Numerical Approximation of Boundary Control for the Wave Equation - with Application to an Inverse Problem

Mariegaard, Jesper Sandvig; Knudsen, Kim; Hansen, Per Christian; Pedersen, Michael

Publication date:
2009

Jesper Sandvig Mariegaard

Numerical Approximation of Boundary Control for the Wave Equation

—with Application to an Inverse Problem

Ph.D. Dissertation, July 2009
Supervisors: Associate Professor **Kim Knudsen**
Professor **Per Christian Hansen**
Professor **Michael Pedersen**
Department of Mathematics, Technical University of Denmark.

Abstract

We consider a control problem for the wave equation: Given the initial state, find a specific boundary condition, called a control, that steers the system to a desired final state. The Hilbert uniqueness method (HUM) is a mathematical method for the solution of such control problems. It builds on the duality between the control system and its adjoint system, and these systems are connected via a so-called controllability operator.

In this project, we are concerned with the *numerical approximation* of HUM control for the one-dimensional wave equation. We study two semi-discretizations of the wave equation: a linear finite element method (L-FEM) and a discontinuous Galerkin-FEM (DG-FEM).

The controllability operator is discretized with both L-FEM and DG-FEM to obtain a HUM matrix. We show that formulating HUM in a *sine* basis is beneficial for several reasons: (i) separation of low and high frequency waves, (ii) close connection to the dispersive relation, (iii) simple and effective filtering.

The dispersive behavior of a discretization is very important for its ability to solve control problems. We demonstrate that the *group velocity* is determining for a scheme's success in relation to HUM. The vanishing group velocity for high wavenumbers results in a dramatic decay of the corresponding eigenvalues of the HUM matrix and thereby also in a huge condition number. We show that, provided sufficient filtering, the *phase velocity* decides the accuracy of the computed controls.

DG-FEM shows very suitable for the treatment of control problems. The good dispersive behavior is an important virtue and a decisive factor in the success over L-FEM. Increasing the order of DG-FEM even give results of spectral accuracy.

The field of control is closely related to other fields of mathematics among these are *inverse problems*. As an example, we employ a HUM solution to an inverse source problem for the wave equation: Given boundary measurements for a wave problem with a separable source, find the spatial part of the source term. The reconstruction formula depends on a set of HUM eigenfunction controls; we suggest a discretization and show its convergence. We compare results obtained by L-FEM controls and DG-FEM controls. The reconstruction formula is seen to be quite sensitive to control inaccuracies which indeed favors DG-FEM over L-FEM.

Resumé

Numerisk approksimation af randkontrol for bølgeligningen

Vi betragter et kontrolproblem for bølgeligningen: Givet begyndelsestilstanden, find en særlig randbetingelse, kaldet en kontrol, som styrer systemet til en ønsket sluttilstand. *Hilbert uniqueness method* (HUM) er en matematisk metode til løsning af denne type kontrolproblem. Den bygger på dualiteten mellem kontrolsystemet og dets adjungeret system, og disse systemer er forbundet via en såkaldt kontrollabilitetsoperator.

I dette projekt behandler vi den numeriske approksimation af HUM-kontrol for bølgeligningen i én dimension. Vi studerer to semi-diskretiseringer af bølgeligningen: en lineær *finite element* metode (L-FEM) og en *discontinuous Galerkin*-FEM (DG-FEM).

Ved diskretisering af kontrollabilitetsoperatoren med både L-FEM og DG-FEM opnås en HUM-matrix. Vi viser, at formuleringen af HUM i sinus-basis er fordelagtig af følgende grunde: (i) adskillelse af bølger med lavt og højt bølgetal, (ii) tæt kobling til dispersionsrelationen, (iii) simpel og effektiv filtrering.

En diskretiserings dispersive egenskaber er meget vigtige for dens evne til at løse kontrolproblemer. Vi demonstrerer, hvordan gruppehastigheden er afgørende for en diskretiserings muligheder for succes i forbindelse med HUM. Den forsvindende gruppehastighed for høje bølgetal resulterer i et dramatisk fald i HUM matricens egenværdier og derved også i et enormt konditionstal. Vi viser at fasehastigheden ved passende filtrering bestemmer nøjagtigheden af de udregnede kontroller.

DG-FEM viser sig velegnet til behandlingen af kontrolproblemer. Dens gode dispersive egenskaber er et vigtigt fortrin, som er en afgørende faktor i dens succes over L-FEM. Ved at øge den polynomielle orden kan DG-FEM endda give resultater med spektral nøjagtighed.

Emnet kontrol er tæt knyttet til andre matematiske områder - blandt disse er *inverse problemer*. Som et eksempel på dette anvender vi HUM til løsning af et inverst kildeproblem: givet randdata for et bølgeproblem med en separabel kilde, find den stedslige del af kildeledet.

Rekonstruktionsformlen bygger på en serie HUM egenfunktionskontroller. Vi foreslår en diskretisering og viser dens konvergens. Vi sammenligner ligeledes resultater opnået med L-FEM-kontroller og DG-FEM-kontroller. Rekonstruktionsformlen viser sig følsom overfor kontrolunøjagtigheder, hvilket betyder, at DG-FEM klarer sig markant bedre end L-FEM.

Preface

This dissertation was written as partial fulfillment of the requirements for obtaining the Ph.D. degree. The work has been carried out in the Department of Mathematics at the Technical University of Denmark (DTU Mathematics) from September 2005 to July 2009. The main supervisor until August 2008 was Pofessor Michael Pedersen (now Roskilde University); from August 2008 the main supervisor has been Associate Professor Kim Knudsen (DTU Mathematics). Professor Per Christian Hansen (DTU Informatics) has been co-supervisor.

Prerequisites. The reader is assumed familiar with basic theory of functional analysis, partial differential equations, and numerical analysis. No particular knowledge about control theory or inverse problems is required.

Software. The computer software developed in this project is freely available for download under the GNU GPL license at

http://www.mat.dtu.dk/people/J.S.Mariegaard/software/

The reader is encouraged to download it for use and modification. The developed software has been written in the technical computing environment Matlab.

Acknowledgments. I would foremost like to thank my supervisors Kim Knudsen and Per Christian Hansen for taking interest in my project and discussing its details with me. I would also like to thank my first supervisor Michael Pedersen for introducing the Hilbert uniqueness method to me. Professor Enrique Zuazua (Universidad Autónoma de Madrid, Spain) who I visited from August through December 2006 was kind to point me towards the field of inverse problems and insisted on the close connection to control problems. I am also grateful to Professor Jan S. Hesthaven (Brown University, USA) for teaching me about discontinuous Galerkin-FEM and recommending its use in context with control. Finally, I would like to thank Allan P. Engsig-Karup for helping me out with different numerical issues.

I dedicate this dissertation to Eskild, Eigil and Lise.

Kgs. Lyngby, July 22, 2009

Revised edition 1.1,
Kgs. Lyngby, September 7, 2009 Jesper Sandvig Mariegaard

Contents

Abstract iii

Resume v

Preface vii

1 Introduction 1
- 1.1 What is control? . . . 1
- 1.2 The objective . . . 3
- 1.3 Scope and structure . . . 3

2 HUM for the wave equation 7
- 2.1 The control system . . . 8
  - 2.1.1 Types of controllability . . . 9
  - 2.1.2 An auxiliary system . . . 10
  - 2.1.3 The control area . . . 10
- 2.2 The adjoint system . . . 11
  - 2.2.1 Energy spaces . . . 12
  - 2.2.2 Duality . . . 13
- 2.3 Observability . . . 14
- 2.4 HUM: the operator approach . . . 15
  - 2.4.1 The reconstruction operator . . . 15
  - 2.4.2 The Λ-operator and a main theorem . . . 16
  - 2.4.3 Constructing Λ as a matrix . . . 17
  - 2.4.4 Constructing Φ and Ψ as matrices . . . 18
- 2.5 HUM: The minimization approach . . . 19

3 Approximating solutions to the wave equation 21
- 3.1 The continuous 1-d wave equation . . . 22
  - 3.1.1 The wave equation as a conservation law . . . 23
- 3.2 Classical semi-discretizations . . . 26
  - 3.2.1 The finite difference method (FDM) . . . 27
  - 3.2.2 The finite element method (FEM) . . . 28
  - 3.2.3 A unified formulation . . . 30
  - 3.2.4 Other classical methods . . . 31
- 3.3 Discontinuous Galerkin FEM . . . 31
  - 3.3.1 A DG-scheme for the advection equation . . . 32
  - 3.3.2 The LGL grid and DG-basis functions . . . 34

3.3.3 A DG-scheme for the wave equation . . . 37
3.4 Time integration . . . 40
3.5 Method properties and analysis . . . 43
3.5.1 The dispersion relation and group velocity . . . 43
3.5.2 Convergence . . . 49

4 Numerical HUM **51**
4.1 Discrete control . . . 53
4.1.1 Classical control theory . . . 53
4.1.2 Semi-discrete HUM . . . 54
4.1.3 Discrete HUM . . . 60
4.2 Construction of the $\boldsymbol{L}$ matrix . . . 61
4.2.1 Matrix assembling—procedures and details . . . 61
4.2.2 Discrete HUM by minimization . . . 63
4.2.3 The sine basis . . . 64
4.2.4 Constructing $\boldsymbol{L}$ with L-FEM . . . 66
4.2.5 Constructing $\boldsymbol{L}$ with DG-FEM . . . 79
4.3 Iterative HUM by conjugate gradients . . . 98
4.3.1 The algorithm . . . 98
4.3.2 Filtering by basis truncation . . . 100
4.3.3 A test problem . . . 100
4.4 Concluding remarks . . . 103
4.4.1 Related work . . . 104
4.4.2 Discussion . . . 105

5 The inverse problem—an application of HUM **109**
5.1 An inverse source problem . . . 111
5.1.1 Examination of the inverse problem . . . 112
5.1.2 An auxiliary inverse 'initial data' problem . . . 112
5.2 A HUM solution to the inverse problem . . . 113
5.2.1 Stability . . . 114
5.2.2 Reconstruction . . . 116
5.2.3 Regularization . . . 118
5.3 Discrete reconstruction . . . 119
5.3.1 Discrete (IIDP) . . . 120
5.3.2 Discrete (ISP) . . . 121
5.4 Numerical results . . . 122
5.4.1 Data and the forward problem . . . 122
5.4.2 The degree of ill-posedness . . . 123
5.4.3 Reconstruction with analytic HUM controls . . . 124
5.4.4 Numerical reconstruction with L-FEM . . . 131
5.4.5 Numerical reconstruction with DG-FEM . . . 135
5.4.6 An example with random coefficients . . . 141
5.5 Concluding remarks . . . 142

6 Conclusion **145**
6.1 Results . . . 145
6.1.1 The control problem and numerical HUM . . . 145
6.1.2 The inverse problem and the numerical reconstruction . . . 146
6.1.3 Software contributions . . . 147

6.2 Future work 147

A List of Symbols **149**

B Mathematical details **155**
B.1 Analytic solution to the forward problem 155

C The Matlab package **157**
C.1 Module `WAVE` 158
C.1.1 Function summary 158
C.1.2 Short user guide 158
C.1.3 Examples of use 159
C.2 Module `HUM` 159
C.2.1 Function summary 159
C.2.2 Short user guide 160
C.2.3 Examples of use 160
C.3 Module `DGWAVE` 160
C.3.1 Function summary 161
C.3.2 Short user guide 161
C.3.3 Examples of use 161
C.4 Module `DGHUM` 162
C.4.1 Function summary 162
C.4.2 Short user guide 162
C.4.3 Examples of use 162
C.5 Module `IP` 163
C.5.1 Function summary 163
C.5.2 Short user guide 163
C.5.3 Examples of use 164

Bibliography **169**

CHAPTER 1

Introduction

It has been known to scientists and engineers for centuries that partial differential equations (PDEs) can be used to describe a huge class of physical phenomena. In this dissertation, we shall consider control problems that are governed by PDEs. A systematic method called the Hilbert uniqueness method (HUM) can be used to obtain control of the wave equation. The main objective of the dissertation is the numerical approximation of this HUM control. Before we go into more details about ends and means, we will elaborate on the aspects of the topic control.

1.1 What is control?

Consider a "system" whose "state" can change over time; the state of the system could for example be its temperature or displacement. If we are allowed to act on the system, in some way, and thereby change its state, we call the system a *control system*; our action on the system is called a *control*. We will consider control problems of the following kind: given the *initial state*, find a control that steers the system to a specific *final state*.

If the system is a PDE, then the state is a function typically of the space variable.

The control of PDEs is—to distinguish it from the traditional finite-dimensional control—often designated *control of systems of distributed parameters*. The two fields share the important notions of *controllability* and *observability*. A problem is said to be controllable if there exists a control that drives the system to the desired final state. A problem is observable on a subset of the domain or boundary if the *total* energy of the system can be determined by measurements on this subset only. Controllability and observability are mathematical duals in the sense that a system is controllable if, and only if, its adjoint system is observable.

It should be noted that controllability has only little in common with *optimal* control, since optimal control deals with optimization strategies for control problems. Controllability is, however, closely related to stabilization, homogenization and inverse problems; we shall later consider an example of the relation to the latter.

In most PDE control problems, there are many possible controls that solve the problem, but one control has particular interest: the control of minimal energy.

In the late 1980s, the French mathematician Jacques-Louis Lions announced the *Hilbert uniqueness method* for controllability problems [Lio88]. HUM is an abstract, systematic and constructive method that provides, among all controls, the unique control of minimal L^2-norm. The first example of its use was the wave equation.

We shall consider a control problem for the wave equation in a bounded domain. The control that we seek is a boundary condition for this wave equation which makes the problem a *boundary* control problem.

The concept of the problem is easily conceived in 1-d:

> *An idealized piece of string, which is fixed in one end, experiences wave motion. We have the other end of the string in our hand. If we do nothing, the motion of the idealized string will go on forever, but if we instead move our hand up or down, we create new waves and thus change the state of the string. If the desired final state is zero (no motion), we may move our hand up and down in ways so to drag out the energy of the string motion and thereby putting it to rest. This is boundary control. The control which requires minimal effort, in terms of hand movement, is found by HUM.*

HUM has, however, the ability to deal with boundary control of the wave equation not only in 1-d but in complex geometries in higher dimensions.

Numerical methods have to be applied to obtain a specific control by the HUM. The discretization of the wave equation introduce spurious, non-physical waves of high frequency that threatens to ruin the control. This behavior is closely related to the dispersive properties of the discretization. Different numerical methods have been used in relation to HUM, but these have at large been simple, lower order discretizations due to the much focus on theoretical aspects of the numerical analysis. Only a few authors have engaged in the more practical facets and conducted numerical studies.

The discontinuous Galerkin finite element method is a semi-discretization method which more recently has become popular for the solution of PDEs and

in particular wave problems. It is known to exhibit good dispersive behavior, yet it has not been used for HUM control.

Ever since the introduction of HUM, the method and its underlying ideas have influenced many other fields; one of them is *inverse problems*. An inverse problem concerns finding causes from measured effects which, as mentioned, is a problem closely related to controllability and observability. Masahiro Yamamoto used HUM for the solution of an inverse source problem for the wave equation in his paper [Yam95]. He considered a problem with source term separable in a temporal and a spatial part. The problem consisted, for given temporal part, of finding the spatial part of the source from boundary measurements on part of the boundary. It seems that no one has engaged in the numerical approximation of his method; a lack that motivates this study.

1.2 The objective

This dissertation deals with the numerical approximation of HUM boundary control for the wave equation. The main goal is to find "good" approximate controls for the problem in 1-d. We take a practical approach to this end, and we wish to examine the following topics.

I. *Analyze, identify, and understand the mechanisms that may lead to diverging controls.*
   - *examine the spectral properties of the discretized HUM operators*
   - *trace the consequences of numerical dispersion through each step of the approximation of HUM*
   - *study the effects of formulating the problem in sine basis*

II. *Examine whether and how the discontinuous Galerkin-FEM can be used for the numerical approximation of HUM boundary control.*

III. *Study the numerical approximation of Yamamoto's method for the inverse source problem.*
   - *discretize the problem and assess the degree of ill-posedness*
   - *examine how the known problems from numerical HUM effect the solution of the inverse problem*

IV. *Develop software for HUM boundary control and for the inverse problem and make it freely available to the benefit of others.*

1.3 Scope and structure

The HUM is presented in **Chapter 2** as a method for control for the wave equation. Apart from the physical control system, an auxiliary (Section 2.1) and an adjoint system (Section 2.2) are introduced. These two systems are equipped with energy norms and connected by a Green's formula which establishes a fundamental duality between them. An observation map is assigned to the adjoint system in Section 2.3 and the dual equivalent, a reconstruction map, is assigned

to the auxiliary system in Section 2.4. The observation and the reconstruction are combined to the so-called Λ operator whose inversion provides the solution of the HUM problem. Section 2.4 also features the representation of the HUM operators as infinite matrices in the 1-d case; a construction that marks the opening of the spectral study of the HUM operators. Finally, Section 2.5 presents the HUM as a minimization problem instead of an operator problem.

Chapter 3 deals with the discretization of the wave equation with the aim of finding approximate solutions that can be used for for the numerical approximation of HUM. We shall consider numerical HUM in 1-d and study therefore the 1-d wave equation. The *continuous* 1-d wave equation is studied in Section 3.1. We take the method of lines approach and consider first the semi-discretization of the wave equation. Classical methods, such as the finite difference and finite element method (FEM), are introduced in Section 3.2, and the discontinuous Galerkin-FEM (DG-FEM) in Section 3.3. We rewrite the wave equation as a conservation law with two advection equations to make it consistent with the DG-FEM formulation. This also makes it possible to use an upwind scheme. Section 3.4 is on time discretization which is necessary to obtain a fully discrete scheme. Section 3.5 analyzes the dispersive properties of the presented schemes which becomes important for the examination in Chapter 4. We pick two schemes for further studies: linear FEM (L-FEM) with trapezoidal time integration and DG-FEM with a Runge-Kutta time integration.

With the means for obtaining discrete solutions to the wave equation, we are, in **Chapter 4**, ready to introduce the discretization of HUM. Section 4.1 suggests a semi-discretization of HUM; any semi-discretization makes HUM a finite dimensional control problem for which there exists a vast amount of literature. Considering the control of the wave equation as any old finite dimensional control problem will, however, easily lead to failure.

We proceed with the full discretization before we describe the construction of the HUM operator Λ as a matrix in Section 4.2. This section presents the first results of the numerical study. The construction, which is carried out in sine basis, is carefully analyzed for both L-FEM and DG-FEM. We establish connections to the dispersive properties of the discretizations. The use of the sine basis provides, in addition, a simple filtering procedure.

The explicit construction of Λ as a matrix rapidly becomes infeasible and Section 4.3 presents an iterative solution to the HUM problem with a conjugate gradient algorithm, and convergence is studied for an example. Chapter 4 is concluded in Section 4.4 by a short review of related work and a discussion of the obtained results.

Chapter 5 takes this dissertation in a new direction by introducing an inverse problem. It is the inverse source problem for the wave equation described above (see also Section 5.1), and more importantly M. Yamamoto's solution by HUM presented in Section 5.2. We suggest a discretization of the reconstruction formula in the 1-d case in Section 5.3, and present the obtained numerical results in Section 5.4. The continuous inverse problem is ill-posed; we assess to which degree the discretized problem is ill-posed by considering the singular values of the corresponding forward map.

HUM eigenfunction controls are needed for the reconstruction of Fourier

coefficients for the source term and we consider first analytic HUM controls and then numerical controls obtained with respectively L-FEM and DG-FEM. We attempt to restore 25 random coefficients with these three different sets of controls in Section 5.4.6. The last section of Chapter 5 is a discussion of the obtained results.

The final chapter, **Chapter 6**, features a list of the most important results obtained in this project and a list of recommended future work.

The developed software, all written in Matlab, is considered an important contribution of this work. A brief description of each function together with a short user guide can be found in Appendix C. The complete code including examples may be obtained at

http://www.mat.dtu.dk/people/J.S.Mariegaard/software/

The remaining appendices contain a list of symbols used in the dissertation (Appendix A) and a few mathematical details (Appendix B).

CHAPTER 2

HUM for the wave equation

In this chapter, we consider boundary control for the wave equation. The key concept, apart from controllability, is observability—whether the total energy of a system can be found by partial measurements. Controllability and observability are mathematical duals in the way that the control system is controllable if and only if its adjoint system is observable.

We present the Hilbert uniqueness method (HUM) which is a general, systematic, and constructive method for obtaining the best[1] possible control. The method is formulated abstractly in *Hilbert* spaces and build on *uniqueness* results for the governing PDEs. We use it here in a concrete setting for the classical wave equation in an open, bounded domain in $\mathbb{R}^d$. We will remark on the special 1-d case several times in this chapter, since we shall later study the numerical approximation of 1-d HUM in Chapter 4.

We present HUM in two different formulations, an operator approach in Section 2.4 and a minimization approach in Section 2.5, as these form the bases of different ways of approximating the control, which we will return to in Chapter 4.

[1] It is the "best" control in the sense that it is, among all possible controls, the control with minimal L^2-norm.

The exposition below is primarily inspired by the works of E. Zuazua and co-workers (see, *e.g.,* [Zua05] and [MZ05]) and M. Pedersen [Ped08]. The proofs in this chapter are not by the author and those not central to the exposition (*e.g.,* solvability results) have been omitted. We present proofs of central HUM results in brevity or by a mere outline.

A word on the geometry and notation

Before setting off, we need to introduce some notation. We shall consider systems in an open, bounded domain $\Omega \in \mathbb{R}^d$ with boundary Γ for the time interval $(0,T)$. We call the time-boundary cylinder $\Sigma = (0,T) \times \Gamma$. A part of the boundary will be known as the control boundary and is denoted $\Gamma_0 \subset \Gamma$; the corresponding part of Σ is denoted $\Sigma_0 = (0,T) \times \Gamma_0$.

Note that, when we consider the 1-d case, Ω will denote the line segment $(0,1)$, and the control boundary Γ_0 will be a single point $x = 1$.

Throughout this dissertation, time derivatives $\partial/\partial t$ of a function y will be shown by the superscript y'. This is done to clearly distinct it from the spatial derivatives. We will use Δ for the Laplacian operator.

2.1 The control system

We set off by introducing the main object of our study: a wave equation which we call the *control system*

$$u'' - \Delta u = 0, \qquad \text{in } (0,T) \times \Omega, \tag{2.1a}$$

$$u(t,x) = \begin{cases} \kappa(t,x) & \text{for } x = \Gamma_0, \\ 0 & \text{for } x = \Gamma \setminus \Gamma_0, \end{cases} \qquad t \in (0,T), \tag{2.1b}$$

$$u(0,x) = u^0(x), \quad u'(0,x) = u^1(x), \qquad x \text{ in } \Omega, \tag{2.1c}$$

where the initial data $(u^0, u^1) \in L^2(\Omega) \times H^{-1}(\Omega)$ is given, and κ is a function in $L^2(\Sigma_0)$. The space $H^{-1}(\Omega)$ is the dual of $H_0^1(\Omega)$.

The space $L^2(\Omega) \times H^{-1}(\Omega)$ will be used frequently so to shorten notation we introduce

$$\widetilde{\mathcal{E}}^\star := L^2(\Omega) \times H^{-1}(\Omega). \tag{2.2}$$

Also, for the space of the boundary function κ we introduce the *boundary*[2] *space*

$$\mathcal{B} := L^2(\Sigma_0). \tag{2.3}$$

We equip both spaces with the usual norms. There are good reasons for the choice of these two spaces; they are in fact essential to HUM and the consequence of careful analysis by J.-L. Lions [Lio88]. We will return to this below.

Before being more specific about our main objective—the control problem—we state an existence and uniqueness result for system (2.1) (see, *e.g.,* [Ped00, page 180] for a proof).

[2]Note that, in spite of the name *boundary* space, $\mathcal{B}$ is the L^2-space only over *a part* $\Sigma_0 = (0,T) \times \Gamma_0$ of the complete time-boundary cylinder $\Sigma = (0,T) \times \Gamma$.

Theorem 2.1. *For any $(u^0, u^1) \in \widetilde{\mathcal{E}}^\star$ and any $\kappa \in \mathcal{B}$ there exists a unique weak solution to (2.1) with the regularity*

$$(u, u') \in C([0,T]; \widetilde{\mathcal{E}}^\star), \tag{2.4}$$

moreover, the map $\{u^0, u^1, \kappa\} \mapsto \{u, u'\}$ is linear and there exists a constant $c(T) < 0$ such that

$$\|(u,u')\|_{L^\infty((0,T);\widetilde{\mathcal{E}}^\star)} \leq c(T)\big(\|(u^0,u^1)\|_{\widetilde{\mathcal{E}}^\star} + \|\kappa\|_{\mathcal{B}} \big). \qquad \square$$

Remark 2.2. The wave equation is time-reversible in nature due to the symmetry of the wave operator $\partial_t^2 - \Delta$, that is, replacing t by $T - t$ in (2.1) will lead to the exact same PDE. The above regularity result holds in both directions. ■

The state of the control system (2.1) at time t reads $(u(t,\cdot), u'(t,\cdot))$, and it belongs to the state space $\widetilde{\mathcal{E}}^\star$. Control is action on a system through a control variable which changes the state of the system. We seek to change the state of (2.1) by acting on the control boundary Γ_0 with the control function κ. This type of control is called boundary control.

We shall now define the boundary control problem for system (2.1).

Definition 2.3 (Control problem). Given $(u^0, u^1) \in \widetilde{\mathcal{E}}^\star$, find $\kappa \in \mathcal{B}$ such that (2.1) is steered to zero in time T, *i.e.,*

$$u(T,x) = u'(T,x) = 0, \qquad x \text{ in } \Omega. \qquad \square$$

At this point, it is relevant to ask under which conditions this control problem can be solved for *all* initial data in $\widetilde{\mathcal{E}}^\star$. This is a question about *controllability*.

2.1.1 Types of controllability

Different types of boundary controllability of the wave equation exist. They can conveniently be characterized in terms of the set of reachable final states $R(T;(u^0,u^1))$ which we define as

$$R(T;(u^0,u^1)) := \{(u(T,\cdot), u'(T,\cdot)) \mid u \text{ is the solution to (2.1)}\}.$$

where (2.1) have initial data $(u^0, u^1) \in \widetilde{\mathcal{E}}^\star$ and the control $\kappa \in \mathcal{B}$.

Definition 2.4 (Controllability). Let u be a solution to (2.1) for initial data $(u^0, u^1) \in \widetilde{\mathcal{E}}^\star$. The system is then said to be

Exactly controllable in time T if $R(T;(u^0,u^1))$ is equal to $\widetilde{\mathcal{E}}^\star$ for all initial data $(u^0, u^1) \in \widetilde{\mathcal{E}}^\star$.

Approximately controllable in time T if $R(T;(u^0,u^1))$ is dense in $\widetilde{\mathcal{E}}^\star$ for all initial data $(u^0, u^1) \in \widetilde{\mathcal{E}}^\star$.

Null controllable in time T if the state $(0,0) \in R(T;(u^0,u^1))$ for all initial data $(u^0, u^1) \in \widetilde{\mathcal{E}}^\star$. $\square$

Remark 2.5. For linear systems like (2.1) null controllability and exact controllability are equivalent notions. The null controllability problem related to the control problem, Definition 2.3, is therefore also a problem of exact controllability. ■

Remark 2.6. The wave equation is a partial differential equation with *finite* speed of propagation. *None* of the above types of controllability can therefore be expected to hold unless we allow sufficient large control time T. ■

Approximate controllability, however interesting, is out of the scope of this dissertation. It is only mentioned here for completeness. Readers are referred to [GL95] or [Ped08].

2.1.2 An auxiliary system

Above, we sought a control that—forward in time—would steer the control system from an initial state to a desired final state. Below, we will try to do the opposite, that is, to steer an auxiliary system from a chosen final state back to some desired initial state. This system will prove useful in the following sections.

Let ψ be the solution of the auxiliary system which we solve "backwards" in time from T to 0

$$\psi'' - \Delta\psi = 0, \qquad \text{in } (0,T)\times\Omega, \tag{2.5a}$$

$$\psi(t,x) = \begin{cases} \kappa(t,x) & \text{for } x = \Gamma_0, \\ 0 & \text{for } x = \Gamma \setminus \Gamma_0, \end{cases} \qquad t \in (0,T), \tag{2.5b}$$

$$\psi(T,x) = \psi'(T,x) = 0, \qquad x \text{ in } \Omega, \tag{2.5c}$$

where $\kappa \in \mathcal{B}$. We have from Theorem 2.1 that the state $(\psi(t,\cdot), \psi'(t,\cdot))$ belongs to the state space $\widetilde{\mathcal{E}}^\star$ (Remark 2.2).

A function $\kappa \in \mathcal{B}$ will drive the auxiliary system (2.5) from rest at time T back to some initial state $(\psi(0,\cdot), \psi'(0,\cdot))$. The problem of finding a control κ that steers (2.5) back to $(\psi(0,\cdot), \psi'(0,\cdot)) = (u^0, u^1)$ is equivalent to solving the control problem, Definition 2.3. The found control κ will by construction steer (u^0, u^1) to zero as the two system are identical in this case.

2.1.3 The control area

Up till now we have not said much about the requirements on the control area—the subset Γ_0 of the boundary Γ—on which the control function can act. Naturally, we expect that the control system will be exactly controllable only when the control boundary is a sufficiently large part of the complete boundary.

The traditional way to choose the control boundary is by the following procedure. Let x_0 be an arbitrary chosen point in $\mathbb{R}^d$. Then choose the control boundary

$$\Gamma_0(x_0) = \{x \in \Gamma \mid (x - x_0)\cdot n \geq 0\},$$

where $(x - x_0)$ is the vector from x_0 to the boundary point x and n is the outward unit normal vector at x. A greater control area $\Gamma_1 \subset \Gamma_0(x_0)$ is obviously also admissible. The choice of $\Gamma_0(x_0)$ as control boundary area was important in the original HUM formulation as $(x - x_0)\cdot n$ was used as a multiplier in the

proof by J.L. Lions, [Lio88].

As an alternative to this choice of Γ_0, Bardos, Lebeau, and Rauch proved controllability for C^∞ domains if "*every ray of geometrical optics that propagates in* Ω *and is reflected on its boundary* Γ *must reach* Γ_0 *in time less than* T" [BLR92]. This is called a geometric control condition (GCC) and was proved by the use of micro local analysis. The result was later extended to hold also for C^3 domains. Essentially, the geometric control condition says that the control boundary must be placed such that no rays can be trapped, *e.g.,* between parallel boundary segments.

In general we must expect a decrease in the size of control area will imply an increase in minimal control time T_0 and that a stronger control is needed.

Remark 2.7. In 1-d where the domain $\Omega = (0,1)$, the situation is quite simple. There are two cases:

a) Γ_0 consists of both end points, $\Gamma_0 = \Gamma$, or
b) Γ_0 consists of one end point, say $x = 1$.

Naturally, case b) require twice the control time T_0 compared to case a). In this dissertation, we consider only case b) with $\Gamma_0 = \{1\}$. ■

2.2 The adjoint system

The HUM builds upon the connection between the control system (2.1) and its *adjoint system*

$$\varphi'' - \Delta\varphi = 0, \qquad \text{in } (0,T)\times\Omega, \tag{2.6a}$$
$$\varphi(t,x) = 0, \qquad (t,x) \text{ in } \Sigma, \tag{2.6b}$$
$$\varphi(0,x) = \varphi^0(x), \quad \varphi'(0,x) = \varphi^1(x), \qquad x \text{ in } \Omega, \tag{2.6c}$$

with initial data $(\varphi^0, \varphi^1) \in \mathcal{E}$ which is defined by

$$\mathcal{E} := H_0^1(\Omega) \times L^2(\Omega). \tag{2.7}$$

Note that this space is (almost[3]) the dual of the state space $\widetilde{\mathcal{E}}^\star$ for the control system. Before elaborating on this duality, we state an existence and uniqueness result for (2.6) again without proof (see, *e.g.,* [Ped00, page 178]).

Theorem 2.8. *For any* $(\varphi^0, \varphi^1) \in \mathcal{E}$ *the adjoint system* (2.6) *has a unique weak solution with regularity*

$$(\varphi, \varphi') \in C([0,T]; \mathcal{E}).$$

Furthermore,

$$\frac{\partial\varphi}{\partial n} \in L^2(\Sigma),$$

and there exists a constant $c(T) < 0$ *such that the estimate*

$$\|(\varphi,\varphi')\|_{L^\infty((0,T);\mathcal{E})} \le c(T)\left\|(\varphi^0,\varphi^1)\right\|_{\mathcal{E}} \tag{2.8}$$

holds. □

[3]Strictly, $\mathcal{E}$ is the dual of $H^{-1}(\Omega)\times L^2(\Omega)$ which is isometric isomorphic to $\widetilde{\mathcal{E}}^\star$. At this point it suffices to use $\mathcal{E}$ and $\widetilde{\mathcal{E}}^\star$.

Remark 2.9. The L^2-regularity of the Neumann data $\frac{\partial\varphi}{\partial n}$ is a stronger result than the standard trace results (one-half more regular) that could have been obtained from $\varphi(t,\cdot) \in H_0^1(\Omega)$. This result—a so-called "newer" regularity result—is due to [LLT86], and is known as the *"hidden" regularity* of the wave equation. ■

Specifically, the restriction of the Neumann data $\frac{\partial\varphi}{\partial n}$ to the boundary part Γ_0 belongs to the boundary space $\mathcal{B}$. This "observed" quantity $\frac{\partial\varphi}{\partial n}|_{\Gamma_0}$ will turn out to be essential in connection with the control problem.

2.2.1 Energy spaces

Let us define the mechanical energy E of the adjoint system (2.6) at time t

$$E(t) := \frac{1}{2}\int_\Omega \big(\,|\nabla\varphi(t,x)|^2 + |\varphi'(t,x)|^2\,\big)\mathrm{d}x, \qquad 0 \le t \le T. \tag{2.9}$$

The energy is constant in time (that is, (2.6) is conservative)

$$E(t) = E(0) = \frac{1}{2}\int_\Omega \big(\,\big|\nabla\varphi^0(x)\big|^2 + \big|\varphi^1(x)\big|^2\,\big)\mathrm{d}x, \qquad 0 \le t \le T,$$

due to the homogeneous boundary conditions.

Remark 2.10. In 1-d where $\Omega = (0,1)$ and solutions φ to (2.6) can be expressed in terms of Fourier series by

$$\varphi(t,x) = \sum_{k\in\mathbb{N}} \Big(a_k\cos(k\pi t) + \frac{b_k}{k\pi}\sin(k\pi t)\Big)\sin(k\pi x),$$

where a_k and b_k are defined by the expansion of the initial data

$$\varphi^0(x) = \sum_{k\in\mathbb{N}} a_k\sin(k\pi x), \quad \varphi^1(x) = \sum_{k\in\mathbb{N}} b_k\sin(k\pi x).$$

The energy can be computed straight forwardly by

$$E(0) = \frac{1}{4}\sum_{k\in\mathbb{N}} \big(a_k^2k^2\pi^2 + b_k^2\big).$$ ■

Back to the general case, the energy (2.9) is equivalent to the square of the usual norm on the state space

$$\mathcal{E} = H_0^1(\Omega)\times L^2(\Omega).$$

The norm $\big\|(\varphi^0,\varphi^1)\big\|_{\mathcal{E}}$ does, in other words, describe the mechanical energy of the adjoint system. Consequently, we term $\mathcal{E}$ the *energy space* for the adjoint system. Its dual space

$$\mathcal{E}^\star = H^{-1}(\Omega)\times L^2(\Omega), \tag{2.10}$$

trivially isometric isomorphic to $\widetilde{\mathcal{E}}^\star$ defined in (2.2), is the space of the permuted data set $(u^1,-u^0)$ compared to (u^0,u^1) of the space $\widetilde{\mathcal{E}}^\star$. The change of order and sign can be deduced from a more elaborate definition of the adjoint wave

operator (see [Ras04, page 112] for a thorough deduction). We introduce the related permutation operator $Q\colon \widetilde{\mathcal{E}}^\star \to \mathcal{E}^\star$ defined by the matrix

$$\boldsymbol{Q} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \text{such that} \quad \begin{pmatrix} u^1 \\ -u^0 \end{pmatrix} = \boldsymbol{Q} \begin{pmatrix} u^0 \\ u^1 \end{pmatrix}.$$

The duality product between $(u^1, -u^0) \in \mathcal{E}^\star$ and $(\varphi^0, \varphi^1) \in \mathcal{E}$ is defined by

$$\begin{aligned} \langle (u^1, -u^0), (\varphi^0, \varphi^1) \rangle_{\mathcal{E}^\star, \mathcal{E}} &= \langle u^1, \varphi^0 \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} - \langle u^0, \varphi^1 \rangle_{L^2(\Omega)} \\ &= \langle (u^0, u^1), (\varphi^0, \varphi^1) \rangle_{\tilde{\mathcal{E}}^\star, \mathcal{E}}. \end{aligned} \tag{2.11}$$

where the duality product $\langle \cdot, \cdot \rangle_{H^{-1}(\Omega), H_0^1(\Omega)}$ is the natural extension[4] of the L^2-inner product (see [Ped00, page 124]).

2.2.2 Duality

The connection between the control system and the adjoint system, which is of fundamental importance to HUM, still needs to be made precise. We will see how the systems are tied together by a Green's formula, Green's 2nd identity (see [Ped00, page 148]).

Proposition 2.11 (Green's 2nd identity). *Let φ and ψ be twice differentiable functions, then the following identity holds*

$$\int_\Omega \big(\psi \Delta\varphi - \psi \Delta\varphi\big) \mathrm{d}x = \int_\Gamma \Big(\psi \frac{\partial \varphi}{\partial n} - \varphi \frac{\partial \psi}{\partial n}\Big) \mathrm{d}s. \tag{2.12}$$

□

We proceed by linking the auxiliary and the adjoint system in the following proposition.

Proposition 2.12. *Let φ be the solution to (2.6) with any initial data $(\varphi^0, \varphi^1) \in \mathcal{E}$ and ψ the solution to (2.5) with any $\kappa \in \mathcal{B}$, then the following identity holds*

$$\langle (\psi'(0,\cdot), -\psi(0,\cdot)), (\varphi^0, \varphi^1) \rangle_{\mathcal{E}^\star, \mathcal{E}} = \int_0^T \int_{\Gamma_0} \kappa \frac{\partial \varphi}{\partial n} \mathrm{d}\Gamma \mathrm{d}t. \qquad \square$$

Proof. We will settle with a sketch of the proof—see [Ped08] for actual calculations. First, assume φ and ψ to be smooth and multiply $\psi'' - \Delta\psi = 0$ by φ and integrate over the space-time domain. Proceed with integration by parts while applying relevant boundary conditions and apply next Green's 2nd identity (2.12). The duality identity then follows from density. ■

Note how the "hidden" L^2-regularity of $\frac{\partial \varphi}{\partial n}$ (Theorem 2.8) is truly crucial for the identity in Proposition 2.12.

A key theorem ([Ped08, Theorem 6.3]) follows immediately from Proposition 2.12.

Theorem 2.13. *The control system (2.1) with initial data $(u^1, -u^0) \in \mathcal{E}^\star$ can be steered to zero in time T if and only if there exists a control $\kappa \in \mathcal{B}$ such that*

$$\langle (u^1, -u^0), (\varphi^0, \varphi^1) \rangle_{\mathcal{E}^\star, \mathcal{E}} = \int_0^T \int_{\Gamma_0} \kappa \frac{\partial \varphi}{\partial n} \mathrm{d}\Gamma \mathrm{d}t \tag{2.13}$$

for all $(\varphi^0, \varphi^1) \in \mathcal{E}$, where φ is the solution to (2.6) for the initial data (φ^0, φ^1). □

[4] The duality product—or duality pairing—$\langle f, g \rangle_{H^{-1}(\Omega), H_0^1(\Omega)}$ is, to be fair, not constructed as an *extension* to the L^2-inner product, but it *reduces* to this case for $f \in L^2(\Omega)$.

Proof. The result follows straightly from Proposition 2.12 since u is exactly controllable if and only if there exists a κ such that $(u^1, -u^0) = (\psi'(0,\cdot), -\psi(0,\cdot))$. ■

2.3 Observability

In the introduction of this chapter, we suggested that controllability and observability are mathematical duals. Before showing that the controllability of (2.1) is equivalent to the observability of (2.6), we need to define the notion of observability.

Loosely speaking, system (2.6) is said to be observable on Γ_0 in time T if the total energy of the system can be measured by a partial measurement on Γ_0 for all $t \leq T$. Section 2.2.1 showed how the energy of the—conservative—adjoint system, defined in (2.9), can conveniently be computed using the energy norm on the initial data $E(0) = \left\|(\varphi^0, \varphi^1)\right\|_{\mathcal{E}}$. We wish to measure this energy in terms of the L^2-regular Neumann data $\partial\varphi/\partial n$ at Γ_0 and introduce to this end the linear *observation operator*

$$\Phi\colon \mathcal{E} \to \mathcal{B} \quad \text{defined by} \quad \Phi(\varphi^0, \varphi^1) = \frac{\partial\varphi}{\partial n}\Big|_{\Gamma_0}, \tag{2.14}$$

which maps the initial data of the adjoint system (2.6) on the Neumann data at Γ_0. We seek, hereafter, to bound the energy of the adjoint system $E(0)$ by the partial measurement $\left\|\Phi(\varphi^0, \varphi^1)\right\|_{\mathcal{B}}$.

Note that, the observation operator is continuous by the "hidden" regularity for the adjoint system (Theorem 2.8), *i.e.,* $\exists c > 0$ such that

$$\left\|\Phi(\varphi^0, \varphi^1)\right\|_{\mathcal{B}} \equiv \left\|\frac{\partial\varphi}{\partial n}\Big|_{\Gamma_0}\right\|_{\mathcal{B}} \leq \left\|\frac{\partial\varphi}{\partial n}\right\|_{L^2(\Sigma)} \leq c\left\|(\varphi^0, \varphi^1)\right\|_{\mathcal{E}}. \tag{2.15}$$

Let us define the observability of the adjoint system and state the observability inequality.

Definition 2.14 (Observability inequality). Let $(\varphi^0, \varphi^1) \in \mathcal{E}$ be initial data for the adjoint system (2.6), and let Φ be the operator defined by (2.14). We say that system (2.6) is observable on Γ_0 in time T if there exists a constant $c(T) > 0$ such that the *observability inequality*

$$\left\|(\varphi^0, \varphi^1)\right\|_{\mathcal{E}} \leq c(T)\left\|\Phi(\varphi^0, \varphi^1)\right\|_{\mathcal{B}}, \tag{2.16}$$

holds for all $(\varphi^0, \varphi^1) \in \mathcal{E}$. The constant $c(T)$ depends on the control time T and is known as the observability constant. □

We see that (2.16) implies the injectivity of Φ. As a consequence of the observability inequality, we can define a new norm on the initial data.

Proposition 2.15. *Assume that the observability inequality* (2.16) *holds and let* Φ *be the operator defined by* (2.14). *Then for* $(\varphi^0, \varphi^1) \in \mathcal{E}$ *the mapping*

$$(\varphi^0, \varphi^1) \mapsto \left\|\Phi(\varphi^0, \varphi^1)\right\|_{\mathcal{B}},$$

from $\mathcal{E}$ *to* $\mathbb{R}$ *defines a norm on the space of initial data* $\mathcal{E}$. *Moreover, this norm is equivalent to the usual norm on* $\mathcal{E}$. □

Proof. The positivity and symmetry is straight forward. The norm is equivalent to the usual norm on $\mathcal{E}$ by the inequalities (2.15) and (2.16). ∎

Before making the link between observability and controllability, we remark on the observability in the 1-d case.

Remark 2.16. Recall from Remark 2.10 that the energy (2.9) in terms of the Fourier coefficients becomes $E(0) = \frac{1}{4}\sum_{k\in\mathbb{N}}(a_k^2k^2\pi^2 + b_k^2)$. The observation data $\Phi(\varphi^0, \varphi^1)$ can be computed by

$$\frac{\partial\varphi}{\partial x}\Big|_{x=1} = \sum_{k\in\mathbb{N}}(-1)^k\big(k\pi a_k\cos(k\pi t) + b_k\sin(k\pi t)\big).$$

If we now consider the case $T = 2$ and utilize the orthogonality of $\cos(k\pi t)$ and $\sin(k\pi t)$ on $L^2(0,2)$ it follows that

$$\big\|\Phi(\varphi^0,\varphi^1)\big\|_{\mathcal{B}}^2 = \int_0^2\left|\frac{\partial\varphi}{\partial x}\Big|_{x=1}\right|^2\mathrm{d}t = \sum_{k\in\mathbb{N}}\big(a_k^2k^2\pi^2 + b_k^2\big).$$

The norm $\big\|\Phi(\varphi^0,\varphi^1)\big\|_{\mathcal{B}}$ must be even greater for $T > 2$. Comparing with the energy $E(0)$ above, we see that the observability inequality (2.16) holds for all $T \geq 2$.

For $T < 2$, however, the adjoint system is *not* observable. If $T = 2 - 2\delta$, with $\delta > 0$, then the part of φ^0 on $(1-\delta, 1) \subset \Omega$ which initially travels left will not make it back to the observation boundary $\Gamma_0 = \{1\}$ in time T due to the unit speed of propagation. ∎

2.4 HUM: the operator approach

The original HUM formulation by J.L. Lions, [Lio88], established the connection between the adjoint system and the auxiliary system by multiplier techniques to form an isomorphic mapping from $\mathcal{E}$ into $\mathcal{E}^\star$. Inspired by [Ped08], we will instead present HUM below in terms of observation and reconstruction.

2.4.1 The reconstruction operator

Recall, from Section 2.1.2, the auxiliary ψ-system that was solved backwards in time. We introduce the *reconstruction operator* Ψ associated with this system

$$\Psi\colon \mathcal{B} \to \mathcal{E}^\star \quad \text{defined by} \quad \Psi\colon \kappa \mapsto (\psi'(0,\cdot), -\psi(0,\cdot)). \tag{2.17}$$

It maps the Dirichlet boundary function κ on the set of (perturbed) "end" data $(\psi'(0,\cdot), -\psi(0,\cdot))$. Any boundary function $\kappa \in \mathcal{B}$ that solves the control problem for initial data $(u^1, -u^0) \in \mathcal{E}^\star$ also satisfies the equation

$$\Psi(\kappa) = (u^1, -u^0). \tag{2.18}$$

Before continuing with specification of the control κ, we state an important fact about the relation between the observation and reconstruction operators.

Proposition 2.17. *Assume that the control system* (2.1) *is exactly controllable. Then the operator*

$$\Psi^* = \Phi\colon \mathcal{E} \to \mathcal{B}$$

is the adjoint of $\Psi\colon \mathcal{B} \to \mathcal{E}^\star$. *Conversely,* $\Phi^* = \Psi$ *is the adjoint of* Φ. □

Proof. Insert $\Psi(\kappa) = (\psi'(0,\cdot), -\psi(0,\cdot))$ in Proposition 2.12 and it follows immediately

$$\begin{aligned}\langle \Psi(\kappa), (\varphi^0, \varphi^1)\rangle_{\mathcal{E}^\star,\mathcal{E}} &= \int_0^T \int_{\Gamma_0} \kappa \frac{\partial \varphi}{\partial n} \mathrm{d}\Gamma \mathrm{d}t \\ &= \langle \kappa, \Phi(\varphi^0, \varphi^1)\rangle_{\mathcal{B}}.\end{aligned}$$

■

2.4.2 The Λ-operator and a main theorem

A HUM control is a special control κ that, in addition to satisfying (2.18), is build from the adjoint system by

$$\text{HUM control:} \qquad \kappa = \Phi(\varphi^0, \varphi^1). \tag{2.19}$$

This is possible only because of the "hidden" regularity (Remark 2.9). By picking the control in this way, we make the remaining pieces fit perfectly together.

Connecting the HUM control (2.19) with the requirement (2.18) leads to the equation

$$\Psi\Phi(\varphi^0, \varphi^1) = (u^1, -u^0). \tag{2.20}$$

The composite operator $\Psi\Phi$ is better known as the Λ operator

$$\Lambda \colon \mathcal{E} \to \mathcal{E}^\star \quad \text{defined by} \quad \Lambda = \Psi \circ \Phi. \tag{2.21}$$

Equation (2.20)—with Λ in place of $\Psi\Phi$—is the heart of the original HUM. If equation (2.20) can be solved, its solution provides the specific set of initial conditions $(\bar\varphi^0, \bar\varphi^1)$ to the adjoint system, that by construction leads to a control $\kappa = \Phi(\bar\varphi^0, \bar\varphi^1)$ for system (2.1). It turns out that (2.20) can be solved if and only if the observability inequality (2.16) holds.

By inserting the HUM control (2.19) and $\Lambda(\varphi^0, \varphi^1)$ in place of $(u^1, -u^0)$ in the duality identity (2.13), we get

$$\langle \Lambda(\varphi^0, \varphi^1), (\varphi^0, \varphi^1)\rangle_{\mathcal{E}^\star,\mathcal{E}} = \int_0^T \int_{\Gamma_0} \left|\frac{\partial \varphi}{\partial n}\right|^2 \mathrm{d}\Gamma \mathrm{d}t \equiv \left\|\Phi(\varphi^0, \varphi^1)\right\|_{\mathcal{B}}^2. \tag{2.22}$$

The primary task in the original proof by J.L. Lions was to prove that (2.22) defines a norm on the set of initial data (φ^0, φ^1) and that this norm is equivalent to the usual norm; this was done by the multiplier $(x - x_0) \cdot n$ as explained in Section 2.1.3. The observability inequality (2.16) together with Theorem 2.13 form an alternative to the original proof, and by Proposition 2.15 we have the same norm equivalence for $\left\|\Phi(\varphi^0, \varphi^1)\right\|_{\mathcal{B}}$. This also means that (2.22) forms a norm on $\mathcal{E}$ and Λ is thereby a self-adjoint and positive operator. Furthermore, by Riesz representation theorem, we have that Λ is an isomorphism from $\mathcal{E}$ onto $\mathcal{E}^\star$ (see [Ped00, page 220]).

We summarize the above in a main theorem.

Theorem 2.18. *The control system* (2.1) *is exactly controllable on* $\mathcal{E}$ *in time* T *if and only if its adjoint system* (2.6) *is observable on* Γ_0 *in time* T.

If the adjoint system is observable, then the HUM operator equation (2.20) *has a unique solution* $(\bar\varphi^0, \bar\varphi^1)$ *and* $\kappa = \Phi(\bar\varphi^0, \bar\varphi^1)$ *defines a control for the control problem; this control is of minimal* L^2*-norm. Furthermore,* Λ *is a positive and self-adjoint operator, and it forms an isomorphism from* $\mathcal{E}$ *onto* $\mathcal{E}^\star$. □

We conclude this section by defining a bounded linear *controllability operator* $\Pi\colon \mathcal{E}^\star \to \mathcal{B}$, provided that $\Lambda = \Phi^*\Phi$ is invertible, by

$$\Pi(u^1, -u^0) = (\Phi(\Phi^*\Phi)^{-1})(u^1, -u^0). \tag{2.23}$$

Hence, if the adjoint system is observable the operator Π maps the initial data $(u^1, -u^0)$ of the control system onto the sought control $\kappa = \Pi(u^1, -u^0)$ thereby providing its solution.

2.4.3 Constructing Λ as a matrix

The HUM key-operators Λ, Φ, and Ψ may be represented as *in*-finite dimensional matrices $\mathbf{\Lambda}$, $\mathbf{\Phi}$, and $\mathbf{\Psi}$. We construct these matrices as a first step in the direction of finite dimensional approximation.

Let, for the separable Hilbert spaces $\mathcal{E}$ and $\mathcal{E}^\star$, the vectors $\{e_j\}_{j\in\mathbb{N}}$ form a basis for $\mathcal{E}$ and $\{e'_j\}_{j\in\mathbb{N}}$ a basis for $\mathcal{E}^\star$. We expand the initial data (φ^0, φ^1) and $(u^1, -u^0)$ in these bases

$$\begin{pmatrix}\varphi^0\\ \varphi^1\end{pmatrix} = \sum_{j\in\mathbb{N}} \boldsymbol{\varphi}_j e_j, \qquad \begin{pmatrix}u^1\\ -u^0\end{pmatrix} = \sum_{j\in\mathbb{N}} \boldsymbol{u}_j e'_j,$$

where $\boldsymbol{\varphi}$ and $\boldsymbol{u}$ are infinite column vectors of coefficients. We assume further that we have the orthogonality property

$$\langle e'_j, e_i\rangle_{\mathcal{E}^\star,\mathcal{E}} = \delta_{ij}, \qquad i,j \in \mathbb{N}.$$

The above expansions and the orthogonality give the following equivalence between the operator equation (2.20) and a matrix equation

$$\begin{aligned}
\Lambda\begin{pmatrix}\varphi^0\\ \varphi^1\end{pmatrix} = \begin{pmatrix}u^1\\ -u^0\end{pmatrix} \quad &\iff \quad \sum_{j\in\mathbb{N}} \boldsymbol{\varphi}_j \Lambda e_j = \sum_{i\in\mathbb{N}} \boldsymbol{u}_i e'_i \\
&\iff \quad \langle \sum_{j\in\mathbb{N}} \boldsymbol{\varphi}_j \Lambda e_j, e_i\rangle_{\mathcal{E}^\star,\mathcal{E}} = \langle \sum_{k\in\mathbb{N}} \boldsymbol{u}_k e'_k, e_i\rangle_{\mathcal{E}^\star,\mathcal{E}}, \ \ \forall i \in \mathbb{N} \\
&\iff \quad \sum_{j\in\mathbb{N}} \langle \Lambda e_j, e_i\rangle_{\mathcal{E}^\star,\mathcal{E}} \boldsymbol{\varphi}_j = \boldsymbol{u}_i, \ \ \forall i \in \mathbb{N} \\
&\iff \quad \mathbf{\Lambda}\boldsymbol{\varphi} = \boldsymbol{u},
\end{aligned} \tag{2.24}$$

where the matrix $\mathbf{\Lambda}$ is defined by

$$\mathbf{\Lambda}_{ij} = \langle \Lambda e_j, e_i\rangle_{\mathcal{E}^\star,\mathcal{E}}, \qquad i,j \in \mathbb{N}. \tag{2.25}$$

We call this way of constructing the matrix $\mathbf{\Lambda}$ *direct assembling* to distinguish it from the alternative that we will present below. Direct assembly involves first a solution of the adjoint problem (observation) and then a solution to the auxiliary problem (reconstruction) for each basis function e_j.

If we instead use the factorization $\Lambda = \Phi^*\Phi$ (from (2.20) and Proposition 2.17), we can express the matrix element $\mathbf{\Lambda}_{ij}$ by

$$\mathbf{\Lambda}_{ij} = \langle \Phi e_j, \Phi e_i\rangle_{\mathcal{B}}, \qquad i,j \in \mathbb{N}, \tag{2.26}$$

which we, as a method, denote *inner product assembling*. This alternative procedure only requires the solution of the adjoint system and only computation of half the entries as we have symmetry $\langle \Phi e_j, \Phi e_i\rangle_{\mathcal{B}} = \langle \Phi e_i, \Phi e_j\rangle_{\mathcal{B}}$ by construction.

Remark 2.19. The entries of the infinite matrix $\mathbf{\Lambda}$ can in 1-d be computed analytically in the Fourier basis $\left\{\sqrt{2}\sin(j\pi x)\right\}_{j\in\mathbb{N}}$. The case $T = 2n$ where $n \in \mathbb{N}$ is particularly simple; it results in a diagonal $\mathbf{\Lambda}$ and can be found in Remark 2.22. The computations are cumbersome when $T \neq 2n$ (see [Ras04, page 185] for details) and the result is generally a full matrix $\mathbf{\Lambda}$. ∎

2.4.4 Constructing Φ and Ψ as matrices

We will construct Φ and Ψ as matrices in the same manner as we did for Λ. To this end, we propose an orthonormal basis $\{b_j\}_{j\in\mathbb{N}}$ for $\mathcal{B}$ in which we can expand functions $\kappa \in \mathcal{B}$

$$\kappa = \sum_{j\in\mathbb{N}} \boldsymbol{\kappa}_j b_j,$$

where $\boldsymbol{\kappa}$ is the infinite vector of the coefficients with respect to this basis. We rewrite the operator equation $\kappa = \Phi(\varphi^0, \varphi^1)$ as a matrix equation

$$\begin{aligned}
\Phi\begin{pmatrix}\varphi^0\\ \varphi^1\end{pmatrix} = \kappa \quad &\iff \quad \sum_{j\in\mathbb{N}} \boldsymbol{\varphi}_j \Phi e_j = \sum_{i\in\mathbb{N}} \boldsymbol{\kappa}_i b_i \\
&\iff \quad \langle \sum_{j\in\mathbb{N}} \boldsymbol{\varphi}_j \Phi e_j, b_i\rangle_{\mathcal{B}} = \langle \sum_{k\in\mathbb{N}} \boldsymbol{\kappa}_k b_k, b_i\rangle_{\mathcal{B}} \ \ \forall i \in \mathbb{N} \\
&\iff \quad \sum_{j\in\mathbb{N}} \langle \Phi e_j, b_i\rangle_{\mathcal{B}} \boldsymbol{\varphi}_j = \boldsymbol{\kappa}_i, \ \ \forall i \in \mathbb{N} \\
&\iff \quad \mathbf{\Phi}\boldsymbol{\varphi} = \boldsymbol{\kappa}, \qquad (2.27)
\end{aligned}$$

where $\mathbf{\Phi}$ is defined by

$$\mathbf{\Phi}_{ij} = \langle \Phi e_j, b_i\rangle_{\mathcal{B}}, \qquad i, j \in \mathbb{N}. \qquad (2.28)$$

The same procedure goes for the operator equation $\Psi\,\kappa = (\psi'(0,\cdot), -\psi(0,\cdot))$

$$\begin{aligned}
\Psi\,\kappa = \begin{pmatrix}\psi'(0,\cdot)\\ -\psi(0,\cdot)\end{pmatrix} \quad &\iff \quad \sum_{j\in\mathbb{N}} \boldsymbol{\kappa}_j \Psi b_j = \sum_{i\in\mathbb{N}} \boldsymbol{\psi}_i e_i \\
&\iff \quad \langle \sum_{j\in\mathbb{N}} \boldsymbol{\kappa}_j \Psi b_j, e_i\rangle_{\mathcal{E}^\star,\mathcal{E}} = \langle \sum_{k\in\mathbb{N}} \boldsymbol{\psi}_k e_k, e_i\rangle_{\mathcal{E}^\star,\mathcal{E}} \ \ \forall i \in \mathbb{N} \\
&\iff \quad \sum_{j\in\mathbb{N}} \langle \Psi b_j, e_i\rangle_{\mathcal{E}^\star,\mathcal{E}} \boldsymbol{\kappa}_j = \boldsymbol{\psi}_i, \ \ \forall i \in \mathbb{N} \\
&\iff \quad \mathbf{\Psi}\boldsymbol{\kappa} = \boldsymbol{\psi}, \qquad (2.29)
\end{aligned}$$

where $\mathbf{\Psi}$ is defined by

$$\mathbf{\Psi}_{ij} = \langle \Psi b_j, e_i\rangle_{\mathcal{E}^\star,\mathcal{E}}, \qquad i, j \in \mathbb{N}. \qquad (2.30)$$

Note, that picking $\{\Phi e_j\}_{j\in\mathbb{N}}$ as the basis $\{b_j\}_{j\in\mathbb{N}}$ in (2.28) amounts to constructing Λ by inner product assembly; using the same basis in (2.26) amounts to constructing Λ by direct assembly.

Remark 2.20. The operator $\Phi\colon \mathcal{E} \to \mathcal{B}$ can be considered as two sub-operators

$$\Phi^0\colon H_0^1(\Omega) \to \mathcal{B} \quad \text{defined by} \quad \Phi^0\varphi^0 = \Phi(\varphi^0, 0) \qquad (2.31a)$$

$$\Phi^1\colon L^2(\Omega) \to \mathcal{B} \quad \text{defined by} \quad \Phi^1\varphi^1 = \Phi(0, \varphi^1). \qquad (2.31b)$$

The division may be passed on to the matrix representation $\mathbf{\Phi} = [\mathbf{\Phi}^0, \mathbf{\Phi}^1]$. The same can be done for the operator Ψ and its matrix representation $\mathbf{\Psi}$. ■

Remark 2.21. The 1-d observation of the Fourier sine basis $\{\sqrt{2}\sin(j\pi x)\}_{j\in\mathbb{N}}$ is particularly simple. Let us consider the Φ^0 observation (2.31a) first. The initial data $(\varphi^0, \varphi^1) = (\sqrt{2}\sin(j\pi x), 0)$ leads to the solution

$$\varphi(t,x) = \sqrt{2}\cos(j\pi t)\sin(j\pi x), \qquad j \in \mathbb{N},$$

of the adjoint system (2.6). By taking the normal derivative at $x = 1$ we obtain

$$\Phi^0\big(\sqrt{2}\sin(j\pi x)\big) = (-1)^j\sqrt{2}j\pi\cos(j\pi t), \qquad j \in \mathbb{N}.$$

Correspondingly, we get the solution $\varphi(t,x) = \frac{\sqrt{2}}{j\pi}\sin(j\pi t)\sin(j\pi x)$ for the initial data $(\varphi^0, \varphi^1) = (0, \sqrt{2}\sin(j\pi x))$ which leads to the following Φ^1 observation

$$\Phi^1\big(\sqrt{2}\sin(j\pi x)\big) = (-1)^j\sqrt{2}\sin(j\pi t), \qquad j \in \mathbb{N}.$$

Notice the orthogonality of these observations in the special case $T = 2n$ for $n \in \mathbb{N}$. ■

Remark 2.22. The 1-d sine basis observation from Remark 2.21 has a simple diagonal matrix representation when $T = 2n$ for $n \in \mathbb{N}$. We define the functions

$$b_i^0 = \sqrt{\frac{2}{T}}\cos(i\pi t), \quad b_i^1 = \sqrt{\frac{2}{T}}\sin(i\pi t), \qquad i \in \mathbb{N}$$

and recognize that $\{b_i^0, b_i^1\}_{i\in\mathbb{N}}$ constitute an orthonormal basis for $\mathcal{B}$. We consider the construction of the matrix $\mathbf{\Phi} = [\mathbf{\Phi^0}, \mathbf{\Phi^1}]$. Let in the following index i correspond to the first half of the rows in each of the matrices $\mathbf{\Phi}^0$ and $\mathbf{\Phi}^0$, and let k correspond to the second half of the rows. Then we have the elements

$$\mathbf{\Phi}_{ij}^0 = \Big\langle (-1)^j\sqrt{2}j\pi\cos(j\pi t),\, b_i^0 \Big\rangle_{\mathcal{B}} = (-1)^j j\pi\sqrt{T}\delta_{ij}, \quad i,j \in \mathbb{N}, \tag{2.32a}$$

$$\mathbf{\Phi}_{kj}^1 = \Big\langle (-1)^j\sqrt{2}\sin(j\pi t),\, b_k^1 \Big\rangle_{\mathcal{B}} = (-1)^j\sqrt{T}\delta_{kj}, \quad k,j \in \mathbb{N}. \tag{2.32b}$$

It is trivial to extend this to $\mathbf{\Lambda}$ which also will have diagonal structure. The first half $j \in \mathbb{N}$ of the diagonal becomes $\mathbf{\Lambda}_{jj} = (j\pi)^2 T$ and the second half $k \in \mathbb{N}$ becomes $\mathbf{\Lambda}_{kk} = T$. ■

2.5 HUM: The minimization approach

HUM has in recent years been studied in an alternative formulation to the operator approach described in the Section 2.4. This alternative is formulated as a minimization problem from which the sought initial data set $(\bar{\varphi}^0, \bar{\varphi}^1)$, defining the HUM-control κ, emerges as the minimizer of a certain energy functional. One of the advantages of this more recent formulation is that it only depends on the adjoint system—and not on the auxiliary one. The exposition in this section has been greatly inspired by [MZ05]. Note that in this section the distinction between $\mathcal{E}^\star$ and $\widetilde{\mathcal{E}}^\star$ is not important. For this reason we use initial data in the

usual order, *i.e.*, $(u^0, u^1) \in \widetilde{\mathcal{E}}^\star$.

We introduce the energy functional $\mathcal{J}\colon \mathcal{E} \to \mathbb{R}$

$$\mathcal{J}(\varphi^0, \varphi^1) := \frac{1}{2} \left\| \Phi(\varphi^0, \varphi^1) \right\|_{\mathcal{B}}^2 - \langle (u^0, u^1), (\varphi^0, \varphi^1) \rangle_{\tilde{\mathcal{E}}^\star, \mathcal{E}}, \tag{2.33}$$

where Φ is the observation operator (2.14), $(u^0, u^1) \in \widetilde{\mathcal{E}}^\star$ is the initial data for the control system (2.1), and $\langle \cdot, \cdot \rangle_{\tilde{\mathcal{E}}^\star, \mathcal{E}}$ is the duality product defined by (2.11).

We wish to show that the functional $\mathcal{J}$ attains its minimum at $(\bar{\varphi}^0, \bar{\varphi}^1)$ in $\mathcal{E}$ and that this minimum produces the sought control by $\kappa = \Phi(\bar{\varphi}^0, \bar{\varphi}^1)$.

Theorem 2.23. *Let $(u^0, u^1) \in \widetilde{\mathcal{E}}^\star$ and let $\mathcal{J}$ be the functional defined by (2.33). Suppose that $\mathcal{J}$ has a unique minimum at $(\bar{\varphi}^0, \bar{\varphi}^1) \in \mathcal{E}$. If $\bar{\varphi}$ is the solution of (2.6) for the initial conditions $(\bar{\varphi}^0, \bar{\varphi}^1)$, then $\kappa = \partial\bar{\varphi}/\partial n\big|_{\Gamma_0}$ defines the control that will steer (2.1) to zero for the initial conditions (u^0, u^1).* □

Proof. By assumption $\mathcal{J}$ has its (unique) minimum at $(\bar{\varphi}^0, \bar{\varphi}^1)$. This means that

$$\begin{aligned} 0 &= \lim_{h \to 0} \frac{1}{h} \Big(\mathcal{J}\big((\bar{\varphi}^0, \bar{\varphi}^1) + h(\varphi^0, \varphi^1)\big) - \mathcal{J}(\bar{\varphi}^0, \bar{\varphi}^1) \Big) \\ &= \int_0^T \int_{\Gamma_0} \frac{\partial \bar{\varphi}}{\partial n} \frac{\partial \varphi}{\partial n} \mathrm{d}s\mathrm{d}t - \langle (u^0, u^1), (\varphi^0, \varphi^1) \rangle_{\tilde{\mathcal{E}}^\star, \mathcal{E}} \end{aligned}$$

for any $(\varphi^0, \varphi^1) \in \mathcal{E}$. Then by $\kappa = \partial\bar{\varphi}/\partial n\big|_{\Gamma_0}$ in Theorem 2.13 the proof is complete. ■

Now that we know that a minimizer, if it exists, provide the control that we seek, we need a condition on the existence of a unique minimizer. The requirement is, again, that the adjoint system is observable.

Theorem 2.24. *Let $(u^0, u^1) \in \widetilde{\mathcal{E}}^\star$ be given. Assume that the adjoint system (2.6) is observable on Γ_0 in time T. Then the functional $\mathcal{J}$ defined by (2.33) has a unique minimizer $(\bar{\varphi}^0, \bar{\varphi}^1) \in \mathcal{E}$.* □

Proof. The so-called *direct method of calculus of variations* tells us that $\mathcal{J}$ has a minimum in $\mathcal{E}$ provided that its a) convex, b) lower semi-continuous, and c) coercive; the minimum is unique if $\mathcal{J}$ is strictly convex.

The continuity follows from the "hidden" regularity (Remark 2.9), whereas the observability inequality can be used to establish the coercivity and strict convexity. See [MZ05] for details. ■

Not only do we have a means for obtaining a control, it can also be proved (see [MZ05]) that the found control is of minimal L^2-norm.

Proposition 2.25. *Let $\kappa = \Phi(\bar{\varphi}^0, \bar{\varphi}^1)$ be the control given by minimization of $\mathcal{J}$. Then $\kappa \in \mathcal{B}$ is the control of minimal $\mathcal{B}$-norm.* □

To summarize the findings of this section, system (2.1) can, for any initial data $(u^0, u^1) \in \widetilde{\mathcal{E}}^\star$, be controlled exactly in time T by a control $\kappa \in \mathcal{B}$ if and only if its adjoint system (2.6) is observable on $\mathcal{E}$ in time T on Γ_0. If (2.6) is observable, then the energy functional $\mathcal{J}$ has a unique minimizer $(\bar{\varphi}^0, \bar{\varphi}^1)$ which by $\kappa = \Phi(\bar{\varphi}^0, \bar{\varphi}^1)$ defines the control that will steer the state of (2.1) exactly to zero in time T.

CHAPTER 3

Approximating solutions to the wave equation

The main objective of this dissertation is to seek "good" numerical approximations to the HUM-control of the wave equation. Achieving this goal primarily involves two tasks:

(1) finding approximate solutions to the wave equation,

(2) approximating HUM itself.

Item (2) is the topic of the next chapter; this chapter is devoted to (1).

We find approximate solutions to the wave equation by the so-called *method of lines* in which the discretization of the spatial and temporal part of the system is dealt with separately. In practice, this means that we discretize the spatial part of the wave equation first. The continuous solution $y(t,\cdot)$ is approximated by $y_h(t,\cdot)$, where h denote a characteristic length of the spatial discretization. The function $y_h(t,\cdot)$ is represented discretely by a vector $\boldsymbol{y}_h(t)$ of, *e.g.,* nodal values of $y_h(t,\cdot)$ which turns the PDE into a system of ODEs in $\boldsymbol{y}_h(t)$. Secondly, a time integration, or time-stepping, procedure takes care of the temporal part of the system leading to a fully discrete solution.

In this chapter, we consider the 1-d wave equation as it will be the model problem to which we shall apply control in the following chapter. The existence of analytic solutions for this simple 1-d model, which we shall study in Section 3.1, is a clear advantage when we consider numerical approximation of the control in Chapter 4.

After a general discussion of how to obtain approximate solutions to the 1-d wave equation, we turn to the main subject of this chapter which is the discontinuous Galerkin finite element method. It is a recent method, compared to the classical finite difference and finite element methods, and it is well-suited for wave problems It has, nevertheless, not previously been used in the context of HUM control for the wave equation.

We begin by reviewing the continuous 1-d wave equation and its solution in Section 3.1. The classical semi-discretizations will be treated in Section 3.2 before we, in Section 3.3, consider the discontinuous Galerkin FEM. Section 3.4 deals with time integration and Section 3.5 finally concludes this chapter with an analysis of the dispersive properties of the introduced methods.

3.1 The continuous 1-d wave equation

We need to solve the wave equations (2.6) and (2.5), which have respectively homogeneous and inhomogeneous boundary conditions, in order to solve the HUM control problem. In Chapter 5, we shall consider a wave equation with a forcing term. To encompass them all, we consider the 1-d wave equation on $\Omega = (0, 1)$

$$y'' - \frac{\partial^2}{\partial x^2} y = f, \qquad \text{in } (0,T) \times \Omega \tag{3.1a}$$

$$y(t,0) = g_0(t), \quad y(t,1) = g_1(t), \qquad t \in (0,T) \tag{3.1b}$$

$$y(0,x) = y^0(x), \quad y'(0,x) = y^1(x). \qquad x \in \Omega, \tag{3.1c}$$

where f is a forcing term, the functions (y^0, y^1) are initial data, and g_0 and g_1 are Dirichlet boundary conditions. It is well-known that for sufficiently smooth data $\{y^0, y^1, g_0, g_1, f\}$ the system (3.1) is well-posed and has a unique, classical solution $y \in C^2([0,T] \times \Omega)$ (see, *e.g.,* [Eva98]).

It is also well-known, that the notion of solutions to (3.1) can be extended to so-called weak solutions. These do not hold pointwise as their classical counterparts but in an integrated form instead. The introduction of weak formulations open up for a wide range of numerical methods.

Let us for a moment, to simplify the exposition, assume $g_0 = g_1 = 0$. We multiply (3.1a) by a smooth test function $v \in \mathcal{V}$ and integrate over the domain

$$\int_\Omega (y'' - \frac{\partial^2 y}{\partial x^2}) v \mathrm{d}x = \int_\Omega f v \mathrm{d}x, \quad \forall v \in \mathcal{V}.$$

Integration by parts leads to the variational problem: find y with values $y(t,\cdot) \in H_0^1(\Omega)$ such that

$$\langle y'', v \rangle + a(y,v) = \langle f, v \rangle, \quad \forall v \in H_0^1(\Omega), \tag{3.2}$$

where $H_0^1(\Omega)$ is the space of test functions $\mathcal{V}$ and $a(\cdot,\cdot)$ is a symmetric, bi-linear, coercive form bounded on $H_0^1(\Omega)$ defined by

$$a(y,v) := \Big\langle \frac{\partial y}{\partial x}, \frac{\partial v}{\partial x} \Big\rangle. \tag{3.3}$$

A solution y that satisfies (3.2) is then called a weak solution to (3.1a). A classical solution to (3.1a) is also a weak solution. This variational formulation is the foundation for semi-discretizations such as the finite element method as we shall see in Section 3.2.2.

Non-homogeneous Dirichlet boundary conditions are imposed in the definition of the solution space; the test space $\mathcal{V}$ is defined with the corresponding homogeneous Dirichlet condition.

Solutions for (3.1) with general non-smooth initial and boundary data are naturally defined in weak sense as we saw already in Theorem 2.1 and Theorem 2.8. For wave equations with L^2-regular forcing term and boundary conditions, we generally have unique weak solutions $y \in C([0,T];\mathcal{H})$ when $y_0 \in \mathcal{H}$, that is, if the initial displacement y_0 is $H^1(\Omega)$ we have time-continuous y with values $y(t,\cdot)$ in that space (see, *e.g.,* [Ped00]).

3.1.1 The wave equation as a conservation law

Equation (3.1a) is not the only way to express the wave equation in 1-d. It may, like a huge class of other PDEs, be formulated as a conservation law. A conservation law is a mathematical formulation that originates in a certain symmetry of the corresponding physical system. In scalar form, it reads

$$\frac{\partial y}{\partial t} + \frac{\partial \varrho(y)}{\partial x} = f,$$

where ϱ is a flux function. Note that we here, and in the rest of Section 3.1.1, use $\frac{\partial y}{\partial t}$ to denote the temporal derivative of y instead of y' due to typographical concerns. Common examples conforming to this format are conservation of energy and conservation of angular momentum. The discontinuous Galerkin method—the topic of Section 3.3—and other semi-discretization methods are formulated for conservation laws. We will introduce the wave equation as a system of conservation laws for later use.

Let $(y,z) \in [C^2([0,T]\times\Omega)]^2$ be a solution to

$$\frac{\partial}{\partial t}\begin{bmatrix} y \\ z \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\frac{\partial}{\partial x}\begin{bmatrix} y \\ z \end{bmatrix} = 0, \tag{3.4}$$

where z is an auxiliary variable and the forcing term f has been omitted to simplify the following exposition. A function $y \in C^2([0,T]\times\Omega)$ satisfies (3.4) if and only if y is the solution to (3.1a) which can be verified by differentiation of the first equation with t and the second with x.

This conservation law is a coupled system of advection equations. It can be de-coupled by eigen-decomposition of the symmetric system matrix

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \boldsymbol{S}\boldsymbol{D}\boldsymbol{S}^{\mathsf{T}}, \quad \text{with } \boldsymbol{S} = \frac{\sqrt{2}}{2}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}, \text{ and } \boldsymbol{D} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix},$$

which allows us to write the wave equation de-coupled as one left- and one right-going advection equation

$$\frac{\partial}{\partial t}\begin{bmatrix} p \\ q \end{bmatrix} + \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \frac{\partial}{\partial x}\begin{bmatrix} p \\ q \end{bmatrix} = 0. \tag{3.5}$$

The new variables p and q relate to y and z by the orthogonal transformation

$$\begin{bmatrix} p \\ q \end{bmatrix} = \boldsymbol{S}^{\mathsf{T}} \begin{bmatrix} y \\ z \end{bmatrix}.$$

We use this transformation to connect the initial data (p^0, q^0) for (3.5) with the initial data (y^0, y^1) of (3.1c) via

$$\begin{bmatrix} p^0 \\ q^0 \end{bmatrix} = \boldsymbol{S} \begin{bmatrix} y^0 \\ z^0 \end{bmatrix}, \text{ where } z^0(x) = -\int_0^x y^1(s)\mathrm{d}s + z^0(0). \tag{3.6}$$

The expression for z^0 is deduced from the first equation $\frac{\partial}{\partial t}y + \frac{\partial}{\partial x}z = 0$ in (3.4).

Each advection equation in (3.5) requires one boundary condition. The solution p is left-bound and we therefore need a condition at $x = 1$. The right-bound q requires, conversely, a condition at $x = 0$. By the $\boldsymbol{S}$ transformation and the boundary conditions (3.1b) we have

$$\begin{aligned} y(t,0) &= \tfrac{\sqrt{2}}{2}(-p(t,0) + q(t,0)) = g_0(t), \\ y(t,1) &= \tfrac{\sqrt{2}}{2}(-p(t,1) + q(t,1)) = g_1(t), \end{aligned}$$

which gives us the boundary conditions for (3.5) as

$$q(t,0) = p(t,0) + \sqrt{2}g_0(t), \tag{3.7a}$$

$$p(t,1) = q(t,1) - \sqrt{2}g_1(t). \tag{3.7b}$$

These equations clearly show that we have moved the coupling from equation level in system (3.4) to the level of boundary conditions in system (3.5). This allows geometric construction of the solutions and better numerical solutions, too.

Note, finally, that the notion of solutions to (3.5) may be extended from classical to weak solutions like we saw for the classical wave equation above. Before returning to this point, we shall see how the characteristic solutions p and q can be constructed.

Constructing solutions p and q

Solutions to $\frac{\partial}{\partial t}p - \frac{\partial}{\partial x}p = 0$ are left-bound and have the form $p(t,x) = \widetilde{p}(t+x)$ as can easily be verified by insertion. This means that the solution on any line $t + x = c$ is constant. The line $t + x = 1$ separates the solution area in two domains: the first $t + x < 1$ is the domain of dependence of the initial data $p(0,x) = p^0(x)$, the second $t + x > 1$ is only influenced by the boundary condition $p(t,1) = h_1(t)$. We can write the solution

$$p(t,x) = \begin{cases} p^0(t+x), & t + x < 1 \\ h_1(t+x-1), & 1 < t + x. \end{cases}$$

Solutions to $\frac{\partial}{\partial t}q + \frac{\partial}{\partial x}q = 0$, on the other hand, are right-bound and of the form $q(t-x)$. In this case it is the line $t - x = 0$ that separates initial data $q(0,x) = q^0(x)$ from boundary data $q(t,0) = h_0(t)$. The solution is

$$q(t,x) = \begin{cases} q^0(-(t-x)), & t - x < 0 \\ h_0(t-x), & 0 < t - x. \end{cases}$$

The solutions p and q are coupled via the boundary conditions as we saw in (3.7). Let $g_0 = 0$ as this is the case in the control problem (2.1) and it simplifies the derivation below. This leaves us with the reduced coupling conditions

$$p(t,1) = q(t,1) - \sqrt{2}g_1(t) \qquad \text{and} \qquad q(t,0) = p(t,0).$$

With this information at hand we are now able to construct the solutions ge-

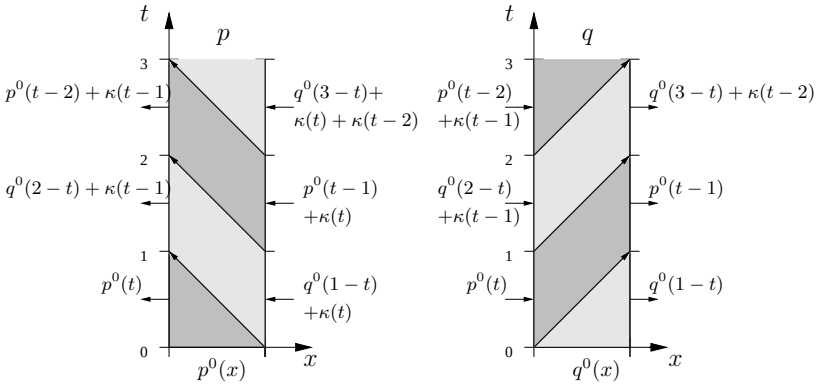


Figure 3.1: Construction of solutions to the advection equations in p (left side) and q (right side). Solution p gets information from the bottom ($t = 0$) and the right endpoint ($x = 1$) and sends this data to the left ($x = 0$) where it is passed to q. Solution q gets data from the bottom ($t = 0$) and the left side ($x = 0$) and sends it all to the right ($x = 1$) where it is passed to p. Dark shaded areas "carries" p^0 whereas the lighter shaded "carries" q^0.

ometrically as sketched in Figure 3.1. At time $t = 0$ the solution p is p^0 on $0 < x < 1$, but we also know that p is constant on any line $t + x = c$ which means that we know p on $t + x < 1$. The knowledge about p on $x = 0$ at $0 < t < 1$ gives us the boundary condition for q on the same line, which again defines q in $0 < t - x < 1$ and so on. In this way we are able to construct p and q step-by-step in the whole time-space domain as follows

$$p(t,x) = \begin{cases} p^0(t+x), & t+x < 1 \\ q^0(2-(t+x)) - \sqrt{2}g_1(t+x-1), & 1 < t+x < 2 \\ p^0(t+x-2) - \sqrt{2}g_1(t+x-1), & 2 < t+x < 3 \\ q^0(4-(t+x)) - \sqrt{2}\big(g_1(t+x-1) + g_1(t+x-3)\big), & 3 < t+x < 4 \\ \cdots & \cdots \end{cases}$$

and

$$q(t,x) = \begin{cases} q^0(-(t-x)), & t-x<0 \\ p^0(t-x), & 0<t-x<1 \\ q^0(2-(t-x)) - \sqrt{2}g_1(t-x-1), & 1<t-x<2 \\ p^0(t-x-2) - \sqrt{2}g_1(t-x-1), & 2<t-x<3 \\ \dots & \dots \end{cases}$$

In the homogeneous case, $g_0 = g_1 = 0$, we can even write the solutions for all $t \geq 0$ in short, closed form

$$p(t,x) = \begin{cases} p^0(t+x-\lfloor t+x \rfloor), & \lfloor t+x \rfloor \text{ even} \\ q^0(\lceil t+x \rceil - (t+x)), & \lfloor t+x \rfloor \text{ odd} \end{cases}$$

and

$$q(t,x) = \begin{cases} q^0(\lceil t-x \rceil - (t-x)), & \lceil t-x \rceil \text{ even} \\ p^0(t-x-\lfloor t-x \rfloor), & \lceil t-x \rceil \text{ odd} \end{cases}$$

where the floor function $x \mapsto \lfloor x \rfloor$ maps $x \in \mathbb{R}$ to largest integer not greater than x and the ceiling function $x \mapsto \lceil x \rceil$ maps $x \in \mathbb{R}$ to the smallest integer not less than x.

We can easily allow L^2-regular initial and boundary data in our construction of p and q. The solutions are unique and constant on the characteristic lines; everything—including possible discontinuities—are propagated along these lines. If we have the data $p^0, q^0 \in L^2(\Omega)$ then p and q will be weak solutions to (3.5).

Let us summarize: If (y^0, y^1) is given as initial data to (3.1) we may obtain the data (p^0, q^0) for the corresponding system in characteristic variables by (3.6). Then, after solving the system in p and q, we restore the solution y by $y = \frac{\sqrt{2}}{2}(-p+q)$.

3.2 Classical semi-discretizations

This section deals with semi-discretizations of the wave equation (3.1) with the purpose of determining approximate solutions to y. Let in the following $y_h \colon (0,T) \times \Omega \to \mathbb{R}$ denote an approximation to y where h refers to a characteristic length of the spatial discretization. This gives rise to two fundamental questions:

1. how do we represent the approximate solution y_h ?

2. in which way should y_h satisfy the PDE ?

Typically, we seek y_h in a finite dimensional subspace of the solution space and express it by a series of either basis functions for the approximation space or by interpolating Lagrange polynomials. To these series correspond solution vectors, marked with bold, $\boldsymbol{y}$ containing coefficients (modes) or nodal values at grid points.

We can require the approximate solution y_h to satisfy the wave equation in several different ways. For this characterization, it is convenient to consider the

residual[1]

$$\mathcal{R}_h(t,x) = y_h''(t,x) - \Delta y_h(t,x) - f(t,x). \tag{3.8}$$

If the residual is zero for all $(t,x) \in (0,T) \times \Omega$ then y_h is the exact solution. *Collocation* is an approach requiring y_h to satisfy $\mathcal{R}_h(t,x_i) = 0$ at a set of discrete points x_i. These interpolation points are called collocation points. We will see an example of this below when considering the finite difference method. A competing approach is the *Galerkin* method and its relatives (*e.g.,* Petrov-Galerkin) where one seeks solutions y_h such that the residual is orthogonal to all test functions v in some finite dimensional test space $\mathcal{V}_h$. The finite element method is most often build on this type of requirement.

We will briefly revise two classical methods, the finite difference method (FDM) and the finite element method (FEM), below. We will also very briefly discuss other important methods such as the spectral method and the finite volume method in Section 3.2.4. All these methods will serve as a source of reference and an introduction to important semi-discretizations concepts before moving on to discontinuous Galerkin FEM in Section 3.3.

3.2.1 The finite difference method (FDM)

The basic idea of the FDM is to replace continuous derivatives in the PDE with linear combinations of discrete function values called finite differences. This simple, intuitive idea is one of the main advantages of this method.

First step is to introduce an equidistant grid on $\Omega = (0,1)$ with N inner grid points

$$0 = x_0, x_1, \ldots, x_N, x_{N+1} = 1, \tag{3.9}$$

with spacing $h = 1/(N+1)$. Then identify the approximate solution y_h with the vector $\boldsymbol{y}$ containing N discrete values for fixed t

$$\boldsymbol{y}(t) = [y_1(t), \ldots, y_N(t)]^{\mathsf{T}},$$

where $y_i(t) = y_h(t,x_i)$ for $i = 1, \ldots, N$. Notice that the values at the endpoints are not included as they are defined by the Dirichlet boundary conditions (3.1b). Local polynomial interpolation, *e.g.,* $y_h(t,x) = \sum_{i=0}^{2} \alpha_i(t)(x - x_k)^i$ for $x \in [x_{k-1}, x_{k+1}]$, may be used for reconstruction. The forcing term f is sampled on the same grid resulting in the vector $\boldsymbol{f}(t) = [f_1(t), \ldots, f_N(t)]^{\mathsf{T}}$, where $f_i(t) = f(t,x_i)$ for $i = 1, \ldots, N$.

The second step is the approximation of the 1-d Laplacian $-\partial^2/\partial x^2$ by finite differences. If we choose 2nd order central finite differences, around x_i, in (3.1a) we obtain the following N difference equations

$$y_i''(t) - \frac{y_{i+1}(t) - 2y_i(t) + y_{i-1}(t)}{h^2} = f_i(t), \qquad i = 1, \ldots, N, \quad t \in (0,T).$$

By assuming homogeneous boundary conditions $y(t,0) = y(t,1) = 0$ for a moment, we may write the total scheme in the form

$$\boldsymbol{y}''(t) + \boldsymbol{A}\boldsymbol{y}(t) = \boldsymbol{f}(t), \tag{3.10}$$

[1] The residual may be defined slightly different with Δ replaced by an approximate Δ_h as is the case for the finite difference method.

where $\boldsymbol{A}$ is the discrete (negative) Laplacian defined by

$$\boldsymbol{A}_{ij} = \frac{1}{h^2}(2\delta_{ij} - \delta_{i,j-1} - \delta_{i,j+1}), \qquad i,j = 1,\ldots,N. \tag{3.11}$$

This is equivalent to requiring the residual—with Δ replaced by its finite difference approximation—to vanish at the collocation points x_i. The ODE-system (3.10) can now be solved via time integration provided a sampling of the initial data (3.1c). We postpone the treatment of non-homogeneous boundary conditions to Section 3.2.3. Time integration will be dealt with in Section 3.4.

Upwinding

If one wishes to take advantage of the characteristic form (3.5) it can be done by using one-sided finite differences. The derivative $\frac{\partial}{\partial x}$ for the advection equation with the left-bound component p is wisely approximated by $\frac{\partial}{\partial x}p(t,x_i) \approx \frac{1}{h}(p_i(t) - p_{i-1}(t))$ as all information is coming from the right side of x_i. Likewise for q we may employ a left-sided finite difference in the approximation of $\frac{\partial}{\partial x}q(t,x_i)$ which reads $\frac{1}{h}(q_{i+1}(t) - q_i(t))$. In this way we could obtain a scheme in a matrix form similar to (3.10). We will return to upwinding in Section 3.3.

3.2.2 The finite element method (FEM)

Let us again assume that $y(t,0) = y(t,1) = 0$. Consider the following (nodal) representation by the set of compact basis functions $\{\psi_i^{\text{L}}\}_{i\leq N}$ with the property $\psi_i^{\text{L}}(x_j) = \delta_{ij}$

$$y_h(t,x) = \sum_{i=1}^{N} y_h(t,x_i)\psi_i^{\text{L}}(x), \tag{3.12}$$

where $y_h(t,x_i)$ is the nodal value of y_h in the node x_i. The nodal values can be collected in a column vector $\boldsymbol{y}(t) = [y_h(t,x_1),\ldots,y_h(t,x_N)]^{\mathsf{T}}$.

We insert y_h in the variational form (3.2) and obtain the Galerkin formulation: find $y_h \in \mathcal{V}_h$ such that

$$\int_\Omega y_h'' v \mathrm{d}x + \int_\Omega \frac{\partial y_h}{\partial x}\frac{\partial v}{\partial x}\mathrm{d}x = \int_\Omega f v \mathrm{d}x, \quad \forall v \in \mathcal{V}_h, \tag{3.13}$$

where we have replaced the infinite test space $\mathcal{V}$ by the N dimensional subspace $\mathcal{V}_h$ spanned by the basis functions

$$\mathcal{V}_h := \operatorname{span}\{\psi_1^{\text{L}},\ldots,\psi_N^{\text{L}}\}.$$

The Galerkin formulation (3.13) is actually equivalent to requiring the residual (3.8) to be orthogonal to all basis functions

$$\int_\Omega \mathcal{R}_h(t,x)\psi_n^{\text{L}}(x)\mathrm{d}x = 0, \qquad n = 1,\ldots,N.$$

We insert (3.12) in the Galerkin form (3.13) to obtain N equations constituting a FEM for the wave equation

$$\boldsymbol{M}\boldsymbol{y}''(t) + \boldsymbol{K}\boldsymbol{y}(t) = \boldsymbol{M}\boldsymbol{f}(t), \tag{3.14}$$

where the mass and stiffness matrices are defined by

$$\boldsymbol{M}_{ij} = \int_\Omega \psi_i^{\text{L}}(x)\psi_j^{\text{L}}(x)\mathrm{d}x, \tag{3.15}$$

$$\boldsymbol{K}_{ij} = \int_\Omega \frac{\partial \psi_i^{\text{L}}}{\partial x}(x)\frac{\partial \psi_j^{\text{L}}}{\partial x}(x)\mathrm{d}x. \tag{3.16}$$

Notice how the stiffness matrix is symmetric and positive definite like its bilinear ancestor $a(\cdot,\cdot)$ of (3.3). The matrix $\boldsymbol{M}^{-1}\boldsymbol{K}$ is an approximation to the negative Laplacian in the basis $\{\psi_n\}_{n\leq N}$.

Consider a basis consisting of piecewise linear polynomials, so-called hat functions,

$$\psi_i^{\text{L}}(x) = \begin{cases} (x - x_{i-1})/(x_i - x_{i-1}) & \text{for } x_{i-1} \leq x \leq x_i, \\ (x_{1+i} - x)/(x_{1+i} - x_i) & \text{for } x_i \leq x \leq x_{i+1}, \\ 0 & \text{otherwise,} \end{cases} \tag{3.17}$$

for $i = 1, \ldots, N$. On an equidistant grid with spacing h, the mass and stiffness matrices (3.15) and (3.16) become

$$\boldsymbol{M}_{ij} = h(\tfrac{2}{3}\delta_{ij} + \tfrac{1}{6}(\delta_{i,j-1} + \delta_{i,j+1})), \qquad i,j = 1,\ldots,N \tag{3.18}$$

$$\boldsymbol{K}_{ij} = \frac{1}{h}(2\delta_{ij} - (\delta_{i,j-1} + \delta_{i,j+1})), \qquad i,j = 1,\ldots,N. \tag{3.19}$$

The stiffness matrix' relation to the finite difference approximation (3.11) of the discrete Laplacian $\boldsymbol{A}$ is simply $\boldsymbol{K} = h\boldsymbol{A}$ for equidistant grids. The only difference between the 2nd order central finite difference method (3.10) and the present linear FEM is thereby the appearance of the mass matrix. Schemes with a non-diagonal mass matrix are sometimes called *implicit* semi-discretizations as they lead to systems of implicit algebraic equations when handled with an explicit time marching approach.

Notice how the FDM approximates the PDE, where FEM instead seeks approximations to its solution. Notice also that FEM have no "direction" which means that upwind-type solutions, like the one sketched in the preceding section, are not possible with FEM.

We still need to specify how to deal with boundary and initial conditions to complete the treatment of the FEM semi-discretization (3.14). We will return to this matter shortly after a brief description about *mixed* finite element methods.

A mixed FEM

A set of basis functions ψ_n^{L} was used to approximate y in the above FEM formulation. The same set of basis functions were, implicitly, used to expand the "velocity" y_h'. It is, however, possible to pick different bases for y_h and y_h'. The formulation is then usually called a *mixed* FEM. This can sometimes be advantageous and even the most natural choice for some PDEs such as the Stokes equation for viscous fluid flow [BS02].

A relevant method with the standard linear splines as basis for y_h and a piecewise constant basis for y_h' suitable for control of the wave equation was described in [CM06, page 419–420]. The resulting scheme is part of a unified formulation which we will present below. The same mixed FEM, which also is known as the "box method", was also used by J.M. Rasmussen in [Ras04].

On an equidistant grid with spacing h the resulting mass matrix becomes [CM06]

$$\boldsymbol{M}_{ij} = h(\tfrac{1}{2}\delta_{ij} + \tfrac{1}{4}(\delta_{i,j-1} + \delta_{i,j+1})), \qquad i,j = 1,\ldots,N.$$

The stiffness matrix is identical to (3.19).

3.2.3 A unified formulation

It can easily be verified that the above FDM, linear FEM, and mixed FEM on equidistant grids are all contained in the semi-discretization (3.14) with the following α-family of mass and stiffness matrices [VB82, page 33]

$$\boldsymbol{M}_{ij} = h((1-2\alpha)\delta_{ij} + \alpha(\delta_{i,j-1} + \delta_{i,j+1})), \qquad i,j = 1,\ldots,N, \tag{3.20}$$

$$\boldsymbol{K}_{ij} = \frac{1}{h}(2\delta_{ij} - (\delta_{i,j-1} + \delta_{i,j+1})), \qquad i,j = 1,\ldots,N, \tag{3.21}$$

where α is a parameter $0 \leq \alpha \leq 1/2$. We have the following special cases:

$\alpha = 0$	2nd order central FDM
$\alpha = 1/12$	Higher order (Störmer-Numerov)
$\alpha = 1/6$	FEM with linear splines
$\alpha = 1/4$	Mixed FEM with piecewise constant basis for y_h',

where the Störmer-Numerov choice is the only leading to truncation errors of order $\mathcal{O}(h^4)$ all other choices of α lead to an $\mathcal{O}(h^2)$ accuracy [VB82].

A unified description like the above, which also was used by J.M. Rasmussen in [Ras04], allows a unified analysis and implementation. It will serve as the main source of comparison when working with the discontinuous Galerkin method which we will see to shortly.

The final unified FEM-FDM scheme reads

$$\boldsymbol{M}\boldsymbol{y}''(t) + \boldsymbol{K}\boldsymbol{y}(t) = \boldsymbol{M}\boldsymbol{f}(t) + \boldsymbol{g}(t) \tag{3.22}$$

where $\boldsymbol{M}$ and $\boldsymbol{K}$ are defined by (3.20) and (3.21) and the boundary contribution vector $\boldsymbol{g}$ is

$$\boldsymbol{g}(t) = [\tfrac{1}{h}g_0(t), 0, \ldots, 0, \tfrac{1}{h}g_1(t)]^{\mathsf{T}}. \tag{3.23}$$

when the boundary conditions (3.1b) are non-homogeneous. Note that non-zero Dirichlet conditions are actually included in the definition of the trial space and the representation of y_h for the FEM case.

In standard ODE first order form, the scheme becomes

$$\begin{bmatrix} \boldsymbol{M} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M} \end{bmatrix} \frac{d}{dt} \begin{bmatrix} \boldsymbol{y}(t) \\ \boldsymbol{y}'(t) \end{bmatrix}' + \begin{bmatrix} \boldsymbol{0} & -\boldsymbol{M} \\ \boldsymbol{K} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}(t) \\ \boldsymbol{y}'(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{M}\boldsymbol{f}(t) + \boldsymbol{g}(t) \end{bmatrix}. \tag{3.24}$$

In both cases, (3.22) and (3.24), the initial data (3.1c) may simply be sampled

$$\boldsymbol{y}(0) = \boldsymbol{y}^0 := [y^0(x_1), \ldots, y^0(x_N)]^{\mathsf{T}},$$
$$\boldsymbol{y}'(0) = \boldsymbol{y}^1 := [y^1(x_1), \ldots, y^1(x_N)]^{\mathsf{T}},$$

to obtain discrete initial conditions. For initial data which is reasonable band-limited in the frequency spectrum this is good choice. For less band-limited

data, however, a simple sampling will introduce an aliasing effect leaving higher frequency components indistinguishable from their low frequency alias [VB82]. If necessary these high frequency parts can be eliminated via a Fourier transform of higher resolution. The same goes for the forcing term f.

3.2.4 Other classical methods

The finite difference and finite element methods are the most widespread among semi-discretization methods for PDEs but two other classical methods deserve mentioning here as they are important to the discontinuous Galerkin FEM. We are thinking about the finite volume method and the spectral method.

The finite volume method, which is particularly important in the field of computational fluid dynamics, is based on the subdivision of the spatial domain into cells. The residual is required to vanish on each cell which leads to a local scheme. Volume integrals over the cell volumes are transformed to surface integrals by Gauss' theorem which results in the need for evaluating fluxes at the cell interfaces. Different fluxes lead to different finite volume methods—design of such fluxes is a science in itself.

Spectral and pseudo-spectral methods are high order methods which in bounded, non-periodic cases depend on special non-equidistant grids such as Chebyshev and Legendre grids (often given a Lobatto or Lobatto-Gauss prefix). The use of these clustered grids avoid the Runge phenomenon known from high order polynomial interpolation on equidistant grids. In this way spectral differentiation with very high accuracy is attained which can be used for highly accurate semi-discretization of PDEs. See [Tre00] and [Boy01] for more on spectral methods.

3.3 Discontinuous Galerkin FEM

Each of the above classical methods for semi-discretizations have their pros and cons. A more recent method called discontinuous Galerkin FEM (abbreviated DG-FEM) combines elements—advantages say its advocates—from several of these classical methods to obtain a highly flexible method of possible high order.

The solution is represented by a sum of K local solutions—one on each of the K conforming, non-overlapping elements D^k, $k = 1, 2, \ldots, K$. The global solution is not required to be continuous across the interfaces, which is the property that justifies "discontinuous" in the name of the method. It should be noted that due to extensive use of superscripts on y in this function, we use $\frac{\partial y}{\partial t}$ instead of the y' like in Section 3.1.1.

DG-FEM, which is closer related to finite volumes than the other classical semi-discretizations, is a method developed for conservation laws. We consider a scalar law in the form

$$\frac{\partial y}{\partial t} + \frac{\partial \varrho(y)}{\partial x} = f. \tag{3.25}$$

where ϱ is a flux function. The most simple form is the unforced ($f = 0$) advection equation with constant speed for which the flux function ϱ is

$$\varrho(y) = ay, \qquad a \text{ constant.}$$

Note that equation (3.25), however simple, is the prototype for the characteristic form of the wave equation (3.5). We present a DG-scheme for the advection equation below.

The exposition here relies heavily on the work of J. Hesthaven and T. Warburton [HW08]. This holds true for the implementation of DG-FEM as well.

3.3.1 A DG-scheme for the advection equation

The local solution y^k is approximated by a polynomial of degree $N_p - 1$ on each element $\mathsf{D}^k = (x_{\mathrm{L}}^k, x_{\mathrm{R}}^k)$. We expand y^k in two ways

$$y_h^k(t,x) = \sum_{n=1}^{N_p} \widehat{y}_n^k(t)\psi_n(x) = \sum_{i=1}^{N_p} y_h^k(t,x_i^k)\ell_i^k(x). \tag{3.26}$$

The first representation is a *modal* one, where $\widehat{y}_n^k(t)$ is the expansion coefficients of N_p local polynomial basis functions $\psi_n, n = 1,\ldots,N_p$. The second is a *nodal* representation in which ℓ_i is the i'th interpolating Lagrange polynomial with $\ell_i(x_j^k) = \delta_{ij}$ where $x_i^k \in \mathsf{D}^k, i = 1,\ldots,N_p$. We have, in either case, a total of $N_p \cdot K$ unknowns.

The residual for the advection equation (3.25) with $f = 0$ and $h(y) = ay$ is

$$\mathcal{R}_h(t,x) = \frac{\partial y_h}{\partial t} + \frac{\partial (a y_h)}{\partial x}.$$

A way to let the approximation fulfill the PDE is to require the residual to "vanish" is some way on each element.

Let ζ_n represent either the modal basis function ψ_n or the nodal ℓ_n^k in the following. The derivation of the DG-scheme holds in both cases resulting respectively in the modal and nodal formulation. We require that the residual is orthogonal to all basis functions on each element D^k, that is,

$$\int_{\mathsf{D}^k} \mathcal{R}_h(t,x)\zeta_n(x)\mathrm{d}x = 0, \quad 1 \le n \le N_p, \tag{3.27}$$

which is the classic Galerkin requirement—just element-wise. Applying integration by parts in the spatial direction yields

$$\int_{\mathsf{D}^k} \Big(\frac{\partial y_h^k}{\partial t}\zeta_n - (a y_h^k)\frac{\partial \zeta_n}{\partial x}\Big)\mathrm{d}x = -\big[(a y_h^k)\zeta_n\big]_{x_{\mathrm{L}}^k}^{x_{\mathrm{R}}^k}, \quad 1 \le n \le N_p,$$

which is N_p pure local equations. No connection to other elements or boundary conditions exist; the equations are consequently not suitable for obtaining global solutions.

This connection may be established, however, by relaxing requirement (3.27) slightly by replacing $(a y_h^k)$ on the right side by a linear combination of values at the local interface points x_{L}^k and x_{R}^k and the interface points of the neighboring elements; that is, for the left end, x_{L}^k, replace $(a y_h^k(x_{\mathrm{L}}^k))$ by a combination of $(a y_h^k(x_{\mathrm{L}}^k))$ and $(a y_h^{k-1}(x_{\mathrm{R}}^{k-1}))$ and similar for the right end.

We therefore replace $(a y_h^k)$ in the equations by $(a y_h)^*$ and denote this new quantity, which is legacy from the finite volume method, the *numerical flux*. The right choice of this flux depends heavily on the dynamics of the underlying PDE. We shall return to this issue shortly. The numerical flux also determines

how to assign the Dirichlet boundary conditions as they will play the role of neighboring interface point at $x = 0$ for D^1 and at $x = 1$ for D^K. Inserting the numerical flux results in the weak DG-scheme for D^k

$$\int_{\mathsf{D}^k} \big(\frac{\partial y_h^k}{\partial t}\zeta_n - (ay_h^k)\frac{\partial \zeta_n}{\partial x}\big)\mathrm{d}x = -\big[(ay_h)^*\zeta_n\big]_{x_{\mathrm{L}}^k}^{x_{\mathrm{R}}^k}, \quad 1 \le n \le N_p. \tag{3.28}$$

The alternative *strong* formulation is obtained by performing integration by parts once more

$$\int_{\mathsf{D}^k} \big(\frac{\partial y_h^k}{\partial t}\zeta_n + \frac{\partial (ay_h^k)}{\partial x}\zeta_n\big)\mathrm{d}x = \big[\big((ay_h^k) - (ay_h)^*\big)\zeta_n\big]_{x_{\mathrm{L}}^k}^{x_{\mathrm{R}}^k}, \quad 1 \le n \le N_p. \tag{3.29}$$

Either case gives us a total of $N_p \cdot K$ equations for the same number of unknowns. The weak and strong formulations are mathematically equivalent but may behave different numerically.

Notice how, in both the weak and strong case, the right hand term is responsible for the flow of information between elements and for boundary conditions as well. This emphasizes the important role of the numerical flux $(ay_h)^*$.

Inserting either the nodal or modal representation of y_h in either the strong or weak formulation will result in a DG-scheme, which can be used for computation. We will from now on concentrate on the nodal form (replace ζ_n with ℓ_n) of the strong scheme which for element k leads us to a system of ODEs

$$\boldsymbol{M}^k \frac{d}{dt}\boldsymbol{y}^k + \boldsymbol{S}^k(a\boldsymbol{y}^k) = \big[\big((ay_h^k) - (ay_h)^*\big)\boldsymbol{\ell}^k(x)\big]_{x_{\mathrm{L}}^k}^{x_{\mathrm{R}}^k}, \tag{3.30}$$

where $\boldsymbol{y}^k$ and $\boldsymbol{\ell}^k(x)$ are respectively the vector of the local nodal solution and the vector of Lagrange polynomials

$$\boldsymbol{y}^k = [y_h^k(t, x_1^k), \ldots, y_h^k(t, x_{N_p}^k)]^{\mathsf{T}}, \qquad \boldsymbol{\ell}^k(x) = [\ell_1^k(x), \ldots, \ell_{N_p}^k(x)]^{\mathsf{T}},$$

and the local mass $\boldsymbol{M}^k$ and stiffness matrices $\boldsymbol{S}^k$ are determined by

$$\boldsymbol{M}_{ij}^k = \langle \ell_i^k, \ell_j^k \rangle_{\mathsf{D}^k} \qquad \boldsymbol{S}_{ij}^k = \Big\langle \ell_i^k, \frac{d\ell_j^k}{dx} \Big\rangle_{\mathsf{D}^k}, \tag{3.31}$$

where $\langle \cdot, \cdot \rangle_{\mathsf{D}^k}$ is the local L^2-inner product on $L^2(\mathsf{D}^k)$.

Notice the local nature of the method: all operators are local and the only exchange of information between elements take place across interfaces via the numerical flux.

Numerical flux

The design of a good numerical flux is a big and important topic—and strongly problem dependent, too. We shall mention only the most basic examples here, but they are, however, adequate for our simple 1-d linear wave equation.

Let y^- denote the local interface point belonging to element D^k and y^+ the interface point of the neighboring element. We define the average and jump across an interface by

$$\{\{y\}\} = \frac{y^- + y^+}{2}, \qquad\qquad [\![y]\!] = \boldsymbol{n}^- y^- + \boldsymbol{n}^+ y^+,$$

where $\boldsymbol{n}^-$ is the outward normal on element D^k and $\boldsymbol{n}^+$ is the outward normal of the neighbor. In the 1-d case the normals reduce to either $+1$ or -1.

If we wish to take information from one side only, which seems like a sensible choice for the advection equation, we can use an *upwind flux*

$$(ay)^* = \{\{ay\}\} + \tfrac{1}{2}\,|a|\,[\![y]\!]. \tag{3.32}$$

If we know that information flows in *both* directions a *central flux*, which averages interior and exterior information, might be a good choice

$$(ay)^* = \{\{ay\}\}. \tag{3.33}$$

We can express both fluxes and any combination of them via the family

$$(ay)^* = \{\{ay\}\} + |a|\,\frac{1-\alpha}{2}[\![y]\!], \quad 0 \le \alpha \le 1, \tag{3.34}$$

which reduces to a pure central flux for $\alpha = 1$ and pure upwinding when $\alpha = 0$.

The numerical flux is fundamental not only for flow of information between elements but also for assigning boundary conditions. When $a > 0$ the advection equation need a boundary condition at the left end of the domain. A Dirichlet condition $y(t, x_{\mathrm{L}}^1) = g^0(t)$ at this point may be considered, when viewed upon from the first element D^1, as the value y^+ and in this way assigned through the numerical flux.

Stability

Stability of (3.30), which is central for the convergence, can be shown by use of an energy method as explained in [HW08, page 25]. Equation (3.30) is multiplied by $\boldsymbol{y}^{\mathsf{T}}$ which gives an expression for the change of local energy over time. By summing over all elements, we require the change of energy to be less than or equal to zero in case of homogeneous Dirichlet conditions. The authors in [HW08] show that (3.30) is stable with the α-flux (3.34) for $0 \le \alpha \le 1$.

3.3.2 The LGL grid and DG-basis functions

We claimed in the beginning of Section 3.3 that the DG method was highly flexible and of possible high order. Furthermore, we have not put any restrictions on the local polynomial order $N_p - 1$ of y_h^k in (3.26). We do know, however, that polynomial interpolation on *equidistant* grids breaks down due to Runge phenomena for even moderate order. Here DG-FEM borrows from the theory of spectral methods. What we need locally is a clustered grid and our preferred choice is a Legendre, a Legendre-Gauss-Lobatto (LGL), grid, since it translates to standard L^2-norms directly instead of weighted L^2-norms as in the Chebyshev case.

We wish to set up a reference element $r \in \mathsf{I} = [-1, 1]$ and introduce for this reason the affine mapping

$$\begin{aligned} x \in \mathsf{D}^k = [x_{\mathrm{L}}^k, x_{\mathrm{R}}^k] : &\qquad x(r) = x_{\mathrm{L}}^k + \frac{h^k}{2}(1+r), &\qquad -1 \le r \le 1, \\ r \in \mathsf{I} = [-1, 1] : &\qquad r(x) = -1 + \frac{2}{h^k}(x - x_{\mathrm{L}}^k), &\qquad x_{\mathrm{L}}^k \le x \le x_{\mathrm{R}}^k, \end{aligned}$$

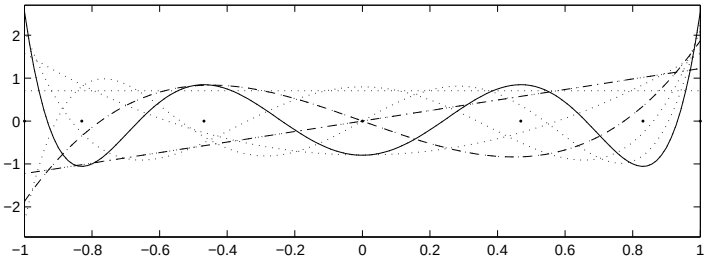


Figure 3.2: The first seven normalized Legendre polynomials $\widetilde{P}_0, \ldots, \widetilde{P}_6$ as function of $r \in \mathsf{I}$. The linear $\widetilde{P}_1$ is marked with a dashed-dotted line, $\widetilde{P}_3$ with dashed line, $\widetilde{P}_6$ with solid line and the remaining with dotted lines.

where $h^k = (x_{\mathrm{R}}^k - x_{\mathrm{L}}^k)$ is the size of element D^k. We introduce an orthonormal basis on I of normalized Legendre polynomials (Figure 3.2 shows the first 7)

$$\psi_n(r) = \widetilde{P}_{n-1}(r) = \frac{P_{n-1}(r)}{\sqrt{\gamma_{n-1}}}, \qquad \gamma_n = \frac{2}{2n+1}, \tag{3.35}$$

where $P_n(r)$ is the Legendre polynomial of order n and γ_n is a normalization factor. This basis is the optimal polynomial basis on I as it reduces the local modal mass matrix $\boldsymbol{M}_{ij} = \langle \psi_i, \psi_j \rangle_{\mathsf{I}}$ to the identity, which means that we can recover the i'th component of $\widehat{y}$ by the L^2-projection $\widehat{y}_i = \langle y_h, \psi_i \rangle_{\mathsf{I}}$. The inner product $\langle \cdot, \cdot \rangle_{\mathsf{I}}$ is the standard inner product over $L^2(\mathsf{I})$. A Gaussian quadrature rule can be used to obtain nodes and weights for an $(2N_p - 1)$'th order accurate quadrature. We will, however, consider it as an interpolation problem

$$y(r_i) = \sum_{n=1}^{N_p} \widehat{y}_n \widetilde{P}_{n-1}(r_i), \qquad i = 1, \ldots, N_p,$$

where r_i are N_p distinct grid points. We define the generalized Vandermonde matrix $\boldsymbol{V}$ by

$$\boldsymbol{V}_{ij} = \widetilde{P}_{j-1}(r_i), \qquad i, j = 1, \ldots, N_p, \tag{3.36}$$

for any set of distinct $r_i \in \mathsf{I}$. This allows us to write

$$\boldsymbol{y} = \boldsymbol{V}\widehat{\boldsymbol{y}},$$

where $\boldsymbol{y} = [y(r_1), \ldots, y(r_{N_p})]^{\mathsf{T}}$ and $\widehat{\boldsymbol{y}} = [\widehat{y}_1, \ldots, \widehat{y}_{N_p}]^{\mathsf{T}}$ by the uniqueness of polynomial interpolation.

By using Lagrange basis polynomials with the property $\ell_i(r_j) = \delta_{ij}$ (shown for $N_p = 7$ on Figure 3.3), we can express the interpolation

$$y(r) \approx y_h(r) = \sum_{n=1}^{N_p} \widehat{y}_n \widetilde{P}_{n-1}(r) = \sum_{i=1}^{N_p} y(r_i) \ell_i(r). \tag{3.37}$$

This gives us the relation

$$\boldsymbol{V}^{\mathsf{T}} \boldsymbol{\ell}(r) = \widetilde{\boldsymbol{P}}(r), \tag{3.38}$$

where $\boldsymbol{\ell}(r) = [\ell_1(r), \ldots, \ell_{N_p}(r)]^{\mathsf{T}}$ and $\widetilde{\boldsymbol{P}}(r) = [\widetilde{P}_0(r), \ldots, \widetilde{P}_{N_p-1}(r)]^{\mathsf{T}}$. Hesthaven and Warburton [HW08] show that the best approximating polynomial of order

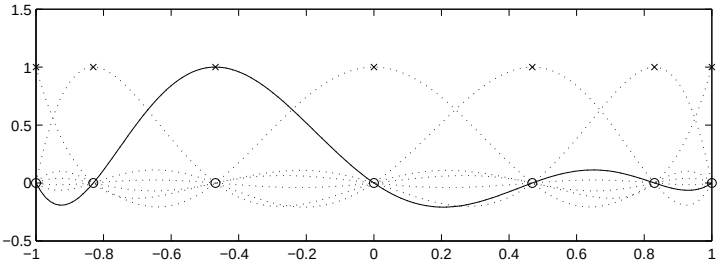


Figure 3.3: Interpolating Lagrange polynomials ℓ_i for which $\ell_i(r_j) = \delta_{ij}$ and $N_p = 7$ on the LGL grid. The grid points r_j are marked by (o) and ℓ_3 is marked with a solid line.

$N_p - 1$ is obtained when the Lebesgue constant $\max \sum_{i=1}^{N_p} |\ell_i(r)|$ is minimized. And furthermore, that this occurs when the grid points are defined as the zeros of

$$f(r) = (1 - r^2)\widetilde{P}'_{N_p-1}(r),$$

known as the Legendre-Gauss-Lobatto (LGL) quadrature points.

Mass and stiffness matrices

The local (nodal) mass and stiffness matrices (3.31) can be related to their equivalents on the reference element I by

$$\boldsymbol{M}_{ij} = \langle \ell_i, \ell_j \rangle_{\mathsf{I}} \qquad \boldsymbol{S}_{ij} = \Big\langle \ell_i, \frac{d\ell_j}{dx} \Big\rangle_{\mathsf{I}},$$

where $\langle \cdot, \cdot \rangle_{\mathsf{I}}$ is the L^2 inner product on I. From (3.38) we know that ℓ_i can be expressed by the Vandermonde matrix $\boldsymbol{V}$ and the Legendre polynomials $\widetilde{P}_n$ and due to the orthogonality of the latter we have

$$\boldsymbol{M} = (\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}})^{-1}. \tag{3.39}$$

Differentiation on the reference element can be done with the matrix $\boldsymbol{D}_r$ with the (ij)'th entry $d\ell_j/dr|_{r_i}$. It may be expressed by

$$\boldsymbol{D}_r = \boldsymbol{V}_r \boldsymbol{V}^{\mathsf{T}}, \tag{3.40}$$

where $\boldsymbol{V}_r$ is defined by $[\boldsymbol{V}_r]_{ij} = d\widetilde{P}_j/dr|_{r_i}$. Finally, the stiffness matrix may be found by the identity

$$\boldsymbol{S} = \boldsymbol{M}\boldsymbol{D}_r. \tag{3.41}$$

On element D^k the matrices $\boldsymbol{M}$ and $\boldsymbol{S}$ become

$$\boldsymbol{M}^k = \frac{h^k}{2}\boldsymbol{M}, \qquad \boldsymbol{S}^k = \boldsymbol{S},$$

where h^k is the width of element D^k.

The surface integral

The right hand side of (3.30) is a surface integral which on the reference element I becomes

$$\big[(ay_h - (ay)^*)\boldsymbol{\ell}(r)\big]_{-1}^{1} = (ay_h - (ay)^*)\big|_{r_{N_p}} \boldsymbol{e}_{N_p} - (ay_h - (ay)^*)\big|_{r_1} \boldsymbol{e}_1$$

where $\boldsymbol{e}_i$ is the i'th coordinate vector in $\mathbb{R}^{N_p}$.

ODE system

With the gained knowledge we can now write the local DG-scheme for the advection equation in standard ODE-form $\frac{d}{dt}\boldsymbol{y} = \mathcal{L}_h(\boldsymbol{y}, t)$ as

$$\frac{d}{dt}\boldsymbol{y}^k = -a(\boldsymbol{M}^k)^{-1}\boldsymbol{S}\boldsymbol{y}^k + (\boldsymbol{M}^k)^{-1}\left[\boldsymbol{\ell}^k(x)\left(ay_h^k - (ay_h)^*\right)\right]_{x_{\mathrm{L}}^k}^{x_{\mathrm{R}}^k}, \tag{3.42}$$

where the mass and stiffness matrices are given via the Vandermonde matrix (3.36) and the differentiation matrix (3.40) as (3.39) and (3.41). The surface term can be determined as explained above and the flux is given by (3.34).

Consistency and convergence

Lax equivalence theorem [LR56] says that consistency and stability implies convergence of a numerical scheme. Consistency is about approximations of functions and operators—are they consistent with their continuous ancestors? Our approximation is consistent if the $t = 0$ error $y(0, x) - y_h(0, x)$ and the truncation error tends to zero for an increasing number of variables $N_p \cdot K$. Hesthaven and Warburton prove [HW08, page 77] that this holds for the just presented DG-approximation with Legendre polynomials. Hence, by recalling the stability result on page 34, we can conclude that the DG-scheme for the advection equation is convergent.

3.3.3 A DG-scheme for the wave equation

We have just seen a DG-formulation for the prototypical advection equation. It is tempting to use the characteristic form (3.5) which consists of exactly two advection equations when seeking a DG method for the wave equation. Alternatively, we may take our starting point in (3.1a) which we can formulate as a system

$$\frac{\partial y}{\partial t} = z, \qquad g = \frac{\partial y}{\partial x}, \qquad \frac{\partial z}{\partial t} = \frac{\partial g}{\partial x}.$$

It may be approximated by the local scheme

$$\frac{d}{dt}\boldsymbol{y}^k = \boldsymbol{z}^k \tag{3.43a}$$

$$\boldsymbol{g}^k = \frac{h^k}{2}\boldsymbol{D}_r\boldsymbol{y}^k \tag{3.43b}$$

$$\frac{d}{dt}\boldsymbol{z}^k = -(\boldsymbol{M}^k)^{-1}\boldsymbol{S}\boldsymbol{g}^k + (\boldsymbol{M}^k)^{-1}\left[\boldsymbol{\ell}^k(x)\left(g_h^k - (g_h)^*\right)\right]_{x_{\mathrm{L}}^k}^{x_{\mathrm{R}}^k}, \tag{3.43c}$$

where $\boldsymbol{g}^k = \frac{h^k}{2}\boldsymbol{D}_r\boldsymbol{y}^k$ is the gradient of $\boldsymbol{y}^k$ in x-direction formed by the differentiation matrix $\boldsymbol{D}_r$, defined in (3.40), and the Jacobian factor $\frac{h^k}{2}$. This gradient serves as the flux function g_h in the third line. The vectors $\boldsymbol{y}^k, \boldsymbol{g}^k$ and $\boldsymbol{z}^k$ contains the nodal values at element D^k for y_h, g_h, and y_h approximating y, g and z. Figure 3.4(left) shows the structure of the resulting right hand side $\mathcal{L}_h$ of the ODE-system $\frac{d}{dt}\boldsymbol{Y}(t) = \mathcal{L}_h\boldsymbol{Y}(t)$ where $\boldsymbol{Y}(t) = [\boldsymbol{y}, \boldsymbol{z}]^{\mathsf{T}}$.

A central flux (3.33) is a simple first choice for the numerical flux $(g_h)^*$. An upwind flux, on the contrary, can not be used as information passes in both directions.

Upwinding is, however, very well-suited for wave problems which suggest us to construct our DG scheme for the characteristic variables p and q (3.5) instead. It consists of an advection equation with $a = -1$ in p and one with $a = 1$ in q leading to the local system

$$\frac{d}{dt}\boldsymbol{p}^k = (\boldsymbol{M}^k)^{-1}\boldsymbol{S}\boldsymbol{p}^k - (\boldsymbol{M}^k)^{-1}\big[\boldsymbol{\ell}^k(x)\big(p_h^k - (p_h)^*\big)\big]_{x_{\mathrm{L}}^k}^{x_{\mathrm{R}}^k}, \tag{3.44a}$$

$$\frac{d}{dt}\boldsymbol{q}^k = -(\boldsymbol{M}^k)^{-1}\boldsymbol{S}\boldsymbol{q}^k + (\boldsymbol{M}^k)^{-1}\big[\boldsymbol{\ell}^k(x)\big(q_h^k - (q_h)^*\big)\big]_{x_{\mathrm{L}}^k}^{x_{\mathrm{R}}^k}, \tag{3.44b}$$

for which we use upwinding (3.32) as numerical flux. The coupling between (3.44a) and (3.44b) occurs on the boundary as defined in (3.7). With homogeneous Dirichlet condition at $x = 0$ and inhomogeneous $y(t, 1) = \kappa(t)$ at $x = 1$, we get

$$\begin{aligned} p_h(t, x_{\mathrm{R}}^K) &= q_h(t, x_{\mathrm{R}}^K) - \sqrt{2}\kappa(t), \\ q_h(t, x_{\mathrm{L}}^1) &= p_h(t, x_{\mathrm{L}}^1), \end{aligned}$$

where $x_{\mathrm{R}}^K = 1$ and $x_{\mathrm{L}}^1 = 0$. These expressions are used respectively in the flux p_h^* as p_{R}^+ on element D^K and in the flux q_h^* as q_{L}^+ on element D^1. Figure 3.4(right) shows the right hand side $\mathcal{L}_h$ of the resulting ODE-system $\frac{d}{dt}\boldsymbol{Y}(t) = \mathcal{L}_h\boldsymbol{Y}(t)$ where $\boldsymbol{Y}(t) = [\boldsymbol{p}, \boldsymbol{q}]^{\mathsf{T}}$.

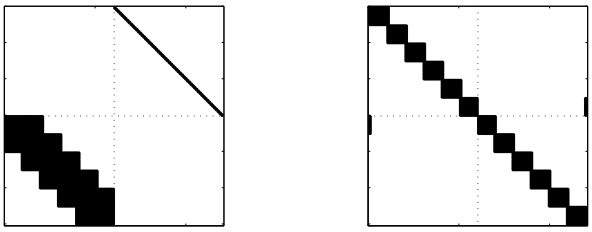


Figure 3.4: The structure (non-zero elements) of the right hand side $\mathcal{L}_h$ of $\frac{d}{dt}Y = \mathcal{L}_h Y$ for (3.43) (left plot) and (3.44) (right plot). Here shown with $K = 6$ elements and $N_p = 10$ nodes per element.

The system has, according to (3.7), the local initial data

$$\begin{aligned} \boldsymbol{p}^k(0) &= \tfrac{\sqrt{2}}{2}\big(-(\boldsymbol{y}^0)^k + (\boldsymbol{z}^0)^k\big), \\ \boldsymbol{q}^k(0) &= \tfrac{\sqrt{2}}{2}\big((\boldsymbol{y}^0)^k + (\boldsymbol{z}^0)^k\big), \end{aligned}$$

where $(\boldsymbol{y}^0)^k$ is a vector with the nodal values $[y^0(x_1^k), \ldots, y^0(x_{N_p}^k)]^{\mathsf{T}}$ of the initial condition y^0 on element D^k. The vector $(\boldsymbol{z}^0)^k$ consists of the nodal values $[z^0(x_1^k), \ldots, z^0(x_{N_p}^k)]^{\mathsf{T}}$ of the k'th element of the anti-derivative $z^0(x) = -\int_0^x y^1(s)\mathrm{d}s + z^0(0)$ (3.6). If y^1 is only known by its nodal values on the LGL grid, we will need a routine for determining its anti-derivative; we present this routine next.

The anti-derivative

The anti-derivative F of a function f defined by its nodal values $\boldsymbol{f} = \oplus_{k=1}^{K} \boldsymbol{f}^k$ on an LGL grid is

$$F(x) = \bigoplus_{k=1}^{K} F^k(x), \quad F^k(x) = \int_{x_{\mathrm{L}}^k}^{x} f^k(s)\mathrm{d}s + F^k(x_{\mathrm{L}}^k), \quad x_{\mathrm{L}}^k \le x \le x_{\mathrm{R}}^k,$$

where $\oplus$ is the direct sum. The local part f^k can be expressed in its modal form on the reference element I by

$$f^k(r) = \sum_{n=1}^{N_p} \widehat{f}_n^k \widetilde{P}_{n-1}(r).$$

We may determine the k'th anti-derivative on I by (omitting the integration constant)

$$F^k(r) = \tfrac{h^k}{2} \int_{-1}^{r} f^k(s)\mathrm{d}s = \sum_{n=1}^{N_p} \widehat{f}_n^k \int_{-1}^{r} \widetilde{P}_{n-1}(s)\mathrm{d}s, \qquad -1 \le r \le 1,$$

where the factor $\frac{h^k}{2}$ emerges from the transformation of the integral to the reference element. For the anti-derivative of the standard Legendre polynomial P_n we have the relation [Asm05, page 315]

$$\int_{-1}^{r} P_n(s)\mathrm{d}s = \frac{1}{2n+1}\big(P_{n+1}(r) - P_{n-1}(r)\big),$$

which can be normalized, just like (3.35),

$$\int_{-1}^{r} \widetilde{P}_n(s)\mathrm{d}s = \frac{\sqrt{\gamma_n}}{2}\big(\sqrt{\gamma_{n+1}}\widetilde{P}_{n+1}(r) - \sqrt{\gamma_{n-1}}\widetilde{P}_{n-1}(r)\big), \tag{3.45}$$

where the normalization still is $\gamma_n = \frac{2}{2n+1}$. We define the integration matrix $\boldsymbol{J}^{\mathsf{I}}$ by

$$\boldsymbol{J}_{ij}^{\mathsf{I}} = \int_{-1}^{r_i} \widetilde{P}_{j-1}(s)\mathrm{d}s, \qquad i,j = 1,\ldots,N_p.$$

We summarize our findings for the anti-derivative F of f represented by the nodal vector $\boldsymbol{f}^k = [f^k(x_1^k),\ldots,f^k(x_{N_p}^k)]^{\mathsf{T}}$ on element D^k with the following expression

$$\boldsymbol{F}^k = \tfrac{h^k}{2}\boldsymbol{J}^{\mathsf{I}}\boldsymbol{V}^{-1}\boldsymbol{f}^k + c,$$

where $\boldsymbol{F}^k = [F^k(x_1^k),\ldots,F^k(x_{N_p}^k)]^{\mathsf{T}}$, $\boldsymbol{V}$ the Vandermonde matrix (3.36) and c is an integration constant.

Notice that if f is not well resolved and has information in its highest mode $\widehat{f}_{N_p}^k$, on one or more elements D^k, this information will be partially lost. We can represent polynomials of order up to $N_p - 1$, yet the anti-derivative F will in this case have polynomial information up to order N_p according to (3.45), that is, an $\widehat{F}_{N_p+1}^k$ component on top of the possible N_p modes. If f, on the contrary, is well resolved and all its highest modes $\widehat{f}_{N_p}^k$ are zero, then the integration is exact.

When possible it is, of course, preferable to use the initial condition z^0 directly instead of finding the anti-derivative of y^1.

3.4 Time integration

In our hands are now several different semi-discretizations of the wave equation all of which can be condensed to at least one of the two forms

(1) a second order system $\boldsymbol{M}\boldsymbol{y}''(t) + \boldsymbol{K}\boldsymbol{y}(t) = \boldsymbol{M}\boldsymbol{f}(t)$ of size N like (3.22)

(2) a first order system $\boldsymbol{Y}'(t) = \mathcal{L}_h\boldsymbol{Y}(t)$ of size $2N$ like (3.44)

where $\boldsymbol{Y}$ may represent $[\boldsymbol{y}, \boldsymbol{y}']^\mathsf{T}$ or the characteristic variables $[\boldsymbol{p}, \boldsymbol{q}]^\mathsf{T}$ from (3.44) corresponding to a choice of right hand side $\mathcal{L}_h$. Although all of the above semi-discretizations could easily fit in (2) there are convenient ways to treat second order systems which suggests keeping both ODE formulations side by side.

It remains to integrate (1) or (2) in time to obtain the fully discrete solution. An abundance of methods exists for this task. We shall focus only on a few of them: The Newmark scheme is a classic choice for the treatment of dynamic system like (1). Runge-Kutta methods are build on quadrature rules are most naturally formulated for first order systems (2). Simple finite difference time-stepping, on the other hand, applies easily to both (1) and (2). The explicit mid-point rule and the trapezoidal rule are two simple schemes for time integration; they will appear below for second order system as special cases of the Newmark algorithm. J. Rasmussen analyzed in [Ras04] these two schemes for first and second order systems and derived discrete energy norms corresponding to the continuous system in both cases.

The simple finite difference methods have their primary advantage when it comes to the numerical analysis. Generalized Newmark, higher-order Runge-Kutta and other advanced schemes produce, in return, more accurate results—sometimes even compensating for incorrect dispersion behavior of semi-discretizations (see *e.g.,* [Kre01, Kre06b, Kre08]) or some other unwanted property. We will return to numerical dispersion and related issues in Section 3.5.

ODE stability can be analyzed by considering the eigenvalues of $\mathcal{L}_h$. The eigenvalues should be in the stability region of the time integration scheme, and an important factor for this stability is the Courant number which is the ratio between the time step and spatial step size $\mu = \Delta t/h$. We shall not go in to the details of stability analysis here but refer to [Ise96] for the analysis of stability of finite difference and Runge-Kutta schemes and to [Kre06a] for analysis of the Newmark method.

The Newmark scheme

The Newmark scheme [New59] is formulated for mechanical systems

$$\boldsymbol{M}\boldsymbol{y}''(t) + \boldsymbol{C}\boldsymbol{y}'(t) + \boldsymbol{K}\boldsymbol{y}(t) = \boldsymbol{M}\boldsymbol{f}(t),$$

where the matrices $\boldsymbol{M}, \boldsymbol{C}$ and $\boldsymbol{K}$ are the mass, damping and stiffness matrix, respectively. In the case of the classical wave equation, treated in this dissertation, there is no (structural) damping, *i.e.,* $\boldsymbol{C} = \boldsymbol{0}$. We will, however, keep $\boldsymbol{C}$ in the formulation for a moment while introducing the basic concepts of the method. Note that the right hand side $\boldsymbol{M}\boldsymbol{f}$ may contain contributions from boundary conditions alongside the sampling of the forcing term f.

In the following, we consider M discrete instances of the time t that is $0 = t_0, t_1, \ldots, t_m, \ldots, t_M = T$ with regular spacing Δt.

The idea behind the Newmark scheme is to use different approximations for the displacement $\boldsymbol{U}^m = \boldsymbol{y}(t_m)$, the velocity $\boldsymbol{V}^m = \boldsymbol{y}'(t_m)$ and the acceleration $\boldsymbol{A}^m = \boldsymbol{y}''(t_m)$. Collecting the above for time step $m+1$ gives

$$\boldsymbol{M}\boldsymbol{A}^{m+1} + \boldsymbol{C}\boldsymbol{V}^{m+1} + \boldsymbol{K}\boldsymbol{U}^{m+1} = \boldsymbol{M}\boldsymbol{f}^{m+1}. \tag{3.46}$$

For $\boldsymbol{U}^{m+1}$ and $\boldsymbol{V}^{m+1}$ consider these modified Taylor approximations, where $\boldsymbol{A}^m$ is replaced by weighted average of $\boldsymbol{A}^m$ and $\boldsymbol{A}^{m+1}$

$$\boldsymbol{U}^{m+1} = \boldsymbol{U}^m + \Delta t \boldsymbol{V}^m + \frac{\Delta t^2}{2}[(1-2\beta)\boldsymbol{A}^m + 2\beta \boldsymbol{A}^{m+1}] \tag{3.47}$$

$$\boldsymbol{V}^{m+1} = \boldsymbol{V}^m + \Delta t[(1-\gamma)\boldsymbol{A}^m + \gamma \boldsymbol{A}^{m+1}], \tag{3.48}$$

where $0 \leq \beta \leq \frac{1}{2}$ and $0 \leq \gamma \leq 1$. The unknowns $\boldsymbol{A}^{m+1}$, $\boldsymbol{V}^{m+1}$ and $\boldsymbol{U}^{m+1}$ can be computed from the three above equations (3.46)–(3.48) since $\boldsymbol{f}^{m+1}$ is known.

A Newmark algorithm when $\boldsymbol{C} = \boldsymbol{0}$

In the case of the wave equation ($\boldsymbol{C} = \boldsymbol{0}$) we can do the Newmark time-stepping as follows.

The algorithm is set off by an initialization procedure involving the initial conditions $\boldsymbol{y}(0)$ and $\boldsymbol{y}'(0)$.

$$\begin{aligned}
&\boldsymbol{U}^0 = \boldsymbol{y}(0)\\
&\boldsymbol{V}^0 = \boldsymbol{y}'(0)\\
&\text{solve } \boldsymbol{M}\boldsymbol{A}^0 = \boldsymbol{M}\boldsymbol{f}^0 - \boldsymbol{K}\boldsymbol{U}^0 \text{ for } \boldsymbol{A}^0
\end{aligned}$$

The consecutive steps $m = 0, 1, \ldots$ depends on the choice of the two Newmark parameters β and γ

$$\begin{aligned}
&\boldsymbol{U}_{\text{tmp}} = \boldsymbol{U}^m + \Delta t \boldsymbol{V}^m + \tfrac{1}{2}\Delta t^2(1-2\beta)\boldsymbol{A}^m\\
&\boldsymbol{V}_{\text{tmp}} = \boldsymbol{V}^m + \Delta t(1-\gamma)\boldsymbol{A}^m\\
&\text{solve } [\boldsymbol{M} + \beta\Delta t^2 \boldsymbol{K}]\boldsymbol{A}^{m+1} = \boldsymbol{M}\boldsymbol{f}^{m+1} - \boldsymbol{K}\boldsymbol{U}_{\text{tmp}}\\
&\boldsymbol{U}^{m+1} = \boldsymbol{U}_{\text{tmp}} + \Delta t^2 \beta \boldsymbol{A}^{m+1}\\
&\boldsymbol{V}^{m+1} = \boldsymbol{V}_{\text{tmp}} + \Delta t \gamma \boldsymbol{A}^{m+1}.
\end{aligned}$$

Certain parameter choices lead to well-known time integration schemes.

Central FD The choice $\beta = 0, \gamma = \frac{1}{2}$ will result in the symmetric and well-known explicit midpoint rule (leap-frog formula if FDM)

$$\boldsymbol{M}\frac{\boldsymbol{U}^{m+1} - 2\boldsymbol{U}^m + \boldsymbol{U}^{m-1}}{\Delta t^2} + \boldsymbol{K}\boldsymbol{U}^m = \boldsymbol{M}\boldsymbol{f}^m.$$

It is second order accurate in time but explicit therefore requiring special care when deciding the time step size Δt.

Trapezoidal rule Choosing $\beta = \frac{1}{4}, \gamma = \frac{1}{2}$ will lead to the so-called trapezoidal rule which also is second order accurate in time but, opposite to the above, implicit and hence unconditionally stable. This scheme is also well-known and is in addition energy conserving [Kre06a]—which is very attractive for conservative systems like the treated.

Runge-Kutta schemes

There exists a rich theory for numerical integration or quadrature which it is more commonly called. The employment of this theory in the field of differential equations of the type

$$\boldsymbol{y}'(t) = \mathcal{L}_h(\boldsymbol{y}(t), t)$$

results in the class of so-called Runge-Kutta methods [Run95, Kut01]. Let us integrate from t_m to $t_{m+1} = t_m + \Delta t$

$$\begin{aligned}\boldsymbol{y}(t_{m+1}) &= \boldsymbol{y}(t_m) + \int_{t_m}^{t_{m+1}} \mathcal{L}_h(\boldsymbol{y}(\tau), \tau)\mathrm{d}\tau \\ &= \boldsymbol{y}(t_m) + \Delta t \int_0^1 \mathcal{L}_h(\boldsymbol{y}(t_m + \Delta t\tau), t_m + \Delta t\tau)\mathrm{d}\tau.\end{aligned}$$

We replace this integral with a quadrature rule on ν nodes and weights c_j and $b_j, j = 1, \ldots, \nu$ like

$$\boldsymbol{y}^{m+1} = \boldsymbol{y}^m + \Delta t \sum_{j=1}^{\nu} b_j \mathcal{L}_h(\boldsymbol{y}(t_m + c_j\tau), t_m + c_j\tau)$$

which leaves open the question on how to approximate $\boldsymbol{y}$ at times later than t_m, that is, $\boldsymbol{y}(t_m + c_j\tau)$ for $j = 2, \ldots, \nu$. The idea of the explicit Runge-Kutta (ERK) method is here to introduce a set of approximants $\boldsymbol{\xi}^j$ for $\boldsymbol{y}(t_m + c_j\tau)$ and then use linear combinations of the preceding approximants for $\boldsymbol{\xi}^{j+1}$.

$$\begin{aligned}\boldsymbol{\xi}^1 &= \boldsymbol{y}^m \\ \boldsymbol{\xi}^2 &= \boldsymbol{y}^m + \Delta t a_{2,1}\mathcal{L}_h(t_m, \boldsymbol{\xi}^1) \\ \boldsymbol{\xi}^3 &= \boldsymbol{y}^m + \Delta t a_{3,1}\mathcal{L}_h(t_m, \boldsymbol{\xi}^1) + \Delta t a_{3,2}\mathcal{L}_h(t_m + c_2\Delta t, \boldsymbol{\xi}^2) \\ &\vdots \\ \boldsymbol{\xi}^\nu &= \boldsymbol{y}^m + \Delta t \sum_{i=1}^{\nu-1} a_{\nu,i}\mathcal{L}_h(t_m + c_i\Delta t, \boldsymbol{\xi}^i).\end{aligned}$$

With these approximants at hand we can complete the quadrature by

$$\boldsymbol{y}^{m+1} = \boldsymbol{y}^m + \Delta t \sum_{j=1}^{\nu} b_j \mathcal{L}_h(\boldsymbol{\xi}^j, t_m + c_j\tau)$$

which constitute an explicit Runge-Kutta method of ν stages. Any such method can be displayed in a RK tableaux

$$\begin{array}{c|cccc} c_1 & & & & \\ c_2 & a_{2,1} & & & \\ \vdots & \vdots & \ddots & & \\ c_\nu & a_{\nu,1} & \cdots & a_{1,\nu-1} & \\ \hline & b_1 & \cdots & b_{\nu-1} & b_\nu. \end{array}$$

Instead of a classic ERK, a low storage ERK (LSERK) with five stages can be used

$$\begin{aligned}
\boldsymbol{\xi}^0 &= \boldsymbol{y}^m \\
&\vdots \\
\boldsymbol{k}^i &= a_i \boldsymbol{k}^{i-1} + \Delta t \mathcal{L}_h(t_m + c_i \Delta t, \boldsymbol{\xi}^{i-1}) \\
\boldsymbol{\xi}^i &= \boldsymbol{\xi}^{i-1} + b_i \boldsymbol{k}^i \\
&\vdots \\
\boldsymbol{y}^{m+1} &= \boldsymbol{\xi}^5
\end{aligned}$$

resulting in a fourth order scheme. We will use this scheme, with the coefficients on page 64 in [HW08], for the DG-FEM semi-discretization.

3.5 Method properties and analysis

We have now introduced several methods for the numerical solution of the wave equation. They will be used, in the next chapter, as an important tool for the numerical solution of the HUM control problem. It is well-known (see, *e.g.,* [Zua05]) that the numerical dispersion and, in particular, the group velocity of a numerical scheme is determining for its ability to deal with control problems. In the next section, we analyze the dispersive properties of some of the schemes introduced above. Section 3.5.2 concludes this chapter with a brief numerical convergence analysis of the schemes verifying the implementation.

3.5.1 The dispersion relation and group velocity

To analyze the presented scheme's dispersive properties, we will assume spatially periodic domains. Let us consider simple periodic solutions of the form

$$y(t, x) = e^{i(\omega t - \xi x)} \tag{3.49}$$

where i is the complex unit, ω is the frequency, and ξ is the wave number. The insertion of this solution in a PDE gives a dispersion relation

$$\omega = \omega(\xi).$$

It is quite easy to see that we for the classical wave equation, $\frac{\partial^2 y}{\partial t^2} - \frac{\partial^2 y}{\partial x^2} = 0$, get the relation $\omega(\xi) = \xi$. And correspondingly for the advection equation, $\frac{\partial y}{\partial t} - a\frac{\partial y}{\partial x} = 0$, we get the relation $\omega(\xi) = a\xi$. These simple linear relationships show that both equations are non-dispersive.

A monochromatic wave travel at the *phase speed*

$$c = \frac{\omega(\xi)}{\xi} \tag{3.50}$$

which by design is $c = 1$ at all wavelengths for our wave equation. A more important concept is, however, the *group velocity* c_g which is defined as

$$c_g = \frac{d\omega(\xi)}{d\xi}. \tag{3.51}$$

The group velocity for the wave equation is again simply $c_g = 1$. A wave packet—and its energy—propagates at the group velocity [Tre82].

We shall now move away from the analysis at PDE level to the analysis of semi-discretizations.

Dispersion analysis of semi-discrete schemes

Let us consider the dispersive properties of the semi-discretization of the wave equation by the unified α-scheme (3.22). We insert the spatially discrete trial solution

$$y(t, x_n) = \widehat{y}_n e^{i(\omega t - \xi x_n)}$$

where $x_n = nh$ and h is the grid spacing. This gives us the numerical dispersion relation [VB82]

$$\omega^\alpha(\xi) = \frac{\sin(\xi h/2)}{\xi h/2} \frac{\xi}{\sqrt{1 - 4\alpha \sin^2(\xi h/2)}},$$

which shows that the scheme, in contrast to the underlying PDE, is dispersive. We get the phase velocity by (3.50)

$$c^\alpha = \frac{\sin(\xi h/2)}{\xi h/2} \frac{1}{\sqrt{1 - 4\alpha \sin^2(\xi h/2)}}.$$

As a measure of the dispersive properties, the phase velocity is plotted in Figure 3.5 for four different choices of the parameter α. Figure 3.6 shows the

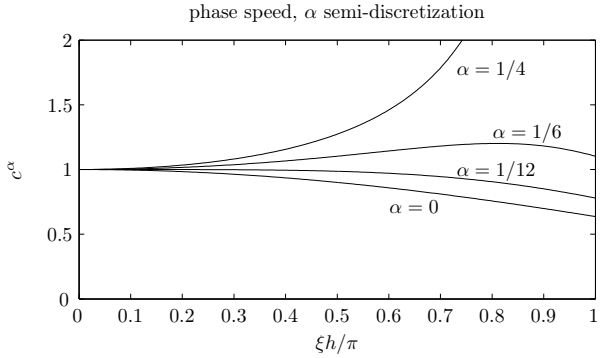


Figure 3.5: Dispersion diagram for the unified scheme showing the phase speed as function of the discrete wavenumber.

corresponding group velocities. Notice that even though the phase speeds are approximated well for $\alpha = 0, 1/12$ and $1/6$, the group velocities behave quite differently; they all tend to zero as $\xi h \to \pi$. It seems that the linear FEM scheme ($\alpha = 1/6$) is best: it is the closest approximant to the correct phase velocity and the group velocity is slightly above the correct value meaning that discretized wave packets travel a little faster than at c_g. This behavior is favorable to group velocities lower than the exact when it comes to control as we shall see in the next chapter.

Things are a little more complicated for the DG-FEM semi-discretization. Let

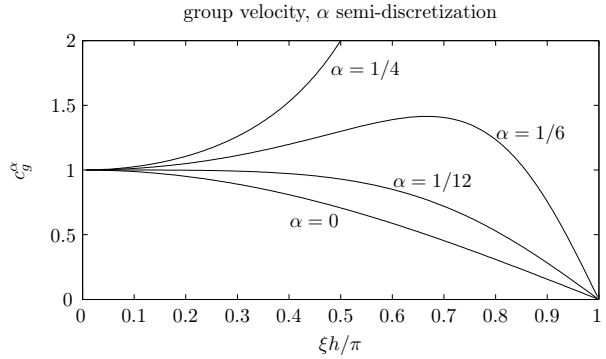


Figure 3.6: The group velocity as function of the discrete wavenumber for the unified semi-discretization.

us consider the advection equation since we use this in the characteristic DG-FEM formulation (3.44). With the upwind flux and $a = 1$, the scheme reads

$$\frac{h^k}{2}\boldsymbol{M}\frac{d\boldsymbol{y}^k}{dt} + \boldsymbol{S}(\boldsymbol{y}^k) = -\boldsymbol{e}_1\big(y_h^k(x_{\mathrm{L}}^k) - y_h^{k-1}(x_{\mathrm{R}}^{k-1})\big),$$

where h^k is the length $(x_{\mathrm{R}}^k - x_{\mathrm{L}}^k)$ of element D^k and $\boldsymbol{e}_i$ is the i'th coordinate vector in $\mathbb{R}^{N_p}$. We assume, in the following, that the element size is constant $h = h^k$. We suggest the trial solution

$$\boldsymbol{y}^k(t, x^k) = \widehat{\boldsymbol{y}}^k e^{i(\omega t - \xi x^k)}$$

with the N_p sized coefficient vector $\widehat{\boldsymbol{y}}^k$. As in [HW08, page 89], we assume periodicity of the solutions and obtain the generalized eigenvalue problem

$$\big(\boldsymbol{S} + \boldsymbol{e}_1(\boldsymbol{e}_1^{\mathsf{T}} - e^{i\xi h}\boldsymbol{e}_{N_p}^{\mathsf{T}})\big)\widehat{\boldsymbol{y}}^k = i\omega\frac{h}{2}\boldsymbol{M}\widehat{\boldsymbol{y}}^k \tag{3.52}$$

It is solved numerically and from the solution we derive the phase velocities c^{DG,N_p} for $N_p = 3, \ldots, 7$ as shown in Figure 3.7. Differentiation gives the corre-

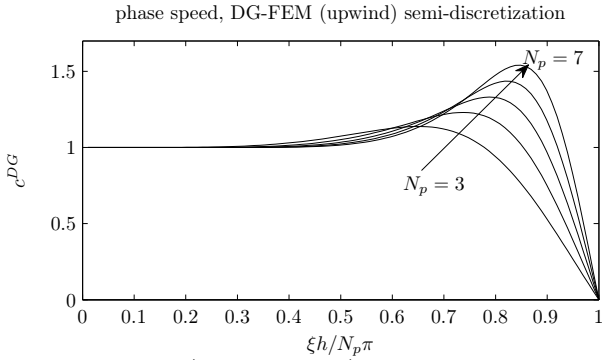


Figure 3.7: The phase speed (physical mode) as function of the discrete wavenumber for the DG-FEM semi-discretization of the advection equation with upwind flux (3.30).

sponding group velocities c_g^{DG,N_p} shown in Figure 3.8. See also the zoom of left

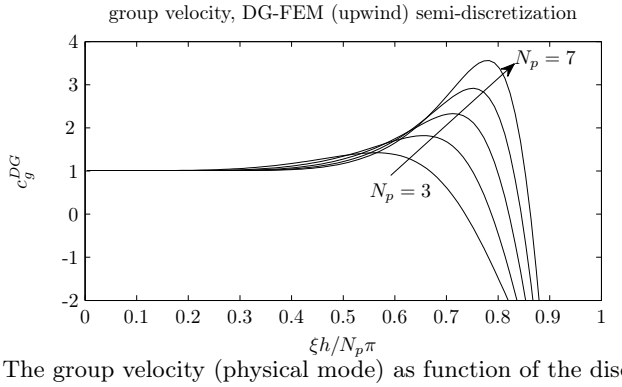


Figure 3.8: The group velocity (physical mode) as function of the discrete wavenumber for the DG-FEM semi-discretization of the advection equation with upwind flux (3.30).

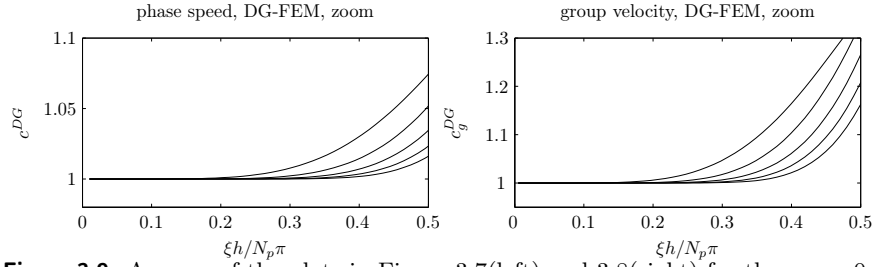


Figure 3.9: A zoom of the plots in Figure 3.7(left) and 3.8(right) for the range $0 \leq \xi h/N_p \leq \pi/2$. In both plots, the graphs range from $N_p = 3$ (uppermost) to $N_p = 7$ (lower-most).

half of these two plots in Figure 3.9.

The plots may be compared with those for the unified scheme in Figure 3.5 and 3.6. Note, however, that Figure 3.5 and 3.6 show the dispersive properties of a semi-discretization of the *wave* equation, while Figures 3.7–3.9 are for a semi-discretization of the *advection* equation. But since the wave equation in essence is two advection equations, it is still meaningful to compare the plots.

The low wavenumbers on Figure 3.7 (and left part of Figure 3.9) show that DG-FEM exhibit accurate dispersive behavior and the higher the polynomial order N_p, the better. For high wavenumbers, especially the group velocity show strongly unphysical behavior going negative after about $\xi h/N_p = 0.85\pi$; energy can be propagated in the opposite direction of the physical one by high wave number components. But what we do *not* see on this plot is the corresponding damping—or dissipation—which, luckily, is very strong in the same region. The dissipation is found as the imaginary part of the complex eigenvalues of (3.52).

Dispersion analysis of fully discrete schemes

We shall now see what happens after we integrate the ODEs of the semi-discrete schemes in time. Let Δt be the time step size in the following. Besides the time integration scheme, the Courant number, $\mu = \Delta t/\Delta x$ where Δx is the spatial grid spacing, plays a significant role for the dispersive behavior.

Consider the fully discrete trial solution

$$y_n^m = e^{i(\omega t_m - \xi x_n)}$$

where $t_m = m\Delta t$. We insert this in a fully discrete scheme, which for, *e.g.,* the unified scheme with central finite difference time discretization yields

$$\boldsymbol{M}\frac{\boldsymbol{y}^{m+1} - 2\boldsymbol{y}^m + \boldsymbol{y}^{m-1}}{\Delta t^2} + \boldsymbol{K}\boldsymbol{y}^m = 0.$$

After insertion, we obtain an expression in ξh with μ as an important parameter [Ras04], in this case

$$\sin^2(\omega\Delta t/2) = \frac{\mu^2 \sin^2(\xi h/2)}{1 - 4\alpha \sin^2(\xi h/2)}.$$

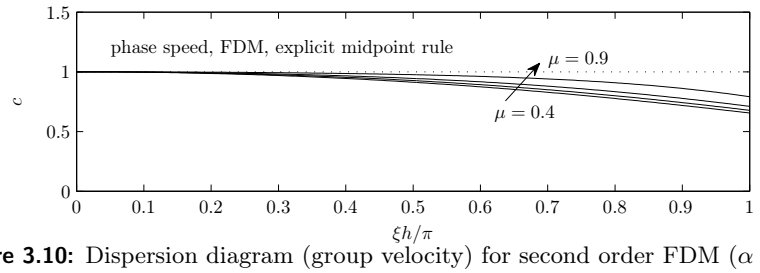


Figure 3.10: Dispersion diagram (group velocity) for second order FDM ($\alpha = 0$) with explicit midpoint rule time integration.

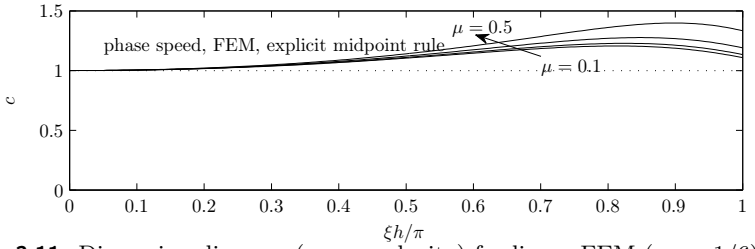


Figure 3.11: Dispersion diagram (group velocity) for linear FEM ($\alpha = 1/6$) with explicit midpoint rule time integration.

We show the corresponding *group velocity* for different Courant numbers μ on Figure 3.12 for the FDM ($\alpha = 0$) and on Figure 3.13 for the FEM ($\alpha = 1/6$). The graphs on both figures resemble quite closely the graphs for $\alpha = 0$ and $\alpha = 1/6$ on Figure 3.6. The time integration has in this case only little effect on the dispersive behavior of the semi-discrete schemes. Increasing the Courant number does, however, have an amplifying effect on the graphs. Conversely, it tends, as expected, to the time continuous case as $\mu \to 0$.

Let us also consider trapezoidal time integration. Integrating (3.22) by the trapezoidal scheme gives [Ras04]

$$\tan^2(\omega\Delta t/2) = \frac{\mu^2 \sin^2(\xi h/2)}{1 - 4\alpha \sin^2(\xi h/2)}.$$

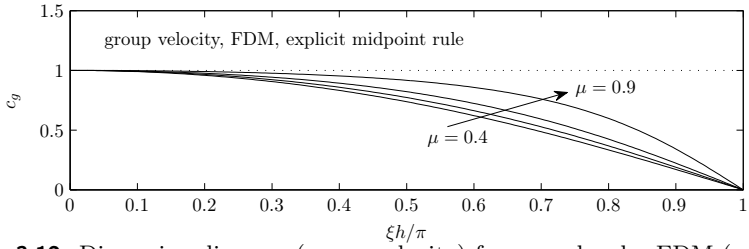


Figure 3.12: Dispersion diagram (group velocity) for second order FDM ($\alpha = 0$) with explicit midpoint rule time integration.

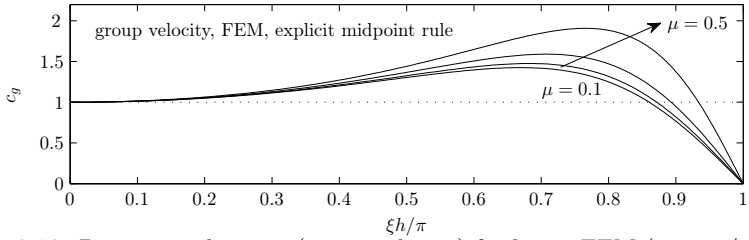


Figure 3.13: Dispersion diagram (group velocity) for linear FEM ($\alpha = 1/6$) with explicit midpoint rule time integration.

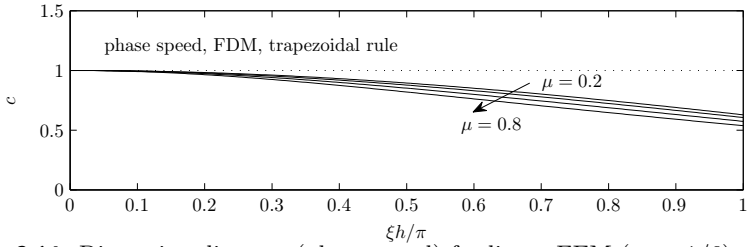


Figure 3.14: Dispersion diagram (phase speed) for linear FEM ($\alpha = 1/6$) with implicit trapezoidal time integration.

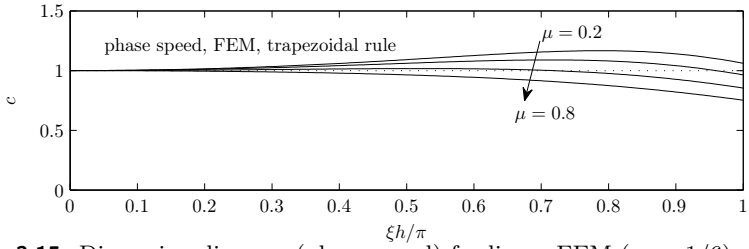


Figure 3.15: Dispersion diagram (phase speed) for linear FEM ($\alpha = 1/6$) with implicit trapezoidal time integration.

We show the phase speed for FDM in Figure 3.14 and for FEM in Figure 3.15 for different Courant numbers μ. Figure 3.16 and 3.17 show the corresponding group velocities. This time integration has, contrary to integration by central finite differences, a strong effect on the dispersive behavior of both semi-discrete schemes. The effect of increasing the Courant number is also opposite; here it reduces the group velocities.

We choose linear FEM with trapezoidal integration for the use with numeri-

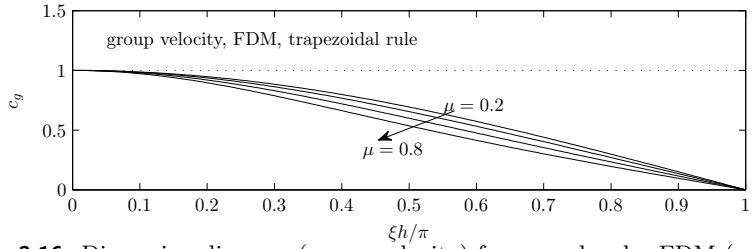


Figure 3.16: Dispersion diagram (group velocity) for second order FDM ($\alpha = 0$) with implicit trapezoidal time integration.

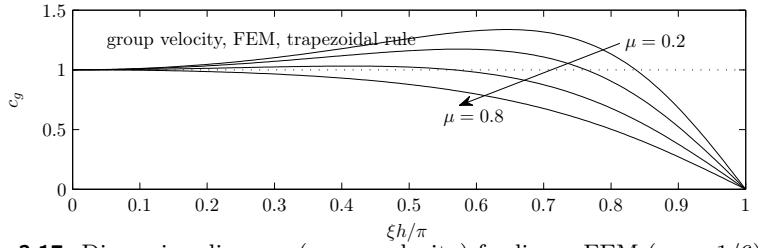


Figure 3.17: Dispersion diagram (group velocity) for linear FEM ($\alpha = 1/6$) with implicit trapezoidal time integration.

cal HUM. Both phase (Figure 3.15) and group velocity (Figure 3.17) is approximated well by this scheme, and it seems that the Courant number $\mu = 0.6$ gives the best approximation. The explicit midpoint rule, on the contrary, takes FEMs semi-discrete dispersion relation (Figurea 3.5–3.6 with $\alpha = 1/6$) in the wrong direction. The FDM with explicit midpoint rule integration seems a worthy contender to FEM with trapezoidal integration, but the scheme is singular in the sense that it consists of the same approximation in space and time whose shortcomings therefore cancel out.

We integrate DG-FEM with a higher order Runge-Kutta scheme (LSERK) which generally has smaller effect on the dispersion relation in the interesting low wavenumber region. The precise analysis can be made with the use of Padé approximants.

3.5.2 Convergence

Finally, we shall very shortly demonstrate the convergence of the DG-FEM scheme (3.44) for smooth data. We consider a homogeneous wave equation in w with the arbitrarily chosen smooth initial data

$$w^0(x) = \sin(2\pi x) - 0.3\sin(4\pi x),$$
$$w^1(x) = \sin(3\pi x) - \tfrac{\pi}{5}\sin(5\pi x).$$

It is integrated in time by an LSERK method [HW08, page 64] (see also page 43) until $T = 1.87\sqrt{2}$ for different values of the polynomial order $N_p - 1$ and number of elements K. We study the so-called h-convergence on Figure 3.18 by fixing $N_p = 3$ and varying the number of elements K. The p-convergence is studied by increasing the order N_p for a fixed number of elements. A convergence plot of this kind is shown on Figure 3.19 with $K = 6$. We will refrain from

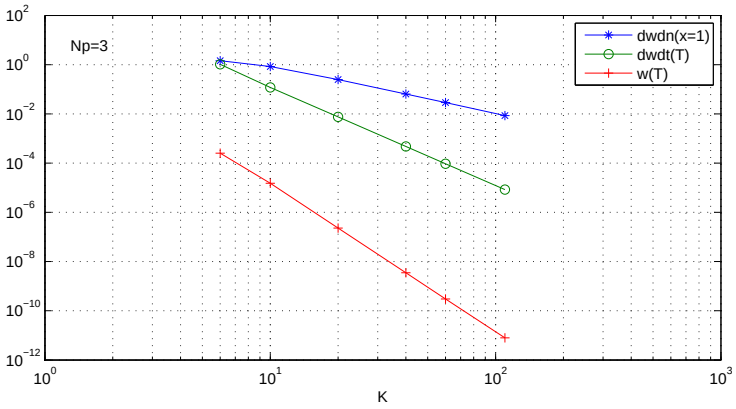


Figure 3.18: h-convergence for the DG-FEM scheme (3.44) with $N_p = 3$. The graphs represent the L^2-norm of the error for respectively the normal derivative $\frac{\partial w_h}{\partial n}$ at $x = 1$, the time derivative $\frac{\partial w_h}{\partial t}$ at $t = T$, and w_h and $t = T$.

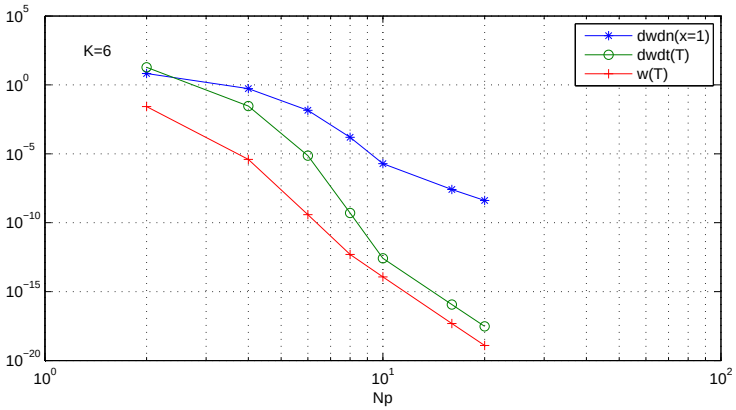


Figure 3.19: p-convergence for the DG-FEM scheme (3.44) with $K = 6$. The graphs represent the L^2-norm of the error for respectively the normal derivative $\frac{\partial w_h}{\partial n}$ at $x = 1$, the time derivative $\frac{\partial w_h}{\partial t}$ at $t = T$, and w_h and $t = T$.

going deeper into this analysis and conclude here only that the scheme (and its implementation) is convergent—at least for smooth data.

CHAPTER 4

Numerical HUM

The topic of this chapter is the numerical approximation of HUM. Chapter 3 introduced discretizations for the wave equation—the main ingredient in numerical HUM. We discretize the wave equation in order to find approximate controls. High-frequency spurious (non-physical) solutions to the wave equation arising from this discretization are known to pose a serious threat to the convergence of the approximate controls [Zua05]. The phenomena is closely related to the dispersion relation and the group velocity of the discretization; using a convergent scheme for the approximation of the wave equation is no guarantee in terms of convergence of controls. The diagram in Figure 4.1 gives schematic view of the challenge. Luckily, it is also known that convergence can

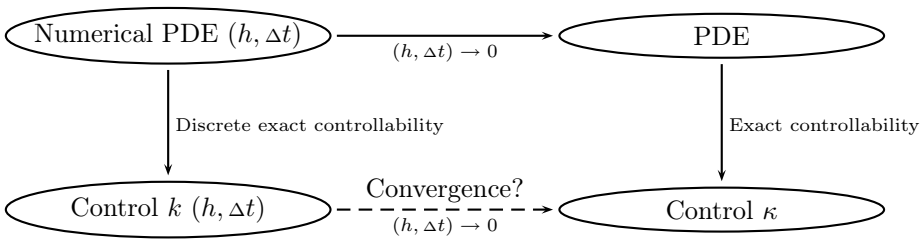


Figure 4.1: Diagram over the convergence of numerical HUM.

be restored by proper filtering or regularization. We will return to this.

HUM is a very general method able to deal with multi-dimensional problems. The numerical approximation of it is, however, not even fully understood in 1-d. Furthermore, working with only one spatial dimension offers some advantages over higher dimensions in terms of reduced (simple characteristic rays) dynamics, fewer technicalities and access to exact solutions. HUM is not necessary for boundary controllability in 1-d, though, other methods could be applied. The motivation is, however, not the control of a 1-d problem, but HUM it-self and, as the ultimate goal, HUM control of multidimensional problems in complex geometries. This is the background for this study of the numerical HUM for the 1-d wave equation.

In this chapter, we study different ways of finding numerical HUM controls. Several theoretical accounts on the relation between numerical dispersion and control can be found in the literature. We will take a more practical approach and study the consequences of inexact phase and group velocities in concrete cases for linear FEM (L-FEM). The discontinuous Galerkin-FEM (DG-FEM) was introduced in Chapter 3 and is very well-suited for wave problems. It has not previously been used for HUM-control. We study the abilities of DG-FEM for numerical HUM in this chapter and compare the results with those obtained by L-FEM.

We wish, furthermore, to examine the effect of discretizing the problem in sine basis as the sine basis is a natural eigenfunction basis for the problem and it relates well to the numerical dispersion of a discretization. This has not previously been described in the literature.

This chapter sets off by a short review of linear finite-dimensional control theory in Section 4.1.1. A finite-dimensional control problem is exactly what we obtain after spatial semi-discretization of our infinite-dimensional HUM control problem. The semi-discrete HUM will be described in Section 4.1.2. We study two semi-discretizations, L-FEM and DG-FEM, in detail; we review their properties here, too. The fully discrete HUM problem is described in Section 4.1.3.

We deal with the direct solution of HUM in Section 4.2 where we present the construction of the $\boldsymbol{L}$ matrix approximating the fundamental Λ operator. This can be done by either direct or inner-product assembly. We also show how the discretization of the minimization formulation of HUM is closely related to the inner-product assembly in Section 4.2.2. Hereafter, Section 4.2.3 shows how the $\boldsymbol{L}$ matrix is constructed in sine basis. We present the numerical study of this construction with L-FEM and DG-FEM in Sections 4.2.4 and 4.2.5, respectively.

The explicit construction of the $\boldsymbol{L}$ matrix rapidly becomes infeasible as the degrees of freedom increase. Section 4.3 is about an iterative alternative to the direct approach. We present the classical conjugate gradient algorithm formulated for HUM by Glowinski, Li and Lions [GLL90] in Section 4.3.1. In Section 4.3.2, we introduce a new filtering step in the algorithm consisting of basis truncation. Section 4.3.3 features a test problem and we examine different values of the truncation parameter as well as convergence for L-FEM and DG-FEM. Finally, Section 4.4 concludes the chapter by a discussion and a short review of related work.

4.1 Discrete control

We need to discretize the system that we want to control in order to obtain approximate controls. The system is a wave equation, a PDE, and we discretize in two steps: first spatially which results in a semi-discrete model, a system of ODEs; secondly, time integration leads to a fully discrete model. The spatial discretization is in many ways the most important step as it determines most of the dynamics of the model, suggests norm approximations, and gives rise to a finite-dimensional control system of ODEs. The ODE control problem is linear and could be approached with the standard tools from classical control theory. But the transition from PDE to ODE also means that the rich Hilbert space structure of the PDE controllability problem, for which the right spaces and norms are truly essential to its solution, is replaced by an ODE setting with much less structure. Suddenly does the control time T, for example, according to classical control theory, not matter, even though we *know* from infinite dimensional theory that it should, due to the finite speed of propagation. In this way, we may obtain non-physical solutions to the control problem if we are not careful. For this reason, we will focus much on semi-discrete systems in this section. We begin with the most simple ideas from the huge area of control of finite-dimensional systems.

4.1.1 Classical control theory

The state-space representation of a control system, which is very common in classical finite dimensional control theory (see, *e.g.,* [CF03]), reads

$$\boldsymbol{x}'(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{v}(t), \qquad 0 \le t \le T \tag{4.1a}$$

$$\boldsymbol{x}(0) = \boldsymbol{x}^0, \tag{4.1b}$$

with state $\boldsymbol{x}(t)$ in the state space $\mathcal{X}$ and control $\boldsymbol{v}(t)$ in the control space $\mathcal{U}$. The $N \times N$ matrix $\boldsymbol{A}$ is denoted the system matrix and $\boldsymbol{B}$ is the input matrix and is of size $N \times \widetilde{N}$ where $\widetilde{N} \le N$.

We define the controllability matrix

$$\boldsymbol{W}_c := [\boldsymbol{A}^{N-1}\boldsymbol{B}, \boldsymbol{A}^{N-2}\boldsymbol{B}, \ldots, \boldsymbol{A}\boldsymbol{B}, \boldsymbol{B}]$$

with the i'th column being the vector $\boldsymbol{A}^{N-i}\boldsymbol{B}$. The exact controllability of (4.1) may be checked with the (Kalman) rank condition.

Theorem 4.1 (Rank Condition). *The pair $\{\boldsymbol{A}, \boldsymbol{B}\}$, i.e., system (4.1), is controllable if and only if the controllability matrix $\boldsymbol{W}_c$ has full rank N. If so the system is controllable to any time $T > 0$.* □

We refer to [CF03] for a proof. The rank condition is easy to test in the special case where $\boldsymbol{A}$ is diagonalizable. We state the following result without proof (see [Ras04]).

Theorem 4.2. *Let $\boldsymbol{A}$ be diagonalizable, $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1}$, and let $\boldsymbol{B}$ be a column vector. The control system (4.1) is controllable if and only if the eigenvalues of $\boldsymbol{A}$ are distinct and the vector $\boldsymbol{V}^{-1}\boldsymbol{B}$ contains no zero values.* □

Writing the control system in eigenvector basis

$$\boldsymbol{V}^{-1}\boldsymbol{x}'(t) = \boldsymbol{D}\boldsymbol{V}^{-1}\boldsymbol{x}(t) + \boldsymbol{V}^{-1}\boldsymbol{B}\boldsymbol{v}(t)$$

shows us why all elements of $\boldsymbol{V}^{-1}\boldsymbol{B}$ must be non-zero; any zero element would leave that specific eigenmode uncontrollable.

The adjoint system associated to (4.1) is

$$\boldsymbol{y}'(t) = \boldsymbol{A}^*\boldsymbol{y}(t), \tag{4.2a}$$
$$\boldsymbol{y}(0) = \boldsymbol{y}^0, \tag{4.2b}$$

with the "output" or observation

$$\boldsymbol{C}\boldsymbol{y}(t)$$

where $\boldsymbol{C} = \boldsymbol{B}^*$. An observability matrix, whose non-singularity is equivalent to the observability of (4.2), may be constructed in a similar way as the controllability matrix above. The pair $\{\boldsymbol{A}, \boldsymbol{B}\}$ is controllable if and only if the pair $\{\boldsymbol{A}^*, \boldsymbol{B}^*\}$ is observable (see [CF03])—just as we saw for the wave equation in Chapter 2.

Much more could be said about the classical control theory, but we refer instead to the works of Sontag [Son98], Corless and Frazho [CF03], and for a more HUM-minded approach to Micu and Zuazua [MZ05].

4.1.2 Semi-discrete HUM

Chapter 3 showed how the wave equation can be approximated by systems of ODEs. Let, in the subsequent exposition, $N \in \mathbb{N}$ be the number of elements in a vector $\boldsymbol{y}$ approximating the function y defined on Ω. The approximation of the infinite dimensional energy space $\mathcal{E} = H_0^1(\Omega) \times L^2(\Omega)$ which we call $\mathcal{X}$ is therefore $2N$-dimensional. We use $\mathcal{X}^\star$ to denote the approximation of $\mathcal{E}^\star = H^{-1}(\Omega) \times L^2(\Omega)$; this approximation is also $2N$-dimensional.

The distinction between the spaces $\mathcal{X}$ and $\mathcal{X}^\star$ may, at this point, seem somewhat artificial since $\mathcal{X} = \mathcal{X}^\star = \mathbb{R}^{2N}$ and all norms are equivalent in $\mathbb{R}^{2N}$. We wish, however, to equip $\mathcal{X}$ and $\mathcal{X}^\star$ with discrete norms reflecting the properties of their infinite dimensional ancestors. Let $(y, z) \in \mathcal{E}$ and $(u, v) \in \mathcal{E}^\star$ and let $[\boldsymbol{y}, \boldsymbol{z}]^\mathsf{T} \in \mathcal{X}$ and $[\boldsymbol{u}, \boldsymbol{v}]^\mathsf{T} \in \mathcal{X}^\star$ be their approximating vectors. The norms of $\mathcal{X}$ and $\mathcal{X}^\star$ are

$$\left\| \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} \right\|_{\mathcal{X}} = \|\boldsymbol{y}\|_1 + \|\boldsymbol{z}\|_0 \tag{4.3}$$
$$\left\| \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} \right\|_{\mathcal{X}^\star} = \|\boldsymbol{u}\|_{-1} + \|\boldsymbol{v}\|_0 \tag{4.4}$$

where $\|\cdot\|_1$ is an approximation of the $H_0^1(\Omega)$ norm, $\|\cdot\|_0$ an approximation to the $L^2(\Omega)$ norm, and $\|\cdot\|_{-1}$ to the $H^{-1}(\Omega)$ norm. We approximate the duality product $\langle \cdot, \cdot \rangle_{\mathcal{E}^\star, \mathcal{E}}$ by

$$\left\langle \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix}, \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} \right\rangle_{\mathcal{X}^\star, \mathcal{X}} = \langle \boldsymbol{u}, \boldsymbol{y} \rangle_{-1,1} + \langle \boldsymbol{v}, \boldsymbol{z} \rangle_0, \tag{4.5}$$

where $\langle \cdot, \cdot \rangle_{-1,1}$ approximates the duality product between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$. The discrete norms and the duality product depend on the choice of semi-discretization. They will be specified below for L-FEM and DG-FEM.

Let us consider a generic semi-discrete model of the wave equation (3.1)

$$\boldsymbol{Y}'(t) = \boldsymbol{\mathcal{L}}_h \boldsymbol{Y}(t) + \boldsymbol{B}_h g_1(t), \qquad t \in [0,T] \tag{4.6a}$$
$$\boldsymbol{Y}(0) = \boldsymbol{Y}^0, \tag{4.6b}$$

where $\boldsymbol{Y}(t)$ consists of two N-sized[1] vectors, *e.g.*, $\boldsymbol{Y}(t) = [\boldsymbol{y}, \boldsymbol{z}]^\mathsf{T}$ with corresponding initial data $\boldsymbol{Y}^0$. The system matrix $\boldsymbol{\mathcal{L}}_h$ of size $2N \times 2N$ is an approximation to the spatial differential operator and the boundary matrix $\boldsymbol{B}_h$ assigns the scalar boundary condition $g_1(t)$ to the system. Both of these matrices have elements according to the choice of semi-discretization (see below).

Based on the model (4.6), we introduce a $2N$-sized approximation to the control system (2.1)

$$\boldsymbol{U}'(t) = \boldsymbol{\mathcal{L}}_h \boldsymbol{U}(t) + \boldsymbol{B}_h k(t), \qquad t \in [0,T] \tag{4.7a}$$
$$\boldsymbol{U}(0) = \boldsymbol{U}^0, \tag{4.7b}$$

where the initial data (u^0, u^1) of the continuous system (2.1) have been sampled, or projected onto $\mathcal{X}^\star$, resulting in the vector $\boldsymbol{U}^0 = [\boldsymbol{u}^0, \boldsymbol{u}^1]^\mathsf{T}$. The function k is a boundary control applied at the right end of the domain. This representation is in state space form (4.1). The state space is $\mathcal{X}$ and the control space is in this case simply $\mathbb{R}$ as k is a scalar function.

We associate a semi-discrete control problem to system (4.7).

Definition 4.3. Given $\boldsymbol{U}^0 \in \mathcal{X}^\star$, find $k \in \mathcal{B}$ such that (4.7) is steered to zero in time $t = T$, *i.e.*, $\boldsymbol{U}(T) = 0$. □

The semi-discrete control problem is a finite dimensional control problem in state space form (4.1). We may examine whether the pair $\{\boldsymbol{\mathcal{L}}_h, \boldsymbol{B}_h\}$ is controllable for a choice of semi-discretization by Theorem 4.2, but a positive result is no guarantee in terms of physically meaningful controls. Applying the methods from classical control will in general not lead to useful controls for the PDE.

We proceed with the HUM approximation by considering the $2N$-sized approximation of the adjoint system (2.6) with the variable $\boldsymbol{W}$ as the semi-discrete counterpart to φ

$$\boldsymbol{W}'(t) = \boldsymbol{\mathcal{L}}_h \boldsymbol{W}(t), \qquad t \in [0,T] \tag{4.8a}$$
$$\boldsymbol{W}(0) = \boldsymbol{W}^0, \tag{4.8b}$$

for which the initial data $\boldsymbol{W}^0 = [\boldsymbol{w}^0, \boldsymbol{w}^1]^\mathsf{T}$ corresponds to the continuous (φ^0, φ^1) of (2.6). Further, let $\boldsymbol{C}_h \boldsymbol{W}(t)$ denote an approximation to the normal derivative $\frac{\partial}{\partial n}$ at $x = 1$

$$\frac{\partial}{\partial n}\varphi(1,t) \approx \boldsymbol{C}_h \boldsymbol{W}(t). \tag{4.9}$$

The output matrix $\boldsymbol{C}_h$ needs to provide a convergent approximation of $\frac{\partial}{\partial n}$ for $h \to 0$.

Finding the observation $\boldsymbol{C}_h \boldsymbol{W}(t)$ from the initial data $\boldsymbol{W}^0$ is called semi-discrete observation. It is now formally defined.

[1] In the case of DG-variables each vector consists of $N_p \cdot K$ elements; we use the convention $N = N_p \cdot K$ in this case.

Definition 4.4 (sd-observation). Let $\boldsymbol{W}^0 \in \mathcal{X}$ be initial data for (4.8), and let $\boldsymbol{W}(t)$ be the solution of (4.8) for this initial data. Computing the Neumann output $\boldsymbol{C}_h\boldsymbol{W}(t)$ is then called *sd-observation*. The related operator

$$P^{\text{sd}} : \mathcal{X} \to \mathcal{B} \quad \text{defined by} \quad P^{\text{sd}} : \boldsymbol{W}^0 \mapsto \boldsymbol{C}_h\boldsymbol{W}(t), \tag{4.10}$$

is called the *sd-observation operator*. □

Note that P^{sd} is a semi-discrete approximation to the observation operator Φ defined in (2.14). Note also that sd-observation requires an *exact* solution of (4.8) as time is still continuous.

The semi-discrete operator equation

Let us also define a semi-discrete version of system (2.5)

$$\boldsymbol{Z}'(t) = \boldsymbol{\mathcal{L}}_h\boldsymbol{Z}(t) + \boldsymbol{B}_h k(t), \qquad t \in [0, T] \tag{4.11a}$$
$$\boldsymbol{Z}(T) = \boldsymbol{0}, \tag{4.11b}$$

which is solved backwards in time, so that the output at $t = 0$ becomes $\boldsymbol{Z}(0)$. Parallel to the reconstruction introduced in Section 2.4, we will define the semi-discrete reconstruction.

Definition 4.5 (sd-reconstruction). Given a function $k \in \mathcal{B}$ assume that (4.11) is solvable for that k. We solve (4.11) to obtain the output $[\boldsymbol{z}'(0), -\boldsymbol{z}(0)]^{\mathsf{T}} \in \mathcal{X}^\star$ and call this operation *sd-reconstruction*. The corresponding operator

$$R^{\text{sd}} : \mathcal{B} \to \mathcal{X}^\star \quad \text{defined by} \quad R^{\text{sd}} : k \mapsto \begin{bmatrix} \boldsymbol{z}'(0) \\ -\boldsymbol{z}(0) \end{bmatrix}, \tag{4.12}$$

is denoted the *sd-reconstruction operator*. □

R^{sd} is a semi-discrete approximation to the reconstruction operator Ψ defined in (2.17).

We seek a specific function $k \in \mathcal{B}$ that satisfies the requirement

$$R^{\text{sd}} k = \begin{bmatrix} \boldsymbol{u}^1 \\ -\boldsymbol{u}^0 \end{bmatrix}.$$

Such a function k will by construction solve the semi-discrete control problem and is hence called a control.

A control k driving the semi-discrete system (4.7) to zero in time $t = T$ is called a HUM-control if it is formed by

$$k = P^{\text{sd}} \begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix} \tag{4.13}$$

where $\boldsymbol{W}^0 = [\boldsymbol{w}^0, \boldsymbol{w}^1]^{\mathsf{T}}$ is a set of initial data for the semi-discrete adjoint system (4.8).

The semi-discrete HUM operator equation becomes

$$L^{\text{sd}} \begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}^1 \\ -\boldsymbol{u}^0 \end{bmatrix}, \tag{4.14}$$

where the operator L^{sd}

$$L^{\text{sd}} : \mathcal{X} \to \mathcal{X}^\star \quad \text{defined by} \quad L^{\text{sd}} := R^{\text{sd}} P^{\text{sd}} \tag{4.15}$$

approximates the Λ operator (2.21). The operator L^{sd} takes discrete initial data $[\boldsymbol{w}^0, \boldsymbol{w}^1]^\mathsf{T}$ and maps it to the approximate Neumann boundary data $\boldsymbol{C}_h \boldsymbol{W}(t)$, then takes this data as Dirichlet boundary condition $k(t) = \boldsymbol{C}_h \boldsymbol{W}(t)$ and maps it onto the $t = 0$ state $[\boldsymbol{z}'(0), -\boldsymbol{z}(0)]^\mathsf{T}$. If the solution $\overline{\boldsymbol{W}}^0$ to the HUM equation (4.14) exists, it provides the sought control by $k = P^{\text{sd}} \overline{\boldsymbol{W}}^0$.

The operator L^{sd} maps between two $2N$ dimensional vector spaces and can therefore be considered as a $2N$ by $2N$ matrix. Finding the matrix elements, however, would require the *exact* solutions of (4.8) and (4.11). Obviously, we need to integrate in time to make the systems fully discrete. Before doing so we will go through some details about the two semi-discretizations, L-FEM and DG-FEM, that we will study throughout this chapter.

L-FEM semi-discretization

The unified scheme (3.24) has the linear FEM (L-FEM) as a special case with $\alpha = 1/6$. Let $\boldsymbol{y}$ be a column vector with N nodal values $y(x_n)$ for $n = 1, \ldots, N$ of the function y defined on Ω. The L-FEM approximation y_h reads

$$y \approx y_h = \sum_{n=1}^{N} y_h(x_n) \psi_n^{\text{L}}(x)$$

where ψ_n^{L} is the linear hat basis function (3.17).

Let furthermore $\boldsymbol{e}_n$ be a coordinate vector in $\mathbb{R}^N$. The system and boundary matrix of the generic model (4.6), where $\boldsymbol{Y}(t) = [\boldsymbol{y}(t), \boldsymbol{y}'(t)]^\mathsf{T}$, become (see also Section 3.2.3)

$$\boldsymbol{\mathcal{L}}_h = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ -\boldsymbol{M}^{-1}\boldsymbol{K} & \boldsymbol{0} \end{bmatrix}, \qquad \boldsymbol{B}_h = \begin{bmatrix} \boldsymbol{0} \\ \frac{1}{h}\boldsymbol{e}_N \end{bmatrix}. \tag{4.16}$$

The mass matrix $\boldsymbol{M}$ is defined in (3.20) with $\alpha = 1/6$ and the stiffness matrix $\boldsymbol{K}$ in (3.21). We check that the ODE system is controllable by Theorem 4.2.

We approximate the normal derivative at the right endpoint with a simple first order finite difference $(\boldsymbol{y}_{N+1} - \boldsymbol{y}_N)/h = -\boldsymbol{y}_N/h$. This gives the output matrix

$$\boldsymbol{C}_h = \begin{bmatrix} -\frac{1}{h}\boldsymbol{e}_N^\mathsf{T} & \boldsymbol{0}^\mathsf{T} \end{bmatrix}. \tag{4.17}$$

The mass and stiffness matrices give rise to natural approximations of the L^2- and H^1-norms. Let u and v be functions in $L^2(\Omega)$ and u_h and v_h their finite element approximations. Let $\boldsymbol{u}$ be an N sized column vector with $u_h(x_n)$ at position n and likewise for $\boldsymbol{v}$.

The approximate L^2-inner product $\langle \cdot, \cdot \rangle_0$ becomes

$$\begin{aligned} \langle u, v \rangle_{L^2(\Omega)} \approx \langle u_h, v_h \rangle_{L^2(\Omega)} &= \sum_{i=1}^{N} \sum_{j=1}^{N} u_h(x_i) v_h(x_j) \langle \psi_i^{\text{L}}, \psi_j^{\text{L}} \rangle_{L^2(\Omega)} \\ &= \boldsymbol{u}^\mathsf{T} \boldsymbol{M} \boldsymbol{v} =: \langle \boldsymbol{u}, \boldsymbol{v} \rangle_0 \end{aligned} \tag{4.18}$$

according to the definition of the mass matrix (3.15). Equivalently, we have for $u, v \in H_0^1(\Omega)$ and their approximants u_h and v_h

$$\langle u, v\rangle_{H^1(\Omega)} \approx \langle u_h, v_h\rangle_{H^1(\Omega)} = \sum_{i=1}^{N}\sum_{j=1}^{N} u_h(x_i)v_h(x_j)\langle \psi_i^{\mathrm{L}}, \psi_j^{\mathrm{L}}\rangle_{H^1(\Omega)} = \boldsymbol{u}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{v} =: \langle \boldsymbol{u}, \boldsymbol{v}\rangle_1 \tag{4.19}$$

c.f. the definition of the stiffness matrix (3.16). We will use the norms $\|\cdot\|_0$ and $\|\cdot\|_1$ induced by the above inner products. This also gives the $\mathcal{X}$-norm by (4.3).

Rasmussen deduced in [Ras04], by energy conservation principles, a discrete norm for $\mathcal{X}^\star$ by the $H^{-1}(\Omega)$-inner product for $u, v \in H^{-1}(\Omega)$

$$\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{-1} := \boldsymbol{u}^{\mathsf{T}}\boldsymbol{M}\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{v} \tag{4.20}$$

where $\boldsymbol{M}^{-1}\boldsymbol{K} = (\boldsymbol{K}^{-1}\boldsymbol{M})^{-1}$ approximates the Laplacian. For the duality product (4.5) Rasmussen used

$$\langle \boldsymbol{u}, \boldsymbol{w}\rangle_{-1,1} := \boldsymbol{u}^{\mathsf{T}}\boldsymbol{M}\boldsymbol{w} \tag{4.21}$$

where $\boldsymbol{u}$ and $\boldsymbol{w}$ are approximating vectors of $u \in H^{-1}(\Omega)$ and $v \in H_0^1(\Omega)$, respectively. We will use these approximations too and have thus by (4.4) and (4.5) specified the $\mathcal{X}^\star$-norm and the duality product between $\mathcal{X}$ and $\mathcal{X}^\star$. This concludes the review of L-FEM.

DG-FEM semi-discretization

Let us consider the DG-FEM semi-discretization in characteristic variables p and q introduced in (3.44) (see also Section 3.1.1).

The LGL-grid and the local higher order polynomial basis on each element are some of the major differences compared to the L-FEM approach. The domain is divided into K elements each of which has N_p interpolation points that are LGL distributed. The local polynomial basis has order $N_p - 1$. Functions y defined on Ω are represented on each element D^k by (3.37)—a sum of N_p interpolating Lagrange basis functions ℓ_i (nodal) or a sum of N_p normalized Legendre polynomials $\widetilde{P}_n$ (modal)

$$y \approx y_h = \bigoplus_{k=1}^{K} y_h^k, \qquad y_h^k(t,x) = \sum_{i=1}^{N_p} y_h^k(t, x_i^k)\ell_i^k(x) = \sum_{n=1}^{N_p} \widehat{y}_n^k(t)\widetilde{P}_{n-1}(x).$$

Recall also that we do not require functions to be continuous across interfaces. Local mass and stiffness matrices $\boldsymbol{M}^k$ and $\boldsymbol{K}^k$ are defined on each element D^k in (3.31) and the K local approximations are connected via the numerical flux which accounts for the flow of information between elements. We will use upwind fluxes (3.32); left-bound for p and right-bound for q.

We fit the DG semi-discretization in the generic model (4.6) by letting $\boldsymbol{Y}(t) = [\boldsymbol{p}(t), \boldsymbol{q}(t)]^{\mathsf{T}}$ where $\boldsymbol{p}$ is the nodal vector representing p and likewise with $\boldsymbol{q}$. The system matrix $\boldsymbol{\mathcal{L}}_h$ is shown in (3.44).

Boundary conditions are assigned by the numerical flux. The choice of upwind flux determines that the only influence of the Dirichlet condition at $x = 1$ will be on p on element K. Let $\mathbf{0}$ be a zero column vector of size N_p and let $\boldsymbol{e}_n$ be a vector of equal size with a 1 at position n and 0 elsewhere. We pre-multiply

with the Jacobian contribution $h^k/2$ and the inverse mass matrix and obtain the boundary matrix

$$\boldsymbol{B}_h^{\mathsf{T}} = \Big[\mathbf{0}, \ldots, \mathbf{0}, \frac{h^K}{2}(\boldsymbol{M}^K)^{-1}\boldsymbol{e}_{N_p} \,\Big|\, \mathbf{0}, \ldots, \mathbf{0}\Big]$$

where $h^K = (x_{\mathrm{R}}^K - x_{\mathrm{L}}^K)$ is the size of element D^K and the vertical bar indicates the center of the vector (and thus dividing the $\boldsymbol{p}$ and $\boldsymbol{q}$ parts). The pair $\{\boldsymbol{\mathcal{L}}_h, \boldsymbol{B}_h\}$ is found controllable by Theorem 4.2.

In the previous example about L-FEM, we used a simple first order finite difference approximation for the approximation of the normal derivative which came naturally from the underlying linear FEM basis. Equivalently, it comes natural for the DG semi-discretization to use the local polynomial basis for approximation of the normal derivative. The calculation of the local derivative $\partial y_h^k/\partial x$ on element D^k is straightforward with the differentiation matrix $\boldsymbol{D}_r$, (3.40), and the N_p nodal values of the derivative become

$$\frac{2}{h^k}\boldsymbol{D}_r\boldsymbol{y}^k,$$

where $\boldsymbol{y}^k$ contains the nodal values of y_h on element D^k and the Jacobian scaling with the element size $h^k = (x_{\mathrm{R}}^k - x_{\mathrm{L}}^k)$ concerns mapping to the reference element. The normal derivative is found by evaluating the derived polynomial of element D^K at the right endpoint, $x = x_{\mathrm{R}}^K = 1$.

$$\frac{\partial y}{\partial n}\Big|_{\Gamma_0} \approx \frac{2}{h^K}\boldsymbol{D}_r(N_p, :)\boldsymbol{y}^K \tag{4.22}$$

where $\boldsymbol{D}_r(N_p, :)$ denotes the bottom row (number N_p) of the differentiation matrix $\boldsymbol{D}_r$. The bottom row is associated with the right endpoint of the reference element. By recalling that $y = \sqrt{2}/2(-p+q)$, we can now form the output matrix $\boldsymbol{C}_h$ belonging to the generic system (4.6)

$$\boldsymbol{C}_h = \Big[\mathbf{0}, \ldots, \mathbf{0}, -\frac{\sqrt{2}}{h^K}\boldsymbol{D}_r(N_p, :) \,\Big|\, \mathbf{0}, \ldots, \mathbf{0}, \frac{\sqrt{2}}{h^K}\boldsymbol{D}_r(N_p, :)\Big]. \tag{4.23}$$

It remains now to specify discrete inner products and norms reflecting the nature of the energy spaces $\mathcal{E}$ and $\mathcal{E}^\star$. Let u and v be functions in $L^2(\Omega)$ and $\boldsymbol{u}$ and $\boldsymbol{v}$ their nodal DG approximation vectors each consisting of the K local nodal vectors $\boldsymbol{u}^k$ and $\boldsymbol{v}^k$ of size N_p. The approximate L^2-inner product $\langle\cdot,\cdot\rangle_0$ becomes

$$\langle \boldsymbol{u}, \boldsymbol{v}\rangle_0 := \sum_{k=1}^{K} (\boldsymbol{u}^k)^{\mathsf{T}}\boldsymbol{M}^k\boldsymbol{v}^k \tag{4.24}$$

due to the definition of the local mass matrix (3.31). Equivalently, for $u, v \in H_0^1(\Omega)$ we get the approximate H^1-inner product $\langle\cdot,\cdot\rangle_1$

$$\langle \boldsymbol{u}, \boldsymbol{v}\rangle_1 := \sum_{k=1}^{K} (\frac{2}{h^k}\boldsymbol{D}_r\boldsymbol{u}^k)^{\mathsf{T}}\boldsymbol{M}^k(\frac{2}{h^k}\boldsymbol{D}_r\boldsymbol{v}^k). \tag{4.25}$$

We will use the norms induced by the inner products. Furthermore, we will use

$$\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{-1,1} := \sum_{k=1}^{K} (\boldsymbol{u}^k)^{\mathsf{T}}\boldsymbol{M}^k\boldsymbol{v}^k \tag{4.26}$$

for the approximation of the duality product $\langle u, v\rangle_{H^{-1},H_0^1}$ as with L-FEM. The norms for $\mathcal{X}$ and $\mathcal{X}^\star$ and the duality product between them follow by (4.3), (4.4), and (4.5).

4.1.3 Discrete HUM

To make the problem fully discrete, we now introduce $M-1$ time steps from 0 to T with uniform step size Δt such that $t_m = m\Delta t$ for $m = 0, 1, \ldots, M-1$. The discretization is necessary for obtaining numerical solutions to the ODE systems (4.8) and (4.11). We replace the continuous boundary space $\mathcal{B}$ with the time discrete $\mathcal{T} = \mathbb{R}^M$ with the inner product

$$\langle \boldsymbol{p}, \boldsymbol{q}\rangle_{\mathcal{T}} = \Delta t \boldsymbol{p}\boldsymbol{q}^{\mathsf{T}} \tag{4.27}$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ are row vectors of size $1 \times M$.

We now introduce a collection of discrete operators for a choice of time integration scheme (see Section 3.4). They all map between finite dimensional vector spaces and are for this reason in essence matrices. We wish, however, to distinguish between applying the operator (matrix-vector multiplication) and explicitly constructing the underlying matrix and therefore keep this distinction. We let capital letters denote the discrete operators and *bold* capital letters their matrix representations.

Definition 4.6 (discrete observation). Given the initial data $[\boldsymbol{w}(0), \boldsymbol{w}'(0)]^{\mathsf{T}} \in \mathcal{X}$ for (4.8) we define the *discrete observation operator*

$$P\colon \mathcal{X} \to \mathcal{T} \quad \text{defined by} \quad P\colon \begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix} \mapsto \boldsymbol{p}, \tag{4.28}$$

where $\boldsymbol{p} = [p(0), \ldots, p(M\Delta t)]$ with $p(m\Delta t) = \boldsymbol{C}_h \boldsymbol{W}(m\Delta t)$. The vector $\boldsymbol{W}(m\Delta t)$ is the numerical solution to (4.8) at time step m. □

Definition 4.7 (discrete reconstruction). Given $\boldsymbol{k} \in \mathcal{T}$ we define the discrete reconstruction operator

$$R\colon \mathcal{T} \to \mathcal{X}^\star \quad \text{defined by} \quad R\colon \boldsymbol{k} \mapsto \begin{bmatrix} \boldsymbol{z}^1 \\ -\boldsymbol{z}^0 \end{bmatrix}, \tag{4.29}$$

where $[\boldsymbol{z}^1, -\boldsymbol{z}^0]^{\mathsf{T}}$ is the state of (4.11) at $t_0 = 0$ after its time integration from T to 0. □

With these fully discrete operators at hand, we now define our discrete approximation to Λ

$$L\colon \mathcal{X} \to \mathcal{X}^\star \quad \text{defined by} \quad L := RP. \tag{4.30}$$

Equivalent to (4.14) we introduce the discrete HUM operator equation

$$L \begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}^1 \\ -\boldsymbol{u}^0 \end{bmatrix}. \tag{4.31}$$

Its solution $[\bar{\boldsymbol{w}}^0, \bar{\boldsymbol{w}}^1]^{\mathsf{T}}$, if it exists, provides the sought approximate control by

$$\boldsymbol{k}^{\mathsf{T}} = P \begin{bmatrix} \bar{\boldsymbol{w}}^0 \\ \bar{\boldsymbol{w}}^1 \end{bmatrix}.$$

The continuous Λ is dependent on T only (for fixed Γ_0). Its approximation L depends, however, also on:

Semi-discretization scheme $\boldsymbol{\mathcal{L}}_h$ (element size h, and approximation order p).

- Approximation of normal derivative $\boldsymbol{C}_h$.
- Assigning the Dirichlet boundary condition with $\boldsymbol{B}_h$.

Time integration method and time step size $\triangle t$.

The discrete HUM equation (4.31) can be solved directly by constructing L as a matrix, which we shall see in Section 4.2, or iteratively—matrix-free—as we shall do in Section 4.3.

4.2 Construction of the L matrix

The operator L may be used to obtain a matrix representation $\boldsymbol{L}$ in the same way as we did for Λ previously. Section 2.4 described two different ways of constructing Λ as a matrix: direct assembly (2.25) and inner-product assembly (2.26). We will apply the same practices for $\boldsymbol{L}$.

Let in the following $\boldsymbol{e}_j$ for $1 \leq j \leq 2N$ be vectors constituting a basis in $\mathcal{X}$. The canonical basis is the simplest choice, but it might not be a sensible one. Later in this chapter, we will use a discrete basis related to the eigenfunctions of the continuous HUM operator.

Direct assembly

We construct $\boldsymbol{L}$ by direct assembly by computing

$$\boldsymbol{L}_{ij} = \langle L\boldsymbol{e}_j, \boldsymbol{e}_i \rangle_{\mathcal{X}^\star,\mathcal{X}}, \qquad i,j = 1,\ldots,2N, \tag{4.32}$$

where L is defined by (4.30) and $\langle \cdot,\cdot \rangle_{\mathcal{X}^\star,\mathcal{X}}$ is the approximation (4.5) to duality product $\langle \cdot,\cdot \rangle_{\mathcal{E}^\star,\mathcal{E}}$ between $\mathcal{E}$ and $\mathcal{E}^\star$.

Inner-product assembly

The inner-product assembly of $\boldsymbol{L}$ is done by

$$\boldsymbol{L}_{ij} = \langle P\boldsymbol{e}_j, P\boldsymbol{e}_i \rangle_{\mathcal{T}}, \qquad i,j = 1,\ldots,2N, \tag{4.33}$$

where P is defined by (4.28) and $\langle \cdot,\cdot \rangle_{\mathcal{T}}$ is the approximation (4.27) to the inner product $\langle \cdot,\cdot \rangle_{\mathcal{B}}$.

It is worth noticing that inner product assembly, in contrast to direct assembly, do not require reconstruction; hence only half the computations are needed.

4.2.1 Matrix assembling—procedures and details

We wish to determine the elements of $\boldsymbol{L}$ by direct assembly (4.32) and need therefore the action of the operator L defined by (4.30). The output of $L[\boldsymbol{w}^0, \boldsymbol{w}^1]^\mathsf{T}$ is denoted $[\boldsymbol{z}^1, -\boldsymbol{z}^0]^\mathsf{T}$.

Let now, and in the rest of this chapter, $\boldsymbol{e}_j$ denote a basis vector in $\mathbb{R}^N$ (instead of in $\mathcal{X}$); the set $\{\boldsymbol{e}_j\}_{j\leq N}$ constitute a basis for $\mathbb{R}^N$. We apply L to the basis vectors $\boldsymbol{e}_j$ in the position of first $\boldsymbol{w}^0$ and then $\boldsymbol{w}^1$, that is,

$$\begin{bmatrix} \boldsymbol{z}_j^{1,0} \\ -\boldsymbol{z}_j^{0,0} \end{bmatrix} = L \begin{bmatrix} \boldsymbol{e}_j^0 \\ \boldsymbol{0} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \boldsymbol{z}_j^{1,1} \\ -\boldsymbol{z}_j^{0,1} \end{bmatrix} = L \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{e}_j^1 \end{bmatrix}, \quad j = 1,\ldots,N$$

where the superscript 0 denotes that $\boldsymbol{z}_j^{i,0}$ is created from $\boldsymbol{w}^0$ and likewise with the 1 on $\boldsymbol{z}_j^{i,1}$. The corresponding 0 or 1 superscript on $\boldsymbol{e}_j$ is only to indicate the position of $\boldsymbol{e}_j$; we use the same basis for $\boldsymbol{w}^0$ and $\boldsymbol{w}^1$.

It is convenient to divide $\boldsymbol{L}$ into four $N \times N$ sub-matrices $\boldsymbol{L}^i$, $i = 1, 2, 3, 4$ enumerated

$$\boldsymbol{L} = \begin{bmatrix} \boldsymbol{L}^1 & \boldsymbol{L}^2 \\ \boldsymbol{L}^3 & \boldsymbol{L}^4 \end{bmatrix}.$$

Each sub-matrix describes the effect of L applied to one initial data vector for the adjoint system $\boldsymbol{w}^0$ or $\boldsymbol{w}^1$ on either $\boldsymbol{z}^0$ or $\boldsymbol{z}^1$ of the control system. The elements of the sub-matrices are determined for $i, j = 1, \ldots, N$ by

$$\left.\begin{aligned} \boldsymbol{L}_{ij}^1 &= \langle \boldsymbol{z}_j^{1,0}, \boldsymbol{e}_i^0 \rangle_{-1,1} & \boldsymbol{L}_{ij}^2 &= \langle \boldsymbol{z}_j^{1,1}, \boldsymbol{e}_i^0 \rangle_{-1,1} \\ \boldsymbol{L}_{ij}^3 &= \langle \boldsymbol{z}_j^{0,0}, \boldsymbol{e}_i^1 \rangle_0 & \boldsymbol{L}_{ij}^4 &= \langle \boldsymbol{z}_j^{0,1}, \boldsymbol{e}_i^1 \rangle_0, \end{aligned}\right\} \tag{4.34}$$

where the discrete inner products depend on the semi-discretization scheme. The duality product $\langle \cdot, \cdot \rangle_{\mathcal{X}^\star, \mathcal{X}}$ has here been split in $\langle \cdot, \cdot \rangle_{-1,1}$ and $\langle \cdot, \cdot \rangle_0$ in accordance with (4.5). In the case of L-FEM, $\langle \cdot, \cdot \rangle_{-1,1}$ and $\langle \cdot, \cdot \rangle_0$ are defined by (4.21) and (4.18) while by (4.26) and (4.18) for DG-FEM.

If we instead approach the construction of $\boldsymbol{L}$ by the inner product assembly technique (4.33), we only need the discrete observation (4.28). Consider the row vectors $\boldsymbol{p}_j^0$ and $\boldsymbol{p}_j^1$ produced by discrete observation

$$(\boldsymbol{p}_j^0)^\mathsf{T} = P \begin{bmatrix} \boldsymbol{e}_j^0 \\ \boldsymbol{0} \end{bmatrix} \quad \text{and} \quad (\boldsymbol{p}_j^1)^\mathsf{T} = P \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{e}_j^1 \end{bmatrix}, \tag{4.35}$$

for $j = 1, \ldots, N$. The superscript 0 on $\boldsymbol{p}_j^0$ indicates the observation of the initial data $\boldsymbol{w}^0$ and equivalently with 1 on $\boldsymbol{p}_j^1$. The sub-matrices of $\boldsymbol{L}$ are now assembled according to (4.33) by the $i, j = 1, \ldots, N$ operations

$$\left.\begin{aligned} \boldsymbol{L}_{ij}^1 &= \langle \boldsymbol{p}_j^0, \boldsymbol{p}_i^0 \rangle_{\mathcal{T}} & \boldsymbol{L}_{ij}^2 &= \langle \boldsymbol{p}_j^1, \boldsymbol{p}_i^0 \rangle_{\mathcal{T}} \\ \boldsymbol{L}_{ij}^3 &= \langle \boldsymbol{p}_j^0, \boldsymbol{p}_i^1 \rangle_{\mathcal{T}} & \boldsymbol{L}_{ij}^4 &= \langle \boldsymbol{p}_j^1, \boldsymbol{p}_i^1 \rangle_{\mathcal{T}}. \end{aligned}\right\} \tag{4.36}$$

where the time discrete L^2-inner product $\langle \cdot, \cdot \rangle_{\mathcal{T}}$ is defined by (4.27).

After computing the elements of $\boldsymbol{L}$ by (4.34) or (4.36), the specific set of initial data $[\bar{\boldsymbol{w}}^0, \bar{\boldsymbol{w}}^1]^\mathsf{T}$ of the adjoint system can be found by solving the $2N \times 2N$ linear system

$$\begin{bmatrix} \boldsymbol{L}^1 & \boldsymbol{L}^2 \\ \boldsymbol{L}^3 & \boldsymbol{L}^4 \end{bmatrix} \begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}^1 \\ -\boldsymbol{u}^0 \end{bmatrix}. \tag{4.37}$$

The approximate control $\boldsymbol{k}$ is then computed by applying P to the solution

$$\boldsymbol{k}^\mathsf{T} = P \begin{bmatrix} \bar{\boldsymbol{w}}^0 \\ \bar{\boldsymbol{w}}^1 \end{bmatrix}. \tag{4.38}$$

The vector $\boldsymbol{k}$ is an approximation of the continuous control κ to which it should converge for $h \to 0$. We will examine the convergence later in this chapter.

Matrices P and R

Let us take the $\mathcal{T}$-inner product between observation vector $\boldsymbol{p}_j^0$ or $\boldsymbol{p}_j^1$ (4.35) and a basis vector $\boldsymbol{b}_i$ in $\mathcal{T}$. This constructs the matrix $\boldsymbol{P}$

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}^0 & \boldsymbol{P}^1 \end{bmatrix}, \quad \boldsymbol{P}_{ij}^0 = \langle \boldsymbol{p}_j^0, \boldsymbol{b}_i \rangle_{\mathcal{T}}, \quad \boldsymbol{P}_{ij}^1 = \langle \boldsymbol{p}_j^1, \boldsymbol{b}_i \rangle_{\mathcal{T}}, \tag{4.39}$$

which has $M \times 2N$ elements. The inner product assembly corresponds to taking $\Delta t \boldsymbol{P}^{\mathsf{T}} \boldsymbol{P}$ as $\boldsymbol{L}$.

$\boldsymbol{P}$ tells us almost everything that we need to know about a discretization in regards to HUM-control and about $\boldsymbol{L}$, too, as $\boldsymbol{P}$—in essence—is the square root of $\boldsymbol{L}$. $\boldsymbol{P}$ even comes as a natural by-product when constructing $\boldsymbol{L}$ by either assembly technique as it simply consists of the discrete observation of the basis vectors (4.35). The observation of the sine basis is a step on the way to $\boldsymbol{L}$ when using direct assembly.

With the matrix $\boldsymbol{P}$, we can determine the approximate control $\boldsymbol{k}$ from the solution $[\bar{\boldsymbol{w}}^0, \bar{\boldsymbol{w}}^1]^{\mathsf{T}}$ of (4.37) by a matrix-vector multiplication

$$\boldsymbol{k}^{\mathsf{T}} = \begin{bmatrix} \boldsymbol{P}^0 & \boldsymbol{P}^1 \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{w}}^0 \\ \bar{\boldsymbol{w}}^1 \end{bmatrix},$$

instead of solving the wave equation as in (4.38).

The discrete reconstruction operator R can be used to construct a corresponding *reconstruction matrix* $\boldsymbol{R}$ similar to what we just did with the discrete observation operator. It is not hereby said that this is a practical thing to do. The analysis, however, provide some useful insight. Let us introduce the construction of $\boldsymbol{R}$ as the discrete equivalent of (2.30)

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}^1 \\ \boldsymbol{R}^0 \end{bmatrix}, \quad \boldsymbol{R}^0_{ij} = \langle \boldsymbol{z}^0_j, \boldsymbol{e}^1_i \rangle_0, \quad \boldsymbol{R}^1_{ij} = \langle \boldsymbol{z}^1_j, \boldsymbol{e}^0_i \rangle_{-1,1}, \tag{4.40}$$

where $i, j = 1, \ldots, N$ and $\boldsymbol{z}^1_j$ and $\boldsymbol{z}^0_j$ are the output vectors of the reconstruction

$$\begin{bmatrix} \boldsymbol{z}^1_j \\ -\boldsymbol{z}^0_j \end{bmatrix} = R\boldsymbol{b}_j, \qquad j = 1, \ldots, N,$$

where R is defined by (4.29). If we use the observation vector $\boldsymbol{p}^0_j$ or $\boldsymbol{p}^1_j$ in place of $\boldsymbol{b}_j$, we recover the directly assembled $\boldsymbol{L}$ which is of course no surprise since $L = RP$.

We will now shortly depart from the construction of $\boldsymbol{L}$ and look at the minimization approach to HUM before proceeding with the remaining assembling details.

4.2.2 Discrete HUM by minimization

In Section 2.5, we showed how HUM, as an alternative to the operator approach, could be formulated as a minimization problem for the functional $\mathcal{J}$ defined in (2.33). Let us replace $\Phi(\varphi^0, \varphi^1)$ with the approximation $P[\boldsymbol{w}^0, \boldsymbol{w}^1]^{\mathsf{T}}$ and the continuous spaces $\mathcal{E}, \mathcal{E}^\star$ and $\mathcal{B}$ with the discrete equivalents in (2.33) to obtain the discretized functional

$$J_h\Big(\begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix} \Big) = \frac{1}{2} \left\| P \begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix} \right\|_{\mathcal{T}} - \langle \begin{bmatrix} \boldsymbol{u}^1 \\ -\boldsymbol{u}^0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix} \rangle_{\mathcal{X}^\star, \mathcal{X}}. \tag{4.41}$$

If we furthermore apply P as the matrix (4.39), use the discrete norm (4.27) and the discrete duality product (4.5) with (4.21), we get

$$J_h\Big(\begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix} \Big) = \frac{1}{2} \Delta t \Big(\boldsymbol{P} \begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix} \Big)^{\mathsf{T}} \Big(\boldsymbol{P} \begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix} \Big) - \begin{bmatrix} \boldsymbol{u}^1 \\ -\boldsymbol{u}^0 \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \boldsymbol{M} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix}$$

Let us now, for brevity, introduce the variables

$$\boldsymbol{w} := \begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix} \qquad \mathcal{A} := \Delta t \boldsymbol{P}^\mathsf{T}\boldsymbol{P}, \qquad \boldsymbol{b} := \begin{bmatrix} \boldsymbol{M} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}^1 \\ -\boldsymbol{u}^0 \end{bmatrix}.$$

This reduces the functional to

$$J_h(\boldsymbol{w}) = \tfrac{1}{2}\boldsymbol{w}^\mathsf{T}\mathcal{A}\boldsymbol{w} - \boldsymbol{w}^\mathsf{T}\boldsymbol{b}$$

where we easily identify a quadratic and a linear term. The matrix $\mathcal{A}$ is symmetric and by the spectral decomposition

$$\mathcal{A} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\mathsf{T},$$

we de-couple the above expression and have

$$\begin{aligned} J_h(\boldsymbol{w}) &= \tfrac{1}{2}\boldsymbol{w}^\mathsf{T}\boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\mathsf{T}\boldsymbol{w} - \boldsymbol{w}^\mathsf{T}\boldsymbol{b} \\ &= \tfrac{1}{2}\widehat{\boldsymbol{w}}^\mathsf{T}\boldsymbol{D}\widehat{\boldsymbol{w}} - \widehat{\boldsymbol{w}}^\mathsf{T}\widehat{\boldsymbol{b}}. \end{aligned}$$

The new vector $\widehat{\boldsymbol{w}}$ is the transformation $\boldsymbol{Q}^\mathsf{T}\boldsymbol{w}$ and $\widehat{\boldsymbol{b}} = \boldsymbol{Q}^\mathsf{T}\boldsymbol{b}$. The de-coupled system consists of $2N$ scalar quadratic expressions of the type $\frac{1}{2}ax^2 - bx$ where $a, b \in \mathbb{R}, a > 0$ for which the minimum is simply $x_{min} = a^{-1}b$. Equivalently, we find the minimum for J_h to be

$$\widehat{\boldsymbol{w}}_{\min} = \boldsymbol{D}^{-1}\widehat{\boldsymbol{b}},$$

which we can transform back to

$$\begin{aligned} \boldsymbol{w}_{\min} &= \boldsymbol{Q}\boldsymbol{D}^{-1}\boldsymbol{Q}^\mathsf{T}\boldsymbol{b} \\ &= \mathcal{A}^{-1}\boldsymbol{b}. \end{aligned}$$

This is not surprising, though, as it is standard practice to rewrite a linear problem $\mathcal{A}\boldsymbol{w} = \boldsymbol{b}$ as the minimization of a quadratic functional $\frac{1}{2}\boldsymbol{w}^\mathsf{T}\mathcal{A}\boldsymbol{w} - \boldsymbol{w}^\mathsf{T}\boldsymbol{b}$. It is included here to show that the minimization of the discrete functional corresponds to using the inner-product assembly technique (4.33) since $\mathcal{A} = \Delta t \boldsymbol{P}^\mathsf{T}\boldsymbol{P}$. Constructing $\boldsymbol{P}$ as a matrix before minimizing the functional (4.41) is, however, not a very practical thing to do. We refer to Section 4.3 for a practical iterative approach to the problem.

4.2.3 The sine basis

We now return to the assembling of matrix $\boldsymbol{L}$ and in particular to the choice of a suitable basis. Remark 2.22 gave for $T = 2$ an analytic expression for the infinite observation matrices $\boldsymbol{\Phi}^0$ and $\boldsymbol{\Phi}^1$ in orthonormal sine basis $\left\{e_j^{\mathrm{s}}(\cdot)\right\}_{j\in\mathbb{N}}$ where

$$e_j^{\mathrm{s}}(x) = \sqrt{2}\sin(j\pi x), \qquad x \in \Omega. \tag{4.42}$$

We shall consider truncated versions of $\boldsymbol{\Phi}^0$ and $\boldsymbol{\Phi}^1$ and sample in the time domain ($t_m = m\Delta t$) to obtain the "exact" discrete observation of e_j^{s} for $j = 1, \ldots, N$

$$\boldsymbol{p}_j^0 = [p_j^0(0\Delta t), \ldots, p_j^0((M-1)\Delta t)], \quad p_j^0(t) = (-1)^j\sqrt{2}j\pi\cos(j\pi t) \tag{4.43a}$$

$$\boldsymbol{p}_j^1 = [p_j^1(0\Delta t), \ldots, p_j^1((M-1)\Delta t)], \quad p_j^1(t) = (-1)^j\sqrt{2}\sin(j\pi t). \tag{4.43b}$$

Collecting these vectors as columns in a matrix would give the observation matrix $\boldsymbol{P}$ in canonical basis. Alternatively, we can take discrete versions of the orthonormal bases $b_j^0(t) = \sqrt{2/T}\cos(j\pi t)$ and $b_j^1(t) = \sqrt{2/T}\sin(j\pi t)$ for $j = 1, \ldots, N$ when $T = 2$

$$\begin{aligned}\boldsymbol{b}_j^0 &= [b_j^0(0\Delta t), \ldots, b_j^0((M-1)\Delta t)],\\ \boldsymbol{b}_j^1 &= [b_j^1(0\Delta t), \ldots, b_j^1((M-1)\Delta t)].\end{aligned}$$

Computing the inner products (4.39) with this basis clearly results in diagonal matrices

$$\begin{aligned}\boldsymbol{P}_{ij}^0 &\equiv \langle \boldsymbol{p}_j^0, \boldsymbol{b}_i^0\rangle_T = (-1)^j\sqrt{T}j\pi\delta_{ij},\\ \boldsymbol{P}_{ij}^1 &\equiv \langle \boldsymbol{p}_j^1, \boldsymbol{b}_i^1\rangle_T = (-1)^j\sqrt{T}\delta_{ij},\end{aligned}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, N$. The diagonal structure is a manifestation of the exact dispersion relation—waves at all wavelengths travel at speed one. When we in a minute turn to *approximations* of $\boldsymbol{P}$, we will see that the matrix is shaped by the numerical dispersion relation for the related scheme.

Let us consider the construction of $\boldsymbol{L}$ by inner product assembly for each of its sub-matrices (4.36) when $T = 2$. Sub-matrices $\boldsymbol{L}^2$ and $\boldsymbol{L}^3$ will be zero (inner product between $\boldsymbol{p}_j^0$ and $\boldsymbol{p}_j^1$). Sub-matrix $\boldsymbol{L}^1 = \Delta t(\boldsymbol{P}^0)^{\mathsf{T}}\boldsymbol{P}^0$ will have the elements $\boldsymbol{L}_{ij}^1 = T(j\pi)^2\delta_{ij}$ for $i, j = 1, \ldots, N$, and sub-matrix $\boldsymbol{L}^4 = \Delta t(\boldsymbol{P}^1)^{\mathsf{T}}\boldsymbol{P}^1$ will have the elements $\boldsymbol{L}_{ij}^4 = T\delta_{ij}$ for $i, j = 1, \ldots, N$.

Also reconstruction is particularly simple when $T = 2$. The periodicity of $\cos(j\pi t)$ and $\sin(j\pi t)$ allows easy reversion of the wave equation and the "exact" reconstruction becomes

$$R_{\text{ex}}(\sqrt{2}\cos(j\pi t)) = \begin{bmatrix} (-1)^j T j\pi e_j^{\text{s}} \\ 0 \end{bmatrix},$$

and

$$R_{\text{ex}}(\sqrt{2}\sin(j\pi t)) = \begin{bmatrix} 0 \\ (-1)^{j+1} T e_j^{\text{s}} \end{bmatrix},$$

for $j = 1, \ldots, N$. The direct assembly would quite clearly result in the same diagonal matrix $\boldsymbol{L}$ as above.

The diagonal structure of the exact $\boldsymbol{L}$ tells us that $T = 2n$ is a very special case where the modes are not mixed. The sub-matrix $\boldsymbol{L}^1$ act in this case like the Laplacian. The discrete Laplacian belonging to a particular semi-discretization is therefore a natural choice for pre-conditioning for iterative solution of the operator equation (4.31). We will return to this issue in Section 4.3.

The sinusoids are closely connected with the Laplacian and the wave equation. It is well-known that the sine functions $\{e_j^{\text{s}}\}_{j\in\mathbb{N}}$ constitute an orthonormal basis for the Laplacian in 1-d. The solutions to the (homogeneous) wave equation (3.1) is "carried" by sine and cosine functions as can be seen from the Fourier or semi-group solution (or the above analysis). Not only this speaks for the use of sinusoids as basis for the problem—but also the separation of waves into low and high-frequency components and the close link to the dispersion relation in this case are strong arguments.

On the other hand, are sinusoids with short wavelengths compared to the grid size not easily approximated by polynomials which both semi-discretizations use

as shape functions. The pros and cons of the use of sinusoids as basis for the construction of $\boldsymbol{L}$ will be discussed further in the end of this chapter.

4.2.4 Constructing L with L-FEM

Let us take a closer look at the construction of $\boldsymbol{L}$ with L-FEM on an equidistant grid with spacing h (see L-FEM on page 57). When nothing else is mentioned in the following, we use trapezoidal time integration (see Section 3.4) with the Courant number $\mu = 0.6$ for the time step $\Delta t = \mu h$. Recall that $N = 1/h - 1$ is the number of inner grid points which also is the number of basis functions $\boldsymbol{e}_j$.

We need a basis $\{\boldsymbol{e}_j\}_{j\leq N}$ in order to construct the matrix $\boldsymbol{L}$. The simplest choice is to use the canonical basis

$$\boldsymbol{e}_j^{\mathrm{e}} = [e_j^{\mathrm{e}}(x_1), \ldots, e_j^{\mathrm{e}}(x_N)]^{\mathsf{T}}, \qquad e_j^{\mathrm{e}}(x_i) = \delta_{ij}, \qquad i,j = 1,\ldots,N,$$

where $x_i = i\,h$ and the superscript e denotes 'canonical'. In the eyes of the L-FEM semi-discretization, this basis is its own hat basis consisting of the functions ψ_n^{L} defined in (3.17). The linear spline representation of the basis vector $\boldsymbol{e}_j^{\mathrm{e}}$ is therefore simply

$$e_j^{\mathrm{Le}}(x) = \psi_j^{\mathrm{L}}(x), \qquad x \in \Omega, \qquad j = 1,\ldots,N,$$

with superscript L indicating the use of L-FEM. We will, however, as explained in the previous section, focus on the sampled sine basis which we mark by the superscript s

$$\boldsymbol{e}_j^{\mathrm{s}} = [e_j^{\mathrm{s}}(x_1), \ldots, e_j^{\mathrm{s}}(x_N)]^{\mathsf{T}}, \qquad e_j^{\mathrm{s}}(x_i) = \sqrt{2}\sin(j\pi x_i), \quad i,j = 1,\ldots,N.$$

By interpolation, we have the piecewise linear function

$$e_j^{\mathrm{Ls}}(x) = \sum_{i=1}^{N} e_j^{\mathrm{s}}(x_i)\psi_i^{\mathrm{L}}(x), \qquad x \in \Omega, \qquad i,j = 1,\ldots,N, \tag{4.44}$$

which is the L-FEM representation of the sine function e_j^{s}.

Using the sine basis makes it necessary to have a procedure for translating between coefficients and nodal values which can be done in several ways. We choose the continuous sine basis as our starting point. Let us consider the sine expansion of a function $f \in L^2(\Omega)$, which is zero in both endpoints,

$$f(x) = \sum_{k=1}^{\infty} \widehat{f}_k\, e_k^{\mathrm{s}}(x), \qquad x \in \Omega.$$

No cosines are needed in the expansions since f is zero at the endpoints. By the orthonormality of e_k^{s}, we can compute the coefficients $\widehat{f}_k$ by the L^2-inner product

$$\widehat{f}_k = \langle f, e_k^{\mathrm{s}} \rangle_{L^2(\Omega)}, \qquad k = 1,2,\ldots. \tag{4.45}$$

We wish to approximate the N first coefficients by

$$\widehat{f}_k^{\mathrm{L}} = \langle \boldsymbol{f}, \boldsymbol{e}_k^{\mathrm{s}} \rangle_0, \qquad k = 1,\ldots,N, \tag{4.46}$$

where the inner product is defined by (4.18) and $\boldsymbol{f}$ consists of N sampled values $f(x_i), i = 1, \ldots, N$ of f. Furthermore, we define the synthesis

$$f(x) \approx \sum_{k=1}^{N} \widehat{f}_k^{\mathrm{L}}\, e_k^{\mathrm{Ls}}(x), \qquad x \in \Omega,$$

which approximates f. Notice that we use index k for coefficients and index j when we refer to a specific basis function.

Before we proceed with the approximation of HUM, we shall briefly consider the difference of using the canonical basis $\boldsymbol{e}_j^{\mathrm{e}}$ and the sine basis $\boldsymbol{e}_j^{\mathrm{s}}$. Both bases can be used for the construction of $\boldsymbol{L}$ and the total result would be the same due to the linearity of the problem and the perfect transformation between the canonical and sine basis—each canonical basis vector $\boldsymbol{e}_j^{\mathrm{e}}$ may be seen as a linear combination of the vectors $\boldsymbol{e}_j^{\mathrm{s}}$, $j = 1, \ldots, N$, and vice versa.

Figure 4.2 shows the coefficients $\widehat{f}_k$ and $\widehat{f}_k^{\mathrm{L}}$ for $k = 1, \ldots, N$ of three hat functions (canonical basis vectors) $e_j^{\mathrm{Le}}, j = 2, 6$ and 27. Notice that almost all

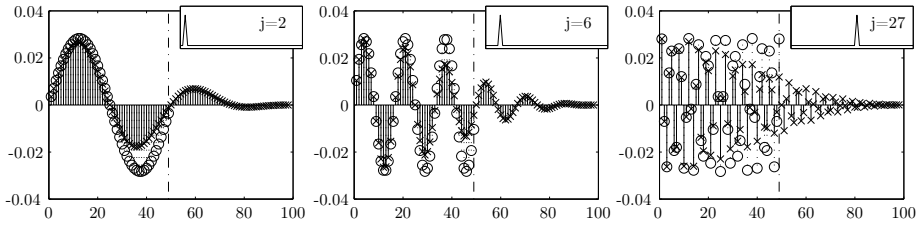


Figure 4.2: The spectrum of three hat basis functions e_j^{Le}, $j = 2, 6$ and 27. The solid stems with (×)-marks show the coefficients $\langle e_j^{\mathrm{Le}}, e_k^{\mathrm{s}} \rangle_{L^2(\Omega)}$ for $k = 1, \ldots, 100$. The dotted stems with (o)-marks show the L-FEM coefficients $\langle \boldsymbol{e}_j^{\mathrm{Le}}, \boldsymbol{e}_k^{\mathrm{Ls}} \rangle_0$ for $k = 1, \ldots, N$. The vertical dashed-dotted line marks N—the number of basis elements in the L-FEM representation. The small inlets show the position of the relevant hat function on Ω.

modes are excited for all three canonical basis function, and notice also the aliasing effect for high wavenumbers. Whereas the sine basis separate well and poorly resolved parts of a function, the canonical basis mixes all that valuable information together. *If* waves at all wavelengths, however, travelled at the correct speed $c = 1$, the compact, *localized* nature of the canonical basis would be kept, thus leading to a sparse observation matrix. But this is, due to numerical dispersion, not the case. We conclude that we keep most information and characteristic behavior by using the sine basis (see also the discussion in Section 4.2.3).

Representation of sinusoids by linear splines

How well we can approximate a sine wave by linear splines depends on the ratio between the wavenumber j and the number of grid points N. For small wavenumbers j compared to N the approximation error will be negligible. Relative large wavenumbers with only a few grid points per wavelength will, on the contrary, lead to significant errors. Consider the Fourier coefficients $\widehat{f}_k$ for $k = 1, \ldots 2N$ of the linear splines e_j^{Ls} for $j = 1, \ldots, N$, shown in Figure 4.3. The left half of the figure contains the sinusoids that we can represent with $N = 49$ L-FEM sine basis functions. The right half shows the higher frequency

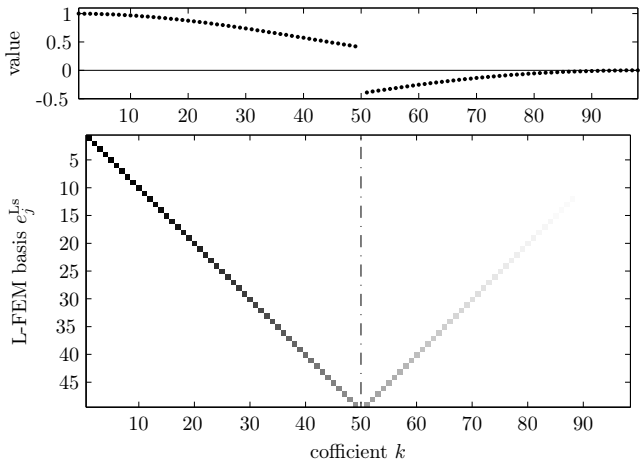


Figure 4.3: The bottom plot shows the spectrum of the N L-FEM sine representations e_j^{Ls}. The gray scaled element in index (j,k) displays the absolute value of the coefficient $\langle e_j^{\mathrm{Ls}}, e_k^{\mathrm{s}} \rangle_{L^2(\Omega)}$; black is 1, white is 0. The top plot shows the value of the corresponding coefficient (notice: only one per column). The vertical dashed-dotted line marks the number of basis vectors $N = 49$—the coefficients to the left are represented in the sampled sine basis, the part to right of the line represents the alias error.

alias corresponding to the modes $k > N$ of the hat basis shown in Figure 4.2. Notice that Figure 4.3 shows that any one sinusoid e_j^{s} with wavenumber $j \leq N$ is represented, unambiguously, by *only one* $e_j^{\mathrm{Ls}}, j \leq N$. This one-to-one correspondence, which is well-known, is a special feature of the equidistant grid. We mention it because this does not hold for the DG-FEM semi-discretization as we shall see in Section 4.2.5.

The linear spline e_j^{Ls} is not well approximated solely by e_j^{s} when j is large; it also has higher frequency components of which the lowest, $e_{(2N-j)}^{\mathrm{s}}$, is shown on the right hand side of Figure 4.3. This is essential for understanding the approximation error made when trying to approximate e_j^{s} by the piecewise linear e_j^{Ls}. This error as a function of j is shown in Figure 4.4 together with the L^2-norm of e_j^{Ls}. The approximation error is significant for large j, say, $j > \frac{N}{2}$. This

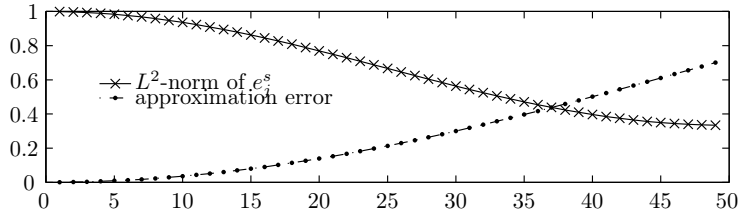


Figure 4.4: L^2-norm of the L-FEM ($N = 49$) sine basis e_j^{Ls} and the (normalized) L^2-norm of the approximation error $\left|e_j^{\mathrm{Ls}} - e_j^{\mathrm{s}}\right|$ both as functions of the index j.

error is important to keep in mind when we later compare our findings with an analytic result obtained from exact e_j^{s}.

Observation of the sine basis

Let us consider the discrete observation P of the L-FEM sine basis e_j^{Ls} for $j = 1, \ldots, N$ with the observation time $T = 2$. The discrete observation (4.28) amounts to (1) solving a wave equation, here with L-FEM semi-discretization (4.16), and (2) finding the approximate normal derivative by $\boldsymbol{C}_h$ defined in (4.17).

The discrete observation can be divided in P^0 observation and P^1 observation as seen in (4.35). We know from (4.43) that the exact P^0 observation of the j'th sinusoid is $p_j^0(m\Delta t) = (-1)^j\sqrt{2}j\pi\cos(j\pi m\Delta t)$ and the exact P^1 observation is $p_j^1(m\Delta t) = (-1)^j\sqrt{2}\sin(j\pi m\Delta t)$ with $m = 0, \ldots, M-1$ when $T = 2$.

We compute normalized temporal Fourier coefficients for $k = 1, \ldots, N$ of the P^0 observation of $\boldsymbol{e}_j^{\mathrm{s}}$ for $j = 1, \ldots, N$ by

$$\beta_{jk}^{\cos} = \left\langle P^0\boldsymbol{e}_j^{\mathrm{s}}, \boldsymbol{p}_k^0\right\rangle_{\mathcal{T}} \Big/ \left\|\boldsymbol{p}_k^0\right\|_{\mathcal{T}}^2, \tag{4.47a}$$

$$\beta_{jk}^{\sin} = \left\langle P^0\boldsymbol{e}_j^{\mathrm{s}}, k\pi\boldsymbol{p}_k^1\right\rangle_{\mathcal{T}} \Big/ \left\|\boldsymbol{p}_k^0\right\|_{\mathcal{T}}^2. \tag{4.47b}$$

Exact observation would give $\beta_{jk}^{\cos} = \delta_{jk}$ and $\beta_{jk}^{\sin} = 0$, but due to numerical dispersion and other numerical effects it is expected that the approximation $P^0\boldsymbol{e}_j^{\mathrm{s}}$ will have other coefficients than $\beta_{jj}^{\cos}$.

We shall first consider two examples: the P^0 observation of $\boldsymbol{e}_7^{\mathrm{s}}$ and that of $\boldsymbol{e}_{32}^{\mathrm{s}}$. Hereafter, we will consider the P^0 observation of all N basis functions collectively. The analysis of P^1 observation has been omitted; we remark only that it exhibits behavior very similar to P^0 observation.

Figure 4.5 shows the L-FEM approximated P^0 observation of the 7'th sinusoid, $\boldsymbol{e}_7^{\mathrm{s}}$, by its Fourier coefficients (4.47).

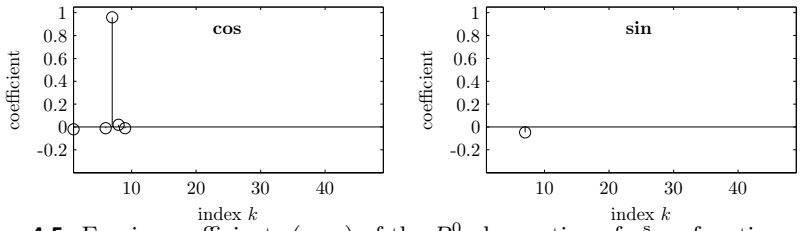


Figure 4.5: Fourier coefficients (4.47) of the P^0 observation of $\boldsymbol{e}_7^{\mathrm{s}}$ as function of the index k. The left plot shows the cosine coefficients (4.47a) and the right plot shows the sine coefficients (4.47b). Observation was made with L-FEM with $N = 49, T = 2, \mu = 0.6$.

The 7'th cosine coefficient comes out very strong, but we already see a beginning deterioration: cosine coefficients other than $j = 7$ are present (small though) and a single sine mode has also appeared.

The observation of the 32'nd mode can be seen in Figure 4.6. The theoretical values, $\beta_{32,k}^{\cos} = \delta_{32,k}$ and $\beta_{32,k}^{\sin} = 0$, are almost unrecognizable from Figure 4.6. This cannot all be explained from the error made from start when approximating the initial data e_j^{s} by linear splines; that is only a tiny piece of the explanation. We will return to the cause of this misfortune after a look at the big picture.

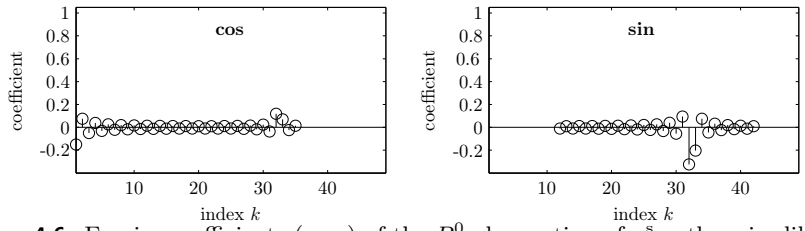


Figure 4.6: Fourier coefficients (4.47) of the P^0 observation of $\boldsymbol{e}^{\mathrm{s}}_{32}$ otherwise like Figure 4.5.

We depict the absolute values of the coefficients (4.47) for all N discrete P^0 observations, $P^0(\boldsymbol{e}^{\mathrm{s}}_j), j = 1, \ldots, N$ in Figure 4.7. The j'th row in the image

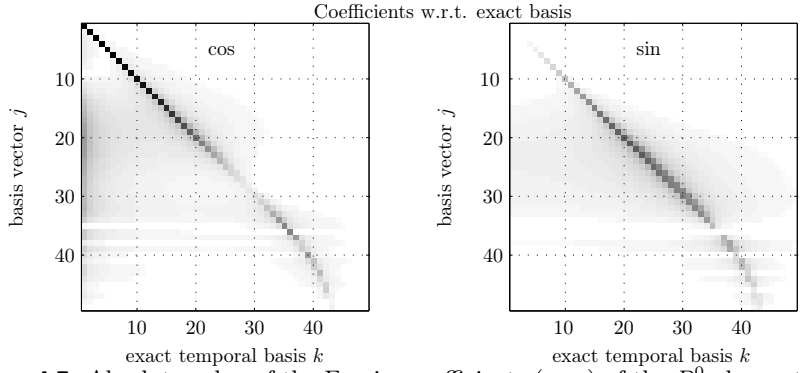


Figure 4.7: Absolute value of the Fourier coefficients (4.47) of the P^0 observation of $\boldsymbol{e}^{\mathrm{s}}_j$ for $j = 1, \ldots, N$. The left side shows the cosine coefficients (4.47a) and the right side shows the sine coefficients (4.47b). Values scale in gray from 0 (white) to 1 (black).

represents the information in a plot like Figure 4.5 or 4.6. Theoretically, the left image should show a (black) diagonal only and the right one should be blank. Figure 4.7 reveals, however, that only the first few L-FEM observation vectors exhibit this "correct" behavior. Roughly, the first half $j < N/2$ of the observation vectors appear at least tolerable, whereas the remaining, $j > N/2$, vectors are particularly problematic as they contain no components of the highest frequencies (notice all columns greater than 42 are approximately zero). What we see here is a consequence of the scheme's *numerical dispersion relation*: the long wavelength components travel at near correct phase speed whereas components at shorter wavelengths travel too slow (see Figure 3.15). This behavior leads to observation of frequencies which are lower than the correct ones.

The close relation between the observation and the dispersion relation is obvious. The observation of the sine basis may, in fact, be considered a practical verification of the dispersion relation shown on Figure 3.15. Figure 4.8 shows the absolute value $|\beta^{\cos}_{jk} + i\beta^{\cos}_{jk}|$ of the coefficients (4.47), where i is the complex unit, for P^0 observations made from L-FEM discretization with four different Courant numbers μ, that is, with different time step sizes Δt. On the images are also shown the phase velocities as functions of j. The largest coefficients on each image follows the phase velocity line quite clearly. Also the *dissipation* plays a role on both Figure 4.8 and on Figure 4.7. A gray "fog" appears in

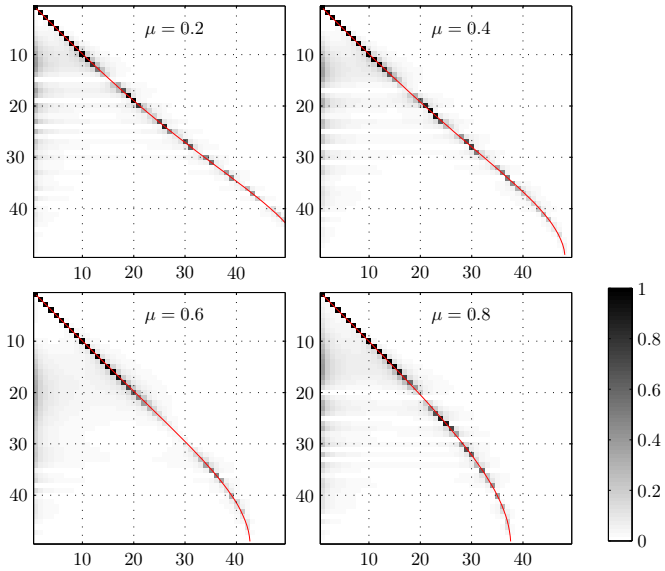


Figure 4.8: Coefficient images similar to Figure 4.7 with P^0 observations of the sine basis. All are discretized with L-FEM and trapezoidal time integration but with different Courant numbers μ. The phase velocity from Figure 3.15 is also shown in each case.

the leftmost part (the lowest coefficients) on each image; this "fog" is caused by dissipation. Another damping effect can be seen by the diminishing strength of diagonal. Finally, we see a smearing effect on the black diagonal already after a few sinusoids; this is also partly due to dissipation. Damping is a well-known property of lower order schemes such as the L-FEM.

Collectively, these effects result in bad conditioning of the observation matrix $\boldsymbol{P}$—and increasingly so for finer grids. Particularly in the situations $\mu = 0.6$ and $\mu = 0.8$ we loose orthogonality on the highest modes in the observation space $P(\mathcal{X})$—the "angle" between the last, high-frequency, observations goes towards 0 for $h \to 0$. This corresponds to—almost—loosing a dimension in the observation space. A similar thing happens in the cases $\mu = 0.2$ and $\mu = 0.4$—at least as long as we do not allow longer control time than $T = 2$. The faster phase velocity leaves small "gaps", or rarefaction areas, in the spectrum which also deteriorates the orthogonality in the observation space. The described loss of orthogonality will lead to ill-posedness when we later consider the *inversion* of matrix $\boldsymbol{L}$ derived from the observation map.

Reconstruction and matrix $\boldsymbol{L}$

We continue with the reconstruction by R from the above found vectors of observation, $P^0(\boldsymbol{e}_j^{\mathbf{s}})$ and $P^1(\boldsymbol{e}_j^{\mathbf{s}})$ for $j = 1, \ldots, N$. Discrete reconstruction (4.29) consists of solving the backwards wave equation (4.11) with a Dirichlet boundary condition applied by $\boldsymbol{B}_h$ which defined in (4.16) for L-FEM.

We compute the elements of $\boldsymbol{L}$ by direct assembly (4.34)—still using the sine basis. Results very similar to the ones below would be seen with inner-product assembly; see Section 4.2.5 for the use of this technique. The inner-products used in direct assembly (4.34) are defined in (4.18) and (4.21) for L-FEM.

Figure 4.9 shows the eigenvalues of the constructed $\boldsymbol{L}$ alongside the analytic values. The lower plot displays the relative error for the approximation of each eigenvalue. The use of the sine basis results—theoretically—in a diagonal $\boldsymbol{L}$.

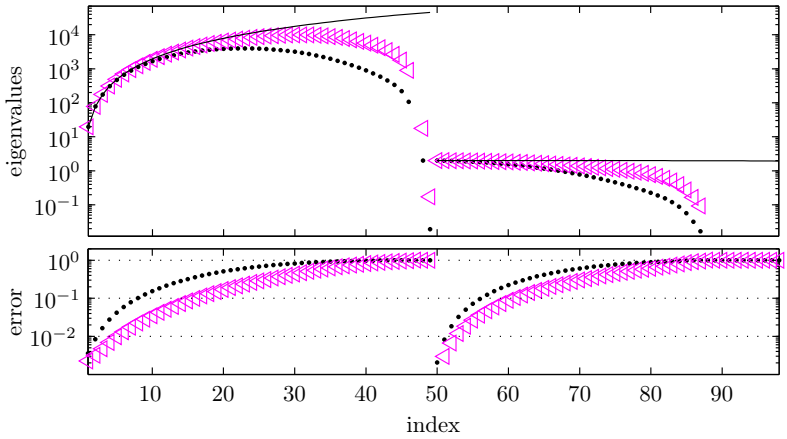


Figure 4.9: Logarithmic plot of the eigenvalues of $\boldsymbol{L}$ for $T = 2$ with L-FEM approximation ($N = 49, \mu = 0.6$) marked with dots. The triangles show the eigenvalues obtained with *normalized* sine basis $\tilde{e}_j^{\mathrm{Ls}} = e_j^{\mathrm{Ls}} / \left\| e_j^{\mathrm{Ls}} \right\|_{L^2(\Omega)}$. The solid line shows the exact eigenvalues. The relative error is shown with dots and triangles, respectively, in a separate plot below.

The eigenvalues of $\boldsymbol{L}$ are therefore the union of the eigenvalues of $\boldsymbol{L}^1$ and $\boldsymbol{L}^4$. In this way, the left half of Figure 4.9 shows the eigenvalues of $\boldsymbol{L}^1$, the right shows the eigenvalues for $\boldsymbol{L}^4$.

Only the first few eigenvalues (long wavelengths) of $\boldsymbol{L}^1$ are approximated reasonable well. The same holds for the eigenvalues of $\boldsymbol{L}^4$. Then up till around 2/3 (long to moderate wavelengths) of the eigenvalues for both $\boldsymbol{L}^1$ and $\boldsymbol{L}^4$ are tolerable. The remaining, though, are quite bad; that holds again for both $\boldsymbol{L}^1$ and $\boldsymbol{L}^4$. And for increasing wavenumber j it gets increasingly worse and the eigenvalues quickly drop to near-zero values. The smallest eigenvalues of $\boldsymbol{L}$ have dropped far below the bottom of the plot in Figure 4.9 and are of the order 10^{-10}.

These rapidly decaying eigenvalues which makes $\boldsymbol{L}$ very ill-conditioned has to do with numerical dispersion, yet it cannot be explained by the phase velocity alone. Consider Figure 4.10 which shows the first N eigenvalues, corresponding to the left half of Figure 4.9, of $\boldsymbol{L}$ obtained with four different Courant numbers $\mu = 0.2, 0.4, 0.6$ and 0.8. On the right side of the figure, we see the group velocities (see Figure 3.17) multiplied by the exact eigenvalues. The close connection between the decaying eigenvalues and the group velocity seems evident, although not as simple as the relation between the phase velocity and P^0 observation. The group velocity tends to zero for $j \to N$ for all L-FEM discretizations presented in Section 3.5.1. This manifests itself in unphysical periodicity seen, *e.g.,* for $P^0(\boldsymbol{e}_N^{\mathrm{s}} - \boldsymbol{e}_{(N-1)}^{\mathrm{s}}) = P^0(\boldsymbol{e}_N^{\mathrm{s}}) - P^0(\boldsymbol{e}_{(N-1)}^{\mathrm{s}})$.

The huge condition number of $\boldsymbol{L}$, which in this case with $N = 49, \mu = 0.6, T = 2$ is $\mathrm{cond}(L) \approx 10^{18}$, is a major obstacle for obtaining useful solutions to the matrix equation (4.37). It is quite easy to see from Figure 4.9 that any short wavelengths in the initial data $[\boldsymbol{u}^1, -\boldsymbol{u}^0]^{\mathsf{T}}$ will be blown up and dominate the solution $[\bar{\boldsymbol{w}}^0, \bar{\boldsymbol{w}}^1]^{\mathsf{T}}$ completely.

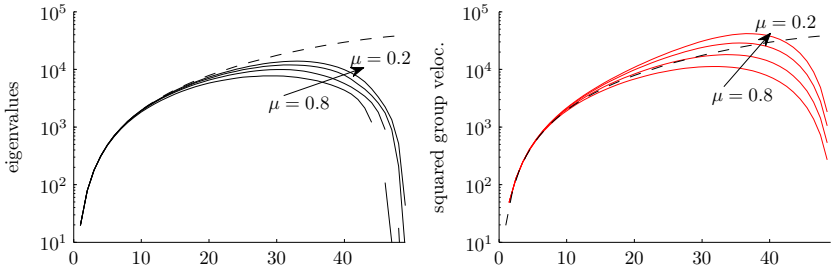


Figure 4.10: Left: Logarithmic plot of the eigenvalues of $\boldsymbol{L}^1$ computed with L-FEM with respectively $\mu = 0.2, 0.4, 0.6$ and 0.8. Right: The exact eigenvalues times the group velocity for the same four schemes.

A simple filter

It is clear by now that some filtering or regularization is needed in order to restore useful solutions. Several things can be done as explained in Zuazua's review [Zua05]; among these are Fourier filtering, Tychonoff regularization or bi-grid algorithms.

We have formulated $\boldsymbol{L}$ in sine basis, and it allows us to construct a family of reduced matrices $\boldsymbol{L}_{(N_c)}$ of size $2N_c \times 2N_c$, where N_c is a cut-off number $1 \leq N_c \leq N$. Each of the four sub-matrices of $\boldsymbol{L}_{(N_c)}$ consists of the N_c first elements of the first N_c columns of $\boldsymbol{L}$. Let $\boldsymbol{L}^n_{(N_c)}, n = 1, 2, 3, 4$ be the four sub-matrices of $\boldsymbol{L}_{(N_c)}$. Then $\boldsymbol{L}_{(N_c)}$ can be constructed from the sine observations by the inner product assembly

$$\left.\begin{aligned} (\boldsymbol{L}^1_{(N_c)})_{ij} &= \langle P^0 \boldsymbol{e}^s_j, P^0 \boldsymbol{e}^s_i \rangle_{\mathcal{T}} \qquad (\boldsymbol{L}^2_{(N_c)})_{ij} = \langle P^1 \boldsymbol{e}^s_j, P^0 \boldsymbol{e}^s_i \rangle_{\mathcal{T}} \\ (\boldsymbol{L}^3_{(N_c)})_{ij} &= \langle P^0 \boldsymbol{e}^s_j, P^1 \boldsymbol{e}^s_i \rangle_{\mathcal{T}} \qquad (\boldsymbol{L}^4_{(N_c)})_{ij} = \langle P^1 \boldsymbol{e}^s_j, P^1 \boldsymbol{e}^s_i \rangle_{\mathcal{T}}. \end{aligned}\right\} \tag{4.48}$$

for $i, j = 1, \ldots, N_c$. It can be defined in the same way for direct assembly.

Consider the family of reduced matrices $\{\boldsymbol{L}_{(N_c)}\}_{N_c \leq N}$. The full matrix $\boldsymbol{L}$ is restored by taking $N_c = N$, that is, $\boldsymbol{L}_{(N)} = \boldsymbol{L}$. Figure 4.11 shows the condition number of the complete family, $N_c = 1, \ldots, N$, of reduced matrices $\boldsymbol{L}_{(N_c)}$. It

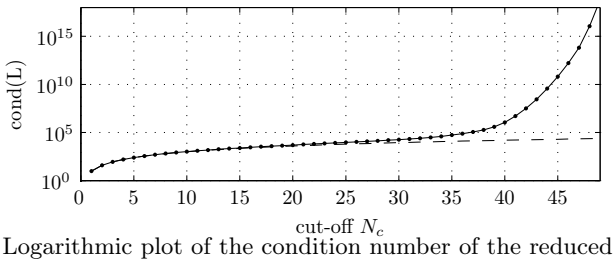


Figure 4.11: Logarithmic plot of the condition number of the reduced matrix $\boldsymbol{L}_{(N_c)}$ as function of the number of modes N_c. The dashed line shows the theoretical value.

illustrates the same problem with the highest modes as we studied in Figure 4.9. We can only expect reasonable behavior of $\boldsymbol{L}$ if we discard the highest modes and take, *e.g.*, $N_c = \lfloor 3/4N \rfloor$ (largest integer less than $3/4N$), when constructing the reduced $\boldsymbol{L}_{(N_c)}$.

Eigenfunction controls

The approximation of $\boldsymbol{L}$ is not our ultimate goal, we are really looking for good approximate controls. How well we have achieved this end can be assessed by considering the set of what we will call *eigenfunction controls*. We define an eigenfunction control η_j for $j \in \mathbb{N}$ as a control for either the data

$$u^0(x) = 0, \qquad u^1(x) = e_j^s(x),$$

or the data

$$u^0(x) = e_j^s(x), \qquad u^1(x) = 0.$$

We will focus here on controls for the latter since will need these in Chapter 5. The controls are found by

$$\eta_j = -\Phi\Lambda^{-1} \begin{bmatrix} 0 \\ -e_j^{\mathrm{s}} \end{bmatrix},$$

where Φ is the continuous observation operator (2.14) and Λ is the HUM operator defined in (2.21). In the 1-d case with $T = 2$, the controls are easily determined analytically as

$$\eta_j^{\mathrm{ex}}(t) = (-1)^{j+1} \frac{\sqrt{2}}{T} \sin(j\pi t), \qquad j = 1, 2, \ldots.$$

We note that these controls constitute a basis on $\mathcal{B}$.

Consider the approximate eigenfunction controls

$$\boldsymbol{\eta}_j = -P_{(N_c)} \boldsymbol{L}_{(N_c)}^{-1} \begin{bmatrix} 0 \\ -\boldsymbol{e}_j^{\mathrm{s}} \end{bmatrix}, \qquad j = 1, \ldots, N_c,$$

computed by the reduced matrix $\boldsymbol{L}_{(N_c)}$ introduced above also for $T = 2$.

We examine the coefficients of the L-FEM eigenfunction controls $\boldsymbol{\eta}_j$ with respect to the basis of the exact eigenfunction controls η_k^{ex} and call these coefficients

$$\beta_{jk}^{\mathrm{eig}} = \langle \boldsymbol{\eta}_j, \boldsymbol{\eta}_k^{\mathrm{ex}} \rangle_{\mathcal{T}} / \left\| \boldsymbol{\eta}_k^{\mathrm{ex}} \right\|_{\mathcal{T}}, \qquad j, k = 1, \ldots, N_c, \tag{4.49}$$

where the vector $\boldsymbol{\eta}_k^{\mathrm{ex}}$ contains M discrete samples of $\eta_j^{\mathrm{ex}}(t)$. We use a filter with cut-off index $N_c = 35$ as it seems reasonable both in terms of the eigenvalues (Figure 4.9) and the condition number of $\boldsymbol{L}$ (Figure 4.11). Approximate controls computed with two different Courant numbers $\mu = 0.6$ and $\mu = 0.2$ are studied.

Figure 4.12 shows the spectrum, that is, the absolute value of the coefficients $\beta_{jk}^{\mathrm{eig}}$, of eigenfunction controls with $\mu = 0.6$. The error of the obtained coefficients is shown on Figure 4.13. Although not perfect, these results are fair especially if we judge by how well the black diagonal is retained in Figure 4.12. Also, the low frequencies show very little error as we see on the left part of the error image in Figure 4.13. The central and lower right region on the error plot does, however, reveal relatively large errors. The l^2-error of the coefficients of each L-FEM eigenfunction control (each row in the image) is plotted as function of the index j in Figure 4.16. We will comment on this error after studying the controls obtained with $\mu = 0.2$.

Let us consider the spectrum of L-FEM eigenfunction controls with $\mu = 0.2$ on Figure 4.14. The smaller Courant number means 3 times as many time steps

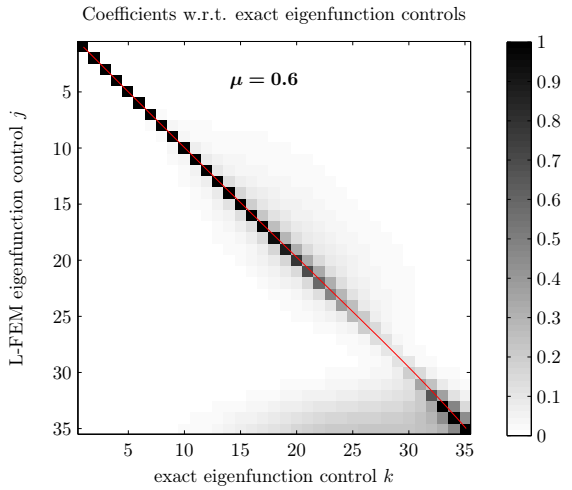


Figure 4.12: The spectrum (4.49) of the eigenfunction controls $\eta_j, j = 1, \ldots, N_c$ obtained by L-FEM with $N = 49, N_c = 35$ and $\mu = 0.6$. The corresponding numerical phase velocity is shown with a thin, solid line.

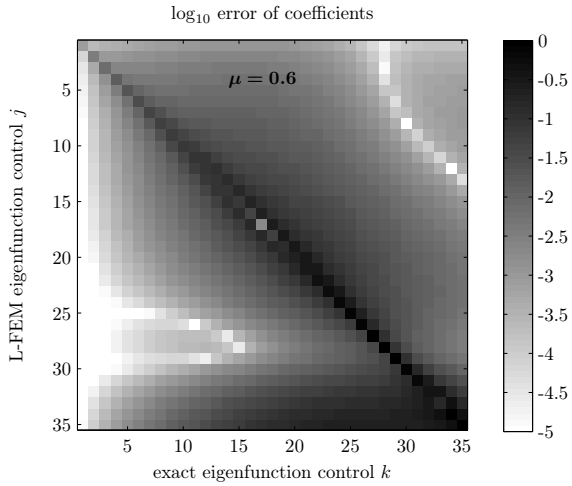


Figure 4.13: $\log_{10}$ of the error on coefficients shown on Figure 4.12 with $\mu = 0.6$. The average l^2 error is 0.476, and the average max-error is 0.371.

which would give more accurate results if it were not for numerical dispersion. The coefficients on Figure 4.14 stray away from the diagonal just as we saw on Figure 4.8(upper left) which is of course no surprise as we again consider the observation of sinusoidal initial data. We see that the coefficients stick very close to the phase velocity curve also shown on the figure. The corresponding error is shown on Figure 4.15. The region around the diagonal is very dark since the controls have a distinct frequency error.

Figure 4.16 displays the l^2-error of the coefficients of each control j (each row in the image) for both $\mu = 0.2$ and $\mu = 0.6$. On this figure, it is quite clear that $\mu = 0.6$ gives a better result than $\mu = 0.2$. This is due to numerical phase velocity of $\mu = 0.2$ is too fast compared to the exact value. Notice that the eigenvalues of $\boldsymbol{L}$ was, in fact, approximated better with $\mu = 0.2$ than with $\mu = 0.6$ (Figure 4.10).

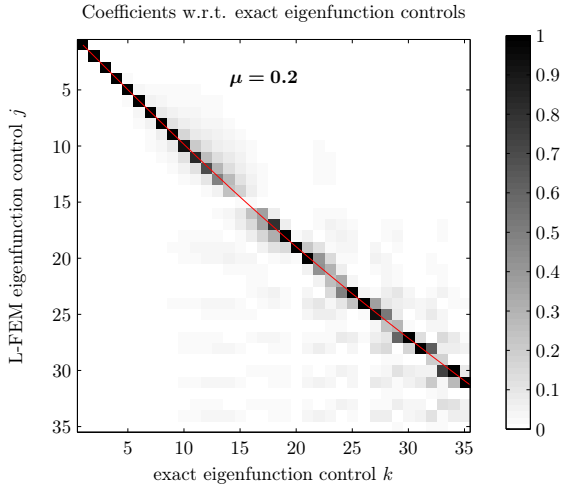


Figure 4.14: Spectrum (4.49) of the eigenfunction controls η_j for $j = 1, \ldots, N_c$ obtained by L-FEM with $N = 49, N_c = 35$ and $\mu = 0.2$. The corresponding numerical phase velocity is shown with a thin, solid line.

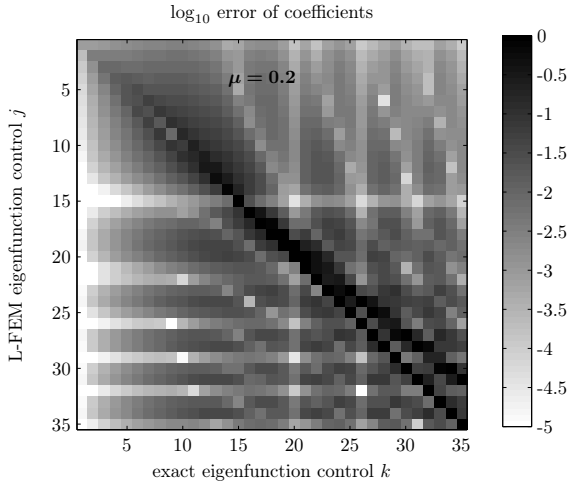


Figure 4.15: $\log_{10}$ of the error on coefficients shown on Figure 4.14 with $\mu = 0.2$. The average l^2 error is 0.999, and the average max-error is 0.834.

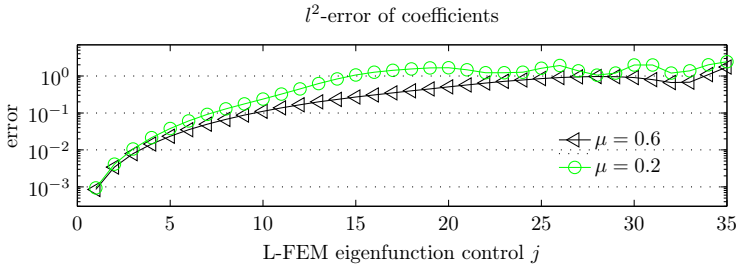


Figure 4.16: l^2 error of coefficients for each eigenfunction control $\boldsymbol{\eta}_j$ as function of the index j for two different approximations with respectively $\mu = 0.2$ and $\mu = 0.6$ both with $N = 49, N_c = 35, T = 2$.

An example with very little noise

Consider an example with $T = \sqrt{11} - e/3 \approx 2.411$. Allowing $T > 2$ relaxes the problem a bit (and reduce the condition number of $\boldsymbol{L}$), but the non-zero elements stray away from the diagonal distinctive for $\boldsymbol{L}$ when $T = 2$. We keep $N = 49$ and $\mu = 0.6$. These choices result in a matrix $\boldsymbol{L}$ with condition number $3.24 \cdot 10^{14}$ and eigenvalues very similar to those displayed in Figure 4.9. Let the functions

$$u^0 = \sqrt{2}\sin(5\pi x), \qquad u^1 = 0,$$

be initial data for a control problem. We perturb u^0 slightly by adding 0.01% of random noise and call the perturbed data u_ε^0. The spectrum of u_ε^0 consists of a 1 at the fifth position and random values around 10^{-5} at all other places.

We sample u_ε^0, solve the HUM problem (4.37) and compute the control (4.38). We do the same—solve (4.37) and compute the control—with a reduced $\boldsymbol{L}_{(N_c)}$

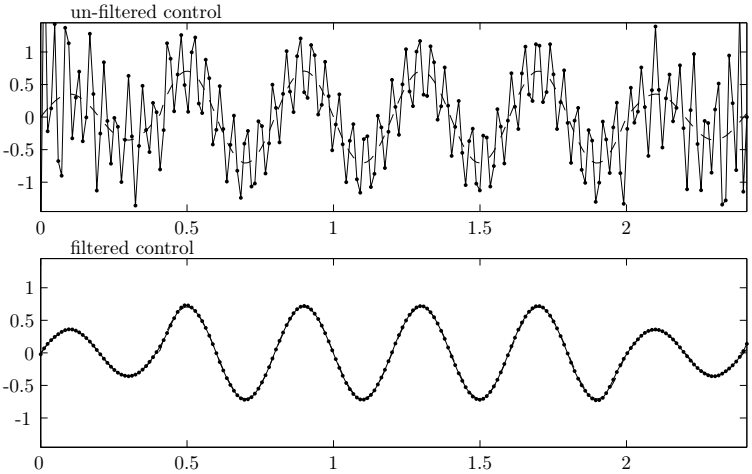


Figure 4.17: Control of $u^0 = \sqrt{2}\sin(5\pi x)$ with 0.01% noise computed by $\boldsymbol{L}$ (upper plot) and by reduced $\boldsymbol{L}_{(N_c)}$ with $N_c = 36$ (lower plot). L-FEM with $N = 49$ and $\mu = 0.6$ was been used in both cases and $T = \sqrt{11} - e/3$. The dashed line is the exact solution.

with $N_c = \lfloor 3/4N \rfloor = 36$. Both computed controls can be seen in Figure 4.17. The strong effect from the high-frequency components of the noise is evident for the un-filtered control. The blow up is caused by the last, near-zero eigenvalues of $\boldsymbol{L}$. If we instead consider the accuracy of computed initial data $\bar{\boldsymbol{w}}^0$ and $\bar{\boldsymbol{w}}^1$ for the adjoint problem, the noise is even more pronounced. It is needless to say that more irregular initial data or shorter control time T would render yet worse approximate controls. For sufficiently bandlimited data, however, *e.g.,* with bandlimit $N_c = \lfloor 3/4N \rfloor$, solutions computed by $\boldsymbol{L}$ or $\boldsymbol{L}_{(N_c)}$ would show no real difference.

Asymptotic properties

The control time T influences the properties of the matrix $\boldsymbol{L}$ in an intuitive way. The longer the time for observation the more we are able to observe and the condition number of $\boldsymbol{L}$ decreases. But only slowly. And we wish not to relax

Table 4.1: Amplitude error of computed solutions with full matrix $\boldsymbol{L}$ and reduced $\boldsymbol{L}_{(N_c)}$ with $N_c = 36$. $\bar{w}_h^0$ and $\bar{w}_h^1$ are the linear splines generated by the vectors $\bar{\boldsymbol{w}}^0$ and $\bar{\boldsymbol{w}}^1$. The norms are computed by the approximations (4.18) and (4.27).

	full, N	reduced, $3/4N$
$\left\|\bar{w}_h^0 - \bar{\varphi}^0\right\|_{L^2(\Omega)}$	9.61e+04	2.86e-02
$\left\|\bar{w}_h^1 - \bar{\varphi}^1\right\|_{L^2(\Omega)}$	3.40e+05	3.90e-02
$\left\|k_h - \kappa_{\text{ex}}\right\|_{L^2(0,T)}$	2.05e+00	3.74e-02

the control time just to pay regards to specific numerical effects; $T = 2$ is in many ways the most interesting case.

Unfortunately the spacing h has a much stronger effect on the conditioning of $\boldsymbol{L}$—the condition number increases dramatically with decreasing h. Furthermore, the cost of computing $\boldsymbol{L}$ rapidly becomes massive when h decrease as it requires the solution of 2 or $4N$ wave equations (depending on the type of assembly procedure) of increasing size.

It becomes infeasible to construct the matrix very quickly—even in this 1-d case. We may instead resort to an iterative method for the solution of a large problem. Section 4.3 is dedicated to the conjugate gradients method for the iterative solution of the problem (4.37). We postpone convergence analysis of approximate controls to that section. We may, at this point, assume convergence of controls for properly bandlimited initial data.

Changing the temporal spacing (for fixed h), that is, the Courant number μ also has an effect on the matrix $\boldsymbol{L}$ and the control. But as we saw in the study of eigenfunction controls computed with two different Courant numbers, this effect is highly dependent on the time integration scheme. Decreasing μ does not necessarily give more accurate results, but neither does it necessarily lead to worse conditioning of $\boldsymbol{L}$. One most consult the scheme's numerical dispersion relation before drawing any conclusions about asymptotic properties for μ.

Concluding remarks

We have used L-FEM as semi-discretization with trapezoidal time integration in this section. Before we could engage in the construction of $\boldsymbol{L}$, we had to choose its *basis*. We chose the sine basis over the canonical due to the former's ability to separate well and poorly resolved waves. The N linear approximants represented the same N continuous sine functions unambiguously in the frequency domain; the amplitude errors were significant, though.

We studied the P^0 *observation* of the sines basis and, we found that much of the frequency behavior was shaped by the numerical phase velocity; the effects of numerical dissipation were also apparent. We considered schemes with different Courant numbers and saw, for different reasons though, that numerical dispersion threatened the orthogonality in the observation space.

The matrix $\boldsymbol{L}$ was constructed by direct assembly, and we compared the numerically found eigenvalues with the exact ones. The eigenvalues corresponding to low wavenumbers were approximated well, the midrange wavenumbers tolerable, but the ones corresponding to the high wavenumbers were very poor and

tended to zero rapidly. The near-zero eigenvalues caused the condition number of $\boldsymbol{L}$ to rocket leaving the solution extremely sensitive to noise (see Figure 4.17). These effects were attributed numerical group velocity which also tends to zero for high wavenumbers. The formulation in sine basis gave rise to a simple *filtering* procedure consisting of truncating the assembly of each of the sub-matrices $\boldsymbol{L}^n$, $n = 1, 2, 3, 4$ after N_c basis functions. This Fourier filter required therefore only N_c/N times the computations of the full matrix.

We used the filter when we computed a set of so-called *eigenfunction controls*. We computed these controls with two different Courant numbers $\mu = 0.2$ and $\mu = 0.6$, and studying the spectrum of the results showed that numerical phase velocity has a major impact on the quality of the controls. Not knowing about the effects of numerical dispersion, one might think that the better temporal resolution of $\mu = 0.2$ would ensure better results. This was, however, not the case. The results from $\mu = 0.6$ were better due to better numerical phase velocity. A filter will eliminate the effect of the vanishing group velocity, but even small inaccuracies in the phase velocity will come out strong as frequency error for the computed controls.

4.2.5 Constructing $\boldsymbol{L}$ with DG-FEM

Section 3.3 introduced the discontinuous Galerkin-FEM as a method for semi-discretization. We shall use the DG-FEM scheme formulated for characteristic variables, which was specified on page 58 ff, with the 4'th order LSERK time integration (see page 43) with Courant number $\mu = 0.6$.

We will primarily use a discretization with $K = 10$ elements and fifth order local polynomial approximation, $N_p = 6$. This semi-discretization has $K \cdot (N_p - 1) - 1 = 49$ inner nodes (recall element endpoints are defined twice) which equals the number of inner nodes used in the L-FEM discretization in the previous section. We use N for the number of inner nodes $K \cdot (N_p - 1) - 1$. Four other discretizations, all with exactly $N = 49$ inner points, will be used too, but to a lesser extend. All five grids are presented below

$$\begin{aligned}
\textbf{grid 0}: &\qquad \boldsymbol{N_p = 6}, &&\qquad \boldsymbol{K = 10},\\
\text{grid } a: &\qquad N_p = 2, &&\qquad K = 50,\\
\text{grid } b: &\qquad N_p = 3, &&\qquad K = 25,\\
\text{grid } c: &\qquad N_p = 11, &&\qquad K = 5,\\
\text{grid } d: &\qquad N_p = 51, &&\qquad K = 1.
\end{aligned}$$

When nothing else is mentioned we use grid 0 (shown in bold face). Let in the following x_i^k denote the i'th node, $1 \leq i \leq N_p$, on the k'th element, $1 \leq k \leq K$.

We wish to represent the continuous sine basis (4.42) on a DG-grid. We have already argued for the use of a sinusoidal basis (see discussion in Section 4.2.3), thus we also acknowledge that other could have been used, *e.g.,* a basis constructed from the local polynomial basis. But the author had only limited success with this.

The sine basis

The DG-formulation suggests two different ways of representing the continuous sine basis (4.42): a nodal and a modal. The nodal is based on interpolation

after sampling of the continuous function e_j^{s} in the grid points x_i^k

$$e_j^{\mathrm{s}}(x_i^k) = \sqrt{2}\sin(j\pi x_i^k), \qquad j = 1, \ldots, N.$$

We collect all these values in the j'th vector

$$\boldsymbol{e}_j^{\mathrm{ss}} = [\boldsymbol{e}_j^{\mathrm{ss},1} \ldots, \boldsymbol{e}_j^{\mathrm{ss},K}]^{\mathsf{T}}, \quad \boldsymbol{e}_j^{\mathrm{ss},k} = [e_j^{\mathrm{s}}(x_1^k), \ldots, e_j^{\mathrm{s}}(x_{N_p}^k)]^{\mathsf{T}}, \tag{4.50}$$

where the superscript ss is for *s*ampled *s*ine. We obtain an order $N_p - 1$ polynomial on each element D^k by the interpolation Lagrangian polynomials ℓ_i from the nodal values of the vector $\boldsymbol{e}_j^{\mathrm{ss},k}$. The collected piecewise polynomial becomes

$$e_j^{\mathrm{ss}} = \bigoplus_{k=1}^{K} e_j^{\mathrm{ss},k} \text{ in } \Omega, \qquad e_j^{\mathrm{ss},k}(x) = \sum_{i=1}^{N_p} e_j^{\mathrm{s}}(x_i^k)\ell_i(x), \quad x \in \mathsf{D}^k. \tag{4.51}$$

We will refer to this sampled approximation of the sine basis function e_j^{s} by the nodal "basis" vector $\boldsymbol{e}_j^{\mathrm{ss}}$ or by the function e_j^{ss}.

The normalized Legendre polynomials $\widetilde{P}_n$ constitute an orthonormal basis on $L^2(-1,1)$. It allows us to expand any L^2-function on element D^k in this basis by mapping to the reference element $\mathsf{I} = [-1,1]$. Consider the sine basis function e_j^{s} on element $\mathsf{D}^k = [x_{\mathrm{L}}^k, x_{\mathrm{R}}^k]$ mapped to reference element I

$$e_j^{\mathrm{s}}\big|_{\mathsf{D}^k}(r) = \sqrt{2}\sin\big(j\pi(x_{\mathrm{L}}^k + \frac{h^k}{2}(1+r))\big), \qquad -1 \le r \le 1.$$

We project this function onto the Legendre basis and obtain the coefficients

$$\widehat{e}_{j,n}^{\mathrm{s},k} = \langle e_j^{\mathrm{s}}\big|_{\mathsf{D}^k}, \widetilde{P}_{n-1}\rangle_{\mathsf{I}}, \qquad n = 1, 2, \ldots$$

for the expansion

$$e_j^{\mathrm{s}}\big|_{\mathsf{D}^k}(r) = \sum_{n=1}^{\infty} \widehat{e}_{j,n}^{\mathrm{s},k}\widetilde{P}_{n-1}(r).$$

The bi-linear form $\langle\cdot,\cdot\rangle_{\mathsf{I}}$ is the L^2-inner product on the reference element. Let us collect the N_p first Legendre coefficients for all K elements in the modal vector

$$\widehat{\boldsymbol{e}}_j^{\mathrm{ps}} = [\widehat{\boldsymbol{e}}_j^{\mathrm{ps},1}, \ldots, \widehat{\boldsymbol{e}}_j^{\mathrm{ps},K}]^{\mathsf{T}}, \qquad \widehat{\boldsymbol{e}}_j^{\mathrm{ps},k} = [\widehat{e}_{j,1}^{\mathrm{s},k}, \ldots, \widehat{e}_{j,N_p}^{\mathrm{s},k}]^{\mathsf{T}}, \tag{4.52}$$

where the superscript ps is for *p*rojected *s*ine. The corresponding nodal vector is obtained by the Vandermonde matrix (3.36) and reads

$$\boldsymbol{e}_j^{\mathrm{ps}} = [\boldsymbol{V}\widehat{\boldsymbol{e}}_j^{\mathrm{ps},1}, \ldots, \boldsymbol{V}\widehat{\boldsymbol{e}}_j^{\mathrm{ps},K}]^{\mathsf{T}}.$$

The truncated Legendre expansion leads to an order $N_p - 1$ polynomial on each element D^k. We collect these in the approximating function

$$e_j^{\mathrm{ps}} = \bigoplus_{k=1}^{K} e_j^{\mathrm{ps},k} \text{ in } \Omega, \qquad e_j^{\mathrm{ps},k}(r) = \sum_{n=1}^{N_p} \widehat{e}_{j,n}^{\mathrm{s},k}\widetilde{P}_{n-1}(r), \quad -1 \le r \le \tag{4.53}$$

We will use the function e_j^{ps} or the vector $\boldsymbol{e}_j^{\mathrm{ps}}$ to refer to this L^2-projection of the sine basis function e_j^{s}.

Locally, on each element, the projection onto the space of Legendre polynomials gives the "best" approximation in terms of minimal L^2-error [HGG07]. The procedure does, however, not necessarily lead to continuous approximations and specifically not to zero values in the domain endpoints. Figure 4.18 shows an example with a projected and a sampled approximation of e^{s}_{32} on a part of the domain. Notice how the projected function deviate from zero at the

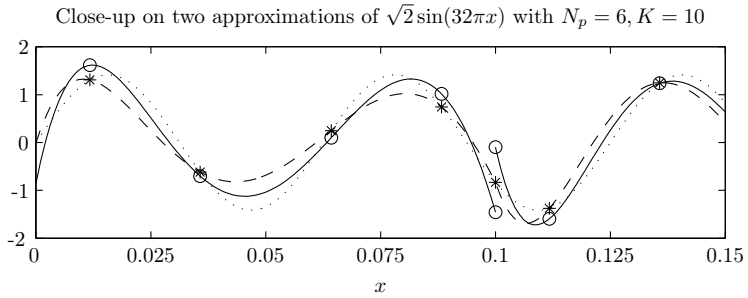


Figure 4.18: A close-up on element D^1 and the left part of D^2 for two approximations of $e^{\mathrm{s}}_{32} = \sqrt{2}\sin(32\pi x)$—one based on sampling e^{ss}_{32} (dashed, $*$) and one based on projection e^{ps}_{32} (solid, o) for grid 0. Compare with the exact function (dotted line).

left endpoint and the significant discontinuity at the interface between D^1 and D^2. It seems reasonable from the plot, though, that this function should have smaller L^2-error compared to the sampled approximation.

We have now defined two piecewise polynomial approximations to the sine basis function e^{s}_j: the sampled e^{ss}_j and the projected e^{ps}_j. Before moving on to using them as basis for the construction of the matrix $\boldsymbol{L}$, we shall assess their quality as approximations to e^{s}_j. Sine functions with long wavelengths (small j) compared to the number of nodes N are expected to be approximated quite well by piecewise polynomials. The theory of polynomial approximation tells us, on the other hand, that approximation of sinusoidals of short wavelengths (large j) compared to N should be handled with caution. In general 4 points-per-wavelength are required to obtain exponential decay of coefficients for approximation with polynomials [HGG07].

The L^2-norm of the approximation error made when approximating the j'th sine basis function for grid 0 is shown on Figure 4.19. The projected data shows smaller error than the sampled as we would expect—the difference is small, though.

Figure 4.20 shows the approximation error for the sampled data on all five grids. From about $j = 30$ onwards the error for all grids are approximately the same. The approximation error of the projected sinusoids for the other four grids behave very similarly; the plots have been omitted for brevity.

Let us now consider the behavior of the two approximations in Fourier domain. We seek polynomial approximations of the j'th sine function which demonstrate corresponding spectral properties. Recall the standard expansion

$$f(x) = \sum_{i=1}^{\infty} \widehat{f}^{\mathrm{s}}_i e^{\mathrm{s}}_i(x), \qquad \widehat{f}^{\mathrm{s}}_i = \langle f, e^{\mathrm{s}}_i \rangle_{L^2(\Omega)}, \quad i = 1, 2, \ldots$$

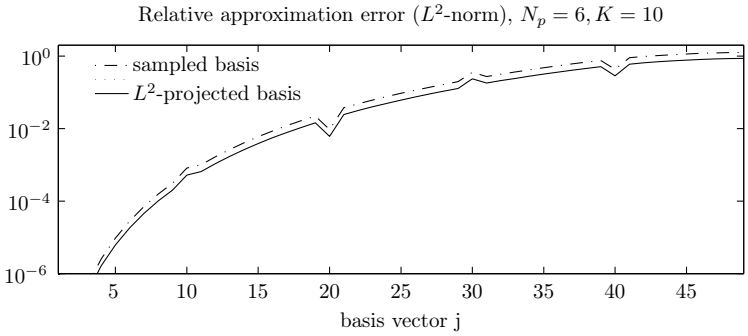


Figure 4.19: The L^2-norm of the error approximating e_j^{s} by sampling, e_j^{ss}, and L^2-projection, e_j^{ps}, for grid 0.

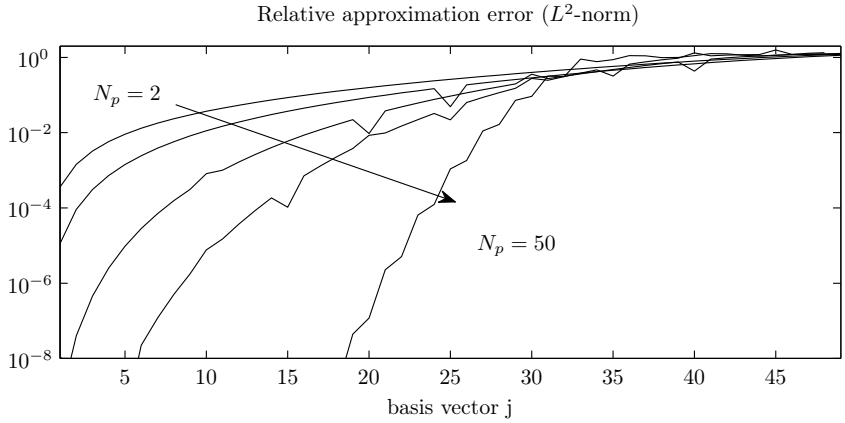


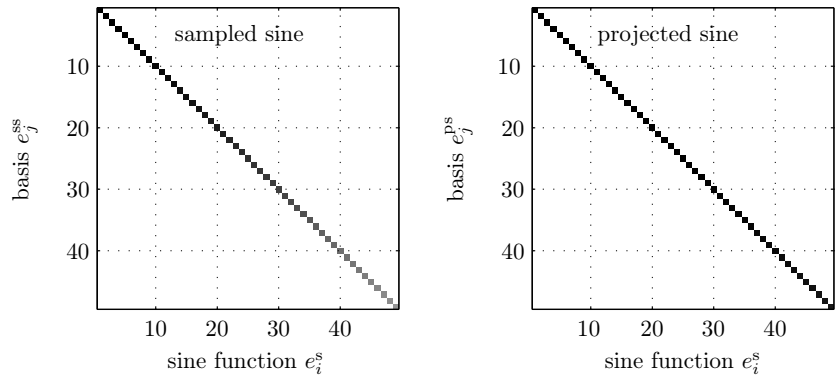
Figure 4.20: The L^2-norm of the error approximating e_j^{s} by sampling for all five grids.

that we saw on page 66. We have determined the coefficients $\widehat{f}_i$ for the approximating functions $f = e_j^{\mathrm{ss}}$ and $f = e_j^{\mathrm{ps}}$ for $i, j = 1, \ldots, N$ for all five grids in Figure 4.21 and 4.22. All polynomials have higher frequency components as well, but they are not relevant here and have been omitted from the plots.

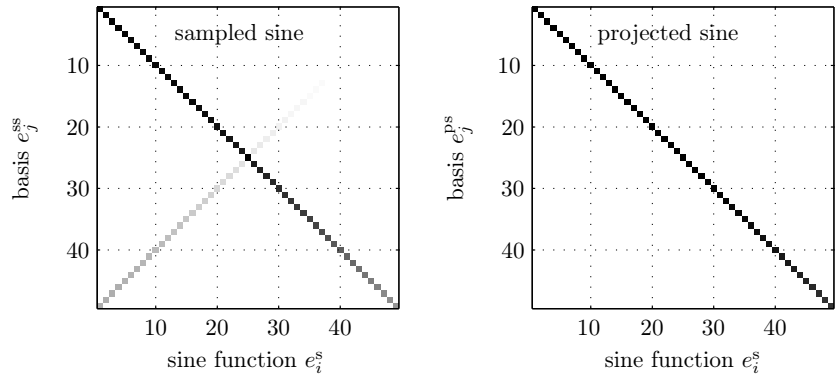
The images can be compared with the left side of Figure 4.3 which showed the spectrum of the L-FEM approximation of the same sine basis. Not surprisingly, the left plot of Figure 4.21(a) shows clear resemblance with Figure 4.3; they both display coefficients found from linear approximation of the sampled sine functions on an equidistant grid with $N = 49$ grid points. We see again only diagonal values and that the values of the coefficients decrease for large j (turns from black to gray on the plots).

The quadratic polynomial approximations of grid b gives rise to unwanted ambiguity for $j > 25$ as seen on left plot on Figure 4.21(b). By ambiguity we mean that the polynomial approximation of sinusoid e_j^{s} with large wavenumber, say $j = 27$ will "look like" a combination of the 27'th and the 23'rd mode. This ambiguity is *not* seen when approximation sinusoids by linear splines on an equidistant grid which Figure 4.3 also showed. We expect the boundary observation, $P(\boldsymbol{e}_{27}^{\mathrm{ss}})$, to contain the two components as well.

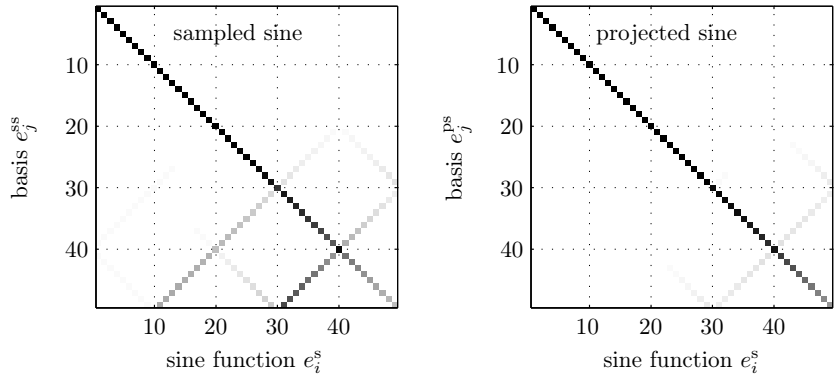
The same phenomena can be seen from about $j = 25$ onwards in Fig-



(a) Grid a: $N_p = 2, K = 50$.

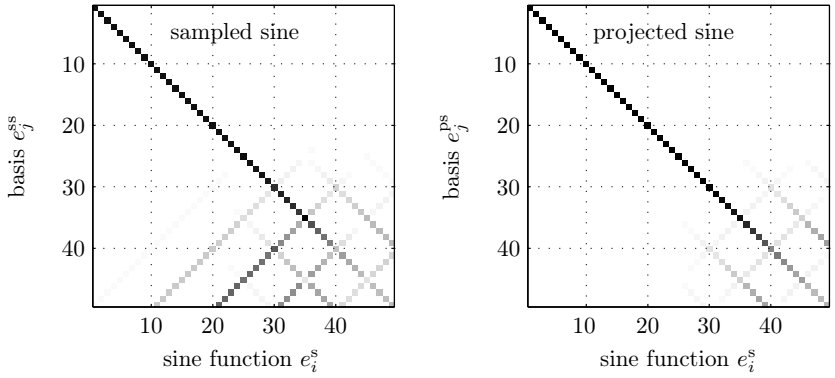


(b) Grid b: $N_p = 3, K = 25$.

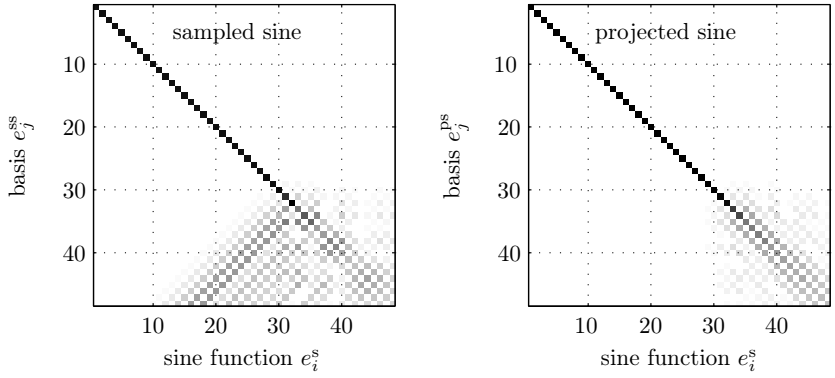


(c) Grid 0: $N_p = 6, K = 10$.

Figure 4.21: Absolute value of the sine spectrum of the functions e_j^{ss} (left side) and e_j^{ps} (right side) for three different grids. The piecewise polynomial e_j^{ss} is induced by sampling (4.51) and e_j^{ps} is made from projection (4.53). On all six images each row shows the coefficients $\langle f_j, e_i^s \rangle_{L^2(\Omega)}$ of the j'th approximate basis function. A black square indicates the value of the corresponding coefficient is 1; smaller values are shown in gray.



(a) Grid c: $N_p = 11, K = 5$.



(b) Grid d: $N_p = 51, K = 1$.

Figure 4.22: The sine spectrum of the functions e_j^{ss} (left side) and e_j^{ps} (right side) for grid c and d. See caption of Figure 4.21 for further explanation.

ure 4.21(b,c) and 4.22(a). The periodicity of the "ambiguity pattern" is related to the number of elements. Instead of having the same wavelength in all of Ω, the wavelength of e_j^{ss} with large j varies over Ω. The phenomena is not as pronounced for e_j^{ps} since the projected sine is not forced to be continuous across element interfaces (see also Figure 4.18). This wavelength variation of the sampled sines will carry over to the observation and cause mixed frequencies.

We need to specify how we determine approximations to the coefficients $\widehat{f}_i^s$ for a given DG-function f identified with its vector of nodal values $\boldsymbol{f}$. We define for $i = 1, \ldots, N$ the approximate coefficients

$$\widehat{f}_i^{\mathrm{ss}} = \langle \boldsymbol{f}, \boldsymbol{e}_i^{\mathrm{ss}} \rangle_0, \qquad \widehat{f}_i^{\mathrm{ps}} = \langle \boldsymbol{f}, \boldsymbol{e}_i^{\mathrm{ps}} \rangle_0,$$

from which we synthesize by

$$f(x) \approx \sum_{i=1}^{N} \widehat{f}_i^{\mathrm{ss}} e_i^{\mathrm{ss}}(x), \qquad f(x) \approx \sum_{i=1}^{N} \widehat{f}_i^{\mathrm{ps}} e_i^{\mathrm{ps}}(x).$$

where the inner product $\langle \cdot, \cdot \rangle_0$ is defined in (4.24).

Regardless the choice of representation technique—sampling or projection—roughly the first half of the sine basis functions appear reasonably approximated by local polynomials. This is in agreement with an asymptotic result in [HGG07] saying 4 points-per-wavelength are required to obtain exponential convergence. Judging from the number of off-diagonal elements in the spectrum images, Figure 4.21 and 4.22, and the smaller L^2-error, we should be in favor of projecting our sine basis onto the local Legendre polynomials instead of sampling. That this projection is not necessarily zero on the domain boundary could, on the other hand, be a problem to the observation of the sine basis.

Observation of the sine basis

Observation consists of (1) solution of the wave equation (4.8) and (2) approximation of the normal derivative, $\boldsymbol{C}_h$. The output matrix $\boldsymbol{C}_h$, which is based on exact differentiation of the underlying polynomial, was specified in (4.23). We will call it the *polynomial derivative*.

Results found by polynomial derivative will be compared to the approximation found by the simple first order finite difference

$$\frac{\partial y_h}{\partial n}\Big|_{\Gamma_0} \approx \frac{y_h(x^K_{N_p}) - y_h(x^K_{N_p-1})}{x^K_{N_p} - x^K_{N_p-1}} \tag{4.54}$$

where $x^K_{N_p} = 1$ is the right endpoint (the last node of element D^K), $x^K_{N_p-1}$ is its nearest neighbor and y_h is the approximate solution. Notice that since we enforce boundary conditions only weakly $y_h(x^K_{N_p})$ is not necessarily zero. We will refer to this approximation by *finite difference* or *fd. derivative*.

Before going into the actual observation, we shall assess the quality of the derivatives for the two sine approximations e^{ss}_j and e^{ps}_j. Approximation by polynomial derivative is highly accurate when applied to the sampled basis with small j as can be seen on the left side of Figure 4.24. For non-smooth functions, on the

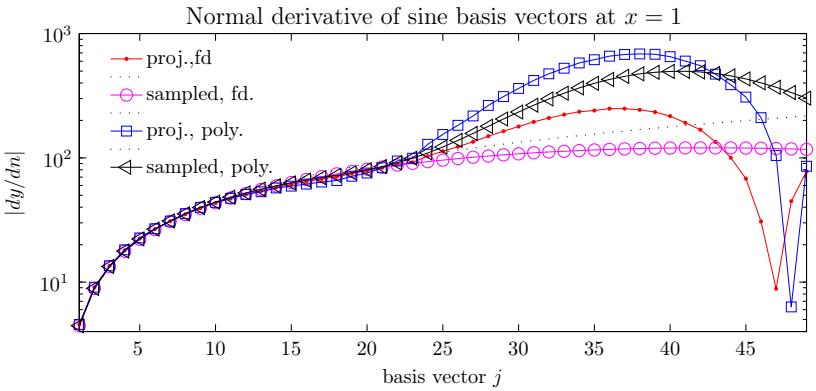


Figure 4.23: The absolute value of four different approximations to the normal derivative of the sine basis vectors $e^{\text{s}}_j, j = 1, \ldots, N$ at $x = 1$. Polynomial differentiation or first order finite differences has been used to approximate the derivative for the projected and the sampled sine basis. The dotted line is the exact derivative.

contrary, the approximation is not good. Higher order polynomial representa-

tions of non-smooth functions are well-known to exhibit undesirable behavior at endpoints. This effect is aggravated further by taking the derivative.

Figure 4.23 shows the absolute value of approximate normal derivatives found from the sampled or projected sine basis with either polynomial or first order finite difference approximation. We see how the polynomial derivatives show strong over-shooting for $j > 25$. The two approximations based on the projected basis show a dramatic drop in size for the highest basis vectors.

We find the relative error of the four approximate derivatives on Figure 4.24. It is remarkable to note that the simple first order approximation of the deriva-

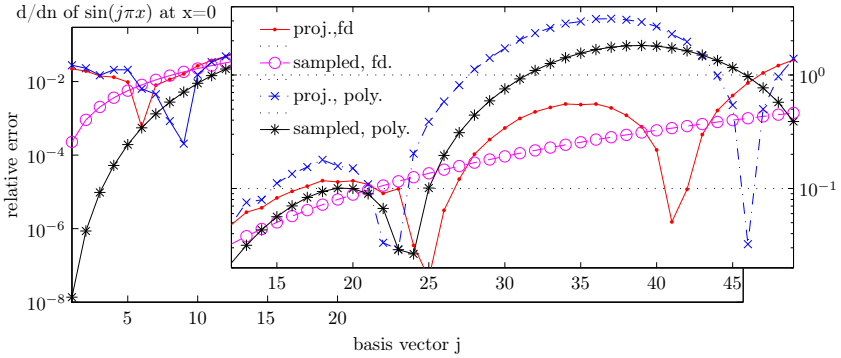


Figure 4.24: The relative error of four different approximations of $\partial/\partial n$ at $x = 1$ for the sine basis. Polynomial differentiation or first order finite differences has been used to approximate the derivative for the projected and the sampled sine basis. The inlet is a zoom of the relative error from 12 to 49 with different scaling on the ordinate axis.

tive gives the most satisfactory result after, say, $j = 25$. Notice also the disappointing accuracy for the projected basis even for quite low numbers of j.

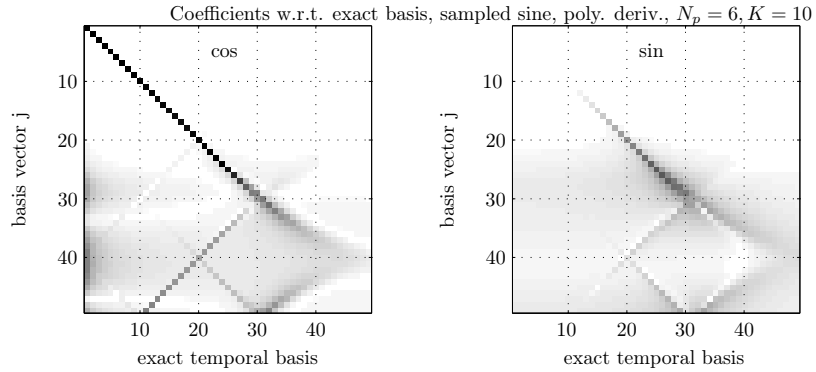
We now return to the *observation* of our discrete sine basis. We consider observation with observation time $T = 2$, and we concentrate on P^0 observation; the results for P^1 observation have been omitted as they are very similar. The computed discrete observation may be compared to the exact observation (4.43). We will do this in two ways: We first examine the "spectrum" of the discrete observation by determining the coefficients with respect to the exact temporal basis, (4.43), for each observation $P^0\boldsymbol{e}_j^{\mathrm{s}}$. Secondly, we will examine the L^2-norm of the amplitude error for each observation $P^0\boldsymbol{e}_j^{\mathrm{s}}$. We compare results obtained from the sampled basis (4.51) and the projected basis (4.53) calculated by polynomial (4.22) and fd. derivative (4.54).

Recall the exact temporal basis vectors $\boldsymbol{p}_i^0$ and $\boldsymbol{p}_i^1$ defined in (4.43) and the following coefficients of P^0 observation

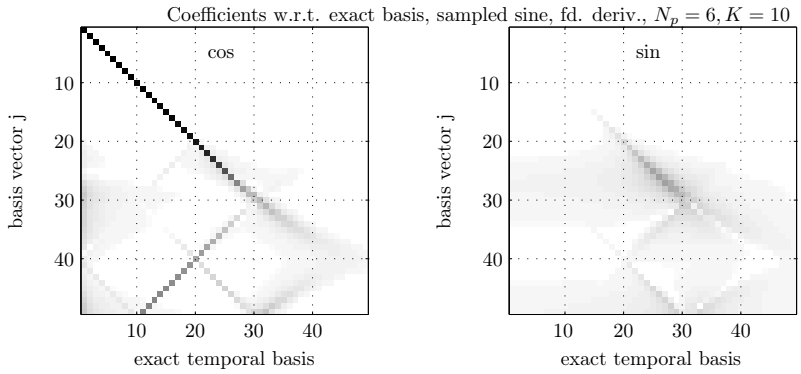
$$\beta_{ji}^{\cos} = \left\langle P^0\boldsymbol{e}_j^{\mathrm{s}}, \boldsymbol{p}_i^0 \right\rangle_{\mathcal{T}} \Big/ \left\|\boldsymbol{p}_i^0\right\|_{\mathcal{T}}^2,$$
$$\beta_{ji}^{\sin} = \left\langle P^0\boldsymbol{e}_j^{\mathrm{s}}, i\pi\boldsymbol{p}_i^1 \right\rangle_{\mathcal{T}} \Big/ \left\|\boldsymbol{p}_i^0\right\|_{\mathcal{T}}^2,$$

for $i, j = 1, \ldots, N$. Consider the spectrum images of Figure 4.25 with the coefficients $\beta_{ji}^{\cos}$ and $\beta_{ji}^{\sin}$ computed for the P^0 observation of the sampled sines $\boldsymbol{e}_j^{\mathrm{ss}}$ for $j = 1, \ldots, N$ with respectively polynomial and finite difference derivative.

The observation of about the first half of the sampled basis vectors seems satisfactory by either differentiation technique. The absolute value of the coefficients



(a) Polynomial derivative

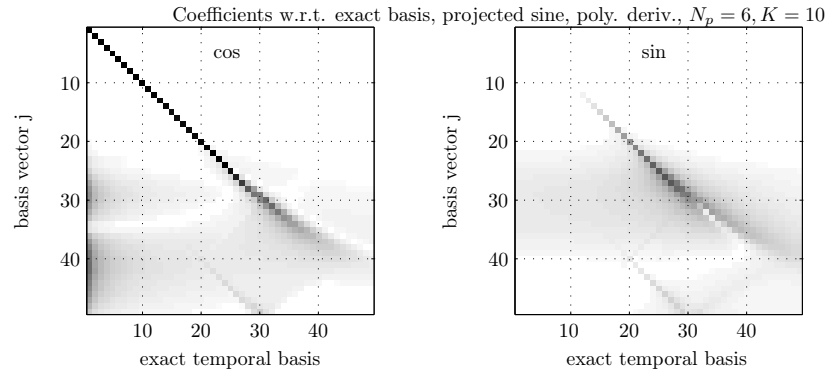


(b) Finite difference derivative

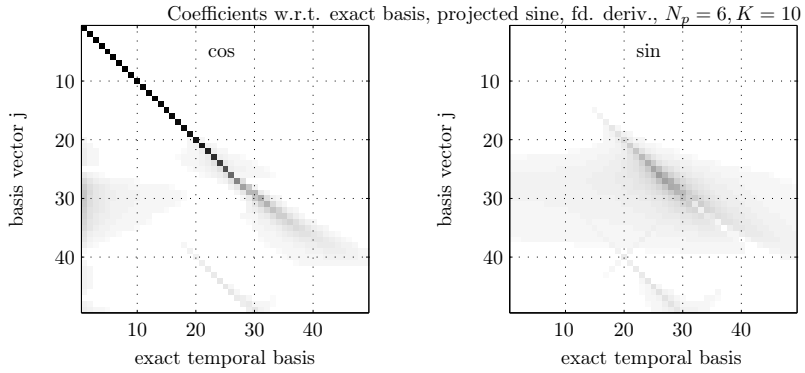
Figure 4.25: Absolute value of $\beta_{ji}^{\cos}$ and $\beta_{ji}^{\sin}$ coefficients of $P^0 \boldsymbol{e}_j^{\text{ss}}$ with respect to the exact temporal cosine basis (left) and sine basis (right) where $\boldsymbol{e}_j^{\text{ss}}$ is the **sampled** sine basis. **(a)** shows the result with polynomial derivative (4.22) and **(b)** with finite difference derivative (4.54).

for finite difference derivative (a) are lower than the ones for polynomial differentiation (b) for $j > 25$ as we would expect from Figure 4.23. Notice in both cases the very close resemblance with the pattern from Figure 4.21(c). Basis vectors with j greater than 30 appear not solely by "their own" wavenumber but as a linear combination of more than one. This combination carry over to the observation as we anticipated above. Figure 4.25 also shows a dissipative smearing effect on the high-frequency components; it is worth noticing that it happens later than we saw for L-FEM on Figure 4.7.

The frequency plots on Figure 4.26 of the discrete observations of the projected sine basis (4.53) show also a pattern inherited from their basis vectors (see right plot in Figure 4.21(c)). The ambiguity patterns are less pronounced than for the sampled basis, but the failure to represent the high-frequency observations is unfortunately intact. The space of observations should cover all frequencies up to N for $\boldsymbol{L}$ to be consistent with Λ (for N-bandlimited initial data). Both Figure 4.25 and 4.26 show, however, unfortunate lacks in the high



(a) Polynomial derivative



(b) Finite difference derivative

Figure 4.26: Absolute value of $\beta_{ji}^{\cos}$ and $\beta_{ji}^{\sin}$ coefficients of $P^0 \boldsymbol{e}_j^{\mathrm{ps}}$ with respect to the exact temporal cosine basis (left) and sine basis (right) where $\boldsymbol{e}_j^{\mathrm{ps}}$ is the **projected** sine basis. **(a)** shows the result with polynomial derivative (4.22) and **(b)** with finite difference derivative (4.54).

frequencies and just like L-FEM we loose orthogonality in the observation space.

Figure 4.27 shows the L^2-norm of the relative amplitude error for each basis vector. Notice that even small errors in frequency comes out very strongly on this plot. What is more interesting, though, is that amplitude errors of the projected and the sampled basis are almost identical in spite of their different behavior in the frequency domain. This holds for both approximate derivatives.

It seems, at this point, that the polynomial derivative delivers better results than the finite difference derivative especially for long wavelengths as we see on Figure 4.27. All approximations showed, however, the same problems for short wavelengths which we saw on the right side of all spectrum plots in Figure 4.25 and 4.26.

Matrix $\boldsymbol{L}$ by inner-product assembly

Let us construct the matrix $\boldsymbol{L}$ by inner product assembly (4.33) with $T = 2$. We need, for this purpose, the P^0 observation vectors examined above together with the corresponding P^1 observations. $\boldsymbol{L}$ is computed for both the sampled (4.50) and the projected sine basis (4.52).

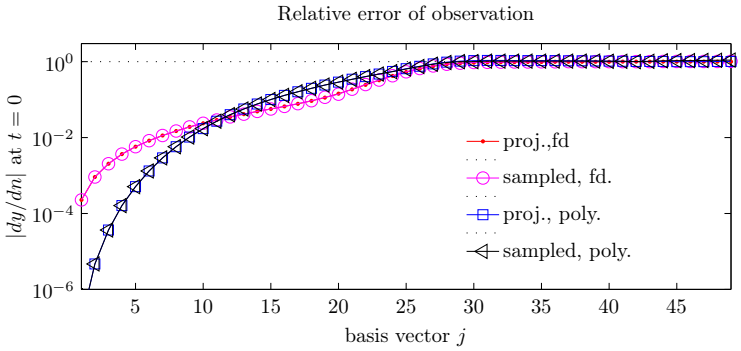


Figure 4.27: L^2-norm of the relative error of the P^0 observation of each basis vector for the sampled and projected basis with polynomial and finite difference derivative.

One way to examine the quality of $\boldsymbol{L}$ is to compare its eigenvalues with the theoretical values. See Figure 4.28 for the eigenvalues of three different approximate $\boldsymbol{L}$. It seems that matrix $\boldsymbol{L}$ computed with the projected sine basis

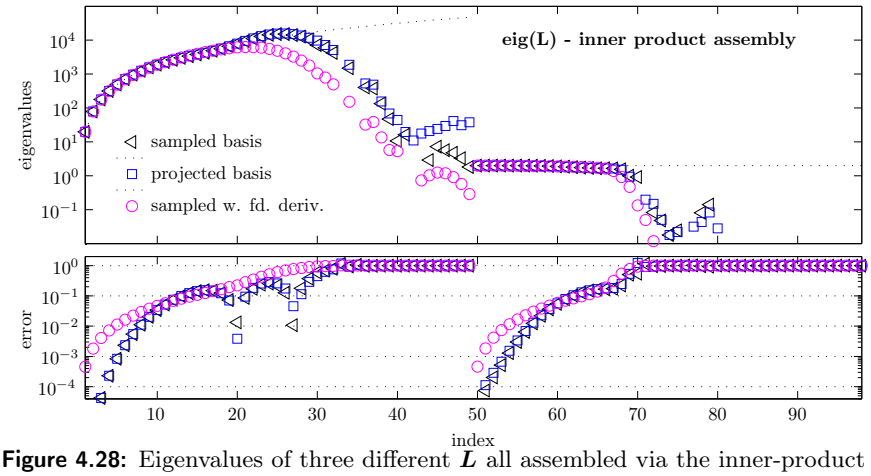


Figure 4.28: Eigenvalues of three different $\boldsymbol{L}$ all assembled via the inner-product technique shown on logarithmic scale. The values from index 1 through 49 correspond to the eigenvalues of $\boldsymbol{L}^1$ whereas index 50 to 98 correspond to the eigenvalues of $\boldsymbol{L}^4$. The dotted line shows the exact eigenvalues. The plot in the bottom shows the relative error also on logarithmic scale.

performs marginally better than the matrix constructed via the sampled basis. Both of these are determined with polynomial derivatives. The finite difference derivative clearly results in less accurate eigenvalues compared to the results obtained by polynomial derivatives (the fd. derivative for the projected basis have been omitted from the plot—it behaves very similar to the fd. derivative of the sampled basis). This poorer behavior is a consequence of the regularizing effect of the finite difference derivative—see Figure 4.23.

The first 10 eigenvalues computed with polynomial derivative are quite accurate compared to those obtained by L-FEM shown on Figure 4.9. The error

levels for moderate wavenumbers are approximately the same, around 10%, for DG-FEM and L-FEM.

Let us again study the family of reduced matrices $\boldsymbol{L}_{(N_c)}$ obtained by (4.48) where N_c is the cut-off index $1 \leq N_c \leq N$.

Consider five reduced matrices $\boldsymbol{L}_{(N_c)}$ all with cut-off $N_c = 20$, but computed on the five different grids that we introduced on page 79. The first N_c eigenvalues of $\boldsymbol{L}_{(N_c)}$, corresponding to the eigenvalues of $\boldsymbol{L}^4_{(N_c)}$, are shown on Figure 4.29; the sub-matrix is of most interest since it will be needed in Chapter 5. The

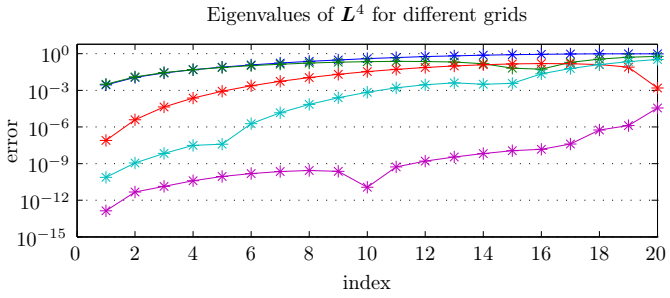


Figure 4.29: The error for the eigenvalues of the inner product assembled $\boldsymbol{L}^4_{(N_c)}$ with $N_c = 20$ for grid a (on top), b, 0, c and d (lowest) that respectively has polynomial order 1,2,5,10 and 50.

higher the order, the better the eigenvalues. The discretization with $K = 1$ polynomial of order $(N_p - 1) = 50$ gives eigenvalues with a general error level as low as 10^{-10}. This "spectral" accuracy has its price in terms of computation time; the computation takes about 10 times as long as with our standard 5'th order discretization due to restriction on the time step size by stability concerns.

The operator Λ is a positive, self-adjoint operator and its matrix approximation should therefore be symmetric and positive definite. Since $\boldsymbol{L}$ is assembled by inner-products (4.33) we are guaranteed symmetry by construction. The positive definiteness is, however, not ensured by construction. Let us use the ratio between the number of non-positive eigenvalues and the total number of eigenvalues as a measure of lack of positive definiteness. Figure 4.30 displays 1 minus this ratio for $\boldsymbol{L}_{(N_c)}$ as a function of N_c. In all three cases shown on the figure, the first negative eigenvalue for $\boldsymbol{L}_{(N_c)}$ occurs for $N_c = 34$.

Figure 4.31 shows the condition number of the reduced $\boldsymbol{L}_{(N_c)}$ as a function of the cut-off number N_c. Until $N_c = 25$ the condition number scale as the theoretical condition number $(\pi N_c)^2$ (from the exact eigenvalues), but for larger N_c it increases dramatically due to the near-zero eigenvalues of $\boldsymbol{L}^4_{(N_c)}$. This is the case for all three approaches. The corresponding condition number of $\boldsymbol{L}_{(N_c)}$ for L-FEM, which we examined in Figure 4.11, followed the theoretical value up to larger N_c.

Matrix $\boldsymbol{L}$ by direct assembly

Let us now construct $\boldsymbol{L}$ by direct assembly (4.32) for the sampled and the projected basis. We do this for polynomial derivatives alone; finite difference derivatives lead to results of poorer quality and are no longer considered. Recall that direct assembly involves a reconstruction process after observation. The

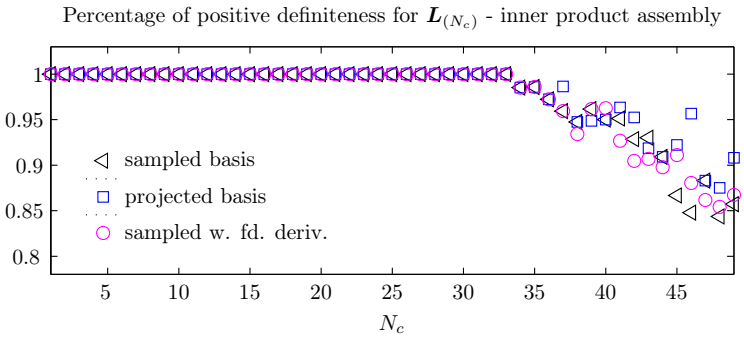


Figure 4.30: The degree of positive definiteness of $\boldsymbol{L}_{(N_c)}$ versus cut-off number N_c. The measure is simply 1 minus the ratio between the number of non-positive eigenvalues of and the total number of eigenvalues.

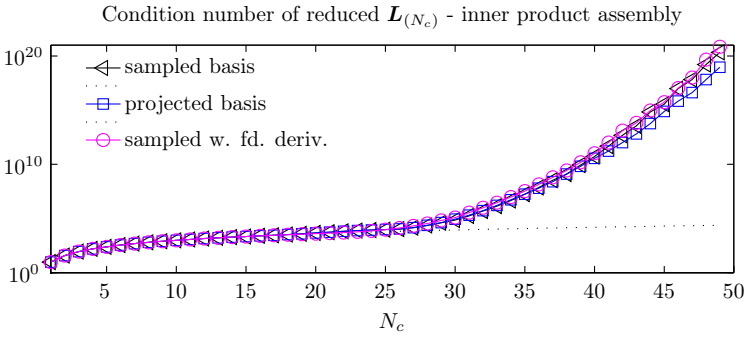


Figure 4.31: Logarithmic plot of the condition number of the reduced matrix $\boldsymbol{L}_{(N_c)}$ as function of the cut-off number N_c. The dotted line depicts the theoretical condition number justified by the corresponding exact eigenvalues.

Dirichlet data is applied through the numerical flux as discussed in Section 3.3.3, that is, in our case with an upwind flux as described by $\boldsymbol{B}_h$ in (4.22). We use the inner product approximations (4.24) and (4.26).

The continuous HUM relies on the special relationship between observation Φ and reconstruction Ψ: they are each other's adjoint. This relation was kept (at large) for approximation with L-FEM due to the "symmetry" between measuring Neumann data and applying Dirichlet data. This simple "symmetry" does not hold for DG—at least not in the DG-formulation used here. This does, however, not necessarily mean that $\boldsymbol{L}$ will not be symmetric, but that the semi-discretization does not guarantee it. The symmetry of $\boldsymbol{L}$ is important, though, not only for the sake of mimicking the properties of Λ but also if we want to solve the control problem iteratively by efficient algorithms like conjugate gradients. We will return to this in Section 4.3.

Before studying the symmetry of the constructed $\boldsymbol{L}$, we shall examine its eigenvalues in Figure 4.32. The eigenvalues are quite similar to those of the inner-product assembled $\boldsymbol{L}$ which we saw on Figure 4.28. Roughly half the eigenvalues—those corresponding to the longest wavelengths—are good whereas the rest drop dramatically in size (far below the bottom figure) causing severe

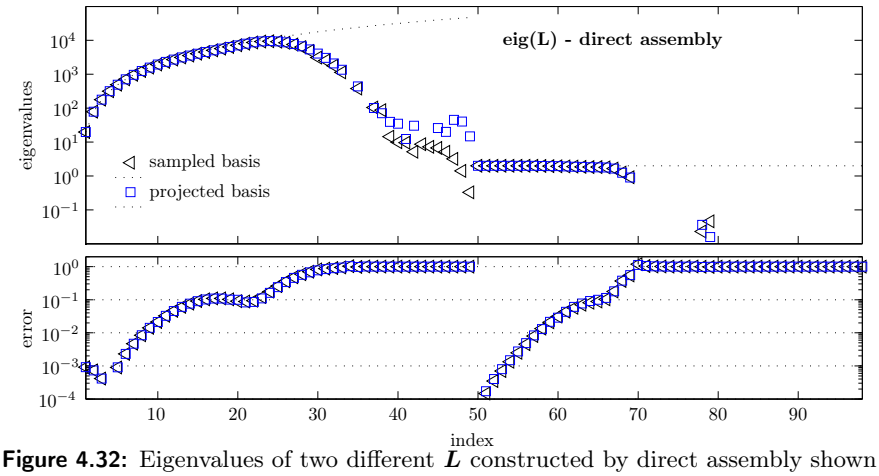


Figure 4.32: Eigenvalues of two different $\boldsymbol{L}$ constructed by direct assembly shown on logarithmic axis. The dotted line shows the exact eigenvalues. The plot in the bottom shows the corresponding relative error also on logarithmic scale.

problems for the solution of the matrix equation (4.37). When "inverted", the matrix will blow up any short wavelength component of the input. Different strategies can be applied to meet this difficulty as we discussed previously. We will once more study the family of reduced matrices $\boldsymbol{L}_{(N_c)}$.

Figure 4.33 shows a simple measure of the degree of positive definiteness. It seems almost identical to the plot for the inner-product assembled matrix on Figure 4.30. The matrices $\boldsymbol{L}_{(N_c)}$ are positive definite for all $N_c \leq 32$.

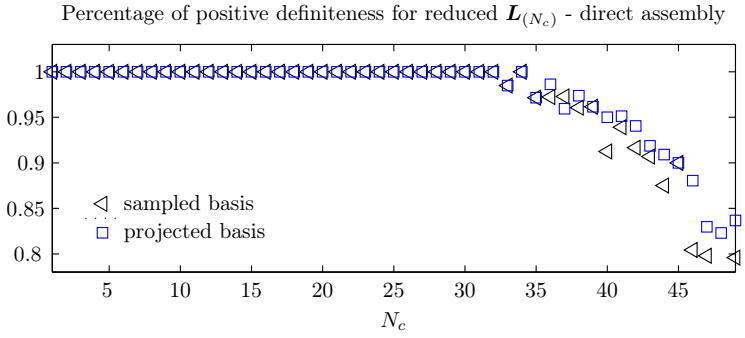


Figure 4.33: The degree of positive definiteness of $\boldsymbol{L}_{(N_c)}$ versus cut-off number N_c. The measure is 1 minus the ratio between the number of non-positive eigenvalues of and the total number of eigenvalues.

We also measure the degree of symmetry of $\boldsymbol{L}_{(N_c)}$ as shown on Figure 4.34. It is a simple measure reasoning that if a matrix $\boldsymbol{A}$ is symmetric the norm $\|\boldsymbol{A} - \boldsymbol{A}^{\mathsf{T}}\|$ is zero, and the norm is one if $\boldsymbol{A}$ is anti-symmetric $\boldsymbol{A}^{\mathsf{T}} = -\boldsymbol{A}$. The measure is

$$\text{Degree of symmetry } = 1 - \frac{\|\boldsymbol{A} - \boldsymbol{A}^{\mathsf{T}}\|}{\|\boldsymbol{A} + \boldsymbol{A}^{\mathsf{T}}\|},$$

which is 1 for symmetric matrices and 0 for the opposite case. The choice of

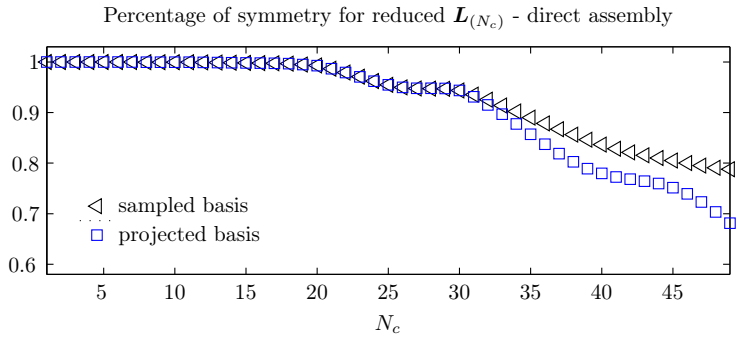


Figure 4.34: Degree of symmetry, $1 - \left\|\boldsymbol{A} - \boldsymbol{A}^{\mathsf{T}}\right\| / \left\|\boldsymbol{A} + \boldsymbol{A}^{\mathsf{T}}\right\|$, of reduced $\boldsymbol{L}_{(N_c)}$ as a function of the cut-off number N_c.

norm obviously effects the measure. We have used the largest singular value-norm.

99% symmetry is retained uptil $N_c = 20$ whereas $N_c = 25$ allows almost 95% symmetry. It decays faster from here and more so for $\boldsymbol{L}$ made from the projected basis than for $\boldsymbol{L}$ made from the sampled basis.

The condition number of $\boldsymbol{L}_{(N_c)}$ grows fast after $N_c = 25$ for increasing N_c—see Figure 4.35—again very similar to the situation for the inner-product assembled $\boldsymbol{L}_{(N_c)}$.

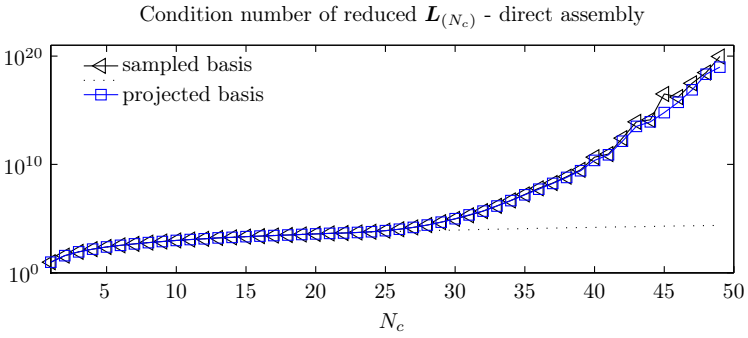


Figure 4.35: Logarithmic plot of the condition number of the reduced matrix $\boldsymbol{L}_{(N_c)}$ as function of the cut-off number N_c. The dotted line shows the theoretical condition number dictated by corresponding exact eigenvalues.

We conclude that there is only very little difference between the results obtained by inner product assembly and by direct assembly when measuring on the eigenvalues and condition numbers. The inner product assembled $\boldsymbol{L}$ is symmetric by construction. This section shows that the direct assembled $\boldsymbol{L}$ is not symmetric. We need to reduce it quite a lot to get a matrix which is almost symmetric. This lack of symmetry is an unfortunate drawback when comparing to L-FEM.

Eigenfunction controls

We shall now study the quality of the eigenfunction controls that DG-FEM produce. Let $\boldsymbol{\eta}_j$ be the control

$$\boldsymbol{\eta}_j = -P_{(N_c)} \boldsymbol{L}_{(N_c)}^{-1} \begin{bmatrix} 0 \\ -\boldsymbol{e}_j^{\mathrm{s}} \end{bmatrix}, \qquad j = 1, \ldots, N_c.$$

We consider two different discretizations, grid 0 and grid c, and use $N_c = 35$ as cut-off index in both cases. The choice of $N_c = 35$ is made for the sake of comparison with L-FEM, if this had not been in mind, the approximation of sines (Figure 4.19) and the eigenvalues of the computed $\boldsymbol{L}$ would have suggested $N_c = 30$.

The $\beta_{jk}^{\mathrm{eig}}$ coefficient for the discretization with $N_p = 6, K = 10$ (grid 0) are shown on Figure 4.36. The corresponding errors can be found on Figure 4.37.

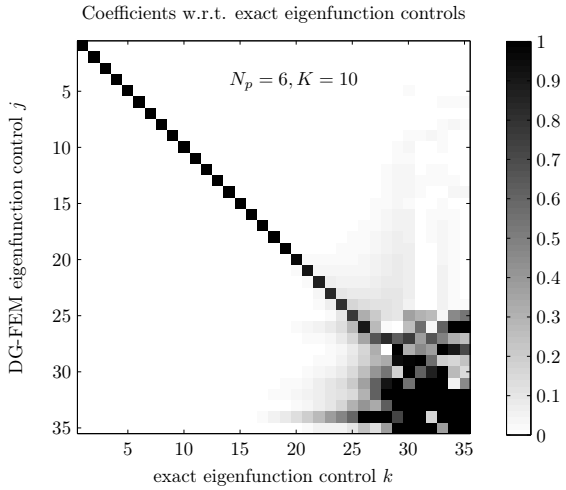


Figure 4.36: Spectrum (4.49) of the eigenfunction controls η_j for $j = 1, \ldots, N_c$ obtained by $N_p = 6, K = 10$ and $N_c = 35$.

We see that the result are quite accurate in most of the spectrum, but also that errors become large for $j, k > 25$. This suggests that a little more filtering would be a good idea here—decrease the cut-off index to an N_c between 25 and 30. Compare the images with those for L-FEM in Figure 4.12–4.15. The results for DG-FEM are clearly better. Notice also the ambiguity patterns on Figure 4.37 inherited from the representation of sines by local polynomials (see Figure 4.21).

Let us consider the control coefficients for the discretization with $N_p = 11, K = 5$ (grid c) in Figure 4.38 and their error in Figure 4.39. The higher order results in greater accuracy, yet it seems again that the filtering with N_c is insufficient.

Finally, Figure 4.40 shows the l^2 error on the coefficients for each eigenfunction control $\boldsymbol{\eta}_j$ as function of j. This plot shows the significantly lower error for the controls from grid c compared to grid 0. Though, if the plot is compared to Figure 4.16, which showed the equivalent l^2-errors for L-FEM, it becomes apparent that both sets of DG-FEM controls are much better than the L-FEM ditto. The good dispersive properties of the DG-FEM discretizations are evident.

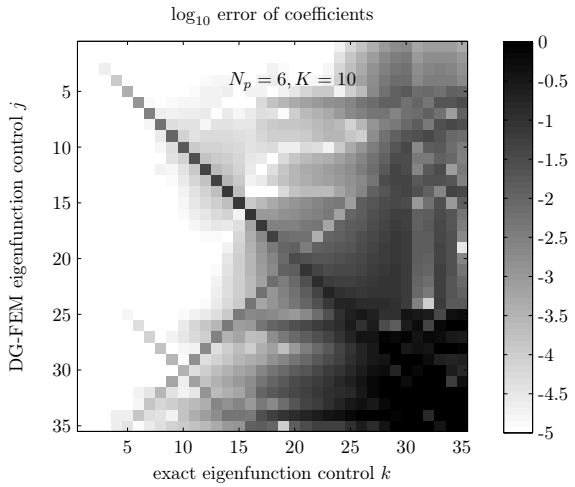


Figure 4.37: $\log_{10}$ of the error on coefficients shown on Figure 4.36 with $N_p = 6, K = 10$.

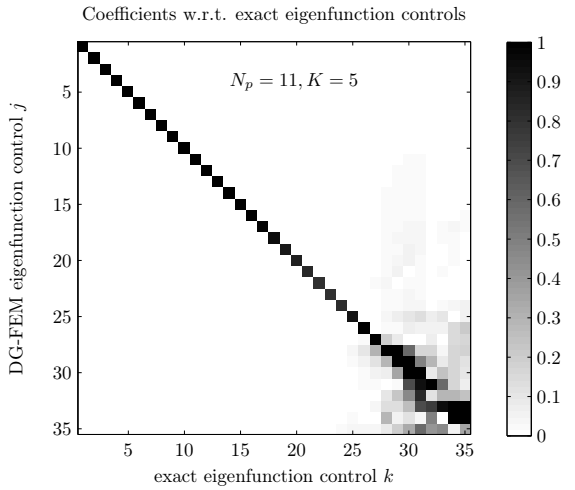


Figure 4.38: Spectrum (4.49) of the eigenfunction controls η_j for $j = 1, \ldots, N_c$ obtained by $N_p = 11, K = 5$ and $N_c = 35$.

Concluding remarks

We have used DG-FEM for the approximation of HUM boundary control in this section. We considered five different discretizations—five grids—ranging from low to high order, but all with $N = 49$ inner grid points. Two different polynomial representations of the *sinusoidal basis* were suggested: a sampled sine and a projected sine. The approximation by higher order polynomials gave rise to, what we called, "ambiguity patterns" in the frequency images. These were more pronounced for sampled than for the projected sines.

The *observation* of both sine representations was carried out with two different approximations for the normal derivate: one based on the differentiation of the underlying local polynomial (called polynomial derivate) and one based on a simple first order finite differences (called fd. derivative). The former is the most natural for the DG-formulation, but it has deficiencies when the order

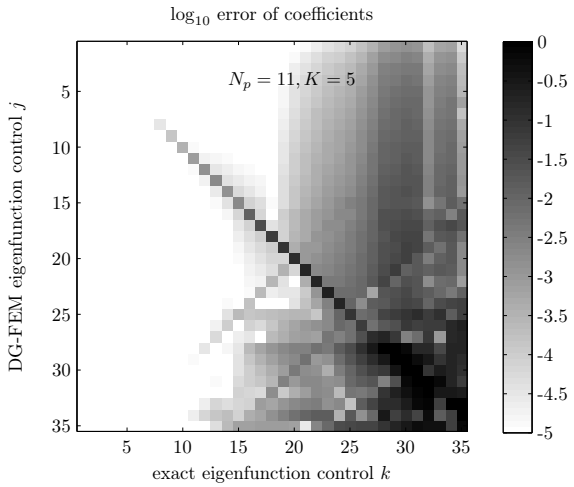


Figure 4.39: $\log_{10}$ of the error on coefficients shown on Figure 4.36 with $N_p = 11, K = 5$.

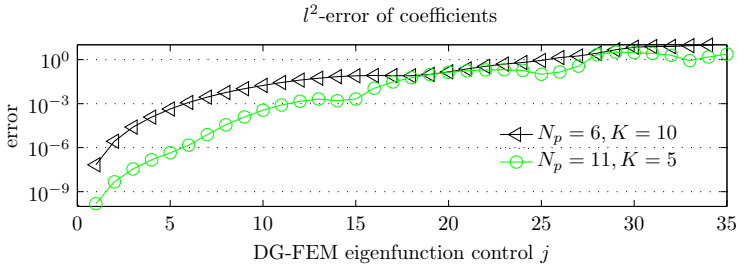


Figure 4.40: l^2 error of coefficients for each eigenfunction control $\boldsymbol{\eta}_j$ as function of the index j for two different discretizations ,respectively, $N_p = 6, K = 10$ and $N_p = 11, K = 5$ both with $N_c = 35, T = 2$.

of the local polynomial goes up—as when approximating sinusoids with moderate to high wavenumbers. However, the fd. derivative did not remedy this and was therefore rejected. The remaining investigations were carried out using polynomial derivatives; determining the derivative remains a weak link, though.

The first half of the sine observations seemed quite good—in agreement with the dispersion relation (Figure 3.9). But as the wavenumber count increased, the situation worsened. The highest frequencies were not sufficiently present (almost missing) which resulted in diminishing orthogonality of the observation space for those frequencies.

We studied also the construction of $\boldsymbol{L}$ and its properties. The first 10 eigenvalues of $\boldsymbol{L}$ were quite accurate, the next 20 were *on par* with the L-FEM results, while the remaining were bad. The convergence of the first 20 eigenvalues of $\boldsymbol{L}^4$ was demonstrated for increasing polynomial order; we even obtained "spectral" accuracy by using grid d.

Both inner product and direct assembly of $\boldsymbol{L}$ were considered. Similar results were obtained and we therefore suggest the use of inner product assembly since it requires only half the computations. It seemed that the problems, in terms of rapidly growing condition number and lack of positive definiteness etc.,

arose earlier for DG-FEM than for L-FEM. This called for a cut-off index no higher than $N_c = 30$.

The eigenfunction controls $\boldsymbol{\eta}_j$ obtained with DG-FEM were far better than those by L-FEM except for $j > 25$. For the discretization with $N_p = 6, K = 10$ the first 10 eigenfunction controls were of good quality (l^2-error less than 10^{-2}), but increasing the order to $N_p = 11, K = 5$ gives at least 15 good eigenfunction controls (l^2-error less than 10^{-3}). It should be noted that, in both these cases, a smaller cut-off index N_c (around 25-30) is necessary for DG-FEM to render reasonable results than for L-FEM (around 35).

Improvements and future work with DG-FEM for HUM

DG-FEM works quite well for HUM in the low wavenumber region, but it could be improved for the midrange wavenumbers. We suggest below a few ideas for improvement.

Prolate spheroidal wave functions. We have argued for the use of sine basis functions in relation to HUM. Global trigonometric functions are, however, only poorly approximated by local polynomials when the wavelengths are short. Prolate spheroidal wave functions [SP61] (hereafter PSWFs) may be used as an alternative to polynomials for local approximation. PSWFs are very well-suited for approximation of band-limited functions [XRY01]. They constitute a complete orthonormal basis in $L^2(-1, 1)$ like the (normalized) Legendre polynomials.

In practice, PSWFs are conveniently determined from series of Legendre polynomials—the coefficients decay superalgebraicaly [XRY01]. By replacing the local polynomial basis in the DG semi-discretization with PSWFs, the method would be much better equipped for the sine basis. Quoting [Boy04] PSWFs "oscillate more uniformly on $x \in L^2(-1, 1)$ than either Chebyshev or Legendre polynomials" which was exactly what we missed in this section for sine functions with relative short wavelengths.

Grid mapping. It is possible that the use of other interpolation points—like the ones suggested in [KTE93]—instead of the LGL points used here, could result in better resolution of trigonometric functions. In the article [KTE93] the authors introduce a set of new interpolation points reducing the extreme values of the differentiation operator and thereby allowing much larger time steps. The new points are much more equally distributed which results in better resolution properties. It is questionable, however, how much better approximations of the sine basis functions this approach could lead to. The underlying basis is *still* polynomial and one way or the other not ideal for the approximation of trigonometric functions.

Use of local polynomial basis. If we let go of the sine basis, using a combination of DG's local Legendre bases as basis for $\boldsymbol{L}$ might be a good idea. Globally continuous combinations which are restricted to zero on the boundary would possibly be a sensible choice. In this way we could still separate modes—only now the basis would be local as well. The highest modes would still be causing

trouble for the observation due to the differentiation (normal derivative at Γ_0) and the numerical dispersion.

4.3 Iterative HUM by conjugate gradients

The numerical solution of a wave equation of size N requires order MN^3 floating point operations (flops) where $M = NT/\mu$ is the number of time steps. We need to solve $2N$ or $4N$ wave equations, depending on assembly technique, to construct the matrix $\boldsymbol{L}$. As the number of DOFs increases, it quickly becomes very expensive to solve the control problem in this way. It is needless to say that solving HUM-control problems by matrix assembling in 2- and 3-d will rapidly become extremely demanding. To meet these difficulties, we need an iterative—matrix-free—HUM solution.

The pioneering work by Glowinski, Li and Lions in [GLL90] on the numerical approximation of HUM featured a conjugate gradient algorithm. Their algorithm is in fact a *preconditioned* conjugate gradient method. This was not explicitly mentioned in the paper, but a solution of a homogeneous Poisson equation works as a sensible preconditioner for the algorithm. No other iterative method for the solution of the HUM problem have been described in the literature. The algorithm from [GLL90] have, however, been used in several other works counting [GKW89], [AL98], and [CMM08].

The method of conjugate gradients [HS52] is an iterative method for solving linear problems like the generic system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

where $\boldsymbol{A}$ is a symmetric, positive definite $N \times N$ matrix. Solution by conjugate gradients is a natural choice of method for HUM as the underlying operator Λ is self-adjoint and positive. A "good" discretization should therefore lead to a symmetric and positive definite matrix. The convergence of conjugate gradients is fast if the eigenvalues of $\boldsymbol{A}$ are clustered (condition number is small), but if this is not the case, a preconditioner $\boldsymbol{M}_{\mathrm{p}}$ which "looks like" $\boldsymbol{A}$ might be helpful. The preconditioned problem

$$\boldsymbol{M}_{\mathrm{p}}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{M}_{\mathrm{p}}^{-1}\boldsymbol{b},$$

should be easier to solve and the matrix $\boldsymbol{M}_{\mathrm{p}}^{-1}\boldsymbol{A}$ better conditioned. The perfect preconditioner is $\boldsymbol{M}_{\mathrm{p}} = \boldsymbol{A}^{-1}$ but this would require solving the full direct problem. Using the diagonal of $\boldsymbol{A}$ (Jacobi preconditioning) or its approximate eigenvalues, if they are known a priori, are other more practical possibilities.

4.3.1 The algorithm

Given discrete initial data $[\boldsymbol{u}^1, -\boldsymbol{u}^0]^{\mathsf{T}}$ for the control problem (4.7), we aim at solving the preconditioned HUM problem

$$\begin{bmatrix} \boldsymbol{M}_{\mathrm{p}}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{L}^1 & \boldsymbol{L}^2 \\ \boldsymbol{L}^3 & \boldsymbol{L}^4 \end{bmatrix} \begin{bmatrix} \boldsymbol{w}^0 \\ \boldsymbol{w}^1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{M}_{\mathrm{p}}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}^1 \\ -\boldsymbol{u}^0 \end{bmatrix}, \tag{4.55}$$

where the preconditioner $\boldsymbol{M}_{\mathrm{p}}$ is an approximation to the Laplacian in a clamped domain Ω as the eigenvalues of $\boldsymbol{L}^1$ are quite similar to those of the Laplacian. As

as example will the L-FEM semi-discretization (3.22) lead to $\boldsymbol{M}_{\mathrm{p}}^{-1} = \boldsymbol{K}^{-1}\boldsymbol{M}$ where $\boldsymbol{K}$ is the stiffness and $\boldsymbol{M}$ the mass matrix.

We do not wish to *construct* the matrix $\boldsymbol{L}$, but solve (4.55) by the following algorithm.

Algorithm 4.1 (CG-HUM). *Conjugate gradient solution of HUM-problem due to [GLL90].*

$(\boldsymbol{w}^0_{[0]}, \boldsymbol{w}^1_{[0]}) \leftarrow (0,0)$
$\boldsymbol{k}_{[0]} \leftarrow P(\boldsymbol{w}^0_{[0]}, \boldsymbol{w}^1_{[0]})$ (observation)
$(\boldsymbol{z}^1, -\boldsymbol{z}^0) \leftarrow R(\boldsymbol{k}_{[0]})$ (reconstruction)
$\boldsymbol{r}^0_{[0]} \leftarrow \boldsymbol{M}_{\mathrm{p}}^{-1}(\boldsymbol{z}^1 - \boldsymbol{u}^1)$ (residual vector, preconditioning)
$\boldsymbol{r}^1_{[0]} \leftarrow -(\boldsymbol{z}^0 - \boldsymbol{u}^0)$ (residual vector)
$\gamma_0 \leftarrow \|(\boldsymbol{r}^0_{[0]}, \boldsymbol{r}^1_{[0]})\|^2_{\mathcal{X}}$ (squared residual norm)
$(\boldsymbol{e}^0_{[0]}, \boldsymbol{e}^1_{[0]}) \leftarrow (\boldsymbol{r}^0_{[0]}, \boldsymbol{r}^1_{[0]})$ (steepest descent)
$j \leftarrow 0$
while $(\gamma_j/\gamma_0) < \mathrm{tol}^2$ **do**
  $\bar{\boldsymbol{k}}_{[j]} \leftarrow P(\boldsymbol{e}^0_{[j]}, \boldsymbol{e}^1_{[j]})$ (observation)
  $(\bar{\boldsymbol{z}}^1, -\bar{\boldsymbol{z}}^0) \leftarrow R(\bar{\boldsymbol{k}}_{[j]})$ (reconstruction)
  $\bar{\boldsymbol{r}}^0_{[j]} \leftarrow \boldsymbol{M}_{\mathrm{p}}^{-1}\bar{\boldsymbol{z}}^1$ (preconditioning)
  $\bar{\boldsymbol{r}}^1_{[j]} \leftarrow -\bar{\boldsymbol{z}}^0$
  $\rho_j \leftarrow \gamma_j \,/\, \langle(\bar{\boldsymbol{r}}^0_{[j]}, \bar{\boldsymbol{r}}^1_{[j]}), (\boldsymbol{e}^0_{[j]}, \boldsymbol{e}^1_{[j]})\rangle_{\mathcal{X}}$
  $(\boldsymbol{w}^0_{[j+1]}, \boldsymbol{w}^1_{[j+1]}) \leftarrow (\boldsymbol{w}^0_{[j]}, \boldsymbol{w}^1_{[j]}) - \rho_j(\boldsymbol{e}^0_{[j]}, \boldsymbol{e}^1_{[j]})$ (update initial data)
  $\boldsymbol{k}_{[j+1]} \leftarrow \boldsymbol{k}_{[j]} - \rho_j\bar{\boldsymbol{k}}_{[j]}$ (update control)
  $(\boldsymbol{r}^0_{[j+1]}, \boldsymbol{r}^1_{[j+1]}) \leftarrow (\boldsymbol{r}^0_{[j]}, \boldsymbol{r}^1_{[j]}) - \rho_j(\bar{\boldsymbol{r}}^0_{[j]}, \bar{\boldsymbol{r}}^1_{[j]})$ (update residual vectors)
  $\gamma_{j+1} \leftarrow \|(\boldsymbol{r}^0_{[j+1]}, \boldsymbol{r}^1_{[j+1]})\|^2_{\mathcal{X}}$ (squared residual norm)
  $(\boldsymbol{e}^0_{[j+1]}, \boldsymbol{e}^1_{[j+1]}) \leftarrow (\boldsymbol{r}^0_{[j+1]}, \boldsymbol{r}^1_{[j+1]}) + (\gamma_{j+1}/\gamma_j)(\boldsymbol{e}^0_{[j]}, \boldsymbol{e}^1_{[j]})$ (search direction)
  $j \leftarrow j + 1$
end while

The action of the L operator is divided in observation by P and reconstruction by R. The main difference between this algorithm and a standard conjugate gradients algorithm is the use of the energy norm $\mathcal{X}$-norm defined in (4.3) and two residual vectors. All computations are done in $\mathcal{X}$ space—the approximation to $\mathcal{E}$—due to the preconditioning step; the residual $(\boldsymbol{r}^0, \boldsymbol{r}^1)$ would otherwise had been an element in $\mathcal{X}^\star$.

The eigenvalue distribution of the discretized Λ operator L suffer from two problems in relation to CG solution.

1. The natural distribution of eigenvalues of $\boldsymbol{L}$ lead to slow convergence. The eigenvalues of $\boldsymbol{L}^1$ scale like $(\pi n)^2$ for $1 \le n \le N$ when T is close to 2; they are approximately constant for $\boldsymbol{L}^4$. This distribution is effectively accounted for by the preconditioner in the above algorithm.

2. The effect of numerical dispersion (incorrect group velocity for short wavelength components) lead to *real* ill-conditioning of $\boldsymbol{L}$. We need a filtering or regularization procedure to account for this effect. Solution by conjugate gradients become very inefficient when $\mathrm{cond}(\boldsymbol{L})$ is huge.

The lack of clustering of the eigenvalues is, however, not the only problem.

The studies of the constructed $\boldsymbol{L}$ in Section 4.2.4 and Section 4.2.5 also revealed lack of symmetry and positive definiteness. The CG method does not even apply in such cases! In practice though, we can try anyway and for smooth data, we will be fine. If higher Fourier modes are excited, however, (*e.g.,* by noise) the algorithm will most likely diverge, and we therefore need a filtering procedure.

4.3.2 Filtering by basis truncation

With the knowledge of the appearance of the assembled $\boldsymbol{L}$ obtained in previous sections, we design a filter which counteracts both the ill-conditioning and the lack of symmetry.

We introduce the projection $\mathcal{P}_{(N_c)}$ onto the space of the first N_c sine basis functions

$$\boldsymbol{y}_{(N_c)} = \mathcal{P}_{(N_c)}\boldsymbol{y} \quad \text{defined by} \quad \boldsymbol{y}_{(N_c)} = \sum_{j=1}^{N_c} \langle \boldsymbol{y}, \boldsymbol{e}_j^{\mathrm{s}} \rangle_0 \boldsymbol{e}_j^{\mathrm{s}}, \tag{4.56}$$

where $\boldsymbol{e}_j^{\mathrm{s}}$ is the sampled sine basis ($\boldsymbol{e}_j^{\mathrm{ss}}$ for DG-FEM) and $1 \leq N_c \leq N$ is the cut-off number. Replace now the observation P and reconstruction R in the above algorithm with the filtered equivalents

$$P_{(N_c)} = P \circ \mathcal{P}_{(N_c)} \tag{4.57}$$

$$R_{(N_c)} = \mathcal{P}_{(N_c)} \circ R. \tag{4.58}$$

The filtered observation and reconstruction gives rise to a modification of Algorithm 4.1 which we call MCG-HUM.

Algorithm 4.2 (MCG-HUM). *Modified conjugate gradient solution of HUM-problem with filtering by sine basis truncation at N_c.*

[like Algorithm 4.1 with P replaced by $P_{(N_c)}$ and R replaced by $R_{(N_c)}$]

This algorithm can be applied to projected initial data $\mathcal{P}_{(N_c)}[\boldsymbol{u}^1, -\boldsymbol{u}^0]^{\mathsf{T}}$. We control hence a projection on the space of the first N_c sine basis functions.

In the following, we will often use the fraction N_c/N, *e.g.,* $1/2$ or $1/3$, to describe the threshold for the basis truncation.

4.3.3 A test problem

We introduce the function

$$f_{\mathrm{tp}}(x) = \mathbb{1}_{(0,1)} \exp(-(5(x-0.35))^6), \qquad x \in \mathbb{R},$$

where the subscript tp is for *t*est *p*roblem. Notice that f_{tp} is 0 outside $(0,1) \subset \mathbb{R}$. A plot of this function can be seen in Figure 4.41.

Let us now consider a control problem with $T = 2.4$ and the following smooth initial data

$$u_{\mathrm{tp}}^0(x) = f_{\mathrm{tp}}(x), \qquad u_{\mathrm{tp}}^1(x) = 0, \qquad x \in \Omega. \tag{4.59}$$

The exact initial data for the corresponding adjoint problem (2.6) becomes

$$\bar{\varphi}_{\mathrm{tp}}^0(x) = 0, \qquad \bar{\varphi}_{\mathrm{tp}}^1(x) = -\tfrac{1}{2} f_{\mathrm{tp}}(x), \qquad x \in \Omega, \tag{4.60}$$

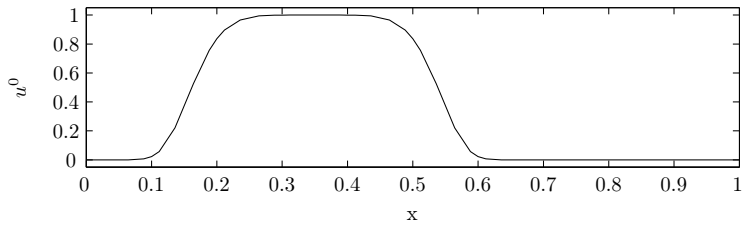


Figure 4.41: The initial data u^0 of the test problem.

and the exact HUM control is

$$\kappa_{\mathrm{tp}}(t) = \tfrac{1}{2} f_{\mathrm{tp}}(1-t) - \tfrac{1}{2} f_{\mathrm{tp}}(t-1), \qquad t \in [0,T]. \tag{4.61}$$

Figure 4.42 displays the HUM-control κ_{tp} as function of time t. The first nine

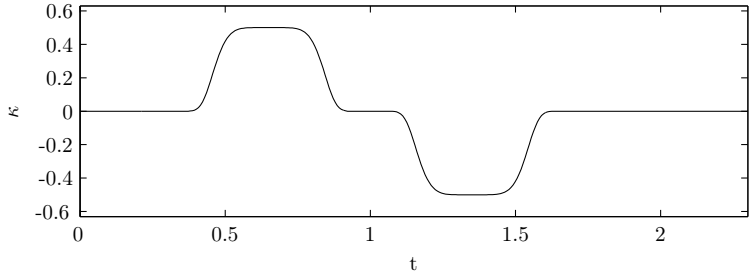


Figure 4.42: The exact HUM-control κ solving the test problem.

digits of the L^2-norm of this control is

$$\|\kappa_{\mathrm{tp}}\|_{L^2(\Sigma_0)} = 0.406572002.$$

Numerical solution (by DG-FEM)

We sample the initial data u^0 in (4.59) at the DG grid points x_i^k for $i = 1, \ldots, N_p$ and $k = 1, \ldots, K$ and collect the values in the vector $\boldsymbol{u}^0_{\mathrm{DG}}$. This nodal representation is then filtered by basis truncation (4.56). We can compare this filtered approximation with the original function (4.59); a good approximation of the initial data is necessary for good approximations of the control.

The DG-setup described on page 59 ff. provides the discretization. We use the sampled sine basis (4.51) and the energy norm (4.3) defined by DG-FEM norms (4.24) and (4.19). After discretization, the MCG-HUM algorithm is used for the numerical solution. The pre-conditioning with $\boldsymbol{M}_{\mathrm{p}}$ corresponds to solving a Poisson equation. DG-FEM is a method for solving conservation laws, but it can be modified to deal with elliptic problems as well. We refer to [HW08, page p. 265] for the details.

We say convergence is attained when the relative residual is 10^{-6}. If the algorithm converges, the obtained initial data for the adjoint problem $\{\bar{\boldsymbol{w}}^0_{\mathrm{DG}}, \bar{\boldsymbol{w}}^1_{\mathrm{DG}}\}$ and the approximate control k_{DG} may be compared with the exact functions (4.60) and (4.61). See Table 4.2 below for such comparison.

Results

Consider our favorite DG-grid with $K = 10$ elements and 5'th order polynomials, $N_p = 6$. We apply the MCG-HUM algorithm with different basis truncation factors. See Figure 4.43 for a plot of the relative residual error as function of iteration number. We observe how the solutions with 1/1 and 1/2 basis trunca-

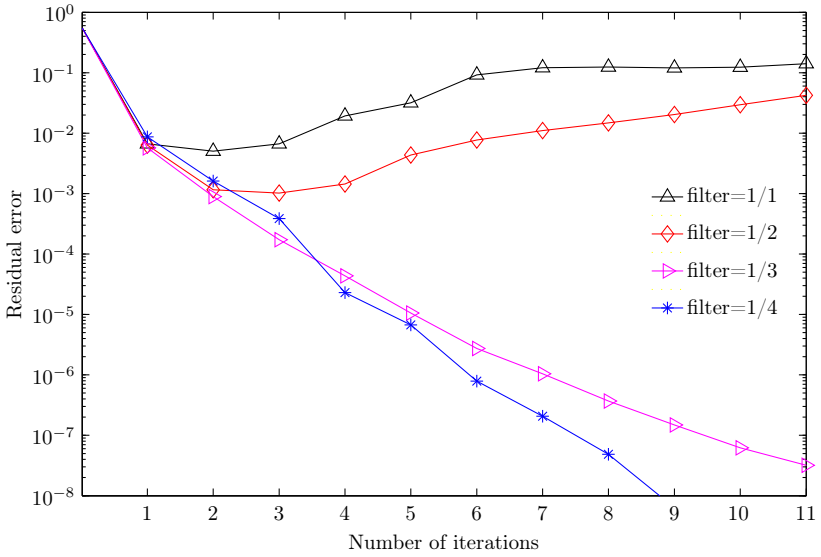


Figure 4.43: Logarithmic plot of the relative error of the residual as function of iteration number for different basis truncations with **DG-FEM** ($K = 10$ and $N_p = 6$).

tion do not converge due to lack of symmetry as we saw for the equivalent $\boldsymbol{L}_{(Nc)}$ on Figure 4.34. For stronger filters with 1/3 and 1/4 truncation the algorithm converges.

Compare Figure 4.43 with L-FEM solutions for the same filter factors on Figure 4.44. Note the different scaling of the x-axis on this plot compared to Figure 4.43; DG-FEM requires fewer iterations than L-FEM.

Convergence

We will now examine the p-convergence of the method with 1/4 basis truncation by changing the polynomial order N_p while keeping the number of elements K fixed. Figure 4.45 shows the norm of the relative residual vs. iteration number of the CG algorithm for four different polynomial orders all with 1/4 truncation. We see that the algorithm converges quite fast for all four discretizations and the higher the order, the faster the convergence. Table 4.2 shows the accuracy of the results obtained after convergence (residual $< 10^{-6}$). We have used the approximate norms $\|\cdot\|_0$ defined in (4.18) for $\|\cdot\|_{L^2(\Omega)}$ and $\|\cdot\|_{\mathcal{T}}$ defined in (4.27) for $\|\cdot\|_{L^2(0,T)}$.

We see how the sought initial data $(\bar{w}^0_{\mathrm{DG}}, \bar{w}^1_{\mathrm{DG}})$ for the adjoint system clearly converges towards $(\bar{\varphi}^0_{\mathrm{tp}}, \bar{\varphi}^1_{\mathrm{tp}})$. The approximate controls k_{DG} converges, likewise, towards the exact control κ_{tp}.

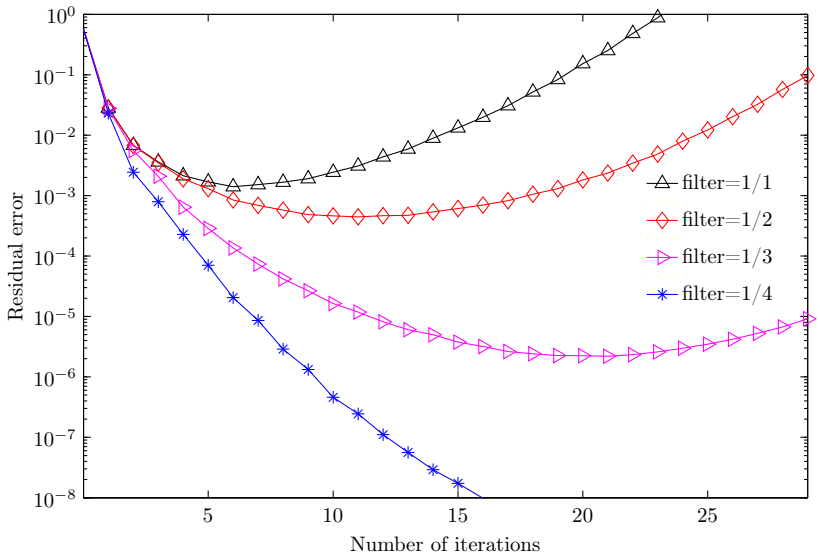


Figure 4.44: Logarithmic plot of the relative error of the residual as function of iteration number for different basis truncations with **L-FEM** ($N = 49$).

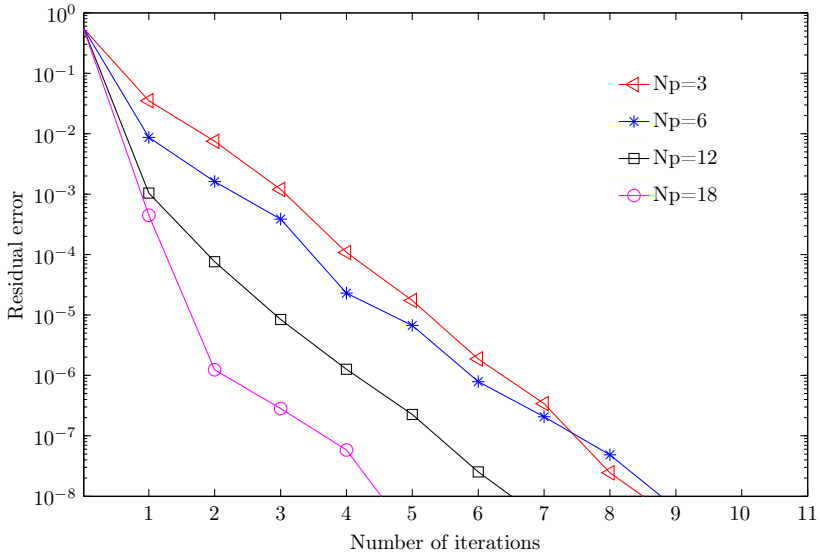


Figure 4.45: Logarithmic plot of the relative error of the residual as function of iteration number for DG-FEM solutions ($K = 10$) with four different polynomial orders N_p. All solutions have been filtered with 1/4 basis truncation.

4.4 Concluding remarks

The proceeding section gives a short review on related work. Hereafter, Section 4.4.2 discusses the results obtained in this chapter.

Table 4.2: CG-algorithm results with 1/4 basis truncation for DG-FEM with $K = 10$ and different polynomial order N_p.

	$N_p = 3$	$N_p = 6$	$N_p = 12$	$N_p = 18$
# iterations	7	6	5	3
$\frac{\left\|u^0_{\mathrm{tp}} - u^0_{\mathrm{DG}}\right\|_{L^2(\Omega)}}{\left\|u^0_{\mathrm{tp}}\right\|_{L^2(\Omega)}}$	2.902e-02	2.010e-04	1.261e-07	2.741e-11
$\left\|\bar{w}^0_{\mathrm{DG}}\right\|_{L^2(\Omega)}$	9.137e-04	2.036e-04	4.313e-05	1.855e-05
$\frac{\left\|\bar{\varphi}^1_{\mathrm{tp}} - \bar{w}^1_{\mathrm{DG}}\right\|_{L^2(\Omega)}}{\left\|\bar{\varphi}^1_{\mathrm{tp}}\right\|_{L^2(\Omega)}}$	1.637e-01	4.526e-02	6.891e-04	1.383e-05
$\frac{\left\|\kappa_{\mathrm{tp}} - k_{\mathrm{DG}}\right\|_{L^2(0,T)}}{\left\|\kappa_{\mathrm{tp}}\right\|_{L^2(0,T)}}$	1.787e-01	4.640e-02	1.253e-03	4.467e-04
$\left\|k_{\mathrm{DG}}\right\|_{L^2(0,T)}$	0.3999170	0.4061986	0.406575	0.4065726

4.4.1 Related work

R. Glowinski, J.L. Lions and C.H. Li were the authors of the pioneering work [GLL90] on the numerical approximation of HUM. The paper presented a conjugate gradient algorithm, described in Section 4.3.1 above, and a bi-grid approach for filtering out the spurious high-frequency solutions. They used a 2nd order central FDM (equal to (3.10)) for their semi-discretization. They exposed the numerical approximation of HUM as a difficult and very sensitive problem with bad asymptotic behavior. In [Glo92], Glowinski extended the algorithm to include a Tychonoff regularization procedure.

E. Zuazua has later been one of the main characters in the further development of numerical HUM. Zuazua and co-workers, S. Micu, J. I. Infante, C. Castro, and M. Negreanu among many others, have put much focus on the mathematical analysis of semi-discrete schemes in relation to HUM.

The lack of so-called *uniform observability* has been pointed out as the main problem. Discrete versions of the observability inequality (2.16) are analyzed for $h \to 0$; for the standard FDM scheme the related observability constant $C \to \infty$ as $h \to 0$. This diverging behavior is denoted non-uniform observability. In the important paper [IZ99] Infante and Zuazua introduced the use of a discrete version of Ingham's inequality in the proof of non-uniform observability and in the recovering of uniform observability after filtering. For another example of theoretical treatment of numerical HUM, see Micu's use of bi-orthogonal sequences in [Mic02].

Zuazua's group has also refined the use of bi-grid procedures in the works [NZ03] and [NZ04a]. The bi-grid algorithm involves the use of a second, coarse grid for some of the computations thereby ruling out the high-frequency waves.

Some focus have been given to the mixed FEM in relation to numerical HUM recently, although already used in [GKW89]. Castro and Micu have analyzed the benefits of using mixed FEM for HUM in [CM06]. The mixed FEM has a dispersion relation which is very well-suited for control since the high-frequency components travel at higher instead of lower than the correct speed.

The work by A. Münch, [Mün04] and [Mün05], inspired by conservative schemes of S. Krenk mentioned in Section 3.4, shows how a family of implicit

schemes for the 1-d wave equation can be used to obtain uniform controllability. The methods are, however, very specific for 1-d problems.

More practical approaches to the numerical approximation of HUM have been sparse. Notable are, nevertheless, the optimization based approaches of M. Gunzburger and co-workers in [JGH03] and [GHJ06], the wavelet filtering technique with computations presented by M. Negreanu and co-workers in [NMS06], and the Ph.D. dissertation of J. M. Rasmussen [Ras04]. The latter contributed some computational efforts with detailed numerical considerations.

4.4.2 Discussion

This chapter began with the semi-discretization of HUM which also marked the transition from *in*finite dimensional control to *finite* dimensional control. We worked out the details for two schemes: L-FEM and DG-FEM, and continued with the full discretization of HUM leading to the discrete operator L as approximation to the fundamental HUM-operator Λ.

Section 4.2 was about constructing L as a matrix $\boldsymbol{L}$ in order to solve the discretized HUM problem directly. We presented two different ways of assembling $\boldsymbol{L}$ and introduced a set of sinusoids as basis. Yet not previously described in the literature, using a sine basis is natural for two reasons: 1) sinusoids are eigenfunctions to the HUM problem, 2) sinusoids are closely linked to the dispersion relation of the scheme which, more or less, defines its ability to deal with control. Trigonometric functions are, on the other hand, not well approximated by polynomials, especially not for short wavelengths. Both semi-discretizations studied in this chapter are build from local polynomials.

The construction of $\boldsymbol{L}$ from L-FEM semi-discretization was the topic of Section 4.2.4. This section showed how the use of the sine basis clarified the close relationship between the dispersion relation and the numerical HUM. By studying the spectrum of the observation of the sine basis, we could see the effect of numerical dispersion for each wavelength component. Changing the Courant number had a significant effect on the observation. The main problem was the diminishing orthogonality of the "highest" directions in the observation space. A consequence was the exponential growth of the condition number of $\boldsymbol{L}$ and that even very small amounts of noise would blow up and ruin the solution of the control problem. A filter was needed. We used a filter based on sine basis truncation which was another benefit of the use of the sine basis. We also computed the eigenfunction controls for L-FEM with two different Courant numbers, and by studying their spectra, we found that they were shaped after their numerical phase velocities. The controls computed with the smallest Courant number (finest temporal resolution) were not as good as those computed with larger Courant number (coarser temporal resolution). We concluded that numerical phase velocity is very important for control.

Section 4.2.5 dealt with the construction of $\boldsymbol{L}$ with DG-FEM semi-discretization. We compared two different ways of representing the sine basis: a *nodal* based on sampling and a *modal* based on projection. The latter had the smallest approximation error in the L^2-norm since the Legendre polynomials, which are the modal basis functions, constitute an orthonormal basis in L^2. The use of a higher order polynomial DG-basis gave some additional challenges in terms of wavenumber ambiguity when representing the sine functions. The problems

were most pronounced for the sampled sines due to the irregular sampling on the LGL grid. It was problematic as *un*ambiguous observation is important for the control—if we do not know which waves we are seeing, then we cannot expect to control them. The higher order polynomial basis allowed accurate computation of the derivative, which we needed for observation, but for low wavenumbers *only*. The normal derivatives of short wavelength sinusoids are only poorly approximated by high order polynomials as is well-known.

The spectrum of the observation of each basis vector was examined for DG-FEM, too. This showed a clear improvement over the L-FEM approximations for the first half of the sine functions. The last sine functions of short wavelength, however, gave poorer results than with L-FEM. It seems that the high order polynomial derivative could have something to do with this, although our attempt to use a low order approximation gave even worse results. Once more the consequence was vanishing orthogonality in the observation space for the higher wavenumbers which again threatened the control.

In the end after constructing $\boldsymbol{L}$ with respectively the sampled and projected sine basis, only very little difference showed. We chose the sampled sine for its simplicity. The matrix $\boldsymbol{L}$ lacked the symmetry that we would prefer it to inherit from Λ. We introduced a family of reduced matrices $\boldsymbol{L}_{(N_c)}$ corresponding to the use of a reduced sine basis, that is, filtering by sine basis truncation, and we restored thereby symmetry. We considered the eigenvalues of $\boldsymbol{L}$ and found that DG-FEM provided accurate results for the eigenvalues that corresponded to low wavenumbers, moderate accuracy was obtained for midrange wavenumbers, the remaining tended to zero. We demonstrated the convergence for the eigenvalues by increasing the polynomial order. It was possible to obtain "spectral" accuracy for the first 20 eigenvalues with grid d. We also computed eigenfunction controls which were superior to those obtained by L-FEM even though a little more filtering was needed.

The HUM problem was also solved iteratively by a conjugate gradients algorithm in Section 4.3. Here we could use the knowledge of the properties of $\boldsymbol{L}$ obtained in the previous sections: eigenvalue distribution, condition number, and lack of symmetry. We proposed a filtering step in the algorithm based on projection onto a reduced set of sine basis functions. We examined different filter factors for a DG-FEM discretized test problem and saw the convergence of the sufficiently filtered algorithm. It seemed that a relatively strong filter, *e.g.,* keeping only the $1/3$ lowest modes, is necessary for convergence.

The DG-FEM discretization gave good results for sines with low wavenumbers, but it needs improvement for waves with midrange wavenumbers. In this region it is comparable to L-FEM—especially if we take in to account that L-FEM is much simpler and only of low order. A promising idea is to replace the local polynomial basis in DG-FEM with prolate spheroidal wave functions (PSWF) which are much better suited for the approximation of sine waves.

In this chapter, we have argued for the use of the sine basis for numerical HUM. This is of course a special possibility for the 1-d problem which does not carry over to higher dimensional control. This does not, however, make the use of the sine basis irrelevant for 1-d problems. It proved very useful for shedding light on the nature of the problem and, in particular, on the connection with numerical dispersion. We can still use the idea of using a modal basis in multi-dimensional problems, *e.g.,* with a PSWF basis. The combination of a local

and a modal basis, which DG-FEM allows, could prove very strong for control problems.

CHAPTER 5

The inverse problem—an application of HUM

So far this dissertation has been about HUM boundary control for the wave equation. This chapter goes in another direction and consider an inverse source problem for the wave equation. We shall later see how HUM can be used for its solution, but let us first introduce the notation of inverse problems.

In the words of J.B. Keller, "two problems are inverses of one another if the formulation of each involves all or part of the solution of the other" [Kel76]. Nowadays, an inverse problem is often specified as a problem of "determining causes for a desired or observed effect" [EHN96]. A certain pattern can be recognized: An operator F maps some "model parameters" x into some "data" y and from this we may formulate two problems

Forward:	Given x, evaluate $F(x)$,
Inverse:	Given y, solve $F(x) = y$ for x.

In this chapter, we are concerned with an inverse source problem of determining an external force x from boundary measurements y. Applying the operator F corresponds, in this case, to solving a wave equation. The problem is called

an inverse *source* problem when x is an external source and not actual *model* parameters.

This chapter is build on the paper "*Stability, reconstruction formula and regularization for an inverse source hyperbolic problem by a control method*", [Yam95], by M. Yamamoto. In this paper, M. Yamamoto presented an inverse source problem for the wave equation and showed how it could be dealt with in a unified manner by HUM. The problem's source term may be separated in a spatial part f and a temporal part σ. The inverse problem consists of finding the unknown f from boundary measurements for given σ.

Yamamoto's method is interesting since it inherits the generality of HUM, meaning that it is applicable for multi-dimensional problems and has great potential for dealing with inverse problems for a wide range of PDEs (see, *e.g.,* [Nic00] and [NZ04b] for results for, respectively, Maxwell's equations and vibrating beams).

The goal of this chapter is the numerical approximation of the reconstruction[1] of the source term's spatial part f. The solution relies on a set of HUM controls for an auxiliary problem. We know, however, that finding numerical HUM controls is difficult, but how does the problems from numerical HUM effect the reconstruction? We wish to examine whether our numerical HUM controls found with L-FEM and DG-FEM in Chapter 4 can be used for the reconstruction and how the use of DG-FEM controls compare to the use of L-FEM controls. We shall also examine the degree of ill-posedness of the problem and, if necessary, apply regularization. How the temporal part σ effect the problem and its reconstruction will be assessed, too.

After introducing the inverse problem in more detail in Section 5.1, we present Yamamoto's results concerning stability, reconstruction and regularization in Section 5.2.

We suggest a discretization of the reconstruction in the 1-d case in Section 5.3. This discretization is followed by a numerical study in Section 5.4. We cover the generation of reliable data in Section 5.4.1 and estimate the degree of ill-posedness by studying the singular values of the forward map in Section 5.4.2.

The reconstruction formula relies on the use of eigenfunction controls obtained by HUM. Section 5.4.3 presents numerical results with analytic HUM controls. Section 5.4.4 and Section 5.4.5 present results with numerical HUM controls obtained after, respectively, L-FEM and DG-FEM semi-discretization. The study is finalized in Section 5.4.6 with the reconstruction of 25 random coefficients by analytic, L-FEM, and DG-FEM controls. We end this chapter by a short discussion in Section 5.5.

Briefly on the geometry and notation

In this chapter, we shall first consider the general case Ω being an open, bounded domain in $\mathbb{R}^d$ with boundary Γ. As in Chapter 2, Σ is the time-boundary cylinder $\Sigma = (0,T) \times \Gamma$. The observation boundary is denoted Γ_0, and the

[1] It should be noted that *reconstruction* in this chapter is not the same as reconstruction in the previous chapters. In this chapter, reconstruction means obtaining the spatial part of a source term, and in particular its Fourier coefficients, from boundary data.

corresponding part of Σ is $\Sigma_0 = (0,T) \times \Gamma_0$. To shorten notation we introduce

$$\partial_{\Gamma_0} v = \frac{\partial v}{\partial n}\Big|_{\Gamma_0}$$

The numerical study in Section 5.3 and 5.4 is carried out in the 1-d case $\Omega = (0,1)$ with $x = 1$ as observation boundary.

5.1 An inverse source problem

Let $v = v_f$ be the solution to the (forced) wave equation

$$v'' - \Delta v = \sigma(t) f(x), \qquad \text{in } (0,T) \times \Omega, \tag{5.1a}$$
$$v(t,x) = 0, \qquad (t,x) \text{ on } \Sigma, \tag{5.1b}$$
$$v(0,x) = 0, \quad v'(0,x) = 0, \qquad x \text{ in } \Omega, \tag{5.1c}$$

for given $f \in L^2(\Omega)$ and $\sigma \in C^1[0,T]$ with $\sigma(0) \neq 0$.

We have the following existence and uniqueness result [Yam95].

Theorem 5.1. *If $\sigma \in C^1[0,T]$, then for any $f \in L^2(\Omega)$ there exists a unique weak solution to (5.1) with regularity*

$$v \in C^1([0,T]; H_0^1(\Omega)) \cap C^2([0,T]; L^2(\Omega)), \tag{5.2}$$
$$\frac{\partial v}{\partial n} \in H^1((0,T); L^2(\Gamma)). \tag{5.3}$$

Moreover, $\exists c > 0$ such that

$$\|(v, v')\|_{L^\infty((0,T); H_0^1(\Omega) \times L^2(\Omega))} \leq c \|f\|_{L^2(\Omega)}. \tag{5.4}$$

and

$$\left\| \frac{\partial v_f}{\partial n} \right\|_{H^1((0,T); L^2(\Gamma))} \leq c \|f\|_{L^2(\Omega)} \tag{5.5}$$

for some constant $c > 0$ independent of f. □

We introduce, for $\Gamma_0 \subset \Gamma$, the following subspace of $H^1((0,T); L^2(\Gamma))$

$$\mathcal{B}^1 := H^1((0,T); L^2(\Gamma_0)), \tag{5.6}$$

which we equip with the H^1 inner product

$$\langle u, v \rangle_{\mathcal{B}^1} = \int_0^T \int_{\Gamma_0} \Big(u(t,x) v(t,x) + \frac{\partial u}{\partial t}(t,x) \frac{\partial v}{\partial t}(t,x) \Big) \mathrm{d}x \mathrm{d}t, \tag{5.7}$$

for all $u, v \in \mathcal{B}^1$; we use the norm induced by the inner product $\|u\|_{\mathcal{B}^1}^2 = \langle u, u \rangle_{\mathcal{B}^1}$.

We are now ready to define the inverse problem.

Definition 5.2 (Inverse source problem). Let $\sigma \in C^1[0,T]$ be given and let $v = v_f$ be the solution to (5.1) for some unknown $f \in L^2(\Omega)$. Then we define the *inverse source problem*:

$$\text{For given data } \partial_{\Gamma_0} v_f \in \mathcal{B}^1 \text{ for (5.1), find } f. \tag{ISP}$$

□

The (ISP) asks whether we can determine the spatial part f of the source term $\sigma(t)f(x)$ for the system (5.1) by the additional information $\partial_{\Gamma_0} v_f$ on a part Γ_0 of the boundary. This problem is the inverse compared to the forward problem: from the known source $\sigma(t)f(x)$ determine the solution v to (5.1). We introduce the map G

$$G\colon L^2(\Omega) \to \mathcal{B}^1 \quad \text{defined by} \quad G(f) = \partial_{\Gamma_0} v_f, \tag{5.8}$$

and summarize the above in the following two lines:

Forward:	Given f, evaluate $G(f)$,
Inverse:	Given $\partial_{\Gamma_0} v$, solve $G(f) = \partial_{\Gamma_0} v$ for f.

Notice how the existence, uniqueness and stability of the forward problem is ensured by Theorem 5.1.

5.1.1 Examination of the inverse problem

We wish to address the following four aspects of the inverse problem.

I. **Identifiability**. The existence of a solution is ensured by considering the data $\partial_{\Gamma_0} v_f \in G(f)$. But what about uniqueness—is the solution to (ISP) unique?

II. **Stability**. Does the solution f depend continuously on the data $\partial v/\partial n$ on Σ_0? Can we estimate the L^2-norm of f by some norm on the boundary data $\partial_{\Gamma_0} v_f$?

III. **Reconstruction**. How can we determine the Fourier coefficients of f in terms of the data $\partial_{\Gamma_0} v_f$ and thereby give an explicit formula for the reconstruction of f?

IV. **Regularization**. The problem is ill-posed. What can we do to regulaize solutions?

The existence and uniqueness of an inverse problem is a little different from dealing with the same questions for the corresponding forward problem. Firstly, we note that we consider only data in the image $G(f)$ of the forward map G. For this reason, existence of solutions is no issue. Secondly, uniqueness of solutions is obtained by M. Yamamoto as a byproduct of the stability analysis (see Section 5.2.1).

We will return to questions II–IV in Section 5.2.

5.1.2 An auxiliary inverse 'initial data' problem

In order to use HUM for the inverse problem (ISP), we will now introduce an auxiliary inverse problem; it will prove very useful in the subsequent sections.

Let $w = w_f$ be the solution, for given $f \in L^2(\Omega)$, to the system

$$w'' - \Delta w = 0, \qquad \text{in } (0,T)\times\Omega \tag{5.9a}$$
$$w(t,x) = 0, \qquad (t,x) \text{ in } \Sigma \tag{5.9b}$$
$$w(0,x) = 0, \quad w'(0,x) = f(x), \qquad x \text{ in } \Omega, \tag{5.9c}$$

which is equal to system (2.6) with $\varphi^0 = 0$ and $\varphi^1 = f$. According to Theorem 2.8 system (5.9) has a unique solution for which $(w, w') \in C([0,T]; H_0^1(\Omega) \times L^2(\Omega))$ and the Neumann data $\frac{\partial w}{\partial n} \in L^2((0,T) \times \Gamma)$.

As the L^2 counterpart of (5.6) we now introduce for $\Gamma_0 \subset \Gamma$

$$\mathcal{B}^0 := L^2((0,T) \times \Gamma_0), \tag{5.10}$$

which is equal to $\mathcal{B}$ of Chapter 2 (we use the superscript 0 in this chapter to emphasize its $L^2 = H^0$ nature compared to the H^1 ditto of $\mathcal{B}^1$).

We note that we can express the solution of (5.9) in terms of the eigensolutions $\{\lambda_k, \phi_k\}_k$, $k \in \mathbb{N}$, by

$$w(t,x) = \sum_{k=1}^{\infty} \langle f, \phi_k \rangle_{L^2(\Omega)} \frac{\sin \lambda_k t}{\lambda_k} \phi_k(x). \tag{5.11}$$

This Fourier series solution will be important for the reconstruction of (IIDP).

For system (5.9) we define the auxiliary inverse problem (IIDP).

Definition 5.3 (inverse 'initial data' problem). Let $w = w_f$ be the solution of (5.9) for some unknown $f \in L^2(\Omega)$. Then we define the *inverse 'initial data' problem*:

$$\text{Given the data } \partial_{\Gamma_0} w_f \text{ for (5.9), find } f, \tag{IIDP}$$

where $\partial_{\Gamma_0} w_f \in \mathcal{B}^0$ is the available Neumann data on Σ_0. □

Or in plain words: can we determine the unknown initial velocity $w'(0,\cdot) = f$ from the additional information $\partial_{\Gamma_0} w_f$?

The claimed relation between (ISP) and (IIDP) is supported by the ensuing proposition which connects the solution of (5.1) to the solution of (5.9).

Proposition 5.4. *Let $\sigma \in C^1(0,T)$ and let v_f be the solution of (5.1) and w_f the solution of (5.9) for $f \in L^2(\Omega)$. Then we have*

$$v_f(t,x) = \int_0^t \sigma(s) w_f(t-s,x) \mathrm{d}s, \quad \text{for } t > 0, x \in \Omega. \qquad \square$$

Proof. The result can be established for $f \in C_0^\infty(\Omega)$ by applying *Duhamel's principle* (see F. Johns classic [Joh82, p.135], also referenced in [Yam95]). It is extended to hold for any $f \in L^2(\Omega)$ by approximating f with a sequence of C_0^∞-functions and using the L^∞ estimates (5.4) and (2.8). ■

5.2 A HUM solution to the inverse problem

The primary message of M. Yamamoto's paper [Yam95] was that HUM can be used for dealing with the stability, reconstruction and regularization of the inverse problem (ISP) in a unified manner. In fact, HUM is used for the solution of (IIDP) which by boundary integral operators is connected to (ISP).

We shall consider the stability of (IIDP) and (ISP) in Section 5.2.1, then reconstruction in Section 5.2.2, and finally regularization in Section 5.2.3.

5.2.1 Stability

To ensure stability of the inverse problem we seek a bound on the size of f by a suitable norm on the Neumann data $\frac{\partial u(f)}{\partial n}$ on Γ_0. HUM provides such bounds and also conditions on the observation time T and the size of the observation boundary $\Gamma_0 \subset \Gamma$.

Stability of (IIDP)

Stability of the inverse problem (IIDP) is a direct consequence of the observability of (5.9).

Proposition 5.5 (Stability of (IIDP)). *Let system* (5.9) *be observable, i.e., it satisfies the observability inequality* (2.16). *Given* $\partial_{\Gamma_0} w_f \in \mathcal{B}^0$ *we have the following estimate*

$$c^{-1} \left\| \partial_{\Gamma_0} w_f \right\|_{\mathcal{B}^0} \leq \left\| f \right\|_{L^2(\Omega)} \leq c \left\| \partial_{\Gamma_0} w_f \right\|_{\mathcal{B}^0},$$

where $c > 0$ *is a constant.* □

Remark 5.6. The inequality on the right side is the so-called *observability inequality* for the system (5.9). The left side inequality is a regularity result—the "hidden" regularity (see (2.15))—for the wave equation. The key point is that mapping $f \mapsto \left\| \partial_{\Gamma_0} w_f \right\|_{\mathcal{B}^0}$ constitutes a norm on the initial data (Proposition 2.15) which, in this case, is $(0, f)$. ■

The observability of system (5.9) puts requirements on the observation time T due to the finite speed of propagation of waves. Furthermore, conditions on the size of Γ_0—see Section 2.1.3—are also consequences of HUM.

It remains to connect this result with the stability of (ISP).

Stability of (ISP)

We are closing in on the main stability theorem, but first we need a few lemmas needed for its proof.

Let us define the boundary integral operator $\mathcal{K}\colon \mathcal{B}^0 \to \mathcal{B}^1$ with the kernel $\sigma \in C^1[0,T]$ with $\sigma(0) \neq 0$ by

$$(\mathcal{K}g)(t,x) = \int_0^t \sigma(t-s) g(s,x) \mathrm{d}s, \quad t \in (0,T), x \in \Gamma_0. \tag{5.12}$$

The following lemma ([Yam95, Lemma 3]) establishes a needed stability result for this central integral operator.

Lemma 5.7. *Let* $\mathcal{K}\colon \mathcal{B}^0 \to \mathcal{B}^1$ *be the operator defined by* (5.12). *Then there exists a constant* $c = c(T, \Omega) > 0$ *such that*

$$c^{-1} \left\| \mathcal{K}g \right\|_{\mathcal{B}^1} \leq \left\| g \right\|_{\mathcal{B}^0} \leq c \left\| \mathcal{K}g \right\|_{\mathcal{B}^1}, \tag{5.13}$$

for any $g \in \mathcal{B}^0$. □

Proof. Firstly, observe that from the integral equation (5.12) we have for some constant $\check{c} > 0$

$$\left\| \mathcal{K}g \right\|_{\mathcal{B}^0} \leq \check{c} \left\| g \right\|_{\mathcal{B}^0}. \tag{5.14}$$

Secondly, take the time derivative of (5.12) for $\sigma \in C^1[0,T]$

$$\frac{\partial \mathcal{K}g}{\partial t}(t) = \sigma(0)g(t) + \int_0^t \sigma'(t-s)g(s)\mathrm{d}s.$$

Since we have assumed that $\sigma(0) \neq 0$, this is a Volterra integral equation of second kind for which we have

$$\widetilde{c}^{-1} \left\| \frac{\partial \mathcal{K}g}{\partial t} \right\|_{\mathcal{B}^0} \leq \|g\|_{\mathcal{B}^0} \leq \widetilde{c} \left\| \frac{\partial \mathcal{K}g}{\partial t} \right\|_{\mathcal{B}^0}. \tag{5.15}$$

Now, by recalling from the definition of the H^1-norm (5.7) that

$$\|g\|^2_{\mathcal{B}^1} = \|g\|^2_{\mathcal{B}^0} + \|\partial g/\partial t\|^2_{\mathcal{B}^0},$$

we easily obtain the left hand side inequality of (5.13); we simply add (5.14) to the left part inequality of (5.15). The right hand side inequality is simply obtained by adding $\|g\|_{\mathcal{B}^0}$ to the right most part of (5.15). ■

By Duhamel's principle, Proposition 5.4, and Lemma 5.7 we can deduce the following result.

Lemma 5.8. *For $f \in C_0^\infty(\Omega)$ let v_f and w_f be the solutions of (5.1) and (5.9), respectively. Then there exists a constant $c > 0$ such that*

$$c^{-1} \|\partial_{\Gamma_0} v_f\|_{\mathcal{B}^1} \leq \|\partial_{\Gamma_0} w_f\|_{\mathcal{B}^0} \leq c \|\partial_{\Gamma_0} v_f\|_{\mathcal{B}^1}. \tag{5.16}$$

□

Proof. Since $f \in C_0^\infty(\Omega)$ the solution $w = w_f$ is sufficiently smooth on $[0,T] \times \overline{\Omega}$ so we have

$$\frac{\partial}{\partial n} \int_0^t \sigma(s) w(t-s,x)\mathrm{d}s = \int_0^t \sigma(s) \frac{\partial w}{\partial n}(t-s,x)\mathrm{d}s, \quad (t,x) \in (0,T) \times \Gamma.$$

By Duhamel's principle,Proposition 5.4, the right hand side equals $\partial_{\Gamma_0} v$, which means

$$\partial_{\Gamma_0} v_f(t) = \big(\mathcal{K} \partial_{\Gamma_0} w_f\big)(t), \quad (t,x) \in (0,T) \times \Gamma_0.$$

Hence, by the estimate (5.13) we are done. ■

We are now ready to state the main stability theorem ([Yam95, Theorem 1]).

Theorem 5.9 (Stability of (ISP)). *Let T and Γ_0 be so that system (5.9) is observable and let v_f be the solution of (5.1). Then there exists a constant $c > 0$ such that*

$$c^{-1} \|\partial_{\Gamma_0} v_f\|_{\mathcal{B}^1} \leq \|f\|_{L^2(\Omega)} \leq c \|\partial_{\Gamma_0} v_f\|_{\mathcal{B}^1},$$

for all $f \in L^2(\Omega)$. □

Proof. We need to extend the result of Lemma 5.8 to hold for $f \in L^2(\Omega)$. This is done by considering a sequence $\{f_n\}_{n\in\mathbb{N}}$ of C_0^∞-functions f_n, for which the inequality (5.16) holds. Then since $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$, we can pick the sequency such that $\|f_n - f\|_{L^2(\Omega)} \to 0$ for $n \to \infty$ and by the use of the bound (5.5) the result (5.16) holds for $f \in L^2(\Omega)$. Finally, the by Proposition 5.5 we conclude the proof. ■

5.2.2 Reconstruction

The reconstruction of f can be done by determining its Fourier coefficients. We expand f in terms of the previously mentioned eigenfunctions $\phi_k, k \in \mathbb{N}$

$$f = \sum_{k=1}^{\infty} \langle f, \phi_k \rangle_{L^2(\Omega)} \phi_k,$$

where $\langle f, \phi_k \rangle_{L^2(\Omega)}$ are the Fourier coefficients $\widehat{f_k}$ of f. As these coefficients are unavailable, we seek to determine them by measurements of $\partial_{\Gamma_0} u(f)$. Initially, we will, however, find the coefficients $\langle f, \phi_k \rangle_{L^2(\Omega)}$ in terms of $\partial_{\Gamma_0} w_f$ from (IIDP).

Reconstruction for (IIDP)

In Chapter 2 we saw how the controllability of a control system was closely linked to the observability of its adjoint system. We can view (5.9) as the adjoint system of some control system. Recall, also from Chapter 2, that a HUM control is a specific control build from the Neumann data $\partial_{\Gamma_0} w$ of the adjoint system on the observation boundary Γ_0.

Note that, from the solution (5.11) of (5.9), we have the following expansion of the Neumann data $\frac{\partial w_f}{\partial n}$ on Γ_0

$$\partial_{\Gamma_0} w_f = \sum_{k=1}^{\infty} \langle f, \phi_k \rangle_{L^2(\Omega)} \frac{\sin \lambda_k t}{\lambda_k} \partial_{\Gamma_0} \phi_k. \tag{5.17}$$

Recall the controllability operator $\Pi \colon \mathcal{E}^\star \to \mathcal{B}^0$, defined in Chapter 2 by (2.23), which gives the control κ of minimal norm for a control problem with the initial data $(u^1, -u^0)$. We will consider a reduced case of this map with the initial data $(0, -u^0)$ and call the corresponding operator

$$\Pi^0 \colon L^2(\Omega) \to \mathcal{B}^0, \qquad \Pi^0(u^0) = \Pi \begin{bmatrix} 0 \\ -u^0 \end{bmatrix}. \tag{5.18}$$

The main idea is now to use a series of HUM eigenfunction controls $\eta_k = \Pi^0(\phi_k)$ formed by the eigenfunctions $u^0 = \phi_k$ to "sample" the Fourier coefficients $\langle f, \phi_k \rangle_{L^2(\Omega)}$ of the expansion of $\partial_{\Gamma_0} w_f$ (5.17) in $\mathcal{B}^0$. We will use this idea first on a single eigenfunction $f = \phi_l$ which gives the following result ([Yam95, Lemma 5]).

Lemma 5.10. *Let system (5.9) be observable and $(\lambda_k, \phi_k), k \in \mathbb{N}$ be the corresponding eigensolutions. Furthermore, let Π^0 be the HUM-controllability operator defined in (5.18). Then we have the identity*

$$\langle \frac{\sin \lambda_l t}{\lambda_l} \partial_{\Gamma_0} \phi_l, -\Pi^0 \phi_k \rangle_{\mathcal{B}^0} = \delta_{kl}, \tag{5.19}$$

$k, l = 1, 2, \ldots$, where δ_{kl} is the Kronecker delta. □

Proof. Observe that with the initial conditions $(\varphi^0, \varphi^1) = (0, \phi_l)$ the system (2.6) has the solution

$$\varphi(t, x) = \frac{\sin(\lambda_l t)}{\lambda_l} \phi_l(x), \qquad (t, x) \text{ in } (0, T) \times \Omega, \tag{5.20}$$

which on Γ_0 has the normal derivative $\partial_{\nu}\varphi(t,x) = \frac{\sin(\lambda_l t)}{\lambda_l}\partial_{\nu}\phi_l(x)$.

Let $\psi = \psi(\kappa)$ be the solution of (2.5) for any $\kappa \in \mathcal{B}^0$ and φ the solution (5.20) to the adjoint system. Then we have

$$\langle \psi(\kappa)(0,\cdot), \phi_l \rangle_{L^2(\Omega)} = -\langle \kappa, \frac{\sin(\lambda_l t)}{\lambda_l}\partial_{\nu}\phi_l \rangle_{\mathcal{B}^0} \quad \text{for } l = 1, 2, \ldots \tag{5.21}$$

which follows from Theorem 2.13.

Now, let $u^0 = \phi_k$ and choose the control function $\kappa \in \mathcal{B}^0$ to be the HUM-control $\kappa = \Pi^0\phi_k$, then (5.19) follows directly from (5.21), since $\psi(\kappa)(0,\cdot) = u^0 = \phi_k$. ∎

Proposition 5.11. *Let system (5.9) be observable. Given $\partial_{\nu} w_f$ in $\mathcal{B}^0$, the function f with the Fourier expansion $f = \sum_{k\in\mathbb{N}} \widehat{f}_k\phi_k$ can be reconstructed by computing*

$$\widehat{f}_k = \langle \partial_{\nu} w_f, -\Pi^0\phi_k \rangle_{\mathcal{B}^0}, \qquad k \in \mathbb{N}, \tag{5.22}$$

where Π^0 is the controllability operator (5.18). □

Proof. Assume that we have an orthonormal basis $\{\widetilde{\theta}_k\}_{k\in\mathbb{N}}$ for $\mathcal{B}^0$ such that

$$\partial_{\nu} w_f = \sum_{l=1}^{\infty} \langle \partial_{\nu} w_f, \widetilde{\theta}_l \rangle_{\mathcal{B}^0} \widetilde{\theta}_l.$$

Now, by introducing (5.17) in this expansion, we get

$$\begin{aligned} \partial_{\nu} w_f &= \sum_{l=1}^{\infty} \Big\langle \sum_{k=1}^{\infty} \langle f, \phi_k \rangle_{L^2(\Omega)} \frac{\sin \lambda_k t}{\lambda_k} \partial_{\nu}\phi_k, \widetilde{\theta}_l \Big\rangle_{\mathcal{B}^0} \widetilde{\theta}_l \\ &= \sum_{l=1}^{\infty} \sum_{k=1}^{\infty} \langle f, \phi_k \rangle_{L^2(\Omega)} \langle \frac{\sin \lambda_k t}{\lambda_k} \partial_{\nu}\phi_k, \widetilde{\theta}_l \rangle_{\mathcal{B}^0} \widetilde{\theta}_l. \end{aligned}$$

Choose $\widetilde{\theta}_l$ as $-\Pi^0\phi_l$. Then by Lemma 5.10

$$\partial_{\nu} w_f = \sum_{k=1}^{\infty} \langle f, \phi_k \rangle_{L^2(\Omega)} \widetilde{\theta}_k,$$

which implies that $\langle f, \phi_k \rangle_{L^2(\Omega)} = \langle \partial_{\nu} w_f, -\Pi^0\phi_k \rangle_{\mathcal{B}^0}$ for all $k \in \mathbb{N}$ and hence completes the proof. ∎

Reconstruction for (ISP)

So far we have found a reconstruction formula for f in (IIDP) in the data $\partial_{\nu} w_f \in \mathcal{B}^0$. It remains to connect this result with (ISP), and we need an integral operator to this end.

Consider the Volterra integral equation of second kind for $(t,x) \in \Sigma_0$

$$\sigma(0)\theta'(t,x) + \int_t^T \big(\sigma'(\xi - t)\theta'(\xi,x) + \sigma(\xi - t)\theta(\xi,x)\big)\mathrm{d}\xi = \eta(t,x), \tag{5.23}$$

where σ is the temporal distribution of the source in (5.1) and $\sigma(0) \neq 0$.

Remark 5.12. The Volterra equation (5.23) is uniquely solvable for $\eta \in \mathcal{B}^0$ and $\theta \in \mathcal{B}^1$ by the so-called *resolvent kernel* and

$$\|\theta\|_{\mathcal{B}^1} \leq c \|\eta\|_{\mathcal{B}^0}$$

where $c > 0$ is a constant. ■

Due to the unique solvability of (5.23), we may define a boundary integral operator Ξ which connects $\eta \in \mathcal{B}^0$ to $\theta \in \mathcal{B}^1$.

Definition 5.13 (Ξ operator). Let $\sigma \in C^1[0,T]$ with $\sigma(0) \neq 0$. Then for $\eta \in \mathcal{B}^0$ we define the bounded operator Ξ

$$\Xi \colon \mathcal{B}^0 \to \mathcal{B}^1 \quad \text{by} \quad \theta = \Xi\eta, \tag{5.24}$$

such that $\theta \in \mathcal{B}^1$ is defined by the solution of (5.23). □

This operator allows us to state the main reconstruction result ([Yam95, Theorem 2]) as follows.

Theorem 5.14. *Let T and Γ_0 be so that system (5.9) is observable. Given $\partial_{\Gamma_0} u(f)$ in $\mathcal{B}^1$, the function $f \in L^2(\Omega)$ with the Fourier expansion $f = \sum_{k\in\mathbb{N}} \widehat{f}_k \phi_k$ can be reconstructed by computing*

$$\widehat{f}_k = \langle \partial_{\Gamma_0} u(f), \Xi(-\Pi^0 \phi_k) \rangle_{\mathcal{B}^1}, \qquad k \in \mathbb{N}, \tag{5.25}$$

where Π^0 is the controllability operator (5.18) and Ξ is the operator defined in Definition 5.13. □

Proof. Proposition 5.11 reduces the proof to a matter of establishing

$$\langle \partial_{\Gamma_0} u(f), \Xi(-\Pi^0 \phi_k) \rangle_{\mathcal{B}^1} = \langle \partial_{\Gamma_0} w_f, -\Pi^0 \phi_k \rangle_{\mathcal{B}^0},$$

for all $k \in \mathbb{N}$. Recall that $\partial_{\Gamma_0} u(f) = \mathcal{K} \partial_{\Gamma_0} w_f$, where $\mathcal{K} \colon \mathcal{B}^0 \to \mathcal{B}^1$ is the operator (5.12). Now, let us, for $g \in \mathcal{B}^0$ and $h \in \mathcal{B}^1$, define theadjoint $\mathcal{K}^* \colon \mathcal{B}^1 \to \mathcal{B}^0$ of $\mathcal{K}$ by

$$\langle \mathcal{K} g, h \rangle_{\mathcal{B}^1} = \langle g, \mathcal{K}^* h \rangle_{\mathcal{B}^0}.$$

It is easy to verify by direct calculations that $\mathcal{K}^*$ is the left inverse of Ξ, that is, $\mathcal{K}^* \Xi \eta = \eta$ for all $\eta \in \mathcal{B}^0$. This leads us to the following simple verification of our initial claim

$$\begin{aligned} \langle \partial_{\Gamma_0} u(f), \Xi(-\Pi^0 \phi_k) \rangle_{\mathcal{B}^1} &= \langle \mathcal{K} \partial_{\Gamma_0} w_f, \Xi(-\Pi^0 \phi_k) \rangle_{\mathcal{B}^1} \\ &= \langle \partial_{\Gamma_0} w_f, \mathcal{K}^* \Xi(-\Pi^0 \phi_k) \rangle_{\mathcal{B}^0} \\ &= \langle \partial_{\Gamma_0} w_f, -\Pi^0 \phi_k \rangle_{\mathcal{B}^0}. \end{aligned}$$

■

5.2.3 Regularization

Regularization is an important topic in the field of inverse problems as almost all inverse problems are ill-posed. Consider the operator $G \colon L^2(\Omega) \to \mathcal{B}^0$ defined

by $Gf = \partial_{\Gamma_0} v_f$. This operator is compact from $L^2(\Omega)$ to $\mathcal{B}^0$. We seek, for given $g_0 \in \mathcal{B}^0$, solutions f_0 to the equation

$$Gf_0 = g_0,$$

which due to the compactness of G is an ill-posed problem [Yam96].

If the available data g_δ is *inexact* with noise level δ, the task is to reconstruct reasonable approximations f_δ to f_0, that is, to solve the problem

$$Gf_\delta = g_\delta,$$

where $\|g_\delta - g_0\|_{\mathcal{B}^0} \leq \delta$. By reasonable we mean that $\|f_\delta - f_0\|_{L^2(\Omega)} \to 0$ for $\delta \to 0$. Notice that g_δ might not be in the range of the forward map $\mathcal{R}(G)$. Furthermore, $G^{-1}\colon \mathcal{R}(G) \to L^2(\Omega)$ is *not* continuous even though G is injective.

Yamamoto considers a Tikhonov-type regularization procedure where the functional

$$\mathcal{H}_\alpha(f) = \|Gf - g_\delta\|^2_{\mathcal{B}^0} + \alpha \|f\|^2_{L^2(\Omega)}$$

should be minimized over $f \in L^2(\Omega)$.

In order to derive concrete convergence rates for this regularization, Yamamoto uses a result from HUM theory on the range of the adjoint map G^* which coincides with the reachable set of a related control system. We will not go further into these details here as discretization itself will have sufficiently regularizing effect to make the problem well-posed (or only very mildly ill-posed)—see Section 5.4.2.

5.3 Discrete reconstruction

We now proceed with the numerical approximation of the reconstruction presented in Section 5.2.2. We consider the 1-d case $\Omega = (0, 1)$ like in the numerical HUM study in Chapter 4. Recall that the normalized eigenfunctions used in the expansion (5.11) in 1-d simply reads

$$\phi_k(x) = e^{\mathrm{s}}_k(x) = \sqrt{2}\sin(k\pi x), \qquad x \in \Omega,\ k = 1, 2, \ldots.$$

Let us introduce a semi-discretization with $N \in \mathbb{N}$ elements such that a function v defined on Ω gets approximated by the vector $\boldsymbol{v}$ with N elements. The space $\mathcal{X}^0$ with the inner product $\langle \cdot, \cdot \rangle_0$ replaces $L^2(\Omega)$. The choice of numerical scheme determines this discrete inner product, *e.g.,* definition (4.18) for L-FEM and (4.24) for DG-FEM.

Let f be a function in $L^2(\Omega)$. We assume that f is N_c-bandlimited, where $N_c \leq N$, such that we may expand it by N_c eigenfunctions $\phi_k, k = 1, \ldots, N_c$

$$f(x) = \sum_{k=1}^{N_c} \widehat{f}_k \phi_k(x), \qquad \widehat{f}_k = \langle f, \phi_k \rangle_{L^2(\Omega)}.$$

Semi-discretization suggests $c_k = \langle \boldsymbol{f}, \boldsymbol{e}^{\mathrm{s}}_k \rangle_0$ as approximation to the Fourier coefficients $\widehat{f}_k$. The vector $\boldsymbol{f}$ consists of N sampled values of f and the vector $\boldsymbol{e}^{\mathrm{s}}_k$ is in the same way a sampling of e^{s}_k. These cannot be found directly, however, as f is unknown, hence we need a reconstruction formula to recover the coefficients.

Note that if f is not N_c bandlimited we will only attempt to restore its first N_c coefficients.

Time discretization is also necessary before we proceed. We introduce a set of M discrete time instances $t = m\Delta t$ for $m = 0, \ldots, M-1$ where Δt is the time step. We define two inner product spaces $\mathcal{T}^0$ and $\mathcal{T}^1$ over $\mathbb{R}$ each of size M but with different inner products. The first, $\mathcal{T}^0$, that approximates $\mathcal{B}^0$, equals $\mathcal{T}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{T}^0}$ defined in (4.27). The second, $\mathcal{T}^1$, approximates $\mathcal{B}^1$ and is equipped with the inner product

$$\begin{aligned} \langle \boldsymbol{g}, \boldsymbol{h} \rangle_{\mathcal{T}^1} &:= \langle \boldsymbol{g}, \boldsymbol{h} \rangle_{\mathcal{T}^0} + \langle \boldsymbol{D}\boldsymbol{g}^{\mathsf{T}}, \boldsymbol{D}\boldsymbol{h}^{\mathsf{T}} \rangle_{\mathcal{T}^0} \\ &\equiv \Delta t\, \boldsymbol{g}\boldsymbol{h}^{\mathsf{T}} + \Delta t\, \boldsymbol{g}(\boldsymbol{D}^{\mathsf{T}}\boldsymbol{D})\boldsymbol{h}^{\mathsf{T}} \end{aligned} \tag{5.26}$$

corresponding to the H^1 inner product $\langle \cdot, \cdot \rangle_{\mathcal{B}^1}$. The vectors $\boldsymbol{g}$ and $\boldsymbol{h}$ are here row vectors in $\mathbb{R}^M$ and $\boldsymbol{D}$ is a temporal differentiation matrix defined by

$$\boldsymbol{D}_{ij} = \begin{cases} 1/\Delta t & j = i \\ -1/\Delta t & j = i-1 \\ 0 & \text{elsewhere,} \end{cases} \tag{5.27}$$

for $i, j = 1, \ldots, M$.

5.3.1 Discrete (IIDP)

We wish to discretize the inverse initial data problem (IIDP) defined on page 113 and, in particular, the reconstruction of f by its Fourier coefficients (5.22).

Consider first a semi-discretization equivalent to the one described in Section 4.1.2 for the HUM problem. We replace a function w defined on Ω by the vector $\boldsymbol{w}$ with N elements. Recall that system (5.9) is a special case of the adjoint HUM system (2.6), and its semi-discretization thus fit in the form (4.8) with $\boldsymbol{W}^0 = [\boldsymbol{0}, \boldsymbol{f}]^{\mathsf{T}}$, where $\boldsymbol{f}$ is the discrete representation of f.

The reconstruction of (IIDP) is based on the measurement of the Neumann data $\partial w_f / \partial n$ at Σ_0 of the auxiliary problem (5.9). We assume that this data is available to us in discrete form in the row vector $\widetilde{\boldsymbol{g}}$

$$\widetilde{\boldsymbol{g}} = [\partial_{\Gamma_0} w_f(0), \partial_{\Gamma_0} w_f(\Delta t), \ldots, \partial_{\Gamma_0} w_f((M-1)\Delta t)]^{\mathsf{T}}.$$

The Neumann data was, in Section 5.2.2, projected onto the space of eigenfunction controls constructed by the controllability operator $\Pi = \Phi\Lambda^{-1}$; we called the eigenfunctions controls $\eta_k = -\Pi(0, -\phi_k)$. The discrete equivalents of these functions are the row vectors

$$\boldsymbol{\eta}_k = -P\boldsymbol{L}^{-1} \begin{bmatrix} \boldsymbol{0} \\ -\boldsymbol{e}_k^{\mathrm{s}} \end{bmatrix}, \qquad k = 1, \ldots, N_c \tag{5.28}$$

where P is defined by (4.28) and $\boldsymbol{L}$ by (4.30). We assume here that $\boldsymbol{L}$ is invertible.

With $\widetilde{\boldsymbol{g}}$ and $\boldsymbol{\eta}_k$ we may approximate the coefficients $\widehat{f}_k = \langle \boldsymbol{f}, \boldsymbol{e}_k^{\mathrm{s}} \rangle_0$ by the following a discrete equivalent to (5.22)

$$\widehat{f}_k \approx \langle \widetilde{\boldsymbol{g}}, \boldsymbol{\eta}_k \rangle_{\mathcal{T}^0} \equiv \Delta t\, \widetilde{\boldsymbol{g}} \boldsymbol{\eta}_k^{\mathsf{T}}, \qquad k = 1, \ldots, N_c.$$

5.3.2 Discrete (ISP)

It remains to make the connection from the eigenfunction controls to the source problem (5.1). In Section 5.2.2 we introduced the solution of a Volterra integral equation to this end.

Discretization of the Volterra integral equation

The Volterra integral equation (5.23) establishes the pivotal connection between (IIDP) and (ISP). Note that the equation has no x-dependence in this simple 1-d case where Γ_0 is a point. We present below a simple discretization of this integral equation. Let in the following $\boldsymbol{\sigma}$ be a given row vector of size M with the m'th element $\sigma(m\Delta t)$; recall that $\sigma(0) \neq 0$. Let, furthermore, $\boldsymbol{\sigma'}$ be a vector of same size holding the derivative information $\sigma'(m\Delta t)$ at $m = 0, \ldots, M-1$. Consider the matrix equation

$$\sigma(0)\boldsymbol{D}\boldsymbol{\theta}^{\mathsf{T}} + \boldsymbol{U}(\boldsymbol{\sigma'})\boldsymbol{D}\boldsymbol{\theta}^{\mathsf{T}} + \boldsymbol{U}(\boldsymbol{\sigma})\boldsymbol{\theta}^{\mathsf{T}} = \boldsymbol{\eta}^{\mathsf{T}}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are M-sized vectors approximating the functions θ and η and the $M \times M$ differentiation matrix $\boldsymbol{D}$ is defined in (5.27). The matrix $\boldsymbol{U}(\boldsymbol{b})$ is an $M \times M$ **u**pper triangular matrix with elements

$$[\boldsymbol{U}(\boldsymbol{b})]_{ij} = \begin{cases} \Delta t\, \boldsymbol{b}_j & j \geq i, \\ 0 & j < i, \end{cases} \qquad i, j = 1, \ldots, M,$$

where $\boldsymbol{b}_j$ is the j'th element in the M sized vector $\boldsymbol{b}$ and Δt is the time step size. Row i of the matrix $\boldsymbol{U}(\boldsymbol{\sigma})$ multiplied the vector $\boldsymbol{\theta}^{\mathsf{T}}$ thus approximates the integral $\int_{i\Delta t}^{T} \sigma(s-t)\theta(s)\mathrm{d}s$ with kernel σ.

We now approximate the Volterra integral equation (5.23) by the matrix equation

$$\boldsymbol{X}_{\boldsymbol{\sigma}}\boldsymbol{\theta} = \boldsymbol{\eta}, \tag{5.29}$$

where $\boldsymbol{X}_{\boldsymbol{\sigma}}$ is an $M \times M$ *Volterra matrix* defined by

$$\boldsymbol{X}_{\boldsymbol{\sigma}} := \sigma(0)\boldsymbol{D} + \boldsymbol{U}(\boldsymbol{\sigma'})\boldsymbol{D} + \boldsymbol{U}(\boldsymbol{\sigma}). \tag{5.30}$$

If it exists, the inverse $\boldsymbol{X}_{\boldsymbol{\sigma}}^{-1}$ approximates the Ξ operator defined in (5.24).

The discrete reconstruction formula

Let $\boldsymbol{g}$ be a row vector of M time discrete Neumann data measurements at the right endpoint of Ω

$$\boldsymbol{g} = [\partial_{\Gamma_0} v_f(0), \partial_{\Gamma_0} v_f(\Delta t), \ldots, \partial_{\Gamma_0} v_f((M-1)\Delta t)]^{\mathsf{T}}$$

corresponding to the function $\partial_{\Gamma_0} v_f$. We approximate the N_c first of the coefficients (5.25) by

$$\widehat{f_k} \approx c_k^r = \left\langle \boldsymbol{g}, -\boldsymbol{X}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\eta}_k \right\rangle_{\mathcal{T}^1}, \qquad k = 1, \ldots, N_c, \tag{5.31}$$

where $\boldsymbol{X}_{\boldsymbol{\sigma}}$ is the Volterra matrix (5.30) and $\boldsymbol{\eta}_k$ is the eigenfunction control (5.28). The discrete inner product $\langle \cdot, \cdot \rangle_{\mathcal{T}^1}$ is defined in (5.26).

5.4 Numerical results

We now have a formula (5.31) for the discrete reconstruction of the Fourier coefficients of the unknown spatial distribution f of the source. The current section presents a numerical study of this reconstruction. Formula (5.31) relies on three components

(i) reliable data $\boldsymbol{g}$ (solution of the forward problem)

(ii) approximation of the HUM eigenfunction controls $\boldsymbol{\eta}_k$

(iii) approximation of the Volterra integral operator $\boldsymbol{X}_{\boldsymbol{\sigma}}$

The first item (i) may seem trivial, but it is not since the wavenumber can be high and we need to solve the forward problem numerically. For some choices of the temporal distribution σ of the source, we may, however, obtain analytic solutions to the forward problem $g = G(f)$ when $f = \phi_k$. Section 5.4.1 describes the generation of accurate and reliable data.

Item (ii) was the subject of Chapter 4. We will again focus on the L-FEM and DG-FEM semi-discretizations for the construction of controls. We know from Chapter 4 that this approximation is difficult, and in order to assess the quality of $\boldsymbol{X}_{\boldsymbol{\sigma}}$ as approximation of Ξ (iii) we will consider the special case $T = 2$ which grants the possibility of using simple analytic HUM eigenfunction controls $\boldsymbol{\eta}_k^{\mathrm{ex}}$. Section 5.4.3 studies the numerical reconstruction with analytic controls. The results of that section will be compared to results obtained with L-FEM generated controls $\boldsymbol{\eta}_k^{\mathrm{L}}$ in Section 5.4.4 and results by DG-FEM generated controls $\boldsymbol{\eta}_k^{\mathrm{L}}$ in Section 5.4.5.

5.4.1 Data and the forward problem

Data g for the inverse problem needs to be generated from f so we can assess the quality of the reconstructed solution. This is done by solving the forced wave equation (5.1) with the right hand side $\sigma(t)f(x)$. For known σ the forward, linar problem reads

$$G\colon f \mapsto \partial_{\Gamma_0} v_f,$$

where v_f is the solution of (5.1) and $\partial_{\Gamma_0} v_f$ is the Neumann data at $x = 1$. We get most information by solving for one eigenfunction $f = \phi_l$ at the time for $l = 1, \ldots, N_c$. We shall, in the following investigation, use three different temporal functions σ which allow analytic solution of the forward problem in 1-d when $f = \phi_l$. The σ functions, which all satisfy the requirement $\sigma(0) \neq 0$, are

$$\begin{aligned}
\sigma_a(t) &= 1, \\
\sigma_b(t) &= \cos(\pi t) + t^2 - 3t + 1, \\
\sigma_c(t) &= \cos(20\pi t).
\end{aligned}$$

These functions behave quite differently and are used to examine the effect of σ on the reconstruction. The different characteristics of these functions form different Volterra matrices $\boldsymbol{X}_{\boldsymbol{\sigma_a}}$, $\boldsymbol{X}_{\boldsymbol{\sigma_b}}$, and $\boldsymbol{X}_{\boldsymbol{\sigma_c}}$.

The simple σ_a has the derivative $\sigma_a'(t) = 0$ which makes the second term $\boldsymbol{U}(\boldsymbol{\sigma'})\boldsymbol{D}$ of (5.30) vanish. The function σ_b is smooth and contain both a polynomial and a trigonometric part. It is zero at $t = T$, but its derivative is

non-zero in both endpoints. The function σ_c oscillates fast and harmonically, and its derivative is zero at both endpoints. It is expected that σ_c requires finer resolution compared to σ_a and σ_b due to its fast variation.

Let us finally remark that for other σ functions for which analytic solution are not available, we need an accurate numerical solution on a very fine grid. It could, *e.g.,* be the Störmer-Numerov semi-discretization (unified scheme (3.22) with $\alpha = 1/12$) which is second order accurate. It is important to use different solvers for the forward and inverse problems. The use of the same solver for both purposes may, reversely, be unrealistically advantageous and is known as *inverse crime* (see [CK98]).

Let us consider a particular situation of the forward problem with $\sigma_a(t) = 1$ and $f = \phi_l$. We seek an analytic solution to this problem. The corresponding auxiliary problem (5.9) has the solution

$$w(t,x) = \frac{\sin(l\pi t)}{l\pi}\phi_l(x),$$

and the Neumann data $\partial_{\Gamma_0} w_f = (-1)^l\sqrt{2}\sin(l\pi t)$. We obtain the Neumann data for the source problem by applying the integral operator $\mathcal{K}$, defined in (5.12),

$$\begin{aligned}\partial_{\Gamma_0} v_f = \mathcal{K}\partial_{\Gamma_0} w_f &= \int_0^t 1(-1)^l\sqrt{2}\sin(l\pi s)\mathrm{d}s \\ &= (-1)^l\frac{\sqrt{2}}{l\pi}\big(-\cos(l\pi t) + \cos(0)\big).\end{aligned}$$

The exact solution to the forward problem $g_l = G(\phi_l)$ with σ_a is therefore

$$g_l^a(t) = (-1)^l\frac{\sqrt{2}}{l\pi}(1-\cos(l\pi t)), \qquad l = 1,\ldots,N_c \tag{5.32}$$

where the superscript a indicates the use of σ_a. We call g_l^a the *eigenfunction data*. Analytic solutions to the forward problem also exist for σ_b and σ_c. The expressions are very long but can be found in Appendix B.1.

5.4.2 The degree of ill-posedness

Let us consider σ_a and store the eigenfunction data g_l^a as rows in a matrix $\boldsymbol{G}^a$ with the l'th row being g_l^a at M discrete times

$$\boldsymbol{G}^a_{ml} = (-1)^l\frac{\sqrt{2}}{l\pi}\big(1-\cos(l\pi m\Delta t)\big), \qquad l = 1,\ldots,N_c, \quad m = 0,1,\ldots,M-1.$$

We study the singular values μ_k of this matrix to get an idea of the degree of ill-posedness of the inverse problem for the situation $\sigma_a(t) = 1$. P. C. Hansen [Han98] uses the following classification: the problem is characterized as *mildly* ill-posed if the singular values $\mu_k = \mathcal{O}(k^{-\alpha})$ for $\alpha \leq 1$, *moderately* ill-posed if $\mu_k = \mathcal{O}(k^{-\alpha})$ for $\alpha > 1$, and *severly* ill-posed if $\mu_k = \mathcal{O}(e^{-\alpha k})$ for $\alpha > 1$. Figure 5.1 shows the singular values of matrix $\boldsymbol{G}^a$. We see that μ_k scale like k^{-1} which corresponds to a mild ill-posedness of the inverse problem according to the above classification. It seems therefore that the discretization itself has a sufficient regularizing effect to make it computationally stable. The problem is

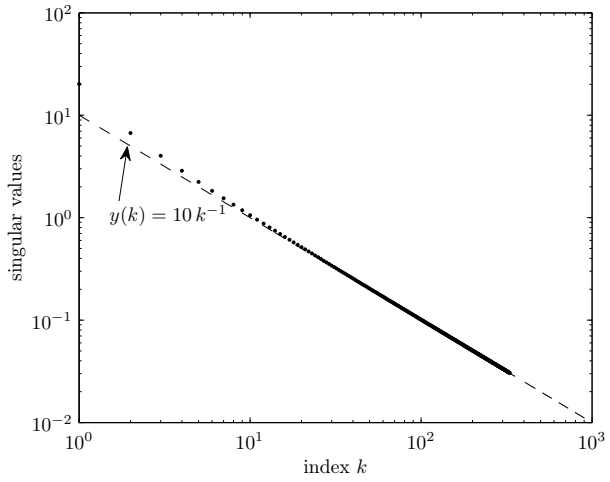


Figure 5.1: The singular values (dots) of the matrix $\boldsymbol{G}^a$ for $\sigma_a(t) = 1$ computed for $l = 1, \ldots, 300$ and $M = 1000$ discrete time steps ($T = 2$). The dashed line, $y(k) = 10\,k^{-1}$, shows the asymptotic behavior.

therefore not very sensible to noise. Notice, however, that this is not necessarily true for the reconstruction too.

The other temporal distributions, σ_b and σ_c, result in matrices $\boldsymbol{G}^b$ and $\boldsymbol{G}^c$ with similar behavior. The plots have for this reason been omitted.

5.4.3 Reconstruction with analytic HUM controls

We now proceed with the primary investigation of $\boldsymbol{X}_{\boldsymbol{\sigma}}$ which we construct for σ_a, σ_b, and σ_c for $T = 2$ by (5.30). This results in three $M \times M$ matrices $\boldsymbol{X}_{\boldsymbol{\sigma_a}}$, $\boldsymbol{X}_{\boldsymbol{\sigma_b}}$, and $\boldsymbol{X}_{\boldsymbol{\sigma_c}}$. We shall study the numerical reconstruction of the corresponding eigenfunction data $\boldsymbol{g}_l^a$,$\boldsymbol{g}_l^b$, or $\boldsymbol{g}_l^c$ but first we describe the construction of the analytic eigenfunction controls $\boldsymbol{\eta}_k^{\mathrm{ex}}$.

Exact HUM eigenfunction controls

We know from Chapter 2 that the special case $T = 2$ allows a simple exact HUM solution to the control problem with the initial data $(u^0, u^1) = (\phi_k, 0)$. The corresponding initial data of the adjoint problem becomes $(\bar{\varphi}^0, \bar{\varphi}^1) = (0, -\frac{1}{T}\phi_k)$ according to Remark 2.19. The eigenfunction control, which we obtain by observation with Φ, reads

$$\eta_k^{\mathrm{ex}}(t) = (-1)^{k+1}\frac{\sqrt{2}}{T}\sin(k\pi t), \qquad k = 1, \ldots, N_c.$$

This analytic result allows us to check the consistency of the discrete Volterra integral operator $\boldsymbol{X}_{\boldsymbol{\sigma}}$ without dealing with the difficult approximation of HUM at the same time. We store M samples of the function η_k^{ex} in the row vector

$$\boldsymbol{\eta}_k^{\mathrm{ex}} = [\eta_k^{\mathrm{ex}}(0), \eta_k^{\mathrm{ex}}(\Delta t), \ldots, \eta_k^{\mathrm{ex}}((M-1)\Delta t)].$$

This eigenfunction control vector will be used for the numerical reconstruction.

Numerical reconstruction

Let the vector $\boldsymbol{g}$ hold discrete data generated from f by the forward map $G(f)$. The first N_c Fourier coefficients $\widehat{f}_k$ of f may be approximated by the reconstruction formula

$$\widehat{f}_k \approx c_k^{\mathrm{r}} = \langle \boldsymbol{g}, -\boldsymbol{X}_{\boldsymbol{\sigma}}^{-1} \boldsymbol{\eta}_k^{\mathrm{ex}} \rangle_{\mathcal{T}^1}, \qquad k = 1, \ldots, N_c,$$

where the superscript r is for **r**econstructed. Let $\boldsymbol{g}_l$ be the data generated from the l'th eigenfunction $f = \phi_l$ which means that the exact Fourier coefficients are $\widehat{f}_k = \delta_{kl}$ for $K = 1, \ldots, N_c$. All N_c original coefficients $k = 1, \ldots, N_c$ for all N_c eigenfunctions $l = 1, \ldots, N_c$ may be collected in a matrix $\boldsymbol{C}^{\mathrm{ex}}$ with elements $\boldsymbol{C}_{kl}^{\mathrm{ex}} = \delta_{kl}$. We assemble a corresponding matrix of reconstructed coefficients $\boldsymbol{C}^{\mathrm{r}}$ with elements

$$\boldsymbol{C}_{kl}^{\mathrm{r}} = \langle \boldsymbol{g}_l, -\boldsymbol{X}_{\boldsymbol{\sigma}}^{-1} \boldsymbol{\eta}_k^{\mathrm{ex}} \rangle_{\mathcal{T}^1}, \qquad k, l = 1, \ldots, N_c.$$

We will compare the coefficient matrix $\boldsymbol{C}^{\mathrm{r}}$ with $\boldsymbol{C}^{\mathrm{ex}}$ for our three concrete choices of σ below. When nothing else is mentioned, we use $N_c = 25$ and $M = 285$ discrete times.

(a) Reconstruction when $\sigma = \sigma_a$

Let the temporal distribution be the constant σ_a which gives us the Volterra matrix $\boldsymbol{X}_{\boldsymbol{\sigma_a}}$ and the exact eigenfunction data $\boldsymbol{g}_l^a$ with discrete values of the function (5.32). We compute the reconstructed coefficients of the matrix $\boldsymbol{C}^{a,\mathrm{r}}$ by

$$\boldsymbol{C}_{kl}^{a,\mathrm{r}} = \langle \boldsymbol{g}_l^a, -\boldsymbol{X}_{\boldsymbol{\sigma_a}}^{-1} \boldsymbol{\eta}_k^{\mathrm{ex}} \rangle_{\mathcal{T}^1}, \qquad k, l = 1, \ldots, N_c.$$

Consider first two examples: $l = 4$ and $l = 23$. We show the reconstructed coefficients $\boldsymbol{C}_{k,4}^{a,\mathrm{r}}$ of the first case in Figure 5.2. Coefficient number 4, which

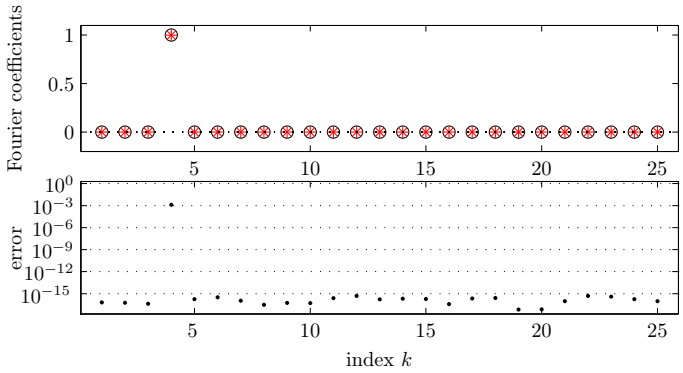


Figure 5.2: Reconstructed (*) and exact (o) Fourier coefficients for $f = \phi_4$ (upper plot) with σ_a and analytic controls $\boldsymbol{\eta}_k^{\mathrm{ex}}$. The lower plot shows the corresponding error for each coefficient on a logarithmic scale.

should be $\boldsymbol{C}_{4,4}^{a,\mathrm{ex}} = 1$, is $\boldsymbol{C}_{4,4}^{a,\mathrm{r}} = 0.9987$ whereas the rest is zero (at the order of machine precision 10^{-15}). The coefficients $\boldsymbol{C}_{k,23}^{a,\mathrm{r}}$ of the second case $l = 23$ is shown in Figure 5.3. The reconstructed coefficient $\boldsymbol{C}_{23,23}^{a,\mathrm{r}}$ is 0.9577 instead of 1 while the zeros are obtained to machine precision again. The less degree of

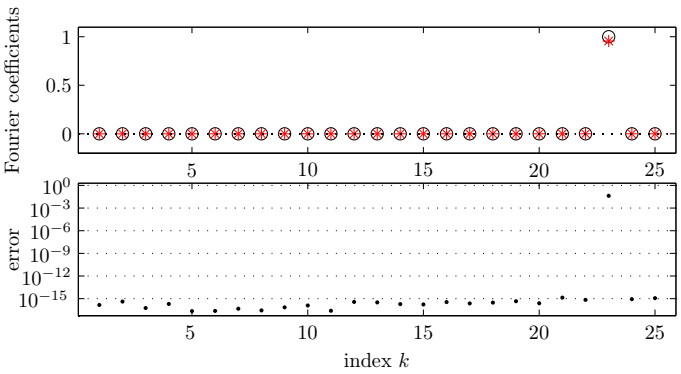


Figure 5.3: Reconstructed (*) and exact (o) Fourier coefficients for $f = \phi_{23}$ (upper plot) with σ_a and analytic controls $\boldsymbol{\eta}_k^{\mathrm{ex}}$. The lower plot shows the corresponding error for each coefficient on a logarithmic scale.

precision in $\boldsymbol{C}^{a,\mathrm{r}}_{23,23}$ compared to $\boldsymbol{C}^{a,\mathrm{r}}_{4,4}$ is probably due to the faster variation of g^a_{23} than of g^a_4.

The $\log_{10}$ of the absolute error for all elements of the complete coefficient matrix $\boldsymbol{C}^{a,\mathrm{r}}$ is plotted in gray scale in Figure 5.4. All off-diagonal elements are

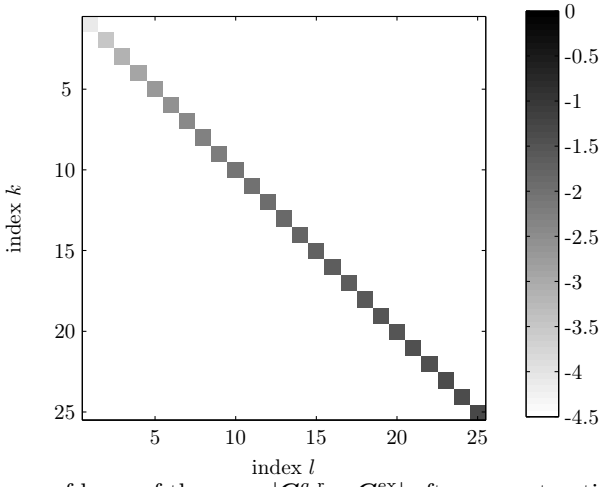


Figure 5.4: Image of $\log_{10}$ of the error $|\boldsymbol{C}^{a,\mathrm{r}} - \boldsymbol{C}^{\mathrm{ex}}|$ after reconstruction with analytic controls $\boldsymbol{\eta}_k^{\mathrm{ex}}$ and σ_a. Dark gray entries show greater error than light gray. A column l shows the error for each of the $N_c = 25$ reconstructed coefficients $k = 1, \ldots, N_c$ from the eigenfunction data $g^a_l = G(\phi_l)$. The average l^2 error over the columns is 0.0177.

machine zeros while the elements of the diagonal increase with the index. The error of the diagonal elements behaves quadratically $|\boldsymbol{C}^{a,\mathrm{r}}_{kk} - \boldsymbol{C}^{\mathrm{ex}}_{kk}| = 8{\cdot}(10)^{-5}k^2$. We measure the l^2 error for each eigenfunction data vector $\boldsymbol{g}^a_l$, that is, for each column in the error matrix, and take the average. In this case, where it has the value 0.0177, it is identical to the average of the diagonal elements as all off-diagonal elements are zero.

By increasing the number of time steps M, we can examine the convergence

of $\boldsymbol{X}_{\boldsymbol{\sigma_a}}$. The convergence plot in Figure 5.5 displays quadratic convergence. The reconstruction formula converge quite clearly in this case with σ_a and an-

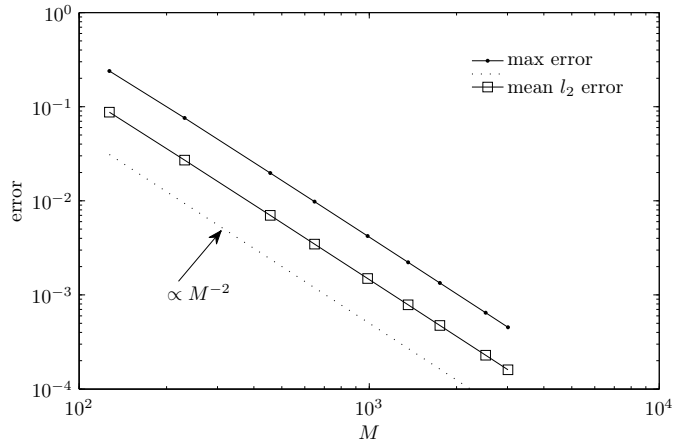


Figure 5.5: The max error, $\max_{kl} |\boldsymbol{C}^{a,\mathrm{r}}_{kl} - \boldsymbol{C}^{\mathrm{ex}}_{kl}|$, and the average l^2 error as functions of temporal resolution M for σ_a and analytic controls $\boldsymbol{\eta}^{\mathrm{ex}}_k$ in log-log axes. The dotted line, $y(M) \propto M^{-2}$, shows the rate of decay.

alytic HUM controls.

(b) Reconstruction when $\sigma = \sigma_b$

The use of the partly trigonometric, partly polynomial σ_b makes the reconstruction a bit more challenging. The corresponding Volterra matrix is $\boldsymbol{X}_{\boldsymbol{\sigma_b}}$. The eigenfunction data $\boldsymbol{g}^b_l$ is a vector of discrete values of the exact solution g^b_l of the forward problem with $f = \phi_l$. It can be seen in full length in Appendix B.1.

Let us again consider the coefficients for $\boldsymbol{g}^b_4$ and $\boldsymbol{g}^b_{23}$. Figure 5.6 shows the coefficients $\langle \boldsymbol{g}^b_4, -\boldsymbol{X}^{-1}_{\boldsymbol{\sigma_b}} \boldsymbol{\eta}^{\mathrm{ex}}_k \rangle_{\mathcal{T}^1}$. We see that the "zero" coefficients are no longer

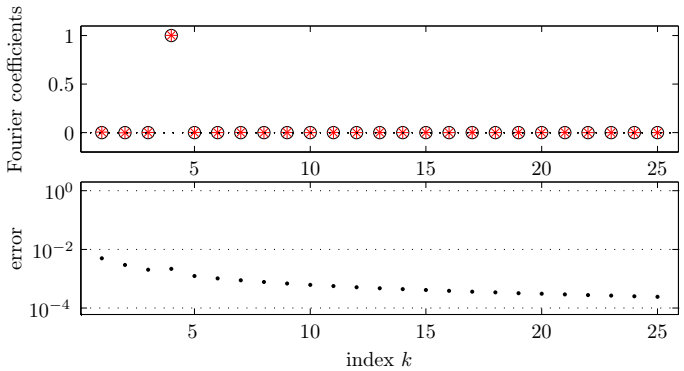


Figure 5.6: Reconstructed (*) and exact (o) Fourier coefficients for $f = \phi_4$ (upper plot) with σ_b and analytic controls $\boldsymbol{\eta}^{\mathrm{ex}}_k$. The lower plot shows the corresponding error for each coefficient on a logarithmic scale.

zero but in the order of 10^{-3}—higher for low k and lower for high k. The error of coefficient four is 0.002156, that is, about the same size as for σ_a.

The reconstructed coefficients for the data $\boldsymbol{g}^b_{23}$ are shown in Figure 5.7. The

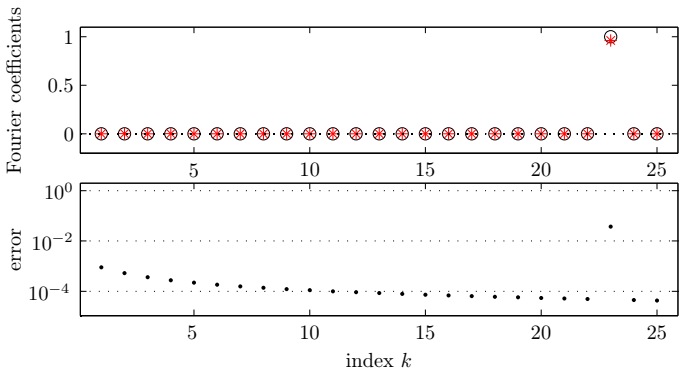


Figure 5.7: Reconstructed (*) and exact (o) Fourier coefficients for $f = \phi_{23}$ (upper plot) with σ_b and analytic controls $\boldsymbol{\eta}_k^{\mathrm{ex}}$. The lower plot shows the corresponding error for each coefficient on a logarithmic scale.

"zero" coefficients show smaller error than for $\boldsymbol{g}_4^b$, about 10^{-4} again decreasing with index k. The 23'rd coefficient has error 0.03722.

We compute the reconstructed coefficients for all eigenfunction data and put the results in the coefficient matrix $\boldsymbol{C}^{b,\mathrm{r}}$ defined by

$$\boldsymbol{C}_{kl}^{b,\mathrm{r}} = \langle \boldsymbol{g}_l^b, -\boldsymbol{X}_{\boldsymbol{\sigma_b}}^{-1} \boldsymbol{\eta}_k^{\mathrm{ex}} \rangle_{\mathcal{T}^1}.$$

We show the absolute error in Figure 5.8 as $\log_{10} \left| \boldsymbol{C}^{b,\mathrm{r}} - \boldsymbol{C}^{\mathrm{ex}} \right|$. The off-diagonal

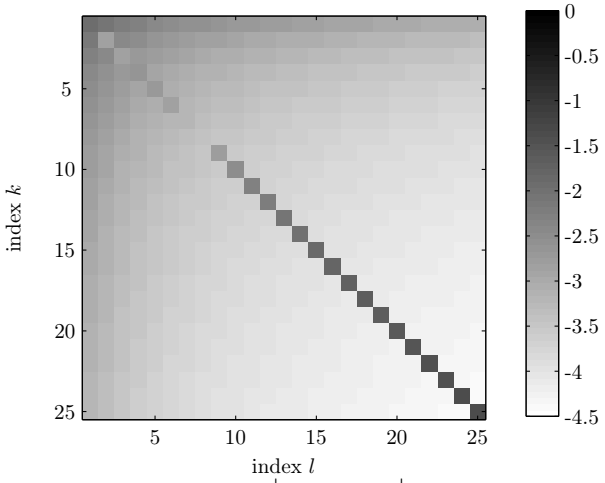


Figure 5.8: Image of $\log_{10}$ of the error $\left| \boldsymbol{C}_{kl}^{b,\mathrm{r}} - \boldsymbol{C}_{kl}^{\mathrm{ex}} \right|$ after reconstruction with analytic controls $\boldsymbol{\eta}_k^{\mathrm{ex}}$ and σ_b. The average l^2 error over the columns is 0.0155.

entries decrease for increasing k and l, whereas the opposite is the case for the diagonal elements that scale roughly like the diagonal elements for σ_a.

Figure 5.9 shows the max error and the average l^2 error for the reconstructed coefficients in log-log axes. The error clearly decreases with increasing temporal resolution M—first quadratically and then linearly for $M > 1000$. We conclude

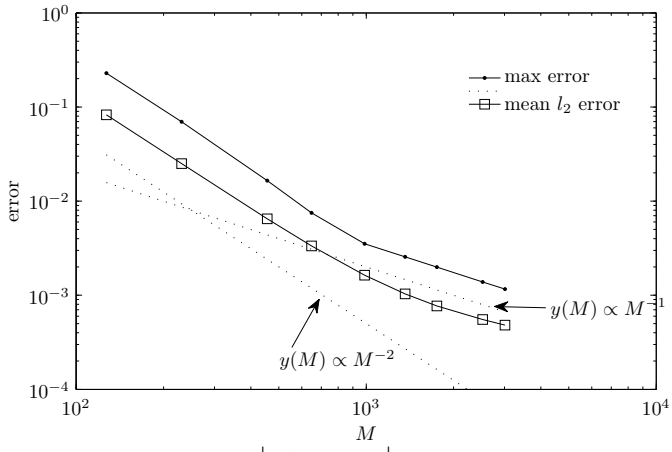


Figure 5.9: The max error, $\max_{kl} \left| \boldsymbol{C}_{kl}^{b,\mathrm{r}} - \boldsymbol{C}_{kl}^{\mathrm{ex}} \right|$, and the average l^2 error as function of temporal resolution M for σ_b and analytic controls $\boldsymbol{\eta}_k^{\mathrm{ex}}$ in log-log axes. The dotted lines, $y(M) \propto M^{-2}$ and $y(M) \propto M^{-1}$, show the approximate rate of decay for low and high M, respectively.

that the reconstruction converges for σ_b as well, though, a bit slower for high M compared to the convergence for σ_a.

(c) Reconstruction when $\sigma = \sigma_c$

We proceed with the examination for $\sigma = \sigma_c$ in the same way as for σ_a and σ_b. Notice, however, that due to the fast oscillations of σ_c, we will need higher temporal resolution compared to the other two functions. For now we stick with $M = 285$, though.

The exact data g_l^c (see Appendix B.1) is sampled and stored in the vector $\boldsymbol{g}_l^c$. We also generate a Volterra matrix $\boldsymbol{X}_{\boldsymbol{\sigma_c}}$ from σ_c.

Figure 5.10 shows the reconstructed coefficients $\langle \boldsymbol{g}_4^c, -\boldsymbol{X}_{\boldsymbol{\sigma_c}}^{-1} \boldsymbol{\eta}_k^{\mathrm{ex}} \rangle_{\mathcal{T}^1}$ for $f = \phi_4$. The error is significant for the 3'rd and the 4'th coefficients. For $k > 5$ it is

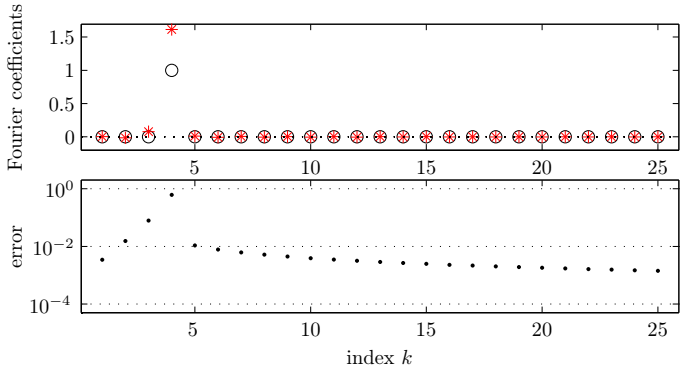


Figure 5.10: Reconstructed (*) and exact (o) Fourier coefficients for $f = \phi_4$ (upper plot) with σ_c and analytic controls $\boldsymbol{\eta}_k^{\mathrm{ex}}$. The lower plot shows the corresponding error for each coefficient on a logarithmic scale.

less than 10^{-2} and decreasing.

Similarly, the reconstructed coefficients of $f = \phi_{23}$ from the data $\boldsymbol{g}^c_{23}$ are shown in Figure 5.11.

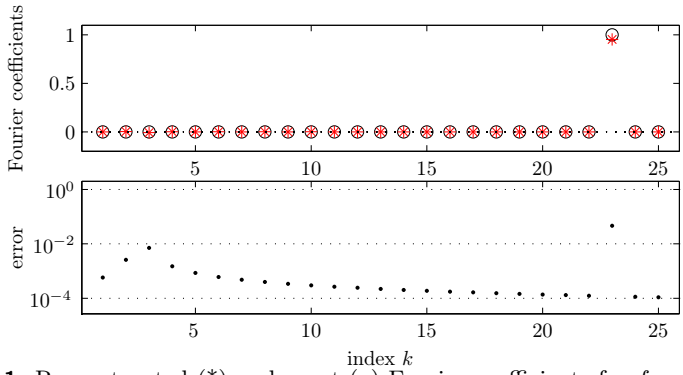


Figure 5.11: Reconstructed (*) and exact (o) Fourier coefficients for $f = \phi_{23}$ (upper plot) with σ_c and analytic controls $\boldsymbol{\eta}^{\mathrm{ex}}_k$. The lower plot shows the corresponding error for each coefficient on a logarithmic scale.

We compute the reconstructed coefficients for all eigenfunction data and put the results in the coefficient matrix $\boldsymbol{C}^{c,\mathrm{r}}$ defined by

$$\boldsymbol{C}^{c,\mathrm{r}}_{kl} = \langle \boldsymbol{g}^c_l, -\boldsymbol{X}^{-1}_{\boldsymbol{\sigma}_c} \boldsymbol{\eta}^{\mathrm{ex}}_k \rangle_{\mathcal{T}^1}.$$

The $\log_{10}$ absolute errors of these coefficients are shown in Figure 5.12. The

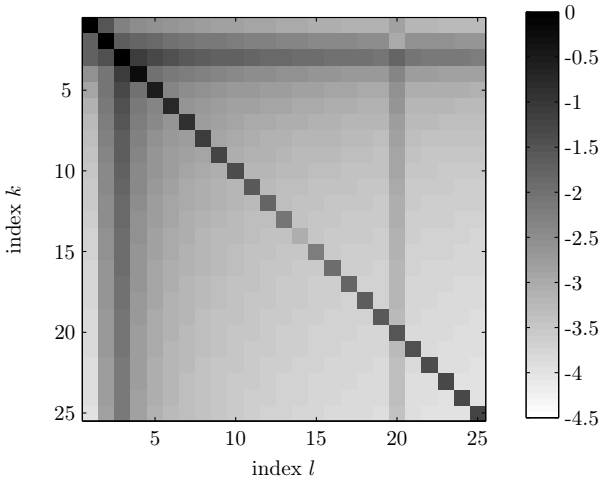


Figure 5.12: Image of $\log_{10}$ of the error $|\boldsymbol{C}^{c,\mathrm{r}}_{kl} - \boldsymbol{C}^{\mathrm{ex}}_{kl}|$ after reconstruction with analytic controls $\boldsymbol{\eta}^{\mathrm{ex}}_k$ and σ_c. The average l^2 error over the columns is 0.307.

errors are generally greater than the ones we saw for σ_b in Figure 5.8. The average l^2 error is 0.307 compared to 0.0155 for σ_b. The first few diagonal coefficients are particularly bad; the first 6 coefficients, which all should be 1, are

$$-0.1984 \quad -1.6015 \quad 3.0223 \quad 1.6141 \quad 1.3128 \quad 1.1899$$

Finer temporal discretization is needed for tolerable results. Notice also that the values of column 20 are a bit off which is most likely due to the singularity of the forward solution g_l^c for $l = 20$ (see Appendix B.1).

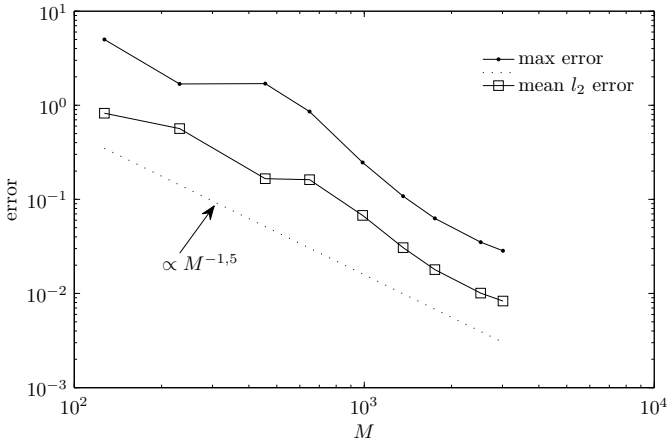


Figure 5.13: The max error, $\max_{kl} |\boldsymbol{C}_{kl}^{c,\mathrm{r}} - \boldsymbol{C}_{kl}^{\mathrm{ex}}|$, and the average l^2 error as function of temporal resolution M for σ_c and analytic controls η_k^{ex} in log-log axes. The dotted line, $y(M) \propto M^{-1.5}$, shows the approximate rate of decay.

We need around $M = 1500$ for σ_c to reach the same error level that we had with just $M = 150$ for σ_a and σ_b. But the reconstruction converges also in this case.

We have now seen that the numerical reconstruction converges with σ_a, σ_b, and σ_c. It seems reasonable, on this basis, to conclude that the numerical reconstruction converges with the use of analytic controls η_k^{ex} and smooth σ. But what happens if we replace the analytic controls with inexact controls obtained with numerical HUM? This is the subject of the next sections.

5.4.4 Numerical reconstruction with L-FEM

We shall now draw on the knowledge about numerical HUM with L-FEM semi-discretization that we gained in Section 4.2.4. Since we need not only a single control but a set of eigenfunction controls, it seems most practical to construct the controllability matrix $\boldsymbol{L}$ at least when the number of coefficients N_c are of the same order as the number of grid points N, *e.g.,* $N_c = \frac{1}{2}N$. If $N_c << N$, however, we prefer instead the conjugate gradient algorithm 4.2 (MCG-HUM) presented on page 100. We will focus here on the use of the matrix $\boldsymbol{L}$. Section 4.2.4 revealed that it is not feasible to use the full matrix, that is, $N_c = N$, but also that the formulation in sine basis allows an easy reduction of $\boldsymbol{L}$.

We need to construct the reduced controllability matrix $\boldsymbol{L}_{(N_c)}$ for $T = 2$. Notice that we, in this chapter, have used N_c for the number of Fourier coefficients that we wish to recover. In the previous chapter, N_c was used to designate the cut-off number for the reduced number of sine basis functions used for $\boldsymbol{L}_{(N_c)}$.

The two uses go well together since we need exactly N_c eigenfunction controls, which we get from $\boldsymbol{L}_{(N_c)}$, for the reconstruction of N_c Fourier coefficients. In this section, we will use $N_c = 25$ and define the L-FEM eigenfunction controls

$$\boldsymbol{\eta}_k^{\mathrm{L}} = -P_{(N_c)}\boldsymbol{L}_{(N_c)}^{-1}\begin{bmatrix}\mathbf{0}\\ -\boldsymbol{e}_j^{\mathrm{s}}\end{bmatrix}, \qquad k = 1, \ldots, N_c,$$

where $P_{(N_c)}$ is the reduced observation operator (4.28) (in sine basis). Both $P_{(N_c)}$ and $\boldsymbol{L}_{(N_c)}$ are constructed in L-FEM semi-discretization introduced on page 57 with $N = 49$ inner nodes. The superscript L on $\boldsymbol{\eta}_k^{\mathrm{L}}$ denotes the use of L-FEM semi-discretization. We use trapezoidal time integration with a Courant number such that $M = 285$.

Notice that the fraction N_c/N, which we denoted the filter fraction in the last chapter, is deciding for the quality of the controls. The smaller the fraction the more well-resolved are the short wavelength components.

Reconstruction

We will in the following examination focus mostly on σ_b as not to make the exposition unnecessarily long. We compute the coefficients by the reconstruction formula

$$\boldsymbol{C}_{kl}^{b,\mathrm{L}} = \langle \boldsymbol{g}_l^b, -\boldsymbol{X}_{\boldsymbol{\sigma_b}}^{-1}\boldsymbol{\eta}_k^{\mathrm{L}}\rangle_{\mathcal{T}^1}, \qquad k, l = 1, \ldots, N_c,$$

where the superscript L again denotes L-FEM; the superscript r has been omitted. As before, we set off by studying reconstruction of the fourth eigenfunction $l = 4$ which can be seen in Figure 5.14. The coefficients are restored quite well

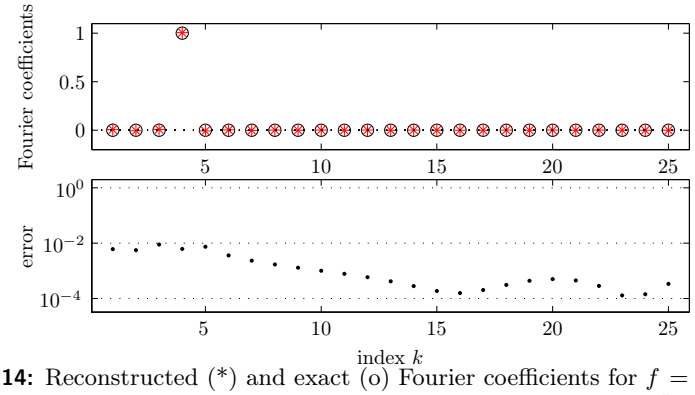


Figure 5.14: Reconstructed (*) and exact (o) Fourier coefficients for $f = \phi_4$ (upper plot) with σ_b and L-FEM ($N = 49$) numerical controls $\boldsymbol{\eta}_k^{\mathrm{L}}$. The lower plot shows the corresponding error for each coefficient on a logarithmic scale.

with a general error level of about 10^{-3}. Figure 5.15 shows the reconstruction of the $l = 23$'rd eigen function. The first 20 zeroes are retained quite well but the 23'rd coefficient, which should be one, is zero and the coefficients near it are off as well. This is most likely due to the numerical dispersion effects which also was seen in the falling eigenvalues of $\boldsymbol{L}^4$ in Figure 4.9.

The $\log_{10}$ error of all coefficients for all eigenfunction data for σ_a, σ_b, and σ_c is shown on Figure 5.16 on page 134. Observe, when comparing with the equivalent images for the analytic controls in Figures 5.4, 5.8, and 5.12, how the plots here are dominated by the error of the numerical HUM. The average

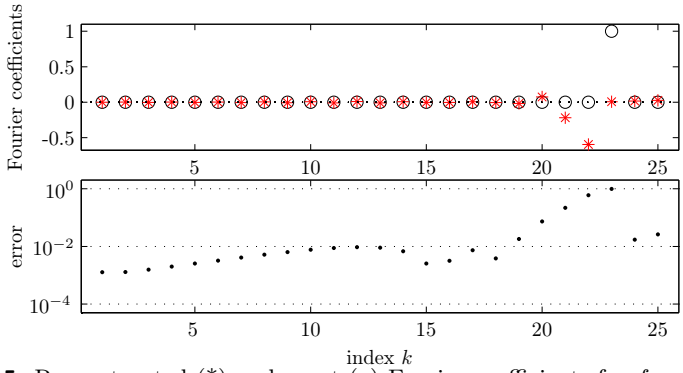
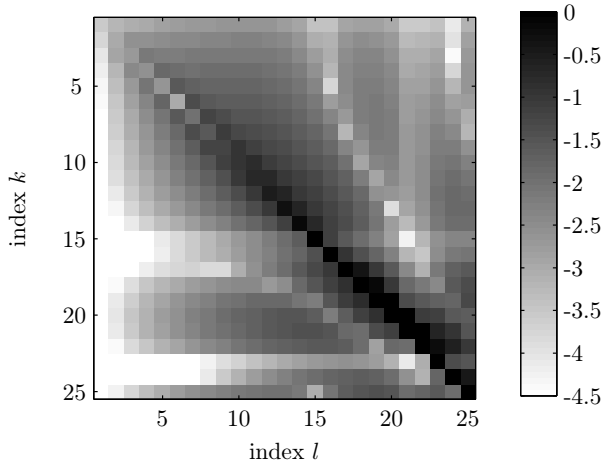
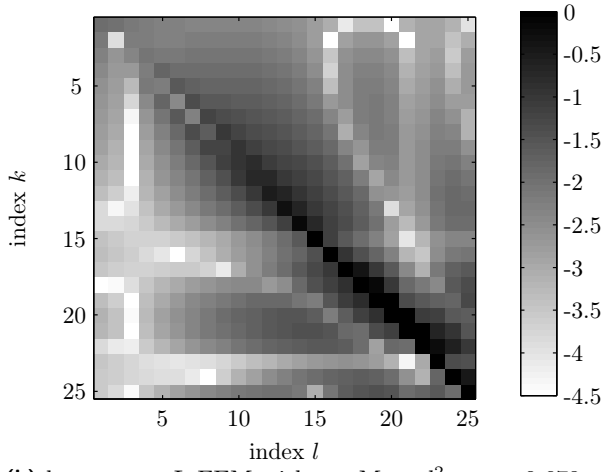


Figure 5.15: Reconstructed (*) and exact (o) Fourier coefficients for $f = \phi_{23}$ (upper plot) with σ_b and L-FEM ($N = 49$) numerical controls $\boldsymbol{\eta}_k^{\mathrm{L}}$. The lower plot shows the corresponding error for each coefficient on a logarithmic scale.

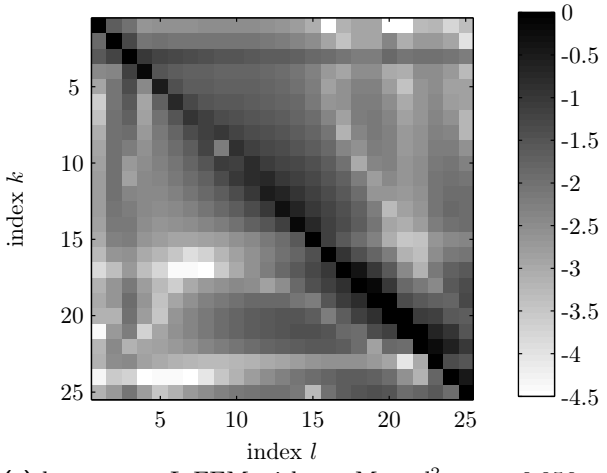
l^2-errors are also large and we conclude that more than $N = 49$ elements are needed for reasonable reconstruction of all $N_c = 25$ coefficients.



(a) $\log_{10}$ error; L-FEM with σ_a. Mean l^2-error=0.676.



(b) $\log_{10}$ error; L-FEM with σ_b. Mean l^2-error=0.678.



(c) $\log_{10}$ error; L-FEM with σ_c. Mean l^2-error=0.956.

Figure 5.16: Images of $\log_{10}$ of the error $\left|\boldsymbol{C}_{kl}^{\cdot,\mathrm{L}} - \boldsymbol{C}_{kl}^{\cdot,\mathrm{ex}}\right|$ after reconstruction with L-FEM ($N = 49$) numerical controls $\boldsymbol{\eta}_k^{\mathrm{L}}$ and three different σ (a,b, and c). The l^2-error showed under each of the above images is calculated as the average of the l^2-error of each column.

Convergence

Even though L-FEM did not give satisfactory results in the study above, it might will if we increase N and thereby decrease the filter fraction N_c/N. We fix the ratio between M and N so the Courant number $\mu = \Delta t/h = 0.6$, that is $M = \frac{T}{0.6}(N+1)+1$. The max error and the average l^2 error are plotted in logarithmic axes as function of M in Figure 5.17 to show the convergence. The

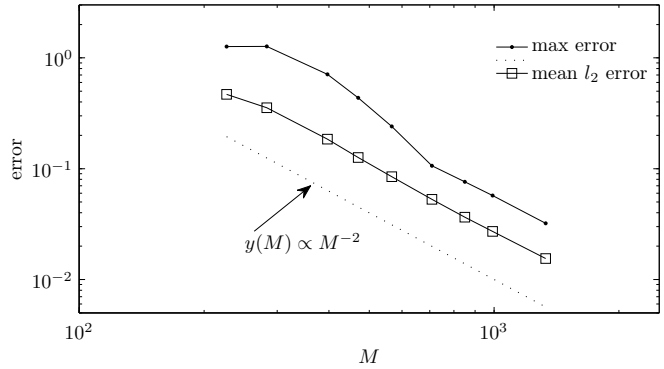


Figure 5.17: The max error, $\max_{kl}\left|\boldsymbol{C}_{kl}^{b,\mathrm{L}} - \boldsymbol{C}_{kl}^{\mathrm{ex}}\right|$, and the average l^2 error as function of temporal resolution M for σ_b and L-FEM controls η_k^{L} in log-log axes. The following numbers of inner grid points were used in the computations $N = 39, 49, 69, 82, 99, 124, 149, 174, 234$ and 289. The dotted line, $y(M) \propto M^{-2}$, shows the approximate rate of decay.

plot shows a quadratic convergence yet with errors well above those shown on Figure 5.9 for analytic controls.

5.4.5 Numerical reconstruction with DG-FEM

Let us consider the numerical reconstruction by eigenfunction controls obtained by DG-FEM semi-discretization. We compute the following set of controls

$$\boldsymbol{\eta}_k^{\mathrm{DG}} = -P_{(N_c)}\boldsymbol{L}_{(N_c)}^{-1}\begin{bmatrix}\boldsymbol{0}\\ -\boldsymbol{e}_k^{\mathrm{ss}}\end{bmatrix}, \qquad k = 1,\ldots,N_c,$$

where $P_{(N_c)}$ and $\boldsymbol{L}_{(N_c)}$ are the reduced matrices obtained with DG-FEM semi-discretization as described in Section 4.2.5 and $\boldsymbol{e}_k^{\mathrm{ss}}$ is the *sampled* sine basis defined in (4.50). Alternatively, we could use the projected sines $\boldsymbol{e}_k^{\mathrm{ps}}$ (4.52) which make up another polynomial representation of the continuous sine basis. The superscript DG on $\boldsymbol{\eta}_k^{\mathrm{DG}}$ denotes the use of *DG*-FEM semi-discretization. The control time is still $T = 2$.

We use our favorite grid with $K = 10$ elements and local polynomial order $N_p = 6$, as in Section 4.2.5, which gives $N = 49$ inner nodes. The Courant number $\mu = 0.6$ is used which gives $M = 285$ discrete times; h is here the *minimal* grid spacing.

Reconstruction

We wish to reconstruct the eigenfunctions ϕ_l for $l = 1,\ldots,N_c$ one at the time. Data $\boldsymbol{g}_l^b$ is generated from $f = \phi_l$ with σ_b. We try to reconstruct the eigenfunc-

tion from this data with the reconstruction formula

$$\boldsymbol{C}_{kl}^{b,\mathrm{DG}} = \langle \boldsymbol{g}_l^b, -\boldsymbol{X}_{\boldsymbol{\sigma_b}}^{-1} \boldsymbol{\eta}_k^{\mathrm{DG}} \rangle_{\mathcal{T}^1}, \qquad k,l = 1,\ldots,N_c.$$

Once more, we study the reconstruction of the fourth eigenfunction $l = 4$ first as may be seen in Figure 5.18. The result is satisfactory, similar to that of

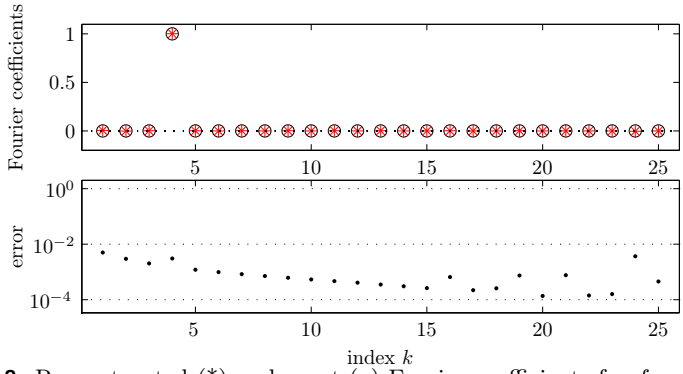


Figure 5.18: Reconstructed (*) and exact (o) Fourier coefficients for $f = \phi_4$ (upper plot) with σ_b and DG-FEM ($N_p = 6, K = 10$) numerical controls $\boldsymbol{\eta}_k^{\mathrm{DG}}$. The lower plot shows the corresponding error for each coefficient on a logarithmic scale.

L-FEM, and the general error level is around 10^{-3}. We see, however, some fluctuations from $k = 15$ and onwards. The error for $l = 23$, which we see on Figure 5.19, is increasing with the index k and it becomes noticeable after

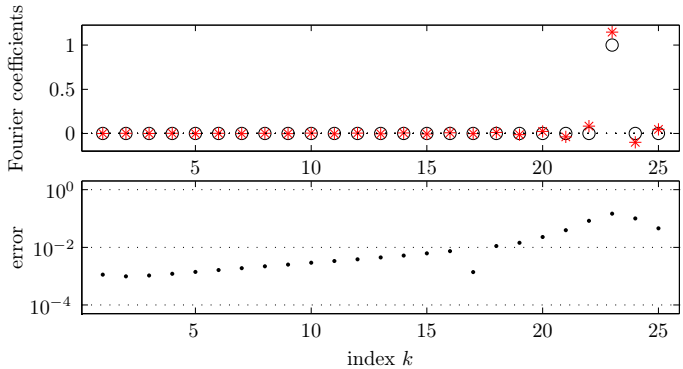
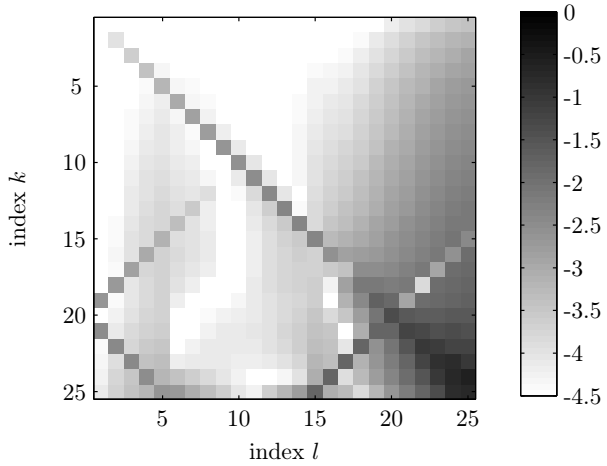


Figure 5.19: Reconstructed (*) and exact (o) Fourier coefficients for $f = \phi_{23}$ (upper plot) with σ_b and DG-FEM ($N_p = 6, K = 10$) numerical controls $\boldsymbol{\eta}_k^{\mathrm{DG}}$. The lower plot shows the corresponding error for each coefficient on a logarithmic scale.

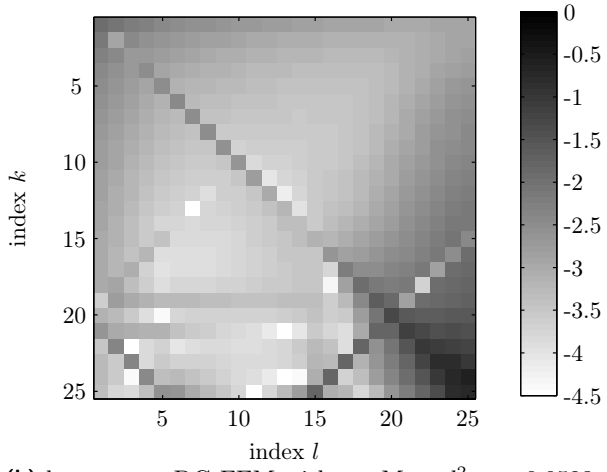
$k = 20$. But compared to the corresponding error for L-FEM (Figure 5.15), the present results are good. The yet inexact results for high k relates again to the decaying eigenvalues for $\boldsymbol{L}^4$ (see Figure 4.32).

The images of the $\log_{10}$ error for all coefficients for σ_a, σ_b, and σ_c, shown in Figure 5.20, reveal a huge performance difference between L-FEM and DG-FEM (compare with Figure 5.16). Especially, the first two images have very light gray tones, testifying small errors, compared to the much darker ones on Figure 5.16. The presented images bear closer resemblance with those made with analytic

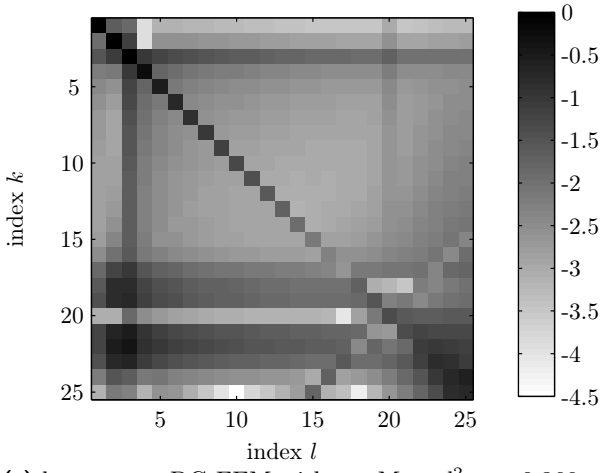
controls in Figures 5.4 and 5.8, at least for the first 16 or 17 coefficients; the highest coefficients are of lower quality. In the lower left corner we see the ambiguity patterns discussed in Section 4.2.5—see, *e.g.,* Figure 4.25(a) and the surrounding text. The dark gray area in the lower right corner reveals greater error than for the rest of the coefficients and is the major contributor to the average l^2-error.



(a) $\log_{10}$ error; DG-FEM with σ_a. Mean l^2-err=0.0492.



(b) $\log_{10}$ error; DG-FEM with σ_b. Mean l^2-err=0.0523.



(c) $\log_{10}$ error; DG-FEM with σ_c. Mean l^2-err=0.366.

Figure 5.20: Images of $\log_{10}$ of the error $\left|\boldsymbol{C}_{kl}^{\cdot,\mathrm{DG}} - \boldsymbol{C}_{kl}^{\cdot,\mathrm{ex}}\right|$ after reconstruction with DG-FEM ($N_p = 6, K = 10$) numerical controls $\boldsymbol{\eta}_k^{\mathrm{DG}}$ and three different σ (a,b, and c). The l^2-error showed under each of the above images is calculated as the average of the l^2-error of each column.

Convergence

We now vary the number of elements K to examine the h-convergence of the reconstruction with DG-FEM controls. Let $N_p = 6$ be fixed and let M scale with K such that we keep the Courant number $\mu = \Delta t/h$. We study the case with σ_b and compute the average l^2-error and the max error. The results are plotted in logarithmic axes in Figure 5.21.

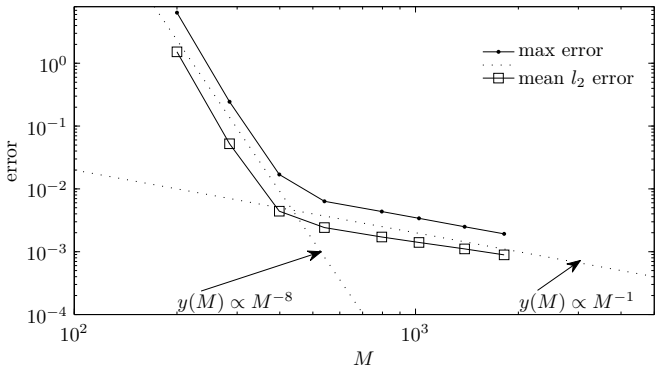


Figure 5.21: h-convergence of the max error, $\max_{kl} \left| \boldsymbol{C}_{kl}^{b,\mathrm{DG}} - \boldsymbol{C}_{kl}^{\mathrm{ex}} \right|$, and the average l^2 error as function of temporal resolution M for σ_b and DG-FEM controls $\boldsymbol{\eta}_k^{\mathrm{DG}}$ in log-log axes. The numbers of elements were $K = 7, 10, 14, 19, 28, 36, 49$ and 64 and the local polynomial order $N_p = 6$. The dotted lines show the approximate rates of decay for low and high M, respectively. See computational details in Table 5.1.

The convergence is very fast, proportional to M^{-8}, in the beginning but seems to hit a hurdle after $K = 14$ which makes the convergence only linear hereafter. This is the error of $\boldsymbol{X}_{\boldsymbol{\sigma_b}}$ that takes over—compare with the numbers in Figure 5.9. It is therefore advisable not to use more elements than $K = 14$ when the polynomial order is $N_p = 6$ as the gain in accuracy for the controls will be consumed by the $\boldsymbol{X}_{\boldsymbol{\sigma}}$ error. The numerical controls are here sufficiently accurate and the simple low order differentiation and integration of $\boldsymbol{X}_{\boldsymbol{\sigma}}$ now becomes the major obstacle for more accurate reconstruction.

Table 5.1 shows the details behind the data points of the plot in Figure 5.21. For each K there is shown a corresponding number of inner nodes $N = K(N_p - 1) - 1$, filter fraction N_c/N, number of discrete times M, and of course the error values.

Let now $K = 10$ be fixed. We vary the local polynomial order N_p to study the p-convergence—Figure 5.22 shows the results. The fast convergence, proportional to M^{-6}, of the reconstruction for low N_p is due to the fast convergence of DG-FEM controls. Again, the error of $\boldsymbol{X}_{\boldsymbol{\sigma_b}}$ (see Figure 5.9) becomes dominating shortly after $M = 500$ ($N_p = 8$). This suggest to use polynomial order no higher than $N_p = 8$ when the number of elements are $K = 10$ for the same reason as before. Table 5.2 displays the details of the p-convergence analysis.

Table 5.1: Values used for the h-convergence analysis of the numerical reconstruction with DG-FEM controls (see Figure 5.21). The local polynomial order is everywhere $N_p = 6$. Below K is the number of elements, N is the number of inner grid points, N_c/N is the filter fraction, M is the number of discrete times, l^2 is short for l^2-error, and max is short for max error.

K	7	10	**14**	19	28	36	49	64
N	34	49	**69**	94	139	179	244	319
N_c/N	0.7353	0.5102	**0.3623**	0.2660	0.1799	0.1397	0.1025	0.0784
M	200	285	**399**	541	796	1023	1392	1818
l^2	1.5282	0.0523	**0.0044**	0.0024	0.0017	0.0014	0.0011	0.0009
max	6.4297	0.2440	**0.0169**	0.0064	0.0043	0.0034	0.0025	0.0019

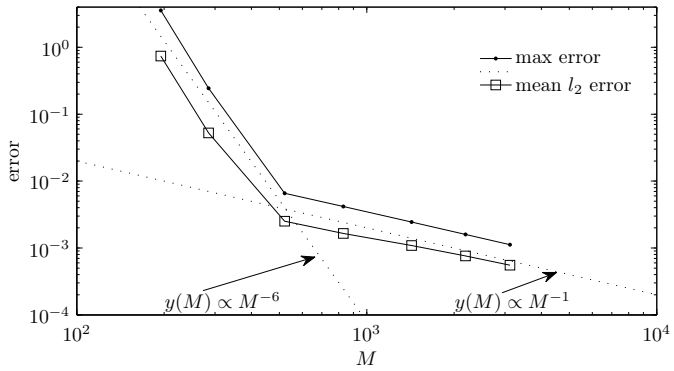


Figure 5.22: p-convergence of the max error, $\max_{kl} \left| \boldsymbol{C}_{kl}^{b,\mathrm{DG}} - \boldsymbol{C}_{kl}^{\mathrm{ex}} \right|$, and the average l^2 error as function of temporal resolution M for σ_b and DG-FEM controls $\boldsymbol{\eta}_k^{\mathrm{DG}}$ in log-log axes. The number of elements was $K = 10$ and the local polynomial orders $N_p = 5, 6, 8, 10, 13, 16$ and 19. The dotted lines show the approximate rates of decay for low and high M, respectively. See computational details in Table 5.2.

Table 5.2: Values used for the p-convergence analysis of the numerical reconstruction with DG-FEM controls (see Figure 5.22). The number of elements is $K = 10$. See Table 5.1 for detailed explanation.

N_p	5	6	**8**	10	13	16	19
N	39	49	**69**	89	119	149	179
N_c/N	0.6410	0.5102	**0.3623**	0.2809	0.2101	0.1678	0.1397
M	195	285	**521**	830	1429	2192	3118
l^2	0.7402	0.0523	**0.0025**	0.0017	0.0011	0.0008	0.0006
max	3.5676	0.2440	**0.0066**	0.0042	0.0024	0.0016	0.0011

5.4.6 An example with random coefficients

So far we have solely considered results for one eigenfunction data vector $\boldsymbol{g}_l$ at the time. We will conclude this numerical study with the numerical reconstruction of 25 random Fourier coefficients. A random sequence of numbers all from the set $[-1;-0.05]\cup[0.05;1]$ has been generated. The forward problem with σ_b is solved analytically (see Appendix B.1) and the Neumann data is stored in the vector $\boldsymbol{g}^b$.

We try to reconstruct the random coefficients from the data $\boldsymbol{g}^b$ with the reconstruction formula (5.31) with three different sets of eigen function controls $\boldsymbol{\eta}_k$.

(i) Analytic HUM controls $\boldsymbol{\eta}_k^{\mathrm{ex}}$ (Figure 5.23)
(ii) Numerical HUM controls with L-FEM $\boldsymbol{\eta}_k^{\mathrm{L}}$ (Figure 5.24)
(iii) Numerical HUM controls with DG-FEM $\boldsymbol{\eta}_k^{\mathrm{DG}}$ (Figure 5.25)

Each of the figures below presents f, its Fourier coefficients and the error of each coefficient. In all cases $M = 285$ has been used. The analytic eigenfunction

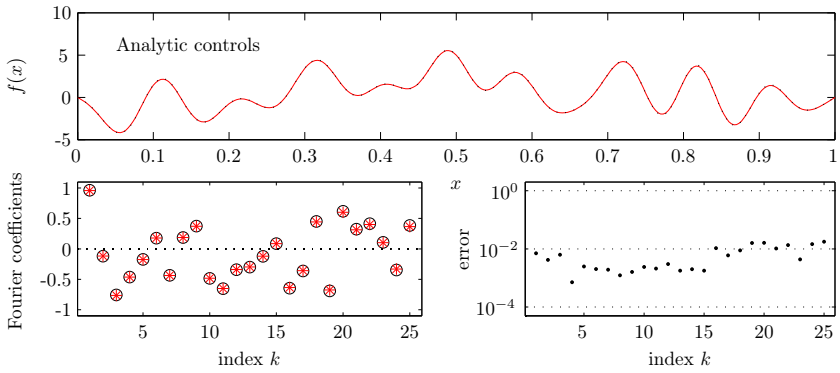


Figure 5.23: Numerical reconstruction performed with analytic HUM controls $\boldsymbol{\eta}_k^{\mathrm{ex}}$ and $M = 285$. Upper plot: the graph of the reconstructed f (solid) and the original f (dotted). Lower left: the reconstructed Fourier coefficients (*) and the original coefficients (o) as function of the index k. Lower right: the absolute error of each coefficient.

controls should, according to the error image of Figure 5.8, lead to a small error as we also see on Figure 5.23. The caption of Figure 5.8 reported an average error level around 10^{-2} which corresponds quite well to the error on the lower right plot of Figure 5.23.

The situation is quite different for the L-FEM numerical HUM controls shown in Figure 5.24. The reconstructed f does not resemble the original f. Only the first half of the coefficients are approximated reasonably.

With DG-FEM controls the results are much better as we see on Figure 5.25. The coefficient errors are less than 10^{-2} except for a few after $k = 20$ and the graph of the original f is followed close by the reconstructed f.

That DG-FEM performed better than L-FEM was after all expected as DG-FEM is a higher order method, yet the difference for the numerical reconstruction is still notable.

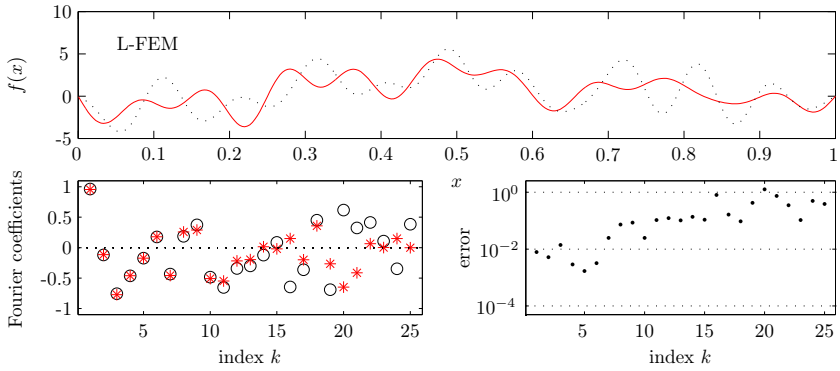


Figure 5.24: Numerical reconstruction made with numerical L-FEM ($N = 49$) HUM controls $\boldsymbol{\eta}_k^{\mathrm{L}}$ and $M = 285$. The graph of f (upper); the Fourier coefficients (lower left); the error of each coefficient (lower right).

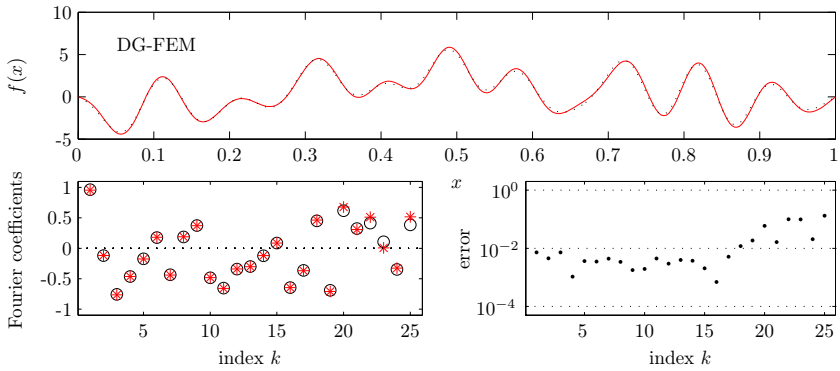


Figure 5.25: Numerical reconstruction made with numerical DG-FEM ($N_p = 6; K = 10$) HUM controls $\boldsymbol{\eta}_k^{\mathrm{DG}}$ and $M = 285$. The graph of f (upper); the Fourier coefficients (lower left); the error of each coefficient (lower right).

5.5 Concluding remarks

In this chapter, we have considered an inverse source problem for the wave equation. The source term was separable in a spatial part f and a temporal part σ. For given σ, the inverse problem consisted of finding f from measured Neumann data on a part of the boundary Γ_0. Yamamoto showed in his paper [Yam95] how the HUM could be used for a unified solution to the problems of stability, reconstruction, and regularization. An auxiliary problem, related to the source problem by a boundary integral operator K, could be considered the adjoint of a controllability problem. By probing this auxiliary problem with eigenfunctions and measuring the output on the boundary by inner products with eigenfunction controls, we restored the Fourier coefficients of f. The final task was to connect the boundary data of the auxiliary problem with the boundary data of the original problem. This was done by the already mentioned boundary integral operator.

Numerical approximations of Yamamoto's solution to the inverse problem have not been published before. We conducted a numerical study of the reconstruc-

tion of f in 1-d in this chapter. A discretization was proposed, and the three basic components of the reconstruction formula, (i) generation of data $\boldsymbol{g}$, (ii) Volterra matrix $\boldsymbol{X}_{\boldsymbol{\sigma}}$, and (iii) eigenfunction controls $\boldsymbol{\eta}_k$, were addressed.

Regarding (i), we used analytic solutions of the forward problem to generate eigenfunction data $g_l = G(\phi_l)$. We introduced a simple approximation of the Volterra integral equation with first order approximations of differentiation and integration which led to the Volterra matrix $\boldsymbol{X}_{\boldsymbol{\sigma}}$. We used three different sets of eigenfunction controls: analytic HUM controls $\boldsymbol{\eta}_k^{\mathrm{ex}}$, numerical HUM controls obtained with, respectively, L-FEM semi-discretization $\boldsymbol{\eta}_k^{\mathrm{L}}$, and DG-FEM semi-discretization $\boldsymbol{\eta}_k^{\mathrm{DG}}$.

Yamamoto showed that the forward map G is compact and the inverse problem therefore ill-posed. We assessed the *degree* of ill-posedness of the inverse problem in 1-d by examining the singular values of the forward problem. The availability of analytic solutions, obtained for the auxiliary problem and mapped to the source problem by an integral equation, made the analysis reliable since we thereby ruled out the effect of numerical errors. The singular values of G scaled like k^{-1} for the index k corresponding to only a mild ill-posedness of the inverse problem. This property was not greatly influenced by changing σ—at least not for the three types of σ studied here. Since the discretized problem was only mildly ill-posed, no regularization was needed.

The temporal distribution of the source σ had an effect on the quality of the numerical reconstruction. This became evident after studying the reconstruction results obtained with analytic HUM controls and three different σ for $M = 285$. The first, σ_a, which was simply the constant 1, reconstructed the zero coefficients to machine precision while providing reasonable reconstruction for the "ones". The smooth σ_b allowed similar reconstruction of the "ones" whereas the "zeros" had an error level of 10^{-3}. The fast, harmonically oscillating σ_c required a higher number of discrete time steps. Convergence was showed for all three σ with different rates of decay, though.

The reconstruction with numerical HUM controls obviously resulted in higher error level compared to the analytic HUM controls. The use of L-FEM with $N = 49$ inner grid points seemed insufficient for the reconstruction of $N_c = 25$ coefficients; only about half the coefficients could be reconstructed with reasonable accuracy. A DG-FEM discretization with the same number of inner grid points, but with more degrees of freedom, though, provided far better results than L-FEM. The reconstruction converged for both semi-discretizations. L-FEM generally lacked far behind DG-FEM until DG-FEM reached the error level of the Volterra matrix which after $K = 14$ (with $N_p = 6$) or $N_p = 8$ (with $K = 10$) became the major source of error. This analysis also showed that even though the inverse problem is not severly ill-posed, the reconstruction process is quite sensible to HUM control inaccuracies. It proved, furthermore, that our simple choices of approximate differentiation and integration for $\boldsymbol{X}_{\boldsymbol{\sigma}}$ were sufficient as the error from inaccurate control is the dominating factor. This holds unless we use HUM controls made with high order DG-FEM which will put $\boldsymbol{X}_{\boldsymbol{\sigma}}$ to the test.

CHAPTER 6

Conclusion

This final chapter will summarize the main results of this work and offer a view of its perspectives as well as suggest future work.

6.1 Results

The first three chapters of this dissertation contained the background necessary for the studies made in Chapter 4 and 5. Chapter 4 treated the numerical approximation of HUM boundary control, and Chapter 5 dealt with the numerical approximation of a reconstruction formula for an inverse problem. Both chapters were concluded by a discussion which we shall not repeat here. We will instead try to condense the main results in just a few paragraphs.

6.1.1 The control problem and numerical HUM

The numerical approximation of HUM boundary control for the wave equation is well-known to be difficult. We have studied numerical HUM in 1-d with mainly two different discretizations of the wave equation: the linear FEM (L-FEM) and the discontinuous Galerkin-FEM (DG-FEM).

Sinusoidal basis. A choice of basis function needs to be made when discretizing HUM. The use a sinusoidal basis has not previously been described in the literature. We presented the use of a sine basis and showed its advantages over the canonical basis: (i) separation of waves with small and large wavenumbers, (ii) a very close connection to the dispersive properties of a discretization, (iii) a simple and effective filtering procedure reducing the number of computations in the construction of matrix $\boldsymbol{L}$ by the factor N_c/N.

L-FEM for numerical HUM. We presented the assembly of matrix $\boldsymbol{L}$ with L-FEM semi-discretization. Our study showed the importance of choosing a scheme with good dispersive properties. Many authors have argued that group velocity is determining for the success of numerical HUM. Group velocity is known to be of significant importance for control. We demonstrated how the vanishing group velocity of waves with highest wavenumbers led to a dramatical decay of the eigenvalues of $\boldsymbol{L}$ which in turn led to huge condition numbers. In this way, *group* velocity is determining for the *success* of the numerical approximation. We found, on the other hand, that it is the *phase* velocity that decides the *quality* and the accuracy of the control after filtering.

DG-FEM for numerical HUM. DG-FEM has not previously been used in the context of HUM boundary control for the wave equation. We applied the method which demonstrated superior results, particularly in the low frequency region, compared to L-FEM. By increasing the order of DG-FEM, we even obtained spectral accuracy for the eigenvalues of $\boldsymbol{L}$. The representation of sinusoids with large wavenumbers by local polynomials does, however, point at the limitations of the used formulation. The "ambiguity patterns", showing the variation of frequencies over the domain, remained in the approximation and could be seen in the resulting eigenfunction controls.

6.1.2 The inverse problem and the numerical reconstruction

Chapter 5 dealt with an inverse source problem for the wave equation. We investigated the numerical aspects of a method by M. Yamamoto and proposed a numerical approximation to the reconstruction in 1-d. The source term consisted of a known temporal part σ and an unknown spatial part f.

The discrete forward problem. By analyzing analytic solutions for the forward problem, we were able to assess the degree of ill-posedness of the inverse problem. We studied the distribution of singular values and found that the inverse problem was only mildly ill-posed; the temporal distribution of the source σ had only negligible effect on this matter.

The discrete reconstruction formula. We suggested a simple discretization of the reconstruction formula and the Volterra integral equation. The discretization converged for increasingly fine temporal discretizations. We also saw that the rate of convergence was influenced by the choice of σ.

Reconstruction by numerical HUM controls. After studying the reconstruction with analytic HUM controls, we considered the eigenfunction

controls found by numerical HUM with respectively L-FEM and DG-FEM. We showed that, in spite of the problem being only mildly ill-posed, the numerical reconstruction was sensible to inaccurate eigenfunction controls. The sensibility was particular pronounced for frequency errors since the Fourier coefficients of f are reconstructed by eigenfunction data, one wavenumber at the time. This certainly favored the DG-FEM controls over the L-FEM ones.

6.1.3 Software contributions

The developed software for the HUM solution of the boundary control for the wave equation and the inverse source problem has been made freely available from http://www.mat.dtu.dk/people/J.S.Mariegaard/software/. It is the intension that scholars of HUM or inverse problems can download the code for studying, modification and extension. The motivation for sharing this software came from the author's own problems with finding code for dealing with HUM control.

6.2 Future work

As with most other scientific endeavors, this project posed at least as many questions as it answered. A few of these and some ideas of improvement, which seem particularly promising, are listed below.

Dispersion and dissipation. The dispersive behavior of a discretization in regards to control was given a lot of attention in this dissertation. Yet, it would be interesting to go even further in this analysis. Can precise predictions be made about the capabilities of a discretization in respect to control alone by considering its dispersive relation? Can we quantify predictions? In terms of required filter index N_c? In terms of convergence rates? Speaking accuracy, the dissipation of a scheme is obviously important as well. We saw the consequences of dissipative behavior when we studied the L-FEM observation of the sine basis vectors, but we did not link it directly to a specific dissipation relation like we did for dispersion. Also in the case of DG-FEM both the dispersive and dissipative properties need to be connected closer to the method's HUM results.

Other bases. Although several of our results relied on the formulation of HUM in sine basis, this basis has its limitations and finding alternative bases of modal-type would be an obvious objective for future investigations. If we use semi-discretizations build on polynomial basis functions, it is natural to utilize those as building blocks for numerical HUM and thereby eliminating the loss that we inevitable sustain by translating to and from trigonometric functions. Furthermore, going into 2- and 3-d would also require alternatives to the sinusoids as these will no longer be eigenfunctions in more general and complex geometries.

DG-FEM derivative. We needed to find the normal derivative at the right end of the domain to find the observation of the adjoint problem. The polynomial basis of DG-FEM constituted a challenge in this regard since derivatives of polynomials are well-known to be of low quality at the end-

points. It is possible that the use of a weak filter on the rightmost element could improve this.

PSWFs for DG-FEM. In the end of Section 4.2.5, we suggested the use of so-called prolate spheroidal wave functions (PSWFs) as alternative to the polynomial basis in the DG-FEM formulations. PSWFs are much better suited for the approximation of sinusoids and would probably improve the sine formulation of HUM significantly.

Uniform observability. One of the great strengths of DG-FEM is its strong theoretical foundation. It is highly desirable to back the numerical findings obtained during this project with rigorous numerical analysis, *e.g.,* a convergence analysis.

2- and 3-d problems. The employment of HUM for the solution of a control problem in 1-d reminds about the saying "take not a musket to kill a butterfly". But as explained previously, the ultimate goal is indeed higher dimensional control problems in complex domains where HUM really has it advantages. The DG-FEM is well-suited for general geometries, too. The main obstacle seems the use of some other basis function, as explained above, which can be generalized to 2- and 3-d.

APPENDIX A

List of Symbols

The following lists of symbols are not exhaustive, but they contain the most important symbols used in this dissertation.

General notation

Symbol	Description
$\equiv$	equivalence by definition
$:=$, $=:$	equality defining left and right hand side, respectively
$\mathbb{R}$	the real numbers
$\mathbb{N}$	the natural numbers
$\square'$	time derivative of a function, *e.g.*, $u' = \partial u / \partial t$
$\langle a, b \rangle_C$	inner product between a and b in C
$\|a\|_A$	A-norm of $a \in A$
$\langle a', a \rangle_{A^\star, A}$	duality product between $a' \in A^\star$ and $a \in A$.
Δ	The Laplacian operator

HUM for the wave equation, Chapter 2

Symbol	Description	Page
Ω	open, bounded subset of $\mathbb{R}^d$—in 1-d $\Omega = (0,1)$	8
Γ	boundary of Ω	8
Γ_0	control boundary $\Gamma_0 \subset \Gamma$—in 1-d $\Gamma_0 = \{1\}$	8
Σ	time-boundary cylinder $\Sigma := (0,T) \times \Gamma$	8
Σ_0	domain of control $\Sigma_0 := (0,T) \times \Gamma_0$—in 1-d $\Sigma_0 = (0,T)$	8
u	solution to the control system (2.1)	8
u^0, u^1	initial data for the control system	8
κ	control function (Dirichlet boundary condition on Σ_0)	8
ψ	solution to the auxiliary system (2.5)	10
φ	solution to the adjoint system (2.6)	11
φ^0, φ^1	initial data for the adjoint problem	11
$\mathcal{E}$	energy space for the adjoint system $\mathcal{E} := H_0^1(\Omega) \times L^2(\Omega)$	11
$\mathcal{E}^\star$	dual $\mathcal{E}^\star = H^{-1}(\Omega) \times L^2(\Omega)$ of the energy space $\mathcal{E}$	12
$\widetilde{\mathcal{E}}^\star$	control system energy space $\widetilde{\mathcal{E}}^\star := L^2(\Omega) \times H^{-1}(\Omega)$	8
$\mathcal{B}$	boundary space $\mathcal{B} = L^2(\Sigma_0)$	8
$\langle \cdot, \cdot \rangle_{\mathcal{E}^\star, \mathcal{E}}$	duality product between the spaces $\mathcal{E}^\star$ and $\mathcal{E}$	13
$E(t)$	mechnical energy of the adjoint system at time t	12
Φ	observation operator defined in (2.14)	14
Ψ	reconstruction operator $\Psi = \Phi^*$ defined in (2.17)	15
Λ	HUM operator $\Lambda := \Psi \circ \Phi$	16
Π	control operator $\Pi := \Phi \Lambda^{-1}$	17
$(\bar{\varphi}^0, \bar{\varphi}^1)$	solution to the equation $\Lambda(\varphi^0, \varphi^1) = (u^1, -u^0)$	20
$\mathbf{\Phi}$, $\mathbf{\Psi}$, $\mathbf{\Lambda}$	matrix representation of above operators	17
$\mathcal{J}$	HUM energy functional defined in (2.33)	20
$(\bar{\varphi}^0, \bar{\varphi}^1)$	unique minimum of the functional $\mathcal{J}$	20

Approximating solutions to the wave equation, Chapter 3

Symbol	Description	Page
y	solution to the model wave equation (3.1)	22
y^0, y^1	initial data for the wave equation	22
g_0, g_1	Dirichlet boundary data for the wave equation	22
f	right hand side for the wave equation	22
z	auxiliary variable (advection system)	23
p, q	characteristic wave variables	24
y_h	approximate solution to (3.1)	26
$\mathcal{R}_h$	residual for y_h in (3.1)	27
N	number of inner grid points	27
h	uniform grid spacing $h = 1/(N+1)$	27

Continued on the next page

Symbol	Description	Page
$\boldsymbol{y}$	vector of nodal values (or coefficients) of y_h	27
ψ_n^{L}	basis function for linear FEM (hat basis)	29
$\boldsymbol{M}$	mass matrix	29
$\boldsymbol{K}$	stiffness matrix	29
α	parameter for the unified semi-discretization	29
D^k	k'th element $\mathsf{D}^k = (x_{\text{L}}^k, x_{\text{R}}^k)$ in DG-FEM formulation	31
K	number of elements	31
h^k	length $h^k = x_{\text{R}}^k - x_{\text{L}}^k$ of element D^k	34
N_p	number of nodes per element (polynomial order is $N_p - 1$)	32
y_h^k	local approximate solution	32
ℓ_i^k	i'th Lagrange basis polynomial on element D^k	32
$\boldsymbol{y}^k$	vector of nodal values on element D^k	33
$\widehat{\boldsymbol{y}}^k$	vector of modes on element D^k	35
$\boldsymbol{M}^k$	local DG-FEM mass matrix	33
$\boldsymbol{S}^k$	local DG-FEM stiffness matrix	33
$(ay)^*$	numerical flux	34
I	reference element $\mathsf{I} = (-1, 1)$	35
$\widetilde{P}_{n-1}$	normalized Legendre polynomial	35
$\boldsymbol{V}$	Vandermonde matrix	35
$\boldsymbol{D}_r$	differentiation matrix on reference element I	36
$\boldsymbol{p}^k, \boldsymbol{q}^k$	nodal vectors of the approximate characteristics p_h, q_h	38
Δt	time step size	41
M	number of discrete instances of time	41
$\mathcal{L}_h$	right hand side of ODE	40
μ	Courant number $\mu = \Delta t / h$	40
ω	frequency of trial solution	43
ξ	wavenumber of trial solution	43
c	phase velocity	43
c_g	group velocity	43

Numerical HUM, Chapter 4

Symbol	Description	Page
$\boldsymbol{u}(t)$	approximation of $u(t)$ sized $N \times 1$	55
$\boldsymbol{U}(t)$	control system state $\boldsymbol{U}(t) = [\boldsymbol{u}(t), \boldsymbol{u}'(t)]^{\mathsf{T}}$	55
$\boldsymbol{U}^0$	initial data $[\boldsymbol{u}^0, \boldsymbol{u}^1]^{\mathsf{T}}$ for the control system	55
k	control function for the semi-discrete system	55
$\boldsymbol{\mathcal{L}}_h$	system matrix sized $2N \times 2N$	55
$\boldsymbol{B}_h$	boundary matrix sized $2N \times 1$	55
$\boldsymbol{z}(t)$	column vector of size $N \times 1$ approximating $\psi(t)$	56
$\boldsymbol{Z}(t)$	state $\boldsymbol{Z}(t) = [\boldsymbol{z}(t), \boldsymbol{z}'(t)]^{\mathsf{T}}$ for the auxiliary system	56
$\boldsymbol{w}(t)$	column vector of size $N \times 1$ approximating $\varphi(t)$	55
$\boldsymbol{W}(t)$	state $\boldsymbol{W}(t) = [\boldsymbol{w}(t), \boldsymbol{w}'(t)]^{\mathsf{T}}$ for the adjoint system	55

Continued on the next page

Symbol	Description	Page
$\boldsymbol{W}^0$	initial data $[\boldsymbol{w}^0, \boldsymbol{w}^1]^\mathsf{T}$ for the adjoint system	55
$\boldsymbol{C}_h$	output matrix sized $1 \times 2N$	55
$\mathcal{X}$	finite dimensional state space approximating $\mathcal{E}$	54
$\mathcal{X}^\star$	finite dimensional state space approximating $\mathcal{E}^\star$	54
P^{sd}	semi-discrete observation operator approximating Φ	56
R^{sd}	semi-discrete reconstruction operator approximating Ψ	56
L^{sd}	semi-discrete controllability operator approximating Λ	57
$\mathcal{T}$	discrete time space $\mathbb{R}^M$ w norm (4.27) approximating $\mathcal{B}$	60
P	discrete observation operator approximating Φ	60
R	discrete reconstruction operator approximating Ψ	60
L	discrete controllability operator approximating Λ	60
$(\bar{\boldsymbol{w}}^0, \bar{\boldsymbol{w}}^1)$	solution to the HUM equation $\boldsymbol{L}[\boldsymbol{w}^0, \boldsymbol{w}^1]^\mathsf{T} = [\boldsymbol{u}^1, -\boldsymbol{u}^0]^\mathsf{T}$	60
$\boldsymbol{k}$	time discrete control sized $M \times 1$	60
$\boldsymbol{L}$	matrix representations of L assembled by (4.32) or (4.33)	61
$\boldsymbol{L}^i$	submatrix $i = 1, 2, 3$ or 4 of matrix $\boldsymbol{L}$	61
$\boldsymbol{L}_{(N_c)}$	reduced $\boldsymbol{L}$ by N_c sine basis functions	73
$\boldsymbol{P}$	matrix representation $\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}^0 & \boldsymbol{P}^1 \end{bmatrix}$ of operator P	62
$\boldsymbol{R}$	matrix representation $\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}^0 & \boldsymbol{R}^1 \end{bmatrix}$ of operator R	63
e_j^s	j'th sine basis function $e_j^s(x) = \sqrt{2}\sin(j\pi x)$	64
p_j^0, p_j^1	exact P^0 and P^1 observation of e_j^s	64
$\widehat{f}_k$	k'th Fourier sine coefficient of f	66
$\boldsymbol{e}_j^s$	nodal vector with N equidistant samples of e_j^s	66
e_j^{Ls}	linear spline approximation to e_j^s	66
$\widehat{f}_k^{\text{L}}$	linear approximation to $\widehat{f}_k$ by	66
$\boldsymbol{e}_j^{ss}$	nodal vector (DG-FEM) with samples of e_j^s	80
e_j^{ss}	discontinuous piecewise polynomial interpolating $\boldsymbol{e}_j^{ss}$	80
$\widehat{f}_k^{ss}$	k'th Fourier sine coefficient by inner product with $\boldsymbol{e}_j^{ss}$	84
$\boldsymbol{e}_j^{ps}$	nodal vector of values obtained by projection of e_j^s	80
e_j^{ps}	discontinuous piecewise polynomial interpolating $\boldsymbol{e}_j^{ps}$	80
$\widehat{f}_k^{ps}$	k'th Fourier sine coefficient by inner product with $\boldsymbol{e}_j^{ps}$	84
$\boldsymbol{M}_{\text{p}}$	conjugate gradient pre-conditioner	98

The inverse problem, Chapter 5

Symbol	Description	Page
v or v_f	solution to the source problem (5.1)	111
σ	temporal part (known)of the forcing term in (5.1)	111
f	spatial part (unknown) of the forcing term	111
G	forward map $G(f) = \partial_{\Gamma_0} v_f$ for the source problem	112
w or w_f	solution to the auxiliary problem (5.9)	113
λ_k, ϕ_k	k'th eigensolution to the eigenvalue prob. associated (5.9)	113

Continued on the next page

Symbol	Description	Page
$\mathcal{B}^1$	boundary space for the source problem system	111
$\mathcal{B}^0$	boundary space for the auxiliary system	113
$\mathcal{K}$	boundary integral operator with kernel σ	114
Ξ	bounded operator defined by the Volterra equation (5.23)	118
$\mathcal{T}^1$	discrete approximation to $\mathcal{B}^1$ with inner product (5.26)	120
$\mathcal{T}^0$	discrete approximation to $\mathcal{B}^0$ with inner product (4.27)	120
$\boldsymbol{D}$	temporal differentiation matrix sized $M \times M$	120
$\boldsymbol{X_\sigma}$	matrix approximation to Ξ^{-1} sized $M \times M$	121
$\sigma_a, \sigma_c, \sigma_c$	three examples of σ	122
g_l^a, g_l^b, g_l^c	exact soln. to forward problem with above σ and $f = \phi_l$	123
$\boldsymbol{C}^{\cdot,r}$	matrix of reconstructed coefficients by exact controls	125
$\boldsymbol{C}^{\cdot,\mathrm{L}}$	the same but by L-FEM controls	132
$\boldsymbol{C}^{\cdot,\mathrm{DG}}$	the same but by DG-FEM controls	136

APPENDIX B

Mathematical details

B.1 Analytic solution to the forward problem

We consider the forward map G defined by (5.8). The forward problem is solved analytically in 1-d with two different σ for the eigenfunction $f = \phi_k = \sqrt{2}\sin(k\pi x)$. We solve the auxiliary problem (5.9) with $f = \phi_k$ first which gives

$$\partial_{\Gamma_0} w_{\phi_k} = (-1)^k \sqrt{2}\sin(k\pi t), \qquad t \in (0,T),\ k \in \mathbb{N}.$$

Then we map the boundary data by the boundary integral opererator (5.12) by

$$g_k = \int_0^t \sigma(t-s)\partial_{\Gamma_0} w_{\phi_k} \mathrm{d}s, \qquad t \in (0,T),\ k \in \mathbb{N},$$

with the two different σ. The computations of the integrals are done in Maple[1].

[1] http://www.maplesoft.com

Solution with σ_b

Let $\sigma = \sigma_b = \cos(\pi t) + t^2 - 3t + 1$ then the above integral gives the following result.

$$\begin{aligned} g_1^b(t) =& \frac{\sqrt{2}}{2\pi^3}\Big(2\,\cos(\pi\,t)\,\pi^2 - 2\,\pi^2 - 2\,t^2\pi^2 + 4 + 6\,t\pi^2 \\ & - 4\,\cos(\pi\,t) - \sin(\pi\,t)\,\pi^3 t - 6\,\pi\,\sin(\pi\,t)\Big) \end{aligned}$$

for $k = 1$ and

$$\begin{aligned} g_k^b(t) =& \frac{(-1)^k\sqrt{2}}{\pi^3 k^3\,(k^2-1)}\Big(t^2k^4\pi^2 + \cos(\pi\,t)\,\pi^2k^4 + 2 - 2\,k^2 - k^2\pi^2 + 3\,tk^2\pi^2 \\ & - t^2k^2\pi^2 + k^4\pi^2 - 3\,tk^4\pi^2 + 3\,k^3\pi\,\sin(tk\pi) - 2\,\cos(tk\pi)\,\pi^2k^4 \\ & - 3\,k\pi\,\sin(tk\pi) + k^2\pi^2\cos(tk\pi) + 2\,\cos(tk\pi)\,k^2 - 2\,\cos(tk\pi)\Big) \end{aligned}$$

for $k = 2, \ldots$.

Solution with σ_c

For $\sigma_c = \cos(20\pi t)$ we get

$$\begin{aligned} g_k^c(t) =& -\frac{\sqrt{2}\,(-1)^k\,k}{\pi\,(k^2-400)}\Big(-524288\,(\cos(\pi\,t))^{20} + 2621440\,(\cos(\pi\,t))^{18} \\ & - 5570560\,(\cos(\pi\,t))^{16} + 6553600\,(\cos(\pi\,t))^{14} - 4659200\,(\cos(\pi\,t))^{12} \\ & + 2050048\,(\cos(\pi\,t))^{10} - 549120\,(\cos(\pi\,t))^{8} + 84480\,(\cos(\pi\,t))^{6} \\ & - 6600\,(\cos(\pi\,t))^{4} + 200\,(\cos(\pi\,t))^{2} - 1 + \cos(tk\pi)\Big) \end{aligned}$$

for $k \neq 20$. In the special case $k = 20$ we have

$$\begin{aligned} g_{20}^c(t) =& 2\sqrt{2}t\sin(\pi\,t)\cos(\pi\,t)\Big(131072\,(\cos(\pi\,t))^{18} - 589824\,(\cos(\pi\,t))^{16} \\ & + 1114112\,(\cos(\pi\,t))^{14} - 1146880\,(\cos(\pi\,t))^{12} + 698880\,(\cos(\pi\,t))^{10} \\ & - 256256\,(\cos(\pi\,t))^{8} + 54912\,(\cos(\pi\,t))^{6} - 6336\,(\cos(\pi\,t))^{4} \\ & + 330\,(\cos(\pi\,t))^{2} - 5\Big). \end{aligned}$$

APPENDIX C

The Matlab package

IPHUM1DWAVE is the name of the Matlab[1] package that accompany this dissertation. During the course of this project more than 200 Matlab files have been developed; a few of these are collected in IPHUM1DWAVE.

IPHUM1DWAVE is organized in 5 modules.

Module WAVE: for the solution of the 1-d wave equation by the unified discretization (including L-FEM).

Module HUM: for the solution of HUM boundary control by construction of $\boldsymbol{L}$ or by conjugate gradients. HUM depends on WAVE.

Module DGWAVE: for the solution of the 1-d wave equation by DG-FEM. DGWAVE depends on Matlab module CODES1D[2] from [HW08].

Module DGHUM: as HUM but with DG-FEM discretization. DGHUM depends on DGWAVE.

[1] http://www.mathworks.com
[2] http://www.caam.rice.edu/~timwar/NUDG/Book/Software.html

Module `IP`**:** for the solution of the inverse source problem with HUM. `IP` depends on `WAVE`+`HUM` and/or `DGWAVE`+`DGHUM`.

We will describe the basic functionality of these modules below. For each module we give a short user guide, examples of use, and a one-line summary of the core functions. The complete documentation and all files can be obtained at

http://www.mat.dtu.dk/people/J.S.Mariegaard/software/

The Matlab files comes bundled in a `.zip`-file (including the necessary files from [HW08]—see copyright file), and after download and extraction, `IPHUM1DWAVE` is ready to use. The user starts Matlab, change directory to the folder with the extracted files, and types

```
>> startup
```

and instructions for further use follows. The user types `>> help [function name]` to get information about the use a particular function or module.

C.1 Module `WAVE`

The purpose of `WAVE` is to discretize and solve the 1-d wave equation by the unified scheme (3.22). A number of different time integration schemes are provided (see Section 3.4).

C.1.1 Function summary

`PLOTWAVE`	Plot solution of 1-d wave equation
`RHSWAVE`	Compute rhs of 1st order syst. wave eq $Y' = f(t, Y)$
`SOLVEU`	Solve 1-d wave equation with $u(t, 1) = k(t)$
`SOLVEW`	Solve 1-d wave eq with homogeneous BCs
`SOLVEWAVE`	Solve 1-d wave eq with Dirichlet BCs
`WAVEGLOBALS`	Declare all global variables for wave solver
`WAVESTARTUP`	Discretization for the 1-d wave equation

It should be noted that `SOLVEU` and `SOLVEW` only are "shells" passing data to the actual solver `SOLVEWAVE`; function `SOLVEU` find approximate solution to u-system (2.1) and `SOLVEW` find approximate solution to w-system (2.6). Before solving the user needs to discretize space and time by `WAVESTARTUP`.

C.1.2 Short user guide

(a) Discretize space and time

1) Declare globals by calling `WAVEGLOBALS`
2) Enter final time `T`, grid spacing `h` and Courant number `mu`
3) Decide spatial discretization by parameter `alp` ($= \alpha$)

`alp`=0: 2nd order central finite difference (FDM)
`alp`=1/12: higher order (Störmer-Numerov)
`alp`=1/6: linear FEM (L-FEM)
`alp`=1/4: mixed FEM

4) Run `WAVESTARTUP` (sets up geometry, matrices, etc.)

(b) Solve wave equation

1) Enter discrete initial data $\boldsymbol{w}^0$ and $\boldsymbol{w}^1$ as "functions" of `xs`
2) Enter boundary conditions (if non-zero) $\boldsymbol{g}_1$ as "function" of `tvec`
3) Choose time integration method `odemthd`

`odemthd='cfd'`: explicit mid-point rule
`odemthd='trapez'`: trapezoidal rule (implicit),
`odemthd='newmark'`: Newmark method (options `bet` and `gam`),
`odemthd='RK5'`: 5 stage ERK (option `Mass`),
`odemthd='ode45'`: build-in ode45 (options: `Mass`, `AbsTol`, etc.)

4) Solve with `SOLVEWAVE`, `SOLVEW` or `SOLVEU`

(c) Post-process

1) Plot solution with `PLOTWAVE`
2) Examine error, *e.g.,* using mass matrix `Mh` etc.

C.1.3 Examples of use

Matlab code C.1: A simple wave equation with homogeneous boundary conditions by FDM and trapezoidal time integration (default)

```
%% (a) discretize space and time
waveglobals();                          % declare all globals
h=0.05;  T=4;                           % grid space h=dx; final time T
mu=0.5;                                 % dt = CFL*dx
alp=0;                                  % FDM
wavestartup(mu);                        % set-up geometry, matrices, etc

%% (b,c) wave equation + plot
w0 = sin(2*pi*xs);                      % initial data w(0,x) = w0
w1 = 0*xs;                              % initial data w'(0,x)= w1
[ts,W]=solveW(w0,w1);                   % solve wave equation
plotwave(W,ts)                          % plot solution
```

C.2 Module HUM

The purpose of `HUM` is to solve the discretized boundary control problem for the 1-d wave equation by numerical HUM. The module allows solving the HUM problem by either construction the matrix $\boldsymbol{L}$ (see Section 4.2) or iteratively by conjugate gradients (see Section 4.3).

C.2.1 Function summary

`BSFILT`	Filter by projection onto set of basis functions
`CGHUM`	Solve HUM by conjugate gradients
`HUMLAM`	Solve HUM by construction of L matrix
`LAMSIN`	Construct L matrix in sine basis
`OBSSIN`	Compute discrete observation P of sine basis vectors
`SOLVEPSI`	Solve backward wave equation (Psi)

The most important functions for the end-user are `CGHUM` (algorithm MCG-HUM) and `HUMLAM` (HUM by $\boldsymbol{L}$ construction) which both solves the HUM control problem given the initial data.

C.2.2 Short user guide

(a) Discretize space and time as in C.1.2(a).

(b) Control problem

1) Enter discrete initial data $\boldsymbol{u}^0$ and $\boldsymbol{u}^1$ as "functions" of `xs`
2) Choose time integration method as global `godemthd` (as `odemthd` in C.1.2(b))
3) Solve control problem with `HUMLAM` or `CGHUM`

(c) Post-process

1) Use `SOLVEU` to test found control, plot result with `PLOTWAVE`.
2) Compute norms on output $u(T,x)$ and $u'(T,x)$ with mass matrix `Mh`

C.2.3 Examples of use

Matlab code C.2: HUM solution to boundary control problem; solution by construction of L-FEM discretized matrix $\boldsymbol{L}$ in sine basis.

```
%% (a) discretize space and time
global godemthd;
waveglobals();                          % declare all globals
h=0.02;  T=2.13;                        % grid space h=dx; final time T
mu=0.5;                                 % Courant number dt = mu*dx
alp=1/6;                                % L-FEM
wavestartup(mu);                        % set-up geometry, matrices, etc.

%% (b) control problem
u0 = sin(pi*xs).^4.*sin(5*pi*xs);       % initial data
u1 = 0.*xs;                             % -
godemthd = 'trapez';                    % time integration methods
Lmthd = 'sin';    Nc = floor(N/2);      % type of L construction, filter
[e0,e1,k,L,P]=humlam(u0,u1,Lmthd,Nc);% HUM solution

%% (c) post-processing
[ts,U] = solveU(u0,u1,k);               % test found control
uT = U(1:N,end);                        % u(T,x)
L2uT= sqrt(uT'*Mh*uT),                  % L2 norm of output
L2k = sqrt(L20Tinprod(k,k)),            % L2 norm of control
```

Matlab code C.3: Changes to be made in the above code to use MCG-HUM algorithm instead.

```
tol = 1e-6;     filtfrac = 1/2;         % CG-tolerance and filter
[e0,e1,k,resid] = cghum(u0,u1,tol,[],[],filtfrac); % MCG-HUM algo.
```

C.3 Module `DGWAVE`

The purpose of `DGWAVE` is to discretize and solve the 1-d wave equation by DG-FEM (see Section 3.3.3). The standard wave equation is transformed to a

system of two de-coupled advection equations in characteristic variables p and q (see Section 3.1.1).

C.3.1 Function summary

DGEVAL	Evaluate DG-function f on points xx
DGH1IP	Compute $H^1(\Omega)$-inner product for two DG-functions
DGINT	Compute anti-derivative of DG-function
DGL2IP	Compute $L^2(\Omega)$-inner product for two DG-functions
DGSTARTUP	Set up DG-FEM discretization
DGWAVEPQ	Solve pq-wave equation by DG-FEM
DGWAVERHSPQ	Compute rhs of pq-system with DG-FEM

For the end-user the functions DGSTARTUP and DGWAVEPQ are the most important; they correspond to WAVESTARTUP and SOLVEWAVE of the WAVE module. Note that the function DGINT can be used to find the anti-derivative of the initial data $y'(0,x) = y^1(x)$ needed for the initial data for the p,q-system.

C.3.2 Short user guide

(a) Discretize space and time

1) Declare globals by WAVEGLOBALS (from module WAVE) and GLOBALS1D (from module Codes1D)
2) Enter final time T and Courant number mu
3) Enter number of elements K and points per element Np
4) Run DGSTARTUP (sets up geometry, matrices, etc.)

(b) Wave equation

1) Enter discrete initial data $\boldsymbol{y}^0$ and $\boldsymbol{y}^1$ as "functions" of x
2) Enter boundary conditions (if non-zero) $\boldsymbol{g}_1$ as "function" of tvec
3) Solve with DGWAVEPQ (uses LSERK time integration)

(c) Post-process

1) Plot solution (use, *e.g.,* DGEVAL)
2) Examine error (use, *e.g.,* DGL2IP or DGH1IP), etc.

C.3.3 Examples of use

Matlab code C.4: A wave equation with homogeneous boundary conditions solved by DG-FEM.

```
%% (a) Discretize space and time
clear all;
waveglobals;    Globals1D();          % load all wave globals
T = 2;          mu=0.6;               % final time and Courant num
dgstartup(6,10,mu);                   % DG-FEM discretiz. Np=6; K=10

%% (b) Wave equation
w0 = sin(2*pi*x);                     % initial data w(0,x) = w0
w1 = 0*x;                             % initial data w'(0,x)= w1
[ts,W]=dgwavepq(w0,w1);               % solve wave equation
```

C.4 Module DGHUM

The purpose of DGHUM is to solve the DG-FEM discretized boundary control problem for the 1-d wave equation by numerical HUM. It has functionality parallel to HUM but due to the different types of spatial discretization; module HUM cannot be used with DGWAVE.

C.4.1 Function summary

DGCGHUM	Solve HUM by conjugate gradients (DG-FEM)
DGBSFILT	Filter by projection onto set of basis functions
DGHUMLAM	Solve HUM by construction of L matrix
DGLAMSIN	Construct L matrix in sine basis by DG-FEM
DGOBSSIN	Compute discrete observation P of sine basis vectors
DGPSI	Solve backward wave equation (Psi)

The most important functions for the end-user are DGCGHUM (DG-FEM implementation of algorithm MCG-HUM) and DGHUMLAM (HUM by DG-FEM $\boldsymbol{L}$ construction) which both solves the HUM control problem given the initial data.

C.4.2 Short user guide

(a) Discretize space and time as with DGWAVE C.3.2(a).

(b) Control problem

1) Enter discrete initial data $\boldsymbol{u}^0$ and $\boldsymbol{u}^1$ as "functions" of x
2) Solve control problem with DGHUMLAM or DGCGHUM

(c) Post-process

1) Use DGWAVEPQ to test found control
2) Compute L^2 norm of final state $(u(T,x), u'(T,x))$ by use of DGL2IP

C.4.3 Examples of use

Matlab code C.5: DG-FEM discretized HUM solution to the boundary control problem; solution by construction of matrix $\boldsymbol{L}$ in sine basis.

```
%% (a) Discretize space and time
clear all;
waveglobals;    Globals1D();          % load all wave globals
T = 2.13;      mu=0.6;                % final time and Courant num
dgstartup(6,10,mu);                   % DG-FEM discretiz. Np=6; K=10

%% (b) control problem
u0 = sin(pi*x).^4.*sin(5*pi*x);       % initial data
u1 = 0.*x;                            % -
Nc = floor(N/3);                      % type of L construction, filter
[e0,e1,k,L,P]=dghumlam(u0,u1,Nc);     % HUM solution

%% (c) post-processing
[ts,U] = dgwavepq(u0,u1,k);           % test found control
uT = U(1:Np*K,end);                   % u(T,x)
L2uT= sqrt(dgL2ip(uT,uT)),            % L2 norm of output
L2k = sqrt(L20Tinprod(k,k)),          % L2 norm of control
```

Matlab code C.6: Changes to be made in the above code to use MCG-HUM algorithm instead.

```
10 tol = 1e-6;     filtfrac = 1/3;       % CG-tolerance and filter
11 [e0,e1,k,resid]=dgcghum(u0,u1,tol,[],[],filtfrac); % MCG-HUM algo
```

C.5 Module IP

The purpose of IP is to solve the inverse source problem (ISP) for the 1-d wave equation. The solution consists of reconstructing the Fourier coefficients of the unknown spatial part f of the source term.

C.5.1 Function summary

H1OTINPROD	Compute $H^1(0,T)$-inner product for two functions
IPDATA	Generate random Fourier coefs for inverse problem
IPFORWARD	Compute numerical solution to forward problem
IPFORWARDEX	Compute exact solution to forward problem
IPHUMBASIS	Construct HUM eigenfunction control-basis
IPPLOTCOEF	Plot reconstructed and original coefficients
IPSIGMA	Return function handle for a sigma function
IPVOLTERRA	Construct Volterra matrix for Volterra BIE

C.5.2 Short user guide

(a) Discretize space and time by either WAVE (see Section C.1.2(a)) or DGWAVE (see Section C.3.2(a))

(b) Forward problem

1) Provide original f, *e.g.,* randomly by IPDATA
2) Provide σ and its derivative as function handles, *e.g.,* with IPSIGMA
3) Generate g by solving forward problem by IPFORWARD or IPFORWARDEX

(c) Reconstruction formula

1) Decide cut-off index Nc
2) Compute eigenfunction controls eta by IPHUMBASIS w method hummthd
   hummthd='anal': analytic controls (when $T = 2$)
   hummthd='lamsin': $\boldsymbol{L}$ in sine basis
   hummthd='cghum': MCG-HUM for WAVE
   hummthd='dglamsin': $\boldsymbol{L}$ in sine basis (DG-FEM)
   hummthd='dgcghum': MCG-HUM for DG-FEM
3) Construct Volterra matrix Xsg by IPVOLTERRA
4) Compute "IP-basis" tht by solving Xsg*tht = eta
5) Compute reconstructed coefficients by $H^1(0,T)$-inner product between g and tht

(d) Post-process

1) Compare reconstructed and original coefficients with IPPLOTCOEF

C.5.3 Examples of use

Matlab code C.7: An inverse problem with 25 coefficients, σ_b and exact generated boundary data. Reconstruction by L-FEM eigenfunction controls.

```
%% (a) Set-up wave-environment
clear all;
global godemthd phi;
waveglobals();                          % load all wave globals
T = 2;      h = 0.02;     CFL=0.353;    % discretization
alp = 1/6;    godemthd='trapez';        % L-FEM + trapezoidal rule
wavestartup(CFL);                       % set-up system

%% (b) Generate data for forward problem
Nc = 25;                                % number of fourier coefs.
[cex] = ipdata(Nc);                     % coeff of original f
sgm='b';                                % sigma-b
[sgfun,dsgfun]=ipsigma(sgm);            % sigma function handles
g = ipforwardex(cex,sgm,tvec);          % forward problem

%% (c) Reconstruction formula
%Nc = 25;                               % cut-off index = num of coefs
[eta,phi]=iphumbasis(Nc,'lamsin');      % eigenfunction controls
[Xsg]=ipvolterra(sgfun(tvec),dsgfun(tvec));  % Volterra matrix
tht = (Xsg\(-eta'))';                   % "IP-basis"
cr = H10Tinprod(g,tht);                 % reconstructed coefficients

%% (d) Plot results
ipplotcoef(cex,cr);                     % compare coefficients
```

Bibliography

[AL98] M. Asch and G. Lebeau. Geometrical aspects of exact boundary controllability for the wave equation—a numerical study. *ESAIM Control Optim. Calc. Var.*, 3:163–212 (electronic), 1998. (Cited on page 98.)

[Asm05] N. H. Asmar. *Partial differential equations*. Pearson Prentice Hall, New Jersey, second edition, 2005. with Fourier series and boundary value problems. (Cited on page 39.)

[BLR92] C. Bardos, G. Lebeau, and J. Rauch. Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary. *SIAM J. Control Optim.*, 30(5):1024–1065, 1992. (Cited on page 11.)

[Boy01] J. P. Boyd. *Chebyshev and Fourier spectral methods*. Dover Publications Inc., Mineola, NY, second edition, 2001. (Cited on page 31.)

[Boy04] J. P. Boyd. Prolate spheroidal wavefunctions as an alternative to Chebyshev and Legendre polynomials for spectral element and pseudospectral algorithms. *J. Comput. Phys.*, 199(2):688–716, 2004. (Cited on page 97.)

[BS02] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2002. (Cited on page 29.)

[CF03] M. J. Corless and A. E. Frazho. *Linear Systems and Control: An Operator Perspective*. CRC Press, 2003. (Cited on pages 53 and 54.)

[CK98] D. Colton and R. Kress. *Inverse acoustic and electromagnetic scattering theory*, volume 93 of *Applied Mathematical Sciences*. Springer-Verlag, Berlin, second edition, 1998. (Cited on page 123.)

[CM06] C. Castro and S. Micu. Boundary controllability of a linear semi-discrete 1-D wave equation derived from a mixed finite element method. *Numer. Math.*, 102(3):413–462, 2006. (Cited on pages 29, 30 and 104.)

[CMM08] C. Castro, S. Micu, and A. Münch. Numerical approximation of the boundary control for the wave equation with mixed finite elements in a square. *IMA J. Numer. Anal.*, 28(1):186–214, 2008. (Cited on page 98.)

[EHN96] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996. (Cited on page 109.)

[Eva98] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998. (Cited on page 22.)

[FG94] H. G. Feichtinger and K. Gröchenig. Theory and practice of irregular sampling. In *Wavelets: mathematics and applications*, Stud. Adv. Math., pages 305–363. CRC, Boca Raton, FL, 1994. (Not cited.)

[GHJ06] M. D. Gunzburger, L. S. Hou, and L. Ju. A numerical method for exact boundary controllability problems for the wave equation. *Comput. Math. Appl.*, 51(5):721–750, 2006. (Cited on page 105.)

[GKW89] R. Glowinski, W. Kinton, and M. F. Wheeler. A mixed finite element formulation for the boundary controllability of the wave equation. *Internat. J. Numer. Methods Engrg.*, 27(3):623–635, 1989. (Cited on pages 98 and 104.)

[GL95] R. Glowinski and J.-L. Lions. Exact and approximate controllability for distributed parameter systems (II). In *Acta numerica, 1995*, Acta Numer., pages 159–333. Cambridge Univ. Press, Cambridge, 1995. (Cited on page 10.)

[GLL90] R. Glowinski, C. H. Li, and J.-L. Lions. A numerical approach to the exact boundary controllability of the wave equation. I. Dirichlet controls: description of the numerical methods. *Japan J. Appl. Math.*, 7(1):1–76, 1990. (Cited on pages 52, 98, 99 and 104.)

[Glo92] R. Glowinski. Ensuring well-posedness by analogy: Stokes problem and boundary control for the wave equation. *J. Comput. Phys.*, 103(2):189–221, 1992. (Cited on page 104.)

[Han98] P. C. Hansen. *Rank-deficient and discrete ill-posed problems*. SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Numerical aspects of linear inversion. (Cited on page 123.)

[HGG07] J. S. Hesthaven, S. Gottlieb, and D. Gottlieb. *Spectral methods for time-dependent problems*, volume 21 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2007. (Cited on pages 81 and 85.)

[HS52] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49:409–436 (1953), 1952. (Cited on page 98.)

[HW08] J. S. Hesthaven and T. Warburton. *Nodal discontinuous Galerkin methods*, volume 54 of *Texts in Applied Mathematics*. Springer, New York, 2008. Algorithms, analysis, and applications. (Cited on pages 32, 34, 35, 37, 43, 45, 49, 101, 157 and 158.)

[Ise96] A. Iserles. *A first course in the numerical analysis of differential equations*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 1996. (Cited on page 40.)

[IZ99] J. A. Infante and E. Zuazua. Boundary observability for the space semi-discretizations of the 1-D wave equation. *M2AN Math. Model. Numer. Anal.*, 33(2):407–438, 1999. (Cited on page 104.)

[JGH03] L. Ju, M. D. Gunzburger, and L. S. Hou. Approximation of exact boundary controllability problems for the 1-D wave equation by optimization-based methods. In *Recent advances in scientific computing and partial differential equations (Hong Kong, 2002)*, volume 330 of *Contemp. Math.*, pages 133–153. Amer. Math. Soc., Providence, RI, 2003. (Cited on page 105.)

[Joh82] F. John. *Partial differential equations*, volume 1 of *Applied Mathematical Sciences*. Springer-Verlag, New York, fourth edition, 1982. (Cited on page 113.)

[Kel76] J. B. Keller. Inverse problems. *Amer. Math. Monthly*, 83(2):107–118, 1976. (Cited on page 109.)

[Kre01] S. Krenk. Dispersion-corrected explicit integration of the wave equation. *Computer Methods in Applied Mechanics and Engineering*, 191(8-10):975 – 987, 2001. (Cited on page 40.)

[Kre06a] S. Krenk. Energy conservation in Newmark based time integration algorithms. *Comput. Methods Appl. Mech. Engrg.*, 195(44-47):6110–6124, 2006. (Cited on pages 40 and 41.)

[Kre06b] S. Krenk. State-space time integration with energy control and fourth-order accuracy for linear dynamic systems. *Internat. J. Numer. Methods Engrg.*, 65(5):595–619, 2006. (Cited on page 40.)

[Kre08] S. Krenk. Extended state-space time integration with high-frequency energy dissipation. *Internat. J. Numer. Methods Engrg.*, 73(12):1767–1787, 2008. (Cited on page 40.)

[KTE93] D. Kosloff and H. Tal-Ezer. A modified Chebyshev pseudospectral method with an $O(N^{-1})$ time step restriction. *J. Comput. Phys.*, 104(2):457–469, 1993. (Cited on page 97.)

[Kut01] W. Kutta. Beitrag zur näherungsweisen integration totaler differentialgleichungen. *Zeitschr. für Math. u. Phys.*,, 46:435–453, 1901. (Cited on page 42.)

[Lio88] J.-L. Lions. Exact controllability, stabilization and perturbations for distributed systems. *SIAM Rev.*, 30(1):1–68, 1988. (Cited on pages 2, 8, 11 and 15.)

[LLT86] I. Lasiecka, J.-L. Lions, and R. Triggiani. Nonhomogeneous boundary value problems for second order hyperbolic operators. *J. Math. Pures Appl. (9)*, 65(2):149–192, 1986. (Cited on page 12.)

[LR56] P. D. Lax and R. D. Richtmyer. Survey of the stability of linear finite difference equations. *Comm. Pure Appl. Math.*, 9:267–293, 1956. (Cited on page 37.)

[Mic02] S. Micu. Uniform boundary controllability of a semi-discrete 1-D wave equation. *Numer. Math.*, 91(4):723–768, 2002. (Cited on page 104.)

[Mün04] A. Münch. Famille de schémas implicites uniformément contrôlables pour l'équation des ondes 1-D. *C. R. Math. Acad. Sci. Paris*, 339(10):733–738, 2004. (Cited on page 104.)

[Mün05] A. Münch. A uniformly controllable and implicit scheme for the 1-D wave equation. *M2AN Math. Model. Numer. Anal.*, 39(2):377–418, 2005. (Cited on page 104.)

[MZ05] S. Micu and E. Zuazua. An introduction to the controllability of partial differential equations. In T. Sari, editor, *Quelques questions de théorie du contrôle*, Collection Travaux en Cours, pages 69–157. Hermann, 2005. (Cited on pages 8, 19, 20 and 54.)

[New59] N. M. Newmark. A method of computation for structural dynamics. *Journal of Engineering Mechanics Division*, pages 67–94, 1959. (Cited on page 40.)

[Nic00] S. Nicaise. Exact boundary controllability of Maxwell's equations in heterogeneous media and an application to an inverse source problem. *SIAM J. Control Optim.*, 38(4):1145–1170 (electronic), 2000. (Cited on page 110.)

[NMS06] M. Negreanu, A.-M. Matache, and C. Schwab. Wavelet filtering for exact controllability of the wave equation. *SIAM J. Sci. Comput.*, 28(5):1851–1885 (electronic), 2006. (Cited on page 105.)

[NZ03] M. Negreanu and E. Zuazua. A 2-grid algorithm for the 1-d wave equation. In *Mathematical and numerical aspects of wave propagation—WAVES 2003*, pages 213–217. Springer, Berlin, 2003. (Cited on page 104.)

[NZ04a] M. Negreanu and E. Zuazua. Convergence of a multigrid method for the controllability of a 1-d wave equation. *C. R. Math. Acad. Sci. Paris*, 338(5):413–418, 2004. (Cited on page 104.)

[NZ04b] S. Nicaise and O. Zaïr. Determination of point sources in vibrating beams by boundary measurements: identifiability, stability, and reconstruction results. *Electron. J. Differential Equations*, pages No. 20, 17 pp. (electronic), 2004. (Cited on page 110.)

[Ped00] M. Pedersen. *Functional analysis in applied mathematics and engineering*. Studies in Advanced Mathematics. Chapman & Hall/CRC, Boca Raton, FL, 2000. (Cited on pages 8, 11, 13, 16 and 23.)

[Ped08] M. Pedersen. Boundary control of plates. Preprint, 2008. (Cited on pages 8, 10, 13 and 15.)

[Ras04] J. M. Rasmusssen. *Boundary Control of Linear Evolution PDEs - Continuous and Discrete*. PhD thesis, DTU Informatics, Technical University of Denmark, 2004. (Cited on pages 13, 18, 29, 30, 40, 47, 53, 58 and 105.)

[Run95] C. Runge. Ueber die numerische Auflösung von Differentialgleichungen. *Math. Ann.*, 46(2):167–178, 1895. (Cited on page 42.)

[Son98] E. D. Sontag. *Mathematical control theory*, volume 6 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 1998. Deterministic finite-dimensional systems. (Cited on page 54.)

[SP61] D. Slepian and H. O. Pollak. Prolate spheroidal wave functions, Fourier analysis and uncertainty. I. *Bell System Tech. J.*, 40:43–63, 1961. (Cited on page 97.)

[Tre82] L. N. Trefethen. Group velocity in finite difference schemes. *SIAM Rev.*, 24(2):113–136, 1982. (Cited on page 44.)

[Tre00] L. N. Trefethen. *Spectral methods in MATLAB*, volume 10 of *Software, Environments, and Tools*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000. (Cited on page 31.)

[VB82] R. Vichnevetsky and J. B. Bowles. *Fourier analysis of numerical approximations of hyperbolic equations*, volume 5 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1982. (Cited on pages 30, 31 and 44.)

[XRY01] H. Xiao, V. Rokhlin, and N. Yarvin. Prolate spheroidal wavefunctions, quadrature and interpolation. *Inverse Problems*, 17(4):805–838, 2001. Special issue to celebrate Pierre Sabatier's 65th birthday (Montpellier, 2000). (Cited on page 97.)

[Yam95] M. Yamamoto. Stability, reconstruction formula and regularization for an inverse source hyperbolic problem by a control method. *Inverse Problems*, 11(2):481–496, 1995. (Cited on pages 3, 110, 111, 113, 114, 115, 116, 118 and 142.)

[Yam96] M. Yamamoto. On ill-posedness and a Tikhonov regularization for a multidimensional inverse hyperbolic problem. *J. Math. Kyoto Univ.*, 36(4):825–856, 1996. (Cited on page 119.)

[Zua05] E. Zuazua. Propagation, observation, and control of waves approximated by finite difference methods. *SIAM Rev.*, 47(2):197–243 (electronic), 2005. (Cited on pages 8, 43, 51 and 73.)