	14,27	5,94	2,80	1,33	4,73	1,39	4,05
	0,50			43,35	21,65	7,56	4,53	1,52	4,13	1,91	6,37
	0,75			56,21	29,78	8,70	4,85	1,61	4,96	2,53	6,05
	1,00			74,08	39,37	9,49	5,85	1,94	4,66	3,58	9,57
h=0,01		k=0,0005		k=0,005		k=0,005		k=0,005 h=0,02			
	0,25	5,94	2,05	4,46	1,21	0,21	0,07	0,20	0,65		
	0,50	13,17	4,22	7,01	2,15	0,38	0,11	0,24	0,76		
	0,75	21,08	6,36	10,08	3,09	0,57	0,17	0,27	0,76		
	1,00	28,66	8,13	13,26	3,83	0,74	0,21	0,38	0,96		

Next, the rational scheme is compared to a self-adaptive conservative scheme (SACS) by Sanz-Serna [41] in Table 4.6.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

TABLE 4.6 : L_∞(× 1000) AT T = 1

SACS			(3,1) Rational		
h	k	L_∞	h	k	L_∞
0,02	0,005	36,1	0,05	0,025	26,99
0,01	0,0005	9,76	0,033	0,01	4,66
0,01	0,001	7,84	0,02	0,005	0,96

Secondly, consider the equation

$$u_t + 6uu_x + u_{xxx} = 0, \quad -\infty < x < \infty$$

with solitary-wave initial data

$$u(x,0) = A\ \mathrm{sech}^2(cx)$$

and

$$A = 2c^2.$$

The exact solution on the infinite interval is

$$u(x,t) = A\ \mathrm{sech}^2(cx - \omega t),$$

where

$$\omega = 4c^3.$$

In Table 4.7 the numerical results are compared to the results of the numerical schemes reported by Taha, et al [46] and Bona, et al [3].

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

TABLE 4.7 : Comparison of different schemes with different amplitudes on the interval [-20,20] at time T = 1

A	Zabusky-Kruskal	Taha and Ablowitz local scheme	Pseudospectral by Fornberg and Whitham	Calahan method with quadratic splines	(3,1) rational
A = 1	N=231 J=50 L_∞=0,00469	N=251 J=8 L_∞=0,00173	N=65 J=104 L_∞=0,00113	N=96 J=25 L_∞=0,00178	N=96 J=25 L_∞=0,000982
A = 2	N=501 J=5263 L_∞=0,00930	N=401 J=10 L_∞=0,00332	N=129 J=238 L_∞=0,00474	N=144 J=45 L_∞=0,00288	N=145 J=200 L_∞=0,00317
A = 4		N=801 J=37 L_∞=0,01747	N=129 J=870 L_∞=0,01752	N=172 J=140 L_∞=0,0171	N=175 J=200 L_∞=0,0116

From Table 4.7 and Figure 4.3 it is clear that the rational scheme performs satisfactorily and its accuracy can only be attributed to the better approximation abilities of a rational function.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

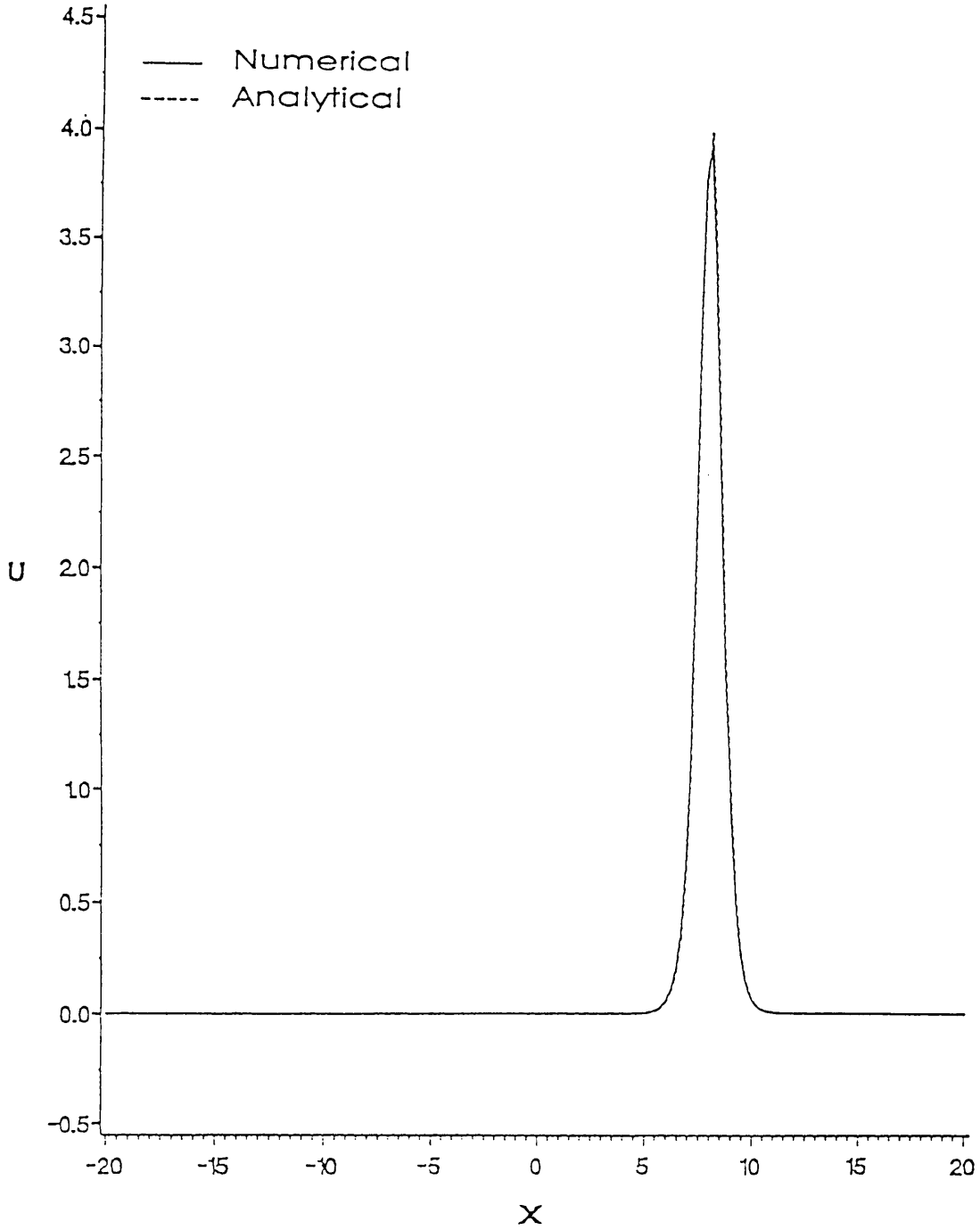

FIGURE 4.3 : (3,1) Rational scheme with $\Lambda = 1$, $J = 200$ and $N = 175$

Although the method is not a conservative scheme in the linearised sense, that is [42]

$$\frac{d}{dt}(\bar{U}^T M \bar{U}) = 0,$$

the L_2-energy does not vary much for the (3,1) Hermite rational approximation as shown in Table 4.8.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

TABLE 4.8 : Energy of (3,1) Hermite rational approximation

t	h/k=0,05/0,025	h/k=0,033/0,01	h/k=0,02/0,005	Exact energy
0,25	0,0863	0,0866	0,0867	0,0868
0,50	0,0857	0,0864	0,0867	0,0868
0,75	0,0852	0,0862	0,0867	0,0868
1,00	0,0848	0,0861	0,0866	0,0868

4.6 CONCLUSIONS

The numerical results obtained by the rational approximation method for the Burgers equation demonstrate that the technique is accurate in the vicinity of steep gradients of the solution. In the numerical experiments conducted with the explicit rational approximation it seems that the scheme is remarkably stable. Although the elements of the tridiagonal matrices A, B and C involve the integration of rational functions it presents no difficulties since the integration is obtained accurately by numerical integration. In addition, the matrix A^{-1} in the explicit rational computational algorithm stays constant throughout the iteration sequence, thus providing a very efficient algorithm. Both implicit and explicit schemes were found to perform satisfactorily at steep gradients and for different Reynolds numbers. The approximation ability of the rational function was clearly demonstrated by the approximation of the wavefronts in the examples with Burgers equation.

Furthermore, the extension to higher-order rational basis functions provides an excellent method to increase the accuracy of the approxi-

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

mation as clearly demonstrated by the results. The method is unique in the sense that additional node-points are not necessary and hence, the tridiagonal structure of the matrices is maintained. Another distinct advantage is that these rational schemes do not possess an unknown parameter that has to be chosen to vary the magnitude of upwinding desired. In conclusion, it is evident that the rational scheme obtained from Galerkin's method with rational basis functions provides an efficient method for solving the Burgers equation.

The rational scheme for the Korteweg-de Vries equation in most cases presents a significant improvement over the methods considered. The error norms reflect the accuracy and efficiency of the new method and suggest that the order of accuracy is better than the first-order accuracy predicted by a linearised truncation error analysis. Although the scheme is not conservative in the linearised sense various numerical experiments have indicated that the energy $\int_{-\infty}^{\infty} u^2 dx$, for computational purposes, remains constant. Finally, it must be emphasised that this method gives a consistent numerical scheme that on account of the influence of the rational functions, has better approximation abilities than most other existing numerical methods.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

CHAPTER 5

NONLINEAR HYPERBOLIC CONSERVATION EQUATIONS

5.1 INTRODUCTION

In this chapter nonlinear scalar equations of the form

$$u_t + f(u)_x = 0 \tag{5.1a}$$

$$u(x,0) = u_0(x), \tag{5.1b}$$

where u is a real valued function, $u_0 \in L_\infty(\mathbb{R})$, f a twice continuous differentiable function, and $(x,t) \in \mathbb{R} \times \mathbb{R}_+$ are studied. These equations are commonly called conservation laws by anology with examples which arise in the study of nonlinear wave phenomena when dissipation effects, such as viscosity are neglected. These equations admit discontinuous solutions, namely shock waves.

There are several difficulties in solving such equations numerically. Typically, numerical schemes produce oscillations behind a shock. Furthermore, all difference schemes have numerical diffusion, dispersion or both due to the truncation error. As a result, there will be diffusion across the discontinuity with each time step. Therefore, it is necessary to have a numerical method that is accurate in smooth regions of the solution while giving a sharp resolution at a discontinuity or shock.

Considering the properties of rational functions, it is envisaged that the rational approximation method would adhere to the above requirements. Firstly, a brief introduction to the mathematical theory is presented, followed by the development of the numerical scheme and

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

proof of the convergence. Numerical results are also presented to show the efficiency of the method.

5.2 MATHEMATICAL SURVEY

The solution of hyperbolic conservation laws of the form

$$u_t + f(u)_x = 0 \tag{5.2a}$$

$$u(x,0) = u_0(x) \tag{5.2b}$$

contains discontinuities, that is the mathematical representation of shock waves. To allow discontinuous solutions one considers weak solutions of (5.2) that satisfy

$$\int_0^\infty \int_{-\infty}^\infty (\phi_t u + \phi_x f(u))dx\,dt + \int_{-\infty}^\infty \phi(x,0)u(x,0)dx = 0, \tag{5.3}$$

where $\phi(x,t)$ is any C_0^1 test function [43]. The integral equation (5.3) imposes a constraint on the value of u on both sides of the discontinuity, namely the Rankine-Hugoniot condition, which states that

$$S[u_\ell - u_r] = [f(u_\ell) - f(u_r)],$$

where $S = \frac{dx}{dt}$ is the speed of the discontinuity and u_ℓ and u_r are the states on the left and the right of the discontinuity respectively. In the effort to solve the initial-value problem, which classically could not be solved, the class of solutions was extended by (5.3). Therefore, several physically unacceptable discontinuous solutions can arise, risking the uniqueness of the problem. An additional condition has to be imposed on weak solutions of conservation laws in order to select the physically relevant solution. Usually this condition identifies the relevant solution u as a limit of solutions, $u(\epsilon)$, with some dissipation present for the viscous equation

$$u_t + f(u)_x = \epsilon u_{xx}, \quad \epsilon > 0. \tag{5.4}$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Oleinik [20,43] has shown that as $\epsilon \to 0$, the solution converges to a solution of the scalar conservation law obeying the entropy condition

$$\frac{f(u) - f(u_\ell)}{u - u_\ell} \geq S \geq \frac{f(u) - f(u_r)}{u - u_r} \tag{5.5}$$

for all u between u_ℓ and u_r. Such limit functions can be characterised by an entropy function $\eta(u)$ [20,52], defined as follows:

(i) η satisfies $\eta_u f_u = F_u$,

where F is the entropy flux.

(ii) η is a convex function of u, i.e. $\eta_{uu} > 0$.

The multiplication of (5.1a) by η_u then yields

$$\begin{aligned} &\eta_u u_t + \eta_u f(u)_x \\ &= \eta_t + F_x \\ &= 0. \end{aligned}$$

It was shown in [19] that if a sequence of solutions $u(\epsilon)$ of (5.4) converges, boundedly and almost everywhere, to a limit u as $\epsilon \to 0$, then the limit satisfies

$$\eta(u)_t + F(u)_x \leq 0 \tag{5.6a}$$

in the weak sense, i.e.

$$-\int_0^\infty \int_{-\infty}^\infty [\phi_t \eta + \phi_x F]dx\,dt - \int_{-\infty}^\infty \phi(x,0)\eta(u(x,0))dx \leq 0 , \tag{5.6b}$$

for all non-negative smooth test functions $\phi(x,t)$ of compact support. Relations (5.6) are called entropy inequalities.

Therefore, in accordance with the above, the numerical approximation of (5.4) will be considered.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

5.3 DISCRETISING THE HYPERBOLIC EQUATION

In order to solve (5.1), where $f(u) = \frac{1}{2}u^2$, the equation

$$u_t + f(u)_x - \epsilon u_{xx} = 0 \qquad (5.7)$$
$$u(x,0) = u_0(x), \quad x \in I = [a,b]$$

will be considered. Since this is identical to the nonlinear Burgers equation, the discrete system (4.4) is applicable, namely

$$\begin{aligned}&(\psi_{j-1},\psi_j)\dot{U}_{j-1} + (\psi_j,\psi_j)\dot{U}_j + (\psi_{j+1},\psi_j)\dot{U}_{j+1}\\ &\quad - \tfrac{1}{2}\{(\psi_{j-1},\psi_j')U_{j-1}^2 + (\psi_j,\psi_j')U_j^2 + (\psi_{j+1},\psi_j')U_{j+1}^2\}\\ &\quad + \epsilon\{(\psi_{j-1}',\psi_j')U_{j-1} + (\psi_j',\psi_j')U_j + (\psi_{j+1}',\psi_j')U_{j+1}\}\\ &= 0 \qquad\qquad j=1,\ldots,N-1.\end{aligned}$$

Using the properties of the (T,T) rational basis functions,

$$-(\psi_{j-1},\psi_j') = (\psi_{j+1},\psi_j') = -\frac{1}{2},$$
$$(\psi_j,\psi_j') = 0,$$
$$(\psi_{j-1},\psi_j) = (\psi_{j+1},\psi_j)$$

and

$$(\psi_{j-1},\psi_j) + (\psi_j,\psi_j) + (\psi_{j+1},\psi_j) = h$$

and discretising the system forward in time the result is

$$\begin{aligned}&AU_{j-1}^{n+1} + BU_j^{n+1} + AU_{j+1}^{n+1}\\ &= AU_{j-1}^{n} + BU_j^{n} + AU_{j+1}^{n}\\ &\quad - \lambda[\tfrac{1}{4}(U_{j+1}^n)^2 - \tfrac{1}{4}(U_{j-1}^n)^2 - CU_{j-1}^n + 2CU_j^n - CU_{j+1}^n], \qquad (5.8)\end{aligned}$$

where

$$A = (\psi_{j-1},\psi_j)/h,$$
$$B = (\psi_j,\psi_j)/h,$$
$$C = -\epsilon(\psi_{j-1}',\psi_j'),$$
$$2A + B = 1$$

and $\lambda = k/h$.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

5.4 CONVERGENCE

In order to investigate the convergence of the numerical scheme (5.8), define the difference approximation $U_{h,k}(x,t)$ to (5.1) as piecewise constant functions,

$$U_{h,k}(x,t) = U_j^n \quad \text{for} \quad (x,t) \in [jh,(j+1)h) \times [nk,(n+1)k).$$

Introduce the following definitions [18,20,52].

Definition 5.4.1.

A difference scheme is consistent with the conservation law

$$u_t + f(u)_x = 0, \quad u(x,0) = u_0(x), \quad x \in I$$

if it can be written in the conservation form

$$\begin{aligned} &AU_{j-1}^{n+1} + BU_j^{n+1} + AU_{j+1}^{n+1} \\ &= AU_{j-1}^n + BU_j^n + AU_{j+1}^n - \lambda[h_f(U_{j+1}^n,U_j^n) - h_f(U_j^n,U_{j-1}^n)] \ , \end{aligned} \tag{5.9}$$

where h_f is a Lipschitz continuous function, i.e.

$$\exists \ K > 0; \quad |h_f(U_{j+1}^n,U_j^n) - h_f(U_j^n,U_j^n)| \le K|U_{j+1}^n - U_j^n| \ ,$$

and satisfies $h_f(u,u) = f(u)$.

In the case of (5.8) the function becomes

$$h_f(U_{j+1}^n,U_j^n) = \frac{1}{4}((U_{j+1}^n)^2 + (U_j^n)^2) + C(U_j^n - U_{j+1}^n)$$

so that

$$h_f(u,u) = f(u) = \frac{1}{2}u^2 \ .$$

Definition 5.4.2.

A difference scheme is consistent with the entropy inequality

$$\eta(u)_t + F(u)_x \le 0$$

if a numerical entropy flux, h_F, exists which satisfies

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

(i) $h_F(u,u) = F(u)$,

(ii) h_F is a Lipschitz continuous function,

(iii) $A\eta(U_{j-1}^{n+1}) + B\eta(U_j^{n+1}) + A\eta(U_{j+1}^{n+1})$
$- A\eta(U_{j-1}^{n}) - B\eta(U_j^{n}) - A\eta(U_{j+1}^{n})$
$- \lambda[h_F(U_{j+1}^{n},U_j^{n}) - h_F(U_j^{n},U_{j-1}^{n})] \leq 0.$

Definition 5.4.3

(a) The total variation (TV) of a mesh function $U_{h,k}$ is defined as

$$TV(U^n) = \sum_j |U_{j+1}^n - U_j^n|.$$

(b) A numerical scheme is total-variation stable if the total variation in x of a sequence of numerical approximations $U_{h,k}(x,t)$ is uniformly bounded in t and h. This means that a constant K^* independent of k and h exists, such that

$$TV(U^n) \leq K^* \quad \forall\, n.$$

Therefore, a constant K can be found independently of k and h, such that $TV(U^n) \leq K\, TV(U^0)$.

(c) A numerical scheme

$$U_j^{n+1} = S \cdot U_j^n ,$$

where S is a difference operator, is total-variation diminishing (TVD) if

$$TV(U^{n+1}) \leq TV(U^n).$$

To establish the convergence of the numerical scheme it is necessary to prove the following lemmas.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Lemma 5.4.4

If the numerical scheme (5.8) is consistent with the conservation law and total-variation stable, then for all p and q

$$\sum_{j\leq\frac{|b-a|}{h}} |U_j^p - U_j^q|h \leq \frac{M}{\lambda}(p - q)k, \tag{5.10}$$

where M is a constant, $\lambda = k/h$ a constant and $[a,b]$ a finite interval of the real line.

Proof. It can be shown after tedious manipulation that

$$\begin{aligned}
&U_j^{p+1} - U_j^{p-1} \\
&= -\lambda[h_f(U_{j+1}^p, U_j^p) - f(U_j^p)] + \lambda[h_f(U_j^p, U_{j-1}^p) - f(U_j^p)] \\
&\quad -\lambda[h_f(U_{j+1}^{p-1}, U_j^{p-1}) - f(U_j^{p-1})] + \lambda[h_f(U_j^{p-1}, U_{j-1}^{p-1}) - f(U_j^{p-1})] \\
&\quad +A(U_j^{p+1} - U_{j-1}^{p+1}) + A(U_j^{p+1} - U_{j+1}^{p+1}) - A(U_j^{p-1} - U_{j-1}^{p-1}) - A(U_j^{p-1} - U_{j+1}^{p-1}).
\end{aligned}$$

Therefore, using the Lipschitz condition,

$$\begin{aligned}
&|U_j^{p+1} - U_j^{p-1}| \\
&\leq \lambda[|U_{j+1}^p - U_j^p| + |U_{j-1}^p - U_j^p| + |U_{j+1}^{p-1} - U_j^{p-1}| + |U_{j-1}^{p-1} - U_j^{p-1}|] \\
&\quad + A[|U_j^{p+1} - U_{j-1}^{p+1}| + |U_j^{p+1} - U_{j+1}^{p+1}| + |U_j^{p-1} - U_{j-1}^{p-1}| + |U_j^{p-1} - U_{j+1}^{p-1}|].
\end{aligned}$$

From the total-variation stability of U_j^p and assuming $\lambda = k/h$ constant, it follows that

$$|U_j^{p+1} - U_j^{p-1}| \leq M ,$$

where M is a constant, so that

$$\sum_{j\leq\frac{|b-a|}{h}} |U_j^{p+1} - U_j^{p-1}|h \leq Mh.$$

The triangle inequality yields

$$\begin{aligned}
\sum_{j\leq\frac{|b-a|}{h}} |U_j^p - U_j^q|h &\leq \sum_{i=q}^{p-2} \sum_j |U_j^{i+2} - U_j^i|h \\
&\leq (p - q)Mh \\
&= \frac{M}{\lambda}(p - q)k.
\end{aligned}$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Lemma 5.4.5

If the numerical scheme satisfies the assumptions of the previous lemma and is total-variation stable, then it is also stable in the norm

$$\|U\| = TV(U) + \max_j |U_j|. \tag{5.11}$$

Proof. The lemma is proved for periodic boundary conditions. The assumption of total-variation stability means a constant K exists, not depending on h or k, such that

$$TV(U^n) \leq K\, TV(U^0).$$

By definition of total variation

$$\begin{aligned} \max_j U_j^n - \min_j U_j^n &\leq \sum_j |U_{j+1}^n - U_j^n| \\ &= TV(U^n) \\ &\leq K\, TV(U^0). \end{aligned}$$

Using periodic boundary conditions, that is $U_0^n = U_N^n$, and consistency with the conservation law, gives

$$\sum_j U_j^n = \sum_j U_j^0 \quad \text{for all} \quad n \geq 0.$$

Now

$$\min_j U_j^n \leq \frac{1}{N} \sum_j U_j^n \leq \max_j U_j^n ,$$

where N is the number of mesh points in the space interval. From the above

$$U_j^n \geq \min_j U_j^n \geq -K\, TV(U^0) + \max_j |U_j^n| \geq -K\, TV(U^0) + \frac{1}{N} \sum_j U_j^0 ,$$

$$U_j^n \leq \max_j U_j^n \leq K\, TV(U^0) + \min_j |U_j^n| \leq K\, TV(U^0) + \frac{1}{N} \sum_j U_j^0 .$$

Thus,

$$U_j^n - \frac{1}{N} \sum_j U_j^0 \geq -K\, TV(U^0) \quad \forall\ j$$

and

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$U_j^n - \frac{1}{N}\sum_j U_j^0 \leq K\ TV(U^0) \quad \forall\ j$$

implies that

$$|U_j^n - \frac{1}{N}\sum_j U_j^0| \leq K\ TV(U^0) \quad \forall\ j.$$

But

$$\big||U_j^n| - \frac{1}{N}|\sum_j U_j^0|\big| \leq |U_j^n - \frac{1}{N}\sum_j U_j^0| \leq K\ TV(U^0)$$

so that for all j

$$\begin{aligned}|U_j^n| &\leq K\ TV(U^0) + \frac{1}{N}|\sum_j U_j^0| \\ &\leq K\ TV(U^0) + \max_j |U_j^0|. \end{aligned} \tag{5.12}$$

Hence

$$\max_j |U_j^n| \leq K\ TV(U^0) + \max_j |U_j^0| \tag{5.13}$$

and

$$TV(U^n) + \max_j |U_j^n| \leq 2K\ TV(U^0) + \max_j |U_j^0|. \tag{5.14}$$

From (5.12), (5.13) and (5.14) it is clear that the numerical scheme is uniformly bounded in the L_1 norm, L_∞ norm and the norm (5.11) respectively.

The convergence of the numerical scheme is outlined in the following theorem [18,43].

Theorem 5.4.6

Assume that the numerical scheme (5.8) is consistent with the conservation law and its entropy inequality. If the resulting numerical approximation is TV stable, then a sequence $\{U_{h_i,k_i}\} \subset \{U_{h,k}\}$ exists which converges to a measurable function $u(x,t)$ in the sense that

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$\int_a^b |U_{h_i,k_i}(x,t) - u(x,t)|dx \longrightarrow 0$$

and

$$\int_0^T \int_a^b |U_{h_i,k_i}(x,t) - u(x,t)|dx\, dt \longrightarrow 0, \quad T > 0,$$

as $i \longrightarrow \infty$, i.e. $(h_i,k_i) \longrightarrow (0,0)$. Furthermore, its limit is the weak solution of (5.1) that satisfies the entropy inequality.

Proof. The first part is analogous to that of [43].

Based on lemmas 5.4.4 and 5.4.5, the set of functions $\{U_{h,k}\}$, considered as functions of x, is uniformly bounded, and by assumption has uniformly bounded total variation on each bounded interval on any line t = constant, $t > 0$. According to Helley's theorem [43] a subsequence $\{U'_{h,k}\}$ exists which converges at each point on any bounded interval of this line to some function of finite variation. Let $\{t_m\}$ be a countable dense subset of the interval $(0,T]$. By means of a diagonal process a subsequence $\{U_{h_i,k_i}\}$ from $\{U'_{h,k}\}$ is found which converges at every point on each line $t = t_m$, $m = 1,2,\ldots$ as $i \longrightarrow \infty$.

Let $U_i = U_{h_i,k_i}$ and show next that for every $t \in (0,T]$, $\{U_i(\cdot,t)\}$ is a Cauchy sequence in $L_1([a,b])$. For $t \in (0,T]$ find a subsequence $\{t_{m_s}\} \subset \{t_m\}$ such that $t_{m_s} \longrightarrow t$ as $s \longrightarrow \infty$. Then

$$\begin{aligned} I_{ij}(t) &= \int_a^b |U_i(x,t) - U_j(x,t)|dx \\ &\leq \int_a^b |U_i(x,t) - U_i(x,t_{m_s})|dx + \int_a^b |U_i(x,t_{m_s}) - U_j(x,t_{m_s})|dx \\ &\quad + \int_a^b |U_j(x,t_{m_s}) - U_j(x,t)|dx \\ &= I_1 + I_2 + I_3. \end{aligned}$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Since $|U_i(x,t_{m_s}) - U_j(x,t_{m_s})| \leq M$, for all $x \in [a,b]$, M a real constant, then as a result of the Lebesque bounded convergence theorem [38], $I_2 \rightarrow 0$ as $i,j \rightarrow \infty$.

Denote by means of $[\sigma]$ the greatest integer value of σ, then

$$U_i(x,t) = U_i(x,[t/k_i]k_i].$$

Using lemma 5.4.4

$$\begin{aligned} I_1 &= \int_a^b |U_i(x,[t/k_i]k_i) - U_i(x,[t_{m_s}/k_i]k_i)|dx \\ &\leq \sum_{j \leq (b-a)/h_i} \int_{jh_i}^{(j+1)h_i} |U_i(x,[t/k_i]k_i) - U_i(x,[t_{m_s}/k_i]k_i)|dx \\ &\leq L|[t/k_i] - [t_{m_s}/k_i]|k_i \\ &= L|t - t_{m_s}|, \end{aligned}$$

where L is a constant. Similarly, $I_3 \leq L|t - t_{m_s}|$, thus

$$I_1 + I_3 \leq 2L|t - t_{m_s}|.$$

Now, for $\epsilon > 0$ given, select t_{m_s} such that $4L|t - t_{m_s}| < \epsilon$. For this fixed s select i and j large enough so that $2I_2 < \epsilon$. Then for these s, i and j

$$I_{ij}(t) < \epsilon.$$

Since ϵ is arbitrary, $I_{ij}(t) \rightarrow 0$ as $i,j \rightarrow \infty$. Thus, for every fixed $t \in (0,T]$ the sequence $\{U_i(x,t)\}$ is a Cauchy sequence in $L_1([a,b])$. Since L_1 is complete, it follows that $\{U_i(x,t)\}$ has a limit $u(x,t)$ for each fixed t, $0 < t \leq T$.

It is necessary to show now that $I_{ij}(t) \rightarrow 0$ as $i,j \rightarrow \infty$, uniformly in t, $0 < \tau \leq t \leq T$. Let $\epsilon > 0$ be given and choose a finite subset $F \subset \{t_m\}$ with the property that if $0 \leq t \leq T$, then

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

a $t_m \in F$ exists satisfying $2L|t - t_m| < \frac{\epsilon}{2}$. Select i and j large enough such that $2I_2 < \epsilon$ for all $t_m \in F$. Then $I_{ij}(t) < \epsilon$ for these i and j. Since ϵ is independent of t, $I_{ij}(t) \to 0$ as $i,j \to \infty$ uniformly in t.

Thus, for any τ, $0 < \tau \leq T$,

$$\int_{\tau}^{T} I_{ij}(t)dt \to 0$$

as $i,j \to \infty$.

Now, $$\int_0^T \int_a^b |U_i(x,t) - U_j(x,t)|dx\,dt$$

$$= \int_0^{\tau} \int_a^b |U_i(x,t) - U_j(x,t)|dx\,dt + \int_{\tau}^{T} I_{ij}(t)dt \,. \qquad (5.15)$$

Let $M = \max_i |U_i(x,0)|$, then from (5.12) it follows that $|U_i(x,t)| \leq M$, therefore

$$\int_0^{\tau} \int_a^b |U_i(x,t) - U_j(x,t)|dx\,dt$$

$$\leq \int_0^{\tau} \int_a^b |U_i(x,t)|dx\,dt + \int_0^{\tau} \int_a^b |U_j(x,t)|dx\,dt$$

$$\leq 2(b - a)M\tau. \qquad (5.16)$$

Select τ so small that $2(b - a)M\tau < \epsilon/2$ and i and j large enough that the second integral in (5.15) is at the most $\epsilon/2$. Consequently, for these i and j

$$\int_0^{T} I_{ij}(t)dt < \epsilon$$

and $\int_0^T I_{ij}(t)dt \to 0$ as $i,j \to \infty$. Hence, $U_i(x,t)$ is a Cauchy sequence in $L_1((0,T] \times [a,b])$ and it follows that $\{U_i(x,t)\}$ converges to a limit $u(x,t)$, which is measurable because the sequence $\{U_i\}$ is measurable [39].

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

According to Rudin [39], if $\{U_i\}$ is a Cauchy sequence in $L_1((0,T] \times [a,b])$ with limit u, then a subsequence exists which converges pointwise almost everywhere to $u(x,t)$. Since $|U_i| \leq M$, then $|u| \leq M$ and, if necessary, redefine u on a set of measure zero.

The second part of the theorem proves that $u(x,t)$ is a weak solution.

Define the following piecewise constant functions:

$$\begin{aligned} U^h(x,t) &= U_j^n \ , \\ \phi^h(x,t) &= \phi_j^n \ , \\ f^h(x,t) &= h_f(U_{j+1}^n, U_j^n) \ , \\ \phi_t^h(x,t) &= (\phi_j^{n+1} - \phi_j^n)/k \ , \\ \phi_x^h(x,t) &= (\phi_{j+1}^n - \phi_j^n)/h \ , \\ F^h(x,t) &= h_F(U_{j+1}^n, U_j^n) \ , \end{aligned} \tag{5.17}$$

for $(x,t) \in [jh,(j+1)h) \times [nk,(n+1)k)$.

Multiply (5.9) by a smooth periodic scalar function, $\phi(x,t) \in C_0^1(I)$, i.e. $\phi_j^n = \phi(jh,nk)$ and sum over j and n, to obtain

$$\begin{aligned} &- \sum_j \phi_j^0 \{AU_{j-1}^0 + BU_j^0 + AU_{j+1}^0\}h \\ &- \sum_{n,j} \frac{(\phi_j^{n+1} - \phi_j^n)}{k} \{AU_{j-1}^{n+1} + BU_j^{n+1} + AU_{j+1}^{n+1}\}hk \\ &- \sum_{n,j} \frac{(\phi_{j+1}^{n+1} - \phi_j^{n+1})}{h} h_f(U_{j+1}^{n+1}, U_j^{n+1})hk \\ &= 0 \end{aligned} \tag{5.18}$$

using summation by parts and periodicity of the boundary conditions.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Owing to the definitions (5.17),

$$-\int_I \phi^h(x,0)(AU^h(x+h,0) + BU^h(x,0) + AU^h(x-h,0))dx$$

$$-\int_0^\infty \int_I \phi_t^h(x,t)(AU^h(x+h,t) + BU^h(x,t) + AU^h(x-h,0))dx\,dt$$

$$-\int_0^\infty \int_I \phi_x^h(x,t)f^h(x,t)dx\,dt = 0, \qquad (5.19)$$

where $I = [a,b]$. Since h_f is K-Lipschitz continuous, f continuous, ϕ continuous and U_j^n bounded, then

$$\begin{aligned}|f^h(x,t)| &\le |h_f(U_{j+1}^n, U_j^n) - f(U_j^n)| + |f(U_j^n)| \\ &\le K|U_{j+1}^n - U_j^n| + M \\ &\le M\ ,\end{aligned}$$

$$|U^h(x \pm h,0)| \le M\ ,$$

$$|\phi^h(x,t)| \le M\ ,$$

$$|\phi_t^h(x,t)| \le M$$

and

$$|\phi_x^h(x,t)| \le M$$

where M is an arbitrary constant. Moreover, if

$$\lim_{j\to\infty} U_j^n = U^n = u(x,t),$$

then

$$\lim_{h\to 0} f^h(x,t) = \lim_{j\to\infty} h_f(U_{j+1}^n, U_j^n) = h_f(U^n, U^n) = f(U^n).$$

Finally, based on Lebesque's dominated convergence theorem [38] and the fact that $2A + B = 1$,

$$\int_I \phi(x,0)u(x,0)dx + \int_0^\infty \int_I \left(\frac{\partial \phi}{\partial t}u + \frac{\partial \phi}{\partial x}f(u)\right)dx\,dt = 0, \quad x \in I,$$

which proves that $u(x,t)$ is a weak solution.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

In order to show that the numerical scheme satisfies the entropy inequality, consider the discrete entropy inequality

$$\begin{aligned} &A\eta(U_{j-1}^{n+1}) + B\eta(U_j^{n+1}) + A\eta(U_{j+1}^{n+1}) \\ &- A\eta(U_{j-1}^{n}) - B\eta(U_j^{n}) - A\eta(U_{j+1}^{n}) \\ &+ \lambda[h_F(U_{j+1}^{n},U_j^{n}) - h_F(U_j^{n},U_{j-1}^{n})] \le 0. \end{aligned} \tag{5.20}$$

Multiply (5.20) by $h\phi_j^n$, $\phi_j^n \ge 0$, sum over n and j,

$$\begin{aligned} &- \sum_j h\phi_j^0(A\eta(U_{j-1}^0) + B\eta(U_j^0) + A\eta(U_{j+1}^0)) \\ &- \sum_{n,j} \frac{(\phi_j^{n+1} - \phi_j^n)}{k}\{A\eta(U_{j-1}^{n+1}) + B\eta(U_j^{n+1}) + A\eta(U_{j+1}^{n+1})\}hk \\ &- \sum_{n,j} \frac{(\phi_{j+1}^{n+1} - \phi_j^{n+1})}{k} h_F(U_{j+1}^{n+1},U_j^{n+1})hk \\ &\le 0. \end{aligned}$$

Again this gives

$$\begin{aligned} &- \int_I \phi^h(x,0)(A\eta(U^h(x+h,t)) + B\eta(U^h(x,t)) + A\eta(U^h(x-h,t)))dx \\ &- \int_0^\infty \int_I [\{A\eta(U^h(x+h,t)) + B\eta(U^h(x,t)) + A\eta(U^h(x-h,t))\} \\ &\times \phi_t^h(x,t) + \phi_x^h(x,t)F^h(x,t)]dx\,dt \le 0. \end{aligned}$$

Further,

$$\begin{aligned} \lim_{h\to 0} F^h(x,t) &= \lim_{j\to\infty} h_F(U_{j+1}^n,U_j^n) \\ &= F(U^n). \end{aligned}$$

If η is a continuous convex function in $\mathbb{R}$, then owing to Lebesque's dominated convergence theorem

$$- \int_I \phi(x,0)\eta(u(x,0))dx - \int_0^\infty \int_I \{\eta(u)\frac{\partial\phi}{\partial t} + \frac{\partial\phi}{\partial x}F(u)\}dx\,dt \le 0,$$

for $x \in I$. This proves the theorem.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

5.5 BOUNDS FOR THE ADDED VISCOSITY

Linearise the nonlinear discrete equation (5.8) by replacing the nonlinear terms

$$\frac{1}{4}(U^n_{j+1})^2 \simeq \frac{S}{2} U^n_{j+1}$$

and

$$\frac{1}{4}(U^n_{j-1})^2 \simeq \frac{S}{2} U^n_{j-1} ,$$

where S is the velocity of the shock wave. By a Taylor series expansion, the linearised truncation error $T_{j,n}$ at (jh,nk) is

$$\begin{aligned} T_{j,n} = \frac{1}{hk}\{&hA(2u + 2ku_t + h^2u_{xx} + k^2u_{tt} + h^2ku_{txx} + \frac{k^3}{3} U_{ttt}) \\ &+ hB(kU_t + \frac{k^2}{2}u_{tt} + \frac{k^3}{6}U_{ttt}) - hA(2u + h^2u_{xx}) \\ &+ \frac{kS}{2}(2hu_x + \frac{h^3}{3}u_{xxx}) - kCh^2u_{xx}\} + O(h^3,h^2k,hk^2,k^3). \end{aligned}$$

Using $2A + B = 1$, the above simplifies to

$$\begin{aligned} T_{j,n} &= u_t + \frac{k}{2}u_{tt} + \frac{k^2}{6}u_{ttt} + Ah^2u_{txx} \\ &\quad + Su_x + \frac{Sh^2}{6}u_{xxx} - Chu_{xx} + O(h^3,h^2k,hk^2,k^3) \\ &= u_t + Su_x + O(h,k), \end{aligned}$$

which shows that the numerical scheme is first-order accurate in h and k. The term Chu_{xx} represents the contribution due to the diffusion term. Neglect this term temporarily and denote the truncation error by

$$T^{II}_{j,n} = u_t + \frac{k}{2}u_{tt} + \frac{k^2}{6}u_{ttt} + Ah^2u_{txx} + Su_x + \frac{Sh^2}{6}u_{xxx}.$$

Now,

$$\begin{aligned} u_{txx} &= (u_t)_{xx} \\ &= (-Su_x)_{xx} \\ &= -Su_{xxx}. \end{aligned}$$

Similarly,

$$u_{tt} = S^2u_{xx}$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

and

$$u_{ttt} = -S^3 u_{xxx}.$$

The replacement of the terms in the truncation error yields

$$T^{II}_{j,n} = u_t + Su_x + \frac{S^2k}{2}u_{xx} + (-\frac{S^3k^2}{6} - SAh^2 + \frac{h^2S}{6})u_{xxx}.$$

The term $\frac{S^2k}{2}$ is the coefficient of numerical dissipation, while $(-\frac{S^3k^2}{6} - SAh^2 + h^2\frac{S}{6})$ is the coefficient of numerical dispersion. Inclusion of the added viscosity or diffusion yields

$$T_{j,n} = u_t + Su_x + (\frac{S^2k}{2} - Ch)u_{xx} + (-\frac{S^3k^2}{6} - SAh^2 + h^2\frac{S}{6})u_{xxx}$$

The coefficient of numerical dispersion $(-\frac{S^3k^2}{6} - SAh^2 + h^2\frac{S}{6})$ stays intact for both the discrete hyperbolic and parabolic equations. The dispersive coefficient indicates Fourier components travelling at different speeds, resulting in oscillations. A steep gradient, such as a shock, contains high-frequency Fourier components. Dissipation damps the high-frequency terms more strongly than the low-frequency terms. This results in smoothing shocks. High-frequency components move more slowly than low-frequency terms. Thus, if there is only a small amount of numerical dissipation present, these high-frequency components will lag behind the low-frequency components, resulting in oscillations trailing the shock wave [36].

In order for the coefficient of numerical dissipation to be representative of the viscosity of a "real physical fluid", the amount of dissipation or diffusion must be negative or else the amount of diffusion would be "unphysical". This means that

$$\frac{S^2k}{2} - Ch \leq 0.$$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

From this a lower bound for the diffusion coefficient ϵ is obtained. Since $C = \frac{7\epsilon}{6h}$ for a $(0,1)$ rational basis function, it follows that

$$\epsilon \geq \frac{12}{7}S^2k \simeq 0,43S^2k. \tag{5.21}$$

For $(0,2)$ and $(0,3)$ rational basis functions, it follows respectively that

$$\epsilon \geq \frac{1}{1,1046}\,\frac{S^2k}{2} \simeq 0,45S^2k$$

and (5.22)

$$\epsilon \geq \frac{1}{1,0764}\,\frac{S^2k}{2} \simeq 0,46S^2k.$$

This suggests that the diffusion coefficient increases with increasing order of rational basis functions.

The stability of the linearised discrete equation was also examined by a standard Fourier stability analysis in Chapter 3 and the following conditions for the $(1,1)$, $(2,2)$ and $(3,3)$ rational basis functions were found, namely

$$\epsilon \leq 0,156\,\frac{h^2}{k},$$

$$\epsilon \leq 0,168\,\frac{h^2}{k} \tag{5.23}$$

and

$$\epsilon \leq 0,174\,\frac{h^2}{k}$$

respectively. By using the above to select ϵ, an upper bound is obtained which simultaneously satisfies the stability properties of the numerical scheme.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

5.6 NUMERICAL RESULTS

Consider the linear hyperbolic equation

$$u_t = cu_x \ , \quad -\infty < x < \infty, \quad t > 0 \ ,$$

with initial condition

$$u(x,0) = u_0(x) = \begin{cases} 1 \ , & 0{,}05 < x < 0{,}15 \\ 0 \ , & \text{elsewhere.} \end{cases}$$

The exact solution is $u(x,t) = u_0(x - ct)$. Restrict t to $0 \leq t \leq 0{,}75$ and x to $-1 \leq x \leq 2$. Select $c = -1$, $h = 0{,}02$ and $h = 0{,}01$, while the time step is taken as $k = 0{,}001$. The Courant condition, $|c|\frac{k}{h} \leq 1$, is satisfied for both choices of h, namely $0{,}05$ and $0{,}1$.

The dependency of the numerical solution of (5.8) upon ϵ is investigated and the results are given in Table 5.1.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

TABLE 5.1: L_2 error norm at time t = 0,75

(1,1) Rational approximation				
h\ϵ	0,0004	0,0005	0,00075	0,001
0,02	0,01545	0,01498	0,01618	0,01812
0,01	0,00884	0,00921	0,01125	0,01289
(2,2) Rational approximation				
0,02		0,01485	0,01579	0,01762
0,01		0,00908	0,01094	0,01254
(3,3) Rational approximation				
0,02				0,01740
0,01				0,01238

Table 5.1 shows the same tendency as predicted by the lower and upper bounds (5.21), (5.22) and (5.23). From the table it is clear that the smaller space step yields better results owing to the better approximation of the initial data. By increasing the order of the rational basis functions one has to increase ϵ in order to avoid oscillations. This is so because the large wave number frequencies are not damped by the small numerical dissipation and as a result, a trail of oscillations is left behind the wave. An increase in the order of approximation has a distinct advantage as can be seen from Figures 5.1 and 5.2. The graphs also verify that any phase errors are very small.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

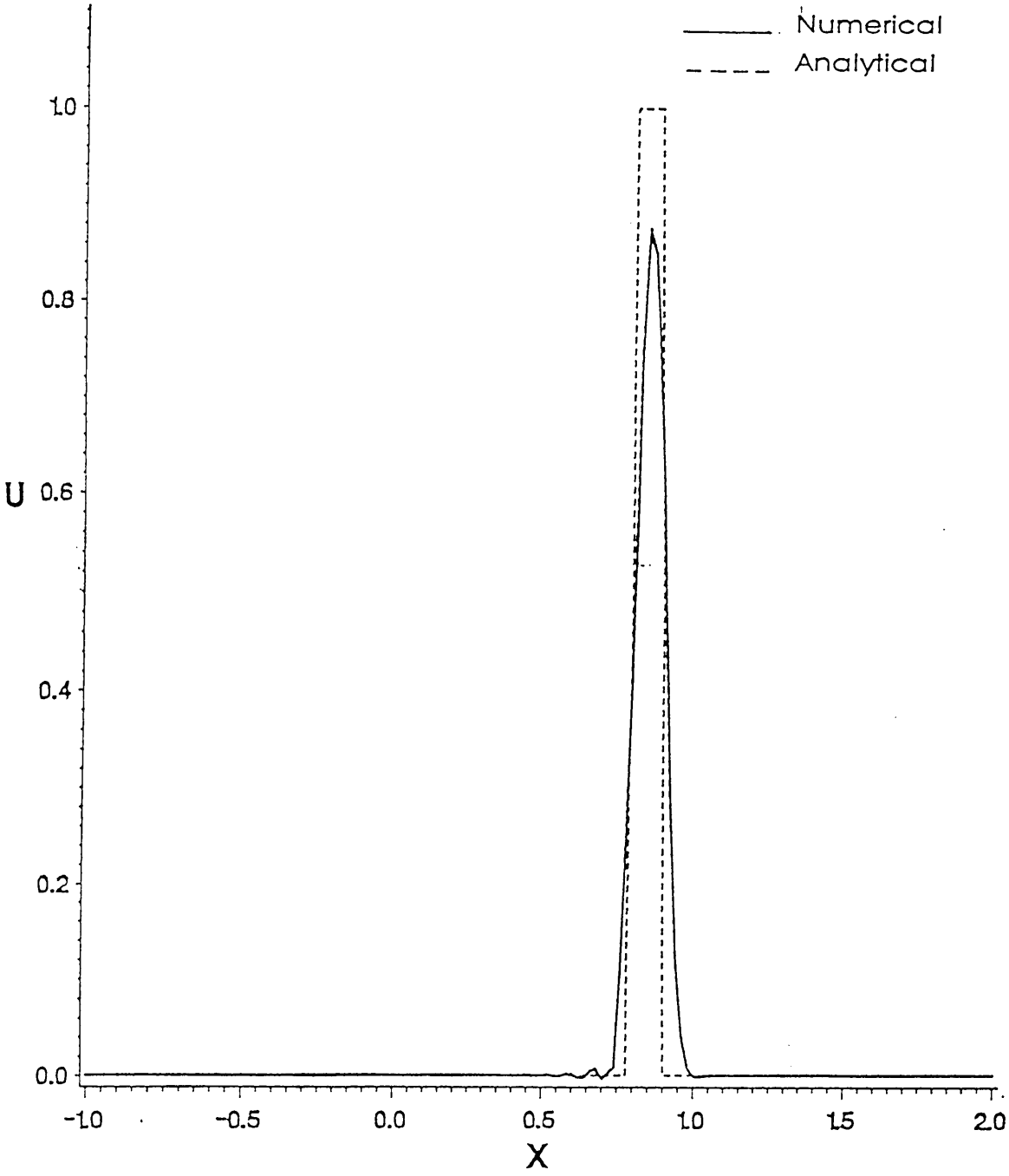

FIGURE 5.1 : (1,1) Rational approximation with $h = 0,02$ and $\epsilon = 0,001$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

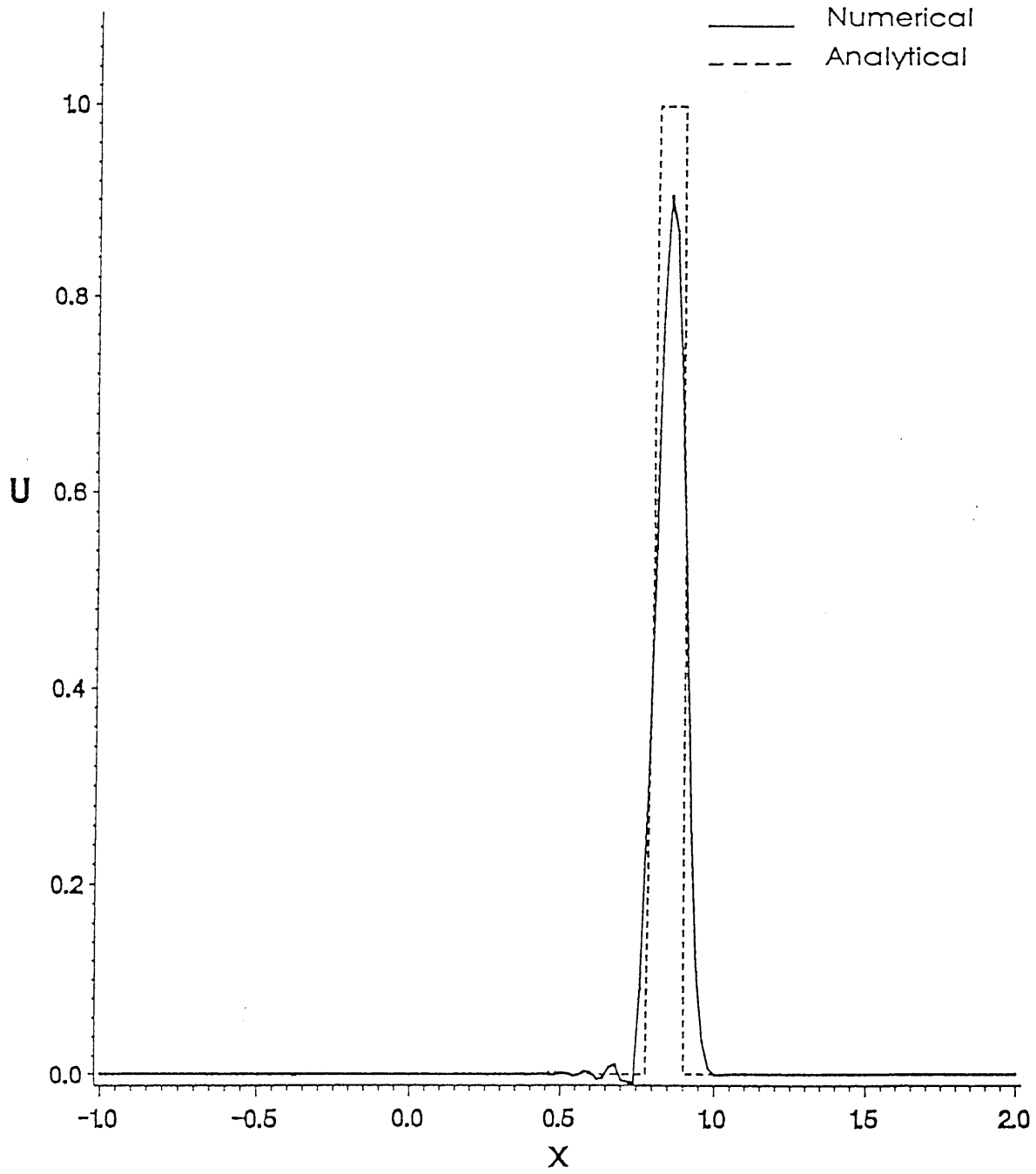

FIGURE 5.2 : (3,3) Rational approximation with $h = 0,02$ and $\epsilon = 0,001$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Secondly, consider

$$u_t + \frac{1}{2}(u^2)_x = 0, \quad -\infty < x < \infty, \quad t > 0$$

with the continuous initial condition

$$u(0,t) = \begin{cases} 1 & , \quad x < 0 \\ 1-x & , \quad 0 \leq x \leq 1 \\ 0 & , \quad x > 1 \end{cases}$$

The exact solution for $t < 1$ is

$$u(x,t) = \begin{cases} 1 & , \quad x < t \\ \frac{x-1}{t-1} & , \quad t \leq x \leq 1 \\ 0 & , \quad x \geq 0 \quad , \end{cases}$$

which describes a wave, consisting of a fan of characteristics for $0 \leq x \leq 1$. At time $t = 1$, however, a shock wave is formed which is described by the solution

$$u(x,t) = \begin{cases} 1 & , \quad x > (t+1)/2 \\ 0 & , \quad x < (t+1)/2 \end{cases}$$

for $t \geq 1$. To avoid boundary conditions, x and t are restricted to $-1 \leq x \leq 3$ and $0 \leq t \leq 2{,}5$ respectively. The time step was fixed, namely $k = 0{,}001$. The numerical scheme (5.8) is tested by using $h = 0{,}0444$ and $h = 0{,}0222$ for different orders of approximation (see Table 5.2).

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

TABLE 5.2 : L_2 error norm at t = 2,5

(1,1) Rational approximation				
h\ϵ	0,01	0,009	0,0075	0,001
0,0444	0,44810	0,47033	0,44377	0,57296
0,0222	0,63652	0,60408	0,55208	0,32754
(2,2) Rational approximation				
0,0444	0,47804	0,46109	0,43589	0,59586
0,0222	0,61935	0,58781	0,53725	0,33643
(3,3) Rational approximation				
0,0444	0,47345	0,45689	0,43236	0,60687
0,0222	0,61137	0,58025	0,53035	0,34094

The results indicate that higher-order rational approximations yield better results than lower orders except for the lowest artificial viscosity where a growth in the L_2 norm was measured. This can be ascribed to the sudden change in gradient of the numerical solution. Introducing a bigger viscosity enlarges the length of the transition in which the shock changes its gradient, thereby smoothing the solution considerably. Although the L_2 norms for the larger space step are better than that of the smaller space step, the results obtained by the latter are very smooth and inhibit the formation and growth of numerical oscillations as seen in Figures 5.3 and 5.4.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

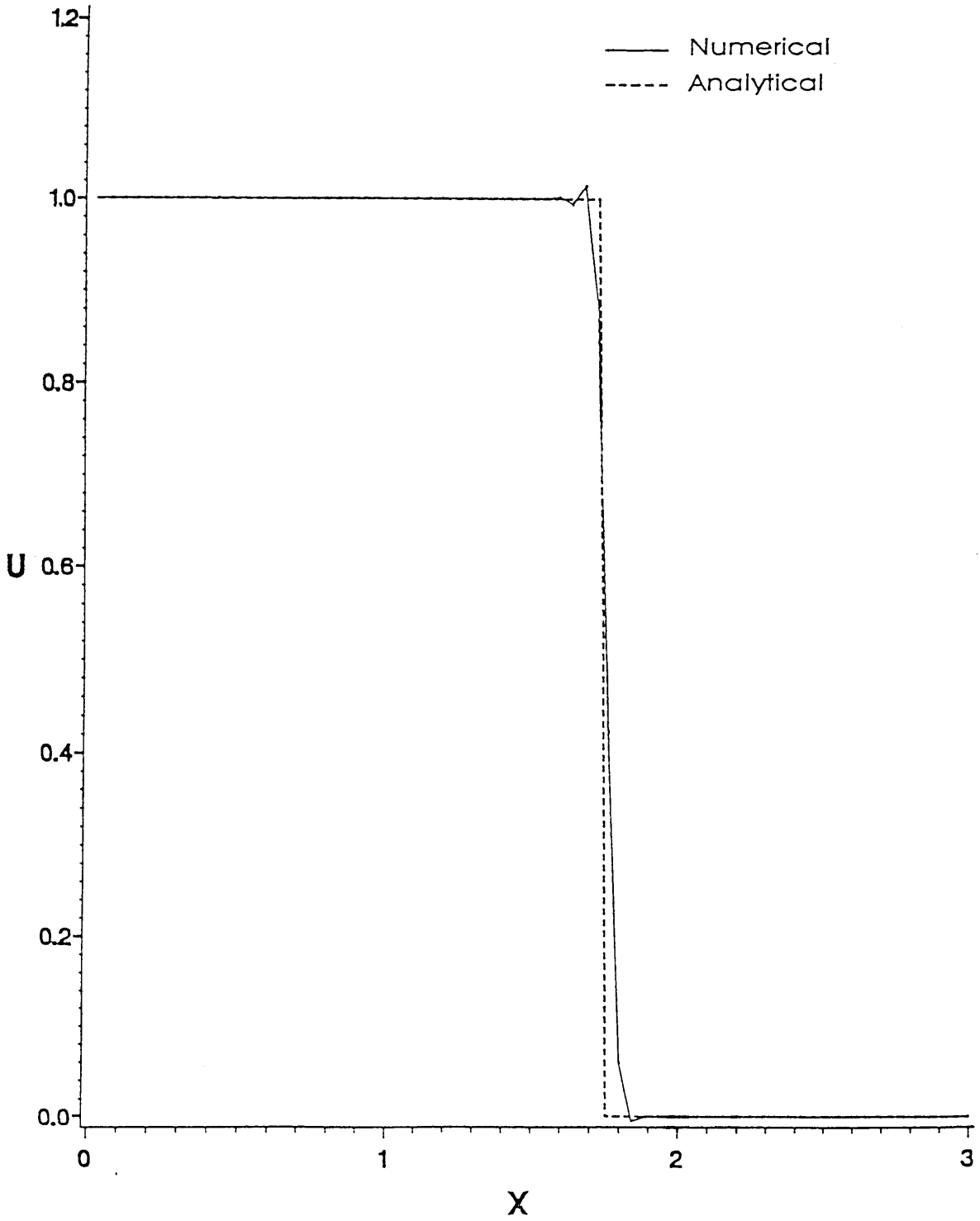

FIGURE 5.3: (1,1) Rational approximation with
ϵ = 0,0075 and h = 0,01444

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

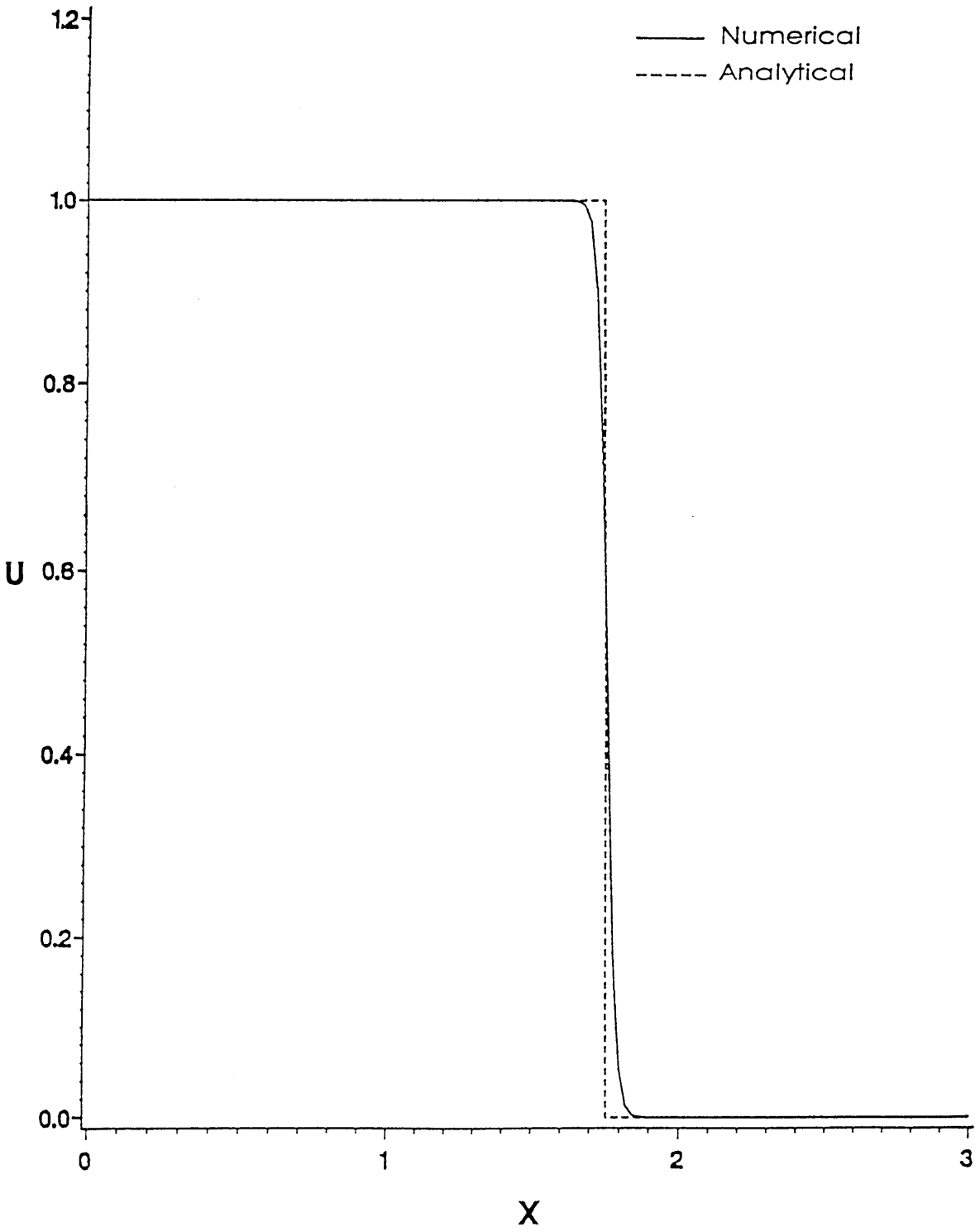

FIGURE 5.4 : (1,1) Rational approximation with
ϵ = 0,0075 and h = 0,0222

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The results obtained by the previous problems compare favourably with those reported by SOD [44, 45].

Next, the rational scheme is compared to a uniformly second-order non-oscillatory scheme (UNO2) by Harten, et al [18] and a typical second-order total-variation diminishing scheme (TVD2), [18]. Both schemes can be written as

$$U_j^{n+1} = U_j^n - \lambda(\hat{f}_{j+1/2} - \hat{f}_{j-1/2}),$$

$$\hat{f}_{j+1/2} = \begin{cases} f(U_j^n) + \frac{1}{2}a_{j+1/2}(1-\lambda a_{j-1/2})S_j^n/[1+\lambda(a_{j+1/2}-a_{j-1/2})] & \text{if } a_{j+1/2} \geq 0 \\ f(U_{j+1}^n) - \frac{1}{2}a_{j+1/2}(1+\lambda a_{j+3/2})S_{j+1}^n/[1+\lambda(a_{j+3/2}-a_{j+1/2})] & \text{if } a_{j+1/2} < 0, \end{cases}$$

where

$$a_{j+1/2} = \begin{cases} (f(U_{j+1}^n) - f(U_j^n))/(U_{j+1}^n - U_j^n) & \text{if } U_j^n \neq U_{j+1}^n \\ \dfrac{df(U_j^n)}{du} & \text{if } U_j^n = U_{j+1}^n \end{cases}$$

and

$$S_j^n = m(S_j^+, S_j^-) = \begin{cases} S \cdot \min\{|S_j^+|, |S_j^-|\} & \text{if } \operatorname{sgn}(S_j^+) = \operatorname{sgn}(S_j^-) = S \\ 0 & \text{otherwise.} \end{cases}$$

The schemes differ only in their calculation of $S_j^{\pm}$:

$$\text{TVD2} : S_j^{\pm} = d_{j\pm1/2}U^n$$

$$\text{UNO2} : S_j^{\pm} = d_{j\pm1/2}U^n \pm \frac{1}{2}D_{j\pm1/2}U^n,$$

where

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$$d_{j+1/2}U = U_{j+1} - U_j \ ,$$
$$D_{j+1/2}U = m(D_jU,\ D_{j+1}U)$$

and

$$D_jU = U_{j+1} - 2U_j + U_{j-1}.$$

In Table 5.3 solutions of UNO2, TVD2 and an explicit rational scheme are presented for the constant coefficient case

$$u_t + u_x = 0, \quad u(x,0) = \sin \pi x, \quad -1 \leq x \leq 1,$$

with periodic boundary conditions at time, $t = 2$. In [18] the time - step was chosen according to $k/h = 0{,}8$ for UNO2 and TVD2, while a fixed time step $k = 0{,}001$ was used for the explicit rational scheme with $\epsilon = 0$.

TABLE 5.3 : L_1 and L_∞ error norms (× 1000) at time $t = 2$

L_∞ (× 1000)				L_1 (× 1000)		
h	UNO2	TVD2	(1,1)	UNO2	TVD2	(1,1)
0,1	7,09	81,19	10,99	8,94	67,78	14,19
0,05	1,61	34,77	9,89	2,04	20,33	12,60
0,025	0,39	14,53	9,82	0,49	5,63	12,51
0,0125	0,09	59,75	9,82	0,12	1,53	12,50

The table illustrates the approximation ability of a rational scheme on a coarse mesh. The UNO2 method performs better than the general explicit rational method. This can be explained by the specialised nature of the algorithm to solve hyperbolic equations. Figure 5.5 shows that the rational approximation at $t = 2$ adapts itself satisfactorily at the extreme turning points.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

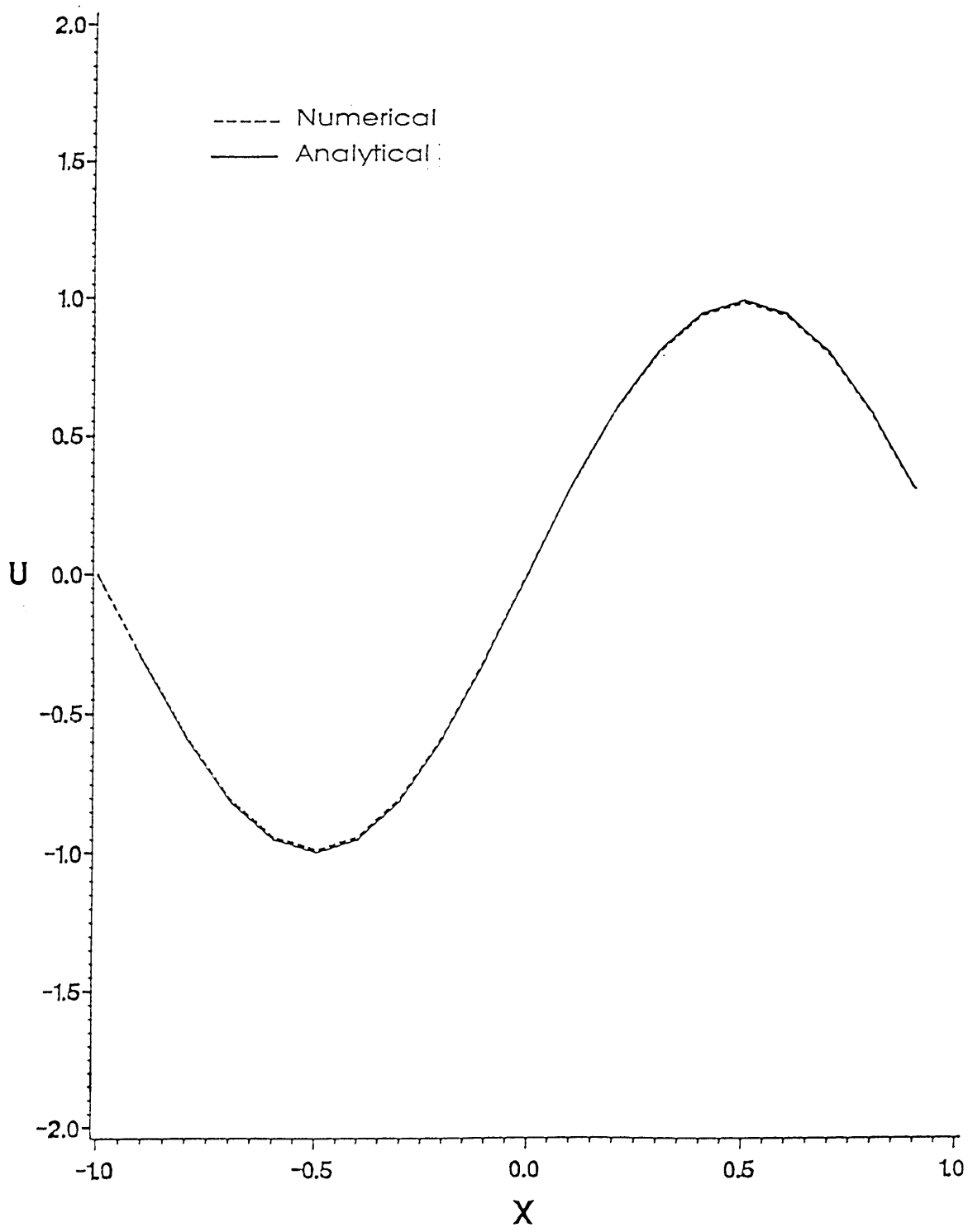

FIGURE 5.5 : (1,1) rational with h = 0,1 and $\epsilon = 0$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Next, the inviscid Burgers equation

$$u_t + \frac{1}{2}(u^2)_x = 0,\ u(x,0) = \sin \pi x,\quad -1 \leq x \leq 1,$$

with periodic boundary conditions is considered. For UNO2 and TVD2, the time step was chosen according to $k/h = 0,5$, while $k/h = 0,1$ and $\epsilon = 0$ are used for the rational scheme. Due to the periodic boundary conditions an implicit rational scheme arises. The L_∞ error and L_1 error for a refinement sequence with $N = 20$, 40 and 80 are listed in Table 5.4 for all the numerical schemes under consideration.

TABLE 5.4 : L_1 and L_∞ error norms at $t = 0,15$

L_∞ (× 1000)					
h	UNO2	TVD2	(1,1)	(2,2)	(3,3)
0,1	18,90	22,38	7,82	7,78	7,76
0,05	5,71	10,54	3,87	3,87	3,87
0,025	1,55	4,42	2,36	2,33	2,31
L_1 (× 1000)					
h	UNO2	TVD2	(1,1)	(2,2)	(3,3)
0,1	10,90	18,54	5,82	5,74	5,71
0,05	3,03	5,05	3,50	3,47	3,46
0,025	0,78	1,34	1,83	1,82	1,82

It can be seen from the table that the numerical method which couples only three adjoining nodes performs satisfactorily in comparison to the more sophisticated algorithms in [18] and [40].

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The solution develops a shock at $t = \frac{1}{\pi}$. The implicit scheme was applied with parameters $h = 0,05$, $\epsilon = 0,01$ and $k = 0,005$ and Figure 5.6 shows the numerical results at time $t = 0,35$. This result compares favourably to that of Sanders [40]. The influence of the viscosity term can be appreciated when $\epsilon = 0$ as shown in Figure 5.7.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

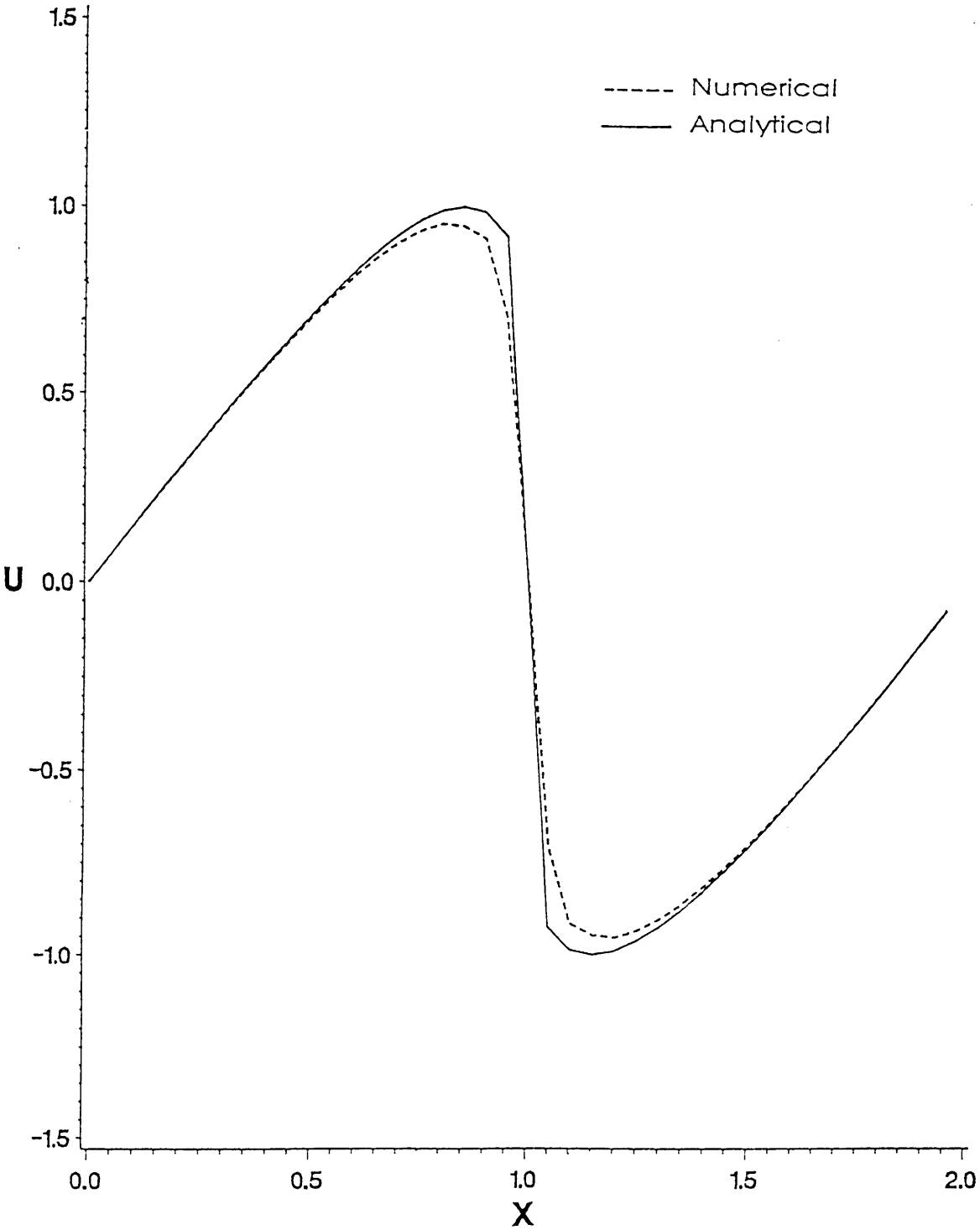

FIGURE 5.6 : (1,1) Rational approximation with $\epsilon = 0{,}01$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

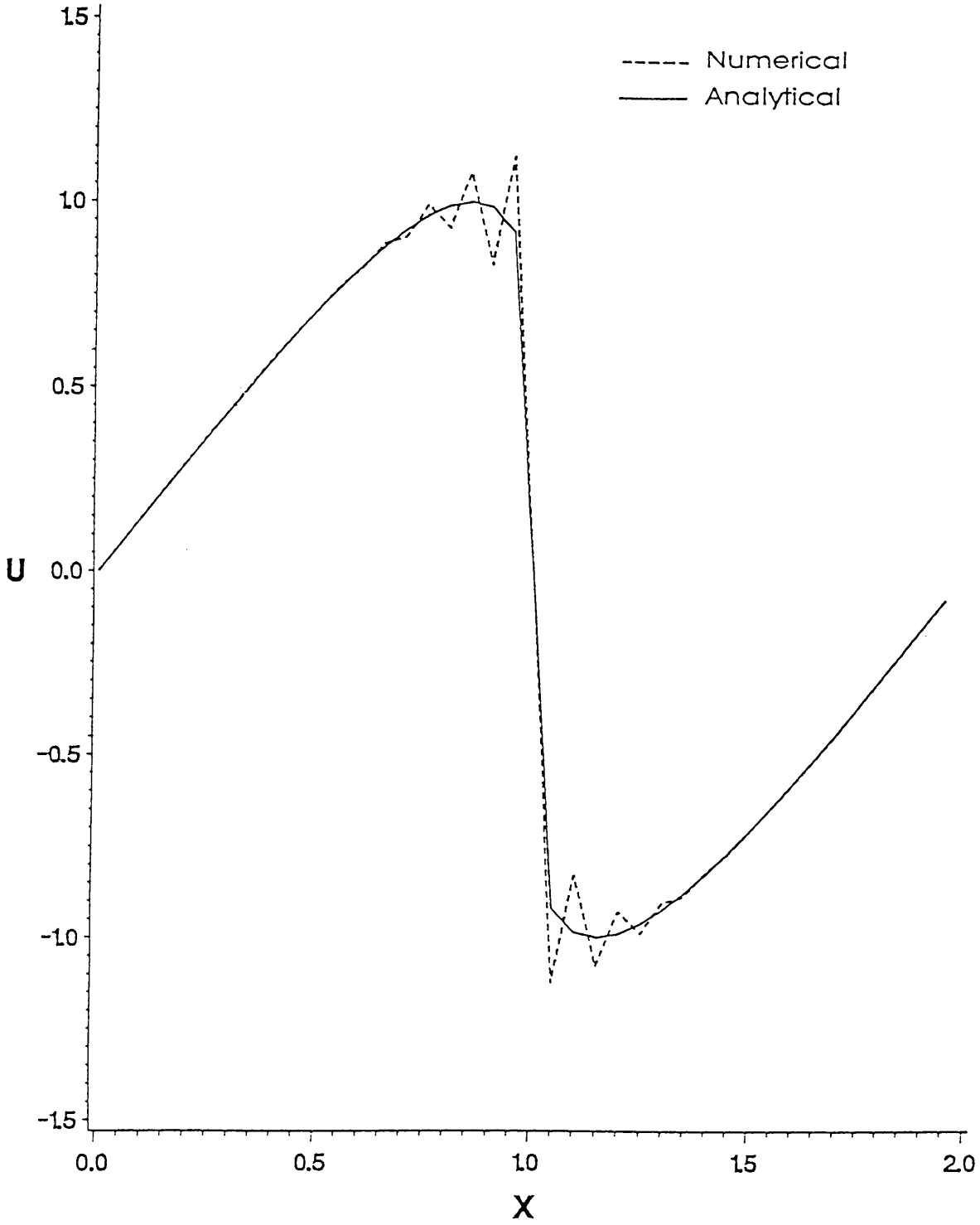

FIGURE 5.7 : (1,1) Rational approximation with $c = 0,0$

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

5.6 CONCLUSIONS

A rational approximation scheme was proposed and it was found to perform satisfactorily in comparison with existing methods. The convergence of the nonlinear scheme was established and an upper bound obtained for the added viscosity to ensure that the transition length in which the solution changes its gradient was not unrealistically enlarged.

This method is able to follow a discontinuity owing to the approximation ability of a rational function. Convection also seems to be correctly modelled by this rational scheme.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

CHAPTER 6

CONCLUSIONS

In almost all the numerical experiments conducted it was found that the rational approximation was more accurate than its polynomial counterpart. This increase in accuracy can only be attributed to the approximation ability of a rational function. To capitalise on this property a coarse computational mesh was used to obtain the rational solution, in comparison to the finer mesh of the finite difference or finite element methods using polynomials.

The shape of the rational basis function was constructed to be biased in the upstream direction. In so doing , the free parameter present in upwinded schemes, which has to be chosen in advance and has a significant effect on the accuracy of the solution, is eliminated. The investigation conducted in Chapter 3 suggested that the effect of convection was modelled satisfactorily, since the numerical scheme yielded eigenvalues with reduced imaginary components. This resulted in numerical solutions where the numerical oscillations were greatly suppressed.

Another distinct advantage of using rational basis functions in the Galerkin method was that the numerical scheme couples only three adjoining node-points in the computational mesh, giving tridiagonal matrices. This leaves only the computation of the inverse of a tridiagonal matrix for an explicit forward-difference scheme initially, or at each time step for an implicit backward-difference scheme. The numerical scheme for the Korteweg-de Vries equation has

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

five-banded matrices due to the choice of the test function which consists of a linear combination of Hermite rational basis functions. Thus, it is clear that by using rational basis functions the computational efficiency in the finite element algorithm is maintained.

The concept of using rational functions as basis functions represents a new method for numerical analysts using the finite element method to solve partial differential equations. The formulation of rational basis functions and the extension to higher-order rational basis functions are unique and this contribution makes a significant impact on the accuracy and ease of the computational algorithm which is used to obtain the numerical solution of these equations. Finally, the advantages of using rational basis functions in the finite element method are twofold; firstly they have better approximation abilities, owing to the influence of the rational function, and secondly these functions simulate upwinding and therefore satisfies an important property of fluid flow in nonlinear parabolic and hyperbolic problems.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

REFERENCES

[1] W.F. Ames, Numerical Methods for Partial Differential Equations. McGraw-Hill, New York 1966.

[2] E.D. Becker, G.F. Carey and J.T. Oden, Finite Elements An Introduction: Volume I. Prentice Hall, New Jersey 1981.

[3] J.L. Bona, V.A. Dougalis and O.A. Karakashian, Fully Discrete Galerkin Methods for the Korteweg-de Vries Equation. Comput. Math. Applic., 12A No 7 (1986) 859-884.

[4] D Braess, Nonlinear Approximation Theory. Springer-Verlag, Heidelberg 1986.

[5] A.N. Brooks and T.J.R. Hughes, Streamline Upwind/Petrov-Galerkin Formulations for Convection Dominated Flows with Particular Emphasis on the Incompressible Navier-Stokes Equations. Comput. Meths. Appl. Mech. Engrg., 32(1982) 199-259.

[6] R.L. Burden, J.D. Faires and A.C. Reynolds, Numerical Analysis. Prindle, Weber and Schmidt, Boston 1978.

[7] G.F. Carey and J.T. Oden, Finite Elements: Fluid Mechanics vol. VI. Prentice-Hall, New Jersey 1986.

[8] E.W. Cheney, Introduction to Approximation Theory. McGraw-Hill, New York 1966.

[9] I. Christie, D.F. Griffiths, A.R. Mitchell and J.M. Sanz-Serna, Product Approximation for Nonlinear Problems in the Finite Element Method. IMA J. Num. Analysis, 1(1981) 253-266.

[10] J.D. Cole, On a Quasi-linear Parabolic Equation Occurring in Aerodynamics. Quarterly of Applied Mathematics, IX No 3(1951) 225-236.

[11] A. Cuyt and L. Wuytack, Nonlinear Methods in Numerical Analysis. North-Holland, New York 1987.

[12] R.K. Dodd, J.C. Eilbeck, J.D. Gibbon and H.C. Morris, Solitons and Nonlinear Wave Equations. Academic Press, London 1982.

[13] Douglas Jr. and T.F. Russell, Numerical Methods for Convection-Dominated Diffusion Problems based on combining the Method of Characteristics with Finite Element or Finite Difference Procedures. SIAM J. Numer. Anal., 19 No 5 (1972), 871-885.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

[14] G. Fairweather, Finite Element Galerkin Methods for Differential Equations. Marcel Dekker, New York 1978.

[15] S.O. Fatunla, Nonlinear Multistep Methods for Initial Value Problems. Comput. Math. Applic., 8 No 3 (1982) 231-239.

[16] R. Frank and C.W. Ueberhuber, Iterated Defect Dorrection for the Efficient Solution of Stiff Systems of Ordinary Differential Equations. BIT 17(1977) 146-159.

[17] D.F. Griffiths and J. Lorenz, An Analysis of the Petrov-Galerkin Finite Element Method. Comput. Meths. Appl. Mech. Engrg., 14(1978) 39-64.

[18] A. Harten, On a Class of High Resolution Total-Variation-Stable Finite-Difference Schemes. SIAM J. Numer. Anal., 21 No 1 (1984) 1-23.

[19] A. Harten, J.M. Hyman and P.D. Lax, On Finite Difference Approximations and Entropy Conditions for Shocks. Comm. Pure Appl. Math., 29(1976) 297-322.

[20] A Harten and P.D. Lax, A Random Choice Finite Difference Scheme for Hyperbolic Conservation Laws. SIAM J. Numer. Anal., 18 No 2 (1981) 289-315.

[21] A. Jeffrey and T. Kakutani, Weak Nonlinear Dispersive Waves: A Discussion Centered Around the Korteweg-de Vries Equation. SIAM Review, 14 No 4 (1972) 582-643.

[22] C. Johnson and A. Szepessy, On the Convergence of a Finite Element Method for a Nonlinear Hyperbolic Conservation Law. Math. Comp., 49 No 187 (1987) 427-444.

[23] E. Kreyszig, Introductory Functional Analysis with Applications. John Wiley, New York 1978.

[24] G.G. Lorentz, Approximation of Functions. Holt, Rinehart and Winston, New York 1966.

[25] Y.L. Luke, W. Fair and J. Wimp, Predictor-Corrector Formulas based on Rational Interpolants. Comput. Math. Applic., 1(1975) 3-112.

[26] A.R. Mitchell and D.F. Griffiths, Semidiscrete Generalized Galerkin Methods for Time-dependent Conduction-Convection Problems. The Mathematics of Finite Elements and Applications III 1978, Academic Press, New York 1979.

[27] A.R. Mitchell and D.F. Griffiths, The Finite Difference Method in Partial Differential Equations. John Wiley, New York 1980.

[28] A.R. Mitchell and R. Wait, The Finite Element Method in Partial Differential Equations. John Wiley, London (1977).

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

[29] M.J. NG-Stynes, E. O'Riordan and M. Stynes, Numerical methods for Time-dependent Convection-Diffusion Equations. J. Comp. Appl. Maths., 21(1988) 289-310.

[30] J.T. Oden and G.F. Carey, Finite Elements: Mathematical Aspects Volume IV. Prentice-Hall, New Jersey 1983.

[31] V.A. Popov, Contributions of Geza Freud to the Theory of Rational Approximation of Functions. J. Approx. Theory, 46(1986) 111-118.

[32] M.J.D. Powell, Approximation Theory and Methods. Cambridge University Press, London 1981.

[33] R. Peyret and T.D. Taylor, Computational Methods for Fluid Flow. Springer-Verlag, New York 1983.

[34] P.M. Prenter, Splines and Variational Methods. John Wiley, New York 1975.

[35] R.D. Richtmyer and K.W. Morton, Difference Methods for Initial-Value Problems. John Wiley, New York 1967.

[36] P.J. Roache, Computational Fluid Dynamics. Hermosa, Albuquerque 1972.

[37] E.E. Rosinger, Nonlinear-Equivalence, Reduction of PDEs to ODEs and Fast Convergent Numerical Methods. Pitman, London 1982.

[38] H.L. Royden, Real Analysis : Second Edition. MacMillan Publishing Co., New York 1968.

[39] W. Rudin, Real and Complex Analysis. McGraw-Hill, New York 1974.

[40] R. Sanders, A Third-Order Accurate Variation Nonexpansive Difference Scheme for Single Nonlinear Conservation Laws. Math. Comp., 51 No 184(1988) 535-558.

[41] J.M. Sanz-Serna, An Explicit Finite-Difference Scheme with Exact Conservation Properties. J. Comput. Phys., 47(1982) 199-210.

[42] J.M. Sanz-Serna and I. Christie, Petrov-Galerkin Methods for Nonlinear Dispersive Waves. J. Comput. Phys., 39(1981) 94-102.

[43] J. Smoller, Shock Waves and Reaction-Diffusion Equations. Springer-Verlag, New York 1983.

[44] G.A. Sod, A Survey of Several Difference Methods for Systems of Nonlinear Hyperbolic Conservation Laws. J. Comput. Phys., 27(1978) 1-31.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

[45] G.A. Sod, Numerical Methods in Fluid Dynamics: Initial and Initial Boundary-Value Problems. Cambridge University Press, New York 1985.

[46] T.R. Taha and M.J. Ablowitz, Analytical and Numerical Aspects of Certain Nonlinear Evolution Equations. III. Numerical, Korteweg-de Vries Equation. J. Comput. Phys., 55(1984) 231-253.

[47] A. van Niekerk and F.D. van Niekerk, An Explicit Finite Element Method for Convection-Diffusion Equations Using Rational Basis Functions. Comput. Math. Applic., to appear.

[48] F.D. van Niekerk, Eindige-Element-Metodes vir Tydafhanklike Parsiele Differensiaalvergelykings. D.Sc.-Thesis, University of Pretoria, Republic of South Africa 1981.

[49] F.D. van Niekerk, Non-linear One Step Methods for Initial Value Problems. Comput. Math. Applic., 13(1987) 367-371.

[50] F.D. van Niekerk and A. van Niekerk, A Galerkin Method Using Rational Basis Functions. Comput. Math. Applic., 17 No 7 (1989) 1085-1093.

[51] F.D. van Niekerk and A. van Niekerk, A Galerkin Method With Rational Basis Functions for Burgers Equation. Comput. Math. Applic., to appear.

[52] J.P. Villa, High-Order Schemes and Entropy Condition for Non-linear Hyperbolic Systems of Conservation Laws. Math. Comp., 50(1988) 53-73.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The development of rational basis functions for the finite element method

by

Andre van Niekerk

Promotor : Prof. F.D. van Niekerk

Department : Mathematics and Applied Mathematics

Degree : Ph.D.

SUMMARY

This thesis is concerned with the development of rational basis functions for the finite element method to obtain the numerical solution of time-dependent partial differential equations. The concept of a rational basis function is introduced and it is found to be biased in the upstream direction. This, together with the fact that a rational function possesses better approximation abilities than a polynomial, motivated the application of rational basis functions in the Galerkin method to solve convection-dominated phenomena. This method gives rise to a rational difference scheme that approximates steep gradients and discontinuities satisfactorily without a stringent mesh refinement. The method is extended to higher-order rational basis functions and continuous Hermite rational basis functions with continuous first derivatives.

The investigation of the new method is mainly of a numerical and experimental nature. If the analytical solutions of the problems considered are available they are compared to the numerical solutions.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

If the solutions are not available, the numerical results are compared to the solutions obtained by existing numerical methods.

The rational method is applied to:

(i) a stiff ordinary differential equation,

(ii) the convection-diffusion equation with Dirichlet, periodic and Neumann boundary conditions,

(iii) the Korteweg-de Vries equation, and

(iv) linear and nonlinear hyperbolic equations.

Convergence, consistency and stability properties of the rational difference schemes are investigated.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Die ontwikkeling van rasionale basisfunksies vir die eindige-element-metode

deur

Andre van Niekerk

Promotor : Prof. F.D. van Niekerk

Departement : Wiskunde en Toegepaste Wiskunde

Graad : Ph.D.

SAMEVATTING

Hierdie proefskrif verteenwoordig 'n studie van die ontwikkeling van rasionale basisfunksies vir die eindige-element-metode om die numeriese oplossing van tydafhanklike parsiële differensiaalvergelykings te verkry. Die begrip van 'n rasionale basisfunksie word geformuleer, en daar is gevind dat die funksie na die stroomopwaartse rigting geweeg is. Hierdie neiging, en die feit dat 'n rasionale funksie beter benaderingseienskappe as 'n polinoom besit, het gelei tot die gedagte om rasionale basisfunksies in die Galerkin-metode te gebruik om die verskynsel van konveksie op te los. Die metode gee aanleiding tot 'n rasionale verskilskema wat in staat is om steil hellings of diskontinuïteite in 'n oplossing te hanteer sonder om die oplossingsrooster aansienlik te verfyn. Die metode is ook uitgebrei na hoër orde rasionale basisfunksies en kontinue Hermitiese rasionale basisfunksies met kontinue eerste afgeleides.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Die ondersoek van die nuut geformuleerde metode is hoofsaaklik numeries en eksperimenteel van aard. Indien analitiese oplossings van die probleme onder beskouing wel bestaan, word dit gebruik om die benaderde oplossing numeries te toets. Indien die oplossings nie bestaan nie, word die resultate numeries vergelyk met oplossings wat deur ander numeriese metodes verkry is.

Die rasionale metode word toegepas op:

(i) 'n stram gewone differensiaalvergelyking,

(ii) 'n konveksie-diffusie-vergelyking met periodiese, Dirichlet- en Neumann-randvoorwaardes,

(iii) die Korteweg-de Vries-vergelyking, en

(iv) lineêre en nie-lineêre hiperboliese vergelykings.

Konvergensie, steekhoudendheid en stabiliteitseienskappe van die rasionale verskilskemas word ondersoek vir die gelineariseerde vergelykings.

Digitised by the Department of Library Services in support of open access to information, University of Pretoria, 2022