ASPECTS OF INTERVAL ANALYSIS APPLIED TO INITIAL-VALUE PROBLEMS FOR ORDINARY DIFFERENTIAL EQUATIONS AND HYPERBOLIC PARTIAL DIFFERENTIAL EQUATIONS

by

ROUMEN ANGUELOV ANGUELOV

submitted in accordance with the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in the subject

MATHEMATICS

at the

UNIVERSITY OF SOUTH AFRICA

PROMOTER: PROF N T BISHOP

JOINT PROMOTER: PROF C J WRIGHT

SEPTEMBER 1998

Student number: 900-129-8

Declaration

I declare that

ASPECTS OF INTERVAL ANALYSIS APPLIED TO INITIAL VALUE PROBLEMS FOR ORDINARY DIFFERENTIAL EQUATIONS AND HYPERBOLIC PARTIAL DIFFERENTIAL EQUATIONS

is my own work and that all sources that I have used or quoted have been indicated and acknowledged by means of complete references.

Signature:  Date: 25-03-1999

R A Anguelov

Acknowledgements

My deepest thanks to the following people:

Prof N Bishop and Prof C J Wright, promoters of the thesis, for their guidance and support.

Prof S Markov, who initiated me in this area of research and with whom we had a fruitful collaboration for many years.

My colleagues and friends from Vista University for their assistance and understanding.

Members of the Interval Community around the world who provided a framework of communication and databases on the web which is of great help to anyone conducting research in this area.

My wife Valentina and my son Bogomil for their love, patience and support.

ASPECTS OF INTERVAL ANALYSIS APPLIED TO INITIAL-VALUE PROBLEMS FOR ORDINARY DIFFERENTIAL EQUATIONS AND HYPERBOLIC PARTIAL DIFFERENTIAL EQUATIONS

by R A Anguelov
Degree: Doctor of Philosophy
Subject: Mathematics
Promoter: Prof N Bishop
Joint Promoter: Prof C J Wright

Summary

Interval analysis is an essential tool in the construction of validated numerical solutions of Initial Value Problems (IVP) for Ordinary (ODE) and Partial (PDE) Differential Equations. A validated solution typically consists of guaranteed lower and upper bounds for the exact solution or set of exact solutions in the case of uncertain data, i.e. it is an interval function (enclosure) containing all solutions of the problem.

IVP for ODE: The central point of discussion is the wrapping effect. A new concept of wrapping function is introduced and applied in studying this effect. It is proved that the wrapping function is the limit of the enclosures produced by any method of certain type (propagate and wrap type). Then, the wrapping effect can be quantified as the difference between the wrapping function and the optimal interval enclosure of the solution set (or some norm of it). The problems with no wrapping effect are characterized as problems for which the wrapping function equals the optimal interval enclosure. A sufficient condition for no wrapping effect is that there exist a linear transformation, preserving the intervals, which reduces the right-hand side of the system of ODE to a quasi-isotone function. This condition is also necessary for linear problems and "near" necessary in the general case.

Hyperbolic PDE: The Initial Value Problem with periodic boundary conditions for the wave equation is considered. It is proved that under certain conditions the problem is an operator equation with an operator of monotone type. Using the established monotone properties, an interval (validated) method for numerical solution of the problem is proposed. The solution is obtained step by step in the time dimension as a Fourier series of the space variable and a polynomial of the time variable. The numerical implementation involves computations in Fourier and Taylor functoids. Propagation of discontinuous waves is a serious problem when a Fourier series is used (Gibbs phenomenon, etc.). We propose the combined use of periodic splines and Fourier series for representing discontinuous functions and a method for propagating discontinuous waves. The numerical implementation involves computations in a Fourier hyper functoid.

Key terms: Validated methods, Interval methods, Enclosure methods, Ordinary differential equations, Partial differential equations, Wrapping effect, Wrapping function, Functoid, Fourier hyper functoid.

Contents

Chapter 1

Introduction

Solving differential equations is one of the main topics in Numerical Analysis and a large variety of numerical methods has been developed, e.g. single- and multi-step methods for Initial Value Problems (IVP) of Ordinary Differential Equations (ODE), difference schemes and finite elements for Partial Differential Equations (PDE). Typically, if certain smoothness conditions are satisfied, these methods produce in theory sufficiently good approximations for suitable values of the parameters of the method (e.g. step size, number of terms in a series, etc.). However, in practice, one seldom knows these values. Even by controlling the parameters of the method, the global accuracy is not really controlled. Although methods like those mentioned above are robust and reliable for most applications, usually it is not difficult to find examples where they return very inaccurate results without any warning to the user. This is impressively demonstrated by Adams *et al.* in [1], [3] with examples of IVP for ODE in which the computed numerical solution does not even reflect the qualitative stability of the exact solution.

In contrast, validated methods produce verified numerical results which carry within themselves assurances of their quality. More precisely, a validated method has two major characteristics:

(1.1) it verifies the existence of a unique solution to the problem and

(1.2) produces guaranteed bounds for this solution.

The bounds produced by a validated method are guaranteed in the sense that all sorts of errors are taken into account (e.g. truncation, rounding, etc.) so that the numerical result has no conditions attached to it and does not require any further analysis. The bounds are indeed bounds of the exact solution.

A validated method may fail in verifying the existence or may produce bounds that are unacceptably wide. Then the user is informed accordingly. However, a validated method never yields a false result.

Apart from the obvious comfort in working with a validated numerical solution in any mathematical application where the exact solution to a problem is not available, there

are situations where validated numerical solutions are particularly desired or needed. We will mention here three such situations.

- Critical computations: when the computational result may be critical to the safety of a system. For example, in the conditions of use of LANCELOT, one of the leading packages for large-scale nonlinear optimization, it is stated that "it should not be relied on as a basis to solve a problem whose incorrect solution could result in injury to a person or property" [24]. Software which implements correctly a validated method need not be accompanied by such a condition.

- Uncertain Data: Most differential equations arising from models in applications in science or engineering typically contain parameters whose values are only approximately known. Since validated methods use interval operations, they can produce bounds guaranteed to enclose solutions arising from any combination of parameter values in the range of interest. Implementing a validated method can be much faster and more informative then repeated simulation runs using a standard method.

- Proving theorems: A validated numerical solution of a mathematical problem can be used in deriving mathematical statements about the solution(s). Such applications are presented for example in [2].

Validated numerical methods have not been very popular in the past mainly because the implementation typically requires more computational time than standard methods. Now, however, when computational resources are readily available, it seems natural to shift the burden of determining the reliability of a numerical solution from the user to the computer.

Validated methods are designed and implemented using interval arithmetic and interval functions. The bounds mentioned in (1.2) are the end points of an interval enclosing the exact solution. That is why often validated methods are referred to as interval methods or enclosure methods.

In most of the applications it is not enough to produce bounds or enclosure of the solution but also

$$\text{to produce close bounds (tight enclosure) of this solution} \tag{1.3}$$

satisfying some acceptable tolerance. In [79], when characterizing the purpose of validated scientific computations, Neumaier specifies that (1.1), (1.2) and (1.3) are three separate issues associated with such computations. A priori estimates on closeness of the bounds produced by a validated method for a particular class of problems characterize, in general, the quality of the method indicating its area of applicability and the expected accuracy. In obtaining close bounds one has to combat the truncation error of the method, the rounding errors, as well as problems related to the type of selected enclosures (intervals) like the wrapping effect.

In the thesis we consider some aspects of the construction of validated solutions of Ordinary Differential Equations (ODE) and Hyperbolic Partial Differential Equations (HPDE).

The idea of solving the Initial Value Problem (IVP) for ODE by constructing lower and upper bounds was proposed by Chaplygin [21], [22] in 1919, but Moore [64] in 1965 formulated enclosure methods using interval arithmetic in the description and implementation. Since then a variety of methods has been developed (see [86] and [78] for recent surveys). In chapter 3, which is devoted to IVP for ODE, we mainly concentrate on studying the wrapping effect associated with the construction of interval enclosures of the set of solutions of IVP with interval initial conditions and the implications concerning the convergence of the enclosures produced by a certain type of method. The wrapping effect is widely discussed in the literature and methods to combat it have been proposed [67], [55], [37], [59], [81], [78]. We study the wrapping effect using the new concept of wrapping function [9] which is the key to understanding the behaviour of interval enclosures when the wrapping effect is in force. For example, because of the wrapping effect the enclosures often do not converge to the optimal interval enclosure of the solution. In that case, what do they converge to? We prove that the limit of the interval enclosures produced by a certain type of method is the wrapping function associated with the particular IVP. In this way, the wrapping effect can be quantified as a certain norm of the difference between the wrapping function and the optimal interval enclosure of the solution set. There is no wrapping effect if and only if the wrapping function equals the optimal enclosure. These results lead not only to a better understanding of what the wrapping effect is, but also to a complete characterization of the problems where no wrapping effect occurs [11] and therefore no complicated procedures (e.g.[37], [59]) need be applied.

Operators of monotone type [23] play an important part in the construction of bounds for the solutions of differential equations. The conditions for the operator of the IVP for ODE to be an operator of monotone type have been known for a long time [21], [68], [92]. It is also well know that there is no wrapping effect when the operator of the IVP for ODE is an operator of monotone type. In chapter 3 we further establish that "essentially" all problems with no wrapping effect are those that are either problems with operator of monotone type or can be transformed into such problems by an interval preserving transformation.

Validated solutions of partial differential equations is a relatively new area with most of the important results obtained only during the past decade [34], [50], [69]-[77]. The proposed validated methods use the fixed point theorem and computable error estimates in obtaining validated enclosures and are mainly applied to problems with point (non-interval) initial conditions. Our approach is based on establishing monotone properties of the problem and using them in designing methods producing validated enclosures. The main advantage of this approach is that it works equally well (same convergence properties) with point and interval initial conditions. Naturally, there is no wrapping effect. In chapter 4 we first find conditions providing for the the operator of the periodic

IVP for HPDE to be an operator of monotone type. Since the initial condition of the problem involves both the solution and its derivative, the established monotone property is not directly applicable to constructing enclosures step by step in the time dimension. For that reason a different monotone property, which involves the time derivative of the functions, is derived. We use the established monotone properties in constructing a method producing interval enclosures for the solution (in the case of a point initial condition) or the set of solutions (in the case of interval initial condition) of the problem. The bounds for the solutions are represented as a Fourier series of the space variable with coefficients that are polynomials of the time variable. This results in computations in the Fourier functoid and the Taylor functoid. In addition to the roundings discussed in [49], we use an easily computable self-adaptive error estimate for rounding of functions with infinite Fourier series.

The application of the method discussed in chapter 4 is reduced to problems with data functions which have at least one derivative with integrable square. This is due to the fact that otherwise no sharp bounds for this functions can be obtained using a Fourier series (Gibbs phenomenon). Furthermore, when the data functions are not smooth enough, sharp bounds can only be obtain when using a large number of terms in the Fourier series. Considering problems with not smooth (or not smooth enough) data functions is important because

- such problems are very common in applications
- the periodic problem, and the transformation of an initial boundary value problem into a periodic initial value problem, often lead to discontinuities at the places where the boundary conditions are initially given.

In order to be able to deal with discontinuities of the data functions or their derivatives we consider in chapter 5 the Fourier hyper functoid [48]. The Fourier hyper functoid is a finite space with basis which, in addition to the standard Fourier basis, includes infinite Fourier series. The basis chosen in chapter 5 includes infinite series of periodic splines. Since we work with these splines explicitly (not with their series) we refer to these approximations as Spline-Fourier approximation. Explicit formulas for computations related to the solution of the wave equation in this hyper functoid are derived. It is demonstrated that the use of the hyper functoid not only enlarges the area of applicability of the method described in chapter 4, but also produces high quality results with smaller computational effort.

Chapter 2

Preliminaries

2.1 Interval spaces.

Interval structures were first introduced by Sunaga [89] but the real development in this area took place after interval arithmetic and interval analysis were introduced in practical applications, showing that intervals are a powerful tool for the design of a new kind of numerical method. The beginning of this development is associated with the name of R Moore [66], [67].

In this section we will briefly introduce some interval spaces and discuss some basic facts in a way similar to [7]. However, we will follow the approach in [14], [60] and define first an interval space over a vector lattice and then consider the real and functional interval spaces as particular cases.

2.1.1 Abstract Interval Space.

Let $\Re$ be a real continuous vector lattice with the operations addition (+) and multiplication by a real number (.) and a partial ordering $\leq$ [47], [6]. This means that

- $\Re$ is a linear space with operations + and . over the field of real numbers $\mathcal{R}$;
- $\Re$ is a lattice with a partial ordering $\leq$;
- $u \leq v \Longrightarrow u + w \leq v + w \quad , \quad u, v, w \in \Re$;
- $u \leq v \Longrightarrow \lambda u \leq \lambda v \quad , \quad u, v \in \Re,\ \lambda \in \mathcal{R}^{+}$;
- $\sup\{U\}$ and $\inf\{U\}$ are defined for every bounded subset U of $\Re$.

Modulus $|.|$ is defined in a vector lattice as $|x| = \sup\{x, 0\} - \inf\{x, 0\}, \ \ x \in \Re$.
Let $\underline{x}, \overline{x} \in \Re$ and $\underline{x} \leq \overline{x}$. The set

$$[\underline{x}, \overline{x}] = \{x \in \Re : \underline{x} \leq x \leq \overline{x}\}$$

is called an interval.

By $\mathcal{IR}$ we denote the set of all intervals

$$\mathcal{IR} = \{X = [\underline{x}, \overline{x}] : \underline{x}, \overline{x} \in \Re, \underline{x} \leq \overline{x}\}$$

For every $x \in \Re$ the interval $[x, x]$ is identified with x. In this way $\Re \subset \mathcal{IR}$. The elements of $\Re$ will be called point intervals. Let U be a bounded subset of $\Re$. Optimal interval enclosure of U is defined as

$$[U] = [\inf(U), \sup(U)]$$

The optimal interval enclosure $[U]$ is the tightest interval enclosure of U in the sense that for any other interval Y, $U \subset Y$ implies $[U] \subset Y$. The intersection of two intervals $X, Y \in \mathcal{IR}$ (considered as subsets of $\Re$), if not empty, is also an interval

$$X \bigcap Y = [\sup\{\underline{x}, \underline{y}\}, \inf\{\overline{x}, \overline{y}\}]$$

while for the union of intervals this is not true. The operation joint ($\vee$) is introduced as follows. If $X, Y \in \mathcal{IR}$ then

$$X \vee Y = [X \bigcup Y] \ .$$

In $\mathcal{IR}$ we consider the following operations

- addition: $X + Y = [\underline{x} + \underline{y}, \overline{x} + \overline{y}]$, $X, Y \in \mathcal{IR}$,
- subtraction: $X \stackrel{-}{} Y = [(\underline{x} - \underline{y}) \vee (\overline{x} - \overline{y})]$, $X, Y \in \mathcal{IR}$,
- multiplication by a scalar: $\alpha X = (\alpha \underline{x}) \vee (\alpha \overline{x})$, $X \in \mathcal{IR}$, $\alpha \in \mathcal{R}$.

The operation subtraction defined above is known as nonstandard or inner subtraction of intervals and we use for it the same notation as in [61], [62]. It can not be generated by the other two operations if we stay within the realm of the elements of $\mathcal{IR}$. The plain subtraction sign $-$ will be used (as in most of the literature on intervals) for the standard subtraction defined by

$$X - Y = X + (-1)Y \ .$$

In a similar way inner addition can be defined as

$$X \stackrel{-}{+} Y = X \stackrel{-}{-} (-1)Y$$

Let us note that

$$A \stackrel{-}{+} B \subset A + B \ ,$$
$$A \stackrel{-}{-} B \subset A - B \ , \quad A, B \in \mathcal{IR}$$

Definition 2.1 *The set $\mathcal{IR}$ with operations $+$, $\stackrel{-}{}$ and . is called interval space generated by the vector lattice $\Re$.*

Let $A, B, C \in I\mathbb{R}$ and $\alpha, \beta \in \mathbb{R}$. The following properties hold true

$$(A + B) + C = A + (B + C) \tag{2.1}$$
$$A + B = B + A \tag{2.2}$$
$$A + 0 = A \tag{2.3}$$
$$\text{If } A + C = B + C \text{ then } A = B \tag{2.4}$$
$$\text{If } \alpha\beta \geq 0 \text{ then } (\alpha + \beta)A = \alpha A + \beta A \text{ and } (\alpha - \beta)A = \alpha A \stackrel{-}{} \beta A \tag{2.5}$$
$$\alpha(A + B) = \alpha A + \alpha B\ , \quad \alpha(A \stackrel{-}{} B) = \alpha A \stackrel{-}{} \alpha B \tag{2.6}$$
$$(\alpha\beta)A = \alpha(\beta A) \tag{2.7}$$
$$1.A = A \tag{2.8}$$

Let us note that $I\mathbb{R}$ is not a linear space with regard to the operations $+$ and $.$ since the non point intervals have no opposite elements and the distributive law $(\alpha+\beta)A = \alpha A+\beta A$ is satisfied only in some cases. The condition for the existence of an opposite element is replaced by a cancellation law of the form (2.4), and a distributive law of the form (2.5) is satisfied. Such a space is called quasi linear. The concept of quasi linear space was introduced by Mayer [63] and further developed in [85] and [62]. The inner operations are needed precisely because the interval space is not strictly linear. For example, the inner subtraction satisfies the following properties

$$A \stackrel{-}{} B = 0 \iff A = B$$
$$(A + C) \stackrel{-}{} (B + C) = A \stackrel{-}{} B$$

which are not true in general for the standard subtraction.

Many of the identities in a linear space are satisfied in an interval space only as inclusions, e.g.

$$A \subset (A \stackrel{-}{} B) + B \tag{2.9}$$
$$(A \stackrel{-}{} B) \subset (A \stackrel{-}{} C) + (C \stackrel{-}{} B) \tag{2.10}$$

In $I\mathbb{R}$ we also consider the operators

- modulus: $|.| : I\mathbb{R} \mapsto \mathbb{R}$ defined by

$$|X| = \sup\{|\underline{x}|, |\overline{x}|\}\ , \quad X \in I\mathbb{R}$$

- width: $w : I\mathbb{R} \mapsto \mathbb{R}$ defined by

$$w(X) = \overline{x} - \underline{x}\ , \quad X \in I\mathbb{R}$$

The range of both modulus and width is the positive cone $\Re^+$ with respect to the ordering $\leq$. Let $X, Y \in I\Re$ and $\alpha \in \mathcal{R}$. We have

$$\begin{aligned} &|X| \geq 0 \ , \quad |X| = 0 \Longleftrightarrow X = 0 \\ &|X + Y| \leq |X| + |Y| \ , \quad |X -^{-} Y| \geq \big| \, |X| - |Y| \, \big| \\ &|\alpha X| = |\alpha||X| \end{aligned} \tag{2.11}$$

$$\begin{aligned} &w(X) \geq 0 \ , \quad w(X) = 0 \Longleftrightarrow X \in \Re \\ &w(X + Y) = w(X) + w(Y) \ , \quad w(X -^{-} Y) = |w(X) - w(Y)| \\ &w(\alpha X) = |\alpha| w(X) \end{aligned} \tag{2.12}$$

If a norm $\|.\|$ is defined in the linear space $\Re$, a metric in $I\Re$ can be introduced by defining the distance between intervals $A, B \in I\Re$ as

$$\wp(A, B) = \| |A -^{-} B| \| \tag{2.13}$$

This definition satisfies the axioms for distance:

$$\begin{aligned} &(i) \quad \wp(A, B) \geq 0 \text{ and } \wp(A, B) = 0 \Leftrightarrow A = B \ , \ A, B \in I\Re \\ &(ii) \quad \wp(A, B) = \wp(B, A) \ , \ A, B \in I\Re \\ &(iii) \quad \wp(A, B) \leq \wp(A, C) + \wp(C, B) \ , \ A, B, C \in I\Re \end{aligned}$$

Conditions (i) and (ii) follow immediately from the definitions of $\wp$ and $-^{-}$. We will prove (iii). Using inclusion (2.10) we have

$$|A -^{-} B| \ \leq \ |(A -^{-} C) + (C -^{-} B)| \leq |(A -^{-} C)| + |(C -^{-} B)|$$

Therefore

$$\begin{aligned} \wp(A, B) &= \| |A -^{-} B| \| \ \leq \ \| |(A -^{-} C)| + |(C -^{-} B)| \| \\ &\leq \| |(A -^{-} C)| \| + \| |(C -^{-} B)| \| \\ &= \wp(A, C) + \wp(C, B) \end{aligned}$$

2.1.2 The Real Interval Space $I\mathcal{R}$.

Let $\Re = \mathcal{R}$. Then the interval space $I\mathcal{R}$ consists of all compact intervals on the real line with operations as defined in the previous section.

Usually in $I\mathcal{R}$ the operation multiplication and division are also defined

$$XY = \{xy : x \in X, y \in Y\} = \left[\inf_{x \in X, y \in Y}(xy), \sup_{x \in X, y \in Y}(xy) \right] \ , \ X, Y \in I\mathcal{R} \tag{2.14}$$

$$X \div Y = \left\{ \frac{x}{y} : x \in X, y \in Y \right\} = \left[\inf_{x \in X, y \in Y} \left(\frac{x}{y} \right), \sup_{x \in X, y \in Y} \left(\frac{x}{y} \right) \right], \ X, Y \in I\mathcal{R}, \ 0 \notin Y \tag{2.15}$$

These operations are important in the computation of enclosures of rational expressions.

The modulus in $\mathcal{R}$ is also a norm in $\mathcal{R}$. Therefore the distance between intervals is

$$\wp(X, Y) = |X -^{-} Y|$$

and coincides with the Hausdorff distance [43] between intervals considered as subsets of $\mathcal{R}$.

2.1.3 The n-Dimensional Interval Space $\mathcal{IR}^n$.

This is the interval space over the n-dimensional linear space $\mathcal{R}^n$. The elements of $\mathcal{IR}^n$ are hyper rectangles of the form

$$\begin{aligned} X &= [\underline{x}, \overline{x}] = \{x \in \mathcal{R}^n : \underline{x} \leq x \leq \overline{x}\} \\ &= \{(x_1, x_2, \ldots, x_n)^T : x_i \in \mathcal{R}, x_i \in [\underline{x}_i, \overline{x}_i]\} \end{aligned}$$

where $\underline{x} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)^T$ and $\overline{x} = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n)^T$ are vectors in $\mathcal{R}^n$ such that $\underline{x}_i \leq \overline{x}_i$, $i = 1, 2, \ldots, n$.

The intervals $X_i = [\underline{x}_i, \overline{x}_i]$ are elements of $\mathcal{IR}$. Since X is a Cartesian product of $X_1, X_2, \ldots, X_n$, it is also represented in the form

$$X = (X_1, X_2, \ldots, X^n)^T$$

It is easy to see that in $\mathcal{IR}^n$ the operations and operators defined in an interval space can be represented using the corresponding operations and operators in $\mathcal{IR}$ coordinate-wise. Let $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_n)$ be intervals in $\mathcal{IR}^n$ and $\alpha \in \mathcal{R}$. We have

$$\begin{aligned} &X + Y = (X_1 + Y_1, X_2 + Y_2, \ldots, X_n + Y_n)^T \\ &X -^{-} Y = (X_1 -^{-} Y_1, X_2 -^{-} Y_2, \ldots, X_n -^{-} Y_n)^T \\ &\alpha X = (\alpha X_1, \alpha X_2, \ldots, \alpha X_n)^T \\ &|X| = (|X_1|, |X_2|, \ldots, |X_n|)^T \\ &w(X) = (w(X_1), w(X_2), \ldots, w(X_n))^T \end{aligned}$$

Let $\mathcal{A} = (\alpha_{ij})$ be a real $n \times n$ matrix and $A = (A_1, A_2, \ldots, A_n)^T \in \mathcal{IR}^n$. The product $\mathcal{A}A$ is defined by

$$\mathcal{A}A = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn} \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ \ldots \\ A_n \end{pmatrix} = \begin{pmatrix} \alpha_{11}A_1 + \alpha_{12}A_2 + \ldots + \alpha_{1n}A_n \\ \alpha_{21}A_1 + \alpha_{22}A_2 + \ldots + \alpha_{2n}A_n \\ \ldots \quad \ldots \quad \ldots \quad \ldots \\ \alpha_{n1}A_1 + \alpha_{n2}A_2 + \ldots + \alpha_{nn}A_n \end{pmatrix}$$

Some of the properties of the matrix multiplication are listed below

$$\begin{aligned}\boldsymbol{\mathcal{A}}(A+B) &= \boldsymbol{\mathcal{A}}A + \boldsymbol{\mathcal{A}}B \\ \boldsymbol{\mathcal{A}}(\beta A) &= \beta(\boldsymbol{\mathcal{A}}A) \\ w(\boldsymbol{\mathcal{A}}A) &= |\boldsymbol{\mathcal{A}}|w(A)\end{aligned}$$

where $|\boldsymbol{\mathcal{A}}| = (|\alpha_{ij}|)$.

For every norm $\|.\|$ in $\mathcal{R}^n$ the function $\wp(A,B) = \|\|A -^- B\|\|$ defines a distance in $\mathcal{IR}^n$. However, if the norm in $\mathcal{R}^n$ is the maximum norm

$$\|a\| = \max\{|a_1|, |a_2|, \ldots, |a_n|\} \ , \quad a \in \mathcal{R}^n$$

the distance induced in the interval space $\mathcal{IR}^n$ by this norm is the Hausdorff distance. We will prove this below.

The Hausdorff distance between two sets A and B is defined by

$$\wp_h(A,B) = \max\{\sup_{a\in A}\inf_{b\in B}\|a-b\|, \sup_{b\in B}\inf_{a\in A}\|a-b\|\}$$

The unit ball relative to the maximum norm is

$$U = \{x \in \mathcal{R}^n : \|x\| \leq 1\} = ([-1,1],[-1,1],\ldots,[-1,1])^T = [-e,e]$$

where $e = (1,1,\ldots,1)$. Using the fact that the unit ball U is an element of $\mathcal{IR}^n$ we can define Hausdorff distance between $A, B \in \mathcal{IR}^n$ in the following equivalent way

$$\wp_h(A,B) = \inf\{\alpha \in R^+ : A \subset B + \alpha U \ \text{ and } \ B \subset A + \alpha U\} \tag{2.16}$$

We also have

$$C \subset \|C\|U \ , \quad C \in \mathcal{IR}^n$$

From the above inclusion and (2.9) we obtain

$$\begin{aligned}A &\subset (A -^- B) + B \subset B + \|\|A -^- B\|\|U \\ B &\subset (B -^- A) + A \subset A + \|\|A -^- B\|\|U\end{aligned}$$

Therefore from the definition (2.16) it follows that

$$\wp_h(A,B) \leq \|\|A -^- B\|\| \tag{2.17}$$

From the inclusions $A \subset B + \wp_h(A,B)U$ and $B \subset A + \wp_h(A,B)U$ the following inequalities can be derived

$$\begin{aligned}\underline{a} &\geq \underline{b} + \wp_h(A,B)e \ , \ \underline{b} \geq \underline{a} + \wp_h(A,B)e \ , \\ \overline{a} &\leq \overline{b} + \wp_h(A,B)e \ , \ \underline{b} \leq \overline{a} + \wp_h(A,B)e \ .\end{aligned}$$

Hence

$$|A -^{-} B| = \sup\{|\underline{a} - \underline{b}|, |\overline{a} - \overline{b}|\} \leq \wp_h(A, B)e$$

Therefore

$$\||A -^{-} B|\| \leq \|\wp_h(A, B)e\| = \wp_h(A, B) \tag{2.18}$$

Inequalities (2.17) and (2.18) imply

$$\wp_h(A, B) = \||A -^{-} B|\|$$

Throughout the thesis we will only use the Hausdorff distance in $\mathcal{IR}^n$ and will denote it simply by $\wp$, i.e. $\wp = \wp_h$, if not otherwise stated.

2.1.4 Interval Space over a Space of Real Functions.

Let $\mathcal{F}(\Omega, \mathcal{R}^n)$ be the set of all bounded functions defined in Ω with values in $\mathcal{R}^n$. $\mathcal{F}(\Omega, \mathcal{R}^n)$ is a linear space with the operations:

- addition

$$f + g \ : \ (f + g)(x) = f(x) + g(x) \ , \ \ x \in \Omega, \ f, g \in \mathcal{F}(\Omega, \mathcal{R}^n)$$

- multiplication by a scalar

$$\alpha f \ : \ (\alpha f)(x) = \alpha f(x) \ , \ \ x \in \Omega, \ f \in \mathcal{F}(\Omega, \mathcal{R}^n), \ \alpha \in \mathcal{R} \ .$$

It is also a vector lattice considering the ordering between functions $f, g \in \mathcal{F}(\Omega, \mathcal{R}^n)$ defined by

$$f \leq g \ \Leftrightarrow f(x) \leq g(x) \ , \ \ x \in \Omega$$

Therefore we can consider the interval space $\mathcal{IF}(\Omega, \mathcal{R}^n)$ over $\mathcal{F}(\Omega, \mathcal{R}^n)$. The operations and operators in this space can be represented in the following form using the corresponding operations and operators in $\mathcal{IR}^n$.

$$\begin{aligned}
F + G &: (F + G)(x) = F(x) + G(x) \ , \ \ x \in \Omega \\
F -^{-} G &: (F -^{-} G)(x) = F(x) -^{-} G(x) \ , \ \ x \in \Omega \\
\alpha F &: (\alpha F)(x) = \alpha F(x) \ , \ \ x \in \Omega \\
|F| &: (|F|)(x) = |F(x)| \ , \ \ x \in \Omega \\
w(F) &: (w(F))(x) = w(F(x)) \ , \ \ x \in \Omega
\end{aligned}$$

Let $F = [\underline{f}, \overline{f}] \in \mathcal{F}(\Omega, \mathcal{R}^n)$. Then $F(x) = [\underline{f}(x), \overline{f}(x)] \in \mathcal{IR}^n$, $x \in \Omega$. Hence $F \in \mathcal{F}(\Omega, \mathcal{IR}^n)$ where $\mathcal{F}(\Omega, \mathcal{IR}^n)$ is the set of all bounded functions defined in Ω with values in $\mathcal{IR}^n$. If $G \in \mathcal{F}(\Omega, \mathcal{IR}^n)$ then G can be written as $G = [\underline{g}, \overline{g}]$ where $\underline{g}, \overline{g} \in \mathcal{F}(\Omega, \mathcal{R}^n)$. Therefore $\mathcal{IF} = \mathcal{IF}(\Omega, \mathcal{R}^n) = \mathcal{F}(\Omega, \mathcal{IR}^n)$.

Every function $F = [\underline{f}, \overline{f}] \in \mathcal{IF}$ can be also written in the form $F = (F_1, F_2, \ldots, F_n)^T$ where $F_i = [\underline{f}_i, \overline{f}_i] \in \mathcal{F}(\Omega, \mathcal{IR})$, $\underline{f} = (\underline{f}_1, \underline{f}_2, \ldots, \underline{f}_n)^T$, $\overline{f} = (\overline{f}_1, \overline{f}_2, \ldots, \overline{f}_n)^T$ and the interval operations and operators can be represented coordinate-wise in the same way as in $\mathcal{IR}^n$.

When $\Omega \in \mathcal{IR}^m$ the space $\mathcal{IF} = \mathcal{F}(\Omega, \mathcal{IR}^n)$ consists of interval-valued function of an interval argument.

Definition 2.2 *Let $F \in \mathcal{F}(\Omega, \mathcal{IR}^n), \Omega \subset \mathcal{IR}^m$. Function F is called inclusion-isotone if $X \subset Y$, $X, Y \in \Omega$ implies $F(X) \subset F(Y)$.*

Let us note that the standard operations in any interval space are inclusion-isotone, i.e. if $A \subset C$ then

$$A + B \subset C + B$$
$$\alpha A \subset \alpha C$$
$$A - B \subset C - B$$
$$B - A \subset B - C$$

The nonstandard (inner) operations $-^-$ and $+^-$ are not inclusion-isotone, in general.

Particularly useful in the applications of interval analysis are interval functions of an interval argument that are obtained as extensions of real functions of a real argument. These functions are discussed in the following section.

2.1.5 Interval Extensions of Real Functions of a Real Argument.

Consider $\mathcal{F}(\Omega, \mathcal{R}^n)$ where $\Omega \subset R^m$. Denote by $\mathcal{I}\Omega$ the set of all intervals $X \in \mathcal{R}^n$ such that $X \subset \Omega$.

Definition 2.3 *Let $f \in \mathcal{F}(\Omega, \mathcal{R}^n)$. Function $F \in \mathcal{F}(\mathcal{I}\Omega, \mathcal{IR}^n)$ is called an interval extension of function f if*

(i) F is inclusion-isotone

(ii) $F(x) = f(x)$, $x \in \Omega$

(iii) $x \in X$ implies $f(x) \in F(X)$, $X \in \mathcal{I}\Omega$

Definition 2.4 *Let $f \in \mathcal{F}(\Omega, \mathcal{R}^n)$. Function $f^* \in \mathcal{F}(\mathcal{I}\Omega, \mathcal{IR}^n)$ defined by*

$$f^*(X) = [\inf_{x \in X} f(x), \sup_{x \in X} f(x)]$$

is called optimal interval extension of function f.

Let us note that $f^*(X)$ can also be represented in the form

$$f^*(X) = \bigvee_{x \in X} f(x) = \left[\bigcup_{x \in X} f(x) \right]$$

i.e. $f^*(X)$ is the optimal interval enclosure in $\mathcal{IR}^n$ of all values $f(x)$ when $x \in X$. It is easy to see that the optimal interval extension is an interval extension in the sense of definition 2.3 and also $f^*(X) \subset F(X)$, $X \in \mathcal{I\Omega}$ for every other interval extension F of f.

Let us consider the function $f(x, y) = \alpha x + \beta y$, where $x, y \in \mathcal{R}^n$, $\alpha, \beta \in \mathcal{R}$. Obviously $f \in \mathcal{F}(\mathcal{R}^{2n}, R^n)$. The optimal interval extension of f in terms of the operations in $\mathcal{IR}^n$ is

$$f^*(X, Y) = \alpha X + \beta Y$$

which implies that the standard operations in $\mathcal{IR}^n$ are optimal interval extensions of the linear operations in $\mathcal{R}^n$.

Computing the optimal interval enclosure in the case of nonlinear functions is generally an optimization problem. Simply replacing the reals by intervals leads to an interval extension (sometimes called naive interval extension) which is not necessarily the optimal one. Let us take for example $f(x) = x^2$, $x \in \mathcal{R}$. $F(X) = X.X$ is an interval extension while the optimal one is

$$f^*(X) = \begin{cases} [0, |X|^2] & \text{if} \quad 0 \in X \\ [\underline{x}^2 \vee \overline{x}^2] & \text{if} \quad 0 \notin X \end{cases}$$

We have $F([-2, 2]) = [-4, 4]$ while $f^*([-2, 2] = [0, 4]$.

Computing interval enclosures is a separate area of interval analysis. In the thesis we will use only the optimal interval extension of real functions and we will refer to it simply as interval extension. When $X = x \in \mathcal{R}^n$ the values of f^* and f are the same. Therefore, using the same notation for f and f^* does not lead to confusion. We will denote both functions by f (or the symbol used for the original function).

2.2 Advanced Computer Arithmetic.

In general, computers are equipped with floating-point arithmetic to approximate mathematical operations with real numbers. Details of the representation of floating-point numbers as their radix, number of digits, exponent range, were in the past central points of the analysis and the implementation of a computer arithmetic. A significant breakthrough was the axiomatic definition [57], [58] of computer arithmetic which is independent of such details.

Let $\mathcal{R}$ be the set of real numbers and R be a finite set of floating-point numbers. A mapping $\square : \mathcal{R} \mapsto R$ is called rounding if it satisfies the following conditions:

$$\square x = x \text{ when } x \in R \qquad \text{(projection)} \qquad (2.19)$$

$$x \leq y \Rightarrow \square x \leq \square y\ , \quad x, y \in \mathcal{R} \qquad \text{(monotonicity)} \qquad (2.20)$$

If a rounding $\square$ satisfies also

$$\square(-x) = -(\square x) \tag{2.21}$$

it is called antisymmetric.

Floating-point operations are defined by

$$x \boxed{\circ} y = \square(x \circ y) \ , \quad \circ \in \{+, -, *, /\} \ , \quad (y \neq 0 \ \text{ when } \ \circ = /) \tag{2.22}$$

A mapping $\square$ satisfying axioms (2.19)–(2.22) is said to be a semimorphism. Axiom (2.21) also defines the unary negation operator $\boxminus$ in R.

$$\boxminus x = \square(-x) = -(\square x)$$

Definition 2.5 *A floating-point number ξ is said to be an approximation of maximum quality [84] to a real number x if there is no floating point number between ξ and x, i.e.*

$$[\xi \vee x] \cap R = \emptyset$$

It is easy to see that a rounding $\square$, which is a semimorphism, provides a maximum quality representation of every real number as well as maximum quality floating point arithmetic operations over the set of floating point real numbers R.

The standard operations in the space $\mathcal{IR}$ of real intervals are inclusion isotone. In order to preserve this property for the corresponding floating point interval operations directed roundings are used. These roundings are denoted by $\bigtriangledown$ (downward) and $\bigtriangleup$ (upward) and we have

$$\bigtriangledown x \leq x \quad \text{and} \quad \bigtriangleup x \geq x$$

The directed rounding are required to satisfy (2.19)–(2.20), the floating point operations are defined by (2.22) but (2.21) is replaced by

$$\bigtriangledown(-x) = -(\bigtriangleup x) \ , \quad \bigtriangleup(-x) = -(\bigtriangledown x)$$

Let $\mathcal{IR}$ be the real interval space and IR be the set of floating point intervals, i.e. the real intervals with end points in R. The mapping $\diamondsuit : \mathcal{IR} \mapsto IR$ defined by

$$\diamondsuit[\underline{x}, \overline{x}] = [\bigtriangledown \underline{x}, \bigtriangleup \overline{x}]$$

satisfies the axioms

$$\diamondsuit X = X \ \text{ when } \ X \in IR \qquad \text{(projection)} \tag{2.23}$$

$$X \subset Y \Rightarrow \diamondsuit X \subset \diamondsuit Y \ , \quad x, y \in \mathcal{IR} \qquad \text{(monotonicity)} \tag{2.24}$$

We also have

$$\diamondsuit(-X) = -(\diamondsuit X) \qquad (antisymmetry) \tag{2.25}$$

The floating point interval operations are defined by

$$\begin{aligned} X \diamondsuit_{\circ} Y &= \diamondsuit(X \circ Y)\,, \quad X, Y \in IR, \quad \circ \in \{+, -, -^{\_}, +^{\_}, *, \div\} \\ \alpha \diamondsuit_{\cdot} X &= \diamondsuit(\alpha X)\,, \quad X \in IR,\ \alpha \in R \end{aligned} \tag{2.26}$$

and can be implemented using directed roundings.

Definition 2.6 *A floating-point interval $\Xi = [\underline{\xi}, \overline{\xi}]$ is said to be a maximum quality interval enclosure to a real interval X if every floating point which belongs to the interior of Ξ belongs to X as well, i.e.*

$$(\underline{\xi}, \overline{\xi}) \cap R \subset X \cap R$$

The rounding $\diamondsuit$ provides a maximum quality representation of every real interval as well as maximum quality floating point arithmetic operations over the set of floating point real intervals IR.

The problem of validation will be solved if it were possible to compute a maximum quality floating point approximation or interval enclosure to the exact answer of each mathematical problem. While this is not a reasonable goal for most of the problems there should be a minimum standard that is required.

Traditionally the computer arithmetic provides maximum quality computations only in $\mathcal{R}$ which can be extended to maximum quality computations in $\mathcal{IR}$ as discussed above. However, the minimum standard set by the advanced computer arithmetic is maximum quality computations in the real vector ($\mathcal{VR}$) and matrix ($\mathcal{MR}$) spaces, the space of complex number $\mathcal{C}$ and the complex vector ($\mathcal{VC}$) and matrix ($\mathcal{MC}$) space as well as the interval spaces over $\mathcal{VR}$, $\mathcal{MR}$, $\mathcal{C}$, $\mathcal{VC}$, $\mathcal{MC}$ denoted by $\mathcal{IVR}$, $\mathcal{IMR}$, $\mathcal{IC}$, $\mathcal{IVC}$, $\mathcal{IMC}$, respectively. These spaces are listed in the first column of table 2.1. The corresponding subsets that can be represented in the computer are denoted by the symbols listed in the second column.

We will refer to the spaces $\mathcal{R}$, $\mathcal{VR}$, $\mathcal{MR}$, $\mathcal{C}$, $\mathcal{VC}$, $\mathcal{MC}$ as point spaces and to $\mathcal{IR}$, $\mathcal{IVR}$, $\mathcal{IMR}$, $\mathcal{IC}$, $\mathcal{IVC}$, $\mathcal{IMC}$ as interval spaces. We will need to distinguish between the above two sets of spaces because (as will be seen below) the ordering used in the definition of rounding in a point space is $\leq$ while the ordering used in an interval space is $\subset$.

Let $\mathcal{U}$ be any of the spaces in the first column of table 2.1 and U be the corresponding space in the second column. Furthermore, let a rounding $\square : \mathcal{U} \mapsto U$ which satisfies the following axioms be given.

Basic Spaces of Computation	Subsets Representable on the computer
$\mathcal{R}$	R
$\mathcal{VR}$	VR
$\mathcal{MR}$	MR
$\mathcal{IR}$	IR
$\mathcal{IVR}$	IVR
$\mathcal{IMR}$	IMR
$\mathcal{C}$	C
$\mathcal{VC}$	VC
$\mathcal{MC}$	MC
$\mathcal{IC}$	IC
$\mathcal{IVC}$	IVC
$\mathcal{IMC}$	IMC

Table 2.1: *Table of the spaces of computation and the corresponding subspaces representable on a computer.*

$$\Box x = x \text{ when } x \in U \tag{2.27}$$

$$x \leq y \Rightarrow \Box x \leq \Box y\,, \quad x, y \in \mathcal{U} \quad \text{(for point spaces)}$$

$$x \subset y \Rightarrow \Box x \subset \Box y\,, \quad x, y \in \mathcal{U} \quad \text{(for interval spaces)} \tag{2.28}$$

$$\Box(-x) = -(\Box x) \tag{2.29}$$

For an interval space $\mathcal{U}$, the rounding $\Box$ is also required to satisfy

$$x \subset \Box x\,, \quad x \in \mathcal{U} \tag{2.30}$$

The theory developed in [56], [57] shows that this rounding is uniquely defined.

Maximum quality approximation (enclosure) is defined in general as follows:

Definition 2.7 *An element $\xi \in U$ is a maximum quality approximation (enclosure) of $x \in \mathcal{U}$ if there are no other elements of U between ξ and x, i.e*

- *for point spaces: there is no* $\eta \in U$ *such that* $\xi \leq \eta \leq x$ *or* $x \leq \eta \leq \xi$;
- *for interval spaces: there is no* $\eta \in U$ *such that* $x \subset \eta \subset \xi$.

It is shown in [57] that a rounding $\square$ satisfying (2.27)–(2.29) ((2.27)–(2.30)) provides maximum quality representation of $\mathcal{U}$ in U. Furthermore, the operations in U, defined by

$$x \boxed{\circ} y = \square(x \circ y) \tag{2.31}$$

are of maximum quality in the sense that $x \boxed{\circ} y$ is the maximum quality approximation (enclosure) of $x \circ y$. The operation $\circ$ is any operation from the set of operation specific for the space $\mathcal{U}$.

The cornerstone in the implementation of the advanced computer arithmetic is the maximum quality scalar product of real floating point vectors $x, y \in VR$ defined by

$$x \boxed{\cdot} y = \square(x.y) = \square\left(\sum_{i=1}^{n} x_i y_i\right)$$

The traditional implementation of the scalar product

$$x_1 \boxed{\cdot} y_1 \boxplus x_2 \boxed{\cdot} y_2 \boxplus \ldots \boxplus x_n \boxed{\cdot} y_n$$

is with $2n - 1$ roundings, while the maximum quality scalar product produces a result with a single rounding. On the basis of the maximum quality scalar product all operations in vector and matrix spaces are implemented with maximum quality. For example, the multiplication of two $n \times n$ matrices is equivalent to computing n^2 scalar products.

In addition to the maximum quality representation and operations in the spaces listed in table 2.1, the advanced computer arithmetic requires a maximum quality evaluation of all standard mathematical functions (modulus, square root, exponential, logarithmic, trigonometric,hyperbolic function, etc.).

At present, a large variety of software products supporting advanced computer arithmetic exists. In [26], [27] a number of such software products are compared. Some of these products are ACRITH-XSC [44], [90], C-XSC [52], Fortran-XSC [91], INTLIB [51], INTPACK [25], PASCAL-XSC [53]. ACRITH-XSC, C-XSC and PASCAL-XSC are production quality commercial products supporting all aspects of the advanced computer arithmetic. For the numerical experiments presented in the thesis we use PASCAL-XSC.

PASCAL-XSC is a programming language providing the following features:

- Explicit language support for directed roundings and the corresponding operations.
- Maximum quality scalar product for vectors of arbitrary length.
- Interval types and and interval arithmetic operations.
- A universal operator concept.

- Overloading of function identifiers and operators.
- Dynamic and structured numerical types.
- A large number of mathematical functions with high accuracy for all numerical types.

The PASCAL-XSC programming in any of the spaces listed in table 2.1 is as easy as, say, PASCAL programming in R. For example, the product of a matrix $M \in R^n$ and an interval vector $V \in IR^n$ is written in the form $M * V$. This operator produces a maximum quality result which is an interval vector.

2.3 Initial Value Problems and Operators of Monotone Type.

In the thesis, we consider the initial value problem (IVP) for ordinary differential equations (ODE) and the periodic initial value problem for hyperbolic partial differential equations (specifically the wave equation). For both problems we consider the case of an initial condition which is an interval: n-dimensional real interval, or set of two interval functions as the case may be. The problems will be formulated in detail in the following sections. The numerical techniques used for these problems are quite different. However, in the computation of validated numerical solutions the two problems share a lot of similarities, particularly related to the validation of the the solution. One such similarity is the connection between the monotone properties of the problem, in terms of the concept of operators of monotone type [23] (recalled below), and the construction of interval enclosures for the exact solution.

Definition 2.8 *Let Ω and $\mathcal{W}$ be lattices and let $\leq$ denote the partial ordering in each of them. An operator $T : \Omega \mapsto \mathcal{W}$ is called an operator of monotone type if*

$$Tu \leq Tv \Rightarrow u \leq v\,, \quad u, v \in \Omega$$

An initial value problem can be formulated generally in the following way:

Let Ω be a space of functions defined on a domain D and let $\mathcal{K}$ and I be operators defined on Ω with ranges in some spaces $\mathcal{W}$ and $\mathcal{W}_0$. Then, the problem is

$$\begin{aligned} &\text{Find } u \in \Omega \text{ such that} \\ &\mathcal{K}u = 0 \\ &Iu = u^0 \end{aligned} \tag{2.32}$$

where $u^0 \in \mathcal{W}_0$ is given.

For example, if $\Omega = C^1[t_0, \overline{t}]$, $\mathcal{W} = C^0[t_0, \overline{t}]$, $\mathcal{W}_0 = \mathcal{R}$, $Iu = u(t_0)$ and $\mathcal{K}u(t) = \frac{du(t)}{dt} - f(t,u)$, where f is a given continuous function of t and u, we obtain an IVP for ODE.

Taking $D = \mathcal{R} \times [t_0, \overline{t}]$, $\Omega = \{u = u(x,t) : D \mapsto \mathcal{R} : u \in C^2(D), u\text{—periodic on } x\}$, $\mathcal{W} = C^0(D)$, $\mathcal{W}_0 = C^2(\mathcal{R}) \times C^1(\mathcal{R})$, $Iu(x) = (u(x,t_0), u_t(x,t_0))$ and $\mathcal{K}u(x,t) = u_{tt}(x,t) - u_{xx}(x,t) - f(x,t,u)$, where f is a given continuous function of t, x and u, problem (2.32) becomes a periodic initial value problem for the wave equation.

The spaces Ω, $\mathcal{W}$ and W_0 are assumed to be vector lattices and have norms so that the corresponding interval spaces with metric are defined.

Operator I in the formulation of problem (2.32) represents an initial condition. In practical applications, very often the value of $Iu = u^0$ is not exactly known. Even in the rare case, when the value of u^0 is known exactly, this value is not necessarily exactly representable on a computer using the available data types. An interval rounding will produce an interval $\diamondsuit(u^0)$ and we will only know that $u^0 \in \diamondsuit(u^0)$. Therefore it is important to consider initial conditions of the form

$$Iu = u^0 \in U^0 \tag{2.33}$$

where $U^0 = [\underline{u}^0, \overline{u}^0] \in \mathcal{I}\mathcal{W}_0$ is given.

If $u(u^0; y)$, $y \in D$, denotes a solution of (2.32) then the set of functions

$$u(U^0) \equiv u(U^0; .) = \{u(u^0; .) : u^0 \in U^0\}$$

is considered a solution of problem (2.32)–(2.33). In general $u(U^0)$ is not an interval function in $\mathcal{I}\Omega$. In interval terms (as discussed in section 2.1),

$$[u(U^0; y)] = [\underline{u}(U^0; y), \overline{u}(U^0; y)],\ y \in D\ .$$

represents the optimal interval enclosure of $u(U^0; y)$, $y \in D$.

Using the above notations, the numerical problem can be formulated as:

$$\begin{array}{l} \text{Construct } S \in \mathcal{I}\Omega \text{ such that} \\ [u(U^0; y)] \subset S(y)\ , y \in D, \text{ and} \\ \wp([u(U^0)], S) < \textit{tolerance}. \end{array} \tag{2.34}$$

The design of methods producing enclosures S as required in (2.34) depends significantly on the structure of $[u(U^0)]$. When the operator $T = (\mathcal{K}, I)$ is an operator of monotone type this structure is very simple.

Indeed, for such an operator T, every solution $u(u^0; y)$ of problem (2.32) such that $u^0 \in U^0 = [\underline{u}^0, \overline{u}^0]$ satisfies

$$T(u(\underline{u}^0)) = (0, \underline{u}^0) \leq (0, u^0) = T(u(u^0)) = (0, u^0) \leq (0, \overline{u}^0) = T(u(\overline{u}^0))$$

Therefore

$$u(\underline{u}^0; y) \leq u(u^0; y) \leq u(\overline{u}^0; y),\ y \in D$$

This implies that the optimal interval enclosure can be represented in the form

$$[u(U^0; y)] = [u(\underline{u}^0; y), u(\overline{u}^0; y)],\ y \in D$$

Since the end points of the optimal interval enclosure are solutions of problem (2.32), an enclosure S of the type required in (2.34) can be computed by computing a lower bound of $u(\underline{u}^0; y)$ and an upper bound of $u(\overline{u}^0; y)$, $y \in D$. This is the approach adopted in chapter 4 for constructing tight enclosures for the solution of the wave equation.

A major problem associated with interval methods for initial value problems is the wrapping effect. It is caused by the set of additional points (called wrapping excess) included in an interval which encloses (wraps) a noninterval set. The result may be a significant inflation of the computed enclosures. The wrapping effect for IVP for ODE is explained and studied in detail in chapter 3 and it is shown that when operator T is of monotone type the problem has no wrapping effect. In chapter 4, where we consider mainly wave equations with a differential operator T which is of monotone type, we observe that no wrapping effect appears.

2.4 Initial Value Problem for ODE: Wrapping Effect.

2.4.1 The Problem.

We consider the initial value problem for ordinary differential equations in the form

$$\dot{x} = f(t, x) \tag{2.35}$$
$$x(t_0) = x^0 \in X^0 \tag{2.36}$$

where $t \in [t_0, \overline{t}] \subset \mathcal{R}$, $x^0 \in \mathcal{R}^n$, $D \subset \mathcal{R}^n$ is an open set, $f : [t_0, \overline{t}] \times D \to \mathcal{R}^n$ and

$$X^0 = ([\underline{x}_1^0, \overline{x}_1^0], [\underline{x}_2^0, \overline{x}_2^0], \ldots, [\underline{x}_n^0, \overline{x}_n^0])^T$$

is an n-dimensional interval vector, $X^0 \subset D$. We assume that a solution is sought in the interval $[t_0, \overline{t}]$. A validated numerical method, if returns an answer, produces an n-dimensional interval function $S(h; t)$, $t \in [t_0, \overline{t}]$ with the assurance that for every $x^0 \in X^0$ a unique solution $x(t_0, x^0; t)$ exists in $[t_0, \overline{t}]$ and $x(t_0, x^0; t) \in S(h; t)$, $t \in [t_0, \overline{t}]$. The parameter h is a parameter of the method.

Our primary task is a study of the wrapping effect associated with validated (interval) methods for problem (2.35)–(2.36). When applying a validated method to a particular problem of the form (2.35)–(2.36) we do not need to make or verify any assumptions for f or X^0 because the existence and uniqueness of the solution is verified automatically.

However, here we will prove statements about the convergence (a priori) of the enclosures produced by a class of methods characterizing the quality of these methods rather than the quality of the enclosures produced for a particular problem. In order to do this, we need to make assumptions providing for existence and uniqueness of the solution. We will assume that in the region $[t_0, \overline{t}] \times D$ the function f

$$\begin{aligned} &i) \quad \text{is bounded: } |f_i(t,x)| \le m_i \in \mathcal{R},\ m = (m_1, m_2, \ldots, m_n)^T \in \mathcal{R}^n\ ; \\ &ii) \quad \text{is continuous about } t\ ; \\ &iii) \quad \text{satisfies the Lipschitz condition about } x\text{:} \\ &\qquad |f_i(t,y) - f_i(t,z)| \le \sum_{j=1}^{n} \lambda_{ij} |y_j - z_j|,\ \lambda_{ij} \in \mathcal{R},\ i,j = 1, \ldots, n\ . \end{aligned} \tag{2.37}$$

We would like to note that all the proofs can be actually carried out under more general assumptions providing only for existence and uniqueness of a solution $x(t_0, x^0; t)$ in a weaker sense leading to a continuous function satisfying $x(t) = x^0 + \int_{t_0}^{t} f(\theta, x(\theta))d\theta,\ t \in [t_0, \overline{t}]$. But we feel that such assumptions will only make the proofs more technical and difficult to read without making any major contribution to the presentation and clarification of the main ideas.

We will also assume that all solutions $x(t_0, x^0; t),\ x^0 \in X^0$, are defined in the whole interval $[t_0, \overline{t}]$.

The set-valued function $x(t_0, X^0; t) = \{x(t_0, x^0; t) : x^0 \in X^0\},\ t \in [t_0, \overline{t}]$ is considered a solution to the problem (2.35)–(2.36). For every $t \in [t_0, \overline{t}]$ by $[x(t_0, X^0; t)]$ we denote the optimal (tightest) interval containing the set $x(t_0, X^0; t)$ (which is not necessarily an interval). The interval function

$$[x(t_0, X^0; .)] : [t_0, \overline{t}] \to \mathcal{IR}^n$$

is called the optimal interval enclosure of the solution. Therefore a validated method produces interval functions $S(h; t)$ such that $[x(t_0, X^0; t)] \subset S(h; t)$.

2.4.2 Historical Overview.

The idea of solving problem (2.35)–(2.36) by constructing lower and upper approximations (bounds) for the solution $x(t_0, X^0; t)$ was proposed by Chaplygin [21], [22] in 1919. He proposed an iterative method with quadratic convergence. See [32] for a more contemporary presentation. He also proved a theorem characterizing the monotonicity of problem (2.35)–(2.36) in the case of one equation. The monotonicity in the case of a system was studied by Müller [68]. He proved that the operator of problem (2.35)–(2.36) is of monotone type under the assumption that the right hand side f is a quasi-isotone function of x.

Definition 2.9 *A function $f = (f_1, f_2, \ldots, f_n)^T : D \to \mathcal{R}^n$ is called quasi-isotone if for every $i = 1, \ldots, n$, $f_i = f_i(x_1, x_2, \ldots, x_n)$ is non-decreasing with respect to all x_j, $j \neq i$.*

His theorems were generalized by Kamke in 1932 and later by Walter [92].

The work of Chaplygin was further developed in the Soviet Union by Luzin and Babkin [17]. It is interesting to note that these mathematicians did not know the results of the German mathematicians Müller and Kamke. For example, Babkin essentially formulates the theorem of Müller for two equations. On the other side, Müller and Kamke did know the results of Chaplygin.

We will use the theorem of Müller in the following form.

Theorem 2.1 *Let function f in equation (2.35) be quasi-isotone with respect to x. If functions $u, v : [t_0, \overline{t}] \mapsto D$ are differentiable in $[t_0, \overline{t}]$ and satisfy the inequalities*

$$\begin{aligned} &\dot{u}(t) - f(t, u(t)) \leq \dot{v}(t) - f(t, v(t)) \,, \quad t \in [t_0, \overline{t}] \,, \\ &u(t_0) \leq v(t_0) \end{aligned}$$

then $u(t) \leq v(t)$, $t \in [t_0, \overline{t}]$.

This theorem means that the operator T defined by $T(x) = (\dot{x} - f(t, x), x(t_0))$ where $x : [t_0, \overline{t}] \mapsto D$ and is differentiable in $[t_0, \overline{t}]$, is an operator of monotone type.

As a direct consequence of theorem 2.1 we obtain

Theorem 2.2 *Let f be quasi-isotone on x and let $S = [\underline{s}, \overline{s}]$ be an interval function defined on $[t_0, \overline{t}]$ such that $S(t) \subset D$, $t \in [t_0, \overline{t}]$ and $\underline{s}, \overline{s}$ are differentiable in $[t_0, \overline{t}]$. If*

$$\begin{aligned} &\dot{\underline{s}}(t) - f(t, \underline{s}(t)) \leq 0 \leq \dot{\overline{s}}(t) - f(t, \overline{s}(t)) \,, \quad t \in [t_0, \overline{t}] \,, \\ &\underline{s}(t_0) \leq \underline{x}^0, \quad \overline{x}^0 \leq \overline{s}(t_0) \end{aligned}$$

then $x(t_0, X^0; t) \subset S(t)$, $t \in [t_0, \overline{t}]$.

In 1965, Moore [64] described an enclosure method for ODE using interval arithmetic for the first time. He used the Taylor series expansion of the solution to construct interval enclosure. This approach remains the most common one until now [55], [37], [59]. Other methods are interval analogs of standard methods (e.g. Runge-Kutta, Adams, etc.) assuming that an interval function containing the error of the corresponding method is available [46]. Moore [65] also introduced Pickard-Lindelöf iteration. Bauch and others [18] modified the method and suggested Newton iteration.

In the eighties Nickel [82], Stetter [88], Bauch *et al.* [19], Corliss [28], Kalmykov *et al.* [46] gave surveys on interval methods available then. Corliss *et al.* [29] contains a very extensive bibliography.

In the area of single-step methods, progress during the last decade was made mainly in improving the computed enclosures by using different strategies for step-size and order

control, defect-correction, improved provisional enclosures and others. Recent surveys are given in [86] and [78].

Every interval method has to face the wrapping effect which is a severe obstacle and has been a central issue in the construction of interval methods since the introduction of interval analysis to the numerical solution of the IVP for ODE. Moore [65], Krückeberg [55], Lohner [59] present modifications of the Taylor method which can reduce the wrapping effect. The method of Lohner is the most famous one. A further modification proposed by Rihm [86] improves the computed enclosures. In the thesis we study the wrapping effect associated with single-step interval methods.

2.4.3 Wrapping Effect and Wrapping Function.

We will consider methods that generate interval enclosure $S(h;t)$ using a mesh $\{t_0, t_1, \ldots, t_p = \bar{t}\}$ where $h = (h_1, h_2, \ldots, h_n)$, $h_k = t_k - t_{k-1}$, $k = 1, \ldots, \overline{k}$.

Naturally, convergence of the form

$$\lim_{\|h\| \to 0} S(h;t) = [x(t_0, X^0; t)]$$

is desirable. However, due to the wrapping effect, in most of the cases such convergence is not observed even when the method has very good local approximation properties. This is demonstrated in the following example.

Example 2.1 Consider the problem

$$\begin{aligned} \dot{x}_1 &= -2x_1\,, & x_1(0) &= x_1^0 \in X_1^0 = 1 + [-\varepsilon_1, \varepsilon_1]\,, \\ \dot{x}_2 &= 2x_1 - x_3\,, & x_2(0) &= x_2^0 \in X_2^0 = 1 + [-\varepsilon_2, \varepsilon_2]\,, \\ \dot{x}_3 &= 2x_1 - x_2\,, & x_3(0) &= x_3^0 \in X_3^0 = 1 + [-\varepsilon_3, \varepsilon_3]\,. \end{aligned} \tag{2.38}$$

in the interval $[0, 1]$. We apply a method based on the Taylor series of the solution with local approximation error $O(h^5)$. In every interval $[t_k, t_{k+1}]$ the already computed enclosure $S(h; t_k)$ is considered as an initial condition and we have

$$\wp(S(h;t), [x(t_k, S(h, t_k); t)]) = O(h_k^5)\,, \quad t \in [t_k, t_{k+1}]\,.$$

When

$$\varepsilon_1 = 0.2,\ \varepsilon_2 = \varepsilon_3 = 0.05 \tag{2.39}$$

the optimal interval enclosure and enclosures, computed for various values of h, are represented graphically on figure 2.1. Since the corresponding enclosures for x_2 and x_3 are the same, they are represented by the same graphs. While the numerically computed enclosures for x_1 are visually indistinguishable from the optimal one, the computed enclosures for x_2 and x_3 clearly diverge from the optimal. We can see that reducing the step size

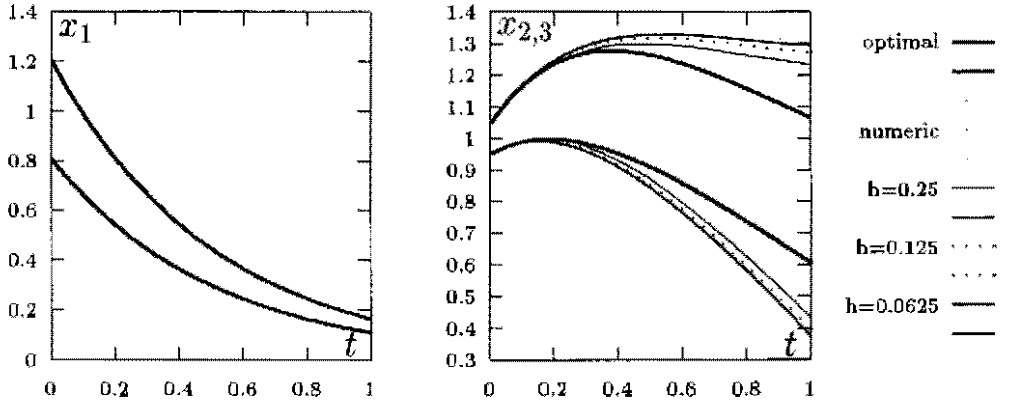


Figure 2.1: *Problem (2.38) with* $\varepsilon_1 = 0.2$, $\varepsilon_2 = \varepsilon_3 = 0.05$. *Optimal enclosure and enclosures computed numerically for various step sizes* h.

makes the matter only worse. Increasing the order of local approximation also does not help.

However, when the same method is applied to the problem (2.38) with

$$\varepsilon_1 = 0,\ \varepsilon_2 = \varepsilon_3 = 0.05 \tag{2.40}$$

we obtain a very good approximation of the optimal interval enclosure in all three variables x_1, x_2 and x_3. The numerical results are graphically represented in figure 2.2. At the top part of the figure the optimal enclosure and the enclosures computed for various values of h are plotted. Since the computed enclosures are very close to the optimal enclosure, they are visually indistinguishable from the optimal one. In order to see their accuracy, the error functions

$$\wp\left(S_i(h;t), [x_i(t_0, X^0; t)]\right)\ ,\quad i = 1, 2, 3$$

are plotted on a logarithmic scale at the bottom part of the figure. A rate of convergence, consistent with the expected rate of global convergence $O(h^4)$ is revealed.

The divergence of the computed interval enclosures away from the optimal enclosure when $h \to 0$ which is observed for x_2 and x_3 on figure 2.1 is due to the wrapping effect. While a detailed explanation of the wrapping effect will be given below, we can say roughly that it manifests itself as divergence of the computed enclosures away from the optimal enclosure when $h \to 0$ irrespective of the order of local approximation. We can see from the numerical experiments with example 2.1 that there are problems (e.g. problem (2.38)–(2.40)) where the wrapping effect does not appear at all and the computed enclosures behave in a "regular" way, i.e. converge to the optimal enclosure when $h \to 0$.

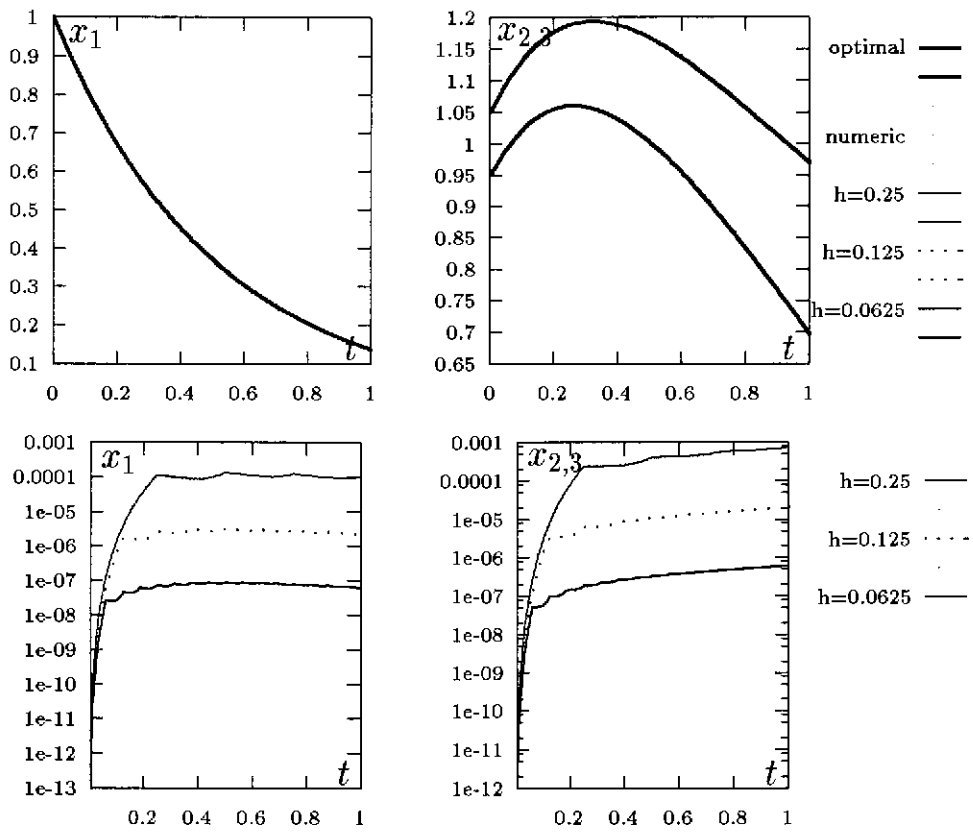


Figure 2.2: *Problem (2.38) with* $\varepsilon_1 = 0$, $\varepsilon_2 = \varepsilon_3 = 0.05$. *Optimal enclosure and enclosures computed numerically for various step sizes h (top) and errors of the computed enclosures on a logarithmic scale (bottom).*

We will explain the wrapping effect using Jackson's [45] propagate and wrap approach. Suppose that we can compute the optimal interval enclosure in any interval $[t_k, t_{k+1}]$. Then interval enclosures can be computed by the following procedure which we call Idealized Propagate and Wrap Algorithm (IPWA):

$$\begin{aligned} &\boldsymbol{S}(h;t_0) = X^0 \\ &\boldsymbol{S}(h;t) = [x(t_k, \boldsymbol{S}(h;t_k);t)],\ t \in [t_k, t_{k+1}],\ k = 0, 1, \ldots, n-1 \end{aligned} \tag{2.41}$$

This method has no local error but does not always produce the optimal enclosure. The solution at t_1 is the set $x(t_0, X^0; t_1)$ which is not necessarily an n-dimensional interval. It is wrapped by an interval $\boldsymbol{S}(h;t_1) = [x(t_0, X^0; t_1)]$, thereby including extra points called wrapping excess. In the interval $[t_1, t_2]$ all solutions starting from the points of $\boldsymbol{S}(h;t_1)$ (including the wrapping excess) are propagated and the set $x(t_1, \boldsymbol{S}(h;t_1);t_2)$ is again wrapped by an interval $\boldsymbol{S}(h;t_2) = [x(t_1, \boldsymbol{S}(h;t_1);t_2)]$ with certain wrapping excess and so on. The wrapping excess at the points of the mesh is what causes, in some cases, blowing

up of the enclosures as observed for variables $x_{2,3}$ on figure 2.1, and referred to as the wrapping effect. In other cases, despite the wrapping excess, the computed enclosures converge to the optimal one (figure 2.2), i.e. there is no wrapping effect.

Moore [67] noticed the problems associated with the wrapping excess at the points of the mesh and proposed coordinate transformations to counter the wrapping effect. A large number of papers on validated (interval) methods for ordinary differential equations deal with the wrapping effect, and [55], [37], [59] mark some major developments in the area. See also [78] for a recent survey.

A well known case of problems with no wrapping effect is when the function f in (2.35) is quasi-isotone. There is also no wrapping effect when the initial condition is a point $X^0 = x^0 \in \mathcal{R}$. However, problem (2.38)-(2.40) is of neither of those types, but there is still no wrapping effect.

We study the wrapping effect by introducing a new concept of wrapping function.

We consider single-step methods producing enclosures $S(h;t)$ such that in every interval $[t_k, t_{k+1}]$ the interval function $S(h;t)$ encloses all solutions propagated from the points of the interval $S(h;t_k)$. We call these methods methods of propagate and wrap type.

Definition 2.10 *A function $X : [t_0, \overline{t}] \to \mathcal{IR}^n$ is said to satisfy a wrapping property with respect to equation (2.35)-(2.36) at the point $\theta \in [t_0, \overline{t})$ if all solutions $x(\theta, u; t)$, $u \in X(\theta)$, exist in the interval $[\theta, \overline{t}]$ and $x(\theta, X(\theta); t) \subset X(t)$, $t \in [\theta, \overline{t}]$.*

Using definition 2.10 we can also define the methods of propagate and wrap type as follows:

Definition 2.11 *A single-step numerical method producing enclosures $S(h;t)$ for the solution of problem (2.35)-(2.36) is called a method of propagate and wrap type if the enclosures $S(h;t)$ satisfy the wrapping property at all points of the mesh.*

The Idealized Propagate and Wrap Algorithm (IPWA) discussed earlier is a method of propagate and wrap type. It produces the tightest enclosure $\boldsymbol{S}(h;t)$ that can be produced by a method of propagate and wrap type because its local error is zero.

In considering methods of propagate and wrap type we will ignore the method used by the numerical procedure for producing enclosures (e.g. Taylor series, Runge-Kutta, etc.). We will only discuss the convergence of the computed bounds $S(h;t)$. In all cases of wrapping effect convergence of the computed enclosures (although not to the optimal enclosure) is observed (see the graphs for $x_{2,3}$ in figure 2.1). Then, it is logical to ask:

If the computed enclosures do not converge to the optimal one, what do they converge to?

We prove that the enclosures computed by any method of propagate and wrap type converge to the wrapping function discussed below.

Since the enclosures produced by methods of propagate and wrap type satisfy the wrapping property at every point of the mesh, we may expect that the limit of such

interval functions when $h \to 0$ satisfies the wrapping property at every point of the interval $[t_0, \overline{t}]$. Therefore, we define the concept of wrapping function as follows:

Definition 2.12 *A function $X : [t_0, \overline{t}] \to I\mathcal{R}^n$ is called wrapping function for problem (2.35)-(2.36) if:*

i) $X(t_0) = X^0$;

ii) X satisfies the wrapping property at every point of the interval $[t_0, \overline{t})$;

iii) for every other function $Y : [t_0, \overline{t}] \to I\mathcal{R}^n$ satisfying i) and ii) we have $X(t) \subset Y(t)$, $t \in [t_0, \overline{t}]$ (i.e. the wrapping function is the optimal function with i) and ii)).

The wrapping function of problem (2.35)–(2.36) is unique. Indeed, if Y and Z are wrapping functions then function X defined by $X(t) = Y(t) \cap Z(t)$, $t \in [t_0, \overline{t}]$ satisfies conditions i) and ii) of the definition and $X(t) \subset Y(t)$, $X(t) \subset Z(t)$, $t \in [t_0, \overline{t}]$. Moreover, the converse inclusions hold from condition iii) in the definition. Therefore $X(t) = Y(t) = Z(t)$. We denote the wrapping function of problem (2.35)–(2.36) by $\widehat{X}$.

In Chapter 3 we prove the existence and some properties of the wrapping function by representing it as a solution to a certain initial value problem. We also prove that, in general, the limit of the enclosures produced by methods of propagate and wrap type is the wrapping function and not the optimal interval enclosure. Therefore, the computed enclosures converge towards the optimal interval enclosure if and only if the wrapping function is equal to the optimal interval enclosure. In this way the wrapping effect can be considered as an inherent property of the problem and can be quantified as the distance between the wrapping function and the optimal interval enclosure. Apart from the theoretical value of this result in studying and understanding the wrapping effect it has a practical application in characterizing problems with no wrapping effect. In [30], [78] it is stated that a complete set of tools for validated solving of IVP for ODE should include software for recognizing problems with quasi-isotone right-hand side and solving them by a straight forward procedure instead of using complicated algorithms [37], [81], [59]. We prove that there is a larger class of initial value problems that have no wrapping effect and we believe that such software should recognize the problems with no wrapping effect as they are specified by the theorems proved in chapter 3.

2.5 Wave Equation: Monotone Properties.

2.5.1 Interval Methods for Partial Differential Equations.

Advances in the development of interval methods for PDEs were reported as early as the seventies. In 1972 Appelt [15] obtained error bounds for an approximate solution of a class

of elliptic problems using interval methods. His approach was also discussed by Moore [67]. A method for construction of bounds for the characteristic initial value problem for hyperbolic equations is proposed in [33], [34]. The problem is formulated as a fixed-point equation, and bounds for the solution of this equation are obtained iteratively. Validation of the bounds is obtained by using the fixed-point theorem approach.

A significant contribution to the development of validated methods for PDE is the introduction of the concept of functoid [49] with applications to PDE discussed in [50]. Particularly applicable to the wave equation is the Fourier functoid which is discussed later in this section. The concept is generalized to a Fourier hyper functoid in [48].

An important development in the last decade is the use of the method of finite elements for computing validated solutions of PDE [69]-[77]. The methods proposed by Nakao and his colleagues use a finite element solution and a computable error estimate for obtaining enclosures. The validation is also based on Schauder's fixed point theorem.

In the thesis we consider the wave equation which is a hyperbolic problem. Our task is to establish monotone properties of the problem and construct methods based on these monotone properties.

2.5.2 The Problem.

We consider the nonlinear wave equation

$$u_{tt}(x,t) - u_{xx}(x,t) = f(x,t,u(x,t)) \,, \quad -l < x < l \,, \quad t > 0 \tag{2.42}$$

$$u(x,0) = g_1(x) \,, \; u_t(x,0) = g_2(x) \,, \quad -l < x < l \tag{2.43}$$

$$u(-l,t) = u(l,t) \,, \; u_x(-l,t) = u_x(l,t) \,, \quad t > 0 \tag{2.44}$$

Condition (2.44) implies that the solution is a function which has a smooth $2l$-periodical extension about x. The periodic boundary condition is essential for the monotone properties of the problem and the construction of numerical methods as discussed in the following sections. However, it is not a very restrictive assumption because a large number of problems can be reduced to problems with periodic boundary conditions of the form (2.42)–(2.44). For example if the problem is given in the more common initial boundary value form:

$$u_{tt}(x,t) - u_{xx}(x,t) = f(x,t,u(x,t)) \,, \quad 0 < x < l \,, \quad t > 0$$

$$u(x,0) = g_1(x) \,, \; u_t(x,0) = g_2(x) \,, \quad 0 < x < l$$

$$u(0,t) = \varphi(t) \,, \; u(l,t) = \phi(t) \,, \quad t > 0$$

substituting

$$v(x,t) = u(x,t) + \frac{1}{l}((x-l)\varphi(t) - x\phi(t))$$

we obtain a problem with zero boundary conditions

$$\begin{aligned} v_{tt}(x,t) - v_{xx}(x,t) &= \hat{f}(x,t,v(x,t)) \,, \quad 0 < x < l \,, \quad t > 0 \\ v(x,0) = \hat{g}_1(x) \,, \; v_t(x,0) &= \hat{g}_2(x) \,, \quad 0 < x < l \\ v(0,t) = v(l,t) &= 0 \,, \qquad t > 0 \end{aligned}$$

where

$$\begin{aligned} \hat{f}(x,t,v) &= f\left(x,t,v - \frac{1}{l}((x-l)\varphi(t) - x\phi(t))\right) + \frac{1}{l}((x-l)\varphi_{tt}(t) - x\phi_{tt}(t)) \\ \hat{g}_1(x) &= g_1(x) + \frac{1}{l}((x-l)\varphi(0) - x\phi(0)) \\ \hat{g}_2(x) &= g_2(x) + \frac{1}{l}((x-l)\varphi_t(0) - x\phi_t(0)) \end{aligned}$$

Defining all functions for $x \in (-l, 0)$ as odd functions of x, the zero boundary conditions at $x = 0$ and $x = l$ can be replaced by periodic boundary conditions at $x = -l$ and $x = l$. Problems concerning the differentiability of the solution that may arise in this transformation will be dealt with using a weaker formulation which is discussed in chapter 5. Let $\Omega[\underline{t}, \overline{t}]$ be the set of all functions $u = u(x,t) : \mathcal{R} \times [0, \infty) \mapsto \mathcal{R}$ which are $2l$-periodical about x and have continuous second derivatives. Assuming that functions f, g_1 and g_2 are extended periodically about x (period $2l$) we can formulate problem (2.42)–(2.44) in the following way:

Find $u \in \Omega[0, \overline{t}]$ such that

$$u_{tt}(x,t) - u_{xx}(x,t) = f(x,t,u(x,t)) \,, \quad x \in \mathcal{R} \,, \quad t > 0 \tag{2.45}$$

$$u(x,0) = g_1(x) \,, \; u_t(x,0) = g_2(x) \,, \quad x \in \mathcal{R} \tag{2.46}$$

The solution of the above problem we denote by $u(0, g; x, t)$.

We also consider an interval initial condition of the form

$$\begin{aligned} u(x,0) &= g_1(x) \in G_1(x) = [\underline{g}_1(x), \overline{g}_1(x)] \,, \\ u_t(x,0) &= g_2(x) \in G_2(x) = [\underline{g}_2(x), \overline{g}_2(x)] \,, \; x \in \mathcal{R} \end{aligned} \tag{2.47}$$

where G_1 and G_2 are given interval functions. The set-valued function

$$u(0, G; x, t) = \{u(0, g; x, t) : g \in G\}, \; x \in \mathcal{R}, \; t \in [0, \overline{t}]$$

is considered a solution of problem (2.45)–(2.47)

In chapter 4 we assume that f is a continuous function of all arguments, $g_{1,2}$ ($\underline{g}_{1,2}, \overline{g}_{1,2}$) are differentiable and $g'_{1,2} \in L_2(-l, l)$ ($\underline{g}_{1,2}, \overline{g}_{1,2} \in L_2(-l, l)$). In addition, we also make the assumption that f is a non-decreasing function of u and the monotone properties of the problem are used in the construction of enclosures. In chapter 5 we consider means of dealing with discontinuities of f, g_1 and g_2 or their derivatives.

2.5.3 Monotone Properties.

The monotone properties of an initial value problem play an important role in the design of interval methods (as discussed in section 2.3). Let us denote the operator $u_{tt} - u_{xx}$ shortly by Lu, i.e.

$$Lu = u_{tt} - u_{xx} \ , \quad u \in \Omega[0, \overline{t}] \ .$$

and let $T(t_\alpha)$ be an operator defined in $\Omega[t_\alpha, \overline{t}]$, $t_\alpha \in [0, \overline{t})$ as

$$T(t_\alpha, u; x, t) = (Lu(x,t) - f(x,t,u), u(x,t_\alpha), u_t(x,t_\alpha)) \ , \ x \in \mathcal{R}, \ t \in [t_\alpha, \overline{t}] \ .$$

Then problem (2.45)–(2.46) can be written as

$$T(0, u) = (0, g_1, g_2)$$

In chapter 4 we prove that

$$\begin{array}{l} \textit{If } f \textit{ is a non–decreasing function of } u \textit{ then} \\ T(t_\alpha) \textit{ is an operator of monotone type} \end{array} \tag{2.48}$$

Therefore, when f is non-decreasing about u, the optimal enclosure $[u(0, G; x, t)]$ of the solution of problem (2.45)–(2.47) can be represented in the form

$$[u(0, G; x, t)] = [u(0, \underline{g}; x, t), u(0, \overline{g}; x, t)]$$

and problem (2.45)–(2.47) is reduced to two problems with point initial conditions given by $\underline{g} = (\underline{g}_1, \underline{g}_2)$ and $\overline{g} = (\overline{g}_1, \overline{g}_2)$ as follows:

$$\begin{array}{l} u_{tt}(x,t) - u_{xx}(x,t) = f(x,t,u(x,t)) \ , \quad x \in \mathcal{R} \ , \quad t \in [0, \overline{t}] \\ u(x,0) = \underline{g}_1(x) \ , \ u_t(x,0) = \underline{g}_2(x) \ , \quad x \in \mathcal{R} \end{array} \tag{2.49}$$

$$\begin{array}{l} u_{tt}(x,t) - u_{xx}(x,t) = f(x,t,u(x,t)) \ , \quad x \in \mathcal{R} \ , \quad t \in [0, \overline{t}] \\ u(x,0) = \overline{g}_1(x) \ , \ u_t(x,0) = \overline{g}_2(x) \ , \quad x \in \mathcal{R} \end{array} \tag{2.50}$$

However, the practical application of the monotonicity of the form (2.48) to the construction of enclosures has a significant shortcoming when the enclosures are constructed step-by-step using a mesh $\{t_0 = 0, t_1, \ldots, t_{\overline{k}} = \overline{t}\}$ in the time dimension. When this approach is used, the numerical solution (or enclosure in our case) computed at t_1 gives the initial condition for the computation of a numerical solution (enclosure) in the interval $[t_1, t_2]$. However, this initial condition does not have the same properties as the condition G at t_0. We will explain this in more detail.

Denote by $S(h, N; x, t) = [\underline{s}(h, N; x, t), \overline{s}(h, N; x, t)]$ the enclosures produced by a method using a mesh $\{t_0 = 0, t_1, \ldots, t_{\overline{k}} = \overline{t}\}$ in the time direction, where $h = (h_1, h_2, \ldots, h_{\overline{k}})$,

$h_k = t_k - t_{k-1}$ and N is another parameter of the method resulting from a discretization in the space dimension. Let also $S(h,N)$ be constructed in the interval $[t_0,t_1]$ in a such way that

$$\begin{aligned} &L\underline{s}(h,N;x,t) \le f(x,t,\underline{s}(h,N;x,t)) \ , \ \ x \in \mathcal{R} \ , \ \ t_0 < t < t_1 \\ &\underline{s}(h,N;x,t_0) = \underline{g}_1(x) \ , \ \ \underline{s}_t(h,N;x,t_0) = \underline{g}_2(x) \ , \ x \in \mathcal{R} \end{aligned} \tag{2.51}$$

$$\begin{aligned} &L\overline{s}(h,N;x,t) \ge f(x,t,\overline{s}(h,N;x,t)) \ , \ \ x \in \mathcal{R} \ , \ \ t_0 < t < t_1 \\ &\overline{s}(h,N;x,t_0) = \overline{g}_1(x) \ , \ \ \overline{s}_t(h,N;x,t_0) = \overline{g}_2(x) \ , \ x \in \mathcal{R} \end{aligned} \tag{2.52}$$

Then in the interval $[t_0,t_1]$ we have

$$T(t_0,\underline{s}(h,N)) \le T(t_0,u(0,\underline{g})) \quad \text{and} \quad T(t_0,\overline{s}(h,N)) \ge T(t_0,u(0,\overline{g})) \ . \tag{2.53}$$

The monotone property (2.48) implies that

$$\underline{s}(h,N;x,t) \le u(0,\underline{g};x,t) \quad \text{and} \quad \overline{s}(h,N,x,t) \ge u(0,\underline{g};x,t) \ \ , \ \ \ x \in \mathcal{R}, \ t \in [t_0,t_1]$$

and therefore

$$[u(0,G;x,t) = [u(0,\underline{g};x,t), u(0,\overline{g};x,t)] \subset S(h,N,x,t) \ , \ \ x \in \mathcal{R}, \ t \in [t_0,t_1].$$

In the interval $[t_1,t_2]$ we consider the pair of problems

$$\begin{aligned} &u_{tt}(x,t) - u_{xx}(x,t) = f(x,t,u(x,t)) \ , \ \ x \in \mathcal{R} \ , \ \ t_1 < t < t_2 \\ &u(x,t_1) = \underline{s}(h,N;x,t_1) \ , \ u_t(x,t_1) = \underline{s}_t(h,N;x,t_1) \ , \ \ x \in \mathcal{R} \end{aligned} \tag{2.54}$$

$$\begin{aligned} &u_{tt}(x,t) - u_{xx}(x,t) = f(x,t,u(x,t)) \ , \ \ x \in \mathcal{R} \ , \ \ t_1 < t < t_2 \\ &u(x,t_1) = \overline{s}(h,N;x,t_1) \ , \ u_t(x,t_1) = \overline{s}_t(h,N;x,t_1) \ , \ \ x \in \mathcal{R} \end{aligned} \tag{2.55}$$

If the interval function $S(h,N)$ is constructed in the interval $[t_1,t_2]$ in such a way that

$$\begin{aligned} &L\underline{s}(h,N;x,t) \le f(x,t,\underline{s}(h,N;x,t)) \ , \ \ x \in \mathcal{R} \ , \ \ t_1 < t < t_2 \\ &L\overline{s}(h,N;x,t) \ge f(x,t,\overline{s}(h,N;x,t)) \ , \ \ x \in \mathcal{R} \ , \ \ t_1 < t < t_2 \end{aligned}$$

and $\underline{s}(h,N)$, $\overline{s}(h,N)$ satisfy the initial conditions of problems (2.54) and (2.55) respectively then the inequalities

$$T(t_1,\underline{s}(h,N)) \le T(t_1,u(0,\underline{g})) \quad \text{and} \quad T(t_1,\overline{s}(h,N)) \ge T(t_1,u(0,\overline{g}))$$

are not necessarily true because at $t = t_1$ the inequalities

$$\underline{s}_t(h,N;x,t_1) \le u_t(0,\underline{g};x,t_1) \quad \text{and} \quad \overline{s}_t(h,N,x,t) \ge u_t(0,\underline{g};x,t_1) \ \ , \ \ \ x \in \mathcal{R}$$

may be violated. The above inequalities do not follow from (2.53) and the monotonicity of the operator $T(t_0)$ and one can easily show examples where they are not true.

In order to establish a suitable monotone property for the problem (2.45)-(2.47) we define a new operator associated with problem (2.45)-(2.47). Operator $\mathcal{T}(t\alpha)$ is defined on $\Omega[t_\alpha, \overline{t}]$, $t\alpha \in [t_0, \overline{t})$ as follows:

$$\mathcal{T}(t_\alpha, u; x, t, y, z)) = (Lu - f(x,t,u), \Phi(u, t_\alpha; y, z)) \ , \ \ t \in [t_\alpha, \overline{t}], \ x, y, z \in \mathcal{R}, \ y \le z \quad (2.56)$$

where

$$\Phi(u, t; y, z) = u(y,t) + u(z,t) + \int_y^z u_t(x,t)dx \ , \ \ y, z \in \mathcal{R}, \ y \le z$$

In Chapter 4 we prove that

$$\mathcal{T}(t_\alpha, u) \le \mathcal{T}(t_\alpha, v) \implies \Phi(u,t) \le \Phi(v,t) \ , \ \ t_\alpha \le t \le \overline{t} \quad (2.57)$$

Let us note that for any $u, v \in \Omega[t_0, \overline{t}]$ and $t \in [t_0, \overline{t}]$ we have

$$(u(x,t) \le v(x,t), \ u_t(x,t) \le v_t(x,t), \ x \in \mathcal{R}) \implies (\Phi(u,t;y,z) \le \Phi(v,t;y,z), \ y, z \in \mathcal{R}, y \le z)$$
$$(\Phi(u,t;y,z) \le \Phi(v,t;y,z), \ y, z \in \mathcal{R}, y \le z) \implies (u(x,t) \le v(x,t), \ x \in \mathcal{R})$$

but the implication

$$(\Phi(u,t;y,z) \le \Phi(v,t;y,z), \ y, z \in \mathcal{R}, y \le z) \implies (u_t(x,t) \le v_t(x,t) \ x \in \mathcal{R})$$

is false. Then, it is easy to see that the operator $\mathcal{T}(t_\alpha)$ is an operator of monotone type according to the usual definition, but it is actually more then that since the inequality $\Phi(u,t) \le \Phi(v,t)$ contains more information then $u(x,t) \le v(x,t)$, $x \in \mathcal{R}$.

Let the enclosure $S(h,N)$ be constructed in the interval $[t_0, t_1]$ in such a way that the inequalities (2.51), (2.52) are satisfied. Then in the interval $[t_0, t_1]$ we have

$$\mathcal{T}(t_0, \underline{s}(h,N)) \le \mathcal{T}(t_0, u(0, \underline{g})) \quad \text{and} \quad \mathcal{T}(t_0, \overline{s}(h,N)) \ge \mathcal{T}(t_0, u(0, \overline{g})) \ .$$

The monotone property (2.57) implies not only

$$\underline{s}(h, N; x, t) \le u(0, \underline{g}; x, t) \quad \text{and} \quad \overline{s}(h, N, x, t) \ge u(0, \underline{g}; x, t) \ , \ \ x \in \mathcal{R}, \ t \in [t_0, t_1]$$

but also

$$\Phi(\underline{s}(h,N), t_1) \le \Phi(u(0, \underline{g}), t_1) \quad \text{and} \quad \Phi(\overline{s}(h,N), t_1) \ge \Phi(u(0, \underline{g}), t_1) \ .$$

Therefore, if the interval function $S(h,N)$ is constructed in the interval $[t_1, t_2]$ in such a way that

$$L\underline{s}(h, N; x, t) \le f(x, t, \underline{s}(h, N; x, t)) \ , \ \ x \in \mathcal{R} \ , \ \ t_1 < t < t_2$$
$$L\overline{s}(h, N; x, t) \ge f(x, t, \overline{s}(h, N; x, t)) \ , \ \ x \in \mathcal{R} \ , \ \ t_1 < t < t_2$$

and $\underline{s}(h,N)$, $\overline{s}(h,N)$ satisfy the initial conditions of problems (2.54) and (2.55), respectively, then

$$\mathcal{T}(t_1,\underline{s}(h,N)) \leq \mathcal{T}(t_1,u(0,\underline{g})) \quad \text{and} \quad \mathcal{T}(t_1,\overline{s}(h,N)) \geq \mathcal{T}(t_1,u(0,\overline{g}))\,.$$

From (2.57) we obtain

$$[u(0,G;x,t) = [u(0,\underline{g};x,t),u(0,\overline{g};x,t)] \subset S(h,N,x,t)\,,\ \ x \in \mathcal{R},\ t \in [t_1,t_2].$$

and also

$$\Phi(\underline{s}(h,N),t_2) \leq \Phi(u(0,\underline{g}),t_2) \quad \text{and} \quad \Phi(\overline{s}(h,N),t_2) \geq \Phi(u(0,\underline{g}),t_2)\,.$$

Hence we can proceed in the same manner in the interval $[t_2,t_3]$ and further along the mesh.

2.6 Functoids.

2.6.1 The Concept of Functoid.

Functoid is a structure resulting from the ultra-arithmetical approach to the solution of problems in functional spaces. The aim of the ultra arithmetic is the development of structures, data types and operations corresponding to functions for direct digital implementation. On a digital computer equipped with ultra-arithmetic, problems associated with functions will be solvable, just as now we solve algebraic problems [38]. Ultra-arithmetic is developed in analogy with the development of computer arithmetic.

Let $\mathcal{M}$ be a space of functions and let M be a finite dimensional subspace spanned by $\Phi_N = \{\varphi_k\}_{k=0}^N$. Every function $f \in \mathcal{M}$ is approximated by $\tau_N(f) \in M$. The mapping τ_N is called rounding (in analogy with the rounding of numbers) and the space M is called a screen of $\mathcal{M}$. Every rounding must satisfy the following requirement (invariance of rounding on the screen):

$$\tau_N(f) = f \ \text{ for every } \ f \in M$$

Every function $f = \sum_{i=0}^{N} \alpha_i\varphi_i \in M$ can be represented by its coefficient vector $\nu(f) = (\alpha_0,\alpha_1,\ldots,\alpha_N)$. Therefore the approximation of the functions in $\mathcal{M}$ is realized through the mappings

$$\mathcal{M} \xrightarrow{\tau_N} M \xleftrightarrow{\nu} K^{N+1}$$

where K is the scalar field of $\mathcal{M}$ (i.e. $K = \mathcal{R}$ or $K = \mathcal{C}$). Since ν is a bijection we can identify M and K^{N+1} and consider only the rounding τ_N.

In $\mathcal{M}$ we consider the operations $+, -, ., /, \int$ defined in the conventional way. By the semimorphism principle τ_N induces corresponding operations in M:

$$f \boxed{\circ} g = \tau_N(f \circ g) \ , \quad \circ \in \{+, -, \times, /\}$$
$$\boxed{\int} f = \tau_N\left(\int f\right)$$

The structure $(M, \boxplus, \boxminus, \boxdot, \boxslash, \boxed{\int})$ is called an (ultra-arithmetical) functoid.[49]

Let IM be the set of all linear combinations of the basis Φ_N taken with interval coefficients, i.e.

$$IM = Isp(\Phi_N) = \left\{\sum_{i=0}^{N} A_k \varphi_i : A_i \in \mathcal{I}K\right\}$$

Every $F \in IM$ can also be considered as an interval function and can be identified as follows

$$F = \sum_{i=0}^{N} A_k \varphi_i = \left\{f \in \mathcal{M} : f(x) \in \sum_{i=0}^{N} A_k \varphi_i(x), x \in D\right\}$$

where D is the domain of the functions in $\mathcal{M}$. In this way IM belongs to the power set $P\mathcal{M}$ of $\mathcal{M}$. A mapping $I\tau_N : P\mathcal{M} \mapsto IM$ is called an interval rounding if

$$f \in I\tau_N(f) \text{ for every } f \in \mathcal{M}$$
$$F \subset I\tau_N(F) \text{ for every } F \in P\mathcal{M}$$
$$F = I\tau_N(F) \text{ for every } F \in IM$$

The operations in $P\mathcal{M}$ are defined by

$$F \circ G = \{f \circ g : f \in F, g \in G\} \ , \quad F, G \in P\mathcal{M} \ , \quad \circ \in \{+, -, ., /\}$$
$$\int F = \left\{\int f : f \in F\right\} \ , \quad F \in P\mathcal{M}$$

Operations in IM are defined using again the semimorphism property:

$$F \diamondsuit G = I\tau_N(F \circ G) \ , \quad F, G \in IM \ , \quad \circ \in \{+, -, ., /\}$$
$$\diamondsuit F = I\tau_N\left(\int F\right)$$

The structure $\left(IM, \diamondsuit, \diamondsuit, \diamondsuit, \diamondsuit, \diamondsuit\right)$ is called an interval functoid.

Typically, in $\mathcal{M}$ we have a basis $\Psi = \{\psi_k\}_{k=0}^{\infty}$ and τ_N and $I\tau_N$ are linear projections. Let

$$\tau_N(\psi_i) = \sum_{j=0}^{N} a_{ij} \varphi_j$$

Then for every function $f = \Psi.\alpha = \sum_{i=0}^{\infty} \alpha_i \psi_i \in \mathcal{M}$ we have

$$\tau_N(f) = \sum_{i=0}^{\infty} \alpha_i \tau_N(\psi_i) = \Phi_N.(\alpha A)$$

where A is a $\infty \times (N+1)$ matrix with entries a_{ij}, $i = 0, \ldots, \infty$, $j = 0, \ldots, N$ and $\alpha = (\alpha_0, \alpha_1, \ldots)$.

Methods of approximation theory provide estimates of the form

$$\max_{x \in D} |\psi_i - \tau_N(\psi_i)| \leq \sigma_i \ , \quad i = 0, 1, \ldots$$

for many classes of bases. From such an estimate interval rounding may be defined as follows

$$I\tau_N(\psi_i) = \tau_N(\psi_i) + [-1, 1]\sigma_i \ , \quad i = 0, 1, \ldots$$

$$I\tau_N(f) = \sum_{i=0}^{\infty} (\alpha_i \tau_N(\psi_i) + [-1, 1]\sigma_i) = \Phi_N.(\alpha A) + [-1, 1](\alpha.\sigma)$$

where $\sigma = (\sigma_0, \sigma_1, \ldots)$. In the canonical case, where $\psi_i = \varphi_i$, $i = 0, 1, \ldots, N$, the rounding τ_N is often defined by

$$\tau_N(\psi_i) = \begin{cases} \varphi_i & , \quad i \leq N \\ 0 & , \quad i > 0 \end{cases}$$

which is equivalent to terminating the series, i.e.

$$\tau_N(f) = \sum_{i=0}^{N} \alpha_i \varphi_i$$

For interval rounding in this case we have $\sigma_i = 0$, $i = 0, 1, \ldots, N$, $\max_{x \in D} |\psi_i(x)| \leq \sigma_i$, $i = N+1, N+2, \ldots$. and

$$I\tau_N(f) = \sum_{i=0}^{N} \alpha_i \varphi_i + [-1, 1] \sum_{i=N+1}^{\infty} \alpha_i \sigma_i$$

A typical example of a functoid is the Taylor functoid where $\mathcal{M}$ is the set of all entire functions considered on some compact interval (e.g.$[-1, 1]$) and M is the span of $\{1, x, x^2, \ldots, x^N\}$. The Taylor rounding is defined by $\tau_N(x^k) = x^k$ for $k = 0, 1, \ldots, N$ and $\tau_N(x^k) = 0$ for $k > N$. If the domain of the function is the interval $[-1, 1]$ the interval Taylor rounding is defined by using $\sigma_k = 0$ for $k = 0, 1, \ldots, N$ and $\sigma_k = 1$ for $k > N$. The Taylor functoid M can also be considered as a screen for the set of all functions with a bounded $N+1$ derivative on $[-1, 1]$. In what follows the symbol τ denotes the Taylor rounding if not otherwise stated. The symbol ρ will be used for the rounding in the Fourier functoid discussed in the next subsection.

2.6.2 Fourier Functoid.

Let $f \in L_2(-1,1)$. The Fourier series of f can be represented as

$$f(x) \sim a_0 + \sum_{k=1}^{\infty}(a_k \cos(k\pi x) + b_k \sin(k\pi x)) = \sum_{k=0}^{\infty}(a_k \cos(k\pi x) + b_k \sin(k\pi x))$$

where

$$a_0 = a_0(f) = \frac{1}{2}\int_{-1}^{1} f(x)dx, \;\; b_0 = b_0(f) = 0$$
$$a_k = a_k(f) = \int_{-1}^{1} f(x)\cos(k\pi x)dx, \;\; b_k = b_k(f) = \int_{-1}^{1} f(x)\sin(k\pi x)dx, \;\; k = 1,2,\ldots$$

The complex form of the Fourier series will be used in some applications:

$$f \sim \sum_{k=-\infty}^{\infty} c_k e^{ik\pi x}$$

where $c_0 = a_0$, $c_k = \frac{1}{2}(a_k - ib_k)$, $c_{-k} = \frac{1}{2}(a_k + ib_k) = conj(c_k)$, $k = 1,2,\ldots$.

Fourier functoid $\mathcal{F}_N$ is defined as the span of $\{\cos(k\pi x), \sin(k\pi x)\}_{k=0}^{N}$, i.e we have

$$\mathcal{F}_N = \left\{ \sum_{k=0}^{N}(a_k \cos(k\pi x) + b_k \sin(k\pi x)) : a_k, b_k \in \mathcal{R} \right\} .$$

The mapping $\rho_N : L_2(-1,1) \mapsto \mathcal{F}_N$ defined by

$$\rho_N(f) = \sum_{k=0}^{N}(a_k \cos(k\pi x) + b_k \sin(k\pi x))$$

is a rounding of $L_2(-1,1)$ onto the screen $\mathcal{F}_N$. Addition and multiplication by a scalar is implemented in $\mathcal{F}_N$ without rounding

$$\left(\sum_{k=0}^{N}(a_k^{(1)} \cos(k\pi x) + b_k^{(1)} \sin(k\pi x))\right) + \left(\sum_{k=0}^{N}(a_k^{(2)} \cos(k\pi x) + b_k^{(2)} \sin(k\pi x))\right)$$
$$= \sum_{k=0}^{N}((a_k^{(1)} + a_k^{(2)}) \cos(k\pi x) + (b_k^{(1)} + b_k^{(2)}) \sin(k\pi x))$$
$$\alpha\left(\sum_{k=0}^{N}(a_k \cos(k\pi x) + b_k \sin(k\pi x))\right) = \left(\sum_{k=0}^{N}(\alpha a_k \cos(k\pi x) + \alpha b_k \sin(k\pi x))\right)$$

The product of two functions of $\mathcal{F}_N$ can be represented in the following way using the complex form of the Fourier series

$$\left(\sum_{k=-N}^{N} c_k e^{ik\pi x}\right)\left(\sum_{k=-N}^{N} d_k e^{ik\pi x}\right) = \left(\sum_{k=-2N}^{2N} \sum_{j=\max\{-N,k-N\}}^{\min\{N,k+N\}} c_{k-j} d_j e^{ik\pi x}\right)$$

and after rounding we have

$$\left(\sum_{k=-N}^{N} c_k e^{ik\pi x}\right) \boxdot \left(\sum_{k=-N}^{N} d_k e^{ik\pi x}\right) = \left(\sum_{k=-N}^{N} \sum_{j=\max\{-N,k-N\}}^{\min\{N,k+N\}} c_{k-j} d_j e^{ik\pi x}\right)$$

This leads to the following formula for multiplication in $\mathcal{F}_N$ in real form

$$\left(\sum_{k=0}^{N}(a_k^{(1)}\cos(k\pi x) + b_k^{(1)}\sin(k\pi x))\right) \boxdot \left(\sum_{k=0}^{N}(a_k^{(2)}\cos(k\pi x) + b_k^{(2)}\sin(k\pi x))\right)$$

$$= \sum_{k=0}^{N}\left(\left(\sum_{j=k-N}^{N}(a_{k-j}^{(1)}a_j^{(2)} - b_{k-j}^{(1)}b_j^{(2)})\right)\cos(k\pi x) + \left(\sum_{j=k-N}^{N}(a_{k-j}^{(1)}b_j^{(2)} + b_{k-j}^{(1)}a_j^{(2)})\right)\sin(k\pi x)\right)$$

where $a_{-j} = a_j$ and $b_{-j} = -b_j$, $j = 1, 2, \ldots, N$.

There is no explicit formula for division and an iterative procedure for computing the quotient is proposed [49]. Since division will not be used in the numerical methods discussed in the thesis we will omit it.

Integration of functions $f \in \mathcal{F}_N$ such that $\int_{-1}^{1} f(x)dx = 0$ (i.e. the constant term in the series is zero) is implementable without rounding.

$$\int\left(\sum_{k=1}^{N}(a_k\cos(k\pi x) + b_k\sin(k\pi x))\right)dx = d_0 + \sum_{k=1}^{N}(-\frac{b_k}{k\pi}\cos(k\pi x) + \frac{a_k}{k\pi}\sin(k\pi x))$$

The integration of a constant is a problem. The set $\mathcal{F}_N$ consists of periodical functions but a constant has no periodical (period 2) antiderivative. We have $\int_0^x 1d\xi = x$, $x \in [-1, 1]$. Denote by s_1 the periodical extension (period 2) of the function x over the whole real line, i.e. s_1 is periodical with period 2 and $s_1(x) = x$ for $x \in (-1, 1]$. Then for any $\alpha \in (-1, 1]$ the function $s_1(x + \alpha)$ can be considered as a generalized antiderivative of 1. The value of α (i.e. the position of the jump) must be determined from additional conditions. This problem can really be resolved only by introducing the Fourier hyper functoid discussed in the next section. However, we can still evaluate definite integrals on $[-1, 1]$ without difficulties. For example, using

$$s_1(x) = 2\sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{k\pi}\sin(k\pi x)$$

we have

$$\left\langle\!\!\int\right\rangle_0^x \left(a_0 + \sum_{k=1}^{N}(a_k\cos(k\pi\xi) + b_k\sin(k\pi\xi))\right)d\xi$$

$$= \rho_N\left(2a_0\sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{k\pi}\sin(k\pi x) + \sum_{k=1}^{N}(-\frac{b_k}{k\pi}\cos(k\pi x) + \frac{a_k}{k\pi}\sin(k\pi x)) + \sum_{k=1}^{N}\frac{b_k}{k\pi}\right)$$

$$= 2\sum_{k=1}^{N}\frac{b_k}{k\pi} + \sum_{k=1}^{N}\left(-\frac{b_k}{k\pi}\cos(k\pi x) + \frac{a_k + 2a_0(-1)^{k-1}}{k\pi}\sin(k\pi x)\right),\ x \in [0, 1]$$

The interval Fourier functoid is

$$\mathcal{IF}_N = \left\{ \sum_{k=0}^{N} (A_k \cos(k\pi x) + B_k \sin(k\pi x)) : A_k, B_k \in \mathcal{IR} \right\}$$

In order to define an interval rounding $I\rho_N$ we need to have an estimate of the form

$$\max_{x\in[-1,1]} |f(x) - \rho_N(f)(x)| \leq \sigma_N(f) \to 0 \text{ when } N \to \infty .$$

This implies that $\rho_N(f)$ converges uniformly to f. However, in general, $\rho_N(f)$ converges to $f \in L_2(-1,1)$ only in the L_2 norm. Therefore $\mathcal{IF}$ can not be an interval screen of $L_2(-1,1)$ but it can be an interval screen of a subset of $L_2(-1,1)$, consisting of functions with uniformly convergent Fourier series.

If the Fourier series of $f \in \mathcal{M}$ is finite, it is easy to define interval rounding using the general approach from the previous section. Let $f(x) = \sum_{k=0}^{P} (a_k \cos(k\pi x) + b_k \sin(k\pi x))$. Then

$$|f(x) - \rho_N(f)(x)| \leq \sum_{k=N+1}^{P} \sqrt{a_k^2 + b_k^2}$$

and

$$I\rho_N(f)(x) = a_0 + [-1,1] \sum_{k=N+1}^{P} \sqrt{a_k^2 + b_k^2} + \sum_{k=0}^{N} (a_k \cos(k\pi x) + b_k \sin(k\pi x))$$

The above formula can be extended for rounding from $\mathcal{IF}_P$ to $\mathcal{IF}_N$, $P > N$.

$$I\rho_N \left(\sum_{k=0}^{P} (A_k \cos(k\pi x) + B_k \sin(k\pi x)) \right)$$
$$= A_0 + [-1,1] \sum_{k=N+1}^{P} \sqrt{A_k^2 + B_k^2} + \sum_{k=0}^{N} (A_k \cos(k\pi x) + B_k \sin(k\pi x)) \quad (2.58)$$

Addition, multiplication by interval and integration of a sum with no constant term are implemented without rounding using the same formulas as in $\mathcal{F}$ replacing a_k by A_k, b_k by B_k and α by $[\underline{\alpha}, \overline{\alpha}]$. The product of $F(x) = \sum_{k=0}^{N} (A_k^{(1)} \cos(k\pi x) + B_k^{(1)} \sin(k\pi x))$ and $G(x) = \sum_{k=0}^{N} (A_k^{(2)} \cos(k\pi x) + B_k^{(2)} \sin(k\pi x))$ is

$$(FG)(x) = \sum_{k=0}^{2N} (A_k^{(3)} \cos(k\pi x) + B_k^{(3)} \sin(k\pi x)) \in \mathcal{F}_{2N}$$

where

$$A_k^{(3)} = \sum_{j=k-N}^{N} (A_{k-j}^{(1)} A_j^{(2)} - B_{k-j}^{(1)} B_j^{(2)})$$
$$B_k^{(3)} = \sum_{j=k-N}^{N} (A_{k-j}^{(1)} B_j^{(2)} + B_{k-j}^{(1)} A_j^{(2)})$$

assuming that $A_{-k} = A_k$ and $B_{-k} = -B_k$, $k = 1, 2, \ldots, N$. Using the rounding in the form (2.58) we have

$$\left(F \diamondsuit G\right)(x) = A_0^{(3)} + [-1,1]\sigma_N(FG) + \sum_{k=1}^{N} (A_k^{(3)} \cos(k\pi x) + B_k^{(3)} \sin(k\pi x))$$

where $\sigma_N(FG) = \sum_{k=N+1}^{2N} \sqrt{\left(A_k^{(3)}\right)^2 + \left(B_k^{(3)}\right)^2}$

The integral of a constant is not a function which has a uniformly convergent Fourier series. Therefore integration is defined in $\mathcal{IF}$ only for functions with zero constant term. However, definite integrals in $[-1, 1]$ can still be evaluated similarly to the definite integrals in $\mathcal{F}$.

Determining a subset $\mathcal{M}$ of $L_2(-1, 1)$ which has $\mathcal{IF}$ as interval screen and extension of the definition of $I\rho_N$ over $\mathcal{M}$ and $P\mathcal{M}$ will be discussed in chapter 4.

Let us note that the interval rounding $I\rho_N$ also defines directed rounding $\underline{\rho}_N$ and $\overline{\rho}_N$ in $\mathcal{M}$. We have

$$I\rho_N(f) = [\underline{\rho}_N(f), \overline{\rho}_N(f)]$$

where

$$\underline{\rho}_N(f) = a_0 - \sigma_N(f) + \sum_{k=1}^{N} (a_k \cos(k\pi x) + b_k \sin(k\pi x))$$
$$\overline{\rho}_N(f) = a_0 + \sigma_N(f) + \sum_{k=1}^{N} (a_k \cos(k\pi x) + b_k \sin(k\pi x))$$

Using directed roundings the operations in $\mathcal{F}$ can also be defined with rounding to the left or to the right.

2.6.3 Application to the Wave Equation.

One of the inconveniences in using the interval Fourier functoid is that the upper and the lower bounds of

$$\sum_{k=0}^{N} (A_k \cos(k\pi x) + B_k \sin(k\pi x)) \in I\mathcal{F}$$

are not functions in $\mathcal{F}$. In other words $\mathcal{IF}$ is not the interval space over $\mathcal{F}$. Let $\underline{f}, \overline{f} \in \mathcal{M}$, $\underline{f} \leq \overline{f}$ and $F = [\underline{f}, \overline{f}]$. Obviously, $F \in P\mathcal{M}$. We have

$$\begin{aligned}
\underline{\rho}_N(\underline{f}) &= a_0 - \sigma_N(\underline{f}) + \sum_{k=1}^{N} \left(\underline{a}_k \cos(k\pi x) + \underline{b}_k \sin(k\pi x)\right) \\
\overline{\rho}_N(\overline{f}) &= a_0 + \sigma_N(\overline{f}) + \sum_{k=1}^{N} \left(\overline{a}_k \cos(k\pi x) + \overline{b}_k \sin(k\pi x)\right)
\end{aligned}$$

Since the inequalities $\underline{a}_k \leq \overline{a}_k$, $\underline{b}_k \leq \overline{b}_k$ are not necessarily true the interval function $[\underline{\rho}_N(\underline{f}), \overline{\rho}_N(\overline{f})]$ is not, in general, an element of $\mathcal{IF}$. It is easy to see that

$$\begin{aligned}
[\underline{\rho}_N(\underline{f}), \overline{\rho}_N(\overline{f})] \subset\ & (\underline{a}_0 - \sigma_N(\underline{f})) \vee (\overline{a}_0 + \sigma_N(\overline{f})) \\
& + \sum_{k=1}^{N} \left((\underline{a}_k \vee \overline{a}_k)\cos(k\pi x) + (\underline{b}_k \vee \overline{b}_k)\sin(k\pi x)\right) \subset I\rho_N(F)
\end{aligned}$$

We have

$$\begin{aligned}
w([\underline{\rho}_N(\underline{f}), \overline{\rho}_N(\overline{f})]) &= \overline{\rho}_N(\overline{f}) - \underline{\rho}_N(\underline{f}) \\
&= \overline{a}_0 - \underline{a}_0 + \sigma_N(\underline{f}) + \sigma_N(\overline{f}) + \sum_{k=1}^{N} \left((\overline{a}_k - \underline{a}_k)\cos(k\pi x) + (\overline{b}_k - \underline{b}_k)\sin(k\pi x)\right)
\end{aligned}$$

while

$$w(I\rho_N(F)) \geq |\overline{a}_0 - \underline{a}_0 + \sigma_N(\underline{f}) + \sigma_N(\overline{f})| + \sum_{k=1}^{N} \left(|\overline{a}_k - \underline{a}_k||\cos(k\pi x)| + |\overline{b}_k - \underline{b}_k||\sin(k\pi x)|\right)$$

It is obvious that there is, in general, a significant difference between the width of $[\underline{\rho}_N(\underline{f}), \overline{\rho}_N(\overline{f})]$ and the width of $I\rho_N(F)$ which increases when N increases.

For that reason, we will not apply the interval Fourier functoid $\mathcal{IF}$ for approximation of interval functions. Instead, we will use the directed roundings $\underline{\rho}_N$ and $\overline{\rho}_N$ to obtain lower and upper bounds $\underline{\rho}_N(\underline{f})$, $\overline{\rho}_N(\overline{f})$ which are elements of $\mathcal{F}$.

The formulation of problem (2.45)–(2.47) as two problems (2.49) and (2.50) facilitates the above approach. We need to calculate a lower bound for the solution of (2.49) and an upper bound for the solution of (2.50).

In chapter 4 we consider a numerical method which uses this approach for producing lower and upper bounds $\underline{s}(h, N)$, $\overline{s}(h, N)$ for the solution $u(0, G; x, t)$ in the form of Fourier series about x:

$$\begin{aligned}
\underline{s}(h, N; x, t) &= \sum_{k=0}^{N} \left(\underline{a}_k(t)\cos(k\pi x) + \underline{b}_k(t)\sin(k\pi x)\right) \\
\overline{s}(h, N; x, t) &= \sum_{k=0}^{N} \left(\overline{a}_k(t)\cos(k\pi x) + \overline{b}_k(t)\sin(k\pi x)\right) \qquad (2.59)
\end{aligned}$$

where the coefficients $\underline{a}_k, \overline{a}_k, \underline{b}_k, \overline{b}_k$ are piece-wise polynomials about t. The bounds are constructed step by step using a mesh $\{t_j = jh : j = 0, 1, \ldots, \bar{j}\}$ in the time dimension.

In every interval $[t_j, t_{j+1}]$ we consider a pair of problems

$$\begin{aligned} &u_{tt}(x,t) - u_{xx}(x,t) = \underset{\sim}{f}(x,t,u(x,t)) \,, \quad x \in \mathcal{R} \,, \quad t > 0 \\ &u(x,t_j) = \underset{\sim}{g}_{j1}(x) \,, \; u_t(x,t_j) = \underset{\sim}{g}_{j2}(x) \,, \quad x \in \mathcal{R} \end{aligned} \tag{2.60}$$

$$\begin{aligned} &u_{tt}(x,t) - u_{xx}(x,t) = \tilde{f}(x,t,u(x,t)) \,, \quad x \in \mathcal{R} \,, \quad t > 0 \\ &u(x,0) = \tilde{g}_{j1}(x) \,, \; u_t(x,0) = \tilde{g}_{j2}(x) \,, \quad x \in \mathcal{R} \end{aligned} \tag{2.61}$$

where for $j = 0$

$$\underset{\sim}{g}_{01} = \underline{\rho}_N(\underline{g}_1) \,, \; \tilde{g}_{01} = \overline{\rho}_N(\overline{g}_1) \,, \; \underset{\sim}{g}_{02} = \underline{\rho}_N(\underline{g}_2), \; \tilde{g}_{02} = \overline{\rho}_N(\overline{g}_2)$$

and for $j \geq 1$

$$\begin{aligned} &\underset{\sim}{g}_{j1}(x) = \underline{s}(h, N; x, t_j) \,, \; \tilde{g}_{j1}(x) = \overline{s}(h, N; x, t_j) \,, \\ &\underset{\sim}{g}_{j2}(x) = \underline{s}_t(h, N; x, t_j), \; \tilde{g}_{j2}(x) = \overline{s}_t(h, N; x, t_j) \,, \quad x \in \mathcal{R} \end{aligned}$$

assuming that the bounds $\underline{s}(h, N; x, t)$, $\overline{s}(h, N; x, t)$ are already computed for $t \leq t_j$. The functions $\underset{\sim}{f}(x,t,u(x,t))$ and $\tilde{f}(x,t,u(x,t))$ are bounds for $f(x,t,u(x,t))$ which are of the same form as the required form (2.59) for the bounds of the solution. We obtain $\underset{\sim}{f}(x,t,u(x,t))$ and $\tilde{f}(x,t,u(x,t))$ using directed Fourier rounding about x and then using directed Taylor rounding for the coefficients.

The solutions $\underset{\sim}{u}$ and $\tilde{u}$ of problems (2.60) and (2.61) are approximated by iterative procedures producing sequences

$$\underset{\sim}{u}^{(0)}, \underset{\sim}{u}^{(1)}, \underset{\sim}{u}^{(2)}, \ldots \quad , \quad \tilde{u}^{(0)}, \tilde{u}^{(1)}, \tilde{u}^{(2)}, \ldots$$

where $\underset{\sim}{u}^{(0)}$ and $\tilde{u}^{(0)}$ are some suitable initial approximations. For every $r \geq 0$ the functions $\underset{\sim}{u}^{(r)}$ and $\tilde{u}^{(r)}$ are substituted in the right hand sides of (2.60) and (2.61) respectively and $\underset{\sim}{u}^{(r+1)}$ and $\tilde{u}^{(r+1)}$ are the solutions of the problems obtained in this way. The practical implementation of this procedure involves computations in the Cartesian product of the Taylor Functoid and the Fourier Functoid. An essential part of the computations is the evaluation of integrals of the form

$$\iint\limits_{\Gamma(x,\Delta,t)} \frac{\theta^q}{q!} \cos(k\pi y) dy d\theta \quad \text{and} \quad \iint\limits_{\Gamma(x,\Delta,t)} \frac{\theta^q}{q!} \cos(k\pi y) dy d\theta$$

where $\Gamma(x, \Delta, t)$ is the triangle with vertices $(x, t+\Delta)$, $(x-\Delta, t)$ and $(x+\Delta, t)$. Suitable formulae for the evaluation of the above integrals are derived in chapter 4.

Using the monotone properties discussed in section 1.6 we prove that

$$\underline{u}^{(0)} \leq \underline{u}^{(1)} \leq \ldots \leq \underline{u}^{(l)} \leq \ldots \leq u(0, \underline{g}) \leq u(0, \overline{g}) \leq \ldots \leq \tilde{u}^{(l)} \leq \ldots \tilde{u}^{(1)} \leq \tilde{u}^{(0)}$$

After a sufficient number of iterations r^* we take

$$\underline{s}(h, N; x, t) = \underline{u}^{(r^*)}(x, t)\ ,\ \ \overline{s}(h, N; x, t) = \tilde{u}^{(r^*)}(x, t)\ ,\ \ x \in \mathcal{R}\ ,\ \ t \in [t_j, t_{j+1}].$$

An a priori estimate, although not needed to determine the accuracy of a particular numerical solution, characterizes the quality of the method. Using standard techniques one can see that the global error is

$$o(N^{\frac{1}{2}-j}) + O(h^{m+1})$$

provided Taylor rounding τ_m and Fourier rounding ρ_N are applicable and functions g_1, g_2, f have derivatives of order j in $L_2(-1, 1)$. The accuracy of the bounds obtained in the numerical examples is consistent with that estimate.

2.7 Fourier Hyper Functoid.

We saw in the previous section that the Fourier functoid has an internal problem with the integration of a constant. In addition to that, the interval Fourier functoid is a screen only for differentiable functions and the rounding error $\sigma_N(f)$ converges towards zero very slowly when the first or second derivative is discontinuous. These problems led to the introduction of the Fourier hyper functoid [48]. In general, we call a functoid M a *hyper functoid* if it involves infinite series of the basis $\{\psi_i\}$ of $\mathcal{M}$. A Fourier hyper functoid is defined in [48] using an ansatz for the coefficient space. Here we will apply a more direct approach giving the functions in the basis of M explicitly. We have already considered the function s_1 defined by $s_1(x) = x$ for $x \in (-1, 1]$ and s_1 periodical with period 2 on $\mathcal{R}$. The sequence of functions $s_1, s_2, s_3, \ldots$ is defined recursively by

$$s_j' = s_{j-1}\ ,\ \ \int_{-1}^{1} s_j(x)dx = 0\ ,\ \ j = 2, 3, \ldots$$

Since these functions are periodical and piece-wise polynomial we will refer to them as periodic splines. We define the screen M as the span of $\{s_j\}_{j=1}^{p} \cup \{\cos(k\pi x), \sin(k\pi x)\}_{k=0}^{N}$. M is a hyper functoid because $s_j, j = 1, \ldots, p$ have infinite Fourier series. We will refer to M as a *Spline-Fourier functoid*. The rounding ρ_{Np} in M is a continuation of ρ_N. We have

$$\rho_{Np}(\sin(k\pi x)) = \rho_N(\sin(k\pi x)),\ \rho_{Np}(\cos(k\pi x)) = \rho_N(\cos(k\pi x)),\ \ k = 0, 1, 2, \ldots$$

$$\rho_{Np}(s_j(x)) = s_j(x)\ ,\quad j = 1,\ldots,p$$

$$\rho_{Np}(s_j(x)) = \sum_{\substack{k=-N\\k\neq 0}}^{N} \frac{(-1)^{k-1}}{(ik\pi)^j} e^{ik\pi} = \begin{cases} 2\sum_{k=1}^{N} \frac{(-1)^{k-1-\frac{j-1}{2}}}{(k\pi)^j} \sin(k\pi x) & ,\ j \text{ - odd} \\ 2\sum_{k=1}^{N} \frac{(-1)^{k-1-\frac{j}{2}}}{(k\pi)^j} \cos(k\pi x) & ,\ j \text{ - even} \end{cases} \quad ,\ j > p$$

In chapter 5 we consider M as a screen for the space

$$\mathcal{M} = H^p(-1,1) = \{f \in C^{p-1}[-1,1] : \frac{d^p f}{dx^p} \in L_2(-1,1)\}$$

If $f \in \mathcal{M}$ this means that the function f and its first $p-1$ derivatives, when extended periodically on $(-\infty,\infty)$, may be discontinuous only at the points $2k+1$, $k \in \mathcal{Z}$. The function $f \in \mathcal{M}$ has a unique representation of the form

$$f(x) = a_0 + \sum_{j=1}^{p} \alpha_j s_j(x) + \sum_{k=1}^{\infty} (a_k \cos(k\pi x) + b_k \sin(k\pi x))$$

where $\sum_{k=1}^{\infty} (a_k \cos(k\pi x) + b_k \sin(k\pi x)) \in C^{p-1}(-\infty,\infty)$.

The rounding $\rho_{Np} : \mathcal{M} \mapsto M$ is defined by

$$\rho_{Np}(f) = a_0 + \sum_{j=1}^{p} \alpha_j s_j(x) + \sum_{k=1}^{N} (a_k \cos(k\pi x) + b_k \sin(k\pi x))$$

For the interval rounding $I\rho_{Np}$ we have

$$I\rho_{Np}(f) = a_0 + [-1,1]\sigma_{Np}(f) + \sum_{j=1}^{p} \alpha_j s_j(x) + \sum_{k=1}^{N} (a_k \cos(k\pi x) + b_k \sin(k\pi x))$$

where $\sigma_{Np}(f) = \sigma_N \left(\sum_{k=N+1}^{\infty} (a_k \cos(k\pi x) + b_k \sin(k\pi x)) \right) = o\left(N^{\frac{1}{2}-p}\right)$. In chapter 5 suitable formulae for computations in the Spline-fourier functoid are derived.

The Fourier hyper functoid is applied to problems of the form (2.45)-(2.47) with data functions that either themselves, or their derivatives, have isolated discontinuities. Discontinuities at points other then $x = 1$ are represented by shifting the interval $[-1,1]$ to the left or to the right. For example, $s_j(x+\gamma)$ will be used to represent a jump of the $j-1$ derivative of a function at the point $x = 1-\gamma$. Lower and upper bounds for the solution $u(0,G;x,t)$ are obtained in the form

$$\underline{s}(h,N;x,t) = \underline{a}_0 + \sum_{l=1}^{\bar{l}} \sum_{j=1}^{p} \sum_{\delta=-1}^{1} \underline{\alpha}_{lj\delta} s_j(x + \delta t + \gamma_l) + \sum_{k=1}^{N} (\underline{a}_k \cos(k\pi x) + \underline{b}_k \sin(k\pi x))$$

$$\overline{s}(h,N;x,t) = \overline{a}_0 + \sum_{l=1}^{\bar{l}} \sum_{j=1}^{p} \sum_{\delta=-1}^{1} \overline{\alpha}_{lj\delta} s_j(x+\delta t+\gamma_l) + \sum_{k=1}^{N} (\overline{a}_k \cos(k\pi x) + \overline{b}_k \sin(k\pi x))$$

The bounds are obtained using an iteration procedure similar to the iteration procedure in chapter 4. In addition to the formulas in chapter 4, formulas for the integrals of the form

$$\iint_{\Gamma(x,\Delta,t)} \theta^q s_j(y+\gamma+\delta\theta)dyd\theta\ ,\quad \delta=-1,0,1,\quad \gamma\in(-1,1]$$

are derived, since the integration over the characteristic triangle $\Gamma(x,\Delta,t)$ is an essential part of the implementation of this procedure.

Chapter 3

Wrapping Effect and Wrapping Function

We consider the initial value problem for ODE as introduced in section 2.4.1, i.e.

$$\dot{x} = f(t, x) \tag{3.1}$$
$$x(t_0) = x^0 \in X^0 \tag{3.2}$$

where $t \in [t_0, \overline{t}] \subset \mathcal{R}$, $x^0 \in \mathcal{R}^n$, $D \subset \mathcal{R}^n$ is an open set, $f : [t_0, \overline{t}] \times D \to \mathcal{R}^n$ and

$$X^0 = ([\underline{x}_1^0, \overline{x}_1^0], [\underline{x}_2^0, \overline{x}_2^0], \ldots, [\underline{x}_n^0, \overline{x}_n^0])^T$$

is an n-dimensional interval vector, $X^0 \subset D$. We assume that f satisfies conditions (2.37).

We consider methods of propagate and wrap type (definition 2.11) producing enclosures $S(h, t)$ for the solution $x(t_0, X^0; t)$ of problem (3.1)–(3.2) using a mesh $\{t_0, t_1, \ldots, t_{\overline{k}} = \overline{t}\}$ where $h = (h_1, h_2, \ldots, h_n)$, $h_k = t_k - t_{k-1}$, $k = 1, \ldots, \overline{k}$. By $\boldsymbol{\mathcal{S}}$ we denote the enclosures produced by the Jackson's IPWA discussed in section 2.4.3.

3.1 Wrapping Function.

The wrapping function $\widehat{X}$ of problem (3.1)–(3.2) was defined in section 2.4.3 as the optimal (tightest) interval function satisfying the wrapping property at every point of the interval $[t_0, \overline{t}]$ and the condition $X(t_0) = X^0$ (definition 2.12).

Here we will represent it as a solution of an initial value problem involving the (natural) interval extension of function f:

$$f(t, X) = (f_1(t, X), f_2(t, X), \ldots, f_n(t, X))^T$$
$$f_i(t, X) = [\underline{f}_i(t, X), \overline{f}_i(t, X)] = [\inf_{x \in X} f_i(t, x), \sup_{x \in X} f_i(t, x)],\ i = 1, ..., n$$

where $X \in \mathcal{I}D = \{X \in \mathcal{I}\mathcal{R}^n : X \subset D\}$.

Conditions (2.37) of function f imply similar properties for its interval extension. In the region $\mathcal{ID}$ the interval extension f

$$\begin{aligned} &i) \quad \text{is bounded: } |f_i(t,X)| \le m_i \in \mathcal{R},\ m = (m_1, m_2, \ldots, m_n)^T \in \mathcal{R}^n\ ; \\ &ii) \quad \text{is continuous about } t\ ; \\ &iii) \quad \text{satisfies Lipschitz condition about } X \text{ in the form:} \\ &\qquad |f(t,Y) \stackrel{-}{-} f(t,Z)| \le \Lambda |Y \stackrel{-}{-} Z| \quad \text{where } \Lambda = (\lambda_{ij}) \in \mathcal{R}^{n\times n}\ . \end{aligned} \tag{3.3}$$

We will also use the following notation. Let $Y \in \mathcal{IR}^n$, then

$$\underline{Y}^i = (Y_1, \ldots, Y_{i-1}, \underline{y}_i, Y_{i+1}, \ldots, Y_n)^T\ ,$$
$$\overline{Y}^i = (Y_1, \ldots, Y_{i-1}, \overline{y}_i, Y_{i+1}, \ldots, Y_n)^T\ .$$

Let $X : [t_0, \overline{t}] \to \mathcal{ID}$ be an interval function. The interval operator $\mathcal{L}$ is defined by

$$\mathcal{L}X(t) = \begin{pmatrix} [\dot{\underline{x}}_1(t) - \underline{f}_1(t, \underline{X}^1(t))\ ,\ \dot{\overline{x}}_1(t) - \overline{f}_1(t, \overline{X}^1(t))] \\ [\dot{\underline{x}}_2(t) - \underline{f}_2(t, \underline{X}^2(t))\ ,\ \dot{\overline{x}}_2(t) - \overline{f}_2(t, \overline{X}^2(t))] \\ \ldots\ldots\ldots\ldots\ldots \\ [\dot{\underline{x}}_n(t) - \underline{f}_n(t, \underline{X}^n(t))\ ,\ \dot{\overline{x}}_n(t) - \overline{f}_n(t, \overline{X}^n(t))] \end{pmatrix}\ ,\quad t \in [t_0, \overline{t}]$$

We consider the following initial value problem

$$\mathcal{L}X = 0 \tag{3.4}$$
$$X(t_0) = X^0 \tag{3.5}$$

From (3.3) it follows that this problem has a unique solution in some interval $[t_0, \sigma]$. For simplicity we will assume that $\sigma = \overline{t}$

Theorem 3.1 *a) The solution of problem (3.4)–(3.5) is the wrapping function of problem (3.1)–(3.2)*

b) The interval enclosures $\mathcal{S}(h;t)$ produced by IPWA converge to the wrapping function of problem (3.1)–(3.2) when $h \to 0$ i.e. $\lim_{h\to 0} \mathcal{S}(h;t) = \widehat{X}(t)$.

Proof. Denote the solution of problem (3.4)–(3.5) by

$$X(t_0, X^0; t) = [\underline{x}(t_0, X^0; t), \overline{x}(t_0, X^0; t)]$$

First we will prove the following inclusions

$$\mathcal{S}(h;t) \subset \widehat{X}(t) \subset X(t_0, X^0; t)\ ,\quad t \in [t_0, \overline{t}]\ ,\ h > 0 \tag{3.6}$$

The first inclusion follows directly from the definition of wrapping function. We will use some monotone properties of the interval extension of f to show the second one. Every component $f_i(t,X) = [\underline{f}_i(t,X), \overline{f}_i(t,X)]$ of $f(t,X)$ satisfies

$$\underline{f}_i(t,X) = \underline{f}_i(t,[\underline{x}_1,\overline{x}_1],[\underline{x}_2,\overline{x}_2],\ldots,[\underline{x}_n,\overline{x}_n]) \text{ is } \begin{matrix}\text{non-decreasing about } \underline{x}_i \\ \text{non-increasing about } \overline{x}_i\end{matrix} \quad , \; i=1,\ldots,n$$

$$\overline{f}_i(t,X) = \overline{f}_i(t,[\underline{x}_1,\overline{x}_1],[\underline{x}_2,\overline{x}_2],\ldots,[\underline{x}_n,\overline{x}_n]) \text{ is } \begin{matrix}\text{non-increasing about } \underline{x}_i \\ \text{non-decreasing about } \overline{x}_i\end{matrix} \quad , \; i=1,\ldots,n$$

Therefore function $g=(g_1,g_2,\ldots,g_{2n})$ defined in the region

$$\{(t,y): t\in[t_0,\overline{t}], y\in D\times D, y_j+y_{n+j}\geq 0,\ j=1,\ldots,n\}$$

by

$$\begin{aligned} g_i(t,y_1,y_2,\ldots,y_{2n}) &= -\underline{f}_i(t,Y_1,\ldots,Y_{i-1},y_i,Y_{i+1},\ldots,Y_n)\ ,\ i=1,\ldots,n \\ g_{n+i}(t,y_1,y_2,\ldots,y_{2n}) &= \overline{f}_i(t,Y_1,\ldots,Y_{i-1},y_{n+i},Y_{i+1},\ldots,Y_n)\ ,\ i=1,\ldots,n \\ \text{where } Y_j &= [-y_j,y_{n+j}],\ j=1,\ldots,n \end{aligned}$$

is a quasi-isotone function.

Consider the equation

$$\dot{y}=g(t,y) \tag{3.7}$$

A well known property of equations with a quasi-isotone right-hand side is that if $y(t)$ and $z(t)$ are two solutions of (3.7) such that $y(\theta)\leq z(\theta)$ for some $\theta\in[t_0,\overline{t})$ then $y(t)\leq z(t)$, $t\in[\theta,\overline{t}]$ [81].

Let $\theta\in[t_0,\overline{t})$ and let $x(\theta,u;t)$ be any solution of equation (2.35) satisfying $x(\theta)=u\in X(t_0,X^0;\theta)$. It is easy to see that the two $2n$-dimensional functions $(-x(\theta,u;t),x(\theta,u;t))$ and $(-\underline{x}(t_0,X^0;t),\overline{x}(t_0,X^0;t))$ are solutions of equation (3.7). At the point θ we have

$$(-x(\theta,u;\theta),x(\theta,u;\theta))=(-u,u)\leq(-\underline{x}(t_0,X^0;\theta),\overline{x}(t_0,X^0;\theta))$$

Therefore $(-x(\theta,u;t),x(\theta,u;t))\leq(-\underline{x}(t_0,X^0;t),\overline{x}(t_0,X^0;t))$, $t\in[\theta,\overline{t}]$ which implies that $\underline{x}(t_0,X^0;t)\leq x(\theta,u;t)\leq\overline{x}(t_0,X^0;t)$, $t\in[t_0,\overline{t}]$. Hence

$$x(\theta,u;t)\in[\underline{x}(t_0,X^0;t),\overline{x}(t_0,X^0;t)]=X(t_0,X^0;t)\ ,\ t\in[t_0,\overline{t}]$$

Since the last inclusion is true for every $u\in X(t_0,X^0;\theta)$ and $\theta\in[t_0,\overline{t}]$ it follows that $X(t_0,X^0;t)$ satisfies the wrapping property at every point of the interval $[t_0,\overline{t})$. But the wrapping function is the optimal function that satisfies $X(t_0)=X^0$ and the wrapping property at every point of $[t_0,\overline{t})$. Therefore $\widehat{X}(t)\subset X(t_0,X^0;t)$, $t\in[t_0,\overline{t}]$. This concludes the proof of inclusion (3.6).

Now we will prove that $\lim_{h\to 0}\mathcal{S}(h;t)=X(t_0,X^0;t)$. This together with (3.6) implies both a) and b) in the theorem.

Let $[t_k,t_{k+1}]$ be an arbitrary subinterval and let $t\in[t_k,t_{k+1}]$. Then $\mathcal{S}(h;t)$ is defined by $\mathcal{S}(h;t)=[\underline{s}(h;t),\overline{s}(h;t)]=[x(t_k,\mathcal{S}(h;t_k);t)]$ where

$$\underline{s}_i(h;t)=\min_{u\in\mathcal{S}(h;t_k)} x_i(t_k,u;t)\ ,\ \overline{s}_i(h;t)=\max_{u\in\mathcal{S}(h;t_k)} x_i(t_k,u;t)\quad ,\ i=1,\ldots,n$$

Using

$$|x(t_k,u;t)-u| = \left|\int_{t_k}^{t} f(\theta,x(t_k,u;\theta)d\theta\right| \leq (t-t_k)m$$

it can be shown that

$$\left|S(h;t) -^{-} S(t;t_k)\right| \leq (t-t_k)m\ . \tag{3.8}$$

Every solution $x(t_k,u;t)$ can be represented in the form

$$\begin{aligned} x(t_k,u;t) &= u+\int_{t_k}^{t} f(\theta,x(t_k,u;\theta))d\theta \\ &= u+\int_{t_k}^{t} f(\theta,u)d\theta+\int_{t_k}^{t} f(x(t_k,u;\theta)-f(\theta,u))d\theta \\ &= \phi(u)+\varepsilon \end{aligned}$$

where $\phi(u) = u+\int_{t_k}^{t} f(\theta,u)d\theta$ and

$$\begin{aligned} |\varepsilon| &= \left|\int_{t_k}^{t} f(\theta,x(t_k,u;\theta)-f(\theta,u))d\theta\right| \leq \int_{t_k}^{t}\Lambda|x(t_k,u;\theta)-u|d\theta \\ &\leq \int_{t_k}^{t}\Lambda(\theta-t_k)md\theta = \frac{1}{2}(t-t_k)^2\Lambda m \leq \frac{1}{2}h^2\Lambda m\ . \end{aligned}$$

Therefore for every $i=1,2,\ldots,n$ we have

$$\phi_i(u)-\frac{1}{2}h^2\Lambda_{i*}m \leq x_i(t_k,u;t) \leq \phi_i(u)+\frac{1}{2}h^2\Lambda_{i*}m$$

where Λ_{i*} is the ith row of matrix Λ. Taking the maximum over $u\in\boldsymbol{\mathcal{S}}(h;\theta)$ of every part in the above inequality we obtain

$$\overline{\phi}_i(\boldsymbol{\mathcal{S}}(h;t_k))-\frac{1}{2}h^2\Lambda_{i*}m \leq \overline{\boldsymbol{s}}(h;t) \leq \overline{\phi}_i(\boldsymbol{\mathcal{S}}(h;t_k))+\frac{1}{2}h^2\Lambda_{i*}m$$

which can be also written in the form

$$|\overline{\boldsymbol{s}}_i(h;t)-\overline{\phi}_i(\boldsymbol{\mathcal{S}}(h;t_k))| \leq \frac{1}{2}h^2\Lambda_{i*}m\ . \tag{3.9}$$

Let's note that for sufficiently small h function $\phi=\phi(u)$ is such that ϕ_i is non-decreasing about u_i, $i=1,\ldots,n$. Hence, for the interval extension of ϕ at $S(h;t_k)$ we have

$$\overline{\phi}_i(\boldsymbol{\mathcal{S}}(t_k)) = \max_{u\in\boldsymbol{\mathcal{S}}(h;t_k)}\{u_i+\int_{t_k}^{t} f_i(\theta,u)d\theta\} = \overline{\boldsymbol{s}}_i(h;t_k)+\int_{t_k}^{t} f_i(\theta,\overline{\boldsymbol{\mathcal{S}}}^{i}(h;t_k))d\theta\ . \tag{3.10}$$

Therefore inequality (3.9) can be written in the form

$$\left|\overline{\boldsymbol{s}}_i(h;t)-\overline{\boldsymbol{s}}_i(h;t_k)-\int_{t_k}^{t} f_i(\theta,\overline{\boldsymbol{\mathcal{S}}}^{i}(h;t_k))d\theta)\right| \leq \frac{1}{2}h^2\Lambda_{i*}m\ . \tag{3.11}$$

Using (3.11), (3.3) and (3.8) we obtain

$$\begin{aligned}
&\left|\overline{s}_i(h;t) - \overline{s}_i(h;t_k) - \int_{t_k}^{t} \overline{f}_i(\theta, \overline{\mathcal{S}}^i(h;\theta))d\theta\right| \\
&\leq \left|\overline{s}_i(h;t) - \overline{s}_i(h;t_k) - \int_{t_k}^{t} f_i(\theta, \overline{\mathcal{S}}^i(h;t_k))d\theta)\right| + \left|\int_{t_k}^{t} \left(\overline{f}_i(\theta, \overline{\mathcal{S}}^i(h;t_k)) - \overline{f}_i(\overline{\mathcal{S}}^i(h;\theta))\right) d\theta\right| \\
&\leq \frac{1}{2}h^2\Lambda_{i*}m + \int_{t_k}^{t} \Lambda_{i*}\left|\overline{\mathcal{S}}^i(h;t_k)) - \overline{\mathcal{S}}^i(h;\theta)\right| d\theta \\
&\leq \frac{1}{2}h^2\Lambda_{i*}m + \frac{1}{2}h^2\Lambda_{i*}m = h^2\Lambda_{i*}m \ .
\end{aligned}$$

Let now $t \in [t_0, \overline{t}]$. There exists an interval $[t_r, t_{r+1}]$ such that $t \in [t_r, t_{r+1}]$. Applying the above inequality for the intervals $[t_0, t_1]$, $[t_1, t_2], \ldots, [t_r, t]$ we have

$$\begin{aligned}
&\left|\overline{s}_i(h;t) - \overline{s}_i(h;t_0) - \int_{t_0}^{t} \overline{f}_i(\overline{\mathcal{S}}^i(h;\theta))d\theta\right| \\
&\leq \sum_{k=0}^{r-1} \left|\overline{s}_i(h;t_{k+1}) - \overline{s}_i(h;t_k) - \int_{t_k}^{t_{k+1}} \overline{f}_i(\theta, \overline{\mathcal{S}}^i(h;\theta))d\theta\right| \\
&\quad + \left|\overline{s}_i(h;t) - \overline{s}_i(h;t_r) - \int_{t_r}^{t} \overline{f}_i(\theta, \overline{\mathcal{S}}^i(h;\theta))d\theta\right| \leq (r+1)\, h^2\Lambda_{i*}m
\end{aligned}$$

which yields

$$\left|\overline{s}_i(h;t) - \overline{s}_i(h;t_0) - \int_{t_0}^{t} \overline{f}_i(\theta, \overline{\mathcal{S}}^i(h;\theta))d\theta\right| \leq h(\overline{t} - t_0)\Lambda_{i*}m \ , \quad i = 1, \ldots, n \ . \tag{3.12}$$

In a similar way we obtain

$$\left|\underline{s}_i(h;t) - \underline{s}_i(h;t_0) - \int_{t_0}^{t} \underline{f}_i(\theta, \underline{\mathcal{S}}^i(h;\theta))d\theta\right| \leq h(\overline{t} - t_0)\Lambda_{i*}m \ , \quad i = 1, \ldots, n \ . \tag{3.13}$$

It is easy to see that the functions in each of the sets $\{\underline{s}(h;.)\}$ and $\{\overline{s}(h;.)\}$ are uniformly bounded and equicontinuous. Then the theorem of Arzelá-Ascoli implies that $\{\underline{s}(h;.)\}$ and $\{\overline{s}(h;.)\}$ considered as generalized sequences of h, $h \to 0$, have subsequences $\{\underline{s}(h_\alpha;.)\}$ and $\{\overline{s}(h_\alpha;.)\}$ that are uniformly convergent to continuous functions $\underline{s}$ and $\overline{s}$ respectively. Obviously $\underline{s} \leq \overline{s}$. Let $\mathcal{S} = [\underline{s}, \overline{s}]$. From (3.12) and (3.13) when $h = h_\alpha \to 0$ it follows that

$$\begin{aligned}
\underline{s}_i(t) &= \underline{s}_i(t_0) + \int_{t_0}^{t} \underline{f}_i(\theta, \underline{\mathcal{S}}^i(\theta))d\theta \ , \ i = 1, \ldots, n \\
\overline{s}_i(t) &= \overline{s}_i(t_0) + \int_{t_0}^{t} \overline{f}_i(\theta, \overline{\mathcal{S}}^i(\theta))d\theta \ , \ i = 1, \ldots, n
\end{aligned}$$

which implies that $\mathcal{S}$ is differentiable and

$$\begin{aligned}
&\mathcal{L}\mathcal{S}(t) = 0 \ , \quad t \in [t_0, \overline{t}] \\
&\mathcal{S}(t_0) = X^0 \ .
\end{aligned}$$

Therefore $\boldsymbol{S}(t) = X(t_0, X^0; t)$, $t \in [t_0, \overline{t}]$.

Since this is true for any other convergent subsequences of $\{\underline{s}(h;.)\}$ and $\{\overline{s}(h;.)\}$ then $\underline{x}(t_0, X^0;.)$ is the only accumulation point of $\{\underline{s}(h;.)\}$ and $\overline{x}(t_0, X^0;.)$ is the only accumulation point of $\{\overline{s}(h;.)\}$. Therefore

$$\lim_{h\to 0} \boldsymbol{S}(h;t) = X(t_0, X^0; t)\ . \tag{3.14}$$

This concludes the proof because both statements of the theorem follow from (3.6) and (3.14).

Theorem 3.2 *Let a numerical method produce interval enclosures $S(h;t)$ of the solution of problem (3.1)-(3.2) such that $S(h;t)$ satisfy the wrapping property at the points of the mesh $\{t_0, t_1, \ldots, t_n\}$ and the local error is*

$$\left| S(h;t) -^{-} [x(t_k, S(h;t_k); t)] \right| = o(h)\ ,\ t \in [t_k, t_{k+1}]\ ,\ k = 0, 1, \ldots, n-1$$

then

$$\lim_{h\to 0} S(h;t) = \widehat{X}(t)\ ,\ t \in [t_0, \overline{t}]\ .$$

Proof. Using standard techniques one can show that the limit of $S(h;t)$ is the same as the limit of $\boldsymbol{S}(h,t)$ and then the statement follows from Theorem 3.1

Theorem 3.2 shows that, in general, the interval enclosures produced by a method of the considered type do not converge to the optimal interval enclosure $[x(t_0, X^0; t)]$ of the solution but to the wrapping function $\widehat{X}(t)$. Convergence to $[x(t_0, X^0; t)]$ is obtained if and only if $[x(t_0, X^0; t)] = \widehat{X}(t)$. More precise analysis can reveal that when $[x(t_0, X^0; t)] \neq \widehat{X}(t)$ the rate of convergence is $O(h)$ irrespective of the rate of the local approximation while if $[x(t_0, X^0; t)] = \widehat{X}(t)$ the rate of convergence corresponds to the rate of local approximation.

3.2 Quantifying the Wrapping Effect.

Using the concept of wrapping function we can quantify the wrapping effect associated with problem (3.1)–(3.2) in the following way. Let $S(h;t)$ be interval enclosures of the solution of (3.1)–(3.2) produced by a method of a propagate and wrap type. The limit of the error of approximation when $h \to 0$ is

$$\lim_{h\to 0} \wp(S(h;t), [x(t_0, X^0; t)]) = \wp(\widehat{X}(t), [x(t_0, X^0; t)]).$$

The quantity

$$\wp(\widehat{X}(t), [x(t_0, X^0; t)]) = \left\| |\widehat{X}(t) -^{-} [x(t_0, X^0; t)])| \right\| \tag{3.15}$$

does not depend on the method and characterizes problem (3.1)–(3.2) with respect to the occurrence of a wrapping effect and its magnitude. In this way it is a measure of the wrapping effect associated with problem (3.1)–(3.2). The vector function

$$|\widehat{X}(t) -^{-} [x(t_0, X^0; t)])| \tag{3.16}$$

provides more detailed information about the wrapping effect because its coordinates give the magnitude of the wrapping effect in the corresponding coordinate directions

$$|\widehat{X}_i(t) -^{-} [x_i(t_0, X^0; t)])| = \wp(\widehat{X}^i(t), [x_i(t_0, X^0; t)]) \ , \ i = 1, \ldots, n \ .$$

Since $[x(t_0, X^0; t)] \subset \widehat{X}(t)$ we have

$$\begin{aligned} \frac{1}{2}\left(w(\widehat{X}(t)) - w([x(t_0, X^0; t)])\right) &\leq |\widehat{X}_i(t) -^{-} [x_i(t_0, X^0; t)])| \\ &\leq w(\widehat{X}(t)) - w([x(t_0, X^0; t)]) \end{aligned}$$

Therefore the vector function

$$w(\widehat{X}(t)) - w([x(t_0, X^0; t)]) \ . \tag{3.17}$$

provides a measure for the wrapping effect equivalent to (3.16).

Each one of the functions (3.15), (3.16) and (3.17) may be used in characterizing the wrapping effect associated with a particular problem.

If $\widehat{X}(t) = [x(t_0, X^0; t)]$ for a problem of the form (3.1)–(3.2) we say that this problem has no wrapping effect because the enclosures produced by any method of propagate and wrap type converge to the optimal interval enclosure with a rate corresponding to the rate of local approximation provided by the method. A problem with no wrapping effect is characterized by any of the functions (3.15), (3.16) or (3.17) being zero.

Revisiting example 2.1

The exact solution of the system of linear equations (2.38) is

$$x(0, x^0; t) = \begin{pmatrix} e^{-2t} & 0 & 0 \\ 2(e^{-t} - e^{-2t}) & \cosh t & -\sinh t \\ 2(e^{-t} - e^{-2t}) & -\sinh t & \cosh t \end{pmatrix} x^0 \ .$$

Therefore for the optimal interval enclosure $[x(0, X^0; t)] = [\underline{x}(0, X^0; t), \overline{x}(0, X^0; t)]$ we have

$$\underline{x}(t_0, X^0; t)| = \begin{pmatrix} \underline{x}_1^0 e^{-2t} \\ 2\underline{x}_1^0(e^{-t} - e^{-2t}) + \underline{x}_2^0 \cosh t - \overline{x}_3^0 \sinh t \\ 2\underline{x}_1^0(e^{-t} - e^{-2t}) - \overline{x}_2^0 \sinh t + \underline{x}_3^0 \cosh t \end{pmatrix} \tag{3.18}$$

and

$$\overline{x}(t_0, X^0; t)| = \begin{pmatrix} \overline{x}_1^0 e^{-2t} \\ 2\overline{x}_1^0(e^{-t} - e^{-2t}) + \overline{x}_2^0 \cosh t - \underline{x}_3^0 \sinh t \\ 2\overline{x}_1^0(e^{-t} - e^{-2t}) - \underline{x}_2^0 \sinh t + \overline{x}_3^0 \cosh t \end{pmatrix} . \tag{3.19}$$

The right-hand side of the equation in example 2.1 has the following interval extension

$$f(t, [\underline{x}, \overline{x}]) = \begin{pmatrix} [-2\overline{x}_1 \ , \ -2\underline{x}_1] \\ [2\underline{x}_1 - \overline{x_3} \ , \ 2\overline{x}_1 - \underline{x_3}] \\ [2\underline{x}_1 - \overline{x_2} \ , \ 2\overline{x}_1 - \underline{x_2}] \end{pmatrix}$$

Therefore problem (3.4)–(3.5) can be written in the form

$$\begin{aligned} &\dot{\underline{x}}_1 = -\underline{x}_1 && \dot{\overline{x}}_1 = -2\overline{x}_1 \\ &\dot{\underline{x}}_2 = 2\underline{x}_1 - \overline{x}_3 && \dot{\overline{x}}_2 = 2\overline{x}_1 - \underline{x}_3 \\ &\dot{\underline{x}}_3 = 2\underline{x}_1 - \overline{x}_2 && \dot{\overline{x}}_3 = 2\overline{x}_1 - \underline{x}_2 \\ &\underline{x}_i(0) = 1 - \varepsilon_i \ , \ i = 1,2,3 && \overline{x}_i(0) = 1 + \varepsilon_i \ , \ i = 1,2,3 \end{aligned}$$

The above problem can be solved using standard techniques and its solution gives the wrapping function $\widehat{X} = [\widehat{\underline{x}}, \widehat{\overline{x}}]$. We have

$$\widehat{\underline{x}}(t) = \begin{pmatrix} \underline{x}_1^0 e^{-2t} \\ \frac{1}{3}\underline{x}_1^0(e^t + 3e^{-t} - 4e^{-2t}) - \frac{1}{3}\overline{x}_1^0(e^t - 3e^{-t} + 2e^{-2t}) + \underline{x}_2^0 \cosh t - \overline{x}_3^0 \sinh t \\ \frac{1}{3}\underline{x}_1^0(e^t + 3e^{-t} - 4e^{-2t}) - \frac{1}{3}\overline{x}_1^0(e^t - 3e^{-t} + 2e^{-2t}) - \overline{x}_2^0 \sinh t + \underline{x}_3^0 \cosh t \end{pmatrix} \tag{3.20}$$

and

$$\widehat{\overline{x}}(t) = \begin{pmatrix} \overline{x}_1^0 e^{-2t} \\ \frac{1}{3}\overline{x}_1^0(e^t + 3e^{-t} - 4e^{-2t}) - \frac{1}{3}\underline{x}_1^0(e^t - 3e^{-t} + 2e^{-2t}) + \overline{x}_2^0 \cosh t - \underline{x}_3^0 \sinh t \\ \frac{1}{3}\overline{x}_1^0(e^t + 3e^{-t} - 4e^{-2t}) - \frac{1}{3}\underline{x}_1^0(e^t - 3e^{-t} + 2e^{-2t}) - \underline{x}_2^0 \sinh t + \overline{x}_3^0 \cosh t \end{pmatrix} . \tag{3.21}$$

Using (3.18), (3.19), (3.20) and (3.21) we can obtain the measure of the wrapping effect (3.16)

$$\begin{aligned} \left| \widehat{X}(t) -^- [x(0, X^0; t)] \right| &= \begin{pmatrix} 0 \\ \frac{1}{3}(e^t - 3e^{-t} + 2e^{-2t})(\overline{x}_1^0 - \underline{x}_1^0) \\ \frac{1}{3}(e^t - 3e^{-t} + 2e^{-2t})(\overline{x}_1^0 - \underline{x}_1^0) \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ \frac{2}{3}(e^t - 3e^{-t} + 2e^{-2t})\varepsilon_1 \\ \frac{2}{3}(e^t - 3e^{-t} + 2e^{-2t})\varepsilon_1 \end{pmatrix} . \end{aligned}$$

From the above form of the wrapping effect measure we can make the following observations:

1. There is no wrapping effect in x_1 (see figure 2.1). This is not surprising because x_1 is obtained only from the first equation and the right hand side of a single equation is always quasi-isotone.

2. The wrapping effect in x_2 and x_3 depends only on the width of X_1^0. Therefore there is no wrapping effect if $w(X_1^0) = 0$ (see figure 2.2).

On figure 3.1, where the computed enclosures for problem (2.38)–(2.39) are plotted together with the wrapping function, convergence of these enclosures to the wrapping function can be observed.

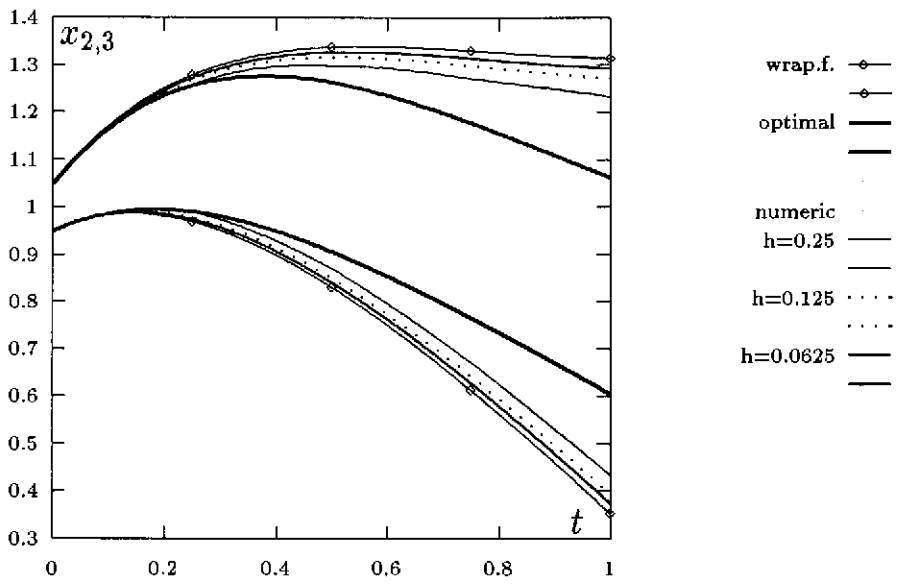


Figure 3.1: *Problem (2.38) with* $\varepsilon_1 = 0.2$, $\varepsilon_2 = \varepsilon_3 = 0.05$. *Wrapping function, optimal enclosure and enclosures computed numerically for various step sizes* h.

Moore's example.

The following example was considered by Moore [67] and discussed in many publications on the wrapping effect.

$$\begin{aligned} x_1 &= x_2 \quad , \quad x_1(0) = x_1^0 \in X_1^0 = [-\delta, \delta] \\ x_2 &= -x_1 \quad , \quad x_2(0) = x_2^0 \in X_2^0 = [1-\delta, 1+\delta] \end{aligned} \tag{3.22}$$

Moore showed that at $t = 2\pi$ the computed interval enclosures are inflated by a factor of approximately $e^{2\pi} \approx 535$. We will obtain this result using the wrapping function of problem (3.22).

The exact solution of this problem is

$$x(0, x^0; t)] = \begin{pmatrix} \cos t & \sin t \\ -\sin t & cost \end{pmatrix} x^0 .$$

Hence the optimal interval enclosure can be represented in the form

$$[x(0;x^0;t)] = \begin{pmatrix} (\cos t)X_1^0 + (\sin t)X_2^0 \\ -(\sin t)X_1^0 + (\cos t)X_2^0 \end{pmatrix}$$

and its width is

$$\begin{aligned} w([x(0;x^0;t)]) &= \begin{pmatrix} |\cos t|w(X_1^0) + |\sin t|w(X_2^0) \\ |\sin t|w(X_1^0) + |\cos t|w(X_2^0) \end{pmatrix} \\ &= (|\cos t| + |\sin t|)\begin{pmatrix} 2\delta \\ 2\delta \end{pmatrix}. \end{aligned} \tag{3.23}$$

Problem (3.4)–(3.5) for the system (3.22) can be written in the following form

$$\begin{aligned} \dot{\underline{x}}_1 &= \underline{x}_2\,, & \underline{x}_1(0) &= -\delta \\ \dot{\overline{x}}_1 &= \overline{x}_2\,, & \overline{x}_1(0) &= \delta \\ \dot{\underline{x}}_2 &= -\overline{x}_1\,, & \underline{x}_2(0) &= 1-\delta \\ \dot{\overline{x}}_2 &= -\underline{x}_1\,, & \overline{x}_2(0) &= 1+\delta \end{aligned}$$

Solving this problem we obtain the wrapping function $\widehat{X}(t) = [\underline{\widehat{x}}(t), \overline{\widehat{x}}(t)]$ of Moore's example:

$$\underline{\widehat{x}}(t) = \frac{1}{2}\begin{pmatrix} (\cos t+\cosh t)\underline{x}_1^0 + (\cos t-\cosh t)\overline{x}_1^0 + (\sin t+\sinh t)\underline{x}_2^0 + (\sin t-\sinh t)\overline{x}_2^0 \\ (-\sin t+\sinh t)\underline{x}_1^0 - (\sin t+\sinh t)\overline{x}_1^0 + (\cos t+\cosh t)\underline{x}_2^0 + (\cos t-\cosh t)\overline{x}_2^0 \end{pmatrix}$$

and

$$\overline{\widehat{x}}(t) = \frac{1}{2}\begin{pmatrix} (\cos t-\cosh t)\underline{x}_1^0 + (\cos t+\cosh t)\overline{x}_1^0 + (\sin t-\sinh t)\underline{x}_2^0 + (\sin t+\sinh t)\overline{x}_2^0 \\ -(\sin t+\sinh t)\underline{x}_1^0 + (-\sin t+\sinh t)\overline{x}_1^0 + (\cos t-\cosh t)\underline{x}_2^0 + (\cos t+\cosh t)\overline{x}_2^0 \end{pmatrix}.$$

Therefore

$$\begin{aligned} w(\widehat{X}(t)) &= \overline{\widehat{x}}(t) - \underline{\widehat{x}}(t) \\ &= \begin{pmatrix} \cosh t(\overline{x}_1^0 - \underline{x}_1^0) + \sinh t(\overline{x}_2^0 - \underline{x}_2^0) \\ \sinh t(\overline{x}_1^0 - \underline{x}_1^0) + \cosh t(\overline{x}_2^0 - \underline{x}_2^0) \end{pmatrix} \\ &= e^t\begin{pmatrix} 2\delta \\ 2\delta \end{pmatrix}. \end{aligned} \tag{3.24}$$

From (3.23) and (3.24) we have

$$w(\widehat{X}(t)) = \frac{e^t}{|\cos t| + |\sin t|} w([x(t_0,x^0;t)]).$$

Since $\widehat{X}$ is the limit of the interval enclosures when $h \to 0$ then these enclosures are inflated at the point t by a factor of approximately $\dfrac{e^t}{|\cos t| + |\sin t|}$ when h is small enough. At $t = 2\pi$ the value of this factor is $e^{2\pi}$.

3.3 Problems Without Wrapping Effect

It is clear from the previous sections that methods of propagate and wrap type can be applied successfully only to problems where no wrapping effect occurs. In this section we will use the concept of wrapping function to characterize such problems. Our approach is to find problems of the form (3.1)-(3.2) such that the wrapping function $\widehat{X}(t)$ equals the optimal interval enclosure $[x(t_0, X^0; t)]$ or equivalently $w(\widehat{X}(t)) = w([x(t_0, X^0; t)])$, $t \in [t_0, \overline{t}]$.

Theorem 3.3 *If a diagonal matrix $Q = diag(q_1, q_2, \ldots, q_n)$, $q_i \in \{-1, 1\}$, $i = 1, \ldots, n$ exists, such that the function $Qf(t, Qx)$ is a quasi-isotone function of $x \in QD = \{Qd : d \in D\}$ then the wrapping function $\widehat{X}$ of problem (3.1)-(3.2) equals the optimal interval enclosure $[x(t_0, X^0; .)]$; i.e. there is no wrapping effect.*

Proof. Let us note that the linear transformation $\mathcal{Q} : \mathcal{R}^n \to \mathcal{R}^n$ defined by $\mathcal{Q}(x) = Qx$ preserves the intervals i.e. if $X \in I\mathcal{R}^n$ then $QX \in I\mathcal{R}^n$ or in general if $X \subset \mathcal{R}^n$ then

$$[QX] = Q[X] \tag{3.25}$$

We consider

$$\dot{y} = g(t, y) \tag{3.26}$$
$$y(t_0) = y_0 \in Y^0 \tag{3.27}$$

where $g(y) = Qf(t, Qy)$, $y \in QD$ is a quasi-isotone function of $y \in QD$ and $Y^0 = [\underline{y}^0, \overline{y}^0] = QX^0$. For the wrapping function $\widehat{Y} = [\underline{\widehat{y}}, \overline{\widehat{y}}]$ of this problem we have

$$\begin{aligned}
&\dot{\underline{\widehat{y}}}_i = \underline{g}_i(t, [\underline{\widehat{y}}_1, \overline{\widehat{y}}_1], \ldots, [\underline{\widehat{y}}_{i-1}, \overline{\widehat{y}}_{i-1}], \underline{\widehat{y}}_i, [\underline{\widehat{y}}_{i+1}, \overline{\widehat{y}}_{i+1}], \ldots, [\underline{\widehat{y}}_n, \overline{\widehat{y}}_n]) \ , \ i = 1, \ldots, n \\
&\dot{\overline{\widehat{y}}}_i = \overline{g}_i(t, [\underline{\widehat{y}}_1, \overline{\widehat{y}}_1], \ldots, [\underline{\widehat{y}}_{i-1}, \overline{\widehat{y}}_{i-1}], \overline{\widehat{y}}_i, [\overline{\widehat{y}}_{i+1}, \overline{\widehat{y}}_{i+1}], \ldots, [\overline{\widehat{y}}_n, \overline{\widehat{y}}_n]) \ , \ i = 1, \ldots, n \\
&\underline{\widehat{y}}(t_0) = \underline{y}^0 \\
&\overline{\widehat{y}}(t_0) = \overline{y}^0
\end{aligned} \tag{3.28}$$

Since g_i is non-decreasing about y_j, $j \neq i$ we have

$$\begin{aligned}
&\underline{g}_i(t, [\underline{\widehat{y}}_1, \overline{\widehat{y}}_1], \ldots, [\underline{\widehat{y}}_{i-1}, \overline{\widehat{y}}_{i-1}], \underline{\widehat{y}}_i, [\underline{\widehat{y}}_{i+1}, \overline{\widehat{y}}_{i+1}] = g_i(t, \underline{\widehat{y}}_1, \ldots, \underline{\widehat{y}}_{i-1}, \underline{\widehat{y}}_i, \underline{\widehat{y}}_{i+1}, \ldots, \underline{\widehat{y}}_n) \ , \\
&\overline{g}_i(t, [\underline{\widehat{y}}_1, \overline{\widehat{y}}_1], \ldots, [\underline{\widehat{y}}_{i-1}, \overline{\widehat{y}}_{i-1}], \overline{\widehat{y}}_i, [\underline{\widehat{y}}_{i+1}, \overline{\widehat{y}}_{i+1}] = g_i(t, \overline{\widehat{y}}_1, \ldots, \overline{\widehat{y}}_{i-1}, \overline{\widehat{y}}_i, \overline{\widehat{y}}_{i+1}, \ldots, \overline{\widehat{y}}_n) \ , \\
&i = 1, \ldots, n \ .
\end{aligned}$$

Then from (3.28) it follows that $\underline{\widehat{y}}(t) = \underline{y}(t_0, \underline{y}^0; t)$ and $\overline{\widehat{y}}(t) = \overline{y}(t_0, \overline{y}^0; t)$ belong to the solution $y(t_0, Y^0; t)$ of problem (3.26)-(3.27) which implies that

$$\widehat{Y}(t) = [y(t_0, Y^0; t] \ , \ t \in [t_0, \overline{t}] \ . \tag{3.29}$$

For every solution $x(t_0, x^0; t)$ of equation (3.1) we have

$$x(t_0, x^0; t) = Qy(t_0, Qx^0; t) .$$

Therefore from (3.25) and (3.29) it follows that

$$[x(t_0, X^0; t)] = [Qy(t_0, QX^0; t)] = Q[y(t_0, Y^0; t)] = Q\widehat{Y}(t) . \tag{3.30}$$

It remains to prove that $Q\widehat{Y}$ is the wrapping function of Problem (3.1)–(3.2).

At $t = t_0$ we have $Q\widehat{Y}(t_0) = Q^2X^0 = X^0$. Let $\theta \in [t_0, \bar{t})$ and let $u \in Q\widehat{Y}(\theta)$. Then $Qu \in \widehat{Y}(\theta)$ and

$$x(\theta, u; t) = Qy(\theta, Qu; t) \in Q\widehat{Y}(t) , \ t \in [\theta, \bar{t}]$$

i.e. $Q\widehat{Y}$ satisfies the wrapping property every $\theta \in [t_0, \bar{t})$. Therefore

$$[x(t_0, X^0; t] \subset \widehat{X}(t) \subset Q\widehat{Y}(t) , \ t \in [t_0, \bar{t}] .$$

Then (3.30) implies

$$[x(t_0, X^0; t] = \widehat{X}(t) = Q\widehat{Y}(t) , \ t \in [t_0, \bar{t}]$$

which concludes the proof of the theorem.

Considering again example 2.1 and Moore's example we can see that theorem 3.3 is not applicable to either of them because a matrix Q with the required properties does not exist. Theorem 3.3 is applicable to following example.

Example 3.1 Consider the problem

$$\begin{aligned} \dot{x}_1 &= -2x_1 , & x_1(0) &= x_1^0 \in X_1^0 = 1 + [-\varepsilon_1, \varepsilon_1] , \\ \dot{x}_2 &= 2x_1 - x_3 , & x_2(0) &= x_2^0 \in X_2^0 = 1 + [-\varepsilon_2, \varepsilon_2] , \\ \dot{x}_3 &= -2x_1 - x_2 , & x_3(0) &= x_3^0 \in X_3^0 = 1 + [-\varepsilon_3, \varepsilon_3] . \end{aligned} \tag{3.31}$$

in the interval $[0, 1]$. Function

$$f(x) = \begin{pmatrix} -2x_1 \\ 2x_1 - x_3 \\ -2x_1 - x_2 \end{pmatrix}$$

can be transformed into a quasi-isotone function using a matrix

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} .$$

Indeed,

$$Qf(t,Qx) = \begin{pmatrix} -2x_1 \\ 2x_1 + x_3 \\ 2x_1 + x_2 \end{pmatrix}$$

is quasi-isotone. Then from theorem (3.3) it follows that problem (3.31) is a problem with no wrapping effect for any initial condition $X^0 \in \mathcal{IR}^n$. This theoretical result is supported by results of numerical experiments. We consider the values of ε_i, $i = 1,2,3$ given by (2.39) and apply the same method as in example 2.1. While in the case of example 2.1 we obtain enclosures which diverge from the optimal enclosure (figure 2.1, right) the enclosures produced by the method in the case of example 3.1, converge to the optimal one. This is demonstrated graphically on figure 3.2. Since the optimal and the numerically computed enclosures for x_1 are the same as in example 2.1 the corresponding graphs are omitted. The graphs for x_2 and x_3 are only presented. The graphs of the computed enclosures are visually indistinguishable from the optimal enclosure. At the bottom part of the figure the error functions

$$\wp\left(S_i(h;t), [x_i(t_0, X^0; t)]\right) \ , \quad i = 1,2,3$$

are plotted on a logarithmic scale. Convergence at a rate consistent with the expected rate of global convergence can be observed.

Let us note that theorem 3.3 provides a sufficient condition for problems with no wrapping effect. An interesting question to consider is whether, and in what form, this condition is also a necessary condition for having no wrapping effect.

3.4 Linear Systems of ODE.

When f is a linear function of x problem (3.1)–(3.2) can be written in the form

$$\dot{x} = A(t)x + b(t) \tag{3.32}$$

$$x(t_0) = x^0 \in X^0 \tag{3.33}$$

where $A(t) = \begin{pmatrix} a_{11}(t) & a_{12}(t) & \dots & a_{1n}(t) \\ a_{21}(t) & a_{22}(t) & \dots & a_{2n}(t) \\ \dots & \dots & \dots & \dots \\ a_{n1}(t) & a_{n2}(t) & \dots & a_{nn}(t) \end{pmatrix}$ and $b(t) = \begin{pmatrix} b_1(t) \\ b_2(t) \\ \dots \\ b_n(t) \end{pmatrix}$.

We assume that A and b are continuous functions of $t \in [t_0, \overline{t}]$. Every solution $x(t_0, x^0; t)$ of equation (3.32) can be represented in the form

$$x(t_0, x^0; t) = M(A;t)\left(x^0 + \int_{t_0}^{t} M(A;\theta)^{-1} b(\theta) d\theta\right)$$

where the $n \times n$ matrix function $M(A;t)$ is the matricant of A defined by

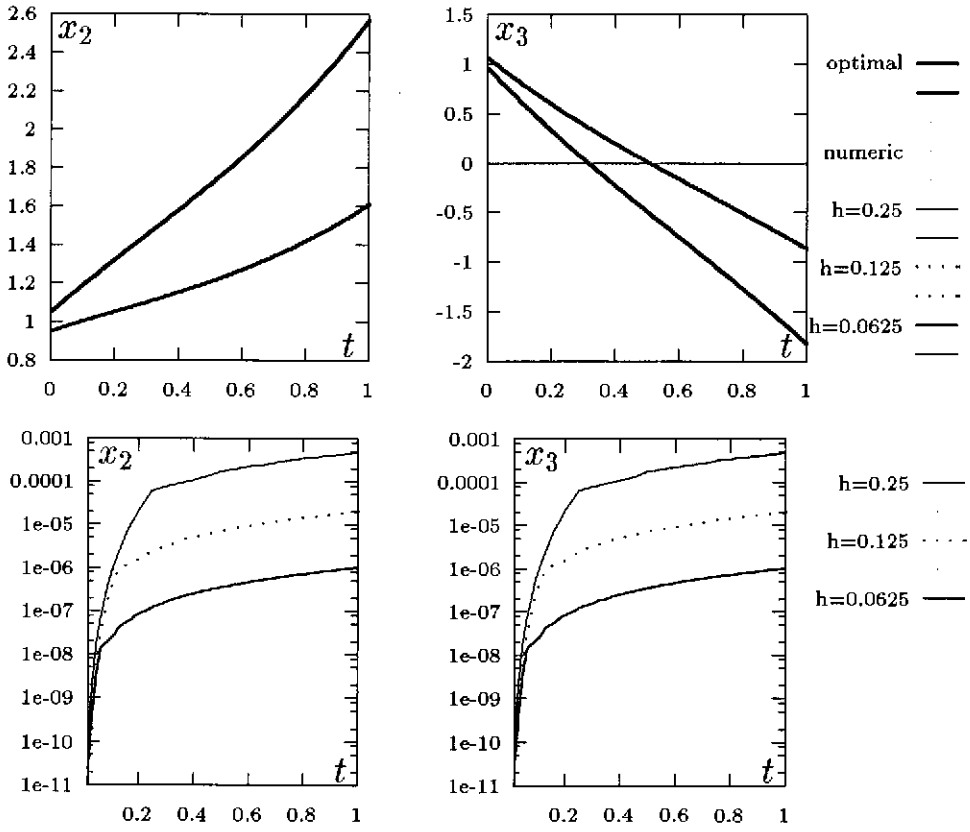


Figure 3.2: *Problem (3.31) with* $\varepsilon_1 = 0.2$, $\varepsilon_2 = \varepsilon_3 = 0.05$. *Optimal enclosure and enclosures computed numerically for various step sizes h (top) and errors of the computed enclosures on a logarithmic scale (bottom).*

$$M(A;t) = \sum_{k=0}^{\infty} M^{(k)}(A;t)$$

where

$$M^{(0)}(A;t) = I \text{ (identity matrix of order } n)$$
$$M^{(k+1)}(A;t) = \int_{t_0}^{t} A(\theta) M^{(k)}(A;\theta)\, d\theta\ , \quad k = 0, 1, \ldots\ .$$

Using interval arithmetic the optimal interval enclosure can be represented as

$$[x(t_0, X^0; t)] = M(A;t)\left(X^0 + \int_{t_0}^{t} M(A;\theta)^{-1} b(\theta)\, d\theta\right)\ .$$

For the width of $[x(t_0, X^0; t)]$ we have

$$w([x(t_0, X^0; t)]) \;=\; w\left(M(A;t)\left(X^0 + \int_{t_0}^{t} M(A;\theta)^{-1} b(\theta)\, d\theta\right)\right)$$

$$
\begin{aligned}
&= w\left(M(A;t)X^0\right) \\
&= |M(A;t)|\, w(X^0) \, .
\end{aligned}
$$

Let $A^+(t)$ denote the matrix $A^+(t) = \begin{pmatrix} a_{11}(t) & |a_{12}(t)| & \dots & |a_{1n}(t)| \\ |a_{21}(t)| & a_{22}(t) & \dots & |a_{2n}(t)| \\ \dots & \dots & \dots & \dots \\ |a_{n1}(t)| & |a_{n2}(t)| & \dots & a_{nn}(t) \end{pmatrix}$.

For the width of the wrapping function we have

$$
\begin{aligned}
\frac{d}{dt} w(\widehat{X}_i(t)) &= \dot{\overline{\widehat{x}}}_i(t) - \dot{\underline{\widehat{x}}}_i(t) \\
&= \max_{x \in \widehat{X}} A_{i*}x - \min_{x \in \widehat{X}} A_{i*}x \\
&= \sum_{j \neq i} |a_{ij}| w(\widehat{X}_j) + a_{ii} w(\widehat{X}_i) \, ,
\end{aligned}
$$

where A_{i*} denotes the ith row of matrix A. Therefore $w(\widehat{X}(t))$ is the solution of

$$
\begin{aligned}
&\dot{y} = A^+(t)y \\
&y(t_0) = w(X^0)
\end{aligned}
$$

and can be represented as

$$
w(\widehat{X}(t)) = M(A^+;t) w(X^0) \, . \tag{3.34}
$$

The wrapping effect measure (3.17) for the linear problem (3.32)–(3.33) is

$$
w(\widehat{X}(t)) - w([x(t_0, X^0; t)]) = \left(M(A^+;t) - |M(A;t)|\right) w(X^0) \, . \tag{3.35}
$$

Using the fact that the solution set at every $t \in [t_0, \overline{t}]$ is an affine transformation of X^0 one can easily see that

$$
|\widehat{X}(t) \mathbin{-^{-}} [x(t_0, X^0; t)]| = \frac{1}{2} \left(w(\widehat{X}(t)) - w([x(t_0, X^0; t)]) \right) \, .
$$

Therefore functions (3.16) and (3.15) can be represented in the form

$$
\begin{aligned}
|\widehat{X}(t) \mathbin{-^{-}} [x(t_0, X^0; t)]| &= \frac{1}{2} \left(M(A^+;t) - |M(A;t)|\right) w(X^0) \\
\wp(\widehat{X}(t), [x(t_0, X^0; t)]) &= \frac{1}{2} \left\| \left(M(A^+;t) - |M(A;t)|\right) w(X^0) \right\| \, .
\end{aligned}
$$

We can see that the linear problem (3.32)–(3.33) has no wrapping effect if and only if

$$
\left(M(A^+;t) - |M(A;t)|\right) w(X^0) = 0 \, , \; t \in [t_0, \overline{t}] \tag{3.36}
$$

Function $f(t,x) = A(t)x + b(t)$ is quasi-isotone if and only if the nondiagonal entries of matrix $A(t)$ are nonnegative for every $t \in [t_0, \overline{t}]$, i.e. $A(t) = A^+(t)$, $t \in [t_0, \overline{t}]$. Therefore, when f is quasi-isotone, condition (3.36) is satisfied and the problem has no wrapping effect for any initial condition.

Suppose now that the condition of theorem 3.3 is satisfied, i.e. there exists a matrix $Q = diag(q_1, \ldots, q_n)$, $q_i \in \{-1, 1\}, i = 1, \ldots, n$ such that $Qf(t, Qx) = QA(t)Qx + Qb(t)$ is quasi-isotone. This means that

$$QA(t)Q = A^+(t). \tag{3.37}$$

Since function (3.35) is nonnegative for any initial condition, the matricants of A and A^+ satisfy the inequality

$$M(A^+; t) \geq |M(A; t)|\ , \quad t \in [t_0, \overline{t}]\ . \tag{3.38}$$

It is easy to prove by induction that

$$M^{(k)}(QAQ; t) = QM^{(k)}(A; t)Q\ . \tag{3.39}$$

Indeed, if (3.39) is true for some k then for $k+1$ we have

$$\begin{aligned} M^{(k+1)}(QAQ; t) &= \int_{t_0}^{t} QA(\theta)QM^{(k)}(QAQ; \theta)d\theta \\ &= \int_{t_0}^{t} QA(\theta)QQM^{(k)}(A; \theta)Qd\theta \\ &= Q\int_{t_0}^{t} A(\theta)M^{(k)}(A; \theta)d\theta Q \\ &= QM^{(k+1)}(A; t)Q\ . \end{aligned}$$

From (3.39) it follows that

$$M(QAQ; t) = QM(A; t)Q\ .$$

Therefore if (3.37) is true we have

$$M(A^+; t) = M(QAQ, t) = QM(A; t)Q \leq |QM(A; t)Q| = |M(A; t)|\ . \tag{3.40}$$

From (3.38) and (3.40) it follows that

$$M(A^+; t) = |M(A; t)|$$

which implies that there is no wrapping effect.

For linear systems of the form (3.32)–(3.33) Theorem 3.3 can be formulated as follows

Theorem 3.4 *If there exists a diagonal matrix* $Q = diag(q_1, q_2, \ldots, q_n)$, $q_i \in \{-1, 1\}$, $i = 1, \ldots, n$ *such that* $QA(t)Q = A^+(t)$, $t \in [t_0, \overline{t}]$ *then problem (3.32)-(3.33) has no wrapping effect.*

Theorem 3.4 essentially coincides with a result in [45] obtained in a different way. The following examples show that the requirements of this theorem are essential.

Example 3.2 Consider the problem

$$\begin{aligned} \dot{x}_1 &= 2(t-1)x_2 \quad , \qquad x_1(0) = x_1^0 \in [0.9, 1.1] \\ \dot{x}_2 &= -2|t-1|x_1 \quad , \qquad x_2(0) = x_2^0 \in [-0.1, 0.1] \end{aligned}$$

for $t \geq 0$. This is a linear problem of the form (3.32)–(3.33) with

$$A(t) = 2\begin{pmatrix} 0 & t-1 \\ -|t-1| & 0 \end{pmatrix} \quad \text{and} \quad b(t) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

In the interval $[0, 1]$ using a matrix

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

we have

$$\begin{aligned} QA(t)Q &= 2\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 0 & t-1 \\ t-1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \\ &= 2\begin{pmatrix} 0 & 1-t \\ 1-t & 0 \end{pmatrix} = 2\begin{pmatrix} 0 & |t-1| \\ |t-1| & 0 \end{pmatrix} = A^+(t)\,. \end{aligned}$$

Therefore there is no wrapping effect.

For $t > 1$ matrix A is

$$A(t) = 2\begin{pmatrix} 0 & t-1 \\ 1-t & 0 \end{pmatrix}$$

and it is easy to see that a matrix Q such that $QA(t)Q = A^+(t)$ does not exist.

We apply a method of propagate and wrap type with local error $O(h^5)$ using a uniform mesh with a step size h to the above problem. The enclosures computed for various values of h as well as the optimal enclosure and the wrapping function are presented in figure 3.3.

In the interval $[0, 1]$ the computed enclosures are visually indistinguishable from $[x(t_0, X^0; t)] = \widehat{X}(t)$ and when the error $\wp(S(h; t), [x(t_0, X^0; t)])$ is plotted on a logarithmic scale (figure 3.4, left) a fast convergence, consistent with the expected $O(h^4)$ is revealed. In the interval $[1, 2.5]$ where the wrapping function is wider then the optimal enclosure the convergence is towards the wrapping function at a rate of $O(h)$ (see figure 3.4, right). On figure 3.4 we can also observe that the error of the interval enclosures approaches the wrapping effect measure (3.15).

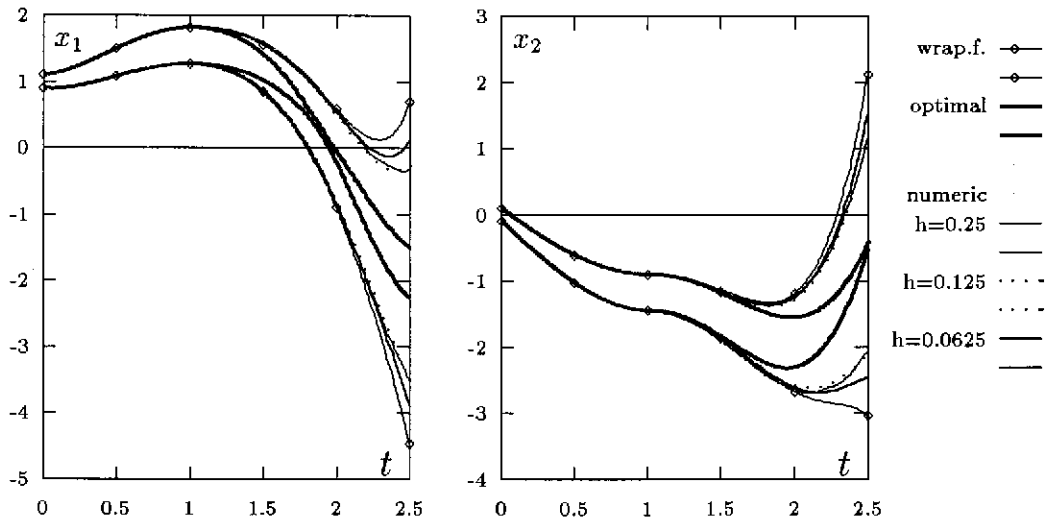


Figure 3.3: *Example 3.2. Wrapping function, optimal enclosure and enclosures computed numerically for various step sizes h.*

Example 3.3 Consider the problem

$$\begin{aligned} \dot{x}_1 &= 2(t-1)x_2 \quad , \qquad x_1(0) = x_1^0 \in [0.9, 1.1] \\ \dot{x}_2 &= 2(t-1)x_1 \quad , \qquad x_2(0) = x_2^0 \in [-0.1, 0.1] \end{aligned}$$

for $t \geq 0$. This is a linear problem of the form (3.32)–(3.33) with

$$A(t) = 2\begin{pmatrix} 0 & t-1 \\ t-1 & 0 \end{pmatrix} \quad \text{and} \quad b(t) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

In the interval $[0, 1]$ matrix A is the same as in example 3.2 and using the same matrix Q we have $QA(t)Q = A^+(t)$. Therefore there is no wrapping effect. For $t > 1$ we have $A(t) = A^+(t)$. Therefore there is also no wrapping effect. However our expectations that this problem has no wrapping effect for $t \geq 0$ are false.

Figure 3.5, where the enclosures computed for various values of h, as well as the optimal enclosure and the wrapping function are plotted, presents a similar situation as in example 3.2, i.e. for $t > 1$ the computed enclosures approach the wrapping function which is wider then the optimal enclosure.

Using standard techniques we obtain the matricants of A and A^+ as follows:

$$M(A;t) = \begin{pmatrix} \cosh(t^2-2t) & \sinh(t^2-2t) \\ \sinh(t^2-2t) & \cosh(t^2-2t) \end{pmatrix}$$

and

$$M(A^+;t) = \begin{pmatrix} \cosh(1+|t-1|(t-1)) & \sinh(1+|t-1|(t-1)) \\ \sinh(1+|t-1|(t-1)) & \cosh(1+|t-1|(t-1)) \end{pmatrix}.$$

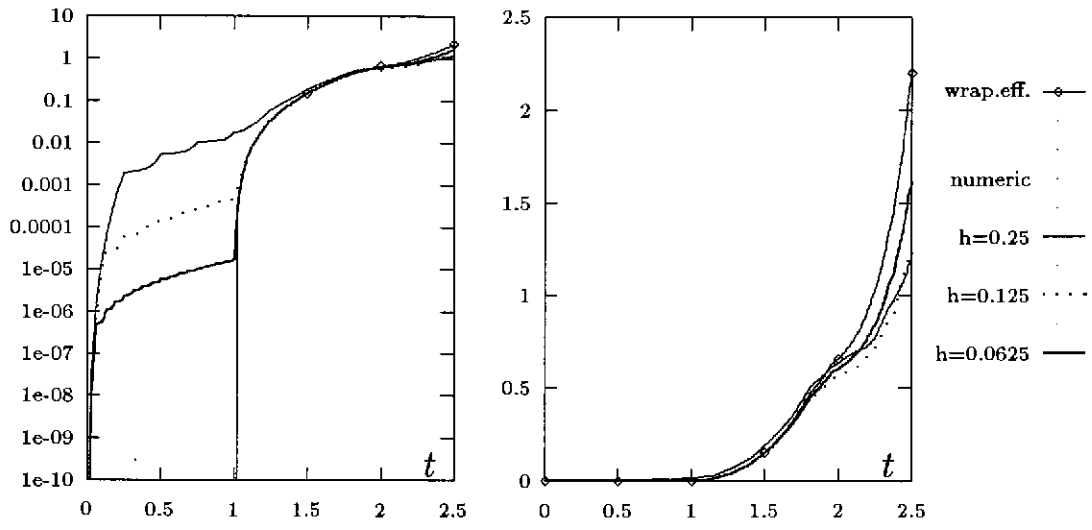


Figure 3.4: *Example 3.2. Wrapping effect measure* $\wp\left(\widehat{X}(t), [x(t_0, X^0; t)]\right)$ *and the errors* $\wp(S(h;t), [x(t_0, X^0; t)])$ *of the enclosures* $S(h;t)$ *computed numerically for various step sizes* h *on a logarithmic scale (left) and standard scale (right).*

Clearly $M(A^+;t) = |M(A;t)|$ for $t \in [0,1]$ and $M(A^+;t) > |M(A;t)|$ for $t > 1$. The wrapping effect measure (3.17) for this problem is

$$w(\widehat{X}(t)) - w([x(t_0, X^0; t)]) = \begin{cases} 0 & , \quad 0 \le t \le 1 \\ 2e \sinh(t-1)^2 \begin{pmatrix} 0.2 \\ 0.2 \end{pmatrix} & , \quad t \ge 1 \end{cases} .$$

The wrapping effect occurring in the interval $(1, 2.5)$ can not be generated in the same interval since, as we know from theorem 3.4, in both intervals $[0,1]$ and $[1,\infty)$ there is no wrapping effect and the wrapping function equals the optimal interval enclosure for any *interval* initial condition. At $t = 1$ we have

$$\widehat{X}(1) = [x(0, X^0; 1)] .$$

However, for $t > 1$ the optimal enclosure encloses the solutions propagated from the set $x(0, X^0; 1)$ which is not necessarily an interval and we have

$$[x(0, X^0; t)] = [x(1, x(0, X^0; 1); t)] .$$

The wrapping function in $[1,\infty)$ equals the optimal interval enclosure of the solutions propagated from the interval $[x(0, X^0; 1)]$ and we have

$$\widehat{X}(t) = [x(1, [x(0, X^0; 1)]; t)] .$$

The difference between the sets $[x(0, X^0; 1)]$ and $x(0, X^0; 1)$ (also called wrapping excess [45]) is what causes the inflation of the enclosures for $t > 1$. Let us note that here the

inflation does not increase with the increase of the number of points in the mesh as usual. The reason is that this inflation (or wrapping effect) results from the wrapping excess at one point only ($t = 1$) while the wrapping excess at other points has no contribution at all.

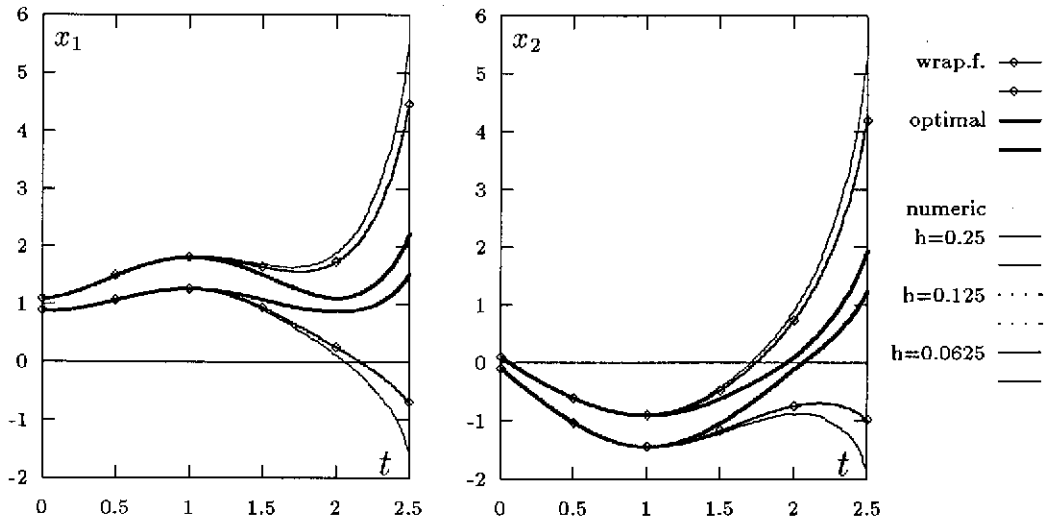


Figure 3.5: *Example 3.3. Wrapping function, optimal enclosure and enclosures computed numerically for various step sizes h.*

We stated earlier that in the intervals where the wrapping function differs from the optimal interval enclosure (i.e. the case of occurrence of wrapping effect), the computed enclosures usually converge to the wrapping function at a rate of $O(h)$ irrespective of the local error of the method. Example 3.3 is an exception. Here the global convergence is of order $O(h^4)$ in the whole interval $[0, 2.5]$ (see figure 3.6).

3.5 Necessary Condition for No Wrapping Effect: Linear Systems.

The condition in theorem 3.4 for linear problems with no wrapping effect is only a sufficient condition. We showed by example 2.1 that it is not a necessary condition. However, if problem (3.32)–(3.33) is irreducible and the initial interval X^0 contains inner points, i.e. $w(X^0) > 0$, the condition in theorem 3.4 is also necessary. Let us recall the definition of irreducible systems of differential equations.

Definition 3.1 *A system of equations of the form (3.1) is called reducible if there exist proper subsets $\mathcal{I}$ and $\mathcal{J}$ of the set $\mathcal{N} = \{1, 2, \ldots, n\}$ such that*

(i) $\mathcal{I} \cup \mathcal{J} = \mathcal{N}$,

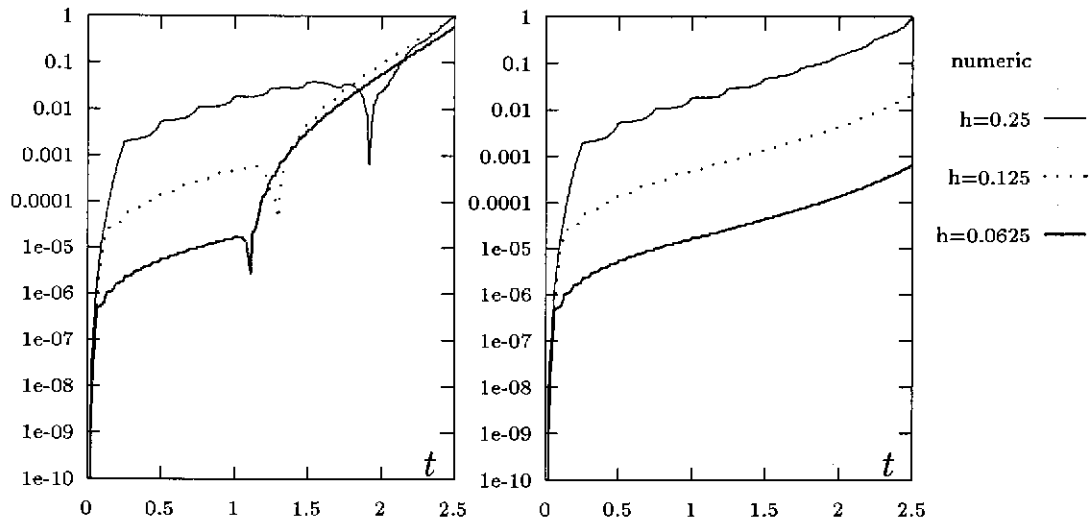


Figure 3.6: *Comparing the rate of convergence of the computed enclosures towards the wrapping function in Example 3.2 and Example 3.3. The distance* $\wp(S(h;t), \widehat{X}(t))$ *is plotted on logarithmic scale for Example 3.2 (left) and example 3.3 (right).*

(ii) if $i \in \mathcal{I}$ *and* $j \in \mathcal{N} \setminus \mathcal{I}$ *then* f_i *does not depend on* x_j*;*

(iii) if $i \in \mathcal{J}$ *and* $j \in \mathcal{N} \setminus \mathcal{J}$ *then* f_i *does not depend on* x_j*.*

A system of equations of the form (3.1) which is not reducible is called irreducible.

Obviously, if a system of differential equations is reducible then its solution reduces to the solution of two or more irreducible systems of smaller dimension. Therefore it is enough to formulate a necessary condition for irreducible systems of differential equations.

In the case of linear systems definition 3.1 assumes the following form:

Definition 3.2 *A system of the form (3.32) is called reducible if there exist proper subsets* $\mathcal{I}$ *and* $\mathcal{J}$ *of the set* $\mathcal{N} = \{1, 2, \ldots, n\}$ *such that*

(i) $\mathcal{I} \cup \mathcal{J} = \mathcal{N}$ *;*

(ii) if $i \in \mathcal{I}$ *and* $j \in \mathcal{N} \setminus \mathcal{I}$ *then* $a_{ij}(t) = 0,\ t \in [t_0, \overline{t}]$ *;*

(iii) if $i \in \mathcal{J}$ *and* $j \in \mathcal{N} \setminus \mathcal{J}$ *then* $a_{ij}(t) = 0,\ t \in [t_0, \overline{t}]$ *.*

A system of the form (3.32) which is not reducible is called irreducible.

We will prove the following theorem.

Theorem 3.5 *Let system (3.32) be irreducible. If problem (3.32)–(3.33) is a problem with no wrapping effect and* $w(X^0) > 0$ *then there exists a diagonal matrix* $Q = diag(q_1, q_2, \ldots, q_n)$, $q_i \in \{-1, 1\}$, $i = 1, \ldots, n$ *such that* $QA(t)Q = A^+(t)$, $t \in [t_0, \overline{t}]$.

Before we discuss the proof we will consider some preliminary results. In analogy with definition 3.2 we define reducible and irreducible matrices.

Definition 3.3 *A real matrix $P = (p_{ij})$ is called reducible if there exist proper subsets $\mathcal{I}$ and $\mathcal{J}$ of the set $\mathcal{N} = \{1, 2, \ldots, n\}$ such that*

(i) $\mathcal{I} \cup \mathcal{J} = \mathcal{N}$;

(ii) if $i \in \mathcal{I}$ *and* $j \in \mathcal{N} \setminus \mathcal{I}$ *then* $p_{ij} = 0$;

(iii) if $i \in \mathcal{J}$ *and* $j \in \mathcal{N} \setminus \mathcal{J}$ *then* $p_{ij} = 0$.

A matrix which is not reducible is called irreducible.

Lemma 3.1 *If a real matrix $P = (p_{ij})$, $p_{ij} \in \{0, 1\}$, $i, j \in \{1, \ldots, n\}$ is irreducible then*

$$\exp(P) = I + P + \frac{1}{2!}P^2 + \frac{1}{3!}P^3 + \ldots$$

contains a row in which none of the entries is zero.

Proof. We consider matrix P as an adjacency matrix of an oriented graph, i.e. $p_{ij} = 1$ implies that there is an arc which issues from the ith vertex and enters the jth vertex of the graph, $p_{ij} = 0$ implies that there is no such arc.

First we will prove that the graph has a vertex connected to any other vertex, i.e. there exist oriented paths from this vertex to any other vertex. Assume the oposite, i.e. every vertex is connected to no more than $k - 1$, $k < n$, verteces and let the vertex i_1 be conected to the verteces $i_2, i_3, \ldots, i_k$. Denote

$$\begin{aligned} \mathcal{N} &= \{1, 2, \ldots, n\} \\ \mathcal{I} &= \{i_1, i_2, \ldots, i_k\}, \\ \mathcal{J} &= \{i \in \mathcal{N} : \text{there exists a vertex } l \in \mathcal{N} \setminus \mathcal{I} \text{ which is connected to vertex } i\} \end{aligned}$$

If $i_1 \in \mathcal{J}$ then there exist a vertex $l \in \mathcal{N} \setminus \mathcal{I}$ connected to i_1 and therefore connected to the verteces $i_1, i_2, \ldots, i_k$. This contradicts the assumption that every vertex is connected to no more than $k - 1$ verteces. Therefore $i_1 \notin \mathcal{J}$ and both $\mathcal{I}$ and $\mathcal{J}$ are proper subsets of $\mathcal{N}$. Now we can see that the sets $\mathcal{I}$ and $\mathcal{J}$ satisfy conditions (i), (ii) and (iii) in definition 3.3.

(i): Since $\mathcal{N} \setminus \mathcal{I} \subset \mathcal{J}$ we have that $\mathcal{I} \cup \mathcal{J} = \mathcal{N}$.

(ii): Let $i \in \mathcal{I}$ and $j \in \mathcal{N} \setminus \mathcal{I}$. If $p_{ij} = 1$ then vertex i_1 is connected to vertex $j \notin \mathcal{I}$ which is a contradiction. Therefore $p_{ij} = 0$.

(iii): Let $i \in \mathcal{J}$ and $j \in \mathcal{N} \setminus \mathcal{J}$. If $p_{ij} = 1$ the definition of the set $\mathcal{J}$ implies that if $i \in \mathcal{J}$ then $j \in \mathcal{J}$. But $j \notin \mathcal{J}$. Therefore $p_{ij} = 0$.

This shows that matrix P is reducible. Since we are given that P is irreducible the assumption that every vertex is connected to no more than $k-1$ verteces, $k < n$, is false. Therefore, there exist a vertex connected to any other vertex of the graph.

Let vertex l be connected to any other vertex of the graph. We will show that the lth row of matrix $\exp(P)$ does not contain zeros. For any $j \in \mathcal{N} \setminus \{l\}$ there exist an oriented path from vertex l to vertex j. Denote by k_j the length of this path. Since P is the adjacency matrix of the graph, $\left(P^k\right)_{lj}$ equals to the number of arc sequences beginning from vertex l and ending at vertex j [42]. This implies

$$\left(P^{k_j}\right)_{lj} > 0 \ , \ j \in \mathcal{N} \setminus \{l\}.$$

Using that the entries of all matrices in the sum

$$\exp(P) = I + P + \frac{1}{2!}P^2 + \frac{1}{3!}P^3 + \dots$$

are nonnegative, we have

$$\begin{aligned} (\exp(P))_{lj} &\geq \frac{1}{k_j!}\left(P^{k_j}\right)_{lj} > 0 \ , \ j \in \mathcal{N} \setminus \{l\}, \\ (\exp(P))_{ll} &\geq (I)_{ll} = 1 \end{aligned}$$

which shows that all entries in the lth row of $\exp(P)$ are strictly positive.

Lemma 3.2 *If an irreducible real matrix $P = (p_{ij})$, $p_{ij} \in \{-1, 0, 1\}$, $i, j \in \{1, \dots, n\}$ is such that $|\exp(P)| = \exp(|P|)$ then there exists a diagonal matrix $Q = diag(q_1, q_2, \dots, q_n)$, $q_i \in \{-1, 1\}$, $i = 1, \dots, n$ such that $QPQ = |P|$.*

Proof. It is easy to see that the following sequence of inequalities holds true:

$$|\exp(P)| = \left|\sum_{k=0}^{\infty} \frac{1}{k!}P^k\right| \leq \sum_{k=0}^{\infty} \frac{1}{k!}|P^k| \leq \sum_{k=0}^{\infty} \frac{1}{k!}|P|^k = \exp(|P|)$$

Since $|\exp(P)| = \exp(|P|)$ all of the above inequalities are satisfied as equalities. This is true if and only if

$$\begin{aligned} &\text{for every } i, j \in \{1, \dots, n\} \text{ the sign of } \left(P^k\right)_{ij}, \text{ if not zero,} \\ &\text{is the same for all } k = 0, 1, 2, \dots \end{aligned} \tag{3.41}$$

$$|P^k| = |P|^k \ , \ k = 1, 2, \dots \tag{3.42}$$

Denote $V = \exp(P)$. Since matrix P is irreducible matrix $|P|$ is also irreducible and lemma 3.1 implies that matrix $|V| = |\exp(P)| = \exp(|P|)$ has a row which does not

contain zeros. Let this be the lth row. Then the lth row of matrix V also does not contain zeros. Denote $q_i = \text{sgn}(v_{li}) = \dfrac{v_{li}}{|v_{li}|}$. We will prove that $Q = diag(q_1, q_2, \ldots, q_n)$ is the required matrix.

Let us consider the matrix

$$VP = \sum_{k=0}^{\infty} \frac{1}{k!} P^{k+1}.$$

From conditions (3.41) and (3.42) it follows that

$$|VP| = \left| \sum_{k=0}^{\infty} \frac{1}{k!} P^{k+1} \right| = \sum_{k=0}^{\infty} \frac{1}{k!} |P^{k+1}| = \sum_{k=0}^{\infty} \frac{1}{k!} |P|^{k+1} = \exp(|P|)|P| = |V||P|$$

Therefore

$$\left| \sum_{k=0}^{\infty} v_{li} p_{ij} \right| = \sum_{k=0}^{\infty} |v_{li} p_{ij}|, \ j = 1, \ldots, n$$

This implies that for every $j = 1, \ldots, n$ the products $v_{li}p_{ij}$, $i = 1, \ldots, n$, if not zero, have all the same sign, equal to the sign of $(VP)_{lj}$. Furthermore, both V and VP are obtained as sum of powers of P with positive coefficients. Then, from condition (3.41) we can see that the entries of VP, if not zero, have the same sign as the entries of matrix V. Therefore the products $v_{li}p_{ij}$, $i, j = 1, \ldots, n$, if not zero, have the same sign as v_{lj}. This means that

$$v_{li} p_{ij} v_{lj} \geq 0 \ , \ i, j = 1, \ldots, n.$$

Hence

$$(QPQ)_{ij} = q_i p_{ij} q_j = \frac{v_{li} p_{ij} v_{lj}}{|v_{li}||v_{lj}|} = \frac{|v_{li} p_{ij} v_{lj}|}{|v_{li}||v_{lj}|} = |p_{ij}| \ , \ i, j = 1, \ldots, n$$

which conludes the proof of the lemma.

Lemma 3.3 *If matrix A in the system (3.32) is such that $|M(A;t)| = M(|A|;t)$, $t \in [t_0, \overline{t}]$ then there exists a diagonal matrix $Q = diag(q_1, q_2, \ldots, q_n)$, $q_i \in \{-1, 1\}$, $i = 1, \ldots, n$ such that $QA(t)Q = |A(t)|$, $t \in [t_0, \overline{t}]$.*

Proof. It is easy to see that the following sequence of inequalities holds true

$$\begin{aligned} |M(A;t)| &= \left| \sum_{k=0}^{\infty} M^{(k)}(A;t) \right| \\ &\leq \sum_{k=0}^{\infty} \left| M^{(k)}(A;t) \right| \\ &\leq \sum_{k=0}^{\infty} M^{(k)}(|A|;t) = M(|A|;t) \, . \end{aligned} \tag{3.43}$$

Since $|M(A;t)| = M(|A|;t)$, $t \in [t_0, \bar{t}]$, all inequalities in (3.43) are equalities for every $t \in [t_0, \bar{t}]$. In particular, we have

$$\left|\sum_{k=0}^{\infty} M^{(k)}(A;t)\right| = \sum_{k=0}^{\infty} \left|M^{(k)}(A;t)\right| \tag{3.44}$$

and

$$\left|M^{(k)}(A;t)\right| = M^{(k)}(|A|;t)\ ,\ k = 1,2,\ldots. \tag{3.45}$$

From (3.45), when $k = 1$, it follows that

$$\left|\int_{t_0}^{\bar{t}} a_{ij}(\theta)d\theta\right| = \int_{t_0}^{\bar{t}} |a_{ij}(\theta)|d\theta\ ,\ \ i,j \in \{1,\ldots,n\}$$

which implies that the entries a_{ij} of matrix A do not change sign in the interval $[t_0, \bar{t}]$, i.e. for every $i,j \in \{1,\ldots,n\}$

$$\text{either } a_{ij}(t) \leq 0\ ,\ \ t \in [t_0,\bar{t}] \quad \text{or} \quad a_{ij}(t) \geq 0\ ,\ \ t \in [t_0,\bar{t}]\ . \tag{3.46}$$

Let ϕ be a real function in the interval $[t_0, \bar{t}]$. Then $\text{sgn}\phi$ is defined (when possible) as

$$\text{sgn}\phi = \begin{cases} 1 & \text{if } \phi(t) \geq 0 \text{ for every } t \in [t_0,\bar{t}] \text{ and } \phi(\tilde{t}) > 0 \text{ for some } \tilde{t} \in [t_0,\bar{t}] \\ -1 & \text{if } \phi(t) \leq 0 \text{ for every } t \in [t_0,\bar{t}] \text{ and } \phi(\tilde{t}) < 0 \text{ for some } \tilde{t} \in [t_0,\bar{t}] \\ 0 & \text{if } \phi(t) = 0 \text{ for every } t \in [t_0,\bar{t}] \end{cases}$$

Condition (3.46) implies that $\text{sgn}a_{ij}$ is well defined for every $i,j \in \{1,\ldots,n\}$.

From (3.44) it follows that for any $i,j = 1,\ldots,n$ all functions $\left(M^{(k)}(A;t)\right)_{ij}$, $k = 0,1,2,\ldots$, if not constant zero, have the same sign which does not change when t varies in the interval $[t_0, \bar{t}]$. Therefore $\text{sgn}\left(M^{(k)}(A;.)\right)_{ij}$ is well defined and

$$\begin{aligned} &\text{for every } i,j \in \{1,\ldots,n\}\ \ \text{sgn}\left(M^{(k)}(A;.)\right)_{ij}, \text{ if not zero,} \\ &\text{is the same for all } k = 1,2,\ldots \end{aligned} \tag{3.47}$$

Denote $P = \text{sgn}A = (\text{sgn}a_{ij})$. Since the system (3.32) is irreducible then matrix P is also irreducible. Furthermore, it is easy to see that $\text{sgn}(P^k) = \text{sgn}(M^{(k)}(A;.))$. Then from (3.47) and (3.45) we obtain that the powers of P satisfy conditions (3.41) and (3.42) which implies that $|\exp(P)| = \exp(|P|)$. Using lemma 3.2 we obtain that there exists a matrix $Q = diag(q_1,\ldots,q_n)$, $q_i \in \{\pm 1\}$, $i = 1,\ldots,n$ such that $QPQ = |P|$. For every $i,j = 1,\ldots,n$ we have

$$\begin{aligned} (QA(t)Q)_{ij} &= q_i a_{ij}(t) q_j = q_i \text{sgn}(a_{ij})|a_{ij}(t)|q_j \\ &= q_i p_{ij} q_j |a_{ij}(t)| = |p_{ij}||a_{ij}(t)| = |a_{ij}(t)|,\ t \in [t_0,\bar{t}] \end{aligned}$$

Therefore $QA(t)Q = |A(t)|$, $t \in [t_0, \overline{t}]$.

Proof of theorem 3.5. Since all entries of the matrix $M(A^+;t) - |M(A;t)|$ are nonnegative and $w(X^0) > 0$ condition (3.36) implies that

$$M(A^+;t) = |M(A;t)| \ , \quad t \in [t_0, \overline{t}].$$

Let

$$\alpha = \max_{i=1,\ldots,n} \max_{t\in[t_0,\overline{t}]} a_{ii}(t) \ .$$

Then the diagonal entries of matrix $C(t) = A(t)+\alpha I$ are all nonnegative and $|C(t)| = C^+(t)$. We have

$$\begin{aligned} |M(C;t)| &= |M(A+\alpha I;t)| = |e^{\alpha t}M(A;t)| = e^{\alpha t}|M(A;t)| \\ &= e^{\alpha t}M(A^+;t) = M(A^+ + \alpha I;t) = M(C^+;t) = M(|C|;t) \ . \end{aligned}$$

Considering a system of the form (3.32) with a matrix C, from lemma 3.3 we obtain that there exists a matrix $Q = diag(q_1, \ldots, q_n)$, $q_i \in \{\pm 1\}$, $i \in \{1, \ldots, n\}$ such that $QC(t)Q = |C(t)|$, $t \in [t_0, \overline{t}]$. Then

$$\begin{aligned} QA(t)Q &= Q(C(t) - \alpha I)Q = QC(t)Q - \alpha I \\ &= C^+(t) - \alpha I = A^+(t) \ , \ t \in [t_0, \overline{t}] \end{aligned}$$

which proves the theorem.

3.6 Necessary Condition for No Wrapping Effect: General Case.

In order to prove a theorem similar to theorem 3.5 in the general case of nonlinear problems of the form (3.1)–(3.2) we will make some additional assumptions for function f. We will assume that function f is differentiable about x and its Jacobian $\dfrac{df}{dx}$

(i) is bounded, i.e. there exists a constant $n \times n$ matrix Λ such that

$$\left|\frac{df(t,x)}{dx}\right| \leq \Lambda \ , \quad t \in [t_0, \overline{t}], \ x \in D \quad \text{and} \tag{3.48}$$

(ii) satisfies a Lipschitz condition of the form

$$\left|\frac{df(t,\xi)}{dx} - \frac{df(t,\psi)}{dx}\right| \leq \|\xi - \psi\|\Gamma \ , \quad t \in [t_0, \overline{t}] \ , \ \xi, \psi \in D \tag{3.49}$$

where Γ is a constant $n \times n$ matrix.

The Jacobian $\frac{df}{dx}$, unlike the linear case, depends not only on t but on x as well. Therefore, it does not seem possible to obtain results about the monotonicity of f, given that the problem has no wrapping effect only for a fixed initial condition. We shall require that the problem has no wrapping effect for all initial conditions X^0 within a certain given interval $G \subset D$. We will assume that all solutions $x(t_0, x^0; t)$, $x^0 \in G$, exist in the interval $[t_0, \overline{t}]$. Obviously, we can derive properties of f only in the area

$$\mathcal{G} = \{(t, x(t_0, x^0; t)) : x^0 \in G\} \subset [t_0, \overline{t}] \times D$$

spanned by the solutions of (3.1) when the initial condition is in G.

The following inequalities

$$w([x(t_0, X^0; t)]) \leq w(\widehat{X}(t)) \leq M(\Lambda; t) w(X^0) \ , \ t \in [t_0, \overline{t}] \tag{3.50}$$

are easy to prove and will be used below.

Theorem 3.6 *Let system (3.1) be irreducible. Let also function f be differentiable about x and let its Jacobian satisfy conditions (3.48) and (3.49). If problem (3.1)–(3.2) has no wrapping effect for every initial condition $X^0 \subset G$ then there exist subsets $Z^{(1)}, Z^{(2)}, \ldots, Z^{(k)}$ of G such that*

(i) $Z^{(1)} \cup Z^{(2)} \cup \ldots \cup Z^{(k)} = G$ *and*

(ii) for every $j = 1, \ldots, k$ there exists a matrix $Q^{(j)} = diag(q_1^{(j)}, \ldots, q_n^{(j)})$, $q_i^{(j)} \in \{\pm 1\}$, $i = 1, \ldots, n$, such that the function $Q^{(j)} f(t, Q^{(j)} x)$ is quasi-isotone about x in any convex subset of $\mathcal{Z}^{(j)} = \{(t, Q^{(j)} x(t_0, x^0; t)) : x^0 \in Z^{(j)}\}$.

Proof. Let u be any interior point of G and let $e = (1, 1, \ldots, 1) \in \mathcal{R}^n$. Let δ be small enough positive real number so that the interval vector

$$X^0 = [u - \delta e, u + \delta e] \tag{3.51}$$

is in G. We consider problem (3.1)–(3.2) with initial condition X^0 given by (3.51).

Let $J(u; t) = \frac{df}{dx}(t, x(t_0, u; t))$ and $b(t, x) = f(t, x) - J(u; t)x$. Function f can be represented as

$$f(t, x) = J(u; t)x + b(t, x) \ .$$

Considering the interval extension of b we have the following estimate for every $t \in [t_0, \overline{t}]$ and interval $X \subset D$ such that $x(t_0, u; t) \in X$.

$$\begin{aligned} w(b(t, X)) &= \max_{\xi, \eta \in X} (f(t, \xi) - f(t, \eta) - J(u; t)(\xi - \eta)) \\ &\leq \max_{\xi, \eta, \psi \in X} \left(\left(\frac{df(t, \psi)}{dx} - \frac{df(t, x(t_0, u; t))}{dx} \right) (\xi - \eta) \right) \\ &\leq \max_{\xi, \eta, \psi \in X} \|\psi - x(t_0, u; t)\| \Gamma(\xi - \eta) \\ &\leq \|w(X)\| \Gamma w(X) \end{aligned}$$

From the above inequality, using (3.50) we have

$$\begin{aligned} w(b(t,[x(t_0,x^0;t)])) &\leq \|w([x(t_0,x^0;t)])\|\Gamma w([x(t_0,x^0;t)]) \\ &\leq \|M(\Lambda;t)w(X^0)\|\Gamma M(\Lambda;t)w(X^0) \\ &\leq 4\delta^2\|M(\Lambda;t)\|\Gamma M(\Lambda;t)e\,. \end{aligned} \tag{3.52}$$

In the same way

$$\begin{aligned} w(b(t,\widehat{X}(t))) &\leq \|w(\widehat{X}(t))\|\Gamma w(\widehat{X}(t)) \\ &\leq 4\delta^2\|M(\Lambda;t)\|\Gamma M(\Lambda;t)e\,. \end{aligned} \tag{3.53}$$

Problem (3.1)–(3.2) can be written as

$$\dot{x} = J(u;t)x + b(t,x) \tag{3.54}$$

$$x(t_0) = x^0 \in X^0 \tag{3.55}$$

Every solution $x(t_0,x^0;t)$ is represented in the form

$$x(t_0,x^0;t) = M(J;t)x^0 + M(J;t)\int_{t_0}^{t}(M(J;\theta))^{-1}b(\theta,x(t_0,x^0;\theta))d\theta\,.$$

Therefore the optimal interval enclosure $[x(t_0,X^0;t)]$ satisfies the inclusion

$$[x(t_0,X^0;t)] \subset M(J;t)X^0 + M(J;t)\int_{t_0}^{t}(M(J;\theta))^{-1}b(\theta,[x(t_0,X^0;\theta)])d\theta\,.$$

Hence, using (3.52), we obtain

$$\begin{aligned} & w\left([x(t_0,X^0;t)]\right) \\ &\leq |M(J;t)|w(X^0) + |M(J;t)|\int_{t_0}^{t}\left|(M(J;\theta))^{-1}\right| w\left(b(\theta,[x(t_0,X^0;\theta)])\right)d\theta \\ &\leq 2\delta|M(J;t)|e + 4\delta^2|M(J;t)|\int_{t_0}^{t}\left|(M(J;\theta))^{-1}\right|\|M(\Lambda;\theta)\|\Gamma M(\Lambda;\theta)ed\theta \\ &= 2\delta|M(J;t)|e + 4\delta^2\phi(t) \end{aligned} \tag{3.56}$$

where

$$\phi(t) = |M(J;t)|\int_{t_0}^{t}\left|(M(J;\theta))^{-1}\right|\|M(\Lambda;\theta)\|\Gamma M(\Lambda;\theta)d\theta\, e$$

is a continuous function of t which does not depend on δ.

In section 3.4 we obtained the width of the wrapping function in an explicit form (3.34). The same method, when applied to problem (3.54)–(3.55) produces a differential inequality of form

$$\frac{d}{dx}(w(\widehat{X}(t))) \geq J^+(u;t)w(\widehat{X}(t)) - w(b(t,\widehat{X}(t))\,.$$

Using (3.53) we obtain

$$\frac{d}{dx}(w(\widehat{X}(t))) \geq J^+(u;t)w(\widehat{X}(t)) - 4\delta^2\|M(\Lambda;t)\|\Gamma M(\Lambda;t)e\,. \tag{3.57}$$

Let us note that the modified Jacobian of the form J^+ is used in [88] where the stability of interval methods is studied using a different approach.

The solution of the linear problem

$$\dot{y} = J^+(u;t)y - 4\delta^2\|M(\Lambda;t)\|\Gamma M(\Lambda;t)e \tag{3.58}$$

$$y(t_0) = w(X^0) \tag{3.59}$$

can be represented as

$$y(t) = 2\delta M(J^+;t)e - 4\delta^2\varphi(t)$$

where

$$\varphi(t) = \|M(\Lambda;t)\| \int_{t_0}^{t} \left(M(J^+;\theta)\right)^{-1} \Gamma M(\Lambda;\theta)d\theta e$$

is a continuous function of t which does not depend on δ.

Since the right-hand side in the system (3.58) is a function which is quasi-isotone about y the differential inequality (3.57) implies that the width of the wrapping function is greater than or equal to the solution of (3.58)–(3.59) for every $t \in [t_0, \overline{t}]$. Hence

$$w(\widehat{X}(t)) \geq 2\delta M(J^+;t) - 4\delta^2\varphi(t)\,. \tag{3.60}$$

Since problem (3.1)–(3.2) is without wrapping effect we have

$$w(\widehat{X}(t)) \;=\; w([x(t_0, X^0;t)])\;,\quad t \in [t_0, \overline{t}]\,.$$

From (3.56) and (3.60) it follows that

$$2\delta M(J^+;t)e - 4\delta^2\varphi(t) \leq w(\widehat{X}(t)) \;=\; w([x(t_0, X^0;t)]) \leq 2\delta|M(J;t)|e + 4\delta^2\phi(t)\,.$$

Using also (3.38) we obtain

$$0 \;\leq\; 2\delta\left(M(J^+;t) - |M(J;t)|\right)e \;\leq\; 4\delta^2(\phi(t) + \varphi(t))$$

and dividing by 2δ

$$0 \;\leq\; \left(M(J^+;t) - |M(J;t)|\right)e \;\leq\; 2\delta(\phi(t) + \varphi(t))\,.$$

Since δ is arbitrary small positive it follows that

$$\left(M(J^+;t) - |M(J;t)|\right)e \;=\; 0\;,\quad t \in [t_0, \overline{t}]\,.$$

The coordinates of e are all positive and the entries of $M(J^+;t) - |M(J;t)|$ are all non-negative. Therefore

$$M(J^+;t) \;=\; |M(J;t)| \;\;,\;\; t \in [t_0, \overline{t}] \;.$$

This implies that a linear system of the form (3.32)–(3.33) with a matrix $A = J$ has no wrapping effect. Then from theorem 3.5 it follows that there exists a matrix

$$Q = diag(q_1, q_2, \ldots, q_n),\; q_i \in \{-1, 1\},\; i = 1, \ldots, n \tag{3.61}$$

such that $QJ(u;t)Q = J^+(u;t)$, $t \in [t_0, \overline{t}]$.

Thus, we proved that

$$\begin{aligned} &\text{for every } u \in G \text{ there exists a matrix } Q = Q(u) \\ &\text{of the form (3.61) such that } QJ(u;t)Q = J^+(u;t). \end{aligned} \tag{3.62}$$

Let $\mathcal{Q}$ be the set of all matrices of the form (3.61). Obviously, $\mathcal{Q}$ is a finite set. For every matrix Q we can consider the set

$$Z(Q) \;=\; \{u \in G : QJ(u;t)Q = J^+(u;t), t \in [t_0, \overline{t}]\} \;.$$

It follows from (3.62) that

$$\bigcup_{Q \in \mathcal{Q}} Z(Q) \;=\; G \;.$$

Excluding from the above union these sets which are empty, we obtain a finite number of sets $Z^{(1)}, Z^{(2)}, \ldots, Z^{(k)}$ such that for every set $Z^{(j)}$ there exists a matrix $Q^{(j)}$ of the form (3.61) such that

$$Q^{(j)} \frac{df(t,x)}{dx} Q^{(j)} \;=\; \left(\frac{df(t,x)}{dx}\right)^+ \;\;,\;\; (t,x) \in \{(t, x(t_0, u; t)) : u \in Z^{(j)}\} \;.$$

This implies that $Q^{(j)} f(t, Q^{(j)} x)$ is quasi-isotone about x in any convex subset of $\mathcal{Z}^{(j)}$ which concludes the proof.

Example 3.4 Consider the problem

$$\begin{aligned} \dot{x}_1 &= -x_1 x_2 \quad , \qquad x_1(0) = x_1^0 \in X_1^0 \\ \dot{x}_2 &= -x_1^2 \quad , \qquad x_2(0) = x_2^0 \in X_2^0 \end{aligned}$$

for $t \geq 0$ where $X^0 \subset G = ([-0.5, 0.5], [0.5, 1.5])^T$. The exact solution $x(0, x_0; t)$ of the above system for a given $x^0 \in X^0$ is

$$\begin{aligned} x_1(0, x^0; t) &= \operatorname{sgn}(x_1^0)\mu(x^0)\operatorname{cosech}(\mu(x^0)t + \eta(x^0)) \\ x_2(0, x^0; t) &= \mu(x^0)\coth(\mu(x^0)t + \eta(x^0)) \quad , \qquad \text{if } x_1^0 \neq 0 \end{aligned}$$

and (3.63)

$$\begin{aligned} x_1(0, x^0; t) &= 0 \\ x_2(0, x^0; t) &= x_2^0 \quad , \qquad \text{if } x_1^0 = 0 \end{aligned}$$

where

$$\mu(x^0) = \sqrt{(x_2^0)^2 - (x_1^0)^2} \quad \text{and} \quad \eta(x^0) = \ln\left(\frac{x_2^0 + \sqrt{(x_2^0)^2 - (x_1^0)^2}}{|x_1^0|}\right) .$$

Using (3.63) to obtain the optimal interval enclosure $[x(0, X^0; t)]$ and problem (3.4)–(3.5) to obtain the wrapping function $\widehat{X}(t)$ we can see that they are equal for any $X^0 \subset G$ Therefore the problem in this example is with no wrapping effect for every $X^0 \subset G$.

The Jacobian of the right-hand side of the system is

$$\frac{df}{dx} = \begin{pmatrix} -x_2 & -x_1 \\ -2x_1 & 0 \end{pmatrix} .$$

Therefore in $\{x \in \mathcal{R}^2 : x_1 \leq 0\}$ function f is quasi-isotone while in $\{x \in \mathcal{R}^2 : x_1 \geq 0\}$ it can be transformed into a quasi-isotone function using matrix $Q = diag(-1, 1)$.

Let $Z^{(1)} = ([-0.5, 0], [0.5, 1.5])^T$ and $Z^{(2)} = ([0, 0.5], [0.5, 1.5])^T$. Since

$$\begin{aligned} x(0, Z^{(1)}; t) &\subset \{x \in \mathcal{R}^2 : x_1 \leq 0\} \ , \quad t \geq 0 \quad \text{and} \\ x(0, Z^{(2)}; t) &\subset \{x \in \mathcal{R}^2 : x_1 \geq 0\} \ , \quad t \geq 0 \ , \end{aligned}$$

$Z^{(1)}$ and $Z^{(2)}$ are the subsets of G that exist according to theorem 3.6.

The necessary condition for no wrapping effect stated in theorem 3.6 is not the same as the sufficient condition in theorem 3.3. Nevertheless, they are quite close. In fact, theorem 3.6 implies that if a problem has no wrapping effect for every initial condition $X^0 \subset G$, the area $\mathcal{G}$ spanned by the solutions can be subdivided into areas where the monotonicity of f does not change (i.e. for every $i, j \in \{1, \ldots, n\}$, $i \neq j$ function f_i is either increasing or decreasing about x_j) and the solutions do not leave or enter any of those areas when t propagates from t_0 to $\bar{t}$. If we can make such a subdivision beforehand we can apply the following theorem which follows directly from theorems 3.3 and 3.6.

Theorem 3.7 *Let a function f be such that f is differentiable about x, its Jacobian satisfies conditions (3.48)–(3.49) and for every $i, j \in \{1, \ldots, n\}$, $i \neq j$ function f_i is either increasing or decreasing about x_j in the area*

$$\mathcal{G} = \{(t, x(t_0, u; t)) : u \in G\} \subset [t_0, \overline{t}] \times D$$

where $G \subset D$ is a given interval.

Then problem (3.1)–(3.2) is a problem with no wrapping effect if and only if there exists a matrix $Q = diag(q_1, \ldots, q_n)$, $q_i \in \{\pm 1\}$, $i = 1, \ldots, n$, such that $Qf(t, Qx)$ is quasi-isotone about x in $\{(t, x) : (t, Qx) \in \mathcal{G}\}$.

Chapter 4

Validated Solution of the Wave Equation.

We consider the nonlinear wave equation

$$u_{tt}(x,t) - u_{xx}(x,t) = f(x,t,u(x,t)) \,, \quad -l < x < l \,, \quad t > 0 \,, \tag{4.1}$$

$$u(x,0) = g_1(x) \,,\; u_t(x,0) = g_2(x) \,, \quad -l < x < l \quad , \tag{4.2}$$

$$u(-l,t) = u(l,t) \,,\; u_x(-l,t) = u_x(l,t) \,, \quad t > 0 \quad . \tag{4.3}$$

Condition (4.3) implies that the solution is a function which has a smooth $2l$-periodical extension about x. A periodic boundary condition is essential for the monotone properties of the problem and the construction of numerical methods discussed in the following sections. However, it is not a very restrictive assumption because a large number of problems can be reduced to problems with periodic boundary conditions of the form (4.1)–(4.3) (see section 2.5.2).

Let $\Omega[\underline{t},\overline{t}]$ be the set of all functions $u = u(x,t) : \mathcal{R} \times [\underline{t},\overline{t}] \mapsto \mathcal{R}$ which are $2l$-periodical about x and have continuous second derivatives. Assuming that functions f, g_1 and g_2 are extended periodically about x (period $2l$) we can formulate problem (4.1)–(4.3) in the following way:

Find $u \in \Omega[0,\overline{t}]$ such that

$$u_{tt}(x,t) - u_{xx}(x,t) = f(x,t,u(x,t)) \,, \quad x \in \mathcal{R} \,, \quad t \in [0,\overline{t}] \,, \tag{4.4}$$

$$u(x,0) = g_1(x) \,,\; u_t(x,0) = g_2(x) \,, \quad x \in \mathcal{R} \quad . \tag{4.5}$$

The solution of the above problem we denote by $u(0,g;x,t)$.

We also consider an interval initial condition of the form

$$\begin{aligned} u(x,0) &= g_1(x) \in G_1(x) = [\underline{g}_1(x), \overline{g}_1(x)] \,, \\ u_t(x,0) &= g_2(x) \in G_2(x) = [\underline{g}_2(x), \overline{g}_2(x)] \,, \; x \in \mathcal{R} \end{aligned} \tag{4.6}$$

where G_1 and G_2 are given interval functions. The set-valued function

$$u(0, G; x, t) = \{u(0, g; x, t) : g \in G\}$$

is considered a solution of problem (4.4)–(4.6).

In this chapter we assume that f is a continuous function of all arguments, $g_{1,2}$ ($\underline{g}_{1,2}, \overline{g}_{1,2}$) are differentiable and $g'_{1,2} \in L_2(-l, l)$ ($\underline{g}_{1,2}, \overline{g}'_{1,2} \in L_2(-l, l)$). In addition we also make the assumption that f is a non-decreasing function of u which implies monotone properties of the problem used in the construction of enclosures.

4.1 Monotone Properties

The importance of the monotone properties of a problem for the design of validated methods was discussed in section 2.3. Here we will establish monotone properties of problem (4.4)-(4.5) which will lead to a representation of the problem as an operator equation involving a suitable operator of monotone type.

Denote by L and Φ the following operators in $\Omega[\underline{t}, \overline{t}]$

$$L(u; x, t) = u_{tt}(x, t) - u_{xx}(x, t) ,$$

$$\Phi(u, t; y, z) = u(y, t) + u(z, t) + \int_y^z u_t(x, t)dx , \ y \le z .$$

Theorem 4.1 *Let $u, v \in \Omega[\underline{t}, \overline{t}]$. If $L(u) \le L(v)$ and $\Phi(u, \underline{t}) \le \Phi(v, \underline{t})$ then $u \le v$ and $\Phi(u, t) \le \Phi(v, t)$ for $t \in [\underline{t}, \overline{t}]$.*

Proof. Denote $w = v - u$. Since L and Φ are linear operators we have

$$\begin{aligned} L(w) &= L(v) - L(u) \ge 0 , \\ \Phi(w, \underline{t}) &= \Phi(v, \underline{t}) - \Phi(u, \underline{t}) \ge 0 . \end{aligned}$$

Let $y, z \in R$, $y \le z$, $t \in [\underline{t}, \overline{t}]$ and let us integrate $L(w)$ over a trapeze Γ with vertices $M(y, t)$, $N(z, t)$, $P(z + t - \underline{t}, \underline{t})$ and $Q(y - t + \underline{t}, \underline{t})$. We have

$$\mathcal{I} = \iint_\Gamma (w_{tt} - w_{xx})\, dxdt \ge 0 .$$

A simple application of Green's theorem gives

$$\begin{aligned} \mathcal{I} &= \oint_{\partial\Gamma} (-w_x dt - w_t dx) \\ &= \int_M^N w_t dx - \int_P^N (w_x dt + w_t dx) - \int_Q^P w_t dx + \int_Q^M (w_x dt + w_t dx) \end{aligned}$$

$$
\begin{aligned}
&= \int_M^N w_t dx + \int_P^N (w_x dx + w_t dt) - \int_Q^P w_t dx + \int_Q^M (w_x dx + w_t dt) \\
&= \int_y^z w_t(x,t)dx + w(z,t) - w(z+t-\underline{t},\underline{t}) \\
&\quad - \int_{y-t+\underline{t}}^{z+t-\underline{t}} w_t(x,\underline{t})dx + w(y,t) - w(y-t+\underline{t},\underline{t}) \\
&= \Phi(w,t;y,z) - \Phi(w,\underline{t};y-t+\underline{t},z+t-\underline{t}) \,.
\end{aligned}
$$

Therefore

$$\Phi(w,t;y,z) \geq \Phi(w,\underline{t};y-t+\underline{t},z+t-\underline{t}) \geq 0 \,.$$

Hence

$$\Phi(u,t) \leq \Phi(v,t) \,,\ t \in [\underline{t},\overline{t}] \,.$$

We also have

$$w(x,t) = \frac{1}{2}\Phi(w,t;x,x) \geq 0$$

which implies

$$u(x,t) \leq v(x,t) \,,\ x \in R \,,\ t \in [\underline{t},\overline{t}] \,.$$

Theorem 4.2 *Let $f(x,t,u)$ be Lipschitzian and monotonically increasing about the last argument and let $u, v \in \Omega[\underline{t},\overline{t}]$. If $L(u) - f(\cdot,\cdot,u) \leq L(v) - f(\cdot,\cdot,v)$ and $\Phi(u,\underline{t}) \leq \Phi(v,\underline{t})$ then $u \leq v$ and $\Phi(u,t) \leq \Phi(v,t)$, $t \in [\underline{t},\overline{t}]$*

Proof. Let f be Lipschitzian with a coefficient k and α be such that $\alpha^2 > k$. Let ε be any positive number. Denote $w = v - u + \varepsilon e^{\alpha(t-\underline{t})}$. We have

$$\Phi(w,\underline{t};y,z) = \Phi(v,\underline{t};y,z) - \Phi(u,\underline{t};y,z) + 2\varepsilon + \varepsilon\alpha(z-y) \geq 2\varepsilon > 0 \,.$$

Therefore $w(x,\underline{t}) = \frac{1}{2}\Phi(w,\underline{t};x,x) \geq \varepsilon > 0$

We wish to show that $w(x,t) > 0$, $x \in R$, $t \in [\underline{t},\overline{t}]$. To do this we assume, to the contrary, that $w(x,t) \leq 0$ for some $x \in R$, $t \in [\underline{t},\overline{t}]$. Denote

$$t' = \inf\{t : w(x,t) \leq 0 \text{ for some } x \in R\} \,.$$

Then

$$w(x,t) = v(x,t) - u(x,t) + \varepsilon e^{\alpha(t-\underline{t})} > 0 \,,\ x \in R,\ t \in [\underline{t},t')$$

and there exists $x' \in R$ such that $w(x',t') = 0$.

For $(x,t) \in R \times [\underline{t},t')$ we have

$$
\begin{aligned}
L(w;x,t) &= L(v;x,t) - L(u;x,t) + \varepsilon\alpha^2 e^{\alpha(t-\underline{t})} \\
&\geq f(x,t,v(x,t)) - f(x,t,u(x,t)) + \varepsilon\alpha^2 e^{\alpha(t-\underline{t})} \\
&\geq f(x,t,u(x,t) - \varepsilon e^{\alpha(t-\underline{t})}) - f(x,t,u(x,t)) + \varepsilon\alpha^2 e^{\alpha(t-\underline{t})} \\
&\geq -k\varepsilon e^{\alpha(t-\underline{t})} + \varepsilon\alpha^2 e^{\alpha(t-\underline{t})} \\
&= \varepsilon(\alpha^2 - k)e^{\alpha(t-\underline{t})} > 0 \,.
\end{aligned}
$$

Let us integrate $L(w)$ over a triangle Γ with vertices $A(x'-t'+\underline{t},\underline{t})$, $B(x'+t'-\underline{t},\underline{t})$ and $C(x',t')$. We have

$$\mathcal{I} = \int\int_{\Gamma}(w_{tt} - w_{xx})dxdt > 0$$

Applying Green's theorem we obtain

$$\begin{aligned}
\mathcal{I} &= \oint_{\partial\Gamma}(-w_x dt - w_t dx) \\
&= \int_A^C (w_x dt + w_t dx) - \int_B^C (w_x dt + w_t dx) - \int_A^B w_t dx \\
&= \int_A^C (w_x dx + w_t dt) + \int_B^C (w_x dx + w_t dt) - \int_A^B w_t dx \\
&= 2w(x',t') - w(x'-t'+\underline{t},\underline{t}) - w(x'+t'-\underline{t},\underline{t}) - \int_{x'-t'+\underline{t}}^{x'+t'-\underline{t}} w_t(x,\underline{t})dx \\
&= 2w(x',t') - \Phi(w,\underline{t};x'-t'+\underline{t},x'+t'-\underline{t}) .
\end{aligned}$$

Therefore $w(x',t') > \frac{1}{2}\Phi(w,\underline{t};x'-t'+\underline{t},x'+t'-\underline{t}) > 0$. But $w(x',t') = 0$, a contradiction. This implies that $w(x,t) > 0$, $x \in R$, $t \in [\underline{t},\overline{t}]$. Hence

$$v(x,t) - u(x,t) > \varepsilon e^{\alpha(t-\underline{t})} , \ x \in R , \ t \in [\underline{t},\overline{t}]$$

for any positive ε. Letting $\varepsilon \to 0$ we conclude that

$$u(x,t) \leq v(x,t) , \ x \in R , \ t \in [\underline{t},\overline{t}] .$$

Using this inequality we have

$$L(v) - L(u) \geq f(\cdot,\cdot,v) - f(\cdot,\cdot,u) \geq 0$$

Then theorem 4.1 implies

$$\Phi(u,t) \leq \Phi(v,t) , \ t \in [\underline{t},\overline{t}]$$

which concludes the proof.

An obvious way of writing problem (4.4)-(4.5) in an operator form is by using operator $T(t_\alpha)$ defined in $\Omega[t_\alpha,\overline{t}]$, $t_\alpha \in [0,\overline{t})$ as

$$T(t_\alpha,u;x,t) = (Lu(x,t) - f(x,t,u), u(x,t_\alpha), u_t(x,t_\alpha)) , \ x \in \mathcal{R}, \ t \in [t_\alpha,\overline{t}] .$$

Then problem (4.4)–(4.5) can be written as

$$T(0,u) = (0,g_1,g_2) .$$

Theorem 4.3 *If f is a non-decreasing function of u then $T(t_\alpha)$ is an operator of monotone type.*

Proof. Let $u, v \in \Omega[t_\alpha, \overline{t}]$ and let $T(t_\alpha, u) \leq T(t_\alpha, v)$. Then we have

$$L(u; x, t) - f(x, t, u(x,t)) \leq L(v; x, t) - f(x, t, v(x,t)) \,, \quad x \in \mathcal{R},\ t \in [t_\alpha, \overline{t}] \tag{4.7}$$

and

$$\begin{aligned} u(x, t_\alpha) &\leq v(x, t_\alpha) \,, \quad x \in \mathcal{R} \\ u_t(x, t_\alpha) &\leq v_t(x, t_\alpha) \,, \quad x \in \mathcal{R} \,. \end{aligned}$$

From the above two inequalities we obtain

$$\begin{aligned} \Phi(u, t_\alpha; y, z) &= u(y, t_\alpha) + u(z, t_\alpha) + \int_y^z u_t(\xi, t_\alpha) d\xi \\ &\leq v(y, t_\alpha) + v(z, t_\alpha) + \int_y^z v_t(\xi, t_\alpha) d\xi \\ &= \Phi(v, t_\alpha; y, z) \,, \quad y, z \in \mathcal{R},\ y \leq z \,. \end{aligned} \tag{4.8}$$

Using theorem 4.2 from (4.7) and (4.8) it follows that

$$\Phi(u, t; y, z) \leq \Phi(v, t; y, z) \,, \quad t \in [t_\alpha, \overline{t}],\ y, z \in \mathcal{R},\ y \leq z \,.$$

Taking $y = z = x$ we have

$$u(x,t) = \frac{1}{2}\Phi(u, t; x, x) \leq \frac{1}{2}\Phi(v, t; x, x) = v(x,t) \,, \quad x \in \mathcal{R},\ t \in [t_\alpha, \overline{t}]$$

which concludes the proof.

Theorem 4.3 implies that when f is non-decreasing about u the optimal enclosure $[u(0, G; x, t)]$ of the solution of problem (4.4)–(4.6) can be represented in the form

$$[u(0, G; x, t)] = [u(0, \underline{g}; x, t), u(0, \overline{g}; x, t)]$$

and problem (4.4)–(4.6) is reduced to two problems with point initial conditions given by $\underline{g} = (\underline{g}_1, \underline{g}_2)$ and $\overline{g} = (\overline{g}_1, \overline{g}_2)$ as follows:

$$\begin{aligned} u_{tt}(x,t) - u_{xx}(x,t) &= f(x, t, u(x,t)) \,, \quad x \in \mathcal{R} \,, \quad t > 0 \\ u(x,0) = \underline{g}_1(x) \,, \ u_t(x,0) &= \underline{g}_2(x) \,, \quad x \in \mathcal{R} \end{aligned} \tag{4.9}$$

$$\begin{aligned} u_{tt}(x,t) - u_{xx}(x,t) &= f(x, t, u(x,t)) \,, \quad x \in \mathcal{R} \,, \quad t > 0 \\ u(x,0) = \overline{g}_1(x) \,, \ u_t(x,0) &= \overline{g}_2(x) \,, \quad x \in \mathcal{R} \end{aligned} \tag{4.10}$$

However, as was shown in the preliminaries (section 2.5.3), the practical application of monotonicity of the form provided by theorem 4.3 to the construction of enclosures has a significant shortcoming when the enclosures are constructed step-by-step using a mesh $\{t_0 = 0, t_1, \ldots, t_{\overline{j}} = \overline{t}\}$ in the time dimension.

Operator $\mathcal{T}(t_\alpha)$ is defined on $\Omega[t_\alpha, \overline{t}]$, $t_\alpha \in [t_0, \overline{t})$ as follows:

$$\mathcal{T}(t_\alpha, u; x, t, y, z)) = (Lu - f(x, t, u), \Phi(u, t_\alpha; y, z)) \,, \quad t \in [t_\alpha, \overline{t}],\ x, y, z \in \mathcal{R},\ y \leq z \,.$$

Theorem 4.4 *For every $u, v \in \Omega[t_\alpha, \overline{t}]$ we have*

$$\mathcal{T}(t_\alpha, u) \leq \mathcal{T}(t_\alpha, v) \implies \Phi(u,t) \leq \Phi(v,t)\,, \quad t_\alpha \leq t \leq \overline{t}\,.$$

Proof. The inequality $\mathcal{T}(t_\alpha, u) \leq \mathcal{T}(t_\alpha, v)$ implies that

$$L(u;x,t) - f(x,t,u(x,t)) \leq L(v;x,t) - f(x,t,v(x,t))\,, \quad x \in \mathcal{R},\ t \in [t_\alpha, \overline{t}]\,,$$
$$\Phi(u, t_\alpha; y, z) \leq \Phi(v, t_\alpha; y, z)\,, \quad y, z \in \mathcal{R},\ y \leq z\,.$$

Then the inequality

$$\Phi(u,t) \leq \Phi(v,t)\,, \quad y \in [t_\alpha, \overline{t}]$$

follows from theorem 4.2.

Let us note that for any $u, v \in \Omega[t_0, \overline{t}]$ and $t \in [t_0, \overline{t}]$ we have

$$(u(x,t) \leq v(x,t),\ u_t(x,t) \leq v_t(x,t),\ x \in \mathcal{R}) \implies (\Phi(u,t;y,z) \leq \Phi(v,t;y,z),\ y,z \in \mathcal{R}, y \leq z)$$
$$(\Phi(u,t;y,z) \leq \Phi(v,t;y,z),\ y,z \in \mathcal{R}, y \leq z) \implies (u(x,t) \leq v(x,t),\ x \in \mathcal{R})$$

but the implication

$$(\Phi(u,t;y,z) \leq \Phi(v,t;y,z),\ y,z \in \mathcal{R}, y \leq z) \implies (u_t(x,t) \leq v_t(x,t)\ x \in \mathcal{R})$$

is false. Then, it is easy to see that the operator $\mathcal{T}(t_\alpha)$ is an operator of monotone type according to the usual definition, but it is actually more than that since the inequality $\Phi(u,t) \leq \Phi(v,t)$ contains more information than $u(x,t) \leq v(x,t)$, $x \in \mathcal{R}$.

Let us define a partial ordering $\preceq$ in $\Omega[\underline{t}, \overline{t}]$ as

$$u \preceq v \overset{def}{\iff} \Phi(u,t) \leq \Phi(v,t)\,, \quad t \in [\underline{t}, \overline{t}]\,.$$

Then theorem 4.4 implies that operator $\mathcal{T}(t_\alpha)$ is an operator of monotone type with regard to the partial ordering $\preceq$ in $\Omega[t_\alpha, \overline{t}]$. Using operator $\mathcal{T}$ problem (4.4)–(4.5) can be written as

$$\mathcal{T}(0,u;x,t,y,z) = \left(0, g_1(y) + g_1(z) + \int_y^z g_2(\xi)d\xi\right)\,, \quad t \in [0,\overline{t}],\ x,y,z \in \mathcal{R},\ y \leq z\,.$$

Why the monotone property provided by operator $\mathcal{T}$ is applicable to construction of bounds for the solution step-by-step in the time dimension, can be explained as follows. In constructing a lower bound $\underline{s}(h, N; x, t)$ in the interval $[t_0, t_1]$ we use

$$\underline{s}(h,N;x,0) = \underline{g}_1 \leq u(0,g;x,0)\,, \quad \underline{s}_t(h,N;x,0) = \underline{g}_2 \leq u_t(0,g;x,0)\,, \quad x \in \mathcal{R},\ g \in G$$

and therefore

$$\Phi(\underline{s}(h,N), t_0) \leq \Phi(u(0,\underline{g}), t_0)\,, \quad g \in G\,.$$

In order to use an already computed bound $\underline{s}(h,N;x,t)$, $x \in \mathcal{R}$, $t \in [t_0,t_j]$ as an initial condition in the next interval $[t_j,t_{j+1}]$ it needs to satisfy

$$\Phi(\underline{s}(h,N),t_j) \leq \Phi(u(0,g),t_j)\ ,\quad g \in G\ . \tag{4.11}$$

If $\underline{s}(h,N)$ is constructed in the interval $[t_0,t_1]$ in such a way that $L\underline{s}(h,N) \leq 0$, then at $t = t_1$ we have

$$\Phi(\underline{s}(h,N),t_1) \leq \Phi(u(0,g),t_1)\ ,\quad g \in G\ .$$

Therefore the condition (4.11) is "self-generating" along the mesh. This is not true for the conditions

$$\underline{s}(h,N;x,t_j) \leq u(0,g,x,t_j)\ \ ,\ \ \underline{s}_t(h,N;x,t_j) \leq u_t(0,g;x,t_j)\ ,\ \ x \in \mathcal{R}\ ,\ g \in G$$

which would be required if the monotone property of the operator T was applied. Similar statements hold true for the construction of an upper bound $\overline{s}(h,N)$.

4.2 General Outline of the Method

The main idea is to construct lower and upper bounds as solutions of initial value problems derived from (4.4)-(4.5). We consider a mesh $\{t_j = jh : j = 0,1,\ldots,\bar{j}\}$ in the time dimension. In every interval $[t_j,t_{j+1}]$ we consider a pair of problems

$$\begin{aligned} u_{tt}(x,t) - u_{xx}(x,t) &= \underset{\sim}{f}(x,t,u(x,t))\ ,\quad x \in \mathcal{R}\ ,\quad t \in [t_j,t_{j+1}] \\ u(x,t_j) = \underset{\sim}{g}_{j1}(x)\ ,\ u_t(x,t_j) &= \underset{\sim}{g}_{j2}(x)\ ,\quad x \in \mathcal{R} \end{aligned} \tag{4.12}$$

$$\begin{aligned} u_{tt}(x,t) - u_{xx}(x,t) &= \tilde{f}(x,t,u(x,t))\ ,\quad x \in \mathcal{R}\ ,\quad t \in [t_j,t_{j+1}] \\ u(x,0) = \tilde{g}_{j1}(x)\ ,\ u_t(x,0) &= \tilde{g}_{j2}(x)\ ,\quad x \in \mathcal{R} \end{aligned} \tag{4.13}$$

where

- functions $\underset{\sim}{f}$ and $\tilde{f}$ are lower and upper bounds of suitable form for f, i.e. we have

$$\underset{\sim}{f}(x,t,u) \leq f(x,t,u) \leq \tilde{f}(x,t,u)\,, x,u \in R,\ t \in [t_j,t_{j+1}]\,, \tag{4.14}$$

- functions $\underset{\sim}{g}_{01}$, $\underset{\sim}{g}_{02}$ are lower bounds of suitable form for $\underline{g}_1$ and $\underline{g}_2$ and $\tilde{g}_{01}$, $\tilde{g}_{02}$ are upper bounds of suitable form for $\overline{g}_1$ and $\overline{g}_2$ respectively, i.e. we have

$$\begin{aligned} \underset{\sim}{g}_{01} \leq \underline{g}_1 \leq \overline{g}_1 \leq \tilde{g}_{01}\ , \\ \underset{\sim}{g}_{02} \leq \underline{g}_2 \leq \overline{g}_2 \leq \tilde{g}_{02}\ , \end{aligned} \tag{4.15}$$

- for $j \geq 1$

$$g_{j1}(x) = \underline{s}(h, N; x, t_j) \ , \ \tilde{g}_{j1}(x) = \overline{s}(h, N; x, t_j) \ ,$$
$$g_{j2}(x) = \underline{s}_t(h, N; x, t_j), \ \ \tilde{g}_{j2}(x) = \overline{s}_t(h, N; x, t_j) \ , \ \ x \in \mathcal{R}$$

assuming that the bounds $\underline{s}(h, N; x, t)$, $\overline{s}(h, N; x, t)$ are already computed for $t \leq t_j$ and satisfy conditions

$$\Phi(\underline{s}(h, N), t_j) \leq \Phi(u(0, g), t_j) \leq \Phi(\overline{s}(h, N), t_j) \ , \ g \in G \ . \tag{4.16}$$

Denote by $\underline{u}$ and $\tilde{u}$ the solutions of problems (4.12) and (4.13), respectively. Using the inequalities (4.15) (when $j = 0$) and the inequalities (4.16) (when $j \geq 1$) we obtain that at the initial point t_j of the interval $[t_j, t_{j+1}]$ functions $\underline{u}$ and $\tilde{u}$ satisfy

$$\Phi(\underline{u}, t_j) \leq \Phi(u(0, g), t_j) \leq \Phi(\tilde{u}, t_j) \ , \ g \in G \ . \tag{4.17}$$

From (4.14) we also obtain

$$L\underline{u} - f(\cdot, \cdot, \underline{u}) \leq L\underline{u} - \underline{f}(\cdot, \cdot, \underline{u}) = 0 = Lu(0, g) - f(\cdot, \cdot, u(0, g)) \ ,$$
$$L\tilde{u} - f(\cdot, \cdot, \tilde{u}) \geq L\tilde{u} - \tilde{f}(\cdot, \cdot, \tilde{u}) = 0 = Lu(0, g) - f(\cdot, \cdot, u(0, g)) \ . \tag{4.18}$$

Then theorem 4.2 and inequalities (4.17), (4.18) imply that the solutions $\underline{u}$ and $\tilde{u}$ of problems (4.12) and (4.13) are lower and upper bounds for every solution $u(0, g)$, $g \in G$ of problem (4.4)–(4.5).

In addition, from Theorem 4.2, we also have

$$\Phi(\underline{u}, t) \leq \Phi(u(0, g), t) \leq \Phi(\tilde{u}, t) \ , \ t \in [t_j, t_{j+1}] \ .$$

Lower and upper bounds for $u(0, G)$ can be obtained from (4.12) and (4.13), provided those problems can be solved in some constructive way. In general, we can obtain only approximations $\underline{u}^{(*)}$, $\tilde{u}^{(*)}$ to the solutions $\underline{u}$, $\tilde{u}$ of (4.12) and (4.13) using certain numerical procedures. In doing so, we must ensure that $\underline{u}^{(*)} \leq u(0, g) \leq \tilde{u}^{(*)}$, $g \in G$ is satisfied. We solve (4.12) and (4.13) iteratively. Given some suitable initial bounds $\underline{u}^{(0)}, \tilde{u}^{(0)} \in \Omega[0, \overline{t}]$, sequences $\{\underline{u}^{(r)}\}, \{\tilde{u}^{(r)}\} \subset \Omega[0, \overline{t}]$ are defined recursively with $\underline{u}^{(r+1)}$ being a solution of

$$u_{tt}(x, t) - u_{xx}(x, t) = \underline{f}(x, t, \underline{u}^{(r)}(x, t))$$
$$u(x, t_j) = \underline{g}_{j1}(x) \ , \ \ u_t(x, t_j) = \underline{g}_{j2}(x) \tag{4.19}$$

and $\tilde{u}^{(r+1)}$ a solution of

$$\begin{aligned} u_{tt}(x,t) - u_{xx}(x,t) &= \tilde{f}(x,t,\tilde{u}^{(r)}(x,t)) \\ u(x,0) = \tilde{g}_{j1}(x)\ ,\ \ u_t(x,0) &= \tilde{g}_{j2}(x)\ \ . \end{aligned} \tag{4.20}$$

Provided the initial functions $\underset{\sim}{u}^{(0)}$, $\tilde{u}^{(0)}$ satisfy the following conditions

$$\begin{aligned} L(\underset{\sim}{u}^{(0)}) &\leq \underset{\sim}{f}(\cdot,\cdot,\underset{\sim}{u}^{(0)}) \\ L(\tilde{u}^{(0)}) &\geq \tilde{f}(\cdot,\cdot,\tilde{u}^{(0)}) \\ \Phi(\underset{\sim}{u}^{(0)},t_j) \leq \Phi(u(0,g),t_j) &\leq \Phi(\tilde{u}^{(0)},t_j)\ ,\ \ g \in G \end{aligned}$$

it can be proved inductively that

$$\begin{aligned} \underset{\sim}{u}^{(r)} \leq u(0,g) &\leq \tilde{u}^{(r)}\ , \qquad\qquad g \in G,\ r = 0,1,2,\ldots \\ \Phi(\underset{\sim}{u}^{(r)},t) \leq \Phi(u(0,g),t) &\leq \Phi(\tilde{u}^{(r)},t)\ ,\ \ t \in [t_j,t_{j+1}],\ g \in G,\ r = 0,1,2,\ldots \end{aligned} \tag{4.21}$$

Indeed, for r=0, using

$$\begin{aligned} L(\underset{\sim}{u}^{(0)}) - f(\cdot,\cdot,\underset{\sim}{u}^{(0)}) \leq L(\underset{\sim}{u}^{(0)}) - \underset{\sim}{f}(\cdot,\cdot,\underset{\sim}{u}^{(0)}) \leq 0\ \ , \\ L(\tilde{u}^{(0)}) - f(\cdot,\cdot,\tilde{u}^{(0)}) \geq L(\tilde{u}^{(0)}) - \tilde{f}(\cdot,\cdot,\tilde{u}^{(0)}) \geq 0\ \ , \\ \Phi(\underset{\sim}{u}^{(0)},t_j) \leq \Phi(u(0,g),t_j) \leq \Phi(\tilde{u}^{(0)},t_j)\ , \end{aligned}$$

from Theorem 4.2 we obtain $\underset{\sim}{u}^{(0)} \leq u(0,g) \leq \tilde{u}^{(0)}$ and $\Phi(\underset{\sim}{u}^{0},t) \leq \Phi(u(0,g),t) \leq \Phi(\tilde{u}^{(0)},t)$, $t \in [t_i,t_{j+1}]$.

Let (4.21) be true for some r. Since $\underset{\sim}{u}^{(r+1)}$ and $\tilde{u}^{(r+1)}$ are solutions to (4.19) and (4.20) we have

$$\begin{aligned} L(\underset{\sim}{u}^{(r+1)}) = \underset{\sim}{f}(\cdot,\cdot,\underset{\sim}{u}^{(r)}) \leq f(\cdot,\cdot,\underset{\sim}{u}^{(r)}) \leq f(\cdot,\cdot,u(0,g)) = Lu(0,g) \\ L(\tilde{u}^{(r+1)}) = \tilde{f}(\cdot,\cdot,\tilde{u}^{(r)}) \geq f(\cdot,\cdot,\tilde{u}^{(r)}) \geq f(\cdot,\cdot,u(0,g)) = Lu(0,g) \\ \Phi(\underset{\sim}{u}^{(r+1)},0) \leq \Phi(u(0,g),0) \leq \Phi(\tilde{u}^{(r+1)},0) \end{aligned}$$

Then Theorem 4.1 implies $\underset{\sim}{u}^{(r+1)} \leq u(0,g) \leq \tilde{u}^{(r+1)}$ and $\Phi(\underset{\sim}{u}^{(r+1)},t) \leq \Phi(u(0,g),t) \leq \Phi(\tilde{u}^{(r+1)},t)$, $t \in [t_j,t_{j+1}]$. Therefore inequalities (4.21) are true for any $r = 0,1,2,\ldots$.

Using similar arguments it can be also proved that if $\underset{\sim}{f}$ and $\tilde{f}$ are increasing about u (this may be expected because f is increasing about u) then $\{\underset{\sim}{u}^{(r)}\}$ is an increasing sequence and $\{\tilde{u}^{(r)}\}$ is a decreasing sequence, i.e.

$$\underset{\sim}{u}^{(0)} \leq \underset{\sim}{u}^{(1)} \leq \ldots \leq \underset{\sim}{u}^{(r)} \leq \ldots \leq u(0,g) \leq \ldots \leq \tilde{u}^{(r)} \leq \ldots \leq \tilde{u}^{(1)} \leq \tilde{u}^{(0)}\ .$$

After a sufficient number r^* of iterations we take $\underline{s}(h,N) = \underline{u}^{(r^*)}$ and $\overline{s}(h,N) = \tilde{u}^{(r^*)}$ as approximate solutions to (4.12) and (4.13).

Problems (4.19) and (4.20) are solved by computations in the Cartesian product of the Taylor functoid and the Fourier functoid. The solutions of problems (4.19) and (4.20) are obtained as Fourier series of x with coefficients that are polynomials of t. Since functions $\underset{\sim}{g}_{j1}$, $\underset{\sim}{g}_{j2}$, $\tilde{g}_{j1}$ and $\tilde{g}_{j2}$, representing the initial conditions in problems (4.19) and (4.20), when $j \geq 1$, result from computations in the previous time interval, they are already functions in the Fourier functoid. Functions $\underset{\sim}{g}_{01}$, $\underset{\sim}{g}_{02}$, $\tilde{g}_{01}$, $\tilde{g}_{02}$ are obtained from $\underline{g}_1$, $\underline{g}_2$, $\overline{g}_1$, $\overline{g}_2$ by using directed Fourier roundings. For functions $\underset{\sim}{f}$ and $\tilde{f}$ it will be difficult to produce a suitable definition for every $x, u \in \mathcal{R}$, $t \in [t_j, t_{j+1}]$. However, for the implementation of the iterative procedure we need only expressions for $\underset{\sim}{f}(x,t,u(x,t))$ and $\tilde{f}(x,t,u(x,t))$ where u is a given function. Such expressions are obtained using directed Fourier and Taylor roundings. In the next section we will revisit the Fourier functoid, particularly considering the directed roundings. In order to simplify the presentation we will consider periodical functions with period 2, i.e. following the notations adopted in this section we take $l = 1$.

4.3 Fourier Functoid: Interval and Directed Roundings.

The Fourier functoid $\mathcal{F}_N$ is defined as the span of $\{\cos(k\pi x), \sin(k\pi x)\}_{k=0}^{N}$, i.e we have

$$\mathcal{F}_N = \left\{ \sum_{k=0}^{N} (a_k \cos(k\pi x) + b_k \sin(k\pi x)) : a_k, b_k \in \mathcal{R} \right\} .$$

The functoid $\mathcal{F}_N$ is a screen of $L_2(-1,1)$. Let $f \in L_2(-1,1)$ have a Fourie series of the form

$$f(x) = \sum_{k=0}^{\infty} (a_k \cos(k\pi x) + b_k \sin(k\pi x))$$

The mapping $\rho_N : L_2(-1,1) \mapsto \mathcal{F}_N$ defined by

$$\rho_N(f) = \sum_{k=0}^{N} (a_k \cos(k\pi x) + b_k \sin(k\pi x))$$

is a rounding from $L_2(-1,1)$ into $\mathcal{F}_N$. The arithmetical operations and integration in $\mathcal{F}_N$ are discussed in the preliminaries (see section 2.6.2).

The interval Fourier functoid is

$$\mathcal{IF}_N = \left\{ \sum_{k=0}^{N} (A_k \cos(k\pi x) + B_k \sin(k\pi x)) : A_k, B_k \in \mathcal{IR} \right\} .$$

Using the general approach described in section 2.6.1 the interval round of f is defined by

$$I\rho(f) = a_0 + [-1,1] \sum_{k=N+1}^{\infty} \sqrt{a_k^2 + b_k^2} + \sum_{k=1}^{N} (a_k \cos(k\pi x) + b_k \sin(k\pi x)) .$$

This definition is really applicable only when the Fourier series of f is finite. Since the operations in $\mathcal{IF}$ considered in the introduction require rounding only from $\mathcal{IF}_{2N}$ onto $\mathcal{IF}_N$, this definition (extended for interval functions) was applied. However, in general, we have at least two problems with that definition:

- Is the series $\sum_{k=N+1}^{\infty} \sqrt{a_k^2 + b_k^2}$ convergent?
- If the series $\sum_{k=N+1}^{\infty} \sqrt{a_k^2 + b_k^2}$ is convergent, how can this sum be obtained constructively?

In order to define an interval rounding $I\rho_N$ for a function $f \in L_2(-1,1)$ we need to have an estimate of the form

$$\max_{x\in[-1,1]} |f(x) - \rho_N(f)(x)| \leq \sigma_N(f) \to 0 \text{ when } N \to \infty .$$

This implies that $\rho_N(f)$ converges uniformly to f. In general, $\rho_N(f)$ converges to $f \in L_2(-1,1)$ only in the L_2 norm. Therefore $\mathcal{IF}$ can not be an interval screen of $L_2(-1,1)$ but it can be an interval screen of a subset of $L_2(-1,1)$ consisting of functions for which the Fourier series converges uniformly.

The Sobolev space $H^m(a,b)$ is defined by

$$H^m(a,b) = \left\{ f \in L_2(a,b) : \frac{d^k f}{dx^k} \in L_2(a,b) \text{ for } 0 \leq k \leq m \right\}$$

where the derivatives are considered in the generalized sense of distributions [4], [31]. Due to the Sobolev Imbedding Theorems this space can also be represented as

$$H^m(a,b) = \left\{ f \in C^{m-1}[a,b] : \frac{d}{dx} f^{(m-1)} \in L_2(a,b) \right\}$$

where only the last derivative is in the sense of distributions. In the analysis of Fourier methods, the natural Sobolev spaces are those of periodic functions:

$$\begin{aligned} H^m_{per}(a,b) &= \left\{ f \in L_2(a,b) : \frac{d^k f}{dx^k} \in L_2(a,b) \text{ for } 0 \leq k \leq m \right\} \\ &= \left\{ f \in C^{m-1}[a,b] : f\text{−periodic (period } b-a) \text{ on } \mathcal{R}, \frac{d}{dx} f^{m-1} \in L_2(a,b) \right\} \end{aligned}$$

where the derivative is in the sense of periodic distribution (period $b-a$) [20]. Obviously we have

$$H^m_{per}(a,b) \subset H^m(a,b).$$

We will show that $\mathcal{IF}$ is an interval screen of $H^1_{per}(-1,1)$, i.e. we consider $\mathcal{M} = H^1_{per}(-1,1)$. Let $f \in \mathcal{M}$. Then

$$f(x) = \sum_{k=0}^{\infty}(a_k \cos(k\pi x) + b_k \sin(k\pi x))$$

for every $x \in \mathcal{R}$ and

$$\int_{-1}^{1} (f'(x))^2\, dx = \sum_{k=0}^{\infty} k^2\pi^2(a_k^2 + b_k^2)\,.$$

We can estimate $|f(x) - \rho_N(f)(x)|$ as follows

$$\begin{aligned}
|f(x) - \rho_N(f)(x)| &= \left|\sum_{k=N+1}^{\infty}(a_k \cos(k\pi x) + b_k \sin(k\pi x))\right| \\
&\le \sum_{k=N+1}^{\infty} |a_k \cos(k\pi x) + b_k \sin(k\pi x)| \\
&\le \sum_{k=N+1}^{\infty} \sqrt{a_k^2 + b_k^2} = \sum_{k=N+1}^{\infty} \frac{1}{k}\sqrt{k^2a_k^2 + k^2b_k^2} \\
&\le \left(\sum_{k=N+1}^{\infty} \frac{1}{k^2}\right)^{\frac{1}{2}} \left(\sum_{k=N+1}^{\infty} (k^2a_k^2 + k^2b_k^2)\right)^{\frac{1}{2}} \\
&\le \frac{1}{\sqrt{N}} \left(\frac{1}{\pi^2}\int_{-1}^{1} (f'(x))^2\, dx - \sum_{k=0}^{N}(k^2a_k^2 + k^2b_k^2)\right)^{\frac{1}{2}}\,.
\end{aligned}$$

Therefore

$$\max_{x\in[-1,1]} |f(x) - \rho_N(f)(x)| \le \varepsilon_N(f) = \frac{1}{\sqrt{N}} \left(\frac{1}{\pi^2}\int_{-1}^{1} (f'(x))^2\, dx - \sum_{k=0}^{N}(k^2a_k^2 + k^2b_k^2)\right)^{\frac{1}{2}} \tag{4.22}$$

The estimate ε_N has two important characteristics:

1. It is easily computable. Since $a_k, b_k, k = 0, 1, \ldots, N$ are already computed with the computation of $\rho_N(f)$ we only need $\displaystyle\int_{-1}^{1} (f'(x))^2\, dx$.

2. The order of convergence of $\varepsilon_N(f)$ towards zero adjusts automatically according to the properties of f. For example, if f has a jth derivative in $L_2(-1,1)$ then $\varepsilon_N(f) = o\left(N^{-j+\frac{1}{2}}\right)$.

If the Fourier series of $f \in \mathcal{M}$ is finite, we do not need to use ε_N. Since the arithmetical operations considered in the introduction involve only rounding of functions in $\mathcal{IF}_{2N}$, in order to have uniformity, we define $I\rho_N : \mathcal{M} \mapsto \mathcal{IF}_N$ in the following way:

$$I\rho_N(f) = a_0 + [-1,1]\sigma_N(f) + \sum_{k=1}^{N}(a_k \cos(k\pi x) + b_k \sin(k\pi x))$$

where $\sigma_N(f) = \sum_{N+1}^{2N} \sqrt{a_k^2 + b_k^2} + \varepsilon_{2N}(f)$. The definition of $I\rho_N$ can be extended over $P\mathcal{M}$ in a natural way. Let $F \in P\mathcal{M}$.

$$I\rho_N(F) = A_0 + [-1,1]\sigma_N(F) + \sum_{k=1}^{N}(A_k \cos(k\pi x) + B_k \sin(k\pi x))$$

where

$$\begin{aligned}
&A_0 = [\{a_0(f) : f \in F\}], \quad B_0 = 0\,, \\
&A_k = [\{a_k(f) : f \in F\}], \quad B_k = [\{b_k(f) : f \in F\}], \quad k = 1,2,\ldots\,, \\
&\sigma_N(F) = [\{\sigma_N(f) : f \in F\}]\,.
\end{aligned}$$

When $F = \sum_{k=0}^{2N}(A_k \cos(k\pi x) + B_k \sin(k\pi x) \in \mathcal{F}_{2N}$ we have

$$I\rho_N(F) = A_0 + [-1,1]\sigma_N(F) + \sum_{k=1}^{N}(A_k \cos(k\pi x) + B_k \sin(k\pi x)) \tag{4.23}$$

where

$$\sigma(F) = \sum_{k=N+1}^{2N} \sqrt{A_k^2 + B_k^2}\,.$$

The interval rounding $I\rho_N$ also defines directed rounding $\underline{\rho}_N$ and $\overline{\rho}_N$ in $\mathcal{M}$. We have

$$I\rho_N(f) = [\underline{\rho}_N(f), \overline{\rho}_N(f)]$$

where

$$\begin{aligned}
\underline{\rho}_N(f) &= a_0 - \sigma_N(f) + \sum_{k=1}^{N}(a_k \cos(k\pi x) + b_k \sin(k\pi x))\,, \\
\overline{\rho}_N(f) &= a_0 + \sigma_N(f) + \sum_{k=1}^{N}(a_k \cos(k\pi x) + b_k \sin(k\pi x))\,.
\end{aligned}$$

Using directed roundings the operations in $\mathbb{F}$ can also be defined with rounding to the left or to the right. In analogy with the advanced computer arithmetic we will denote the operations in $\mathbb{F}$ with rounding to the left by

$$\triangledown\,,\quad \circ \in \{+,-,\times,/\}\,,\quad \nabla$$

and the operations in $\mathbb{F}$ with rounding to the right by

$$\triangle\,,\quad \circ \in \{+,-,\times,/\}\,,\quad \Delta\,.$$

4.4 Some Aspects of the Numerical Implementation of the Method

4.4.1 Formulation of the Pair of Problems (4.12), (4.13).

One of the inconveniences in using the interval Fourier functoid is that the upper and the lower bounds of

$$\sum_{k=0}^{N} \left(A_k \cos(k\pi x) + B_k \sin(k\pi x)\right) \in I\mathbb{F}$$

are not functions in $\mathbb{F}$. In other words $I\mathbb{F}$ is not the interval space over $\mathbb{F}$. Let $\underline{f}, \overline{f} \in \mathcal{M}$, $\underline{f} \leq \overline{f}$ and $F = [\underline{f}, \overline{f}]$. Obviously, $F \in P\mathcal{M}$. We have

$$\begin{aligned}
\underline{\rho}_N(\underline{f}) &= a_0 - \sigma_N(\underline{f}) + \sum_{k=1}^{N} \left(\underline{a}_k \cos(k\pi x) + \underline{b}_k \sin(k\pi x)\right) \\
\overline{\rho}_N(\overline{f}) &= a_0 + \sigma_N(\overline{f}) + \sum_{k=1}^{N} \left(\overline{a}_k \cos(k\pi x) + \overline{b}_k \sin(k\pi x)\right)
\end{aligned}$$

Since the inequalities $\underline{a}_k \leq \overline{a}_k$, $\underline{b}_k \leq \overline{b}_k$ are not necessarily true, the interval function $[\underline{\rho}_N(\underline{f}), \overline{\rho}_N(\overline{f})]$ is not, in general, an element of $I\mathbb{F}$. It is easy to see that

$$\begin{aligned}
[\underline{\rho}_N(\underline{f}), \overline{\rho}_N(\overline{f})] \subset\ & (\underline{a}_0 - \sigma_N(\underline{f})) \vee (\overline{a}_0 + \sigma_N(\overline{f})) \\
& + \sum_{k=1}^{N} \left((\underline{a}_k \vee \overline{a}_k)\cos(k\pi x) + (\underline{b}_k \vee \overline{b}_k)\sin(k\pi x)\right) \subset I\rho_N(F)\,.
\end{aligned}$$

We have

$$\begin{aligned}
w([\underline{\rho}_N(\underline{f}), \overline{\rho}_N(\overline{f})]) &= \overline{\rho}_N(\overline{f}) - \underline{\rho}_N(\underline{f}) \\
&= \overline{a}_0 - \underline{a}_0 + \sigma_N(\underline{f}) + \sigma_N(\overline{f}) + \sum_{k=1}^{N} \left((\overline{a}_k - \underline{a}_k)\cos(k\pi x) + (\overline{b}_k - \underline{b}_k)\sin(k\pi x)\right)
\end{aligned}$$

while

$$w(I\rho_N(F)) \geq |\overline{a}_0 - \underline{a}_0 + \sigma_N(\underline{f}) + \sigma_N(\overline{f})| + \sum_{k=1}^{N} \left(|\overline{a}_k - \underline{a}_k| |\cos(k\pi x)| + |\overline{b}_k - \underline{b}_k| |\sin(k\pi x)| \right) .$$

It is obvious that there is, in general, a significant difference between the width of $[\underline{\rho}_N(\underline{f}), \overline{\rho}_N(\overline{f})]$ and the width of $I\rho_N(F)$, which increases when N increases.

For that reason we will not apply the interval Fourier functoid $\mathcal{IF}$ for approximation of the solution $u(0, G; x, t)$ of problem (4.4)-(4.6). Instead, we will use the directed roundings $\underline{\rho}_N$ and $\overline{\rho}_N$ to obtain lower and upper bounds $\underline{s}(h, N)$, $\overline{s}_N(h, N)$ such that for every $t \in [0, \overline{t}]$

$$\underline{s}(h, N; ., t) \in \mathcal{F} \ , \ \ \overline{s}_N(h, N; ., t) \in \mathcal{F} \ .$$

Remark: Strictly speaking the coefficients of $\underline{\rho}_N(\underline{f})$ and $\overline{\rho}_N(\overline{f})$ will also be intervals when they are represented on a computer, which makes $\underline{\rho}_N(\underline{f})$ and $\overline{\rho}_N(\overline{f})$ elements of $\mathcal{IF}$. However in order to simplify the presentation, we will consider the coefficients as real numbers and use point (noninterval) notations, assuming that in the practical computations they will be represented as narrow intervals according to the advanced computer arithmetic discussed in section 2.2.

The formulation of problem (4.4)-(4.6) on every interval $[t_j, t_{j+1}]$ as two problems, (4.12) and (4.13), facilitates the above approach. We need to calculate a lower bound for the solution of (4.12) and an upper bound for the solution of (4.13).

Since we will use computations in the Fourier functoid, the 'suitable form' of $\underset{\sim}{g}_{01}$, $\tilde{g}_{01}$ and $\underset{\sim}{g}_{02}$, $\tilde{g}_{02}$ are lower and upper Fourier approximations of $\underline{g}_1$, $\overline{g}_1$ and $\underline{g}_2$, $\overline{g}_2$, respectively, i.e.

$$\underset{\sim}{g}_{01} = \underline{\rho}_N(\underline{g}_1) \ , \ \tilde{g}_{01} = \overline{\rho}_N(\overline{g}_1) \ , \ \ \underset{\sim}{g}_{02} = \underline{\rho}_N(\underline{g}_2), \ \ \tilde{g}_{02} = \overline{\rho}_N(\overline{g}_2) \ .$$

Since the bounds $\underline{s}(h, N)$, $\overline{s}(h, N)$ are obtained as Fourier series about x and for every $t \in [0, \overline{t}]$

$$\underline{s}(h, N; x, t), \overline{s}(h, N; x, t) \in \mathcal{F}$$

we also have

$$\underset{\sim}{g}_{j1} = \underline{s}(h, N; x, t_j) \in \mathcal{F} \ , \ \tilde{g}_{j1} = \overline{s}(h, N; x, t_j) \in \mathcal{F} \ ,$$

$$\underset{\sim}{g}_{j2} = \underline{s}_t(h, N; x, t_j) \in \mathcal{F} \ , \ \tilde{g}_{j2} = \overline{s}_t(h, N; x, t_j) \in \mathcal{F}$$

for $j \geq 0$. Therefore for every $j = 0, 1, \ldots, \overline{j} - 1$ functions $\underset{\sim}{g}_{j1}$, $\tilde{g}_{j1}$, $\underset{\sim}{g}_{j2}$, $\tilde{g}_{j2}$ are represented in the following form:

$$\underset{\sim}{g}_{j1}(x) \ = \ \underline{\alpha}_{10} + \sum_{k=1}^{N} \left(\underline{\alpha}_{1k} \cos(k\pi x) + \underline{\beta}_{1k} \sin(k\pi x) \right) \ ,$$

$$
\begin{aligned}
\underset{\sim}{g}_{j2}(x) &= \underline{\alpha}_{20} + \sum_{k=1}^{N} \left(\underline{\alpha}_{2k} \cos(k\pi x) + \underline{\beta}_{2k} \sin(k\pi x) \right) , \\
\tilde{g}_{j1}(x) &= \overline{\alpha}_{10} + \sum_{k=1}^{N} \left(\overline{\alpha}_{1k} \cos(k\pi x) + \overline{\beta}_{1k} \sin(k\pi x) \right) , \\
\tilde{g}_{j2}(x) &= \overline{\alpha}_{20} + \sum_{k=1}^{N} \left(\overline{\alpha}_{2k} \cos(k\pi x) + \overline{\beta}_{2k} \sin(k\pi x) \right) .
\end{aligned}
$$

Functions $\underset{\sim}{f}(x,t,u(x,t))$ and $\tilde{f}(x,t,u(x,t))$ are represented as Fourier series of x with coefficients that are polynomials of t. Using directed Fourier and Taylor roundings $\underset{\sim}{f}(x,t,u(x,t))$ and $\tilde{f}(x,t,u(x,t))$ can be described as follows:

Let

$$
I\rho_N(f(.,.,u))(x,t) = a_0(t) + [-1,1]\sigma_N(f;t) + \sum_{k=0}^{N} \left(a_k(t)\cos(k\pi x) + b_k(t)\sin(k\pi x) \right)
$$

and let

$$
|a_k(t) - \tau_m(a_k)(t)| \leq \breve{\sigma}_m(a_k) , \quad |b_k(t) - \tau_m(b_k)(t)| \leq \breve{\sigma}_m(b_k) , \quad k = 1, \ldots, N .
$$

Then

$$
\begin{aligned}
\underset{\sim}{f}(x,t,u(x,t)) &= \underline{\tau}_m(a_0 - \sigma_N(f)) - \sum_{k=1}^{N} \sqrt{\breve{\sigma}_m^2(a_k) + \breve{\sigma}_m^2(b_k)} \\
&\quad + \sum_{k=0}^{N} \left(\tau_m(a_k)(t)\cos(k\pi x) + \tau_m(b_k)(t)\sin(k\pi x) \right) \\
\tilde{f}(x,t,u(x,t)) &= \overline{\tau}_m(a_0 + \sigma_N(f)) + \sum_{k=1}^{N} \sqrt{\breve{\sigma}_m^2(a_k) + \breve{\sigma}_m^2(b_k)} \\
&\quad + \sum_{k=0}^{N} \left(\tau_m(a_k)(t)\cos(k\pi x) + \tau_m(b_k)(t)\sin(k\pi x) \right) .
\end{aligned}
\tag{4.24}
$$

This is possible under the additional assumption that f has $m+1$ bounded derivatives about t and u. All numerical experiments are performed with the same m ($m = 3$). That is why m is not included in the list of parameters of the method.

For computational reasons, in the interval $[t_j, t_{j+1}]$, it is more convenient to present the coefficients in (4.24) in the form of polynomials of $\Delta = t - t_j$ rather then t. More precisely, functions $\underset{\sim}{f}(x,t,\underset{\sim}{y}^{(r)}(x,t))$ and $\tilde{f}(x,t,\tilde{u}^{(r)}(x,t))$ in (4.19) and (4.20) respectively

are represented as follows:

$$\underline{f}(x,t,\underline{y}^{(r)}(x,t)) = \sum_{q=0}^{m} \underline{a}_{0q}\frac{\Delta^q}{q!} + \sum_{k=1}^{N}\sum_{q=0}^{m}\left(\underline{a}_{kq}\frac{\Delta^q}{q!}\cos(k\pi x) + \underline{b}_{kq}\frac{\Delta^q}{q!}\sin(k\pi x)\right) ,$$

$$\tilde{f}(x,t,\tilde{u}^{(r)}(x,t)) = \sum_{q=0}^{m} \overline{a}_{0q}\frac{\Delta^q}{q!} + \sum_{k=1}^{N}\sum_{q=0}^{m}\left(\overline{a}_{kq}\frac{\Delta^q}{q!}\cos(k\pi x) + \overline{b}_{kq}\frac{\Delta^q}{q!}\sin(k\pi x)\right) .$$

4.4.2 Implementation of the Iterative Procedure

The implementation of the iterative procedure defined by (4.19) and (4.20) requires solving problems of the form

$$\begin{aligned} u_{tt}(x,t) - u_{xx}(x,t) &= \psi(x,\Delta) \\ u(x,t_j) = g_{j1}(x)\ ,\quad u_t(x,t_j) &= g_{j2}(x) \end{aligned} \tag{4.25}$$

where $\Delta = t - t_j$ and functions ψ, g_{j1}, g_{j2} are of the form

$$g_{j1}(x) = \alpha_{10} + \sum_{k=1}^{N}\left(\alpha_{1k}\cos(k\pi x) + \beta_{1k}\sin(k\pi x)\right) , \tag{4.26}$$

$$g_{j2}(x) = \alpha_{20} + \sum_{k=1}^{N}\left(\alpha_{2k}\cos(k\pi x) + \beta_{2k}\sin(k\pi x)\right) , \tag{4.27}$$

$$\psi(x,\Delta) = \sum_{q=0}^{m} a_{0q}\frac{\Delta^q}{q!} + \sum_{k=1}^{N}\sum_{q=0}^{m}\left(a_{kq}\frac{\Delta^q}{q!}\cos(k\pi x) + b_{kq}\frac{\Delta^q}{q!}\sin(k\pi x)\right) . \tag{4.28}$$

Let $\Gamma(x,\Delta,t)$ be the triangle with vertices $(x,t+\Delta)$, $(x-\Delta,t)$ and $(x+\Delta,t)$. Using Green's theorem we have

$$\begin{aligned} \iint_{\Gamma(x,\Delta,t_j)} \psi(y,\theta))\mathrm{d}y\mathrm{d}\theta &= \iint_{\Gamma(x,\Delta,t_j)} (u_{tt}(y,\theta) - u_{xx}(y,\theta))\mathrm{d}y\mathrm{d}\theta \\ &= \oint_{\partial\Gamma(x,\Delta,t_j)} (-u_x(y,\theta)\mathrm{d}\theta - u_t(y,\theta)\mathrm{d}y \\ &= 2u(x,t_j+\Delta) - u(x-\Delta,0) - u(x+\Delta,0) - \int_{x-\Delta}^{x+\Delta} u_t(y,0)\mathrm{d}y \\ &= 2u(x,t) - g_{j1}(x-\Delta) - g_{j1}(x+\Delta) - \int_{x-\Delta}^{x+\Delta} g_{j2}(y)\mathrm{d}y\ . \end{aligned}$$

Therefore, we have

$$u(x,t) = \frac{1}{2}\iint_{\Gamma(x,\Delta,t_j)} \psi(y,\theta)\mathrm{d}y\mathrm{d}\theta + \phi(x,\Delta) \tag{4.29}$$

where

$$\phi(x,\Delta) = \frac{1}{2}\left(g_{j1}(x+\Delta) + g_{j1}(x-\Delta) + \int_{x-\Delta}^{x+\Delta} g_{j2}(y)\mathrm{d}y\right)$$

and $\Delta = t - t_j$.

Using the fact that g_{j1} and g_{j2} have the form (4.26) and (4.27), the function $\phi(x,\Delta)$ can be simplified in the following way:

$$\begin{aligned}
&\frac{1}{2}(g_{j1}(x+\Delta) + g_{j1}(x-\Delta)) = \alpha_{10} \\
&\quad + \frac{1}{2}\sum_{k=1}^{N}(\alpha_{1k}(\cos k\pi(x+\Delta) + \cos k\pi(x-\Delta)) + \beta_{1k}(\sin k\pi(x+\Delta) + \sin k\pi(x-\Delta))) \\
&\quad = \alpha_{10} + \sum_{k=1}^{N}(\alpha_{1k}\cos(k\pi x)\cos(k\pi\Delta) + \beta_{1k}\sin(k\pi x)\cos(k\pi\Delta)) \\
&\quad = \alpha_{10} + \sum_{k=1}^{N}\cos(k\pi\Delta)(\alpha_{1k}\cos(k\pi x) + \beta_{1k}\sin(k\pi x)) \ ,
\end{aligned}$$

$$\begin{aligned}
&\frac{1}{2}\int_{x-\Delta}^{x+\Delta} g_{j2}(y)\mathrm{d}y = \alpha_{20}\Delta \\
&\quad + \frac{1}{2}\sum_{k=1}^{N}\left(\frac{\alpha_{2k}}{k\pi}(\sin k\pi(x+\Delta) - \sin k\pi(x-\Delta)) - \frac{\beta_{2k}}{k\pi}(\cos k\pi(x+\Delta) - \cos k\pi(x-\Delta))\right) \\
&\quad = \alpha_{20}\Delta + \sum_{k=1}^{N}\left(\frac{\alpha_{2k}}{k\pi}\cos(k\pi x)\sin(k\pi\Delta) + \frac{\beta_{2k}}{k\pi}\sin(k\pi x)\sin(k\pi\Delta)\right) \\
&\quad = \alpha_{20}\Delta + \sum_{k=1}^{N}\sin(k\pi\Delta)\left(\frac{\alpha_{2k}}{k\pi}\cos(k\pi x) + \frac{\beta_{2k}}{k\pi}\sin(k\pi x)\right) \ ,
\end{aligned}$$

$$\begin{aligned}
\phi(x,\Delta) \;=\; & \alpha_{10} + \alpha_{20}\Delta + \sum_{k=1}^{N}\left(\left(\alpha_{1k}\cos(k\pi\Delta) + \frac{\alpha_{2k}}{k\pi}\sin(k\pi\Delta)\right)\cos(k\pi x)\right. \\
& \left. + \left(\beta_{1k}\cos(k\pi\Delta) + \frac{\beta_{2k}}{k\pi}\sin(k\pi\Delta)\right)\sin(k\pi x)\right) \ .
\end{aligned}$$

The coefficients in the above Fourier sum are not polynomials. Because

$$\begin{aligned}
\cos(k\pi\Delta) &\in \sum_{q=0}^{m'}\frac{(-1)^q(k\pi\Delta)^{2q}}{(2q)!} + [-1,1]\frac{(k\pi\Delta)^{2m'+2}}{(2m'+2)!} \ , \\
\sin(k\pi\Delta) &\in \sum_{q=1}^{m''}\frac{(-1)^{q-1}(k\pi\Delta)^{2q-1}}{(2q-1)!} + [-1,1]\frac{(k\pi\Delta)^{2m''+1}}{(2m''+1)!} \ ,
\end{aligned}$$

where m' is the largest integer not greater then $\frac{m}{2}$ and m'' is the largest integer not greater then $\frac{m+1}{2}$, we obtain the following inclusion

$$\begin{aligned}\phi(x,\Delta) \;\in\; & \alpha_{10} + [-1,1]\left(\frac{(k\pi h)^{2m'+2}}{(2m'+2)!}\sum_{k=1}^{N}\sqrt{\alpha_{1k}^2+\beta_{1k}^2} + \frac{(k\pi h)^{2m''+1}}{(2m''+1)!}\sum_{k=1}^{N}\frac{\sqrt{\alpha_{2k}^2+\beta_{2k}^2}}{k\pi}\right)\\ + \;& \alpha_{20}\Delta + \sum_{k=1}^{N}\left(\left(\alpha_{1k}\sum_{q=0}^{m'}\frac{(k\pi\Delta)^{2q}}{(2q)!} + \frac{\alpha_{2k}}{k\pi}\sum_{q=1}^{m''}\frac{(k\pi\Delta)^{2q-1}}{(2q-1)!}\right)\cos(k\pi x)\right.\\ & \left. + \left(\beta_{1k}\sum_{q=0}^{m'}\frac{(k\pi\Delta)^{2q}}{(2q)!} + \frac{\beta_{2k}}{k\pi}\sum_{q=1}^{m''}\frac{(k\pi\Delta)^{2q-1}}{(2q-1)!}\right)\sin(k\pi x)\right)\ .\end{aligned}$$

In order to obtain a lower bound for the solution of (4.25) we use in (4.29) the lower bound of the above interval function, and in order to obtain an upper bound for the solution of (4.25) we use in (4.29) the upper bound of this interval function.

Since ψ is a sum of the form (4.28) the double integral in (4.29) is a sum of integrals of the form

$$\iint_{\Gamma(x,\Delta,t_j)}\frac{\theta^q}{q!}dyd\theta\ ,\ \iint_{\Gamma(x,\Delta,t_j)}\frac{\theta^q}{q!}\cos(k\pi y)dyd\theta\ ,\ \iint_{\Gamma(x,\Delta,t_j)}\frac{\theta^q}{q!}\sin(k\pi y)dyd\theta\ .$$

For the evaluation of the above integrals we will derive explicit formulae. We have

$$\begin{aligned}\iint_{\Gamma(x,\Delta,t_j)}\frac{\theta^q}{q!}\mathrm{d}y\mathrm{d}\theta &= \int_0^{\Delta}\int_{x-\Delta+\theta}^{x+\Delta-\theta}\frac{\theta^q}{q!}\mathrm{d}y\mathrm{d}\theta\\ &= 2\int_0^{\Delta}\frac{\theta^q}{q!}(\Delta-\theta)\mathrm{d}\theta\\ &= 2\left[\frac{\theta^{q+1}}{(q+1)!}\Delta - \frac{\theta^{q+2}}{(q+2)!}(q+1)\right]_0^{\Delta}\\ &= 2\left(\frac{\Delta^{q+2}}{(q+2)!}(q+2) - \frac{\Delta^{q+2}}{(q+2)!}(q+1)\right)\\ &= 2\frac{\Delta^{q+2}}{(q+2)!}\ ,\end{aligned}$$

$$\begin{aligned}\iint_{\Gamma(x,\Delta,t_j)}\frac{\theta^q}{q!}e^{ik\pi y}\mathrm{d}y\mathrm{d}\theta &= \int_0^{\Delta}\int_{x-\Delta+\theta}^{x+\Delta-\theta}\frac{\theta^q}{q!}e^{ik\pi y}\mathrm{d}y\mathrm{d}\theta\\ &= \int_0^{\Delta}\frac{\theta^q}{q!}\left(\frac{e^{ik\pi(x+\Delta-\theta)}}{ik\pi} - \frac{e^{ik\pi(x-\Delta+\theta)}}{ik\pi}\right)\mathrm{d}\theta\end{aligned}$$

$$
\begin{aligned}
&= \left[-\sum_{l=0}^{q} \frac{\theta^{q-l}}{(q-l)!} \left(\frac{e^{ik\pi(x+\Delta-\theta)}}{(ik\pi)^{l+2}} + (-1)^l \frac{e^{ik\pi(x-\Delta+\theta)}}{(ik\pi)^{l+2}} \right) \right]_0^{\Delta} \\
&= \frac{e^{ik\pi(x+\Delta)}}{(ik\pi)^{q+2}} + \frac{e^{ik\pi(x-\Delta)}}{(-ik\pi)^{q+2}} - \sum_{l=0}^{q} \frac{(1+(-1)^l)}{(ik\pi)^{l+2}} \frac{\Delta^{q-l}}{(q-l)!} e^{ik\pi x} \\
&= \frac{e^{ik\pi x}}{(ik\pi)^{q+2}} \left(e^{ik\pi\Delta} + (-1)^q e^{-ik\pi\Delta} - \sum_{l=0}^{q} \left(\frac{(ik\pi\Delta)^l}{l!} + (-1)^q \frac{(-ik\pi\Delta)^l}{l!} \right) \right) \\
&= \begin{cases} 2\dfrac{e^{ik\pi x}}{(k\pi)^{q+2}} (-1)^{\frac{q}{2}+1} \left(\cos(k\pi\Delta) - \displaystyle\sum_{l=0}^{\frac{q}{2}} (-1)^l \frac{(k\pi\Delta)^{2l}}{(2l)!} \right) & , \quad q\text{-even} \\ 2\dfrac{e^{ik\pi x}}{(k\pi)^{q+2}} (-1)^{\frac{q+1}{2}} \left(\sin(k\pi\Delta) - \displaystyle\sum_{l=1}^{\frac{q+1}{2}} (-1)^{l-1} \frac{(k\pi\Delta)^{2l-1}}{(2l-1)!} \right) & , \quad q\text{-odd} \end{cases} \\
&= \begin{cases} 2 \displaystyle\sum_{l=\frac{q}{2}+1}^{\infty} (-1)^{l-\frac{q}{2}+1} (k\pi)^{2l-q-2} \frac{\Delta^{2l}}{(2l)!} e^{ik\pi x} & , \quad q\text{-even} \\ 2 \displaystyle\sum_{l=\frac{q+3}{2}}^{\infty} (-1)^{l-\frac{q-1}{2}} (k\pi)^{2l-q-3} \frac{\Delta^{2l-1}}{(2l-1)!} e^{ik\pi x} & , \quad q\text{-odd} \end{cases}
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\iint_{\Gamma(x,\Delta,t_j)} \frac{\theta^q}{q!} \cos(k\pi y) dy d\theta &= Re \left(\iint_{\Gamma(x,\Delta,t_j)} \frac{\theta^q}{q!} e^{ik\pi y} \mathrm{d}y \mathrm{d}\theta \right) \\
&= \begin{cases} -2 \displaystyle\sum_{l=\frac{q}{2}+1}^{\infty} (-1)^{l-\frac{q}{2}} (k\pi)^{2l-q-2} \frac{\Delta^{2l}}{(2l)!} \cos(k\pi x) & , \quad q \text{ - even} \\ 2 \displaystyle\sum_{l=\frac{q+3}{2}}^{\infty} (-1)^{l-\frac{q+1}{2}} (k\pi)^{2l-q-3} \frac{\Delta^{2l-1}}{(2l-1)!} \sin(k\pi x) & , \quad q \text{ - odd} \end{cases} \\
\iint_{\Gamma(x,\Delta,t_j)} \frac{\theta^q}{q!} \sin(k\pi y) dy d\theta &= Im \left(\iint_{\Gamma(x,\Delta,t_j)} \frac{\theta^q}{q!} e^{ik\pi y} \mathrm{d}y \mathrm{d}\theta \right) \\
&= \begin{cases} -2 \displaystyle\sum_{l=\frac{q}{2}+1}^{\infty} (-1)^{l-\frac{q}{2}} (k\pi)^{2l-q-2} \frac{\Delta^{2l}}{(2l)!} \sin(k\pi x) & , \quad q \text{ - even} \\ 2 \displaystyle\sum_{l=\frac{q+3}{2}}^{\infty} (-1)^{l-\frac{q-1}{2}} (k\pi)^{2l-q-3} \frac{\Delta^{2l-1}}{(2l-1)!} \cos(k\pi x) & , \quad q \text{ - odd} \end{cases}
\end{aligned}
$$

The coefficients of $\cos(k\pi x)$ and $\sin(k\pi x)$ in the above expressions need to be rounded to polynomials of degree at most m. We have the following inclusions

$$
\iint_{\Gamma(x,\Delta,t_j)} \frac{\theta^q}{q!} \cos(k\pi y) dy d\theta \in \left[\underline{\iint_{\Gamma(x,\Delta,t_j)}} \frac{\theta^q}{q!} \cos(k\pi y) dy d\theta \, , \, \overline{\iint_{\Gamma(x,\Delta,t_j)}} \frac{\theta^q}{q!} \cos(k\pi y) dy d\theta \right]
$$

$$= \begin{cases} [-2,2]\dfrac{(k\pi h)^{2m'+2}}{(2m'+2)!} - 2\displaystyle\sum_{l=\frac{q}{2}+1}^{m'} (-1)^{l-\frac{q}{2}}(k\pi)^{2l-q-2}\frac{\Delta^{2l}}{(2l)!}\cos(k\pi x) & , \quad q \text{ - even} \\ [-2,2]\dfrac{(k\pi h)^{2m''+1}}{(2m''+1)!} + 2\displaystyle\sum_{l=\frac{q+3}{2}}^{m''} (-1)^{l-\frac{q+1}{2}}(k\pi)^{2l-q-3}\frac{\Delta^{2l-1}}{(2l-1)!}\sin(k\pi x) & , \quad q \text{ - odd} \end{cases} \tag{4.30}$$

$$\iint_{\Gamma(x,\Delta,t_j)} \frac{\theta^q}{q!}\sin(k\pi y)dyd\theta \in \left[\underline{\iint}_{\Gamma(x,\Delta,t_j)} \frac{\theta^q}{q!}\sin(k\pi y)dyd\theta \, , \, \overline{\iint}_{\Gamma(x,\Delta,t_j)} \frac{\theta^q}{q!}\sin(k\pi y)dyd\theta \right]$$

$$= \begin{cases} [-2,2]\dfrac{(k\pi h)^{2m'+2}}{(2m'+2)!} - 2\displaystyle\sum_{l=\frac{q}{2}+1}^{m'} (-1)^{l-\frac{q}{2}}(k\pi)^{2l-q-2}\frac{\Delta^{2l}}{(2l)!}\sin(k\pi x) & , \quad q \text{ - even} \\ [-2,2]\dfrac{(k\pi h)^{2m''+1}}{(2m''+1)!} + 2\displaystyle\sum_{l=\frac{q+3}{2}}^{m''} (-1)^{l-\frac{q-1}{2}}(k\pi)^{2l-q-3}\frac{\Delta^{2l-1}}{(2l-1)!}\cos(k\pi x) & , \quad q \text{ - odd} \end{cases}$$

The lower (upper) bound of the above interval functions is used in (4.29) for computing a lower (upper) bound for the solution of (4.25).

Then, given $\underline{u}^{(r)}$ and $\tilde{u}^{(r)}$, we have

$$\begin{aligned} \underline{u}^{(r+1)}(x,t) &= \frac{1}{2}\underline{\iint}_{\Gamma(x,\Delta,t_j)} \underline{\psi}(y,\theta)dyd\theta + \underline{\phi}(x,\Delta) \, , \\ \tilde{u}^{(r+1)}(x,t) &= \frac{1}{2}\overline{\iint}_{\Gamma(x,\Delta,t_j)} \overline{\psi}(y,\theta)dyd\theta + \overline{\phi}(x,\Delta) \, , \end{aligned} \tag{4.31}$$

where

$$\begin{aligned} \underline{\psi}(y,\theta) &= \underline{f}(y, t_j+\theta, \underline{u}^{(r)}(y,t_j+\theta)) \, , \\ \overline{\psi}(y,\theta) &= \tilde{f}(y, t_j+\theta, \tilde{u}^{(r)}(y,t_j+\theta)) \, , \\ \underline{\phi}(x,\Delta) &= \underline{\alpha}_{10} - \frac{(k\pi h)^{2m'+2}}{(2m'+2)!}\sum_{k=1}^{N}\sqrt{\underline{\alpha}_{1k}^2+\underline{\beta}_{1k}^2} - \frac{(k\pi h)^{2m''+1}}{(2m''+1)!}\sum_{k=1}^{N}\frac{\sqrt{\underline{\alpha}_{2k}^2+\underline{\beta}_{2k}^2}}{k\pi} \\ &+ \underline{\alpha}_{20}\Delta + \sum_{k=1}^{N}\left(\left(\underline{\alpha}_{1k}\sum_{q=0}^{m'}\frac{(k\pi\Delta)^{2q}}{(2q)!} + \frac{\underline{\alpha}_{2k}}{k\pi}\sum_{q=1}^{m''}\frac{(k\pi\Delta)^{2q-1}}{(2q-1)!}\right)\cos(k\pi x)\right. \\ &\left. + \left(\underline{\beta}_{1k}\sum_{q=0}^{m'}\frac{(k\pi\Delta)^{2q}}{(2q)!} + \frac{\underline{\beta}_{2k}}{k\pi}\sum_{q=1}^{m''}\frac{(k\pi\Delta)^{2q-1}}{(2q-1)!}\right)\sin(k\pi x)\right) \, , \\ \overline{\phi}(x,\Delta) &= \overline{\alpha}_{10} + \frac{(k\pi h)^{2m'+2}}{(2m'+2)!}\sum_{k=1}^{N}\sqrt{\overline{\alpha}_{1k}^2+\overline{\beta}_{1k}^2} + \frac{(k\pi h)^{2m''+1}}{(2m''+1)!}\sum_{k=1}^{N}\frac{\sqrt{\overline{\alpha}_{2k}^2+\overline{\beta}_{2k}^2}}{k\pi} \end{aligned}$$

$$+ \quad \overline{\alpha}_{20}\Delta + \sum_{k=1}^{N} \left(\left(\overline{\alpha}_{1k} \sum_{q=0}^{m'} \frac{(k\pi\Delta)^{2q}}{(2q)!} + \frac{\overline{\alpha}_{2k}}{k\pi} \sum_{q=1}^{m''} \frac{(k\pi\Delta)^{2q-1}}{(2q-1)!} \right) \cos(k\pi x) \right.$$
$$\left. + \left(\overline{\beta}_{1k} \sum_{q=0}^{m'} \frac{(k\pi\Delta)^{2q}}{(2q)!} + \frac{\overline{\beta}_{2k}}{k\pi} \sum_{q=1}^{m''} \frac{(k\pi\Delta)^{2q-1}}{(2q-1)!} \right) \sin(k\pi x) \right)$$

and m' is the largest integer not greater then $\frac{m}{2}$, m'' is the largest integer not greater then $\frac{m+1}{2}$.

Let us note that the functions $\underline{\phi}$ and $\overline{\phi}$ are computed once for the interval $[t_j, t_{j+1}]$, since they do not change during the iteration process.

An essential part of the computations in (4.31) is the evaluation of the double integrals with directed rounding which is reduced to evaluating integrals of the form

$$\iint_{\Gamma(x,\Delta,t_j)} \frac{\theta^q}{q!} \cos(k\pi y) dy d\theta \quad , \quad \iint_{\Gamma(x,\Delta,t_j)} \frac{\theta^q}{q!} \sin(k\pi y) dy d\theta \, ,$$
$$\iint_{\Gamma(x,\Delta,t_j)} \frac{\theta^q}{q!} \cos(k\pi y) dy d\theta \quad , \quad \iint_{\Gamma(x,\Delta,t_j)} \theta^q \sin(k\pi y) dy d\theta \, .$$

Suitable formulae for evaluation of the above integrals are provided by (4.30).

4.5 Accuracy

As in all validating methods, the bounds $\underline{s}(h, N)$, $\overline{s}(h, N)$ produced by this method carry within themselves an assurance of their quality.

In the case of a point (not interval) initial condition, i.e. $G(x) = g(x) \in \mathcal{R}$, $x \in \mathcal{R}$, the width

$$w(S(h, N)) = \overline{s}(h, N) - \underline{s}(h, N)$$

results only from computational errors of different sorts. If the computed bounds are too wide the parameters of the method N, h, m can be adjusted accordingly. If a certain accuracy is given in advance this adjustment can be made automatically by the computer program.

In the case of an interval initial condition we have $[u(0, G)] \subset S(h, N)$ where $w[u(0, G)] > 0$. Then the function

$$w(S(h, N)) - w([u(0, G)])$$

is an estimate for the total computational error. However, this function is unknown and although the bounds $\underline{s}(h, N)$, $\overline{s}(h, N)$ are guaranteed we don't know how close they are to the optimal enclosure $[u(0, G)]$ of the solution set $u(0, G)$. The formulation of problem

(4.4)–(4.6) as two problems (4.9) and (4.10) allows us to rectify this situation. We can simply solve problems (4.9) and (4.10) (which are problems with a point initial condition) separately obtaining lower and upper bounds for the solution of each of them and also controlling the accuracy. Then the lower bound for the solution of (4.9) and the upper bound for the solution of (4.10) are the bounds for $u(0, G)$.

An a priori estimate, although not needed to determine the accuracy of a particular numerical solution, can generally characterize the quality of the method. From the form in which the bounds are computed one can easily see that the global error is

$$o(N^{-j+\frac{1}{2}}) + O(h^{m+1}) \tag{4.32}$$

provided g_1, g_2 have j derivatives about x in $L_2(-1,1)$, f has j derivatives about x and u in $L_2(-1,1)$, f has bounded $m+1$ derivatives about t and u and $Nh < const.$

4.6 Numerical Examples

The numerical results presented in this section are produced by a Pascal-XSC program which implements the method described in the previous sections. The maximum degree m of the polynomials of t is 3 for all examples. Therefore the accuracy is

$$o(N^{-j+\frac{1}{2}}) + O(h^4)\,. \tag{4.33}$$

All examples are periodical with period 2 initial value problems for the wave equation. Graphs of solutions, enclosures and errors are plotted for $x \in [-1,1]$.

Example 4.1 We consider the periodic problem for the equation

$$u_{tt} - u_{xx} = \left(\pi^2 + \frac{2}{(t+1)^2}\right) u \tag{4.34}$$

with three different initial conditions.

A. Noninterval Initial Condition in the Fourier Functoid.

We consider equation (4.34) with an initial condition

$$u(x,0) = \sin(\pi x)\ ,\ u_t(x,0) = 2\sin(\pi x) \tag{4.35}$$

The functions prescribed as values of both $u(x,0)$ and $u_t(x,0)$ are trigonometric polynomials so that no rounding of these functions is required. The only function which needs to be rounded is the coefficient of u in the equation (4.34) and this is done in every interval $[t_j, t_{j+1}]$. The exact solution to the problem is $u = (t+1)^2 \sin(\pi x)$. On figure 4.1 where the solution and the enclosure computed with $N = 5$ and $h = 2^{-5}$ are plotted, the solution and the enclosure are visually undistinguishable. Values of the solution and the computed bounds at some points are presented in table 4.1.

Since the solution is a trigonometric polynomial about x we need a relatively small number N of spectral functions ($N = 5$) and further increase of N will not improve the accuracy. The leading term of the error is $0(h^4)$. On figure 4.2 the maximum norm in x of the error of approximation is plotted against the time variable on a logarithmic scale. The obtained numerical results are consistent with the expected rate $O(h^4)$ of convergence.

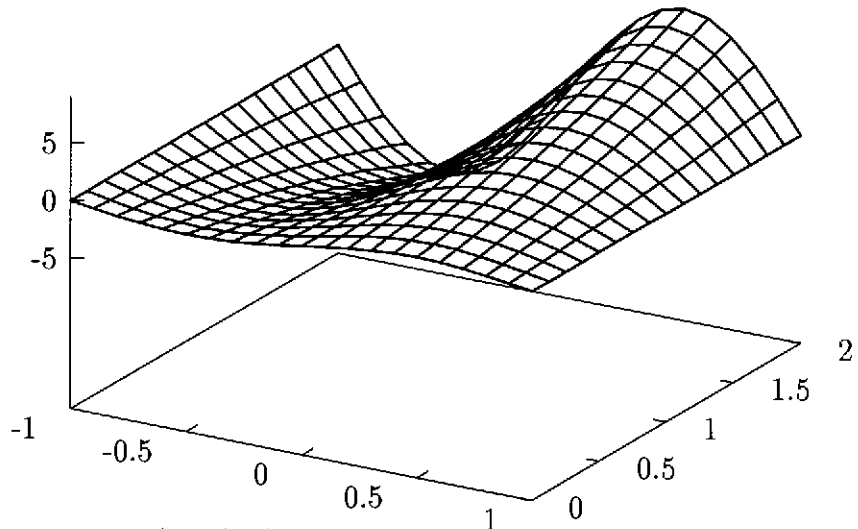


Figure 4.1: *Problem (4.34), (4.35). Exact solution and enclosure computed with* $N = 5$ *and* $h = 2^{-5}$.

$N = 5,\ h = 2^{-5}$						
	$x = 0$			$x = 0.5$		
t	solution	bounds	radius	solution	bounds	radius
0.5	0.0000	$^{+0.0001}_{-0.0001}$	2.3E-5	2.2500	2.2^{501}_{499}	2.3E-5
1.0	0.0000	$^{+0.0002}_{-0.0002}$	1.9E-4	4.0000	$^{4.0002}_{3.9988}$	1.9E-4
1.5	0.0000	$^{+0.0011}_{-0.0011}$	1.1E-3	6.2500	6.2^{511}_{490}	1.1E-3
2.0	0.0000	$^{+0.0053}_{-0.0053}$	5.3E-3	9.0000	$^{9.0052}_{8.9948}$	5.3E-3

Table 4.1: *Problem (4.34), (4.35). Values of the solution and the computed enclosures at some points.*

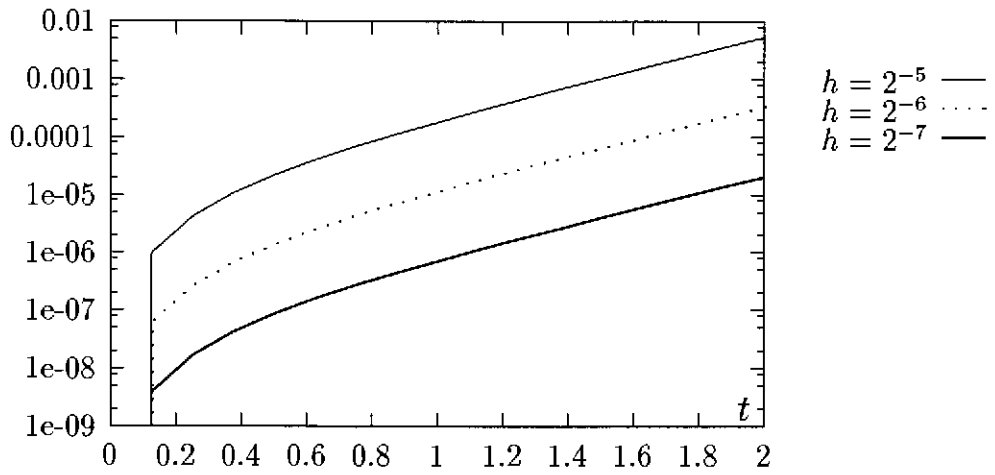


Figure 4.2: *Problem (4.34), (4.35). Maximum norm in x of the radius of the enclosures computed for various step sizes h (logarithmic scale).*

B. Noninterval Initial Condition Which Needs Rounding.

We consider equation (4.34) with an initial condition of the form

$$u(x,0) = \begin{cases} x(1-x)\,, & 0 \le x \le 1 \\ x(1+x)\,, & -1 \le x \le 0 \end{cases} \quad , \quad u_t(x,0) = 0 \tag{4.36}$$

Since the function prescribed as the value of $u(x,0)$ is not in the Fourier functoid it needs to be rounded, i.e. lower and upper bounds in the form of trigonometric polynomials are computed. Therefore a larger number of spectral functions will be required to achieve accuracy similar to the accuracy of the enclosures obtained for the solution of problem (4.34), (4.35) with only $N = 5$. Figures 4.3, 4.4 and 4.5 present the enclosures computed with $N = 10$, $N = 20$ and $N = 40$ respectively. Let us note that we may not allow $N \to \infty$ while h is fixed because $Nh < const$. In the presented numerical experiments we double N and reduce the step size h twice. Numerical values of the enclosures at some points are presented in table 4.2

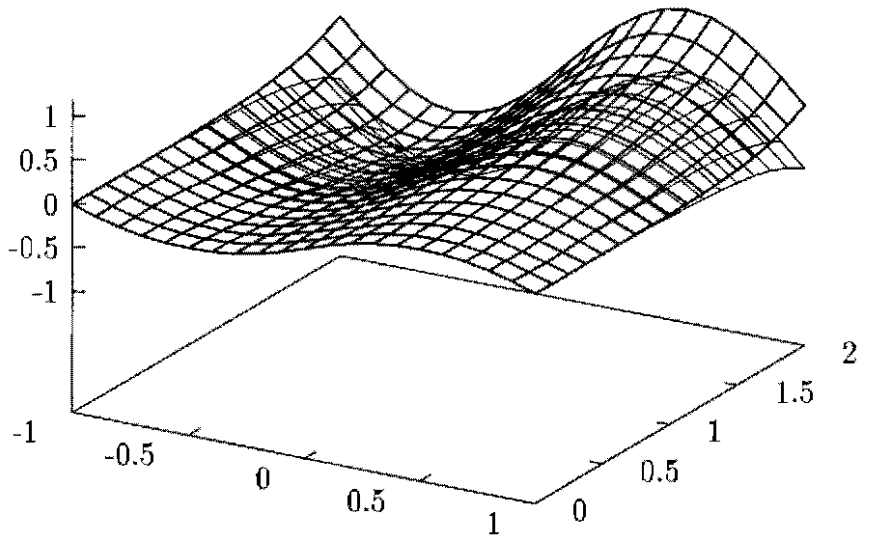


Figure 4.3: *Problem (4.34), (4.36). Enclosure computed with* $N = 10$ *and* $h = 2^{-6}$

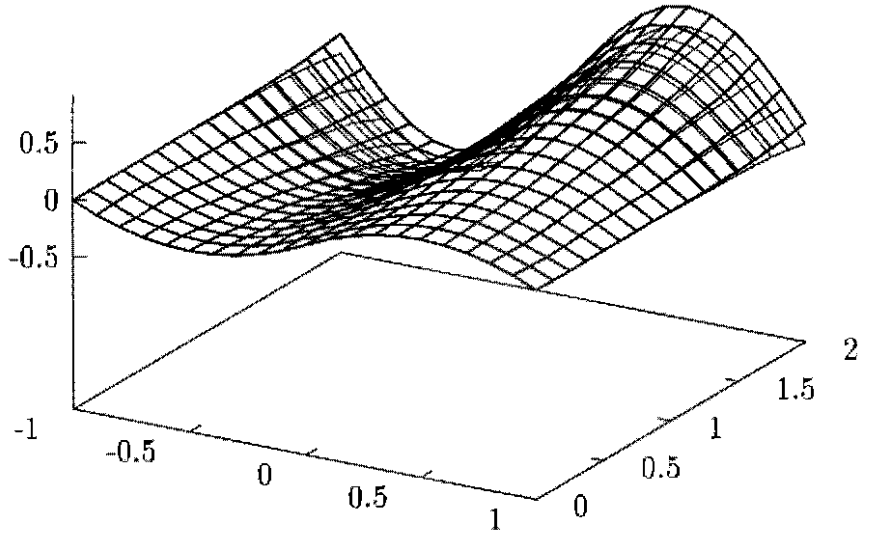


Figure 4.4: *Problem (4.34), (4.36). Enclosure computed with* $N = 20$ *and* $h = 2^{-7}$

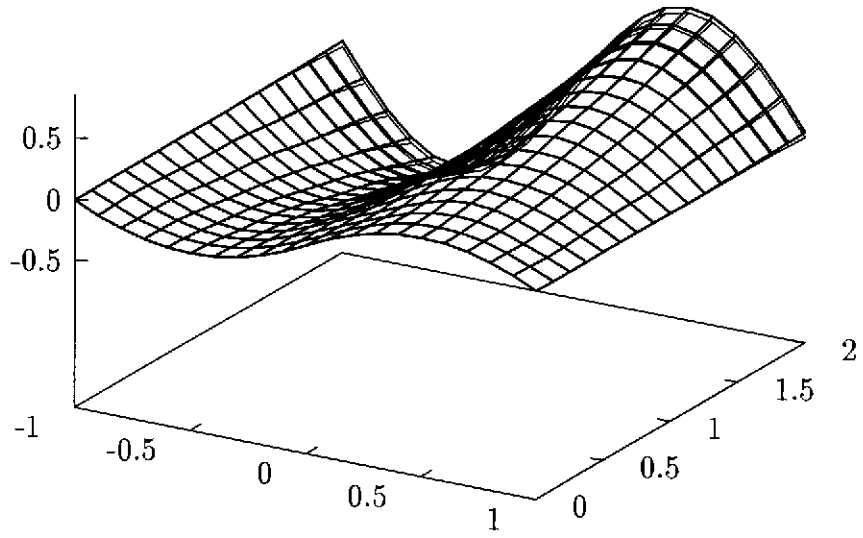


Figure 4.5: *Problem (4.34), (4.36). Enclosure computed with* $N = 40$ *and* $h = 2^{-8}$

	$N = 10, h = 2^{-6}$		$N = 20,\ h = 2^{-7}$		$N = 40,\ h = 2^{-8}$	
	$x = 0.5$		$x = 0.5$		$x = 0.5$	
t	bounds	radius	bounds	radius	bounds	radius
1.0	0.4^{5078}_{2189}	1.5E-2	0.4^{4011}_{3284}	3.7E-3	0.43^{737}_{555}	9.2E-4
2.0	$^{1.1836}_{0.4794}$	3.6E-1	$0.^{91801}_{73932}$	9.0E-2	0.8^{5105}_{0630}	2.3E-2

Table 4.2: *Problem (4.34), (4.36). Values of the computed enclosures at some points .*

Since the initial function is differentiable only twice, the estimate (4.32) gives a rate of convergence $o(N^{-\frac{3}{2}}) + O(h^4)$ where $o(N^{-\frac{3}{2}})$ is the dominating term. On figure 4.6 the maximum norm in x of the radius of the enclosures computed for various values of N and h are plotted against the time variable. The numerical results are consistent with the expected $o(N^{-\frac{3}{2}})$ rate of convergence.

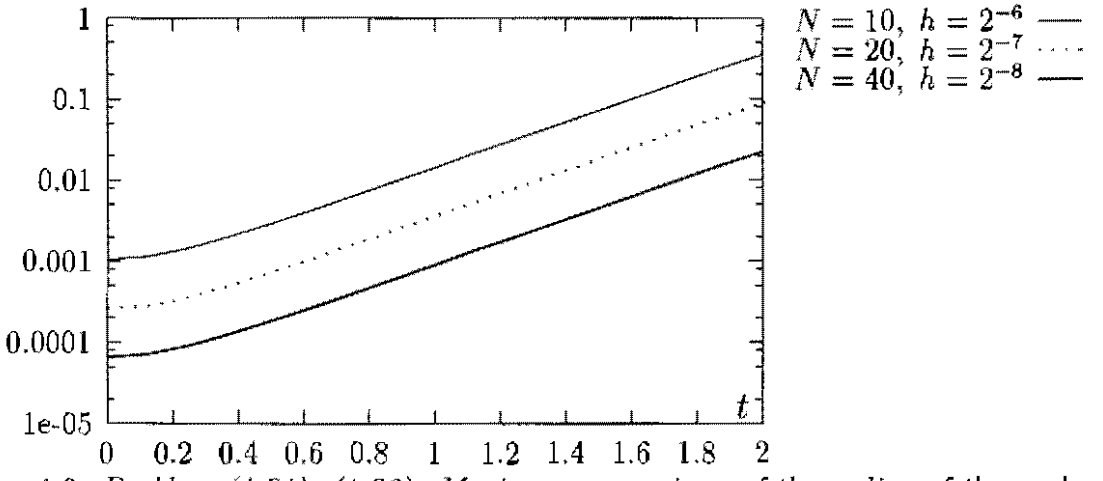


Figure 4.6: *Problem (4.34), (4.36). Maximum norm in x of the radius of the enclosures computed for various values of N and h (logarithmic scale).*

C. Interval Initial Condition.

We consider equation (4.34) with an initial conditions of the form

$$u(x,0) \in G_1(x) = [-0.001, 0.001] + \begin{cases} x(1-x)\,, & 0 \le x \le 1 \\ x(1+x)\,, & -1 \le x \le 0 \end{cases}\,, \quad u_t(x,0) = 0 \qquad (4.37)$$

In this case we compute bounds for a set of solutions

$$u(0, G; x, t) = \{u(0, g; x, t) : g \in G\}$$

where $G = (G_1, 0)$ and $[u(0, G; x, t)]$ denotes its optimal interval enclosure. Since the right-hand side of the equation is an increasing function of u, the corresponding differential operator is an operator of monotone type. In section 4.1 it was shown that

$$u(0, G) = [u(0, G)] = [u(0, \underline{g}), u(0, \overline{g})]\,.$$

The lower and upper bounds prescribed for $u(x, 0)$ are not in the Fourier functoid. Therefore they are rounded and lower and upper bounds in the form of trigonometric polynomials are computed with a certain error of approximation. While this error decreases when N increases, the total width of these bounds remains greater then 0.002 which is the width of G_1. Since the solutions diverge from each other when t increases, this results in a larger width of the bounds at $t = 2$. On figure 4.7 and figure 4.8, where the enclosures computed with $N = 20$ and $N = 40$ respectively are presented, the above phenomenon can be observed. In general, when the initial condition involves interval functions, a large enclosure width does not necessarily indicate poor accuracy because it may result from a large width of the solution set which is being enclosed.

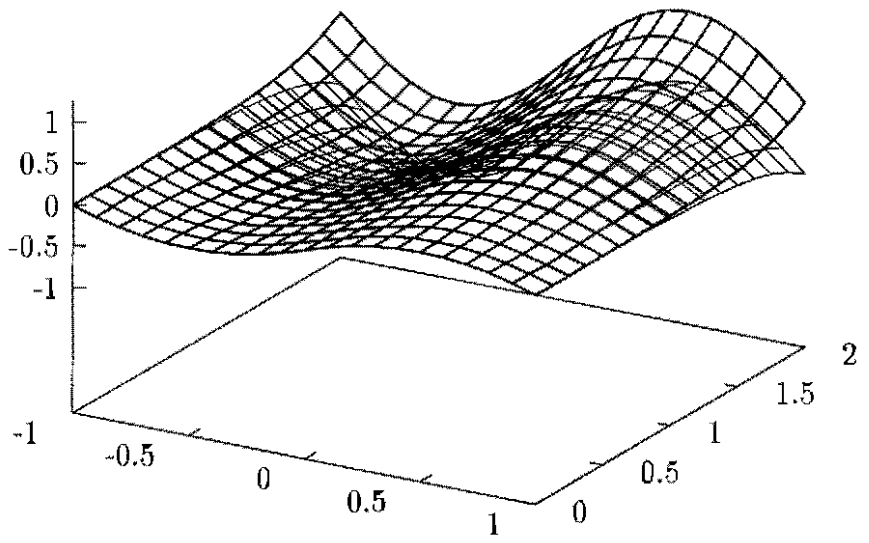


Figure 4.7: *Problem (4.34), (4.37). Enclosure computed with* $N = 20$ *and* $h = 2^{-7}$.

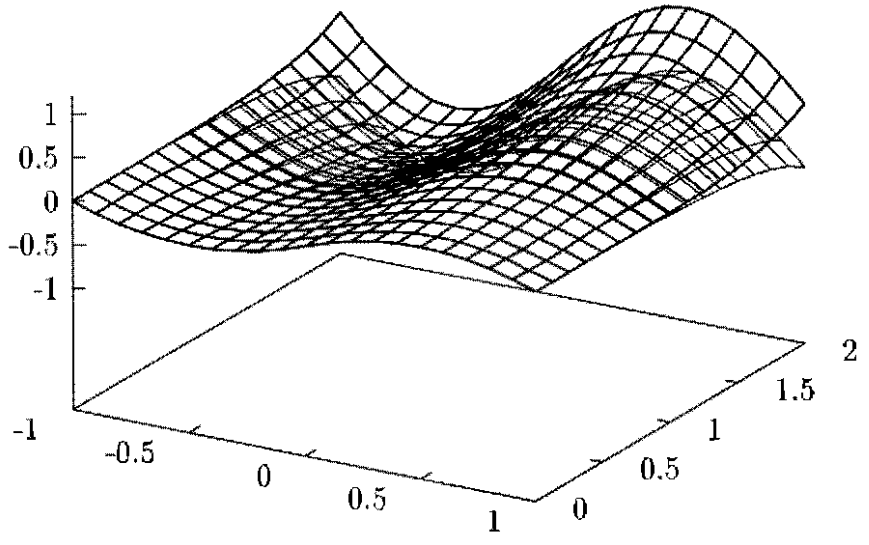


Figure 4.8: *Problem (4.34), (4.37). Enclosure computed with* $N = 40$ *and* $h = 2^{-8}$

In this case, it is important to know how close is the computed enclosure to the set of solutions or, more precisely, how close is the computed enclosure to the optimal interval enclosure of the solution set. The method which we consider allows us not only to compute enclosures but to compute their accuracy as well. We can compute a lower bound for the upper bound of $[u(0,G;x,t)]$ and an upper bound for the lower bound of $[u(0,G;x,t)]$ (see section 4.5). The enclosure $S(h,N;x,t)$ computed with $N=20$ and $h=2^{-7}$ together with the bounds discussed above are presented on figure 4.9. The interval function presented by the two outside surfaces (extreme top and bottom) is the enclosure $S(h,N;x,t)$ of $[u(0,G;x,t)]$ while the two surfaces inside presented an interval function $\check{S}(h,N;x,t)$ which is enclosed by $[u(0,G;x,t)]$, i.e. we have

$$\check{S}(h,N;x,t) \subset [u(0,G;x,t)] \subset S(h,N;x,t)$$

Therefore

$$|S(h,N;x,t) -^{-} \check{S}(h,N;x,t)| \tag{4.38}$$

is an estimate for the accuracy of the enclosure $S(h,N;x,t)$. On figure 4.10 the maximum norm in x of the estimate (4.38) for various values of N and h is plotted on a logarithmic scale against the time variable. The graphs are similar to the graphs on figure 4.6 and are consistent with the expected rate of convergence $o(N^{-\frac{3}{2}})$.

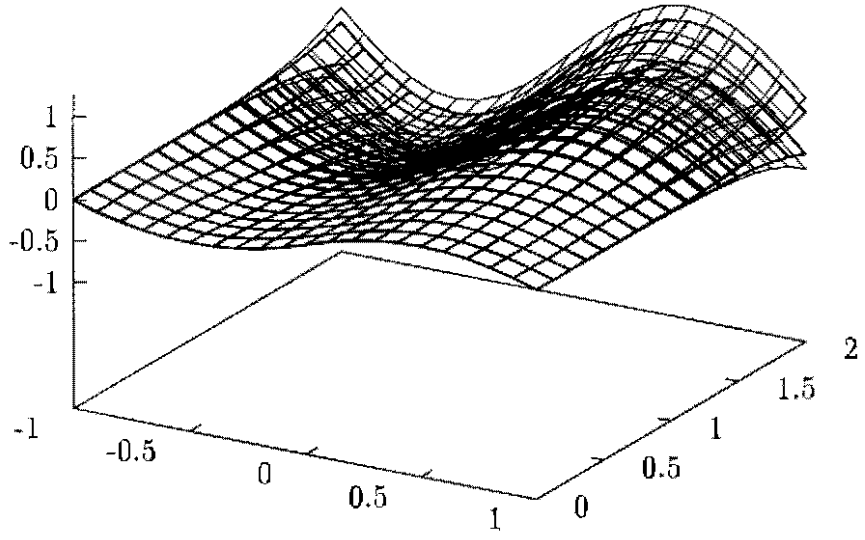


Figure 4.9: *Problem (4.34), (4.37). Enclosure $S(h,N)$ of the solution set $u(0,G)$ and inner approximation $\check{S}(h,N)$ of the solution set $u(0,G)$, both computed with $N=20$ and $h=2^{-7}$.*

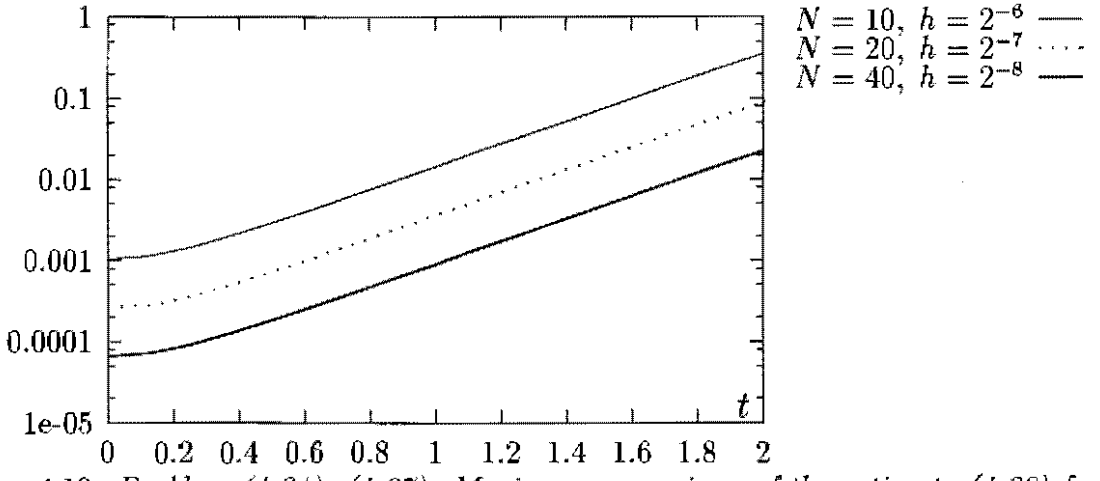


Figure 4.10: *Problem (4.34), (4.37). Maximum norm in x of the estimate (4.38) for the accuracy of the enclosures computed for various values of N and h (logarithmic scale).*

Example 4.2 We consider the following periodic initial value problem:

$$u_{tt} - u_{xx} = \frac{8}{3}u^3 + \pi^2 u + \frac{2\sin(3\pi x)}{3(t+1)^3}$$
$$u(x,0) = \sin(\pi x)\ ,\ u_t(x,0) = -\sin(\pi x) \tag{4.39}$$

The exact solution of this problem is

$$u = \frac{\sin(\pi x)}{t+1}.$$

The exact solution and bounds computed with $N = 5$ and $h = 2^{-5}$ are graphically represented on figure 4.11. The solution and the bounds are visually indistinguishable. Numerical values at some points of the exact solution and the bounds computed with $N = 5$ and several different values of h are presented in table 4.3. Since the exact solution is a trigonometric polynomial of x further increase in the number of spectral functions has little influence on the accuracy of the bounds. In the estimate (4.33) the dominating term is $O(h^4)$. It suggests that if the step size h is halved at least one more correct digit of the solution is obtained. This agrees with the numerical results in table 4.3. The errors of approximation are also represented graphically on figure 4.12 where the maximum norm in x of the radius of the bounds computed for several values of h is plotted on a logarithmic scale against the time variable.

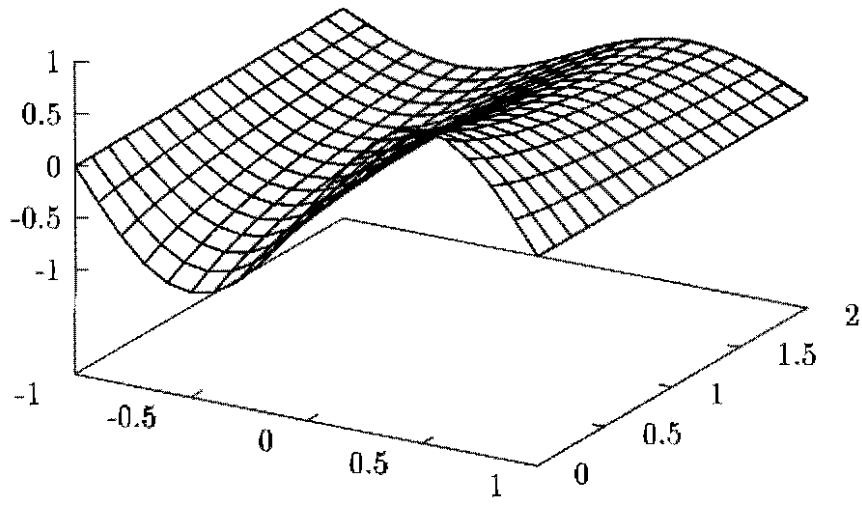


Figure 4.11: *Problem (4.39). Exact solution and enclosure computed with* $N = 5$ *and* $h = 2^{-5}$.

		$N = 5,\ h = 2^{-5}$		$N = 5,\ h = 2^{-6}$		$N = 5,\ h = 2^{-7}$	
		$x = 0.5$		$x = 0.5$		$x = 0.5$	
t	solution	bounds	radius	bounds	radius	bounds	radius
1.0	0.5000	$0.^{50017}_{49982}$	1.7E-4	$0.^{50001}_{49999}$	8.7E-6	$0.^{50001}_{49999}$	5.2E-7
2.0	0.3333	0.3^{3795}_{2871}	4.6E-3	0.333^{56}_{10}	2.3E-4	0.3333^{5}_{2}	1.4E-5

Table 4.3: *Problem (4.39). Values of the solution and the computed enclosures at some points.*

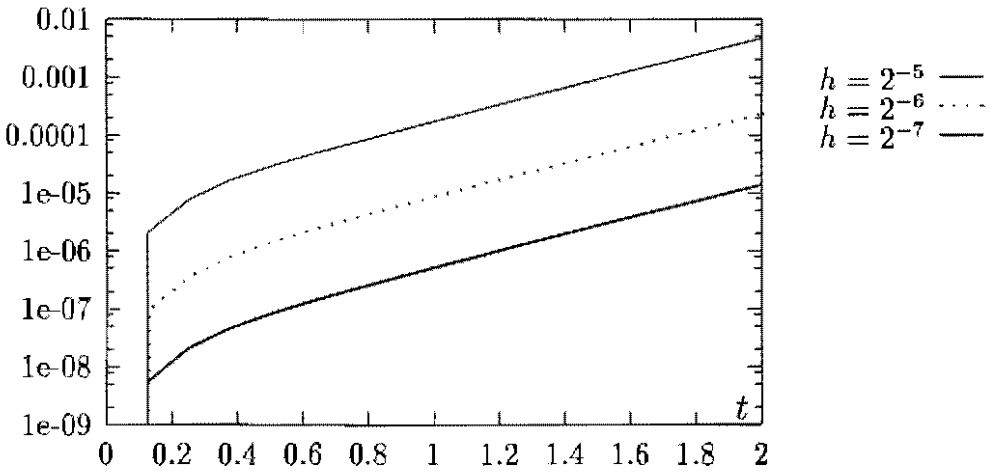


Figure 4.12: *Problem (4.39). Maximum norm in x of the radius of the enclosures computed for various step sizes h (logarithmic scale).*

Example 4.3 We consider an equation similar to the equation in example 2, but with an initial condition which is not in the Fourier functoid and requires rounding. We consider cases of noninterval and interval initial conditions.

$$u_{tt} - u_{xx} = u^3 \tag{4.40}$$

A. Noninterval Initial Condition.

$$u(x,0) = \begin{cases} x(1-x)\,, & 0 \le x \le 1 \\ x(1+x)\,, & -1 \le x \le 0 \end{cases}\,, \quad u_t(x,0) = 0 \tag{4.41}$$

As in example 4.1, case B, the prescribed value of $u(x,0)$ needs to be rounded and a larger number of spectral functions is required to achieve accuracy similar to the accuracy of the enclosures obtained in example 4.2. Figure 4.13 presents the enclosures computed with $N = 5$. The solutions of equation (4.40), unlike example 4.1, case B, do not diverge from each other. That is why the upper and lower bound remain close over the whole domain of the solution and the distance between them can not be observed visually on figure 4.13 (compare with figure 4.3).

We compute also bounds for the solution using $N = 10$ and $N = 20$ respectively. Since $Nh < const$, in the numerical experiments we double N and reduce the step size h twice. Numerical values of the enclosures at some points are presented in table 4.4. As in example 4.1, case B, the rate of convergence is $o(N^{-\frac{5}{2}}) + O(h^4)$ where the first term is dominating. It suggests that at least one more correct digit of the solution is obtained when N is increased by a factor of four. The numerical results presented in table 4.4 support that. In addition, on figure 4.14 the error of approximation is represented graphically. The

maximum norm in x of the radius of the enclosures computed for various values of N and h is plotted on a logarithmic scale against the time variable.

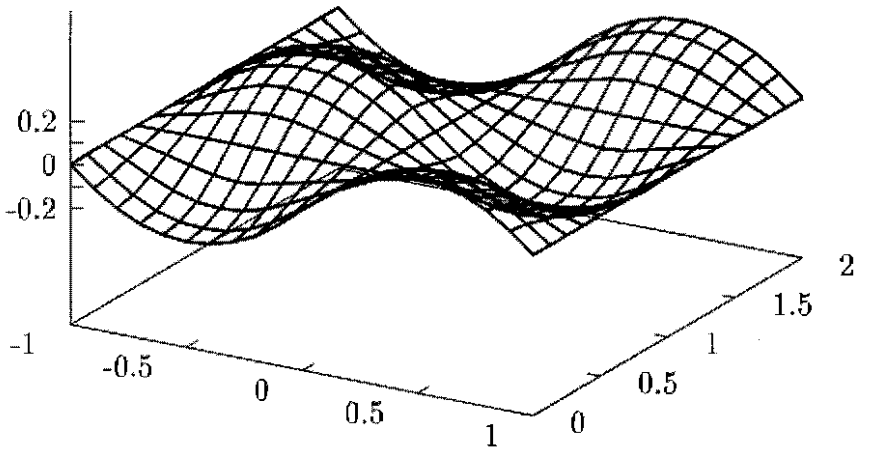


Figure 4.13: *Problem (4.40), (4.41). Enclosure computed with $N = 5$ and $h = 2^{-5}$*

	$N = 5,\ h = 2^{-5}$		$N = 10,\ h = 2^{-6}$		$N = 20,\ h = 2^{-7}$	
	$x = 0.5$		$x = 0.5$		$x = 0.5$	
t	bounds	radius	bounds	radius	bounds	radius
1.0	-0.2^{4729}_{5374}	3.2E-3	-0.2^{4904}_{5119}	1.1E-3	-0.2^{4971}_{5025}	2.7E-4
2.0	0.2^{5403}_{4697}	3.5E-3	0.2^{5126}_{4894}	1.2E-3	0.2^{5025}_{4967}	2.9E-4

Table 4.4: *Problem (4.40), (4.41). Values of the computed enclosures at some points .*

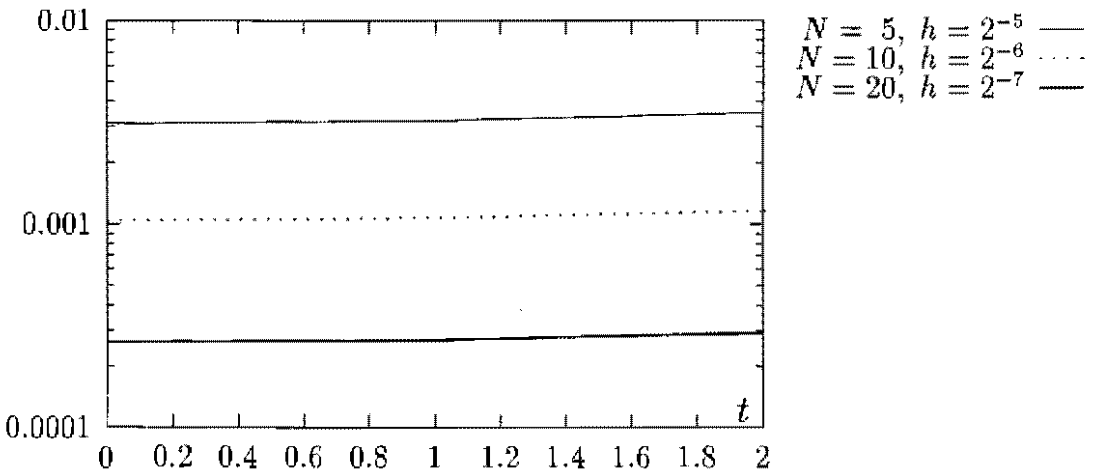


Figure 4.14: *Problem (4.40), (4.41). Maximum norm in x of the radius of the enclosures computed for various values of N and h (logarithmic scale).*

B. Interval Initial Condition.

$$u(x,0) \in G_1(x) = [-0.05, 0.05] + \begin{cases} x(1-x)\ ,\ 0 \le x \le 1 \\ x(1+x)\ ,\ -1 \le x \le 0 \end{cases}\ ,\ u_t(x,0) = 0 \qquad (4.42)$$

Similar to example 4.1, case C, here we compute bounds for a set of solutions

$$u(0, G; x, t) = \{u(0, g; x, t) : g \in G\}$$

where $G = (G_1, 0)$. Since $\underline{g}_1$ and $\overline{g}_1$ are not in the Fourier functoid, they are rounded and lower and upper bounds in the form of trigonometric polynomials are computed with a certain error of approximation. While this error decreases when N increases, the total width of these bounds remains greater then 0.1 which is the width of G_1. On figure 4.15 the enclosure of the solution is computed with $N = 5$ and $h = 2^{-5}$. Unlike example 4.1, case C, the solutions do not diverge from each other when t increases and the width of the enclosures remains of more or less the same magnitude on the whole domain. However, this width does not decrease significantly when N increases and h decreases because it results mainly from the width of the set $u(0, G; x, t)$. Using the same approach as in example 4.1. case C, we can compute an estimate for the accuracy of the computed enclosures. On figure 4.16, the maximum norm in x of this error estimate for enclosures computed with different values of N and h is plotted on a logarithmic scale against the time variable. The graphs are similar to the graphs on figure 4.14 which shows that the method works equally well with interval and noninterval initial conditions.

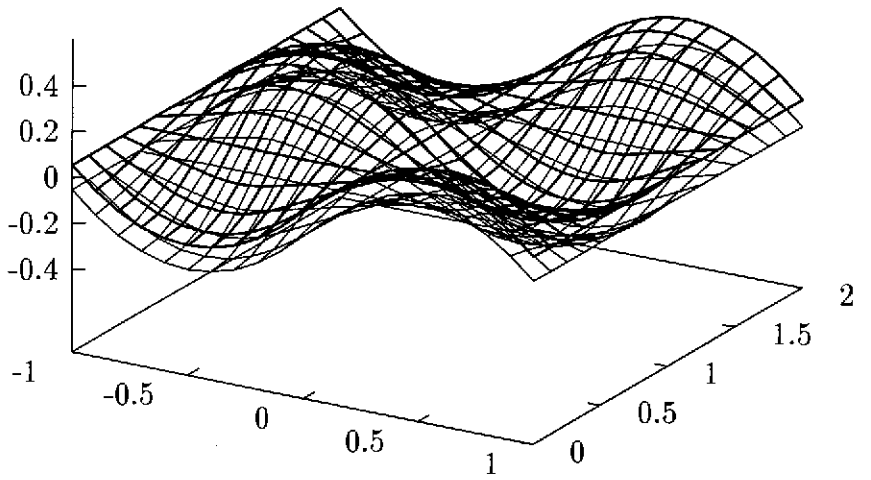


Figure 4.15: *Problem (4.40), (4.42). Enclosure computed with* $N = 5$ *and* $h = 2^{-5}$.

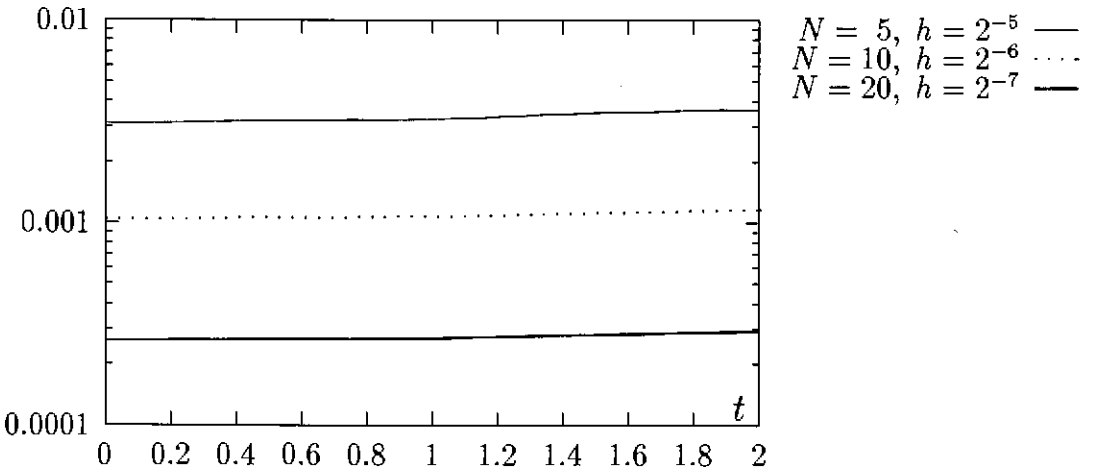


Figure 4.16: *Problem (4.40), (4.42). Maximum norm in* x *of estimate (4.38) for the accuracy of the enclosures computed for various values of* N *and* h *(logarithmic scale).*

Chapter 5

Spline-Fourier Approximations

5.1 The Concept of Hyper Functoid.

Let $\mathcal{M}$ be a separable and arithmetical Hilbert space with a basis $\{\varphi_k : k = 0, 1, \ldots\}$. The space

$$\left\{\sum_{k=0}^{N} c_k \varphi_k\right\}$$

with operations as discussed in section 2.6.1 is called a functoid. In general, a hyper functoid is obtained when infinite series are involved. Since infinitely many coefficients can not be stored individually, the coefficients have to be represented as a function of a coefficient index in a finite dimensional way.

A function $f \in \mathcal{M}$ can be represented as

$$f = \sum_{k=0}^{\infty} c_k \varphi_k.$$

In a functoid the sequence (c_k) representing f is cut off at some $k = N$. In hyper functoid theory the approach is different. The coefficients $c_k = c(k)$ are considered as functions of the coefficient index $k \in D = \mathcal{N} \cup \{0\}$. Let $\mathcal{K}$ denote the number set of the coefficients, i.e. $\mathcal{K} = \mathcal{R}$ or $\mathcal{K} = \mathcal{C}$. Then the space of functions $\mathcal{K}(D) = \{c : D \mapsto K\}$ is called a coefficient space. Let $\{d_0, d_1, \ldots\}$ be a basis in $K(D)$. Then a rounding can be declared in $\mathcal{K}(D)$ by

$$\check{\tau}_N(c) = \check{\tau}_N\left(\sum_{i=0}^{\infty} b_i d_i\right) = \sum_{i=0}^{N} b_i d_i.$$

This rounding maps the space $\mathcal{K}(D)$ onto the screen

$$K(D) = \left\{\sum_{i=0}^{N} b_i d_i : b_i \in \mathcal{K}\right\}$$

This will induce a rounding on $\mathcal{M}$ of the form

$$\tau_N(f) = \tau_N\left(\sum_{k=0}^{\infty} c(k)\varphi_k\right) = \sum_{k=0}^{\infty} \check{\tau}_N(c)(k)\varphi_k$$

The resulting structure

$$M = \left\{\sum_{k=0}^{\infty} c(k)\varphi_k : c(k) = \sum_{i=0}^{N} b_i d_i(k),\ b_i \in \mathcal{K}\right\}$$

with operations induced by the rounding $\check{\tau}_N$ (as in a functoid) is called a hyper functoid.[48]

5.2 Fourier Hyper Functoid.

A function $f \in L_2(-1,1)$ has a Fourier series of the form

$$f(x) \sim \sum_{k=-N}^{N} c(k)e^{ik\pi x}$$

where $c(k) = conj(c(-k)) \in \mathcal{C}$. Therefore the domain of the coefficient index is $D = \mathcal{Z}$ and the coefficient space is

$$\mathcal{K}(D) = \left\{c : D \mapsto \mathcal{C} : c(k) = conj(c(-k)),\ \sum_{k=0}^{\infty} |c(k)|^2 < \infty\right\}.$$

In [48] the following screen $K(D)$ of $\mathcal{K}(D)$ is considered:

$$K(D) = \left\{\sum_{i=-N}^{N} b_i d_i + \sum_{j=1}^{p} a_j d_{N+j} : b_i = conj(b_{-i}) \in \mathcal{C},\ a_j \in \mathcal{R}\right\}$$

where

$$d_i(k) = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{if } k \neq i \end{cases} \quad \text{for } i = 0, \pm 1, \ldots, \pm N \text{ and}$$

$$d_{N+j}(k) = \begin{cases} 0 & \text{if } |k| \leq N \\ \dfrac{(-1)^k}{(ik)^j} & \text{if } |k| > N \end{cases} \quad \text{for } j = 1, \ldots, p.$$

The corresponding Fourier hyper functoid is

$$\mathcal{F}_S = \left\{\sum_{k=-\infty}^{\infty} c(k)e^{ik\pi x} : c \in K(D)\right\}.$$

The coefficient functions $c(k)$ can also be presented in the form

$$c(k) = \begin{cases} b_k & \text{if} \quad |k| \le N \\ (-1)^k \sum_{j=1}^{p} a_j \frac{1}{(ik)^j} & \text{for} \quad |k| > N \end{cases}$$

where $b_k = conj(b_{-k}) \in \mathcal{C}$ and $a_j \in \mathcal{R}$.[48]

Let $f(x) = \sum_{k=-\infty}^{\infty} c(k) e^{ik\pi x} \in \mathcal{F}_S$. Then we have

$$\begin{aligned} f(x) &= \sum_{k=-N}^{N} b_k e^{ik\pi x} + a_j \sum_{j=1}^{p} \sum_{|k|>N} \frac{(-1)^k}{(ik)^j} e^{ik\pi x} \\ &= \sum_{k=-N}^{N} \beta_k e^{ik\pi x} + \alpha_j \sum_{j=1}^{p} \sum_{k=-\infty}^{\infty} \frac{(-1)^k}{(ik\pi)^j} e^{ik\pi x} \end{aligned}$$

where $\beta_0 = b_0$, $\beta_k = b_k - \sum_{j=1}^{p} a_j \frac{(-1)^k}{(ik)^j}$, $k = \pm 1, \pm 2, \ldots, \pm N$ and $\alpha_j = \frac{a_j}{\pi^j}$, $j = 1, 2, \ldots, p$.

Let $s_j(x) = \sum_{k=-\infty}^{\infty} \frac{(-1)^k}{(ik\pi)^j} e^{ik\pi x}$, $x \in \mathcal{R}$. Then $\mathcal{F}_S$ can be represented in the form

$$\mathcal{F}_S = \left\{ f(x) = \sum_{k=-N}^{N} \beta_k e^{ik\pi x} + \sum_{j=1}^{p} \alpha_j s_j(x) : \beta_k = conj(\beta_{-k}) \in \mathcal{C},\ \alpha_j \in \mathcal{R} \right\}. \tag{5.1}$$

In the following sections we will use the above representation of the Fourier hyper functoid for the approximation of periodic functions and solving the wave equation in the case when some of the data functions or their derivatives have discontinuities. We will see later that the functions $\{s_j : j = 1, 2, \ldots\}$ are in fact polynomials on the interval $(-1, 1)$ and therefore they are splines when produced periodically over $(-\infty, \infty)$. That is why we refer to the Fourier hyper functoid $\mathcal{F}_S$ defined above as a Spline-Fourier functoid and we will call the approximations with this functoid Spline-Fourier approximations. We believe that the explicit use of the periodic splines gives some advantages in defining the roundings (left, right, interval), deriving formulas for the operations in the functoid and in the computation of a validated solution of the wave equation. Let us note that the Spline-Fourier functoid is a Fourier hyper functoid but not the only Fourier hyper functoid. Other hyper functoids can be derived using a different basis in the coefficient space.

In the next section we will give a new definition of the periodic splines $\{s_j : j = 1, 2, \ldots\}$ and discuss their properties.

5.3 Definition and Properties of the Periodic Splines.

We consider a set of splines $\{s_j : j = 0, 1, ...\}$ satisfying the following conditions

$$\begin{aligned}
&(i) \quad s_j \text{ is a polynomial of degree } j \text{ on } (-1,1), j = 0,1,\ldots \\
&(ii) \quad s_j \text{ is periodical with a period } 2, j = 0,1,\ldots \\
&(iii) \quad s_j \in C^{j-2}(-\infty,\infty),\ j = 2,3,\ldots \\
&(iv) \quad \frac{ds_{j+1}(x)}{dx} = s_j(x),\ x \in (-1,1),\ j = 0,1,\ldots \\
&(v) \quad s_0(x) = 1,\ x \in (-\infty,\infty)
\end{aligned} \tag{5.2}$$

We can construct the elements of the set inductively using

$$s_{j+1}(x) = \int s_j(x)dx + c \tag{5.3}$$

and determining the value of the constant of integration c from

$$\int_{-1}^{1} s_{j+1}(x)dx = s_{j+2}(1) - s_{j+2}(-1) = 0 \tag{5.4}$$

We find the splines in the form

$$\begin{aligned}
s_{2m} &= c_0\frac{x^{2m}}{(2m)!} + c_1\frac{x^{2m-2}}{(2m-2)!} + \ldots + c_{m-1}\frac{x^2}{2!} + c_m\,, \qquad x \in (-1,1) \\
s_{2m+1} &= c_0\frac{x^{2m+1}}{(2m+1)!} + c_1\frac{x^{2m-1}}{(2m-1)!} + \ldots + c_{m-1}\frac{x^3}{3!} + c_m x\,, \quad x \in (-1,1)
\end{aligned}$$

where the coefficients $c_0, c_1, c_2, \ldots$ are obtained from the following linear system

$$\begin{aligned}
&c_0 = 1 \\
&\frac{c_0}{(2m+1)!} + \frac{c_1}{(2m-1)!} + \ldots + \frac{c_{m-1}}{3!} + c_m = 0\,,\ m = 1,2,\ldots
\end{aligned} \tag{5.5}$$

The splines constructed in this way obviously satisfy conditions *(i),(ii),(iv),(v)* in (5.2). We only need to prove *(iii)*. Relations (5.3) and (5.4) imply that s_j are continuous for $j \geq 2$. Therefore, for $j \geq 2$ *(iv)* is satisfied on the interval $(-\infty,\infty)$. Differentiating $j-2$ times we have $\frac{d^{j-2}s_j(x)}{dx^{j-2}} = s_2(x)$. Hence $s_j \in C^{j-2}(-\infty,\infty)$.

The first few splines in the considered set are listed below

$$\begin{aligned}
s_1(x) &= x\,, & x &\in (-1,1) \\
s_2(x) &= \frac{x^2}{2!} - \frac{1}{6}\,, & x &\in [-1,1] \\
s_3(x) &= \frac{x^3}{3!} - \frac{1}{6}x\,, & x &\in [-1,1] \\
s_4(x) &= \frac{x^4}{4!} - \frac{1}{6}\frac{x^2}{2!} + \frac{7}{360}\,, & x &\in [-1,1] \\
s_5(x) &= \frac{x^5}{5!} - \frac{1}{6}\frac{x^3}{3!} + \frac{7}{360}x\,, & x &\in [-1,1] \\
s_6(x) &= \frac{x^6}{6!} - \frac{1}{6}\frac{x^4}{4!} + \frac{7}{360}\frac{x^2}{2!} - \frac{31}{15120}\,, & x &\in [-1,1]
\end{aligned}$$

Some properties of the periodic splines are presented in the following theorem.

Theorem 5.1

$$(i)\ \ s_j(-x) = (-1)^j s_j(x)\ ,\ \ s_j(1+x) = (-1)^j s_j(1-x)\ ,\ \ j \geq 0$$

$$(ii)\ \ s_{2m+1}(-1) = s_{2m+1}(0) = s_{2m+1}(1)\ ,\ m \geq 0$$

$$(iii)\ \ s_j(x) = -\frac{1}{(i\pi)^j} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \frac{(-1)^k}{k^j} e^{ik\pi x}\ ,\ j \geq 1$$

$$(iv) \int_{-1}^{1} s_p(x) s_q(x) dx = (-1)^{q-1} 2 s_{p+q}(1) = \begin{cases} 0 & ,\ \ p \not\equiv q (\mathrm{mod}\, 2) \\ (-1)^{q-1} 2 s_{p+q}(1), & p \equiv q (\mathrm{mod}\, 2) \end{cases}$$

We assume that $s_1(1) = s_1(-1) = 0$ to ensure that (iii) is satisfied for any x when $j = 1$.

Remark. The periodic splines described in this section are closely related to the monosplines, which are well-known in spline theory, as well as the Bernoulli polynomials. In fact $n! s_n$ and the nth monospline differ by a constant and $s_n(x) = \frac{2^n}{n!} B_n(\frac{x+1}{2})$, $x \in (-1, 1)$, where B_n is the nth Bernoulli polynomial.

5.4 Spline-Fourier Expansion of Real Functions.

In section 4.3 we discussed the Sobolev spaces $H^p(-1,1)$ and $H^p_{per}(-1,1)$. The estimate (4.22) shows that the Fourier series of any function $f \in H^p_{per}(-1,1)$, $p \geq 1$ converges uniformly on $[-1,1]$ to f at a rate of $o(N^{\frac{1}{2}-p})$. This is not true for the space $H^p(-1,1)$. We have

$$H^p_{per}(-1,1) \subset H^p(-1,1)$$

and the essential difference between the two spaces is that the periodical extensions of the functions in $H^p(-1,1)$ and their first $p-1$ derivatives may be discontinuous at the end points of the interval $[-1,1]$. If the periodical extension of $f \in H^p(-1,1)$ is discontinuous at $x = \pm 1$ then the Fourier series of f does not converge uniformly and we have the Gibbs phenomenon. If $f \in H^p(-1,1)$ and its first $j-1$ $(j < p)$ derivatives have continuous periodical extensions the Fourier series of f converges to f uniformly on $[1,1]$ at a rate of $o(N^{\frac{1}{2}-j})$, i.e. at a rate slower then the rate of convergence for functions in $H^p_{per}(-1,1)$. In this section we will consider the use of periodic splines in the Fourier series which leads to a series expansion (Spline-Fourier series) of the functions in $H^p(-1,1)$ with the same qualities (e.g. rate of uniform convergence) as the Fourier series of the functions in $H^p_{per}(-1,1)$.

Theorem 5.2 *Every function $f \in H^p(-1,1)$ has a unique representation in the form*

$$f(x) = a_0 s_0(x) + a_1 s_1(x) + \ldots + a_p s_p(x) + \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} b_k e^{ik\pi x} \tag{5.6}$$

where $\displaystyle\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} b_k e^{ik\pi x} \in H^p_{per}(-1,1).$

The coefficients in (5.6) can be obtained as follows

$$a_j = \frac{1}{2}\left(\frac{d^{j-1}f}{dx^{j-1}}(1-0) - \frac{d^{j-1}f}{dx^{j-1}}(1+0)\right) = \frac{1}{2}\int_{-1}^{1}\frac{d^j f(x)}{dx^j}dx\ ,\ j=1,\ldots,p$$
$$a_0 = \frac{1}{2}\int_{-1}^{1} f(x)dx \tag{5.7}$$
$$b_k = \frac{1}{2(ik\pi)^p}\int_{-1}^{1}\frac{d^p f(x)}{dx^p}e^{-ik\pi}dx\ ,\ k=\pm 1,\pm 2,\ldots$$

Proof. Uniqueness. If the representation (5.6) exists then

$$g(x) = f(x) - a_0 s_0(x) - a_1 s_1(x) - \ldots - a_p s_p(x) = \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} b_k e^{ik\pi x} \in C^{p-1}(-\infty,\infty)$$

Integrating the above two expressions for g we obtain $a_0 = \frac{1}{2}\int_{-1}^{1} f(x)dx$.

The $(j-1)$st derivative of g is

$$\frac{d^{j-1}g(x)}{dx^{j-1}} = \frac{d^{j-1}f(x)}{dx^{j-1}} - a_{j-1}s_0(x) - a_j s_1(x) - \ldots - a_p s_{p-j+1}(x)$$

Since $s_0, s_2, \ldots, s_{p-j+1}$ are continuous in $(-\infty,\infty)$ then $\frac{d^{j-1}f(x)}{dx^{j-1}} - a_j s_1(x)$ must be also continuous in $(-\infty,\infty)$. Therefore

$$\frac{d^{j-1}f}{dx^{j-1}}(1-0) - a_j s_1(1-0) = \frac{d^{j-1}f}{dx^{j-1}}(1+0) - a_j s_1(1+0)$$

which implies

$$a_j = \frac{1}{2}\left(\frac{d^{j-1}f}{dx^{j-1}}(1-0) - \frac{d^{j-1}f}{dx^{j-1}}(1+0)\right),\ j=1,2,\ldots,p.$$

We also have

$$\frac{d^p g(x)}{dx^p} = \frac{d^p f(x)}{dx^p} - a_p = \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty}(ik\pi)^p b_k e^{ik\pi x}$$

Using the formula for the coefficients of the Fourier series expansion of $\frac{d^p g(x)}{dx^p}$ we obtain

$$(ik\pi)^p b_k = \frac{1}{2}\int_{-1}^{1}\frac{d^p g(x)}{dx^p}e^{-ik\pi x}dx = \frac{1}{2}\int_{-1}^{1}\frac{d^p f(x)}{dx^p}e^{-ik\pi x}dx$$

which implies

$$b_k = \frac{1}{2(ik\pi)^p}\int_{-1}^{1}\frac{d^p f(x)}{dx^p}e^{-ik\pi}dx$$

We proved that if the representation (5.6) exists then the coefficients are obtained according to (5.7). This implies that the representation is unique.

Existence. Let

$$g(x) = f(x) - a_0 s_0(x) - a_1 s_1(x) - \ldots - a_p s_p(x)$$

where $a_0, a_1, \ldots, a_p$ are given by (5.7). From the formula for a_0 we obtain that $\int_{-1}^{1} g(x)dx = 0$. Therefore g can be expanded in a Fourier series

$$g(x) = \sum_{k=-\infty}^{\infty} b_k e^{ik\pi x}$$

with $b_0 = 0$. What we have to prove is that $g \in H^p_{per}(-1,1)$. Since $g \in H^p(-1,1)$ we only need to prove that g and its first $p-1$ derivatives are continuous at $x = 1$ (when produced periodically). Note that the only spline that is discontinuous at $x = 1$ is s_1. We have

$$\begin{aligned} g(1-0) - g(1+0) &= f(1-0) - f(1+0) - a_1(s_1(1-0) - s_1(1+0)) \\ &= f(1-0) - f(1+0) - 2a_1 = 0 \end{aligned}$$

Therefore g is continuous at $x = 1$. In a similar way, for the jth derivative of g we have

$$\frac{d^j g}{dx^j}(1-0) - \frac{d^j g}{dx^j}(1+0) = \frac{d^j f}{dx^j}(1-0) - \frac{d^j f}{dx^j}(1+0) - 2a_{j+1}\ ,\ j = 1, 2, \ldots, p-1$$

and using (5.7) for a_{j+1} we obtain

$$\frac{d^j g}{dx^j}(1-0) - \frac{d^j g}{dx^j}(1+0) = 0$$

This completes the proof.

5.5 Spline-Fourier Functoid.

In this section we will consider the Spline-Fourier functoid $\mathcal{FS}$ and the interval Spline-Fourier Functoid $\mathcal{IFS}$ as a screen and interval screen respectively of the space $\mathcal{M} = H^p(-1,1)$. We will also derive suitable formulas for the roundings and the operations in these functoids.

We define in $\mathcal{M}$ a rounding ρ_{Np} in the following way. Let $f \in \mathcal{M}$ and let f be represented in the form (5.6). Then

$$\rho_{Np}(f;x) = a_0 s_0(x) + a_1 s_1(x) + \ldots + a_p s_p(x) + \sum_{\substack{k=-N \\ k \neq 0}}^{N} b_k e^{ik\pi x} \tag{5.8}$$

The rounding error can be estimated as follows

$$
\begin{aligned}
|f(x) - \rho_{Np}(f;x)| &= \left| \sum_{|k|>N} b_k e^{ik\pi x} \right| \leq \sum_{|k|>N} |b_k| \\
&\leq \left(\sum_{|k|>N} (k\pi)^{2p} |b_k|^2 \right)^{\frac{1}{2}} \left(\sum_{|k|>N} \frac{1}{(k\pi)^{2p}} \right)^{\frac{1}{2}} \\
&\leq \left(\frac{1}{2} \int_{-1}^{1} \left(\frac{d^p f(x)}{dx^p} \right)^2 dx - a_p^2 - 2 \sum_{k=1}^{N} (k\pi)^{2p} |b_k|^2 \right)^{\frac{1}{2}} \left(\frac{2}{(2p-1)\pi^{2p} N^{2p-1}} \right)^{\frac{1}{2}} \\
&= o\left(\frac{1}{N^{p-\frac{1}{2}}} \right)
\end{aligned}
$$

The rounding ρ_{Np} maps $\mathcal{M}$ on to the screen

$$M = span\{s_0, s_1, \ldots, s_p, e^{\pm i\pi x}, e^{\pm 2i\pi x}, \ldots, e^{\pm Ni\pi x}\}$$

In $\mathcal{M}$ we consider the operations $\omega = \{+, -, ., /, \int\}$ defined in the conventional way. By the semimorphism principle ρ_{Np} induces corresponding operations in M:

$$
\begin{aligned}
f \boxed{\circ} g &= \rho_{Np}(f \circ g), \quad \circ \in \{+, -, \times, /\} \\
\boxed{\int} f &= \rho_{Np}\left(\int f \right)
\end{aligned}
$$

The structure $\mathcal{F}_S = (M, \boxplus, \boxminus, \boxdot, \boxslash, \boxed{\int})$ is called a Spline-Fourier functoid of $\mathcal{M}$. In order to use this functoid in approximations, e.g. deriving approximate solutions of mathematical problems, we need constructive methods for implementation of the operations. Since $\mathcal{F}_S$ is a linear space then it is closed with respect to the operations addition and scalar multiplication. Those operations are performed by adding the coefficient vectors or multiplying them by a scalar. Other operations in ω may produce a result outside $\mathcal{F}_S$ that has to be rounded. We will consider them one by one.

Multiplication. Let $f_1,\ f_2 \in \mathcal{M}$ be given in the form

$$
\begin{aligned}
f_1(x) &= \sum_{j=0}^{p} a_{1j} s_j(x) + \sum_{\substack{k=-N \\ k\neq 0}}^{N} b_{1k} e^{ik\pi x} \\
f_2(x) &= \sum_{j=0}^{p} a_{2j} s_j(x) + \sum_{\substack{k=-N \\ k\neq 0}}^{N} b_{2k} e^{ik\pi x}
\end{aligned}
$$

Their product can be written as

$$f_1(x) f_2(x) = \Sigma_{ss} + \Sigma_{ee} + \Sigma_{se}$$

where

$$\Sigma_{ss} = \left(\sum_{j=0}^{p} a_{1j}s_j(x)\right)\left(\sum_{j=0}^{p} a_{2j}s_j(x)\right)$$

$$\Sigma_{ee} = \left(\sum_{\substack{k=-N\\k\neq 0}}^{N} b_{1k}e^{ik\pi x}\right)\left(\sum_{\substack{k=-N\\k\neq 0}}^{N} b_{2k}e^{ik\pi x}\right)$$

$$\Sigma_{se} = \left(\sum_{j=0}^{p} a_{1j}s_j(x)\right)\left(\sum_{\substack{k=-N\\k\neq 0}}^{N} b_{2k}e^{ik\pi x}\right) + \left(\sum_{j=0}^{p} a_{2j}s_j(x)\right)\left(\sum_{\substack{k=-N\\k\neq 0}}^{N} b_{1k}e^{ik\pi x}\right)$$

We will consider separately each of the above sums which, for convenience, we will call splines product (Σ_{ss}), exponents product (Σ_{ee}) and mixed product (Σ_{se}).

The splines product. Let $g(x) = s_m(x)s_q(x) \in \mathcal{M}$, $m, q \geq 1$. Function g is a spline of degree $m+q$ and can be represented as a linear combination of $s_0, s_1, \ldots, s_{m+q}$:

$$g(x) = \sum_{j=0}^{m+q} \alpha_j s_j(x)$$

From theorem 5.2 we have $\alpha_j = \frac{1}{2}\left(\frac{d^{j-1}g}{dx^{j-1}}(1-0) - \frac{d^{j-1}g}{dx^{j-1}}(1+0)\right)$ where

$$\frac{d^{j-1}g(x)}{dx^{j-1}} = \sum_{r=\max\{0,j-q-1\}}^{\min\{j-1,m\}} \binom{j-1}{r} s_{m-r}(x)s_{q-j+r+1}(x)$$

Using the fact that s_1 is the only discontinuous function that may appear in the above expression and $s_1(1-0) - s_1(1+0) = 2$ we obtain the following values for the coefficients α_j (assuming $m \leq q$):

$$\begin{aligned}
\alpha_j &= 0 && \text{for } 0 \leq j \leq m-1 \\
\alpha_j &= \tbinom{j-1}{m-1} s_{q+m-j}(1) && \text{for } m \leq j \leq q-1 \\
\alpha_j &= \left(\tbinom{j-1}{m-1} + \tbinom{j-1}{q-1}\right) s_{q+m-j}(1) && \text{for } q \leq j \leq q+m
\end{aligned} \tag{5.9}$$

Hence

$$s_m(x)s_q(x) = \sum_{j=m}^{m+q} \binom{j-1}{m-1} s_{m+q-j}(1)s_j(x) + \sum_{j=q}^{m+q} \binom{j-1}{q-1} s_{m+q-j}(1)s_j(x)$$

Now the splines product can be written in the form

$$\Sigma_{ss} = \sum_{m=0}^{p}\sum_{q=0}^{p} a_{1m}a_{2q}s_m(x)s_q(x)$$

$$
\begin{aligned}
&= a_{10}a_{20} + \sum_{m=1}^{p}\sum_{q=0}^{p}\sum_{j=m}^{m+q}\binom{j-1}{m-1}a_{1m}a_{2q}s_{m+q-j}(1)s_j(x) \\
&\quad + \sum_{q=1}^{p}\sum_{m=0}^{p}\sum_{j=q}^{m+q}\binom{j-1}{q-1}a_{1m}a_{2q}s_{m+q-j}(1)s_j(x) \\
&= a_{10}a_{20} + \sum_{j=1}^{2p}\sum_{m=\max\{1,j-p\}}^{\min\{j,p\}}\sum_{q=j-m}^{p}\binom{j-1}{m-1}a_{1m}a_{2q}s_{m+q-j}(1)s_j(x) \\
&\quad + \sum_{j=1}^{2p}\sum_{q=\max\{1,j-p\}}^{\min\{j,p\}}\sum_{m=j-q}^{p}\binom{j-1}{q-1}a_{1m}a_{2q}s_{m+q-j}(1)s_j(x)
\end{aligned}
$$

The above sum essentially consists of scalar products and can be conveniently represented in a matrix form. Let

$$a_1 = (a_{10}, a_{11}, \ldots, a_{1p})^T \quad , \quad a_2 = (a_{20}, a_{21}, \ldots, a_{2p})^T$$

and let $\tilde{M}_j$ be a matrix of type $(p+1, p+1)$ defined by

$$\left(\tilde{M}_j\right)_{mq} = \begin{cases} \binom{j-1}{m-1}s_{m+q-j}(1) & \text{if} \quad 1 \le m \le j \le m+q \\ 0 & \text{otherwise} \end{cases} \quad , \; m, q = 0, 1, \ldots, p$$

Then the splines sum is

$$
\begin{aligned}
\Sigma_{ss} &= a_{10}a_{20}s_0(x) + \sum_{j=1}^{2p} a_1^T\tilde{M}_j a_2 s_j(x) + \sum_{j=1}^{2p} a_2^T\tilde{M}_j a_1 s_j(x) \\
&= a_{10}a_{20}s_0(x) + \sum_{j=1}^{2p} a_1^T\left(\tilde{M}_j + \tilde{M}_j^T\right)a_2 s_j(x)
\end{aligned}
$$

or

$$\Sigma_{ss} = a_{10}a_{20}s_0(x) + \sum_{j=1}^{2p} a_1^T M_j^{(s)} a_2 s_j(x) \tag{5.10}$$

where $M_j^{(s)} = \tilde{M}_j + \tilde{M}_j^T$. Using the Fourier series of the splines $s_{p+1}, \ldots, s_{2p}$ we obtain the Spline-Fourier expansion of Σ_{ss} in the form

$$
\begin{aligned}
\Sigma_{ss} &= a_{10}a_{20}s_0(x) + \sum_{j=1}^{p} a_1^T M_j^{(s)} a_2 s_j(x) \\
&\quad + \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} a_1^T\left(\sum_{j=p+1}^{2p}\frac{(-1)^{k-1}}{(ik\pi)^j}M_j^{(s)}\right)a_2 e^{ik\pi x} \\
&= a_{10}a_{20}s_0(x) + \sum_{j=1}^{p} a_1^T M_j^{(s)} a_2 s_j(x) + \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} a_1^T\hat{M}_k^{(s)} a_2 e^{ik\pi x}
\end{aligned}
$$

where

$$\hat{M}_k^{(s)} = \sum_{j=p+1}^{2p} \frac{(-1)^{k-1}}{(ik\pi)^j} M_j^{(s)}$$

Exponents product. Combining the equal powers of e we have

$$\begin{aligned}\Sigma_{ee} &= \sum_{k=-2N}^{2N} \left(\sum_{\substack{m=\max\{-N,k-N\} \\ m\neq 0,k}}^{min\{N,k+N\}} b_{1\,m} b_{2\,k-m} \right) e^{ik\pi x} \\ &= \left(\sum_{\substack{m=-N \\ m\neq 0}}^{N} b_{1m} b_{2\,-m} \right) s_0(x) + \sum_{\substack{k=-2N \\ k\neq 0}}^{2N} \left(\sum_{\substack{m=\max\{-N,k-N\} \\ m\neq 0,k}}^{min\{N,k+N\}} b_{1\,m} b_{2\,k-m} \right) e^{ik\pi x}\end{aligned}$$

Using the following notations

$$\begin{aligned} b_1 &= (b_{1\,-N}, \ldots, b_{1\,-1}, 0, b_{11}, \ldots, b_{1N})^T \\ b_2 &= (b_{2\,-N}, \ldots, b_{2\,-1}, 0, b_{21}, \ldots, b_{2N})^T \end{aligned}$$

the coefficients in the above expression can be presented in a matrix form as follows

$$\Sigma_{ee} = \left(b_1^T E_0 b_2 \right) s_0(x) + \sum_{\substack{k=-2N \\ k\neq 0}}^{2N} \left(b_1^T E_k b_2 \right) e^{ik\pi x}$$

where E_k, $k = 0, \pm 1, \ldots, \pm 2N$ are matrices of type $(2N+1, 2N+1)$ defined by

$$(E_k)_{qr} = \begin{cases} 1 & \text{if} \quad q + r = 2N + 2 + k \\ 0 & \text{otherwise} \end{cases}, \quad q, r = 1, 2, \ldots, 2N+1$$

Mixed product. We will use again Theorem 5.2 to obtain the coefficients in the expansion

$$g(x) = s_m(x) e^{iq\pi x} = \sum_{j=0}^{p} \alpha_{mqj} s_j(x) + \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \beta_{mqk} e^{ik\pi x}$$

Since $s_0(x) e^{iq\pi x} = e^{iq\pi x}$ then

$$\alpha_{0qj} = 0 \qquad \text{and} \qquad \beta_{0qk} = \begin{cases} 1 & k = q \\ 0 & k \neq q \end{cases} \tag{5.11}$$

For $m \geq 1$ we have

$$\frac{d^{j-1} g(x)}{dx^{j-1}} = \sum_{r=0}^{min\{m,j-1\}} \binom{j-1}{r} s_{m-r}(x) (iq\pi)^{j-r-1} e^{iq\pi x}$$

Hence

$$\alpha_{mqj} = 0 \quad \text{for} \quad 0 \le j \le m-1 \tag{5.12}$$

and

$$\alpha_{mqj} = \binom{j-1}{m-1}(iq\pi)^{j-m}e^{iq\pi} = (-1)^q\binom{j-1}{m-1}(iq\pi)^{j-m} \quad \text{for} \quad m \le j \le p \tag{5.13}$$

The coefficients of the exponents in the expansion of g we obtain from its pth derivative

$$\begin{aligned}
\frac{d^p g(x)}{dx^p} &= \sum_{r=0}^{m}\binom{p}{r}s_{m-r}(x)(iq\pi)^{p-r}e^{iq\pi x} \\
&= \binom{p}{m}(iq\pi)^{p-m}e^{iq\pi x} + \sum_{r=0}^{m-1}\binom{p}{r}\left(\sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty}\frac{(-1)^{l-1}}{(il\pi)^{m-r}}e^{il\pi x}\right)(iq\pi)^{p-r}e^{iq\pi x} \\
&= \binom{p}{m}(iq\pi)^{p-m}e^{iq\pi x} + \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty}(-1)^{l-1}(iq\pi)^{p-m}\left(\sum_{r=0}^{m-1}\binom{p}{r}\left(\frac{q}{l}\right)^{m-r}\right)e^{i(q+l)\pi x}
\end{aligned}$$

The term in the above sum corresponding to $l = -q$ is

$$\begin{aligned}
&(-1)^{q-1}(iq\pi)^{p-m}\sum_{r=0}^{m-1}\binom{p}{r}(-1)^{m-r} \\
&= (-1)^q(iq\pi)^{p-m}(-1)^{m-1}\sum_{r=0}^{m-1}\binom{p}{r}(-1)^r \\
&= (-1)^q(iq\pi)^{p-m}\binom{p-1}{m-1} = \alpha_p
\end{aligned}$$

Substituting in the sum $k = l + q$ we obtain

$$\begin{aligned}
\frac{d^p g(x)}{dx^p} &= \alpha_p + \binom{p}{m}(iq\pi)^{p-m}e^{iq\pi x} + \\
&+ \sum_{\substack{k=-\infty \\ k\neq 0,q}}^{\infty}(-1)^{k-q-1}(iq\pi)^{p-m}\left(\sum_{r=0}^{m-1}\binom{p}{r}\left(\frac{q}{k-q}\right)^{m-r}\right)e^{ik\pi x}
\end{aligned}$$

Therefore

$$\begin{aligned}
\beta_{mqk} &= \frac{(-1)^{k-q-1}(iq\pi)^{p-m}}{(ik\pi)^p}\sum_{r=0}^{m-1}\binom{p}{r}\left(\frac{q}{k-q}\right)^{m-r} \\
&= \frac{(-1)^{k-q-1}q^{p-m}}{k^p(i\pi)^m}\sum_{r=0}^{m-1}\binom{p}{r}\left(\frac{q}{k-q}\right)^{m-r}, \quad k \neq 0, q \qquad (5.14) \\
\beta_{mqq} &= \binom{p}{m}\frac{(iq\pi)^{p-m}}{(iq\pi)^p} = \binom{p}{m}(iq\pi)^{-m}
\end{aligned}$$

Then the mixed product is

$$\Sigma_{se}=\sum_{j=1}^{p}\left(\sum_{m=0}^{p}\sum_{\substack{q=-N\\q\neq 0}}^{N}a_{1m}b_{2q}\alpha_{mqj}\right)s_j(x)+\sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\left(\sum_{m=0}^{p}\sum_{\substack{q=-N\\q\neq 0}}^{N}a_{1m}b_{2q}\beta_{mqk}\right)e^{ik\pi x}$$
$$+\sum_{j=1}^{p}\left(\sum_{m=0}^{p}\sum_{\substack{q=-N\\q\neq 0}}^{N}a_{2m}b_{1q}\alpha_{mqj}\right)s_j(x)+\sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\left(\sum_{m=0}^{p}\sum_{\substack{q=-N\\q\neq 0}}^{N}a_{2m}b_{1q}\beta_{mqk}\right)e^{ik\pi x}$$

Let matrices $M_j^{(\alpha)}$, $j=1,\ldots,p$ and $M_k^{(\beta)}$, $k=0,\pm1,\pm2,\ldots$ of type $(p+1,2N+1)$ be defined by

$$\left(M_j^{(\alpha)}\right)_{mq}=\alpha_{mqj}\quad,\quad\left(M_j^{(\beta)}\right)_{mq}=\beta_{mqj}$$

where $m=0,1,\ldots,p$, $q=-N,-N+1,\ldots,N$. Note that for convenience in the matrix multiplication the rows of the above matrices are indexed from $-N$ to N as the components of b_1 and b_2 while their columns are indexed from 0 to p as the components of a_1 and a_2. Now the mixed product can be written in the following matrix form:

$$\begin{aligned}\Sigma_{se} &= \sum_{j=1}^{p}\left(a_1^TM_j^{(\alpha)}b_2\right)s_j(x)+\sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\left(a_1^TM_k^{(\beta)}b_2\right)e^{ik\pi x}\\ &+ \sum_{j=1}^{p}\left(a_2^TM_j^{(\alpha)}b_1\right)s_j(x)+\sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\left(a_2^TM_k^{(\beta)}b_1\right)e^{ik\pi x}\end{aligned}$$

or

$$\Sigma_{se}=\sum_{j=1}^{p}\left(a_1^TM_j^{(\alpha)}b_2+a_2^TM_j^{(\alpha)}b_1\right)s_j(x)+\sum_{\substack{k=-\infty\\k\neq 0}}^{\infty}\left(a_1^TM_k^{(\beta)}b_2+a_2^TM_k^{(\beta)}b_1\right)e^{ik\pi x}$$

Adding Σ_{ss}, Σ_{ee} and Σ_{se} we obtain the following Spline-Fourier expansion of f_1f_2:

$$\begin{aligned}f_1(x)f_2(x) &= (a_{10}a_{20}+b_1^TE_0b_2)s_0(x)\\ &+ \sum_{j=1}^{p}\left(a_1^TM_j^{(s)}a_2+a_1^TM_j^{(\alpha)}b_2+a_2^TM_j^{(\alpha)}b_1\right)s_j(x)\\ &+ \sum_{\substack{k=-2N\\k\neq 0}}^{2N}\left(a_1^T\hat{M}_k^{(s)}a_2+b_1^TE_kb_2+a_1^TM_k^{(\beta)}b_2+a_2^TM_k^{(\beta)}b_1\right)e^{ik\pi x}\\ &+ \sum_{|k|>2N}\left(a_1^T\hat{M}_k^{(s)}a_2+a_1^TM_k^{(\beta)}b_2+a_2^TM_k^{(\beta)}b_1\right)e^{ik\pi x}\end{aligned}$$

Therefore

$$f_1(x)\boxed{\times}f_2(x) = (a_{10}a_{20}+b_1^TE_0b_2)s_0(x)$$

$$+ \sum_{j=1}^{p} \left(a_1^T M_j^{(s)} a_2 + a_1^T M_j^{(\alpha)} b_2 + a_2^T M_j^{(\alpha)} b_1 \right) s_j(x)$$

$$+ \sum_{\substack{k=-N \\ k\neq 0}}^{N} \left(a_1^T \hat{M}_k^{(s)} a_2 + b_1^T E_k b_2 + a_1^T M_k^{(\beta)} b_2 + a_2^T M_k^{(\beta)} b_1 \right) e^{ik\pi x}$$

If M is a matrix, denote by $|M|$ a matrix of the same type with entries equal to the modules of the corresponding entries of M. Using this notation, the rounding error can be estimated as follows

$$\begin{aligned} |Error| \;\leq\; & 2 \sum_{k=N+1}^{2N} \left| a_1^T \hat{M}_k^{(s)} a_2 + b_1^T E_k b_2 + a_1^T M_k^{(\beta)} b_2 + a_2^T M_k^{(\beta)} b_1 \right| \\ & + \; |a_1|^T \sum_{|k|>2N} \left| \hat{M}_k^{(s)} \right| |a_2| \\ & + \; |a_1|^T \sum_{|k|>2N} \left| M_k^{(\beta)} \right| |b_2| \; + \; |a_2|^T \sum_{|k|>2N} \left| M_k^{(\beta)} \right| |b_1| \end{aligned}$$

For the infinite sums in the above expression we have

$$\begin{aligned} \sum_{|k|>2N} \left| \hat{M}_k^{(s)} \right| \;&\leq\; \sum_{j=p+1}^{2p} \sum_{k=2N+1}^{\infty} \frac{1}{(k\pi)^j} \left| M_j^{(s)} \right| \\ &\leq\; \sum_{j=p+1}^{2p} \frac{1}{(j-1)\pi^j (2N)^{j-1}} \left| M_j^{(s)} \right| \end{aligned}$$

Therefore

$$\sum_{|k|>2N} \left| \hat{M}_k^{(s)} \right| \;\leq\; \overline{M}^{(s)}$$

where

$$\overline{M}^{(s)} = \sum_{j=p+1}^{2p} \frac{1}{(j-1)\pi^j (2N)^{j-1}} \left| M_j^{(s)} \right|$$

We also have

$$\begin{aligned} \sum_{|k|>2N} |\beta_{mqk}| &\leq \sum_{k=2N+1}^{\infty} \frac{|q|^{p-m}}{k^p \pi^m} \sum_{r=0}^{m-1} \binom{p}{r} \left(\left(\frac{|q|}{2N+q} \right)^{m-r} + \left(\frac{|q|}{2N-q} \right)^{m-r} \right) \\ &\leq \frac{|q|^{p-m}}{(p-1)\pi^m (2N)^{p-1}} \sum_{r=0}^{m-1} \binom{p}{r} \left(\left(\frac{|q|}{2N+q} \right)^{m-r} + \left(\frac{|q|}{2N-q} \right)^{m-r} \right) \\ & \qquad\qquad \text{for } \; m \geq 1 \end{aligned}$$

$$\sum_{|k|>2N} |\beta_{0qk}| \;=\; 0$$

Therefore

$$\sum_{|k|>2N} \left| M_k^{(\beta)} \right| \leq \overline{M}^{(\beta)}$$

where

$$\left(\overline{M}^{(\beta)}\right)_{0q} = 0$$

$$\left(\overline{M}^{(\beta)}\right)_{mq} = \frac{|q|^{p-m}}{(p-1)\pi^m (2N)^{p-1}} \sum_{r=0}^{m-1} \binom{p}{r} \left(\left(\frac{|q|}{2N+q}\right)^{m-r} + \left(\frac{|q|}{2N-q}\right)^{m-r} \right)$$

for $m \neq 0$

Hence

$$\begin{aligned} |Error| \leq\ & 2 \sum_{k=N+1}^{2N} \left| a_1^T \hat{M}_k^{(s)} a_2 + b_1^T E_k b_2 + a_1^T M_k^{(\beta)} b_2 + a_2^T M_k^{(\beta)} b_1 \right| \\ & + |a_1|^T \overline{M}^{(s)} |a_2| + |a_1|^T \overline{M}^{(\beta)} |b_2| + |a_2|^T \overline{M}^{(\beta)} |b_1| \end{aligned}$$

Note that matrices $M_j^{(s)}$, $j = 1, \ldots, p$; $M_k^{(\alpha)}$, $M_k^{(\beta)}$, $\hat{M}_k^{(s)}$, $k = \pm 1, \ldots, \pm 2N$ and $\overline{M}^{(s)}$, $\overline{M}^{(\beta)}$ used in the evaluation of the product of f_1 and f_2 and in the estimation of the error do not depend on f_1 and f_2, so that they can be calculated in advance and used in all products in a certain numerical procedure.

Integration. Let $f \in \mathcal{M}$ and let

$$f(x) = \sum_{j=1}^{p} a_j s_j(x) + \sum_{\substack{k=-N \\ k\neq 0}}^{N} b_k e^{ik\pi x}$$

Then

$$\begin{aligned} \int f(x)dx &= \sum_{j=1}^{p} a_j s_{j+1}(x) + \sum_{\substack{k=-N \\ k\neq 0}}^{N} \frac{b_k}{ik\pi} e^{ik\pi x} \\ &= \sum_{j=2}^{p} a_{j-1} s_j(x) + \sum_{\substack{k=-N \\ k\neq 0}}^{N} \left(\frac{b_k}{ik\pi} - \frac{(-1)^k a_p}{(ik\pi)^{p+1}} \right) e^{ik\pi x} + \sum_{|k|>N} \frac{(-1)^{k-1}}{(ik\pi)^{p+1}} e^{ik\pi x} \end{aligned}$$

Therefore

$$\boxed{\int} f(x)dx = \sum_{j=2}^{p} a_{j-1} s_j(x) + \sum_{\substack{k=-N \\ k\neq 0}}^{N} \left(\frac{b_k}{ik\pi} - \frac{(-1)^k a_p}{(ik\pi)^{p+1}} \right) e^{ik\pi x}$$

with rounding error

$$|Error| = \left| \sum_{|k|>N} \frac{(-1)^{k-1}}{(ik\pi)^{p+1}} e^{ik\pi x} \right| \leq 2|a_p| \sum_{|k|>N} \frac{1}{(k\pi)^{p+1}} \leq \frac{2|a_p|}{p\pi^{p+1}N^p}$$

The integration of $s_0(x)$, when it arises in a particular practical problem, must be handled with special care because $\int s_0(x)dx = s_1(x)$ is true only in $(-1,1)$.

Remark. (*Subtraction and Division*) While subtraction is easy to implement by $f_1 - f_2 = f_1 + (-1)f_2$, division could be complicated. It is defined only if the divisor does not change its sign and then, the quotient $g = f_1 \div f_2$ can be obtained as a solution to $gf_2 = f_1$.

5.6 Approximations of functions with multiple discontinuities

The set $\mathcal{M}$ defined in the beginning of this chapter contains functions that may be discontinuous at the points $x = 2k+1$, $k \in Z$, i.e. one discontinuity in an interval of length 2 is allowed. However, using the same approach an expansion of the type (5.6) can be derived also in the case when the functions have more then one discontinuity or the discontinuity is not at $x = 2k+1$.

Let f and its first $p-1$ derivatives be allowed to be discontinuous at $x = 2k+1-c_l$, $k \in Z$, $l = 1,\ldots,\bar{l}$ (but not at any other points). Then f can be represented in the form

$$f(x) = a_0 + \sum_{l=1}^{\bar{l}} \sum_{j=1}^{p} a_{jl} s_j(x + c_l) + \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} b_k e^{ik\pi x}$$

where

$$\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} b_k e^{ik\pi x} \in H^p_{per}(-1,1)$$

The coefficients are obtained similarly to the coefficients in (5.6):

$$\begin{aligned} a_0 &= \frac{1}{2}\int_{-1}^{1} f(x)dx \\ a_{jl} &= \frac{1}{2}\left(\frac{d^{j-1}}{dx^{j-1}}(1 + c_l - 0) - \frac{d^{j-1}}{dx^{j-1}}(1 + c_l + 0)\right), \quad \begin{matrix} j = 1,\ldots,p \\ l = 1,\ldots,\bar{l} \end{matrix} \\ b_k &= \frac{1}{2(ik\pi)^p}\int_{-1}^{1} \frac{d^p f(x)}{dx^p} e^{-ik\pi x} dx\ , \ k = \pm 1, \pm 2, \ldots \end{aligned}$$

The above statement generalizes theorem 5.2 and the proof is conducted in a similar way.

Function f is approximated by

$$\rho_{Np}(f;x) = a_0 + \sum_{l=1}^{\bar{l}} \sum_{j=1}^{p} a_{jl} s_j(x + c_l) + \sum_{\substack{k=-N \\ k\neq 0}}^{N} b_k e^{ik\pi x} \tag{5.15}$$

with a rounding error

$$
\begin{aligned}
|f(x) - \rho_{Np}(f;x)| &= \left| \sum_{|k|>N} b_k e^{ik\pi x} \right| \leq \sum_{|k|>N} |b_k| \\
&\leq \left(\sum_{|k|>N} (k\pi)^{2p} |b_k|^2 \right)^{\frac{1}{2}} \left(\sum_{|k|>N} \frac{1}{(k\pi)^{2p}} \right)^{\frac{1}{2}} \\
&\leq \left(\frac{1}{2} \int_{-1}^{1} \left(\frac{d^p f(x)}{dx^p} \right)^2 dx - \sum_{l=1}^{\bar{l}} a_{pl}^2 - \sum_{\substack{k=-N \\ k\neq 0}}^{N} (k\pi)^{2p} |b_k|^2 \right)^{\frac{1}{2}} \left(\frac{2}{(2p-1)\pi^{2p} N^{2p-1}} \right)^{\frac{1}{2}} \\
&= o\left(\frac{1}{N^{p-\frac{1}{2}}} \right)
\end{aligned}
$$

Addition and multiplication by a number of expressions of the form (5.15) is obviously not a problem. Integration is performed without any serious difficulties. Let

$$
f(x) = \rho_{Np}(f;x) = \sum_{l=1}^{\bar{l}} \sum_{j=1}^{p} a_{jl} s_j(x + c_l) + \sum_{\substack{k=-N \\ k\neq 0}}^{N} b_k e^{ik\pi x}
$$

Then

$$
\begin{aligned}
\int f(x)dx &= \sum_{l=1}^{\bar{l}} \sum_{j=1}^{p} a_{jl} s_{j+1}(x + c_l) + \sum_{\substack{k=-N \\ k\neq 0}}^{N} \frac{b_k}{ik\pi} e^{ik\pi x} \\
&= \sum_{l=1}^{\bar{l}} \sum_{j=2}^{p} a_{j-1\,l} s_j(x + c_l) + \sum_{\substack{k=-N \\ k\neq 0}}^{N} \left(\frac{b_k}{ik\pi} - (-1)^k \sum_{l=1}^{\bar{l}} \frac{a_{pl} e^{ik\pi c_l}}{(ik\pi)^{p+1}} \right) e^{ik\pi x} \\
&\quad - \sum_{\substack{k=-N \\ k\neq 0}}^{N} \sum_{l=1}^{\bar{l}} \frac{(-1)^k a_{pl}}{(ik\pi)^{p+1}} e^{ik\pi(x+c_l)}
\end{aligned}
$$

Therefore

$$
\int f(x)dx = \sum_{l=1}^{\bar{l}} \sum_{j=2}^{p} a_{j-1\,l} s_j(x + c_l) + \sum_{\substack{k=-N \\ k\neq 0}}^{N} \left(\frac{b_k}{ik\pi} - (-1)^k \sum_{l=1}^{\bar{l}} \frac{a_{pl} e^{ik\pi c_l}}{(ik\pi)^{p+1}} \right) e^{ik\pi x}
$$

where the rounding error is

$$
|Error| \leq \sum_{|k|>N} \frac{1}{(k\pi)^{p+1}} \sum_{l=1}^{\bar{l}} |a_{pl}| \leq \frac{2}{p\pi^{p+1} N^p} \sum_{l=1}^{\bar{l}} |a_{pl}|
$$

With more points of discontinuity, multiplication becomes technically more complicated to implement. Using

$$\begin{aligned} s_m(x+c_{l_1})s_q(x+c_{l_2}) &= \sum_{j=m}^{2p}\binom{j-1}{m-1}s_{m+q-j}(1-c_{l_1}+c_{l_2})s_j(x+c_{l_1}) \\ &+ \sum_{j=q}^{2p}\binom{j-1}{m-1}s_{m+q-j}(1+c_{l_1}-c_{l_2})s_j(x+c_{l_2}) \end{aligned}$$

a product of the form

$$\left(\sum_{m=1}^{p} s_m(x+c_{l_1})\right)\left(\sum_{q=1}^{p} s_q(x+c_{l_2})\right)$$

can be presented as

$$\begin{aligned} \left(\sum_{m=1}^{p} s_m(x+c_{l_1})\right)\left(\sum_{q=1}^{p} s_q(x+c_{l_2})\right) &= \sum_{j=1}^{2p}\left(a_{l_1}^T\tilde{M}_j(c_{l_2}-c_{l_1})a_{l_2}\right)s_j(x+c_{l_1}) \\ &+\sum_{j=1}^{2p}\left(a_{l_2}^T\tilde{M}_j(c_{l_1}-c_{l_2})a_{l_1}\right)s_j(x+c_{l_2}) \end{aligned}$$

where

$$a_{l_1}=(a_{1l_1},\ldots,a_{pl_1})^T \quad , \quad a_{l_2}=(a_{1l_2},\ldots,a_{pl_2})^T$$

$$\left(\tilde{M}_j(y)\right)_{mq}=\begin{cases}\binom{j-1}{m-1}s_{m+q-j}(1-y)\,, & m,q=1,\ldots,p \quad m\le j\le m+q \\ 0 & \text{otherwise}\end{cases}$$

and be processed further as Σ_{ss} in the previous section.

Using the substitution $y=x+c_l$, a mixed product of the form

$$\left(\sum_{m=1}^{p} a_{ml}s_m(x+c_l)\right)\left(\sum_{\substack{q=-N\\ q\neq 0}}^{N} b_q e^{iq\pi x}\right)$$

can be represented as

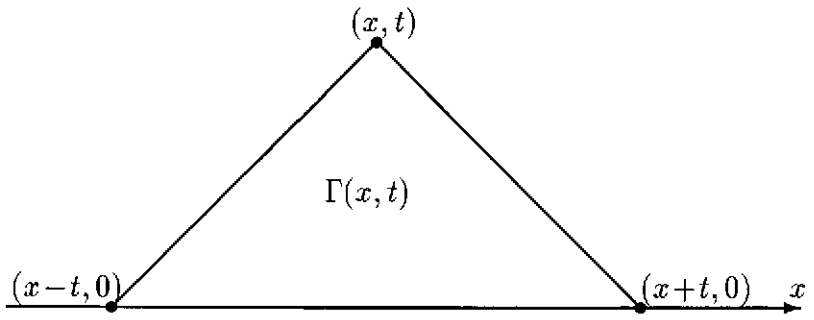
$$\left(\sum_{m=1}^{p} a_{ml}s_m(y)\right)\left(\sum_{\substack{q=-N\\ q\neq 0}}^{N} b_q e^{-iq\pi c_l}e^{iq\pi y}\right)$$

and be dealt with as Σ_{se} in the previous section.

Therefore, the process of multiplication, although technically complicated, can be easily algorithmized and implemented on a computer. Note also that p usually does not need to be very large, e.g. $p=5$ will be probably satisfactory for most of the problems.

5.7 Integral Form of the Wave Equation.

In this section we will consider the wave equation (4.4)–(4.5) when some of the functions g_1, g_2, f or their derivatives have discontinuities. In this case the solution may also be discontinuous and may not be defined by the equation in its standard form (4.4)–(4.5). We will transform this equation in an integral form. Let $\Gamma(x,t)$ be a triangle with vertices $(x-t,0)$, $(x+t,0)$ and (x,t).



Using Green's theorem we have

$$\begin{aligned}
\iint\limits_{\Gamma(x,t)} f(y,\theta,u(y,\theta))\mathrm{d}y\mathrm{d}\theta &= \iint\limits_{\Gamma(x,t)} (u_{tt}(y,\theta) - u_{xx}(y,\theta))\mathrm{d}y\mathrm{d}\theta \\
&= \oint\limits_{\partial\Gamma(x,t)} (-u_x(y,\theta)\mathrm{d}\theta - u_t(y,\theta)\mathrm{d}y \\
&= 2u(x,t) - u(x-t,0) - u(x+t,0) - \int\limits_{x-t}^{x+t} u_t(y,0)\mathrm{d}y \\
&= 2u(x,t) - g_1(x-t) - g_1(x+t) - \int\limits_{x-t}^{x+t} g_2(y)\mathrm{d}y
\end{aligned}$$

Therefore, we have

$$u(x,t) = \frac{1}{2}\iint\limits_{\Gamma(x,t)} f(y,\theta,u(y,\theta))\mathrm{d}y\mathrm{d}\theta + g(x,t) \tag{5.16}$$

where

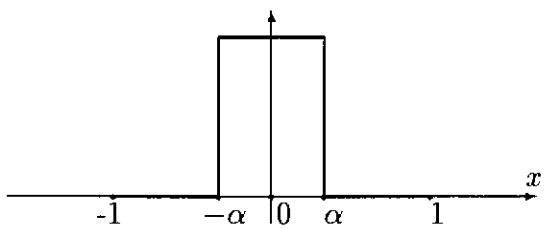
$$g(x,t) = \frac{1}{2}\left(g_1(x+t) + g_1(x-t) + \int_{x-t}^{x+t} g_2(y)\mathrm{d}y\right)$$

We consider problem (4.4)–(4.5) in the form (5.16) which defines a solution also in the case when some of the functions are discontinuous. We use periodic splines to represent

functions with discontinuities. In the following examples some functions that are often considered as initial conditions are presented.

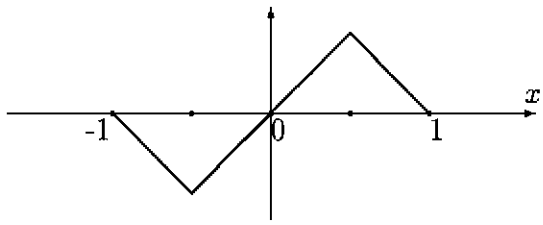
Examples

1. $\phi(x) = \begin{cases} 1 & x \in (-\alpha, \alpha) \\ 0 & x \in (-1, -\alpha) \mathrm{U} (\alpha, 1) \end{cases}$



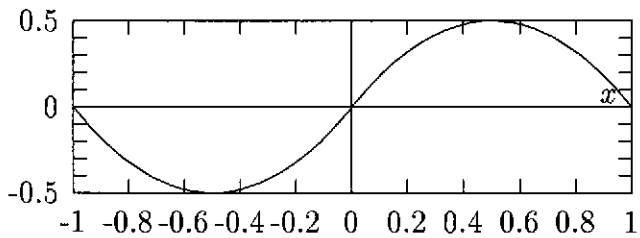
$\phi(x) = \alpha + \frac{1}{2} s_1(x+1-\alpha) - \frac{1}{2} s_1(x-1+\alpha)$

2. $\phi(x) = \begin{cases} -1 - x & x \in (-1, -0.5) \\ x & x \in (-0.5, 0.5) \\ 1 - x & x \in (0.5, 1) \end{cases}$



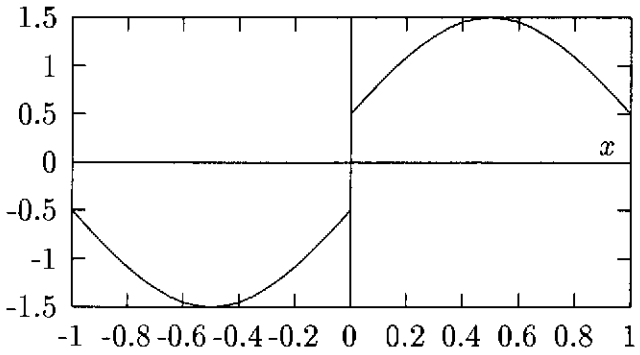
$\phi(x) = s_2(x + 0.5) - s_2(x - 0.5)$

3. $\phi(x) = \begin{cases} 2x(1 + x) & x \in (-1, 0) \\ 2x(1 - x) & x \in (0, 1) \end{cases}$



$\phi(x) = 4s_3(x + 1) - 4s_3(x)$

4. $\phi(x) = \begin{cases} -0.5 + \sin(\pi x) & x \in (-1,0) \\ 0.5 + \sin(\pi x) & x \in (0,1) \end{cases}$



$\phi(x) = \frac{1}{2}s_1(x) - \frac{1}{2}s_1(x+1) + \sin(\pi x)$

5.8 General Outline of the Method

Numerical solution of the wave equation is sought in the form

$$u(x,t) = a_0(t) + \sum_{m=1}^{M}\sum_{j=1}^{p}\sum_{\delta=-1}^{1} a_{mj\delta}(t)s_j(x+\delta t+\alpha'_m) + \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} b_k(t)e^{ik\pi x} \tag{5.17}$$

where $\alpha'_m = 1 - \alpha_m$ and α_m, $m = 1,\ldots,M$ are points in $(-1,1]$ where the data functions or some of their first $p-1$ derivatives may be discontinuous.

The following Newton Type iterative procedure is applied

$$u^{(l+1)}(x,t) = (1-\lambda)u^{(l)}(x,t) + \lambda\left(\frac{1}{2}\iint_{\Gamma(x,t)} f(y,\theta,u^{(l)}(y,\theta))\mathrm{d}y\mathrm{d}\theta + g(x,t)\right)$$

The essential part of each iteration is the evaluation of the integral. This can be done successfully for a number of functions f using the arithmetic in the Spline-Fourier functoid. For example, if $f(x,t,u) = c(t)u^p + \phi(x,t)$ we can obtain an expansion of $f(x,t,u^{(l)}(x,t))$ in the form 5.17 and then integrate. It is particularly easy to do that in the case of linear equations. We also have to chose some form of representation of

the coefficients $a_{mj}(t)$, $b_k(t)$. In this chapter we carry out the computations representing those coefficients as polynomials of t. Assuming that f can be represented in the form

$$f(x,t,u^{(l)}(x,t)) = c_0(t) + \sum_{m=1}^{M}\sum_{j=1}^{p}\sum_{\delta=-1}^{1} c_{mj\delta}(t)s_j(x+\delta t+\alpha_m^l) + \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} d_k(t)e^{ik\pi x}$$

where $c_{mj\delta}(t)$ and $d_k(t)$ are polynomials of t, in each iteration we have to integrate terms of the form

$$\frac{\theta^q}{q!}e^{ik\pi y} \quad \text{and} \quad \frac{\theta^q}{q!}s_j(y+\delta t+\alpha)$$

These terms can be integrated as follows:

$$\begin{aligned}
\int\int_{\Gamma(x,t)} \frac{\theta^q}{q!}s_j(y)\mathrm{d}y\mathrm{d}\theta &= \int_0^t \int_{x-t+\theta}^{x+t-\theta} \frac{\theta^q}{q!}s_j(y)\mathrm{d}y\mathrm{d}\theta \\
&= \int_0^t \frac{\theta^q}{q!}\left(s_{j+1}(x+t-\theta) - s_{j+1}(x-t+\theta)\right)\mathrm{d}\theta \\
&= \left[-\sum_{l=0}^{q}\frac{\theta^{q-l}}{(q-l)!}\left(s_{j+l+2}(x+t-\theta) + (-1)^l s_{j+l+2}(x-t+\theta)\right)\right]_0^t \\
&= s_{j+q+2}(x+t) + (-1)^q s_{j+q+2}(x-t) - \sum_{l=0}^{q}\frac{(1+(-1)^l)t^{q-l}}{(q-l)!}s_{j+l+2}(x) \\
&= s_{j+q+2}(x+t) + (-1)^q s_{j+q+2}(x-t) - \sum_{l=0}^{q}(1+(-1)^{q-l})\frac{t^l}{l!}s_{j+q-l+2}(x)
\end{aligned}$$

$$\begin{aligned}
\int\int_{\Gamma(x,t)} \frac{\theta^q}{q!}s_j(y+\theta)\mathrm{d}y\mathrm{d}\theta &= \int_0^t \int_{x-t+\theta}^{x+t-\theta} \frac{\theta^q}{q!}s_j(y+\theta)\mathrm{d}y\mathrm{d}\theta \\
&= \int_0^t \left(\frac{\theta^q}{q!}s_{j+1}(x+t) - \frac{\theta^q}{q!}s_{j+1}(x-t+2\theta)\right)\mathrm{d}\theta \\
&= \left[\frac{\theta^{q+1}}{(q+1)!}s_{j+1}(x+t) + \sum_{l=1}^{q+1}\left(-\frac{1}{2}\right)^l \frac{\theta^{q+1-l}}{(q+1-l)!}s_{j+l+1}(x-t+2\theta)\right]_0^t \\
&= \left(-\frac{1}{2}\right)^{q+1} s_{j+q+2}(x+t) - \left(-\frac{1}{2}\right)^{q+1} s_{j+q+2}(x-t) + \sum_{l=0}^{q}\left(-\frac{1}{2}\right)^l \frac{t^{q+1-l}}{(q+1-l)!}s_{j+l+1}(x+t) \\
&= \left(-\frac{1}{2}\right)^{q+1} s_{j+q+2}(x+t) - \left(-\frac{1}{2}\right)^{q+1} s_{j+q+2}(x-t) + \sum_{l=1}^{q+1}\left(-\frac{1}{2}\right)^{q-l+1} \frac{t^l}{l!}s_{j+q-l+2}(x+t)
\end{aligned}$$

$$\int_{\Gamma(x,t)}\int \frac{\theta^q}{q!}s_j(y-\theta)\mathrm{d}y\mathrm{d}\theta = \int_0^t \int_{x-t+\theta}^{x+t-\theta} \frac{\theta^q}{q!}s_j(y-\theta)\mathrm{d}y\mathrm{d}\theta$$
$$= \int_0^t \left(\frac{\theta^q}{q!}s_{j+1}(x+t-2\theta) - \frac{\theta^q}{q!}s_{j+1}(x-t)\right)\mathrm{d}\theta$$
$$= -\left[\frac{\theta^{q+1}}{(q+1)!}s_{j+1}(x-t) + \sum_{l=1}^{q+1}\left(\frac{1}{2}\right)^l \frac{\theta^{q+1-l}}{(q+1-l)!}s_{j+l+1}(x+t-2\theta)\right]_0^t$$
$$= \left(\frac{1}{2}\right)^{q+1} s_{j+q+2}(x+t) - \left(\frac{1}{2}\right)^{q+1} s_{j+q+2}(x-t) - \sum_{l=0}^{q}\left(\frac{1}{2}\right)^l \frac{t^{q+1-l}}{(q+1-l)!}s_{j+l+1}(x-t)$$
$$= \left(\frac{1}{2}\right)^{q+1} s_{j+q+2}(x+t) - \left(\frac{1}{2}\right)^{q+1} s_{j+q+2}(x-t) - \sum_{l=1}^{q+1}\left(\frac{1}{2}\right)^{q-l+1} \frac{t^l}{l!}s_{j+q-l+2}(x-t)$$

$$\int_{\Gamma(x,t)}\int \frac{\theta^q}{q!}e^{ik\pi y}\mathrm{d}y\mathrm{d}\theta = \int_0^t \int_{x-t+\theta}^{x+t-\theta} \frac{\theta^q}{q!}e^{ik\pi y}\mathrm{d}y\mathrm{d}\theta$$
$$= \int_0^t \frac{\theta^q}{q!}\left(\frac{e^{ik\pi(x+t-\theta)}}{ik\pi} - \frac{e^{ik\pi(x-t+\theta)}}{ik\pi}\right)\mathrm{d}\theta$$
$$= \left[-\sum_{l=0}^{q}\frac{\theta^{q-l}}{(q-l)!}\left(\frac{e^{ik\pi(x+t-\theta)}}{(ik\pi)^{l+2}} + (-1)^l\frac{e^{ik\pi(x-t+\theta)}}{(ik\pi)^{l+2}}\right)\right]_0^t$$
$$= \frac{e^{ik\pi(x+t)}}{(ik\pi)^{q+2}} + \frac{e^{ik\pi(x-t)}}{(-ik\pi)^{q+2}} - \sum_{l=0}^{q}\frac{(1+(-1)^l)}{(ik\pi)^{l+2}}\frac{t^{q-l}}{(q-l)!}e^{ik\pi x}$$
$$= \frac{e^{ik\pi x}}{(ik\pi)^{q+2}}\left(e^{ik\pi t} + (-1)^q e^{-ik\pi t} - \sum_{l=0}^{q}\left(\frac{(ik\pi t)^l}{l!} + (-1)^q\frac{(-ik\pi t)^l}{l!}\right)\right)$$
$$= \sum_{l=q+2}^{\infty}(1+(-1)^{l-q})(ik\pi)^{l-q-2}\frac{t^l}{l!}e^{ik\pi x}$$

Then for the corresponding sums we have

$$\int_{\Gamma(x,t)}\int \sum_{j=1}^{p}\sum_{q=0}^{\bar{q}}\alpha_{jq0}\frac{\theta^q}{q!}s_j(y)\mathrm{d}y\mathrm{d}\theta$$
$$= \sum_{j=1}^{p}\sum_{q=0}^{\bar{q}}\alpha_{jq0}s_{j+q+2}(x+t) + \sum_{j=1}^{p}\sum_{q=0}^{\bar{q}}\alpha_{jq0}(-1)^q s_{j+q+2}(x-t)$$

$$
\begin{aligned}
&\quad -\sum_{j=1}^{p}\sum_{q=0}^{\bar{q}}\sum_{l=0}^{q}(1+(-1)^{q-l})\alpha_{jq0}\frac{t^l}{l!}s_{j+q-l+2}(x) \\
&= \sum_{j=1}^{p+\bar{q}}\left(\sum_{q=0}^{\min\{\bar{q},j-1\}}\alpha_{j-q\,q\,0}\right)s_{j+2}(x+t)+\sum_{j=1}^{p+\bar{q}}\left(\sum_{q=0}^{\min\{\bar{q},j-1}(-1)^q\alpha_{j-q\,q\,0}\right)s_{j+2}(x-t) \\
&\quad -\sum_{j=1}^{p+\bar{q}}\sum_{q=0}^{\bar{q}}\left(\sum_{l=\max\{q,j+q-p\}}^{\min\{\bar{q},j+q-1\}}(1+(-1)^{l-q})\alpha_{j+q-l\,l\,0}\right)\frac{t^q}{q!}s_{j+2}(x)
\end{aligned}
$$

$$
\begin{aligned}
&\int\!\!\int_{\Gamma(x,t)}\sum_{j=1}^{p}\sum_{q=0}^{\bar{q}}\alpha_{jq1}\frac{\theta^q}{q!}s_j(y+\theta)\mathrm{d}y\mathrm{d}\theta \\
&= \sum_{j=1}^{p}\sum_{q=0}^{\bar{q}}\alpha_{jq1}\left(-\frac{1}{2}\right)^{q+1}s_{j+q+2}(x+t)-\sum_{j=1}^{p}\sum_{q=0}^{\bar{q}}\alpha_{jq1}\left(-\frac{1}{2}\right)^{q+1}s_{j+q+2}(x-t) \\
&\quad -\sum_{j=1}^{p}\sum_{q=0}^{\bar{q}}\sum_{l=1}^{q+1}\alpha_{jq1}\left(-\frac{1}{2}\right)^{q-l+1}\frac{t^l}{l!}s_{j+q-l+2}(x+t) \\
&= \sum_{j=1}^{p+\bar{q}}\left(\sum_{q=0}^{\min\{\bar{q},j-1\}}\left(-\frac{1}{2}\right)^{q+1}\alpha_{j-q\,q\,1}\right)s_{j+2}(x+t)-\sum_{j=1}^{p+\bar{q}}\left(\sum_{q=0}^{\min\{\bar{q},j-1\}}\left(-\frac{1}{2}\right)^{q+1}\alpha_{j-q\,q\,1}\right)s_{j+2}(x-t) \\
&\quad -\sum_{j=1}^{p+\bar{q}}\sum_{q=1}^{\bar{q}+1}\left(\sum_{l=\max\{q-1,j+q-p\}}^{\min\{\bar{q},j+q-1\}}\left(-\frac{1}{2}\right)^{l-q+1}\alpha_{j+q-l\,l\,1}\right)\frac{t^q}{q!}s_{j+2}(x+t)
\end{aligned}
$$

$$
\begin{aligned}
&\int\!\!\int_{\Gamma(x,t)}\sum_{j=1}^{p}\sum_{q=0}^{\bar{q}}\alpha_{j\,q\,-1}\frac{\theta^q}{q!}s_j(y-\theta)\mathrm{d}y\mathrm{d}\theta \\
&= \sum_{j=1}^{p}\sum_{q=0}^{\bar{q}}\alpha_{j\,q\,-1}\left(\frac{1}{2}\right)^{q+1}s_{j+q+2}(x+t)-\sum_{j=1}^{p}\sum_{q=0}^{\bar{q}}\alpha_{j\,q\,-1}\left(\frac{1}{2}\right)^{q+1}s_{j+q+2}(x-t) \\
&\quad -\sum_{j=1}^{p}\sum_{q=0}^{\bar{q}}\sum_{l=1}^{q+1}\alpha_{j\,q\,-1}\left(\frac{1}{2}\right)^{q-l+1}\frac{t^l}{l!}s_{j+q-l+2}(x-t) \\
&= \sum_{j=1}^{p+\bar{q}}\left(\sum_{q=0}^{\min\{\bar{q},j-1\}}\left(\frac{1}{2}\right)^{q+1}\alpha_{j-q\,q\,-1}\right)s_{j+2}(x+t)-\sum_{j=1}^{p+\bar{q}}\left(\sum_{q=0}^{\min\{\bar{q},j-1\}}\left(\frac{1}{2}\right)^{q+1}\alpha_{j-q\,q\,-1}\right)s_{j+2}(x-t) \\
&\quad -\sum_{j=1}^{p+\bar{q}}\sum_{q=1}^{\bar{q}+1}\left(\sum_{l=\max\{q-1,j+q-p\}}^{\min\{\bar{q},j+q-1\}}\left(\frac{1}{2}\right)^{l-q+1}\alpha_{j+q-l\,l\,-1}\right)\frac{t^q}{q!}s_{j+2}(x-t)
\end{aligned}
$$

The integration over $\Gamma(x,t)$ produces splines with larger indices. The rounding of the

splines s_j that are sufficiently smooth (e.g. $j \geq p$) is done as discussed in the previous chapter, i.e. they are replaced by their Fourier series and the error is $O(N^{1-j})$.

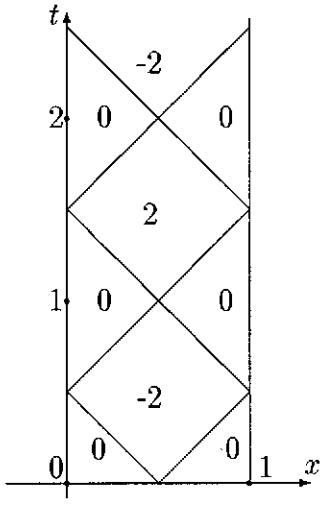
5.9 Numerical Examples.

The numerical results presented in this section are produced by a procedure implementing the method discussed in the previous section. Since the purpose of these examples is only to demonstrate the applicability and some advantages of the proposed Spline-Fourier approach, the procedure is implemented without directed roundings and produces only an approximate solution. Naturally, a procedure, producing validated enclosure can also be developed, but it will require more programming time and effort. We feel that, at this stage, this work should be preceded by further research on the Spline-Fourier functoid, its application to the wave equation and, maybe, the development of user-friendly software for computations in this functoid.

Example 5.1 We consider an initial boundary value problem for the equation:

$$u_{tt} - u_{xx} = \frac{2}{(t+3)^2}u + \pi^2(t+3)^2\sin(\pi x) + 2(t+3)\phi(x,t) \tag{5.18}$$

where $\phi(x,t)$ is a piece-wise constant function defined in $[0,1] \times [0,\infty)$ as shown on the sketch (to the right) and the boundary and initial conditions are given in the following form:

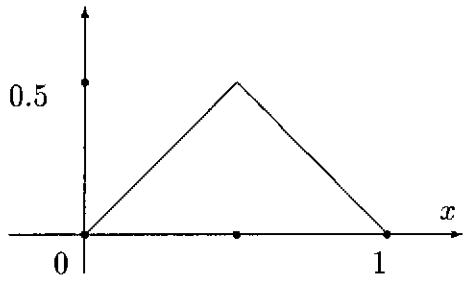


$$\begin{aligned} u(0,t) &= 0 \\ u(1,t) &= 0 \ , \ t \geq 0 \end{aligned}$$

$$\begin{aligned} u(x,0) &= 9\sin(\pi x) + 18\psi(x) \\ u_t(x,0) &= 6\sin(\pi x) + 12\psi(x) \ , \ 0 \leq x \leq 1 \end{aligned} \tag{5.19}$$

where

$$\psi(x) = \begin{cases} x & 0 \le x \le 0.5 \\ 1-x & 0.5 \le x \le 1 \end{cases}$$



Using the method described in section 2.5.2 and the periodic splines, problem (5.18), (5.19) can be written as a periodic (period 2) initial value problem of the form

$$\begin{aligned} u_{tt}-u_{xx} &= \frac{2}{(t+3)^2}u+\pi^2(t+3)^2\sin(\pi x)+2(t+3)(s_1(x+t+0.5) \\ &\quad -s_1(x-t+0.5)-s_1(x+t-0.5)+s_1(x-t-0.5)) \\ u(x,0) &= 9\sin(\pi x)+18(s_2(x+0.5)-s_2(x-0.5)) \\ u_t(x,0) &= 6\sin(\pi x)+12(s_2(x+0.5)-s_2(x-0.5)) \end{aligned}$$

The exact solution of this problem is

$$\begin{aligned} u(x,t) &= (t+3)^2(\sin(\pi x)+s_2(x+t+0.5)+s_2(x-t+0.5) \\ &\quad -s_2(x+t-0.5)-s_2(x-t-0.5)) . \end{aligned}$$

In table 5.1 the values of the numerical solution and the exact solution at some points are presented. A high accuracy of the numerical solution is obtained because the exact solution belongs to the Spline-Fourier functoid in which the computations are performed. All data functions in the problem, except for the coefficient of u, are exactly representable through the adopted data types. The only error in representing the problem on a computer is in representing the function $\frac{2}{(t+3)^2}$. However, it is represented by a Taylor polynomial of degree 10 and this error is very small.

Let us note that, in general, for problems with discontinuous data functions, we can not apply the method considered in chapter 4 because Fourier series of discontinuous functions are not uniformly convergent. In this particular example it is still possible to apply this method (at least theoretically) because the integral over the characteristic triangle $\Gamma(x,t)$ of function ϕ is a continuous function. However, the rate of convergence is only $o(N^{-\frac{1}{2}})$. This implies that a very large N is required. Due to the large number of computations involved and the accumulated rounding error we were not able to obtain any meaningful results, at least not on the PC on which all other numerical experiments are performed.

$x;t$	Numerical Solution $p=5,\ N=5$	Exact Solution
0.5; 0.5	1.2250E+01	1.2250E+01
0.5; 1.0	-7.5921E-11	0.0000
0.5; 1.5	2.0250E+01	2.0250E+01
0.5; 2.0	5.0000E+01	5.0000E+01
0.0; 1.0	-4.0128E-24	0.0000
0.0; 2.0	-5.3254E-21	0.00000

Table 5.1: *Problem (5.18), (5.19): Values of the numerical solution and the exact solution at various points* (x,t).

Example 5.2 We consider the problem

$$\begin{aligned} u_{tt} - u_{xx} &= \left(\pi^2 + \frac{2}{(t+3)^2}\right) u \\ u(x,0) &= 2(s_3(x+1) - s_3(x)) \\ u_t(x,0) &= 0 \end{aligned} \tag{5.20}$$

This problem is formulated directly as a periodic problem and the data functions, where necessary (in this case - in the initial condition), are expressed in terms of periodic splines. The equation is similar to (5.18) but the exact solution is not available. The method discussed in chapter 4 is also applicable to this problem because the function, prescribed as a value of $u(x,0)$, is a function in $H^2_{per}(-1,1)$. Since the exact solution is not known, the numerical solution produced by the method discussed in section 5.8 is compared with the validated numerical solution obtained by the method from chapter 4. Values at some points of both solutions are presented in table 5.2. Since both methods are applicable, the advantage of using the Spline-Fourier functoid and the method from the previous section is mainly in the smaller computational effort required to achieve similar accuracy.

$x;t$	Numerical Solution p=7, N=5	Verified Bounds for the Solution, N=40
0.0; 0.0	0.25000	0.2^{5007}_{4993}
0.5; 1.0	0.29016	0.2^{9018}_{8702}
0.5; 2.0	0.33727	0.3^{5932}_{1847}
0.0; 0.0	2.39E-13	$^{6.58}_{-6.58}$ E-05
0.0; 1.0	-4.79E-10	$^{8.62}_{-8.62}$ E-04
0.0; 2.0	-2.33E-10	$^{2.04}_{-2.04}$ E-02

Table 5.2: *Problem (5.20): Values of the numerical solution and a validated solution obtained by a different method at various points* (x, t).

Chapter 6

Conclusion

In the thesis we consider the following aspects of the construction of interval enclosures for the solutions of Initial Value Problems for

Ordinary Differential Equations:

- the wrapping effect and its implications for the convergence of interval enclosures produced by methods of propagate and wrap type;
- quantifying the wrapping effect;
- necessary and sufficient conditions for no wrapping effect;

Hyperbolic Partial Differential Equations:

- monotone properties of the periodic problem for the wave equation;
- using monotone properties in constructing interval enclosures for the solution (in the case of a point (noninterval) initial condition) or the set of solutions (in the case of an interval initial condition);
- constructing interval enclosures for the solutions of the wave equation in the Cartesian product of a Taylor functoid and a Fourier functoid;
- spline-Fourier approximations (Fourier hyper functoid);
- using Spline-Fourier approximations in representing and propagating discontinuities of the data function of the problem or their derivatives.

The main result with regard to the wrapping effect associated with the construction of interval enclosures of IVP for ODE concerns the convergence of the enclosures produced by methods of propagate and wrap type. We prove (chapter 3) that such enclosures converge to a wrapping function associated with the particular problem. This allows us to quantify the wrapping effect associated with the problem and consider the problems

with no wrapping effect as problems for which the wrapping function equals the optimal interval enclosure of the solutions. The significance of the results obtained in chapter 3 is in

- providing a better understanding of the wrapping effect and the behavior of the computed interval enclosures for a set of solutions;
- characterization of problems with no wrapping effect.

In [30], [78] it is stated that a complete set of tools for validated solving of IVP for ODE should include software for recognizing problems with quasi-isotone right-hand side and solving them by a straightforward procedure instead of using complicated algorithms [37], [81], [59]. In fact such software should recognize the larger class of problems with no wrapping effect as they are specified by the theorems proved in chapter 3.

The central concept in the study of the wrapping effect is the concept of wrapping function. This concept is not associated with a particular method but with the problem to be solved. We believe that this approach can be used in studying the wrapping effect for other IVP. The results in chapter 3 will be a methodological base for such research.

The starting point of our research on validated solutions of the wave equation is the monotone properties of the problem. This is also a new element of our research compared to the existing literature on validated solution of Hyperbolic PDEs. We established in chapter 4 conditions providing for the operator of the Periodic Initial Value Problem for the wave equation to be an operator of monotone type and we also establish monotone properties which facilitate a step-by-step construction of interval enclosures of the solutions. These properties provide a theoretical base for the design of validated methods suitable for both point (noninterval) and interval initial conditions. In support of this statement we propose a method which uses the established monotone properties. The bounds for the solution(s) are computed in the form of Fourier series of the space variable with coefficients which are polynomials of the time variable using a mesh in the time dimension. The implementation of the method requires computations in the Cartesian product of the Taylor functoid and the Fourier functoid. In addition to the roundings and operations discussed in [49] we introduced a new way of rounding the data functions as well as integration over the characteristic triangle. The quality of the enclosures is demonstrated in numerical examples.

Discontinuities are very common particularly in the periodic formulation of the IVP for the wave equation. In order to be able to deal with discontinuities of the data functions we considered in chapter 5 Spline-Fourier approximations which are in fact approximations in the Fourier hyper functoid [48]. We feel that, at least for the problems considered, the explicit use of splines (instead of their Fourier series) simplifies the computations. The proposed approach enlarges the area of applicability of the method discussed in chapter 4 with problems having discontinuous data functions and reduces the computational effort in the case when only derivatives of the data functions are discontinuous. This is also demonstrated in examples.

We think that the results presented in the thesis provide a foundation for future research in the following areas:

- Studying the wrapping effect in the validated solution of PDE using the concept of wrapping function.
- Monotone properties of the wave equation with multidimensional space variable. Using the approach in chapter 4 it is easy to find conditions for the operator of the problem to be an operator of monotone type. However, how to obtain monotone properties suitable for step-by-step construction of enclosures, is not obvious.
- Modification of the method using Spline-Fourier series so that it can be applied on a mesh in the time dimension. If it is applied in the present form the number of discontinuities we have to make provision for will increase at every step. A suitable criterion is to be found for an automatic elimination of some of them.
- Using Hausdorff approximations in computing enclosures of discontinuous functions. For example, we can show that the Fourier series of the jump function converges to its Hausdorff limit at a rate of $O(N^{-\frac{1}{2}})$ and enclosures with the same order of approximation can be constructed.

Bibliography

[1] Adams E, *The Reliability Question for Discretization of Evolution Problems, Part I: Theoretical Consideration of Failures, Part II: Practical Failures*, in Adams E, Kulisch U (eds), Scientific Computing with Automatic Result Verification, Academic Press, San Diego,1993, pp. 423-463, 465-526.

[2] Adams E, *Periodic Solutions: Enclosure, Verification and Applications*, in Computer Arithmetic and Self-Validating Numerical Methods, Ullrich C (ed), Academic Press, San Diego, 1993.

[3] Adams E, Ames W F, Kühn W, Rufeger W, Spreuer H, *Computational Chaos May Be Due to a Single Local Error*, Journal of Computational Physics, 104 (1993), pp. 241-250.

[4] Adams R A, Sobolev Spaces, Academic Press, New York, San Francisco, London, 1975.

[5] Alefeld G,Herzberger J, Introduction to Interval Computations, Academic Press, NY,1983.

[6] Aliprantis C D, Burkinshaw O, Locally Solid Riesz Spaces, Academic Press, New York, San Francisco, London, 1978.

[7] Anguelov R, Markov S, *Extended Segment Analysis*, Freiburger Intervall Berichte 81/10, University of Freiburg, 1981.

[8] Anguelov R., *On the Wrapping Effect*, Comptes Rendus - Bulgarian Academy of Science, Vol.40/8, 1987.

[9] Anguelov R, Markov S, *Wrapping Effect and Wrapping Function*, Reliable Computing 4(4), 1998.

[10] Anguelov R, *Spline-Fourier Approximations of Discontinuous Waves*, Journal of Universal Computer Science (J.UCS), Vol. 4 (2), 1998, pp.110-113.

[11] Anguelov R, *Wrapping Function of the Initial Value Problem for ODE: Applications*, Reliable Computing (to appear).

[12] Anguelov R, *Guaranteed Bounds for the Solution of the Wave Equation*, Quaestiones Mathematicae, Vol. 19 (1-2), 1996, pp. 275-289.

[13] Anguelov R, *Monotone Properties of Hyperbolic Differential Operators*, in Proceedings of the 19th Symposium on Numerical Mathematics, San Lameer, July 1993, University of Natal, 1993.

[14] Apostolatos N, *Allgemeine Intervallarithmetiken und Anwendungen*, Bull. Soc. Math. Grêce (N.S.), 10, 1969, pp. 136-180.

[15] Appelt W, Fehlereinschliessung bei der Numerischen Lösung Elliptischer Differentialgleichungen unter Verwendung eines Intervallarithmetischen Defektverfahrens, PhD Thesis, University of Bonn, 1972.

[16] Asano N., Kato Y, Algebraic and Spectral Methods for Nonlinear Wave Equations, Wiley, New York, 1990.

[17] Babkin B, *Numerical Integration of Systems of Ordinary Differential Equations of First Order (in Russian)*, Izvestija Akademii Nauk SSSR, Serija Mathematika, 18, 1954, pp. 477-484.

[18] Bauch H, *On the Iterative Inclusion of Solutions in Initial Value Problems for Ordinary Differential Equations*, Computing 22, 1979, pp. 339-354.

[19] Bauch H, Jahn K U, Oelschlagel D, Süsse H, Wiebigke V, Intervallmathematik: Theorie und Anwendungen, Teubner, Leibzig, 1987.

[20] Canuto C, Hussaini M Y, Quarteroni A, Zang T A, Spectral Methods in Fluid Dynamics, Springer-Verlag, New York, 1988.

[21] Chaplygin C, *Fundamentals of a New Method for Numerical Integration of Differential Equations, Moskow, 1919 (in Russian)*, in Sobranie Sochinenija I, Gostehizdat, Moskow, 1948, pp.348-368.

[22] Chaplygin C, *Numerical Integration of Differential Equations of First Order, Moskow, 1920 (in Russian)*, in Sobranie Sochinenija I, Gostehizdat, Moskow, 1948,pp. 402-419.

[23] Collatz, L., Functional Analysis and Numerical Mathematics, Springer, New York, 1964.

[24] Conn A R, Gould N I M, Toint P L, LANCELOT: A Fortran Package for Large-Scale Nonlinear Optimization (Release A), Springer Series in Computational Mathematics No.17, Springer-Verlag, Berlin, 1992.

[25] Connel A E, Corless R M, *An Experimental Interval Arithmetic Package in Maple*, Interval Computations, Vol.2, 1993, pp. 120-134.

[26] Corliss G F, *Comparing Software Packages for Interval Arithmetic*, Proceedings of SCAN'93, September 1993, Vienna.

[27] Corliss G F, *Guaranteed Error Bounds for Ordinary Differential Equations*, in Theory of Numerics in Ordinary and Partial Differential Equations, M Ainsworth, J Levesley, W Light, M Marletta, eds., Oxford University Press, 1995, pp. 1-76.

[28] Corliss G F, *Survey of Interval Algorithms for Ordinary Differential Equations*, Appl. Math. Comput. 31, 1989, pp.112-120.

[29] Corliss G F, Krenz G S, Davis P H, *Bibliography on Interval Methods for the Solution of Ordinary Differential Equations*, Technical Report No. 289, Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, 1988.

[30] Corliss G F, *Where is Validated ODE Solving Going*, in Proceedings of the IMACS-GAMM International Symposium on Numerical Methods and Error Bounds, J Herzberger, ed., Academie Verlag, Berlin, 1996, pp. 48-57.

[31] Dautray R, Lions J-L, Mathematical Analysis and Numerical Methods for Science and Technology, Volume 2: Functional and Variational Methods, Springer-Verlag, Berlin, Heidelberg, 1988.

[32] Demidovich B, Maron I, Shuvalova Z, Numerical Methods of Analysis, Nauka, Moskow, 1967.

[33] Dobner H J, Einschliessungalgorithmen für Hyperbolische Differentialgleichungen, PhD Thesis, University of Karlsruhe, 1986.

[34] Dobner H J, *Bounds for the Solution of Hyperbolic Problems*, Computing 38,p.209-218, Springer-Verlag, 1987

[35] Dobner H J, *Verified Solution of the Integral Equations for the Two-Dimensional Dirichlet and Neumann Problem*, Computing, Suppl. 9, p.33-43, Springer-Verlag, 1993.

[36] Dobner H J, Kaucher E, *Self-Validating Computations of Linear and Nonlinear Integral Equations of the Second Kind*, Contributions to Computer Arithmetic, Ed. C.Ullrich, IMACS, 1990.

[37] Eijgenraam P, The Solution of Initial Value Problems Using Interval Arithmetic, Mathematical Centre Tracks No 144, Mathematisch Centrum, Amsterdam, 1981.

[38] Epstein C, Miranker W L, Rivlin T J, *Ultra Arithmetic, Part I: Function Data Types, Part 2: Intervals of Polynomials*, Mathematics and Computers in Simulation, XXIV, 1982, pp. 1-18.

[39] Giannakoukos J, Kaarniadakis G, *Spectral Element - FCT Method for Scalar Hyperbolic Conservation Laws*, International Journal for Numerical Methods in Fluids, Vol.14,707-727, 1992.

[40] Gomez R, Winicour J, *Evolution of Scalar Fields from Characteristic Data*, Journal of Computational Physics 98, p.11-25, 1992.

[41] Hammer R, Hocks M, Kulish U, Ratz D, Numerical Toolbox for Verified Computing I, Springer-Verlag, 1993.

[42] Harary F, Graph Theory, Addison Wesley, 1969.

[43] Hausdorff F, Mengelehren, De Gruyter, Leipzig, 1927.

[44] IBM, *IBM High Accuracy Arithmetic - Extended Scientific Computation*, IBM, Mechanicsburg, Penn., IBM Publication No. SC33-6462-00, 1990.

[45] Jackson L, *Interval Arithmetic error-bounding algorithms*, SIAM J. Numer. Analysis, 12, 1975, pp.223-238.

[46] Kalmykov S, Shokin U, Uldashev Z, Methods of Interval Analysis, Nauka, Novosibirsk, 1986.

[47] Kantorivich L V, On the General Theory of Operations in Partially Ordered Spaces, DAN USSR 1.271-274, 1936 (in Russian).

[48] Kaucher E, Baumhof C, *A Verified Computation of Fourier-Representation of Solutions for Functional Equation*, Computing, Suppl. 9, 101-115(1993), Springer-Verlag, Viena, 1993.

[49] Kaucher E, Miranker W, Self-validating Numerics for Function Space Problems, Academic Press, New York, 1984.

[50] Kaucher E W, *Self-validating Computations for Ordinary and Partial Differential Equations*, in: Kaucher E W, Kulisch U, Ullrich Ch (eds.), Computer Arithmetic, BG Teubner, Stuttgart, 1987.

[51] Kearfott R B, Dawande M, Hu C Y, *ALGORITHM 737: INTLIB, A Portable FORTRAN 77 Interval Elementary Function Library*, ACM Transactions on Mathematical Software, 20, 1994, pp. 447-459.

[52] Klatte R, Kulisch U, Neaga M, Lawo C, Rauch M, Wiethoff A, C-XSC: A C++ Library for Extended Scientific Computation, Springer-Verlag, Berlin, 1993.

[53] Klatte R, Kulisch U, Neaga M, Ratz D, Ullrich C, PASCAL-XSC Language Reference with Examples, Springer-Verlag, Berlin, 1991.

[54] Kok B, Approximation of Discontinuous Solutions of Hyperbolic Partial Differential Equations, DSc thesis, University of Pretoria, 1982.

[55] Krückeberg F, *Ordinary Differential Equations*, in Topics in Interval Analysis, Hansen E, ed., Claredon Press, Oxford,1969, pp.91-97.

[56] Kulisch U, *Grundlagen des Numerishen Rechnens - Mathematische Begründung der Rechnerarithmetik*, Reine Informatik, Band 19, Bibliographisches Institut, Mannheim, 1976.

[57] Kulisch U, Miranker W, Computer Arithmetic in Theory and Practice, Academic Press, NY, 1983.

[58] Kulisch U, Miranker W, *The Arithmetic of a Digital Computer: A new Approach*, SIAM Review, 28, 1986, pp. 1-40.

[59] Lohner R J, *Enclosing the Solutions of Ordinary and Boundary Value Problems*, in Computer Arithmetic: Scientific Computation and Programming Languages, Kaucher E W, Kulisch U W, Ullrich C, eds., Wiley-Teubner Series in Computer Science, Stuttgart, 1987, pp. 255-286.

[60] Markov S, *A Non-standard Subtraction of Intervals*, Serdica, 3, 1977, pp.359-370.

[61] Markov S, *Isomorphic Embeddings of Abstract Interval Systems*, Reliable Computing Vol 3, No 3, Kluwer Academic Publishers, 1997, pp. 199-207.

[62] Markov S, *On the Algebra of Intervals and Convex Bodies*, in Proceedings of SCAN97, J. Universal Computer Science, Vol.4, No.1, 1998, pp. 34-37.

[63] Mayer O, *Algebraische und Metrische Strukturen in der Intervallrechnung und Einige Anwendungen*, Computing, 5, 1970, pp.144-162.

[64] Moore R E, *The Automatic Analysis and Control of Error in Digital Computations Based on the Use of Interval Numbers*, in Rall L B (ed), Error in Digital Computation, Vol. 1, John Wiley and Sons, New York, 1965, pp. 61-130.

[65] Moore R E, *Automatic Local Coordinate Transformation to reduce the Growth of Error Bounds in Interval Computation of Ordinary Differential Equations*,in Rall L B (ed), Error in Digital Computation, Vol. 1, John Wiley and Sons, New York, 1965, pp. 103-140.

[66] Moore R E, Interval Analysis, Prentice Hall, 1966.

[67] Moore R E, Methods and Applications of Interval Analysis, SIAM, Philadelphia, 1979.

[68] Müller M, *Über die Eindeutigkeit der Integrale eines Systems Gemöhnlicher Differentialgleichungen und die Konvergentz einer Gattung von Verfaren zur Approximation dieser Integrale*, Sitzb., Heidelberg, Akad. Wiss., Math.-Naturw., Kl. 9, Abh., 1927.

[69] Nakao M, *Solving Nonlinear Parabolic Problems with Result Verification*, Journal of Computational and Applied Mathematics, 38(1-3), 1991, pp. 323-334.

[70] Nakao M, *Computable Error Estimates for FEM and Numerical Verification of Solutions for Nonlinear PDEs*, in Computational and Applied Mathematics, I (Dublin, 1991), North-Holland, Amsterdam, 1992, pp. 357-366.

[71] Nakao M, *Numerical Verification Methods for the Solutions of Nonlinear Elliptic and Evolution Problems*, in Computer Arithmetic and Enclosure Methods (Oldenburg, 1991), North-Holland, Amsterdam, 1992, pp. 391-399.

[72] Nakao M, Yamamoto N, *Numerical Verifications of Solutions for Elliptic Equations in Nonconvex Polygonal Domains*, Numerische Mathematik, 65(4), 1993, pp. 503-521.

[73] Nakao M, Watanabe Y, *On Computational Proofs of the Existence of solutions to Nonlinear Parabolic Problems*, Journal of Computational and Applied Mathematics, 50(1-3), 1994, pp. 401-410.

[74] Nakao M, *Numerical Verifications of Solutions for Nonlinear Hyperbolic Equations*, Interval Computations, 4, 1994, pp. 64-77.

[75] Nakao M, Yamamoto N, *Linear/Nonlinear Iterative Methods and Verification of Solution*, Journal of Computational and Applied Mathematics, 60(1-2), pp. 271-279.

[76] Nakao M, Watanabe Y, Yamamoto N, *Guaranteed Error Bounds for Finite Element Solutions of Stokes Problem*, in Scientific Computing and Validated Numerics (Wuppertal, 1995), Akademie Verlag, Berlin, 1996, pp. 258-264.

[77] Nakao M, Watanabe Y, Yamamoto N, *Verified Computations of Solutions for Nondifferentiable Elliptic Equations Related to MHD Equilibria*, Nonlinear Analysis: Theory, Methods and Applications, 28(3), 1997, pp. 577-587.

[78] Nedialkov N S, Jackson K R, Corliss G F, *Validated Solutions of Initial Value Problems for Ordinary Differential Rquations*, Technical Report, Department of Computer Science, University of Toronto, 1997.

[79] Neumaier A, Interval Methods for Systems of Equations, Cambridge University Press, Cambridge, 1990.

[80] Nickel K (ed), Interval Mathematics, Academic Press, New York, London, Toronto, 1980.

[81] Nickel K, *How to Fight the Wrapping Effect*, in Interval Mathematics '85, Lecture Notes in Computer Science No 212, Springer, Berlin, 1985, pp. 121-132.

[82] Nickel K, *Using Interval Methods for the Numerical Solution of ODE's*, Z Angew. Math. Mech. 66, 1986, pp. 513-523.

[83] Potter M H, Weinberger H F, Maximum Principles in Differential Equations, Springer-Verlag, Berlin, 1984.

[84] Rall L B, *Tools for Mathematical Computation*, in: Computer Aided Tools in Analysis (Cincinnati, OH, 1989), Springer, New York, 1991, pp. 217-228.

[85] Ratschek H, Ströder G, *Presentation of semi-groups as systems of compact convex sets*, Proc Amer. Math. Soc. 65, 1977, pp. 24-28.

[86] Rihm R, *Interval Methods for Initial Value Problems in ODE*, in Topics in Validated Computations: Proceedings of The IMACS-GAMM International Workshop on Validated Computations, University of Oldenburg, J Herzberger, ed., Elsevier Studies in Computational Mathematics, Elsevier, Amsterdam, New York, 1994.

[87] Schlett M, *A Spectral Method for the Numerical Simulation of Transit-Time Devices*, in Simulation of Semiconductor Devices and Processes, Vol. 5, S Selberherr, ed., Springer, Berlin, 1993, pp. 241-244.

[88] Stetter H J, *Algorithms for the Inclusion of Solutions of Ordinary Initial Value Problems*, in Equadiff 6: Proceedings of the International Conference on Differential Equations and Their Applications (Brno, 1985), Vosmansky J, Zlamal M, eds, Lecture Notes in Mathematics, Vol 1192, Springer-Verlag, Berlin, 1986, pp.85-94.

[89] Sunaga T, *Theory of an Interval Algebra and Its Applications to Numerical Analysis*, RAAG Memoirs 2, 1958, pp.547-564.

[90] Walter W V, *ACRITH-XSC: A FORTRAN-like Language for Verified Scientific Computing*, in Scientific Computing with Automatic Result Verification, E Adams, U Kulisch, eds.,Academic Press, Orlando, Fla., 1992.

[91] Walter W V, *FORTRAN-XSC: A Portable Fortran 90 Module Library for Accurate and Reliable Scientific Computation*, in: Albrecht R, Alefeld G, Stetter H J (eds), Validation Numerics - Theory and Applications, Computing Supplementum 9, Springer-Verlag, Wien/New York, 1993, pp. 265-285.

[92] Walter W, Differential and Integral Inequalities, Springer-Verlag, Heidelberg - Berlin - New York, 1970.