Preface

This dissertation is submitted to the Technical University of Denmark in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

The work has been supervised by Professor Per Christian Hansen, Informatics and Mathematical Modelling, and Associate Professor Michael Pedersen, Department of Mathematics.

I would like to thank my supervisors for their supportive, friendly and easy-going attitude. The subject of my study was fairly new to us all, and they let me move in any direction which I found interesting, and shape my study more or less as I wanted.

A special thanks goes out to Professor Enrique Zuazua from Universidad Autónoma de Madrid, who supervised my seven month stay in Madrid. It was a unique opportunity for me to study the research of Professor Zuazua and his colleagues and students. I learned a lot and found the focus points of my own research during this stay.

I also thank Michael Jacobsen and Jesper Grooss with whom I, in turn, shared a DTU office. Apart from being good friends, they patiently listened to my research problems and theories, whenever I needed some feedback. I furthermore thank Henrik Pilegaard, Mikael Buchholtz and René R. Hansen for good company during lunchtime.

I finally thank both Jesper Grooss and my big brother Thomas for good advice and for helping me proofread this thesis.

Kongens Lyngby, August 2, 2004

Jan Marthedal Rasmussen

2810 04#2

Abstract

Consider a partial differential equation (PDE) of evolution type, such as the wave equation or the heat equation. Assume now that you can influence the behavior of the solution by setting the boundary conditions as you please. This is boundary control in a broad sense.

A substantial amount of literature exists in the area of theoretical results concerning control of partial differential equations. The results have included existence and uniqueness of controls, minimum time requirements, regularity of domains, and many others.

Another huge research field is that of control theory for ordinary differential equations. This field has mostly concerned engineers and others with practical applications in mind.

This thesis makes an attempt to bridge the two research areas. More specifically, we make finite dimensional approximations to certain evolution PDEs, and analyze how properties of the discrete systems resemble the properties of the continuous system.

A common framework in which the continuous systems are formulated will be provided. The treatment includes many types of linear evolution PDEs and boundary conditions. We also consider different types of controllability, such as approximate, null- and exact controllability.

We will consider discrete systems with a viewpoint similar to that used for the continuous systems. Most importantly, we study what is required of a discretization scheme in order for computed control functions to converge to the true, continuous, control function. Examples exist for convergent discretization schemes for which *divergence* of the computed controls occur.

We dig deeper for three specific cases: The heat equation, the wave equation, and a linear system of thermoelasticity. Different aspects of the theory are exemplified through these case studies.

We finally consider how to efficiently implement computer programs for computing controls in practice.

Resumé

Betragt en tidsafhængig, partiel differentialligning (PDE), såsom bølgeligningen eller varmeledningsligningen. Antag nu, at man kan påvirke løsningens opførsel ved at justere på randbetingelserne efter behov. Dette er randkontrol i bred forstand.

En anselig mængde litteratur omhandler teoretiske resultater for kontrol af partielle differentialligninger. Disse resultater omfatter eksistens og entydighed af kontrolfunktioner, tidskrav, domæners regularitet og mange andre.

Et andet stort forskningsområde er kontrol/regulering af ordinære differentialligninger. Dette område optager mest ingeniører og andre, der arbejder med praktiske anvendelser.

Denne afhandling forsøger at bygge bro mellem de to forskningsområder. Vi vil, mere konkret, foretage endelig-dimensionelle tilnærmelser til visse tidsafhængige PDE'er, og analysere hvorledes egenskaber for de diskrete systemer tilnærmer det kontinuerte systems egenskaber.

Vi præsenterer et fælles teoretisk grundlag for de kontinuerte systemer, som vi vil betragte. Dette vil omfatte mange lineære, tidsafhængige PDE'er og forskellige randbetingelser. Vi vil også studere forskellige typer af kontrollérbarhed, såsom nul- og eksakt kontrollérbarhed.

Vi betragter diskrete systemer fra den samme vinkel som for de kontinuerte systemer. Vi studerer også det vigtige spørgsmål om hvad der kræves af en diskretisering, for at beregnede kontrolfunktioner konvergerer mod den korrekte, kontinuerte, kontrolfunktion. Eksempler findes hvor de beregnede kontrolfunktioner, på trods af at en konvergent diskretisering benyttes, *divergerer*.

Vi graver dybere for tre bestemte ligninger: Varmeledningsligningen, bølgeligningen og et lineært termoelasticitets-system. Forskellige aspekter af den etablerede teori vil blive konkretiseret gennem disse eksempler.

Endelig vil vi betragte hvorledes man, på en effektiv måde, kan implementere computerprogrammer, der kan beregne kontrolfunktioner i praksis.

Contents

Chapter 1

Introduction

It's a control freak thing.
I wouldn't let you understand.
— S. H. Underwood

One type of control system is probably already working in your home right now: The thermostat of your refrigerator. How does it work? If you put into your fridge some hot dish, the temperature of the fridge rises. The thermostat senses this and starts cooling. On the other hand, if the temperature gets too low, the cooling system is shut down, and the surroundings of the fridge will make the temperature rise. This kind of control is called *bang-bang control*, and is a simple type of adaptive control, a self-adjusting system.

Many types of control exist. A system exposed to bang-bang control will typically keep on oscillating about a desired state in some way (the state being, for instance, the temperature of the fridge). One could also consider *stabilizing* a system, where a control aims to make all oscillations of a solution disappear as time goes by.

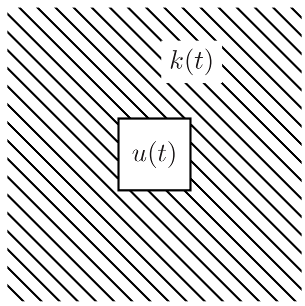


Figure 1.1: The temperature $u(t)$ of a small object is controlled by the surrounding temperature $k(t)$.

This thesis focuses on *exact controllability* for ordinary differential equations, ODEs, and, in particular, partial differential equations, PDEs. Given such a system with some initial state, and where we are allowed to control the system in some way, we want to steer the solution *exactly* to some desired state at a *specific time*. What happens thereafter, is not important.

As an example, let us consider a simple ordinary differential equation:

$$\begin{cases} u'(t) = \alpha(k(t) - u(t)) \,, \\ u(0) = u^0 \,. \end{cases} \tag{1.1}$$

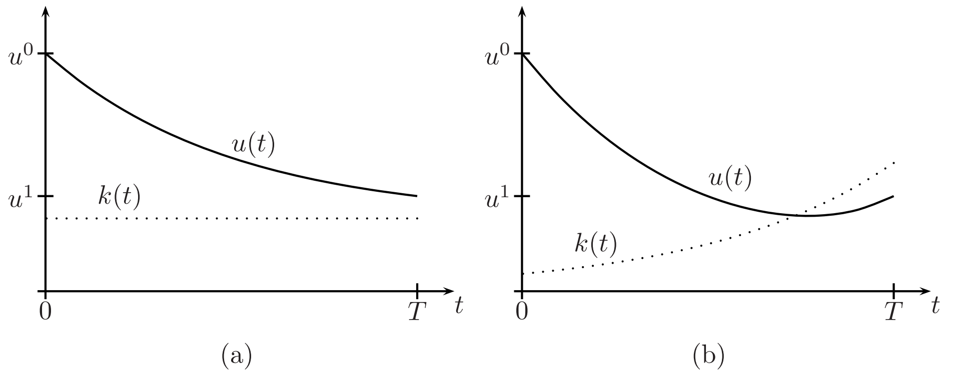


Figure 1.2: Two different controls $k(t)$ that do the same job: Lead the state $u(t)$ from $u(0) = u^0$ to $u(T) = u^1$. Figure (b) on the right furthermore shows the unique control of minimal $L^2(0,T)$-norm.

The state $u(t)$ denotes the temperature of some small object at time t. The whole object is assumed to have the same temperature throughout. The surroundings has everywhere the temperature $k(t)$ at time t, and α is some positive heat transfer constant. Initially, at time $t = 0$, the object has the temperature u^0. See Figure 1.1.

Assume now that we can control the temperature $k(t)$ of the surroundings and that we would like the object at time $t = T$ to have a certain temperature, say $u(T) = u^1$. Since we can write the solution to (1.1) explicitly,

$$u(t) = e^{-\alpha t}u^0 + \alpha e^{-\alpha t}\int_0^t e^{\alpha s}k(s)ds\,,$$

this is not difficult to achieve. Inserting the information we have, we just have to find $k(t)$ such that

$$\alpha\int_0^T e^{\alpha t}k(t)dt = e^{\alpha T}u^1 - u^0\,.$$

For instance, if we seek a constant valued control function, $k(t) = k_0$, we easily arrive at

$$k(t) = k_0 = \frac{e^{\alpha T}u^1 - u^0}{e^{\alpha T} - 1}\,,$$

see Figure 1.2(a).

There are obviously an infinite number of controls $k(t)$ that steer the temperature from u^0 to u^1. One may be interested in finding a control that is *optimal* in some sense. For instance, what if we want to find a control that has minimal $L^2(0,T)$-norm, that is, make the quantity

$$\int_0^T |k(t)|^2 dt$$

the smallest possible? Such a control does exist and is in fact unique. It is furthermore the perhaps easiest optimal control to find. It turns out that in this case, the control function has the form

$$k(t) = v^0 e^{\alpha(t-T)} \quad \text{where} \quad v^0 = 2\frac{u^1 - e^{-\alpha T}u^0}{1 - e^{-2\alpha T}} \,,$$

see Figure 1.2(b). We shall later see how to compute such an optimal control, as part of a much more general theory for partial differential equations.

Note two things in this example: A control can be found no matter how small the final time $T > 0$ is, and the initial and final states can be arbitrary real numbers.

We can clearly extend the previous concepts to higher order differential equations and multidimensional systems. In such cases we can arrive at a common setting of the form

$$\begin{cases} \boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t) + \boldsymbol{B}\boldsymbol{k}(t) \,, \\ \boldsymbol{u}(0) = \boldsymbol{u}^0 \,, \end{cases}$$

where $\boldsymbol{A}$ is a square matrix and $\boldsymbol{B}$ has any number of columns. We now seek a control that steers the system from state $\boldsymbol{u}^0$ to some state $\boldsymbol{u}^1$ at time $t = T$. When $\boldsymbol{A}$ is diagonalizable, each eigenmode of $\boldsymbol{A}$ can either be controlled arbitrarily fast or not at all. This depends on the choice of $\boldsymbol{B}$, of course. The case of a non-diagonalizable $\boldsymbol{A}$ is slightly more complicated, but it is still easily analyzed whether one has controllability or not. The point is, everything is known about controllability of ODEs of the above form.

We now increase the difficulty considerably. We go from an ODE to a PDE, and consider a string on the interval $(0, 1)$. When the transversal oscillations are relatively small, the movements of such a string can be described by a simple linear PDE, typically denoted the *wave equation*:

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = c^2 \dfrac{\partial^2 u}{\partial x^2} \,, \\ u(0, x) = u^0(x), \quad \dfrac{\partial u}{\partial t}(0, x) = \overline{u}^0(x) \,, \\ u(t, 0) = 0, \quad u(t, 1) = k(t) \,, \end{cases}$$

for $0 \le t \le T$ and $0 \le x \le 1$. The constant c is related to the material of the string and represents the speed with which waves of a solution propagate. The initial state, at time $t = 0$, is dictated by u^0 and $\overline{u}^0$. The most interesting quantities here are the boundary conditions. The left end-point is fixed at position 0, but the position of the right end-point is determined by the function $k(t)$. This is the control and since it acts through a boundary condition, it is called a boundary control. (One should of course be precise about the function spaces in which we operate, but we will postpone such details until later.) See Figure 1.3 for an illustration.

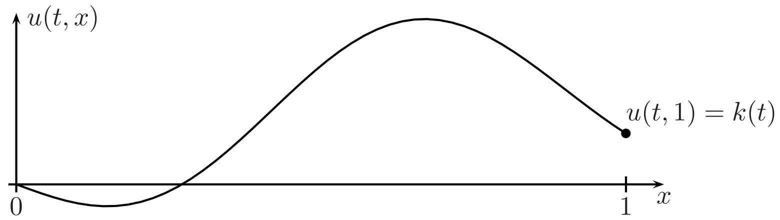


Figure 1.3: The state of a string at some time t. The left end-point is fixed at 0, while the position of the right end-point is determined by $k(t)$.

Our control problem is similar to the previous one: We wish to determine $k(t)$ such that we reach a particular state at time $t = T$,

$$u(T,x) = u^1(x), \quad \frac{\partial u}{\partial t}(T,x) = \overline{u}^1(x)\,.$$

Is this possible no matter the initial conditions, u^0 and $\overline{u}^0$? The answer depends on how much time we have: If $T < 2/c$, the answer is no; if $T \geq 2/c$, we can steer the solution to any final state (again, in appropriate function spaces). This makes sense for the following two reasons: We can only control the solution through a single point, the right end-point, and the fact that waves propagate with constant speed c. How long does it take for a pulse to travel from the right end-point, to be reflected at the left end-point and to travel back again? Exactly $2/c$ time units, and this is the reason for the requirement on the control time.

Let us turn our attention to another, in its formulation, simple PDE. It is the well-known *heat equation*:

$$\begin{cases} \dfrac{\partial u}{\partial t} = c\dfrac{\partial^2 u}{\partial x^2}\,, \\ u(0,x) = u^0(x)\,, \\ u(t,0) = 0, \quad u(t,1) = k(t)\,, \end{cases}$$

for $0 \leq t \leq T$ and $0 \leq x \leq 1$. The quantity $u(t,x)$ denotes the temperature at time t and at position x, in a rod of unit length. The c is a physical constant related to the thermal properties of the material in question.

The heat equation is, however, very different from the wave equation, also when it comes to controllability. Let us again consider the question: Is it possible for any initial state u^0 and final state u^1 to find a control $k(t)$ such that

$$u(T,x) = u^1(x)\,?$$

The answer is no. It is possible, though, always to find a control that steers any initial state u^0 to the *zero state* (this type of controllability is called null-controllability). This can, as opposed to the wave equation, be done *arbitrarily*

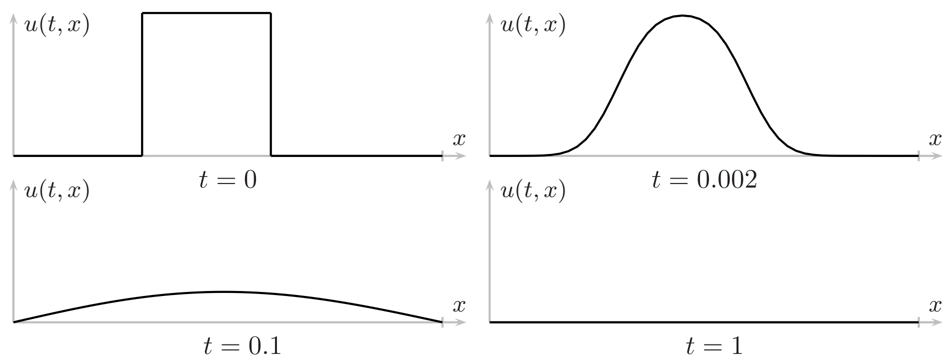


Figure 1.4: Illustration of the smoothing effect of the heat equation. The initial state at $t = 0$ is non-continuous, but every state with $t > 0$ is infinitely smooth. On top of this, the solution quickly approaches the zero state.

fast. Figure 1.4 provides a hint as to why this is so. Here is illustrated a solution to the heat equation at different times, without any control. The initial state is a non-continuous "hat" function, but the solution quickly gets very smooth. Actually, the solution is infinitely smooth for any time $t > 0$. This makes it impossible to steer the solution to any non-smooth state (non-smooth meaning that some derivative is non-continuous).

The heat equation also has a strong damping effect that makes the solution strive towards the zero state. This makes it possible, given a suitable control, always to steer the solution to the zero state.

So when it comes to possible final states, the heat equation is more restrictive than the wave equation. But when it comes to the control time T, the heat equation has no restrictions. Any state can be driven to zero arbitrarily fast. This is possible because for the heat equation, the temperature at one point can affect the temperature at another point arbitrarily fast.

Let us turn our attention to discretizations. In order to fix ideas we will consider a simple finite difference discretization of the wave equation.

A robust and constructive method called HUM (Hilbert Uniqueness Method) exists in the continuous case for finding a control that steers the solution to a given final state. This method also applies for finite dimensional systems and, in particular, for discretizations of the wave equation. One might, quite sensibly, make the following hypothesis: Using a convergent approximation of the wave equation, boundary derivatives and other "ingredients" of HUM, the discrete approximations of the control must converge to the true, continuous one, as time and space steps go to zero. This is not true in general! Understanding why this is so and how to make sure the controls do converge is the *main theme* of this thesis.

Let us give some pointers as to why it can go wrong. Consider Figure 1.5. It illustrates wave propagation according to a finite difference approximation of the one dimensional wave equation. The true solution should be an exact translation to the right of the initial state, the "peak" shown in gray. The numerical scheme

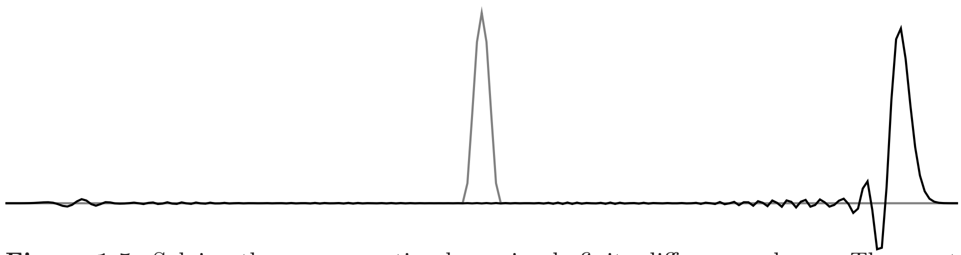


Figure 1.5: Solving the wave equation by a simple finite difference scheme. The exact solution should be a translation to the right of the initial wave, shown in gray. The numerical scheme approximates this solution, but also introduces some spurious waves travelling at wrong speeds and some even in the wrong direction.

approximates this solution, shown in black, but we note some spurious waves. Some highly oscillatory waves travel too slowly, both to the left and right. Actually, if the grid-point distance is h, high frequency waves will travel at a speed of order h in this discrete medium. This is very bad for control! As we have argued for the continuous wave equation, the speed with which waves propagate is essential when it comes to how must time is needed for control. And here, in the discrete setting just described, as $h \to 0$ we are going to need more and more time. This means that if we compute controls for a fixed control time, say T, the controls will *diverge*.

1.1 Structure of the Thesis

Style and Structure are the essence of a book;
great ideas are hogwash.
— Vladimir Nabokov (1899-1977)

Chapter 2 kicks off by establishing the theoretical foundation for this thesis. The type of PDE systems with which we will work, central function definitions and theorems.

Chapter 3 concerns discretizations. First we look at discretizing in space. Especially the Laplace operator in one dimension will be considered. Even though this is a relatively simple subject, we need to obtain various important facts, to be used later on. We then move on to discretizing in time, with focus on two (ODE) discretization schemes. We finally introduce the concept of group velocity for discretizations of hyperbolic equations.

Chapter 4 is to some extent a repeat of Chapter 2, but for discrete systems. Some things are different, though, such as function spaces, norms, how to impose boundary conditions, etc. Important tools for proving that computed controls will converge to the true ones will also be given.

Chapter 5 focuses on an essential part of the Hilbert Uniqueness Method, the controllability operator. Methods for computing a matrix representation of this operator will be given. The controllability operator depends on the control time, the amount of time available for control. For some PDEs, the operator turns out to have some interesting properties as this control time approaches infinity.

Chapter 6 is the first of three chapters to consider specific PDE systems. This chapter concerns the heat equation. Different aspects concerning control for both the continuous systems and discretizations thereof are considered.

Chapter 7 turns to the wave equation. Again, we will apply the theory of Chapter 2 to the continuous system. The treatment of controllability for discretizations of the wave equation, and convergence thereof, is considered in this chapter and is one of the central subjects of the thesis.

Chapter 8 considers a linear system of thermoelasticity, which can be considered a coupling between a heat equation and a wave equation. We consider only the continuous case in one dimension and show, using the knowledge gathered in the previous chapters, how to prove a result concerning boundary controllability of this system.

Chapter 9 goes through an actual implementation of how to compute boundary controls for a two dimensional wave equation. Algorithm complexity, memory usage and other practical aspects will be addressed.

Chapter 10 finalizes by providing an overview of the most important contributions of this thesis. Undoubtably, several areas touched upon are worth digging further into. Several such areas, open questions and also new, but related, research subjects will be mentioned.

Appendix A contain theorems, proofs and derivations that will be referred to from the main text. They have been placed in the appendix in an attempt to not drown the reader in too many detailed derivations that are not essential when reading. They have been included for the interested reader because the results are either not found elsewhere in the literature, or because they are so important that they have been included for completeness.

Appendix B presents a quick reference guide to the notation used throughout the thesis.

CHAPTER 2

Boundary Control of Linear Evolution PDEs

There is nothing more practical than a good theory.
— LEONID ILICH BREZHNEV (1977)

We begin by laying the theoretical foundation for boundary control. This includes introducing the types of evolution equations, whose solutions we wish to control, but also defining the different types of controllability we shall consider.

Many results of this chapter are already known, but most of them have been presented in connection with a particular equation, such as the wave equation or the heat equation. We present here a unified approach, that in an abstract setting formulates a number of results that can easily be applied to a specific PDE system.

Let us note here that Partial Differential Equations, especially in a control context, are often referred to as *Distributed Parameter Systems* in the literature *(Zuazua, 2002b)*.

2.1 Setting the Stage

Let Ω be an open and bounded subset of the d-dimensional Euclidean space $\mathbb{R}^d$. We denote the boundary $\Gamma = \partial\Omega$, and a subset of the boundary, $\Gamma_0 \subset \Gamma$, will be referred to as the *control boundary*. Given $T > 0$, we introduce the time–space cylinders $Q = (0,T) \times \Omega$, $\Sigma = (0,T) \times \Gamma$ and $\Sigma_0 = (0,T) \times \Gamma_0$, for shorter notation.

Consider the linear system of partial differential equations:

$$\begin{cases} u_t = \mathcal{A}u & \text{in } Q \ , \\ \mathcal{B}u = \begin{cases} k & \text{in } \Sigma_0 \ , \\ 0 & \text{in } \Sigma \setminus \Sigma_0 \ , \end{cases} \\ u(0) = u^0 & \text{in } \Omega \ , \end{cases} \tag{2.1}$$

where $u^0 \in H'$. Here, H' is a Hilbert space and u_t denotes the time derivative of u, the state vector. The operator $\mathcal{A}$ is a partial differential operator, generator of a strongly continuous semigroup $e^{t\mathcal{A}} : H' \mapsto H'$, $t \geq 0$; the operator $\mathcal{B} : H' \mapsto \Gamma_0$ is a linear *boundary operator* and $k \in L^2(\Sigma_0)$ is the *control (function)*. We introduce the notation

$$u(t) = \mathcal{L}(u^0, k)(t), \quad t \geq 0\,,$$

to emphasize the solution's dependence on the initial data u^0 and the control k. We assume $\mathcal{L}(u^0, k) \in C([0,T]; H')$ for all $u^0 \in H'$ and $k \in L^2(\Sigma_0)$. We will return to the question of how one can show that the system actually is well posed, as is assumed here. The above system will be referred to as the *control system*.

In this chapter we will consider three different, although related, control problems, where a control k is sought such that the solution is steered to, or towards, a given final state.

2.1.1 The Adjoint System

Before studying controllability we need to introduce the so-called *adjoint system*. It turns out that controllability of the control system is equivalent to certain properties of the adjoint system. This section, and the following two, will present the adjoint system and miscellaneous mappings and relations, which connect the control system and the adjoint system (and thereby providing a reason for the name "adjoint").

The formal adjoint operator of $\mathcal{A}$ is the uniquely defined operator $\mathcal{A}^*$ for which

$$\langle \mathcal{A}u, v\rangle_{H'\times H} = \langle u, \mathcal{A}^* v\rangle_{H'\times H}\,, \qquad \forall (u,v) \in H' \times H\,,$$

where $\mathcal{A}$ is considered as an operator with homogeneous boundary conditions, $k = 0$.

It is, however, often impractical to work with this operator in the sense that $\mathcal{A}^*$ often represents a PDE system in a somewhat indirect way (an illustrative example of this can be found in Chapter 8, where a linear system of thermoelasticity is studied). We will therefore consider another operator $\widetilde{\mathcal{A}}$ that represents a system which is essentially equivalent to that of $\mathcal{A}^*$. This is ensured by requiring the relation

$$\widetilde{\mathcal{A}} = \mathcal{M}^{-1}\mathcal{A}^*\mathcal{M}\,,$$

for an invertible matrix $\mathcal{M}$ containing only scalar entries. The eigenvalues of $\widetilde{\mathcal{A}}$ are easily shown to be equivalent to those of $\mathcal{A}^*$, which means that $\widetilde{\mathcal{A}}$ also generates a strongly continuous semigroup $e^{t\widetilde{\mathcal{A}}} : \widetilde{H} \mapsto \widetilde{H}$, where $\widetilde{H}$ is a Hilbert space equipped with the norm

$$\|v\|_{\widetilde{H}} = \|\mathcal{M}v\|_H\,. \tag{2.2}$$

Defining the norm of the dual space $\widetilde{H}'$ in the usual way, one gets $\|v\|_{\widetilde{H}'} = \|\mathcal{M}^{-T}v\|_{H'}$. For convenience we present the following diagram that relates the function spaces H, H', $\widetilde{H}$ and $\widetilde{H}'$:

$$\begin{tikzcd} H & H' \arrow[l, "\tau"'] \arrow[d, "\mathcal{M}^T"] \\ \widetilde{H} \arrow[u, "\mathcal{M}"] & \widetilde{H}' \arrow[l, "\tilde{\tau}"] \end{tikzcd}$$

Here, $\mathcal{M}^T$ is the usual matrix transpose of $\mathcal{M}$, and τ and $\tilde{\tau}$ are the Riesz canonical isometries defined such that

$$\begin{aligned} &\langle u,v\rangle_{H'\times H} = \langle \tau u, v\rangle_H\,, \quad (u,v)\in H'\times H \\ \text{and}\quad &\langle u,v\rangle_{\widetilde{H}'\times\widetilde{H}} = \langle \tilde{\tau} u, v\rangle_{\widetilde{H}}\,, \quad (u,v)\in \widetilde{H}'\times \widetilde{H}\,. \end{aligned}$$

We now introduce a duality pairing $\{\cdot,\cdot\} : H'\times\widetilde{H}\mapsto\mathbb{R}$ in the following way,

$$\{u,v\} = \langle u,\mathcal{M}v\rangle_{H'\times H} = \langle \mathcal{M}^T u, v\rangle_{\widetilde{H}'\times\widetilde{H}} \quad \text{for all } (u,v)\in H'\times\widetilde{H}\,. \tag{2.3}$$

Note that this implies that $\{\mathcal{A}u,v\} = \{u,\widetilde{\mathcal{A}}v\}$ for all $(u,v)\in H'\times\widetilde{H}$ (here again, $\mathcal{A}$ is considered with homogeneous boundary conditions).

We now introduce what we will refer to as the *adjoint system*:

$$\begin{cases} v_t = -\widetilde{\mathcal{A}}v & \text{in } Q\ , \\ \mathcal{B}v = 0 & \text{in } \Sigma\ , \\ v(T) = v^0 & \text{in } \Omega\ , \end{cases} \tag{2.4}$$

where $v^0\in\widetilde{H}$. Similar to the control system, we will use the notation

$$v(t) = \mathcal{A}(v^0)(t), \quad t\le T\,,$$

to emphasize the solution's dependence on the initial data v^0 (note that the initial conditions are given at $t=T$ and the system is solved backwards in time). We will assume $\mathcal{A}(v^0)\in C([0,T];\widetilde{H})$ for all $v^0\in\widetilde{H}$; this must be shown for every concrete system. In fact, we will assume the following about the growth of the solution of the adjoint system,

$$\|\mathcal{A}(v^0)(t)\|_{\widetilde{H}} \le Ce^{\alpha(T-t)}\|v^0\|_{\widetilde{H}}, \quad 0\le t\le T, \quad \text{for all } v^0\in\widetilde{H}\,, \tag{2.5}$$

for some real constants $C,\alpha>0$. Such a bound is possible whenever $\widetilde{\mathcal{A}}$ is the generator of a strongly continuous semigroup (see *Rudin, 1973*, page 356).

The duality of the control system and the adjoint system, through the duality pairing $\{\cdot,\cdot\}$ and thus the matrix $\mathcal{M}$, is in fact what makes the abstract results of this chapter possible. This is a new approach that emerged from the study of the linear system of thermoelasticity, on which we shall focus in Chapter 8.

2.1.2 The Complementary Boundary Operator

We will now introduce the *complementary* boundary operator. This is a common approach in the field of boundary control, see, e.g., *Pedersen (2000)* for the case of the wave equation. We give here a more general definition.

A linear complementary boundary operator $\mathcal{C} : \widetilde{H} \mapsto L^2(\Gamma_0)$ must exist such that the following Green-like formula holds:

$$\{\mathcal{A}u, v\} - \{u, \widetilde{\mathcal{A}}v\} = \langle \mathcal{B}u, \mathcal{C}v\rangle_{L^2(\Gamma_0)} ,$$

for all $u, v \in C^\infty(\overline{\Omega})$. Note that the range of $\mathcal{C}$ must be $L^2(\Gamma_0)$; in fact, the following is required: A function $K : (0,\infty) \mapsto (0,\infty)$ must exist such that

$$\|\mathcal{C}v(\cdot)\|^2_{L^2(\Sigma_0)} = \int_0^T \|\mathcal{C}v(t)\|^2_{L^2(\Gamma_0)} dt \;\le\; K(T)\|v^0\|^2_{\widetilde{H}} , \tag{2.6}$$

for all $T > 0$ and all $v^0 \in \widetilde{H}$ with corresponding solution v of the adjoint system. We also require that the control fulfills $k \in L^2(\Sigma_0)$. The bound (2.6) is commonly called the *direct inequality.*

We can now present an equality that relates solutions of the control system and the adjoint system. It will turn out to be one of the most useful relations when it comes to boundary controllability.

Theorem 2.1.1. *Let $T > 0$ be fixed. A solution $u \in C((0,T); H')$ of the control system (2.1) with control $k \in L^2(\Sigma_0)$ fulfills $u(0) = u^0 \in H'$ and $u(T) = u^1 \in H'$ if and only if*

$$\langle k, \mathcal{C}v\rangle_{L^2(\Sigma_0)} + \{u^0, v(0)\} - \{u^1, v(T)\} = 0 , \tag{2.7}$$

holds for all solutions $v \in C((0,T); \widetilde{H})$ of the adjoint system (2.4) with initial data $v^0 = v(T) \in \widetilde{H}$.

Proof. Let $T > 0$, $v^0, v^1 \in H'$ and $k \in L^2(\Sigma_0)$ be given.

Assume that $u \in C((0,T); H')$ is a solution of the control system with control k and $u(0) = u^0$, $u(T) = u^1$. Consider now a solution $v \in C((0,T); \widetilde{H})$ of the adjoint system for a fixed, but arbitrary, $v^0 \in \widetilde{H}$. Observe then that

$$\begin{aligned}\big[\{u, v\}\big]_0^T &= \int_0^T \big(\{u_t, v\} + \{u, v_t\}\big) dt = \int_0^T \big(\{\mathcal{A}u, v\} - \{u, \widetilde{\mathcal{A}}v\}\big) dt \\ &= \int_0^T \langle k, \mathcal{C}v\rangle_{L^2(\Gamma_0)} dt ,\end{aligned} \tag{2.8}$$

which is exactly Equation (2.7), since $v^0 \in \widetilde{H}$ was chosen arbitrarily.

Assume now that (2.7) holds for all $v^0 \in \widetilde{H}$. Using initial condition u^0 and control k for the control system, we now get, using (2.8), that

$$\langle k, \mathcal{C}v\rangle_{L^2(\Sigma_0)} + \{u^0, v(0)\} - \{u(T), v(T)\} = 0 ,$$

for all solutions v of the adjoint system. Subtracting this equality from (2.7) yields

$$\begin{aligned} \{u(T)-u^1, v^0\} &= 0, \qquad \forall v^0 \in \widetilde{H} \qquad \Leftrightarrow \\ \langle u(T)-u^1, w^0\rangle_{H'\times H} &= 0, \qquad \forall w^0 \in H\,, \end{aligned}$$

so $u(T) = u^1$. □

We will now return to the question of showing the well-posedness of the control system (2.1). Consider the following expression,

$$\begin{aligned} \|u(T)\|_{H'} &= \sup_{w^0\in H\setminus\{0\}} \frac{\big|\langle u(T), w^0\rangle_{H'\times H}\big|}{\|w^0\|_H} = \sup_{w^0\in H\setminus\{0\}} \frac{\big|\{u(T), \mathcal{M}^{-1}w^0\}\big|}{\|\mathcal{M}^{-1}w^0\|_{\widetilde{H}}} \\ &= \sup_{v^0\in \widetilde{H}\setminus\{0\}} \frac{\big|\{u(T), v^0\}\big|}{\|v^0\|_{\widetilde{H}}} \le \sup_{v^0\in \widetilde{H}\setminus\{0\}} \frac{1}{\|v^0\|_{\widetilde{H}}} \big(|\{u(0), v(0)\}| + \big|\langle k, \mathcal{C}v\rangle_{L^2(\Sigma_0)}\big|\big) \\ &\le \sup_{v^0\in \widetilde{H}\setminus\{0\}} \frac{1}{\|v^0\|_{\widetilde{H}}} \Big(\|u(0)\|_{H'}\|\mathcal{M}v(0)\|_H + \|k\|_{L^2(\Sigma_0)}\|\mathcal{C}v\|_{L^2(\Sigma_0)}\Big) \\ &\le K_1 e^{\alpha T}\|u(0)\|_{H'} + K\|k\|_{L^2(\Sigma_0)}\,. \end{aligned}$$

This shows that the well-posedness of the control system (2.1) can be shown using the solution bound for the adjoint system (2.5), the boundedness of the complementary boundary operator (2.6) and Theorem 2.1.1. This will often be the procedure in practice, see Sections 6.1, 7.1 and 8.1 for examples of this when we study specific control systems.

2.1.3 Important Mappings

For easier notation, we will introduce two bounded and linear maps, L_T and L_T^* for all $T \ge 0$. The map $L_T : H' \mapsto H'$ for $T > 0$ is defined as the solution at $t = T$ for the control system without control,

$$L_T u^0 = \mathcal{L}(u^0, 0)(T)\,,$$

and similarly, $L_T^* : \widetilde{H} \mapsto \widetilde{H}$ for $T > 0$ is defined as the solution at $t = 0$ for the adjoint system,

$$L_T^* v^0 = \mathcal{A}(v^0)(0)\,.$$

Note how L_T and L_T^* are "adjoint" operators in the sense that

$$\{u^0, L_T^* v^0\} = \{L_T u^0, v^0\}\,,$$

for all $(v^0, u^0) \in \widetilde{H} \times H'$ and all $T > 0$, seen easily from Equation (2.7) with $k(t) = 0$.

We introduce two more closely related maps. The first, $G_T : \widetilde{H} \mapsto L^2((0,T)\times\Gamma_0)$ for $T > 0$, applies the complementary boundary operator to a solution of the adjoint system,

$$G_T(v^0) = \mathcal{C}\big(\mathcal{A}(v^0)(\cdot)\big)\,.$$

The second map, $G_T^* : L^2((0,T)\times\Gamma_0) \mapsto \widetilde{H}'$ for $T>0$, takes a control k, applies it to the control system with zero initial conditions, and outputs the state at time $t=T$,

$$G_T^*(k) = \mathcal{M}^T \mathcal{L}(0,k)(T)\,.$$

The premultiplication of $\mathcal{M}^T$ makes sure, as already suggested by the notation, that G_T and G_T^* are adjoint operators for each $T>0$,

$$\begin{aligned}\langle G_T^*(k), v^0\rangle_{\widetilde{H}'\times\widetilde{H}} &= \langle \mathcal{M}^T u(T), v(T)\rangle_{\widetilde{H}'\times\widetilde{H}} = \{u(T), v(T)\}\\ &= \langle k, \mathcal{C}v\rangle_{L^2(\Sigma_0)}\\ &= \langle k, G_T(v^0)\rangle_{L^2(\Sigma_0)}\,,\end{aligned}$$

where relation (2.7) of Theorem 2.1.1 has been used with $u^0=0$.

The authors in *Asch and Lebeau (1998)* (see Appendix A of this paper) introduce maps similar to G_T and G_T^*, which are also observed to be adjoint operators. They were, however, only considering the wave equation.

We finally define the important bilinear form $\gamma_T : \widetilde{H}\times\widetilde{H}\mapsto\mathbb{R}$ as

$$\gamma_T(v^0,w^0) = \langle G_T v^0, G_T w^0\rangle_{L^2(\Sigma_0)} = \int_0^T\int_{\Gamma_0} \mathcal{C}v\,\mathcal{C}w\; d\Gamma\, dt\,. \tag{2.9}$$

This form is easily seen to be indeed bilinear, symmetric, positive semi-definite and bounded (this follows from the inequality (2.6)). An interpretation of this form is that $\gamma_T(v^0,v^0)$ reflects the quantity that is observed from the boundary, through $\mathcal{C}$, during $t\in(0,T)$ of the solution to the adjoint system with initial condition v^0.

Consider the case where

$$\gamma_T(v^0,v^0)=0 \quad\Rightarrow\quad v^0=0 \quad \text{for all } v^0\in\widetilde{H}\,. \tag{2.10}$$

So if $\gamma_T(v^0,v^0)=0$, then $v^0=0$, which in turn means that the corresponding solution is zero at any time $t\in(0,T)$. This property is often called *unique continuation* for the adjoint system. Note that (2.10) is equivalent to the operator G_T having a trivial kernel, $\ker G_T=\{0\}$. Note also that this property depends on both the geometry, the control boundary and the control time T.

2.1.4 Degrees of Controllability

We are now ready to define the three different types of controllability which we will consider. See *Zuazua (2002b)* or *Micu and Zuazua (2004)* for similar definitions.

The first type of controllability, approximate controllability, is the weakest kind and ensures only than we can steer a solution *arbitrarily close* to some desired final state.

Definition 2.1.1. *The control system is* approximately controllable *at time $T>0$ if for every $u^0,u^1\in H'$ and $\epsilon>0$ a control $k\in L^2((0,T)\times\Gamma_0)$ exists such that*

$$\|\mathcal{L}(u^0,k)(T)-u^1\|_{H'}<\epsilon\,.$$

Next is null-controllability, which means that the control system can always be driven exactly to rest, to the zero state.

Definition 2.1.2. *The control system is* null-controllable *at time $T > 0$ if for every $u^0 \in H'$ a control $k \in L^2((0,T) \times \Gamma_0)$ exists such that*

$$\mathcal{L}(u^0, k)(T) = 0\,.$$

Finally, the strongest type of controllability, exact controllability. Here, any initial state can be steered to any final state.

Definition 2.1.3. *The control system is* exactly controllable *at time $T > 0$ if for every $u^0, u^1 \in H'$ a control $k \in L^2((0,T) \times \Gamma_0)$ exists such that*

$$\mathcal{L}(u^0, k)(T) = u^1\,.$$

With the notation introduced by now, let us rewrite relation (2.7) of Theorem 2.1.1:

$$\begin{aligned} &\langle k, G_T v^0 \rangle_{L^2(\Sigma_0)} + \{u^0, L_T^* v^0\} - \{u^1, v^0\} = 0 \quad \Leftrightarrow \\ &\langle G_T^* k, v^0 \rangle_{\widetilde{H}' \times \widetilde{H}} + \{L_T u^0, v^0\} - \{u^1, v^0\} = 0\,, \end{aligned}$$

for all $v^0 \in \widetilde{H}$, which is seen to be just a variational formulation of

$$\begin{aligned} G_T^* k + \mathcal{M}^T L_T u^0 - \mathcal{M}^T u^1 = 0 \quad &\Leftrightarrow \\ G_T^* k = \mathcal{M}^T (u^1 - L_T u^0)\,. & \end{aligned} \tag{2.11}$$

The three types of controllability can now be interpreted as: The equality (2.11) must be satisfied either approximately (approximate controllability), exactly but with $u^1 = 0$ (null-controllability), or for any $u^1 \in H'$ (exact controllability).

2.2 Approximate Controllability

Let us consider approximate controllability in greater detail. Although the topic is not among the main themes of this thesis, it will provide us with some insight that we need later on. Recall the relation (2.11) of the previous section,

$$G_T^* k = \mathcal{M}^T (u^1 - L_T u^0)\,,$$

that holds if and only if the control k steers the control system from u^0 to u^1.

Assume now that the unique continuation property holds, that is, we have $\ker G_T = \{0\}$. This property provides information about the image of the adjoint operator, indeed,

$$\overline{G_T^*(L^2(\Sigma_0))} = (\ker G_T)^\perp = \widetilde{H}'\,, \tag{2.12}$$

where $\overline{\Box}$ denotes set closure (see *Pedersen, 2000*, page 57). Let now $u^0, u^1 \in H'$ and an $\epsilon > 0$ be given. Because of the above relation, a $k \in L^2(\Sigma_0)$ exists such that

$$G_T^* k = \mathcal{M}^T (u^1 - L_T u^0) + r\,,$$

with $\|r\|_{\widetilde{H}'} \leq \epsilon$. The rewrite

$$G_T^* k = \mathcal{M}^T((u^1 + \mathcal{M}^{-T} r) - L_T u^0)\,,$$

now shows that k steers the solution exactly from u^0 to $u^1 + \mathcal{M}^{-T} r$, and thus misses the target u^1 by $\mathcal{M}^{-T} r$ for which $\|\mathcal{M}^{-T} r\|_{H'} = \|r\|_{\widetilde{H}'} \leq \epsilon$.

Since this argument is easily reversed, we have proved the following theorem.

Theorem 2.2.1. *Unique continuation, $\gamma_T(v^0, v^0) = 0 \Rightarrow v^0 = 0$, of the adjoint system is equivalent to having approximate controllability for the control system.*

We will now approach approximate controllability from another angle, namely through the minimization of the functional

$$J_\epsilon(v^0) = \tfrac{1}{2}\gamma_T(v^0, v^0) + \epsilon\|v^0\|_{\widetilde{H}} + \{u^0, L_T^* v^0\} - \{u^1, v^0\}\,. \tag{2.13}$$

The reason for doing this is that some of the following results will be used in later sections.

We first argue that a unique minimizer to J_ϵ exists for every $\epsilon > 0$. This follows if the functional is strictly convex,

$$J_\epsilon(\theta v^1 + (1-\theta)v^2) \;<\; \theta J_\epsilon(v^1) + (1-\theta)J_\epsilon(v^2)\,,$$

for all $v^1, v^2 \in \widetilde{H}$ with $v^1 \neq v^2$ and all $\theta \in (0,1)$, and if it is coercive,

$$J_\epsilon(v_j) \to \infty \text{ for every sequence } \langle v_j \rangle \text{ for which } \|v_j\|_{\widetilde{H}} \to \infty\,,$$

see *Lions (1971)*, page 8.

The strict convexity of J_ϵ, for any $\epsilon \geq 0$, is clearly shown if the functional $v \mapsto \gamma_T(v, v)$ is strictly convex.

Theorem 2.2.2. *If $\gamma_T(v^0, v^0) = 0$ implies $v^0 = 0$ for all $v^0 \in \widetilde{H}$, then γ_T is strictly convex.*

Proof. We wish to show

$$\gamma_T\big(\theta v^1 + (1-\theta)v^2, \theta v^1 + (1-\theta)v^2\big) < \theta\gamma_T(v^1, v^1) + (1-\theta)\gamma_T(v^2, v^2)\,,$$

for every choice of $\theta \in (0,1)$ and all $v^1, v^2 \in \widetilde{H}$ for which $v^1 \neq v^2$. Using the bilinearity of γ_T, this expression is seen to be equivalent to

$$\gamma_T(v^1 - v^2, v^1 - v^2) > 0\,,$$

for all $v^1, v^2 \in \widetilde{H}$ for which $v^1 \neq v^2$. But this is exactly the unique continuation property which is assumed. □

We now turn to show the coercivity, which is a little harder. The proof proceeds as in *Zuazua (1997)*.

Theorem 2.2.3. *Let $\epsilon > 0$ and assume $\gamma_T(v^0, v^0) = 0 \Rightarrow v^0 = 0$ for all $v^0 \in \widetilde{H}$. Then the functional J_ϵ, given by (2.13), is coercive and in fact,*

$$\lim_{\|v^0\|_{\widetilde{H}} \to \infty} \frac{J_\epsilon(v^0)}{\|v^0\|_{\widetilde{H}}} \geq \epsilon \,.$$

Proof. Let $v_1^0, v_2^0, \ldots \in \widetilde{H}$ be a sequence for which $\|v_j^0\|_{\widetilde{H}} \to \infty$ as $j \to \infty$. Let $\overline{v}_1^0, \overline{v}_2^0, \ldots$ be the corresponding normalized sequence,

$$\overline{v}_j^0 = v_j^0 / \|v_j^0\|_{\widetilde{H}} \,.$$

We now have

$$\frac{J_\epsilon(v_j^0)}{\|v_j^0\|_{\widetilde{H}}} = \tfrac{1}{2}\gamma_T(\overline{v}_j^0, \overline{v}_j^0)\|v_j^0\|_{\widetilde{H}} + \epsilon + \{L_T u^0, \overline{v}_j^0\} - \{u^1, \overline{v}_j^0\} \,. \tag{2.14}$$

We will now consider the following two cases separately.

Case 1:

$$\liminf_{j\to\infty} \gamma_T(\overline{v}_j^0, \overline{v}_j^0) > 0 \,.$$

In this case we clearly have $\liminf_{j\to\infty} J_\epsilon(v_j^0)/\|v_j^0\|_{\widetilde{H}} = \infty$.

Case 2:

$$\liminf_{j\to\infty} \gamma_T(\overline{v}_j^0, \overline{v}_j^0) = 0 \,.$$

Since the sequence $\langle \overline{v}_j^0 \rangle$ is bounded, we can extract a weakly convergent subsequence (also indexed by j, for ease of notation),

$$\overline{v}_j^0 \rightharpoonup v^0 \quad \text{weakly in } \widetilde{H} \text{ for } j \to \infty \,,$$

for which, by assumption,

$$\gamma_T(\overline{v}_j^0, \overline{v}_j^0) \to 0 \quad \text{for } j \to \infty \,.$$

The solution corresponding to the limit data v^0 thus fulfills $\gamma_T(v^0, v^0) = 0$, which, using the assumption, implies that $v^0 = 0$. So we have

$$\overline{v}_j^0 \rightharpoonup 0 \quad \text{weakly in } \widetilde{H} \text{ for } j \to \infty \,.$$

This makes the two last terms of (2.14) go to zero, and the result follows. □

We are now ready for the following important theorem. A similar result can be found in *Micu and Zuazua (2004)*, for the specific case of the wave equation.

Theorem 2.2.4. *Let $\gamma_T(v^0, v^0) = 0 \Rightarrow v^0 = 0$ for all $v \in \widetilde{H}$. Then the functional J_ϵ has a unique minimizer $\hat{v}^0$ for every choice of $u^0, u^1 \in H'$ and $\epsilon > 0$. Furthermore, when applying the controls $k = G_T \hat{v}^0$ to the control system, we have*

$$\|u(T) - u^1\|_{H'} \leq \epsilon \,. \tag{2.15}$$

Proof. Let $\epsilon > 0$. The necessary and sufficient conditions for the existence and uniqueness of a minimizer $\hat{v}^0$ have already been established. To show the second part of the theorem, we split into two cases.

The case $\hat{v}^0 = 0$: Let an arbitrary $v^0 \neq 0$ be given. Then for all $\alpha > 0$ we have

$$J_\epsilon(\alpha v^0) = \tfrac{1}{2}\alpha^2\gamma_T(v^0, v^0) + \epsilon\alpha\|v^0\|_{\widetilde{H}} + \alpha\{u^0, L_T^* v^0\} - \alpha\{u^1, v^0\} > J(\hat{v}^0) = 0\,,$$

which, when dividing by α and using the positivity of $\gamma_T(v^0, v^0)$, implies

$$\epsilon\|v^0\|_{\widetilde{H}} + \{u^0, L_T^* v^0\} - \{u^1, v^0\} \geq 0\,. \tag{2.16}$$

Using the null-controls on the control system implies, in particular,

$$\{u^0, L_T^* v^0\} - \{u(T), v^0\} = 0\,.$$

Subtracting this equality from the inequality (2.16) gives

$$\{u^1 - u(T), v^0\} \leq \epsilon\|v^0\|_{\widetilde{H}}\,.$$

We finally get

$$\begin{aligned}
\|u(T) - u^1\|_{H'} &= \sup_{w^0 \in H\setminus\{0\}} \frac{\left|\langle u(T) - u^1, w^0\rangle_{H'\times H}\right|}{\|w^0\|_H} \\
&= \sup_{w^0 \in H\setminus\{0\}} \frac{\left|\{u(T) - u^1, \mathcal{M}^{-1}w^0\}\right|}{\|\mathcal{M}^{-1}w^0\|_{\widetilde{H}}} \\
&= \sup_{v^0 \in \widetilde{H}\setminus\{0\}} \frac{\left|\{u(T) - u^1, v^0\}\right|}{\|v^0\|_{\widetilde{H}}} \leq \sup_{v^0 \in \widetilde{H}\setminus\{0\}} \frac{\epsilon\|v^0\|_{\widetilde{H}}}{\|v^0\|_{\widetilde{H}}} = \epsilon\,.
\end{aligned}$$

The case $\hat{v}^0 \neq 0$: Because of the optimality condition we get by formal differentiation (see Detail 1, page 179),

$$\langle \nabla J_\epsilon(\hat{v}^0), w^0\rangle = \gamma_T(\hat{v}^0, w^0) + \frac{\epsilon}{\|\hat{v}^0\|_{\widetilde{H}}}\langle \hat{v}^0, w^0\rangle + \{u^0, L_T^* w^0\} - \{u^1, w^0\} = 0\,,$$

for all $w^0 \in \widetilde{H}$. Using now

$$\langle \hat{v}^0, w^0\rangle_{\widetilde{H}} = \langle \tilde{\tau}^{-1}\hat{v}^0, \mathcal{M}^{-1}\mathcal{M}w^0\rangle_{\widetilde{H}'\times\widetilde{H}} = \{\mathcal{M}^{-T}\tilde{\tau}^{-1}\hat{v}^0, w^0\}\,,$$

(the operator $\tilde{\tau}^{-1} : \widetilde{H} \mapsto \widetilde{H}'$ satisfies $\langle \tilde{\tau}^{-1}v, w\rangle_{\widetilde{H}'\times\widetilde{H}} = \langle v, w\rangle_{\widetilde{H}}$ for all $w \in \widetilde{H}$) we get that for all $w^0 \in \widetilde{H}$,

$$\gamma_T(\hat{v}^0, w^0) + \{u^0, L_T^* w^0\} - \Big\{u^1 - \frac{\epsilon}{\|\hat{v}^0\|_{\widetilde{H}}}\mathcal{M}^{-T}\tilde{\tau}^{-1}\hat{v}^0, w^0\Big\} = 0\,.$$

Using the result of Theorem 2.1.1 we see that the control induced by $\hat{v}^0$ drives u^0 to the state $u^1 - \epsilon\mathcal{M}^{-T}\tilde{\tau}^{-1}\hat{v}^0/\|\hat{v}^0\|_{\widetilde{H}}$. By computing $\|\epsilon\mathcal{M}^{-T}\tilde{\tau}^{-1}\hat{v}^0\|_{H'}/\|\hat{v}^0\|_{\widetilde{H}} = \epsilon$ we see that (2.15) actually holds with equality. □

Note how the second part of the theorem, the inequality (2.15), implies approximate controllability. So we have now, as promised, shown Theorem 2.2.1 in an alternative way.

2.3 Null-controllability

We now turn to null-controllability, the task of steering a solution exactly to zero at time $t = T$. The following theorem provides sufficient and necessary conditions for null-controllability. A similar result can be found in *Fernández-Cara and Zuazua (2002)* for the one dimensional heat equation with variable coefficients. The proof that follows was communicated to the author by Professor Zuazua *(Zuazua, 2002a)*.

Theorem 2.3.1. *Let $T > 0$ be fixed. A linear and bounded operator $K_T^n : H' \mapsto L^2(\Sigma_0)$ exists for which*

$$\mathcal{L}(u^0, K_T^n(u^0))(T) = 0, \quad \textit{for all } u^0 \in H',$$

if and only if there is a constant $C_n > 0$ such that

$$\|L_T^* v^0\|_{\widetilde{H}}^2 \leq C_n \gamma_T(v^0, v^0), \quad \textit{for all } v^0 \in \widetilde{H}. \tag{2.17}$$

Proof. **Null-controllability $\Rightarrow$ observability inequality**. Observe from relation (2.7) that the following must hold,

$$-\{u^0, L_T^* v^0\} = \langle K_T u^0, G_T v^0 \rangle_{L^2(\Sigma_0)},$$

for all $v^0 \in \widetilde{H}$ and all $u^0 \in H'$. We now get

$$\begin{aligned} \|\mathcal{M} L_T^* v^0\|_H &= \sup_{u^0 \in H' \setminus \{0\}} \frac{\left|\langle u^0, \mathcal{M} L_T^* v^0 \rangle_{H' \times H}\right|}{\|u^0\|_{H'}} = \sup_{u^0 \in H' \setminus \{0\}} \frac{\left|\langle K_T u^0, G_T v^0 \rangle_{L^2(\Sigma_0)}\right|}{\|u^0\|_{H'}} \\ &\leq \sup_{u^0 \in H' \setminus \{0\}} \frac{\|K_T u^0\|_{L^2(\Sigma_0)} \|G_T v^0\|_{L^2(\Sigma_0)}}{\|u^0\|_{H'}} = \|K_T\| \, \gamma_T(v^0, v^0)^{1/2}, \end{aligned}$$

for all $v^0 \in \widetilde{H}$, so (2.17) holds with $C_n = \|K_T\|^2$.

Observability inequality $\Rightarrow$ null-controllability. We will first show the existence of a minimizer for the functional

$$J(v^0) = \tfrac{1}{2} \gamma_T(v^0, v^0) + \{u^0, L_T^* v^0\}. \tag{2.18}$$

This will be done by considering a sequence of minimizers $\langle \hat{v}_\epsilon^0 \rangle$ of the functional

$$J_\epsilon(v^0) = \tfrac{1}{2} \gamma_T(v^0, v^0) + \epsilon \|v^0\| + \{u^0, L_T^* v^0\},$$

for $\epsilon \to 0$.

For fixed $\epsilon > 0$, we know from the previous section that J_ϵ possesses a unique minimizer $\hat{v}_\epsilon^0$ for which the induced control drives the initial state u^0 to a final state where (recall that we aim to hit $u^1 = 0$)

$$\|u(T)\| \leq \epsilon. \tag{2.19}$$

Since $J_\epsilon(0) = 0$, we have

$$\begin{aligned} 0 \geq J_\epsilon(\hat{v}_\epsilon^0) &= \tfrac{1}{2}\gamma_T(\hat{v}_\epsilon^0, \hat{v}_\epsilon^0) + \epsilon\|\hat{v}_\epsilon^0\| + \{u^0, L_T^*\hat{v}_\epsilon^0\} \\ &\geq \tfrac{1}{2}\gamma_T(\hat{v}_\epsilon^0, \hat{v}_\epsilon^0) - \left|\{u^0, L_T^*\hat{v}_\epsilon^0\}\right| \geq \tfrac{1}{2}\gamma_T(\hat{v}_\epsilon^0, \hat{v}_\epsilon^0) - \|u^0\|_{H'}\|\mathcal{M}L_T^*\hat{v}_\epsilon^0\|_H , \end{aligned}$$

which implies, when using the assumption (2.17),

$$\gamma_T(\hat{v}_\epsilon^0, \hat{v}_\epsilon^0)^2 \leq 4\|u^0\|_{H'}^2\|\mathcal{M}L_T^*\hat{v}_\epsilon^0\|_H^2 \leq 4C\|u^0\|_{H'}^2\gamma_T(\hat{v}_\epsilon^0, \hat{v}_\epsilon^0) \quad \Leftrightarrow$$

$$\gamma_T(\hat{v}_\epsilon^0, \hat{v}_\epsilon^0) \leq 4C\|u^0\|_{H'}^2 .$$

From this final expression we see that the $L^2(\Sigma_0)$-norms of the controls are bounded *uniformly* in ϵ (recall that $\gamma_T(v,v)^{1/2}$ precisely is the $L^2(\Sigma_0)$-norm of the corresponding control).

This implies that $\hat{v}_\epsilon^0 \to \hat{v}^0$ as $\epsilon \to 0$, where $\hat{v}^0$ is a (local) minimizer of the functional J, see (2.18). Because of the bound in (2.19) we see that $\hat{v}^0$ indeed induces a control driving the initial state to zero. That $\hat{v}^0$ furthermore is the unique (global) minimizer follows easily from the fact that J is strictly convex. $\square$

The unique minimizer $\hat{v}^0$ obtained by minimizing the functional J in (2.18) induces a control $k = G_T\hat{v}^0$ that solves a given null-controllability problem. We shall see shortly, in Section 2.5, that this control has minimal $L^2(\Sigma_0)$-norm among all controls that solve the same controllability problem.

2.4 Exact controllability

Moving on to the strongest form of controllability, exact controllability, we again consider the statement (2.7) of Theorem 2.1.1,

$$\begin{aligned} \langle k, G_Tv^0\rangle_{L^2(\Sigma_0)} + \{u^0, L_T^*v^0\} - \{u^1, v^0\} = 0 , \\ \langle k, G_Tv^0\rangle_{L^2(\Sigma_0)} - \{u^1 - L_Tu^0, v^0\} = 0 , \end{aligned}$$

that holds for all $v^0 \in \widetilde{H}$, if and only if the control k steers a solution from $u(0) = u^0$ to $u(T) = u^1$. This relation implies that having exact controllability is *equivalent* to being able to steer the zero state to any final state in H' (compare to Definition 2.1.3). More specifically, a control steering from u^0 to u^1 will also steer the system from 0 to $u = u^1 - L_Tu^0$, and vice versa. This is, of course, a consequence of the linearity of the underlying systems.

We now have the following sufficient and necessary conditions for exact controllability. This theorem is well known for the wave equation and can be found in, e.g., *Lions (1988b)*.

Theorem 2.4.1. *Let $T > 0$ be fixed. A linear and bounded operator $K_T^e : H' \mapsto L^2(\Sigma_0)$ exists for which*

$$\mathcal{L}(0, K_T^e(u))(T) = u, \quad \textit{for all } u \in H' ,$$

if and only if there is a constant $C_e > 0$ such that

$$\|v^0\|^2_{\widetilde{H}} \le C_e \gamma_T(v^0, v^0), \quad \textit{for all } v^0 \in \widetilde{H}\,. \tag{2.20}$$

Proof. **Exact controllability $\Rightarrow$ observability inequality** Observe that the following holds,

$$\{u, v^0\} = \langle K_T u, G_T v^0 \rangle_{L^2(\Sigma_0)}\,,$$

for all $v^0 \in \widetilde{H}$ and all $u \in H'$. We now get

$$\begin{aligned}\|\mathcal{M} v^0\|_H &= \sup_{u \in H' \setminus \{0\}} \frac{\left|\langle u, \mathcal{M} v^0 \rangle_{H' \times H}\right|}{\|u\|_{H'}} = \sup_{u \in H' \setminus \{0\}} \frac{\left|\langle K_T u, G_T v^0 \rangle_{L^2(\Sigma_0)}\right|}{\|u\|_{H'}} \\ &\le \sup_{u \in H' \setminus \{0\}} \frac{\|K_T u\|_{L^2(\Sigma_0)} \|G_T v^0\|_{L^2(\Sigma_0)}}{\|u\|_{H'}} = \|K_T\| \gamma_T(v^0, v^0)^{1/2}\,,\end{aligned}$$

so (2.20) holds with $C_e = \|K_T\|^2$.

Observability inequality $\Rightarrow$ exact controllability. Since we know that γ_T is bounded, we have

$$C_e^{-1} \|v^0\|^2_{\widetilde{H}} \le \gamma_T(v^0, v^0) \le \|\gamma_T\| \|v^0\|^2_{\widetilde{H}} \quad \text{for all } v^0 \in \widetilde{H}\,, \tag{2.21}$$

so $\gamma_T(v^0, v^0)^{1/2}$ is a norm *equivalent* to that of $\widetilde{H}$. Let now $u \in H'$ be fixed, but arbitrary. From the Riesz Representation Theorem we deduce that there exists a unique $\hat{v}^0 \in \widetilde{H}$ that fulfills

$$\gamma_T(\hat{v}^0, v^0) = \{u, v^0\}\,, \tag{2.22}$$

for all $v^0 \in \widetilde{H}$ (since $v^0 \mapsto \{u, v^0\}$ is a linear and continuous functional). From Theorem 2.1.1 we see that the control $\hat{k} = G_T \hat{v}$ steers the solution from zero to u.

Observe next that

$$\begin{aligned}\gamma_T(\hat{v}^0, \hat{v}^0) \le \|u\|_{H'} \|\mathcal{M} \hat{v}^0\|_H \le C_e^{1/2} \|u\|_{H'} \gamma_T(\hat{v}^0, \hat{v}^0)^{1/2} \quad &\Rightarrow \\ \|\hat{k}\| = \gamma_T(\hat{v}^0, \hat{v}^0)^{1/2} \le C_e^{1/2} \|u\|_{H'}\,,&\end{aligned}$$

which shows the boundedness of the operator K_T. $\square$

The observability inequality (2.20) is commonly called the *inverse inequality*, as opposed to direct inequality (2.6).

Note that although the above result is well known, the implication *exact controllability $\Rightarrow$ observability inequality* is often not emphasized.

2.5 Hilbert Uniqueness Method

The norm equivalence (2.21) was originally the *central ingredient* of the Hilbert Uniqueness Method (HUM), a method developed by Professor Jacques-Louis Lions.

The method first saw the light in 1988, see *Lions (1988a)* or *Lions (1988b)*. Refer to *Lagnese (1991)* (which includes treatment of first order systems) and *Bensoussan (1993)* for some abstract views on HUM.

We will approach HUM from a slightly different angle, and simply insist that the control function *must* be of the form $k = \mathcal{C}w = G_T w^0$, where w is the solution to the adjoint system with w^0 as initial data. Such a control will be called a *HUM control*.

Equation (2.11) now gets the appearance

$$\begin{aligned} G_T^* G_T w^0 &= \mathcal{M}^T(u^1 - L_T u^0) \quad \Leftrightarrow \\ \Lambda_T w^0 &= f\,, \end{aligned} \tag{2.23}$$

where $\Lambda_T = G_T^* G_T : \widetilde{H} \mapsto \widetilde{H}'$ and $f = \mathcal{M}^T(u^1 - L_T u^0) \in \widetilde{H}'$. Note that the functional f is linear and continuous and depends on both u^0, u^1 and T.

In the variational formulation we get

$$\begin{aligned} \langle G_T w^0, G_T v^0\rangle_{L^2(\Sigma_0)} &= \{u^1, v^0\} - \{u^0, L_T^* v^0\} \quad \Leftrightarrow \\ \gamma_T(w^0, v^0) &= f(v^0)\,, \end{aligned} \tag{2.24}$$

for all $v^0 \in \widetilde{H}$, where the bilinear form $\gamma_T : \widetilde{H} \times \widetilde{H} \mapsto \mathbb{R}$ was introduced in (2.9).

Note that we in (2.23) introduced the very important map Λ_T which we shall call the *controllability operator*. Note also

$$\langle \Lambda_T w^0, v^0\rangle_{\widetilde{H}'\times\widetilde{H}} = \langle G_T^* G_T w^0, v^0\rangle_{\widetilde{H}'\times\widetilde{H}} = \langle G_T w^0, G_T v^0\rangle_{L^2(\Sigma_0)} = \gamma_T(w^0, v^0)\,,$$

for all $w^0, v^0 \in \widetilde{H}$, making Λ_T and γ_T equivalent in a Riesz isomorphism sort of way.

How restrictive is it that the control must be of the form $k = G_T w^0$? Not restrictive at all, as it turns out. When it comes to both null-controllability and exact controllability, a HUM control can *always* be found. This follows from the Theorems 2.3.1 and 2.4.1. In the proof of each theorem, when showing that the observability inequality implies controllability, a HUM control is actually constructed.

Any control obtained through the Hilbert Uniqueness Method has an important optimality property. This is the subject of the following theorem.

Theorem 2.5.1. *Let a HUM control $\hat{k} \in L^2(\Sigma_0)$ exist that steers the initial state $u^0 \in H'$ to the final state $u^1 \in H'$. Then among all controls that steer the control system from u^0 to u^1, the HUM control $\hat{k}$ has the minimal $L^2(\Sigma_0)$-norm.*

Proof. Let the HUM control be given by $\hat{k} = G_T \hat{v}^0$ and let $k \in L^2(\Sigma_0)$ be an arbitrary control solving the same exact controllability problem. These must then fulfill, in particular,

$$\begin{aligned} \langle G_T \hat{v}^0, G_T \hat{v}^0\rangle_{L^2(\Sigma_0)} &= \{u^1, \hat{v}^0\} - \{u^0, L_T^* \hat{v}^0\} \quad \text{and} \\ \langle k, G_T \hat{v}^0\rangle_{L^2(\Sigma_0)} &= \{u^1, \hat{v}^0\} - \{u^0, L_T^* \hat{v}^0\}\,. \end{aligned} \tag{2.25}$$

Combining the two equations we see that

$$\langle k, G_T\hat{v}^0\rangle_{L^2(\Sigma_0)} = \langle G_T\hat{v}^0, G_T\hat{v}^0\rangle_{L^2(\Sigma_0)} \,.$$

We now get

$$\|\hat{k}\|^2_{L^2(\Sigma_0)} = \langle G_T\hat{v}^0, G_T\hat{v}^0\rangle_{L^2(\Sigma_0)} = |\langle k, G_T\hat{v}^0\rangle_{L^2(\Sigma_0)}| \le \|k\|_{L^2(\Sigma_0)}\|\hat{k}\|_{L^2(\Sigma_0)} \,,$$

which immediately leads to $\|\hat{k}\|_{L^2(\Sigma_0)} \le \|k\|_{L^2(\Sigma_0)}$. □

2.5.1 Computing the Controls

Assume that we have exact controllability at time $T > 0$. Given states $u^0, u^1 \in \widetilde{H}'$ we can perform the following steps.

1. Compute $f = \mathcal{M}^T(u^1 - L_T u^0)$.
2. Solve $\Lambda_T w^0 = f$ for $w^0 \in \widetilde{H}$.
3. Set $k = G_T w^0$.

The computed control k now steers the system from u^0 to u^1. The difficult step is clearly inverting the controllability operator in step 2.

2.5.2 Exact Null-Controllability for Reversible Systems

Let us consider the particular case of the wave equation and assume that we wish to drive the state $(u^0, \overline{u}^0)$ at time $t = 0$ to the zero state $(0,0)$ at time $t = T$.

For the wave equation, the controllability operator can be defined in the following way. The adjoint system becomes

$$\begin{cases} v_{tt} = \Delta v & \text{in } Q \ , \\ \mathcal{B}v = 0 & \text{in } \Sigma \ , \\ v(T) = v^0, \quad v_t(T) = \overline{v}^0 & \text{in } \Omega \ , \end{cases} \tag{2.26}$$

where Δ is the Laplace operator, summing up second derivatives in each *space* direction. We then have the control system,

$$\begin{cases} u_{tt} = \Delta u & \text{in } Q \ , \\ \mathcal{B}u = \begin{cases} \mathcal{C}v \\ 0 \end{cases} & \begin{matrix} \text{in } \Sigma_0 \ , \\ \text{in } \Sigma \setminus \Sigma_0 \ , \end{matrix} \\ u(0) = 0, \quad u_t(0) = 0 & \text{in } \Omega \ , \end{cases} \tag{2.27}$$

which defines the controllability operator $\Lambda_T : H^1_0(\Omega) \times L^2(\Omega) \mapsto H^{-1}(\Omega) \times L^2(\Omega)$ as

$$\Lambda_T \begin{pmatrix} v^0 \\ v^1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^T \begin{pmatrix} u(T) \\ u_t(T) \end{pmatrix} = \begin{pmatrix} u_t(T) \\ -u(T) \end{pmatrix} \,.$$

(A derivation of the $\mathcal{M}$ matrix can be found in the beginning of Chapter 7).

As just described in the previous section, the control can be found by doing the following steps.

1. Compute $f = -\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^T L_T \begin{pmatrix} u^0 \\ \overline{u}^0 \end{pmatrix}$.
2. Solve $\Lambda_T \begin{pmatrix} w^0 \\ \overline{w}^0 \end{pmatrix} = f$ for $\begin{pmatrix} w^0 \\ \overline{w}^0 \end{pmatrix}$.
3. Solve (2.26) with $(v^0, \overline{v}^0) = (w^0, \overline{w}^0)$ and use $k(t) = \mathcal{C}v(t)$, $0 \le t \le T$.

We assume that the inversion in step 2 is possible.

The application of L_T in step 1, which involves solving the control system without applying control, can actually be avoided. Let us introduce a similar operator $\widetilde{\Lambda}_T$ which we shall call the *reversed controllability operator*. The adjoint system is as before, but with the time direction reversed,

$$\begin{cases} \tilde{v}_{tt} = \Delta \tilde{v} & \text{in } Q \ , \\ \mathcal{B}\tilde{v} = 0 & \text{in } \Sigma \ , \\ \tilde{v}(0) = \tilde{v}^0, \quad \tilde{v}_t(0) = \overline{\tilde{v}}^0 & \text{in } \Omega \ , \end{cases} \tag{2.28}$$

and reversing the time direction, and introducing a minus in the boundary condition, leads to the following control system,

$$\begin{cases} \tilde{u}_{tt} = \Delta \tilde{u} & \text{in } Q \ , \\ \mathcal{B}\tilde{u} = \begin{cases} -\mathcal{C}\tilde{v} \\ 0 \end{cases} & \begin{matrix} \text{in } \Sigma_0 \ , \\ \text{in } \Sigma \setminus \Sigma_0 \ , \end{matrix} \\ \tilde{u}(T) = 0, \quad \tilde{u}_t(T) = 0 & \text{in } \Omega \ , \end{cases} \tag{2.29}$$

thereby defining $\widetilde{\Lambda}_T : H_0^1(\Omega) \times L^2(\Omega) \mapsto H^{-1}(\Omega) \times L^2(\Omega)$ as

$$\widetilde{\Lambda}_T \begin{pmatrix} \tilde{v}^0 \\ \tilde{v}^1 \end{pmatrix} = \begin{pmatrix} \tilde{u}_t(0) \\ -\tilde{u}(0) \end{pmatrix} .$$

The initially stated control problem of driving $(u^0, \overline{u}^0)$ to $(0,0)$ can now be solved as

1. Solve $\widetilde{\Lambda}_T \begin{pmatrix} w^0 \\ \overline{w}^0 \end{pmatrix} = \begin{pmatrix} \overline{u}^0 \\ -u^0 \end{pmatrix}$ for $\begin{pmatrix} w^0 \\ \overline{w}^0 \end{pmatrix}$.
2. Solve (2.28) with $(\tilde{v}^0, \overline{\tilde{v}}^0) = (w^0, \overline{w}^0)$ and use $k(t) = \mathcal{C}\tilde{v}(t)$, $0 \le t \le T$.

Note that this control, by *construction*, has the wanted property of driving the solution of the control system to zero at time $t = T$. Note furthermore, that if one wishes to reach a non-zero state, then an application of L_T^{-1} is needed. This is possible, since we have a reversible system, but then one might as well use the original method.

The two operators Λ_T and $\widetilde{\Lambda}_T$ are closely related. This can be seen by first considering the adjoint systems (2.26) and (2.28). If $(\tilde{v}^0, \overline{\tilde{v}}^0) = (v^0, -\overline{v}^0)$ we see that $\tilde{v}(t) = v(T - t)$. This, in turn, implies that the solutions to the control systems (2.27) and (2.29) are related by $\tilde{u}(t) = -u(T-t)$. This finally relates the controllability operators in the way that

$$\widetilde{\Lambda}_T = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \Lambda_T \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} .$$

This relation shows that $\widetilde{\Lambda}_T$ is positive (semi-)definite if and only if Λ_T is positive (semi-)definite (obtaining this property is the reason for introducing the minus in the boundary condition in (2.29)).

Computing null-controllability using the reversed controllability operator is the method used by Jacques-Louis Lions in, e.g., *Lions (1988b)* and *Glowinski, Li, and Lions (1990)*, where the wave equation was studied. (Historical note: What was originally called Hilbert Uniqueness Method by Lions actually relied on the *reversed* controllability operator $\widetilde{\Lambda}_T$. Later Lions described RHUM, Reverse Hilbert Uniqueness Method, which made use of the "usual" controllability operator Λ_T (see, e.g., *Lagnese (1991)* or *Bensoussan (1993)* for a mention of RHUM). In today's literature, HUM is used for both types).

2.6 Controlling Projections

What if we only wish to control a *part* of the solution? Is this possible and what do the observability inequalities then look like? That is the subject of this section.

Controllability of projections has previously been considered in the literature, especially in the context of discrete systems, see *Zuazua (2003)* and many of the references therein. But controllability of projections has also been applied to continuous system, see *Lebeau and Zuazua (1998)*, where a linear system of thermoelasticity is considered. A similar system of thermoelasticity will be studied in Chapter 8, where some results of this section will be applied.

Despite their use in the literature, only a few formal results have appeared. When the projections are onto finite dimensional spaces, however, see *Zuazua (1997)* or *Micu and Zuazua (2004)*. We will present some fairly general results in the following.

Consider a Hilbert space H_0', which is a subspace of H', $H_0' \subset H'$. We also introduce the orthogonal projection $\Pi : H' \mapsto H'$, a linear and bounded operator for which

$$H_0' = \Pi(H') = (\ker \Pi)^{\perp} ,$$

and the associated "adjoint" operator $\widetilde{\Pi} : \widetilde{H} \mapsto \widetilde{H}$, also linear and bounded, where

$$\{\Pi U, V\} = \{U, \widetilde{\Pi} V\} ,$$

for all $U \in H'$ and $V \in \widetilde{H}$. We set $\widetilde{H}_0 = \widetilde{\Pi}\widetilde{H}$.

Assume first that the relation

$$\langle k, G_T v^0 \rangle_{L^2(\Sigma_0)} + \{u^0, L_T^* v^0\} - \{u^1, v^0\} = 0 ,$$

holds for all $v^0 \in \widetilde{H}_0$ (compare to (2.7) of Theorem 2.1.1). Let u be the solution of the control system with control k and $u(0) = u^0$. We then know that

$$\langle k, G_T v^0 \rangle_{L^2(\Sigma_0)} + \{u^0, L_T^* v^0\} - \{u(T), v^0\} = 0 \, ,$$

for all $v^0 \in \widetilde{H}$. Subtracting the two expressions above we get that

$$\{u(T) - u^1, \widetilde{\Pi} v^0\} = \{\Pi(u(T) - u^1), v^0\} = 0 \quad \text{for all } v^0 \in \widetilde{H} \, ,$$

implying that $\Pi u(T) = \Pi u^1$. The preceeding statements can be seen as a "projection version" of Theorem 2.1.1.

Let us now consider approximate control of projections, that is, for every $u^0, u^1 \in H'$ and $\epsilon > 0$ we seek a control $k \in L^2((0,T) \times \Gamma_0)$ such that

$$\|\Pi(\mathcal{L}(u^0, k)(T) - u^1)\|_{H'} < \epsilon \, . \tag{2.30}$$

We will repeat the arguments leading up to Theorem 2.2.1 concerning approximate controllability. Assume initially that

$$\gamma_T(v^0, v^0) = \langle G_T v^0, G_T v^0 \rangle_{L^2(\Sigma_0)} = 0 \quad \Rightarrow \quad v^0 = 0 \, ,$$

for all $v^0 \in \widetilde{H}_0$. This is equivalent to

$$\ker G_T \widetilde{\Pi} = \ker \widetilde{\Pi} \, .$$

Obtaining the adjoint of $G_T \widetilde{\Pi}$ is straightforward and we end up with

$$\begin{aligned}
\overline{\mathcal{M}^T \Pi \mathcal{M}^{-T} G_T^*(L^2(\Sigma_0))} = (\ker G_T \widetilde{\Pi})^\perp = (\ker \widetilde{\Pi})^\perp = \widetilde{H}_0' \quad &\Leftrightarrow \\
\overline{\Pi \mathcal{M}^{-T} G_T^*(L^2(\Sigma_0))} = H_0' \quad &\Leftrightarrow \\
\overline{\mathcal{M}^{-T} G_T^*(L^2(\Sigma_0))} \supset H_0' \quad &\Leftrightarrow \\
\overline{\mathcal{L}(0, L^2(\Sigma_0))(T)} \supset H_0' \, , &
\end{aligned}$$

which leads to (2.30). We have thus proved the following theorem.

Theorem 2.6.1. *Let $T > 0$ be fixed. For every choice of $u^0, u^1 \in H'$ and $\epsilon > 0$ there exists a control $k \in L^2(\Sigma_0)$ for which*

$$\left\|\Pi\big(\mathcal{L}(u^0, k)(T) - u^1\big)\right\|_{H'} < \epsilon \, ,$$

if and only if

$$\gamma_T(v^0, v^0) = 0 \quad \Rightarrow \quad v^0 = 0 \, ,$$

for all $v^0 \in \widetilde{H}_0$.

We now turn to null-controllability of projections. What we wish to do is drive a solution's projection onto H_0' to zero. Observe first that Theorems 2.2.2 and 2.2.3 are easily adapted to the (smaller) Hilbert space $\widetilde{H}_0$. With minor modifications, inequality (2.15) of Theorem 2.2.4 becomes

$$\|\Pi(u(T) - u^1)\|_{H'} \leq \epsilon \,.$$

This makes it possible to repeat the arguments of Theorem 2.3.1, if every occurrence of $\widetilde{H}$ is replaced by $\widetilde{H}_0$. We thus arrive at the following theorem.

Theorem 2.6.2. *Let $T > 0$ be fixed. A linear and bounded operator $K_T^{\Pi,n} : H' \mapsto L^2(\Sigma_0)$ exists for which*

$$\Pi\mathcal{L}(u^0, K_T^{\Pi,n}(u^0))(T) = 0, \quad \textit{for all } u^0 \in H' \,,$$

if and only if there is a constant $C_{\Pi,n} > 0$ such that

$$\|L_T^* v^0\|_{\widetilde{H}}^2 \leq C_{\Pi,n}\gamma_T(v^0, v^0), \quad \textit{for all } v^0 \in \widetilde{H}_0 \,. \tag{2.31}$$

Note that the control can be computed by minimizing the functional (2.18), just over the smaller Hilbert space $\widetilde{H}_0$. A control found this way is unique, and has again the optimality condition that its $L^2(\Sigma_0)$-norm is minimal among all controls that solve the same null-controllability of a projection (Theorem 2.5.1 is easily adapted to this case).

Finally, we consider exact controllability of projections. The use of the Riesz Representation Theorem in Theorem 2.4.1 can still be used on the Hilbert subspace $\widetilde{H}_0$, so we have that a $\hat{v}^0 \in \widetilde{H}_0$ exists for which

$$\gamma_T(\hat{v}^0, v^0) = \{u, v^0\}, \quad \text{for all } v^0 \in \widetilde{H}_0 \,, \tag{2.32}$$

with $u = u^1 - L_T u^0$. We now have the following equivalences:

$$\begin{aligned}
\gamma_T(\hat{v}^0, v^0) &= \{u, v^0\} & \forall v^0 \in \widetilde{H}_0 \quad &\Leftrightarrow \\
\{\mathcal{M}^{-T} G_T^* G_T \hat{v}^0, \widetilde{\Pi} v^0\} &= \{u, \widetilde{\Pi} v^0\} & \forall v^0 \in \widetilde{H} \quad &\Leftrightarrow \\
\Pi\mathcal{M}^{-T} G_T^* G_T \hat{v}^0 &= \Pi(u^1 - L_T u^0) & &\Leftrightarrow \\
\Pi\mathcal{L}(u^0, G_T \hat{v}^0)(T) &= \Pi u^1 \,, & &
\end{aligned}$$

where this last relation is exactly what we want. So now we have the following result.

Theorem 2.6.3. *Let $T > 0$ be fixed. A linear and bounded operator $K_T^{\Pi,e} : H' \mapsto L^2(\Sigma_0)$ exists for which*

$$\Pi\mathcal{L}(0, K_T^{\Pi,e}(u))(T) = \Pi u, \quad \textit{for all } u \in H' \,,$$

if and only if there is a constant $C_{\Pi,e} > 0$ such that

$$\|v^0\|_{\widetilde{H}}^2 \leq C_{\Pi,e}\gamma_T(v^0, v^0), \quad \textit{for all } v^0 \in \widetilde{H}_0 \,. \tag{2.33}$$

As was the case for null-controllability, the control obtained by solving (2.32) is unique and provides the control with smallest $L^2(\Sigma_0)$-norm.

Note how the observability inequalities of this section are identical to the cases were control of the *whole* solution was considered. The space over which it must hold is just correspondingly smaller.

2.7 Approximate Solutions

Since the main subject of this thesis is exact controllability, we will only address approximate solutions briefly.

2.7.1 Iterative Solutions and Optimization

As seen in Equation (2.24), a HUM control for exact controllability can be found by determining $w^0 \in \widetilde{H}$ such that

$$\gamma_T(w^0, v^0) = \{u^1 - L_T u^0, v^0\},$$

for all $v^0 \in \widetilde{H}$. This is a variational formulation where γ_T is a symmetric, bilinear form which is also positive definite (provided the conditions of Theorem 2.4.1 are met). An iterative method called the Conjugate Gradient algorithm (CG) is well fitted for such problems. It is an iterative algorithm which progressively finds a better and better approximate solution. The CG algorithm approach to solving control problems has been used extensively by Professor Glowinski and colleagues, see, e.g., *Glowinski, Kinton, and Wheeler (1989)*, *Glowinski and Li (1990)*, *Glowinski, Li, and Lions (1990)*, *Glowinski (1992a)*, *Glowinski (1992b)* and *Carthel, Glowinski, and Lions (1994)*, but also *Asch and Lebeau (1998)* and *Negreanu and Zuazua (2003)* have made use of it. Properties of the CG algorithm in a general Hilbert space setting can be found in *Daniel (1971)*. In a finite dimensional, numerical analysis setting, a large amount of literature can be found, see, e.g., *Golub and Van Loan (1996)* and the references therein. Since the CG algorithm only finds approximate solutions, we will not consider it further. (In the finite dimensional case, CG will solve the problem *exactly* when performing a number of iterations corresponding to the space dimension. In such a case one is better off using a direct method).

Another approach to solving an exact controllability problem is minimizing the functional

$$J(v^0) = \tfrac{1}{2}\gamma_T(v^0, v^0) - \{u^1 - L_T u^0, v^0\},$$

over $\widetilde{H}$. In the case of null-controllability, where $u^1 = 0$, Theorem 2.3.1 (and the proof thereof) provides the necessary and sufficient conditions for the existence of a unique minimizer. For exact controllability, the convexity and coercivity of J (which ensures a unique minimizer) is clear as soon as γ_T is positive in the sense of Theorem 2.4.1. We will not provide a survey of methods for finding the minimizer accurately and efficiently, since it is too far from the subject of this thesis.

An optimization approach was taken in *Park and Lee (2002)* for solving approximate controllability problems for the two dimensional heat equation. They used a CG-type algorithm to minimize J, where the gradient of J was computed using a so-called adjoint variable method. Also in *Eljendy (1992)* was an optimization technique used for solving exact controllability problems for the wave equation in two dimensions.

Another situation in which approximate solutions occur is when *regularization* is used. In practice it may be an unstable process, due to rounding errors, to solve the usual

$$\Lambda_T w^0 = \mathcal{M}^T(u^1 - L_T u^0)\,,$$

if, for instance, the inverse Λ_T^{-1} is unbounded. Instead, one may prefer to solve

$$\Lambda_T^\alpha w^0 = \mathcal{M}^T(u^1 - L_T u^0)\,,$$

where $\Lambda_T^\alpha \to \Lambda_T$ in some sense as $\alpha \to 0$, but where Λ_T^α is considerably more robust to invert (in relation to, e.g., rounding errors). The quantity α is typically called a regularization parameter. The downside to this regularization approach is, of course, that only an approximate solution will be obtained. An example of regularization in conjunction with the wave equation is using

$$\Lambda_T^\alpha = \Lambda_T + \alpha \begin{bmatrix} -\Delta & 0 \\ 0 & I \end{bmatrix}\,,$$

which shifts the spectrum of Λ_T. This type of regularization method was used in *Glowinski and Li (1990)*, *Glowinski, Li, and Lions (1990)* and *Glowinski (1992b)*, together with the CG algorithm for the two-dimensional wave equation. In *Carthel, Glowinski, and Lions (1994)* the authors used so-called Tikhonov regularization to solve approximate controllability problems for the two-dimensional heat equation. See also *Kindermann (1999)* for a similar approach.

2.7.2 By How Much Did We Miss?

Assume now that we have obtained an approximate solution in the sense that

$$\Lambda_T w^0 = \mathcal{M}^T(u^1 - L_T u^0) + r\,,$$

where $r \in \widetilde{H}'$ represents a non-zero residual. Now what happens if we use the control associated to w^0, even though it is not the exact solution? As far as the author knows, this question has not been treated in such a manner before. We answer the question in two different ways.

Case 1: Change in final state: We immediately obtain

$$\Lambda_T w^0 = \mathcal{M}^T((u^1 + \mathcal{M}^{-T} r) - L_T u^0)\,,$$

from which we can conclude that we miss the target exactly by $\mathcal{M}^{-T} r$. From the norm equality $\|\mathcal{M}^{-T} r\|_{H'} = \|r\|_{\widetilde{H}'}$ we see that the norm of the residual r shows exactly by how much we miss the target. This is a somewhat unusual situation, seen in relation to inverse problems in general, where the residual and not the error of the solution itself (here w^0) determines the quality of the solution.

Case 2: Change in initial state: Similarly we get

$$\Lambda_T w^0 = \mathcal{M}^T(u^1 - L_T(u^0 - L_T^{-1}\mathcal{M}^{-T}r))\,.$$

So if we start out with the state $u^0 - L_T^{-1}\mathcal{M}^{-T}r$, we reach the target u^1 exactly. Note, however, that this second case only makes sense for reversible systems, since L_T^{-1} must exist. The case is thus relevant when computing exact controllability for reversible systems, see Section 2.5.2.

2.8 Summary

We were what-what in a what-what?
— Homer Simpson (HOMR, season 12)

Let us collect the most important threads of this chapter. The goal was to steer a solution of the control system (2.1) to a state at time $t = T$ which could be any state (exact controllability), the null state (null-controllability) or sufficiently close to any state (approximate controllability).

The control was exerted on the control boundary $\Gamma_0 \subset \partial\Omega$ through the boundary operator $\mathcal{B}$, leading to Dirichlet or Neumann control, or something else. From this control system, an adjoint system was devised. The boundary conditions of the adjoint system were homogeneous and an associated complementary boundary operator $\mathcal{C}$ had to be determined. Through the operator $\mathcal{C}$ we could define the bilinear form γ_T, where the quantity $\gamma_T(v, v)$ determined, loosely speaking, what could be observed from the boundary of a solution to the adjoint system with initial condition v.

We can now summarize the different types of control as they relate to what is observed through γ_T. The observability inequalities are as follows (each statement must hold for all $v \in \widetilde{H}$):

$$\begin{array}{ccccc}
\|v\|^2_{\widetilde{H}} & \leq & C_e\gamma_T(v,v) & \Leftrightarrow & \text{Exact controllability} \\
 & \Downarrow & & & \\
\|L_T^* v\|^2_{\widetilde{H}} & \leq & C_n\gamma_T(v,v) & \Leftrightarrow & \text{Null controllability} \\
 & \Downarrow & & & \\
v \neq 0 & \Rightarrow & \gamma_T(v,v) > 0 & \Leftrightarrow & \text{Approximate controllability}
\end{array}$$

The vertical implications follow easily, and show how the three types of control are related.

Chapter 3

Discretizations

It is hard to be finite upon an infinite subject,
and all subjects are infinite.
— Herman Melville

To discretize is a way of approximating something infinite dimensional by something finite dimensional. It is typically used when analytical methods become impossible, too hard, or just too time-consuming. Furthermore, representing a solution by a finite number of data makes it possible to visualize the solution on a computer. This can often lead to increased understanding on many different levels, and may even lead to improvements in the analytical methods.

Needless to say, discretizations should lead to good approximations. But most importantly, it must be possible to choose the discretization parameters in such a way that the approximate solution lies as close to the real solution *as one would like*. In other words, it must be *convergent*.

This chapter mostly sets the stage for the following chapters. Apart from a few novel approaches, the material will be well known.

3.1 Discretization in Space

Our focus will here be on the Poisson problem,

$$\begin{aligned} \Delta u &= f, \qquad \text{in } \Omega, \\ u &= 0, \qquad \text{on } \Gamma\,. \end{aligned} \tag{3.1}$$

We will consider discretizations that can be formulated as

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{C}\boldsymbol{f}\,, \tag{3.2}$$

where $\boldsymbol{u}$ and $\boldsymbol{f}$ are vectors of equal length, representing the continuous functions u and f in some way, typically as point-wise samplings or as coefficients with respect

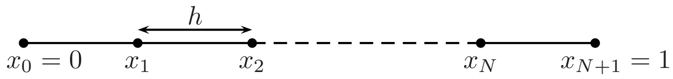


Figure 3.1: An equally spaced grid in the interval $[0, 1]$.

to some basis. The reason for using two matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ is that we wish to use $\boldsymbol{C}$ to approximate the $L^2(\Omega)$-norm in the following sense: If $\boldsymbol{u}$ approximates u then

$$\langle \boldsymbol{u}, \boldsymbol{u} \rangle_{\boldsymbol{C}} = \boldsymbol{u}^T \boldsymbol{C} \boldsymbol{u} \quad \simeq \quad \int_\Omega |u|^2 dx = \langle u, u \rangle_{L^2(\Omega)} \,. \tag{3.3}$$

The formulation (3.2) also makes sense from a Finite Element Method (FEM) point of view. Let B_1, B_2, $\ldots$, B_k be (global) basis functions that are zero on the boundary, $B_i|_\Gamma = 0$, used for a first order FEM discretization,

$$u(x) \simeq \sum_{i=1}^k \boldsymbol{u}_i B_i(x)\,, \qquad f(x) \simeq \sum_{i=1}^k \boldsymbol{f}_i B_i(x)\,.$$

Inserting these expressions into the Poisson problem (3.1) and using also $\langle B_i \rangle_{i=1}^k$ as test functions, we get

$$\boldsymbol{C}(i,j) = \int_\Omega B_i(x) B_j(x) dx, \qquad \boldsymbol{A}(i,j) = -\int_\Omega \nabla B_i(x) \nabla B_j(x) dx\,.$$

Note how relation (3.3) is fulfilled. Traditionally the matrices $\boldsymbol{C}$ and $\boldsymbol{A}$ are called the mass- and stiffness matrix, respectively.

The discussion so far has not been restricted to any particular dimension or type of domain. Let us now consider the simple one dimensional case of $\Omega = (0, 1)$. We use a uniform grid with grid size $h = 1/(N+1)$ and node points $x_j = jh$, $j = 0, 1, \ldots, N+1$, see Figure 3.1.

We introduce the family of discretizations

$$\frac{\boldsymbol{u}_{j+1} - 2\boldsymbol{u}_j + \boldsymbol{u}_{j-1}}{h^2} = \alpha \boldsymbol{f}_{j+1} + (1 - 2\alpha) \boldsymbol{f}_j + \alpha \boldsymbol{f}_{j-1}\,, \tag{3.4}$$

for $j = 1, 2, \ldots, N$ and some real parameter α. For twice continuously differentiable u and continuous f we see, using Taylor series, that this way of discretizing is consistent with (3.1). We shall later see that this scheme also makes sense for less smooth u and f. To bring the scheme into the formulation of (3.2) we set

$$\boldsymbol{A} = \frac{1}{h} \begin{bmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & -2 \end{bmatrix}, \quad \boldsymbol{C}_\alpha = h \begin{bmatrix} 1-2\alpha & \alpha & & \\ \alpha & 1-2\alpha & \alpha & \\ & \ddots & \ddots & \ddots \\ & & \alpha & 1-2\alpha \end{bmatrix}. \tag{3.5}$$

Note how a single power of h has been moved to the other side of the equation, in order for $\boldsymbol{C}_\alpha$ to have the property of (3.3). Note also how the homogeneous Dirichlet boundary conditions have been incorporated implicitly.

Some important values of α should be emphasized. With $\alpha = 0$ we have the well-known finite difference scheme. The case $\alpha = 1/6$ arises if one uses FEM with a hat function basis (such a basis consists of functions B_0, B_1, ..., B_{N+1} that are continuous, linear on each interval (x_i, x_{i+1}) and for which $B_i(x_j) = \delta_{ij}$, that is, 1 for $i = j$ and 0 otherwise). For $\alpha = 1/4$ we call the scheme the *box method* and it turns out to have some very interesting properties, as we shall see later. (The name box method was used in *Vichnevetsky and Bowles (1982)* for the trapezoid rule, see Section 3.2.3 later in this chapter. But (3.4) with $\alpha = 1/4$ is a special case of the trapezoid rule for second order systems, and we will reserve the name box method for this case).

We will call the discretization scheme introduced above for α-discretization. As just mentioned, choosing different values of α covers several well-known ways of discretizing the Poisson operator in one dimension. This possibility of treating several cases at once was first introduced by the author in *Rasmussen (2003)*.

So $\boldsymbol{C}_\alpha^{-1}\boldsymbol{A}$ approximates the Laplacian Δ. Knowledge about its eigenvalues and eigenvectors is essential when it comes to analyzing solutions of evolution equations involving the Laplacian and, in turn, when analyzing control properties of such systems. Luckily both $\boldsymbol{C}_\alpha$ and $\boldsymbol{A}$ of (3.5) are tridiagonal, symmetric and Toeplitz (the diagonal and each off-diagonal of a Toeplitz matrix contain constant entries). This is fortunate since we know the eigenvalues and eigenvectors of such matrices explicitly.

To see this, we start out with the following special case:

$$\boldsymbol{L} = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{N\times N} .$$

The eigensolutions of this matrix,

$$\boldsymbol{L}\boldsymbol{w}_k = \eta_k \boldsymbol{w}_k, \quad k = 1, 2, \ldots, N ,$$

are explicitly known,

$$\boldsymbol{w}_k(j) = \sin(jk\pi h), \qquad \eta_k = 4\sin^2(\tfrac{1}{2}k\pi h) ,$$

which is easily verified by insertion (recall that $h = 1/(N+1)$).

We can now consider general symmetric and tridiagonal Toeplitz matrices. Let therefore

$$\boldsymbol{T}_{\alpha,\beta} = \begin{bmatrix} \beta & \alpha & & \\ \alpha & \beta & \alpha & \\ & \ddots & \ddots & \ddots \\ & & \alpha & \beta \end{bmatrix} \in \mathbb{R}^{N\times N}, \quad \alpha, \beta \in \mathbb{R},\ (\alpha, \beta) \neq (0, 0) ,$$

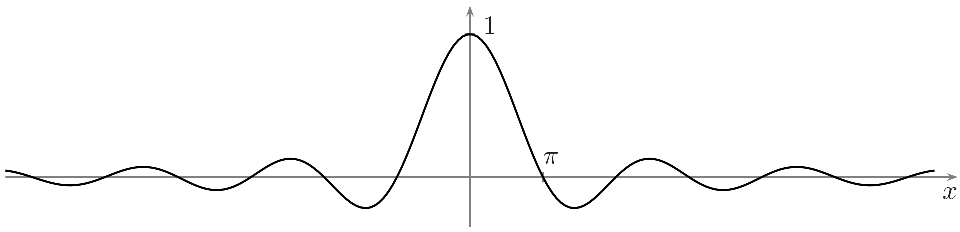


Figure 3.2: A plot of the function $\sin(x)/x$. Important properties of this function are $\sin(x)/x \to 1$ as $x \to 0$, $|\sin(x)/x| \leq 1$ for all $x \in \mathbb{R}$, and that it is decreasing on the interval $0 \leq x \leq \pi$.

and observe that $\boldsymbol{T}_{\alpha,\beta} = (2\alpha + \beta)\boldsymbol{I} - \alpha\boldsymbol{L}$. This shows that the eigenvectors of $\boldsymbol{L}$ and $\boldsymbol{T}_{\alpha,\beta}$ are identical and that the eigenvalues of $\boldsymbol{T}_{\alpha,\beta}$ are $(2\alpha + \beta - \alpha\eta_k)$, $k = 1, 2, \ldots, N$.

We can now compute the eigenvalues λ_k^α of $\boldsymbol{C}_\alpha^{-1}\boldsymbol{A}$,

$$\boldsymbol{C}_\alpha^{-1}\boldsymbol{A}\boldsymbol{w}_k = \lambda_k^\alpha \boldsymbol{w}_k \quad \Leftrightarrow \quad \boldsymbol{A}\boldsymbol{w}_k = \lambda_k^\alpha \boldsymbol{C}_\alpha \boldsymbol{w}_k\,.$$

Since the eigenvectors of $\boldsymbol{T}_{\alpha,\beta}$ are independent of α and β we easily get

$$\lambda_k^\alpha = -\frac{4\sin^2(\frac{1}{2}k\pi h)}{h^2(1 - 4\alpha\sin^2(\frac{1}{2}k\pi h))} = -k^2\pi^2\left(\frac{\sin(\frac{1}{2}k\pi h)}{\frac{1}{2}k\pi h}\right)^2 \frac{1}{1 - 4\alpha\sin^2(\frac{1}{2}k\pi h)}\,. \tag{3.6}$$

The second expression for λ_k^α clearly shows how close they are to the true eigenvalues of the Laplacian (which are $-k^2\pi^2$, $k = 1, 2, \ldots$) whenever $k \ll N$. See Figure 3.3 for an illustration of the eigenvalues when $N = 30$ and for different choices of α.

The expressions above also show that not every α is feasible, since the denominator can become zero if $4\alpha\sin^2(\frac{1}{2}k\pi h)) = 1$. Clearly this can not happen if $\alpha \leq 1/4$. The special border case $\alpha = 1/4$ leads to an interesting expression for the eigenvalues,

$$\lambda_k^{1/4} = -\frac{4}{h^2}\tan^2(\tfrac{1}{2}k\pi h) = -k^2\pi^2\left(\frac{\tan(\frac{1}{2}k\pi h)}{\frac{1}{2}k\pi h}\right)^2\,.$$

Note also that with $\alpha > 1/4$, some eigenvalues will inevitably become *positive* as soon as N get large enough. This would destroy the elliptic nature of $\boldsymbol{C}_\alpha^{-1}\boldsymbol{A}$ and we will henceforth only consider the interval $0 \leq \alpha \leq 1/4$.

The matrices $\boldsymbol{C}_\alpha$ and $\boldsymbol{A}$ that we have just considered were positive and negative definite, respectively. Furthermore, the eigenvectors of $\boldsymbol{C}_\alpha$, $\boldsymbol{A}$ and $\boldsymbol{C}_\alpha^{-1}\boldsymbol{A}$ were identical. This is not true in general, but some useful, general properties are the subject of the following theorem. We here use the discrete inner product $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^T\boldsymbol{v}$.

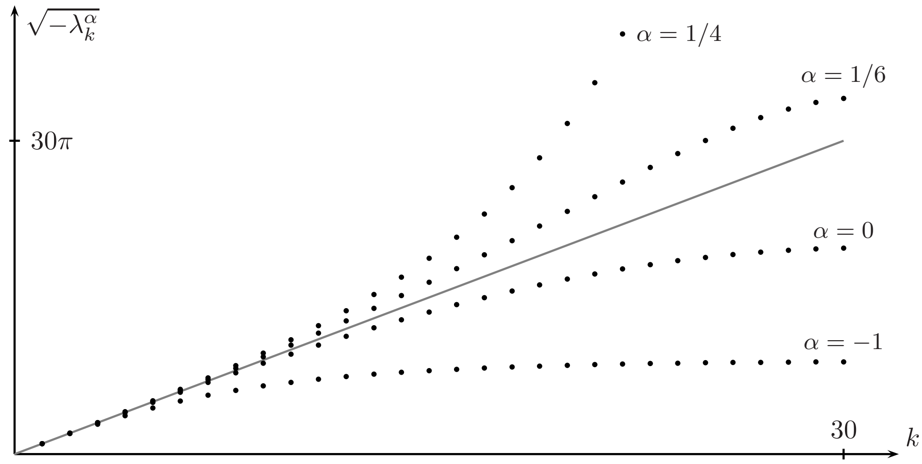


Figure 3.3: Illustration of eigenvalues of $\boldsymbol{C}_\alpha^{-1}\boldsymbol{A}$ for the specific case of $N = 30$. The solid line indicates where the eigenvalues of the continuous operator, Δ, would be.

Theorem 3.1.1. *Let $\boldsymbol{C} \in \mathbb{R}^{N\times N}$ be symmetric and positive definite, let $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ be symmetric and negative definite, and consider the eigenvalue problem,*

$$\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{w}_k = \lambda_k \boldsymbol{w}_k \,.$$

The following then holds:

1. *There exists a full set of eigenvectors $\boldsymbol{w}_k$, $k = 1, 2, \ldots, N$ that spans $\mathbb{R}^N$.*
2. *The matrix $\boldsymbol{C}^{-1}\boldsymbol{A}$ is negative definite, i.e., $\lambda_k < 0$ for $k = 1, 2, \ldots, N$.*
3. *$\langle \boldsymbol{w}_k, \boldsymbol{C}\boldsymbol{w}_l\rangle = 0$ for $\lambda_k \neq \lambda_l$.*
4. *$\langle \boldsymbol{w}_k, \boldsymbol{A}\boldsymbol{w}_l\rangle = 0$ for $\lambda_k \neq \lambda_l$.*

Proof. Consider the following rewrite,

$$\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{w} = \lambda\boldsymbol{w} \quad \Leftrightarrow \quad (\boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{C}^{-\frac{1}{2}})\boldsymbol{v} = \lambda\boldsymbol{v} \,,$$

where $\boldsymbol{w} = \boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{v}$. Since $\boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{C}^{-\frac{1}{2}}$ is symmetric there exists a set of eigenvectors $\{\boldsymbol{v}_k\}_{k=1}^N$ that span $\mathbb{R}^N$. Since $\boldsymbol{w} = \boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{v}_k$ and $\boldsymbol{C}^{-\frac{1}{2}}$ is clearly regular, statement 1 follows.

The eigenvalues of $\boldsymbol{C}^{-1}\boldsymbol{A}$ are equivalent to those of $\boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{C}^{-\frac{1}{2}}$, which is negative definite. This proves statement 2.

Assume now that $\boldsymbol{v}_k$, $\boldsymbol{v}_l$ are eigenvectors of the symmetric matrix $\boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{C}^{-\frac{1}{2}}$, corresponding to different eigenvalues. They are then orthogonal with respect to the inner product $\langle \cdot, \cdot \rangle$, i.e.,

$$0 = \langle \boldsymbol{v}_k, \boldsymbol{v}_l\rangle = \langle \boldsymbol{C}^{\frac{1}{2}}\boldsymbol{w}_k, \boldsymbol{C}^{\frac{1}{2}}\boldsymbol{w}_l\rangle = \langle \boldsymbol{w}_k, \boldsymbol{C}\boldsymbol{w}_l\rangle = \frac{1}{\lambda_l}\langle \boldsymbol{w}_k, \boldsymbol{A}\boldsymbol{w}_l\rangle \,,$$

for all $l, k = 1, 2, \ldots, N$ for which $\lambda_l \neq \lambda_k$. This immediately leads to statements 3 and 4. $\square$

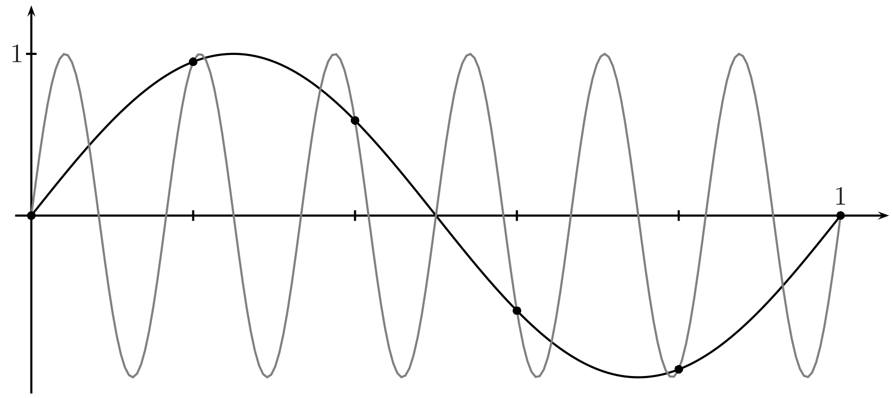


Figure 3.4: Illustration of grid aliasing on an equally spaced grid. Here, the waves $\sin(2x\pi)$ and $\sin(12x\pi)$ cannot be distinguished when sampled on the grid $x = i/5$, $i = 0, 1, \dots, 5$.

3.1.1 Waves in $h\mathbb{Z}$

Consider the illustration in Figure 3.4. Here are shown two clearly distinct sine waves, $\sin(2x\pi)$ and $\sin(12x\pi)$. However, when sampling these continuous functions onto the regular grid $\{0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\}$, they are *indistinguishable*. This common phenomenon is called grid aliasing, and the following theorem describes exactly which waves are "the same" on a regular grid. See *Trefethen (2000)* for some information on grid aliasing.

Theorem 3.1.2. *For $x, y \in \mathbb{R} \setminus \pi\mathbb{Z}$ we have*

$$\sin(xj) = \sin(yj), \qquad \text{for all } j \in \mathbb{Z}\,, \tag{3.7}$$

if and only if 2π divides $x - y$,

$$2\pi \mid x - y\,. \tag{3.8}$$

Proof. First we show that (3.7) implies (3.8). Assume therefore that $\sin(xj) = \sin(yj)$ for all $j \in \mathbb{Z}$. In particular, this means that we must have $\sin(x) = \sin(y)$, implying that either $2\pi \mid x - y$ or $2\pi \mid x + y - \pi$. In the first case, we are done. Assume therefore the second case, that $x + y = (2p + 1)\pi$ for some $p \in \mathbb{Z}$. This implies $2x = -2y + 2(2p + 1)\pi$ so $\sin(2x) = -\sin(2y)$, which is a contradiction.

Assume now that $x = y + 2p\pi$ for some $p \in \mathbb{Z}$. Now for $j \in \mathbb{Z}$ we have $\sin(xj) = \sin(yj + 2jp\pi) = \cos(yj)\sin(2jp\pi) + \sin(yj)\cos(2jp\pi) = \sin(yj)$. □

The following theorem investigates the discrete analog of

$$\int_0^1 \sin(k\pi x)\sin(l\pi x)dx = \tfrac{1}{2}\delta_{kl}\,,$$

for $k, l \in \mathbb{N}$. Note how grid aliasing plays a role in the periodic nature of the result. (The symbol $\nmid$ means "does not divide".)

Theorem 3.1.3. *For every choice of $N \in \mathbb{N}$ and $k, l \in \mathbb{Z}$, the following holds:*

$$\sum_{j=1}^{N} \sin(kj\pi/(N+1)) \sin(lj\pi/(N+1))$$
$$= \begin{cases} -\frac{N+1}{2}\,, & 2(N+1) \mid k+l, \quad 2(N+1) \nmid k-l\,, \\ \frac{N+1}{2}\,, & 2(N+1) \mid k-l, \quad 2(N+1) \nmid k+l\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{3.9}$$

Proof. See Detail 3, page 182. □

We follow up on this result by introducing the matrix $\boldsymbol{W} \in \mathbb{R}^N$ with the entries:

$$\boldsymbol{W}(j,k) = \sin(jkh\pi), \quad \text{for } j,k = 1,2,\ldots,N\,.$$

Note how $\boldsymbol{W}$ clearly is symmetric. An implication of Theorem 3.1.3 is now that

$$\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{W}^2 = \tfrac{1}{2}(N+1)\boldsymbol{I}, \quad \text{implying} \quad \boldsymbol{W}^{-1} = \frac{2}{N+1}\boldsymbol{W}\,. \tag{3.10}$$

We will make use of this result later on.

3.1.2 Semi-Discretizations

Consider a linear system of first order equations,

$$\begin{cases} \boldsymbol{\mathcal{C}}\dot{\boldsymbol{u}}(t) = \boldsymbol{\mathcal{A}}\boldsymbol{u}(t)\,, \\ \boldsymbol{u}(0) = \boldsymbol{u}^0\,, \end{cases} \tag{3.11}$$

where $\boldsymbol{u} : \mathbb{R} \mapsto \mathbb{R}^N$ and $\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{A}} \in \mathbb{R}^{N\times N}$. We assume that $\boldsymbol{\mathcal{C}}^{-1}\boldsymbol{\mathcal{A}}$ exists and is diagonalizable such that

$$\boldsymbol{\mathcal{C}}^{-1}\boldsymbol{\mathcal{A}}\boldsymbol{z}_k = \sigma_k\boldsymbol{z}_k \quad \Leftrightarrow \quad \boldsymbol{\mathcal{A}}\boldsymbol{z}_k = \sigma_k\boldsymbol{\mathcal{C}}\boldsymbol{z}_k, \quad k = 1,2,\ldots,N\,, \tag{3.12}$$

where the eigenvectors $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N$ are linearly independent. If we are now given the initial data in this eigenvector basis,

$$\boldsymbol{u}^0 = \sum_{k=1}^{N} c_k\boldsymbol{z}_k\,,$$

the solution to (3.11) can be written as

$$\boldsymbol{u}(t) = \sum_{k=1}^{N} c_k e^{\sigma_k t}\boldsymbol{z}_k\,.$$

This is the well-known solution formula for first order ODEs.

Consider now a linear ODE of *second* order,

$$\begin{cases} \boldsymbol{C}\ddot{\boldsymbol{u}}(t) = \boldsymbol{A}\boldsymbol{u}(t) \, , \\ \boldsymbol{u}(0) = \boldsymbol{u}^0 , \quad \dot{\boldsymbol{u}}(0) = \overline{\boldsymbol{u}}^0 \, . \end{cases} \tag{3.13}$$

We assume that $\boldsymbol{C}$ and $\boldsymbol{A}$ are symmetric, and positive and negative definite, respectively. From Theorem 3.1.1 it now follows that $\boldsymbol{C}^{-1}\boldsymbol{A}$ is negative definite and with

$$\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{w}_k = \lambda_k \boldsymbol{w}_k \quad \Leftrightarrow \quad \boldsymbol{A}\boldsymbol{w}_k = \lambda_k \boldsymbol{C}\boldsymbol{w}_k , \quad k = 1, 2, \ldots, N \, ,$$

we set $\mu_k^2 = -\lambda_k$ with $\mu_k > 0$. By introducing $\overline{\boldsymbol{u}} = \dot{\boldsymbol{u}}$ we get a first order ODE,

$$\begin{bmatrix} \boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{u}} \\ \dot{\overline{\boldsymbol{u}}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{C} \\ \boldsymbol{A} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \overline{\boldsymbol{u}} \end{bmatrix} . \tag{3.14}$$

We are now interested in the eigensolutions of the matrix governing this system,

$$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{C}^{-1}\boldsymbol{A} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_k \\ \beta_k \boldsymbol{w}_k \end{bmatrix} = \sigma_k \begin{bmatrix} \boldsymbol{w}_k \\ \beta_k \boldsymbol{w}_k \end{bmatrix} .$$

From $\beta_k \boldsymbol{w}_k = \sigma_k \boldsymbol{w}_k$ we immediately see that $\beta_k = \sigma_k$. The relation $\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{w}_k = \sigma_k \beta_k \boldsymbol{w}_k$ now yields $\sigma_k = \pm i\sqrt{-\lambda_k} = \pm i\mu_k$. So the eigensolutions are

$$(\sigma_k, \boldsymbol{z}_k) = \left(i\mu_k , \begin{bmatrix} \boldsymbol{w}_k \\ i\mu_k \boldsymbol{w}_k \end{bmatrix} \right) , \quad (\sigma_{-k}, \boldsymbol{z}_{-k}) = \left(-i\mu_k , \begin{bmatrix} \boldsymbol{w}_k \\ -i\mu_k \boldsymbol{w}_k \end{bmatrix} \right) , \tag{3.15}$$

for $k = 1, 2, \ldots, N$. Let us write this more compactly for later convenience. We set

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 & \cdots & \boldsymbol{w}_N \end{bmatrix} , \qquad \boldsymbol{D} = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_N) \, , \tag{3.16}$$

where $\operatorname{diag}(\cdots)$ is a diagonal matrix with the listed values along the diagonal, and

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1 & \cdots & \boldsymbol{z}_N & \boldsymbol{z}_{-1} & \cdots & \boldsymbol{z}_{-N} \end{bmatrix} = \begin{bmatrix} \boldsymbol{W} & \boldsymbol{W} \\ i\boldsymbol{W}\boldsymbol{D} & -i\boldsymbol{W}\boldsymbol{D} \end{bmatrix} .$$

The inverse of $\boldsymbol{Z}$ is easily seen to be

$$\boldsymbol{Z}^{-1} = \tfrac{1}{2} \begin{bmatrix} \boldsymbol{W}^{-1} & -i\boldsymbol{D}^{-1}\boldsymbol{W}^{-1} \\ \boldsymbol{W}^{-1} & i\boldsymbol{D}^{-1}\boldsymbol{W}^{-1} \end{bmatrix} .$$

We finally have the diagonalization

$$\boldsymbol{Z}^{-1} \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{C}^{-1}\boldsymbol{A} & \boldsymbol{0} \end{bmatrix} \boldsymbol{Z} = \begin{bmatrix} i\boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & -i\boldsymbol{D} \end{bmatrix} .$$

Let us return to writing the solution of the system (3.13) in terms of eigenvectors of $\boldsymbol{C}^{-1}\boldsymbol{A}$. Let the initial conditions of (3.13) be given as

$$\boldsymbol{u}^0 = \sum_{k=1}^{N} a_k \boldsymbol{w}_k , \qquad \overline{\boldsymbol{u}}^0 = \sum_{k=1}^{N} b_k \boldsymbol{w}_k \, , \tag{3.17}$$

where the coefficients are real, $a_k, b_k \in \mathbb{R}$ $(1 \le k \le N)$. We wish to express these initial conditions in the eigenvector basis $\{\boldsymbol{z}_k\}_{1\le|k|\le N}$, that is, find the coefficients $\{c_k\}_{1\le|k|\le N}$ such that

$$\sum_{1\le|k|\le N} c_k \boldsymbol{z}_k = \begin{pmatrix} \boldsymbol{u}^0 \\ \overline{\boldsymbol{u}}^0 \end{pmatrix} .$$

It is easily verified that

$$c_k = \tfrac{1}{2}(a_k - ib_k/\mu_k), \qquad c_{-k} = \tfrac{1}{2}(a_k + ib_k/\mu_k)\,,$$

$k = 1, 2, \ldots, N$, is the unique solution to this problem. This means that the full solution to (3.13) is

$$\begin{bmatrix} \boldsymbol{u}(t) \\ \overline{\boldsymbol{u}}(t) \end{bmatrix} = \sum_{1\le|k|\le N} c_k e^{\sigma_k t} \boldsymbol{z}_k\,,$$

or, written out,

$$\begin{aligned}
\boldsymbol{u}(t) &= \sum_{k=1}^{N} \left[\tfrac{1}{2}(a_k - ib_k/\mu_k)e^{\mu_k t} + \tfrac{1}{2}(a_k + ib_k/\mu_k)e^{-\mu_k t}\right] \boldsymbol{w}_k \\
&= \sum_{k=1}^{N} \left[a_k \cos(\mu_k t) + b_k/\mu_k \sin(\mu_k t)\right] \boldsymbol{w}_k\,, \\
\dot{\boldsymbol{u}}(t) = \overline{\boldsymbol{u}}(t) &= \sum_{k=1}^{N} \left[\tfrac{1}{2}(a_k - ib_k/\mu_k)e^{\mu_k t} - \tfrac{1}{2}(a_k + ib_k/\mu_k)e^{-\mu_k t}\right] i\mu_k \boldsymbol{w}_k \\
&= \sum_{k=1}^{N} \left[-a_k \mu_k \sin(\mu_k t) + b_k \cos(\mu_k t)\right] \boldsymbol{w}_k\,.
\end{aligned} \tag{3.18}$$

Let us finally consider the so-called *energy* of a second order system (3.13),

$$E_h(t) = \tfrac{1}{2}\Big(\langle \dot{\boldsymbol{u}}(t), \boldsymbol{C}\dot{\boldsymbol{u}}(t)\rangle - \langle \boldsymbol{u}(t), \boldsymbol{A}\boldsymbol{u}(t)\rangle\Big)\,. \tag{3.19}$$

It is easily shown that $E_h'(t) = 0$, that is, the energy remains constant, $E_h(t) = E_h(0)$ for all t. Note that since $\boldsymbol{C}$ and $\boldsymbol{A}$ were assumed positive and negative definite, respectively, $E_h(0)$ defines a *norm* on the initial data, $(\boldsymbol{u}^0, \overline{\boldsymbol{u}}^0)$,

$$\left\| \begin{bmatrix} \boldsymbol{u}^0 \\ \overline{\boldsymbol{u}}^0 \end{bmatrix} \right\|_{\widetilde{\boldsymbol{Q}}}^2 = \begin{bmatrix} \boldsymbol{u}^0 \\ \overline{\boldsymbol{u}}^0 \end{bmatrix}^T \widetilde{\boldsymbol{Q}} \begin{bmatrix} \boldsymbol{u}^0 \\ \overline{\boldsymbol{u}}^0 \end{bmatrix}, \qquad \text{where} \quad \widetilde{\boldsymbol{Q}} = \begin{bmatrix} -\boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix}.$$

The above expression for the energy is in fact the primary reason for introducing the extra matrix $\boldsymbol{C}$ into second order systems. Indeed, consider the wave equation $u_{tt} = \Delta u$ in Ω with homogeneous Dirichlet boundary conditions. If now

$$\langle \dot{\boldsymbol{u}}, \boldsymbol{C}\dot{\boldsymbol{u}}\rangle \simeq \int_\Omega |\dot{u}|^2 dx, \quad \text{and} \quad \langle \boldsymbol{u}, \boldsymbol{A}\boldsymbol{u}\rangle \simeq -\int_\Omega |\nabla u|^2 dx\,,$$

then

$$E_h(t) \simeq E(t) = \tfrac{1}{2} \int_\Omega \Big(|\dot{u}|^2 + |\nabla u|^2 \Big) dx \,,$$

where $E(t) = E(0)$ is a natural energy for the continuous wave equation in the sense that

$$E(t) < \infty \quad \Leftrightarrow \quad (u(t), \dot{u}(t)) \in H_0^1(\Omega) \times L^2(\Omega) \,,$$

and the wave equation is well-posed in $H_0^1(\Omega) \times L^2(\Omega)$ (will be shown in Chapter 7).

In the next chapter we will consider controllability of discrete systems. It will here be essential that we use discrete norms that correspond to the continuous ones. To that end we use the convention that the discrete $\widetilde{\boldsymbol{Q}}$-norm approximates the continuous $\widetilde{H}$-norm of Chapter 2 ($\widetilde{H}$ was the Hilbert space in which the adjoint system was well posed).

3.2 Discretization in Time

We apply a *method of lines* approach to time–space discretization. This means we first discretize in space only, thereby obtaining an ODE. Next we apply an ODE solution method to discretize in time.

This section describes two such ODE solution schemes: The explicit midpoint rule and the trapezoid rule. In both cases we will analyze their stability, obtain solution formulas in terms of eigenvectors and eigenvalues, and introduce appropriate discrete norms.

Consider initially an ODE with the general formulation,

$$\dot{u}(t) = f(u(t), t) \,, \tag{3.20}$$

where $u : \mathbb{R} \mapsto V$ for some appropriate vector space V. For use in the discretization, we introduce the time points $t^n = n\Delta t$, where the time step $\Delta t \neq 0$ is constant, and $u^n \simeq u(t^n)$.

3.2.1 Stability of ODEs

The following treatment of stability for ODEs is fairly standard, see, for instance, *Trefethen (1996)*.

Consider the simple case of $f(u,t) = \lambda u$ with $\mathrm{Re}\,\lambda \leq 0$ (we use $\mathrm{Re}\,\lambda$ to refer to the real part of $\lambda \in \mathbb{C}$). The true solution, $u(t) = u(0)e^{\lambda t}$, fulfills $|u(t)| \leq |u(0)|$ for $t \geq 0$ and it is thus a reasonable requirement of an ODE scheme for this f, that $|\boldsymbol{u}^n|$ stays bounded as $n \to \infty$. When this is the case for a particular choice of λ and Δt, we call the scheme *eigenvalue stable*.

Let now $S \subset \mathbb{C}$ be a subset of the complex plane for which $\lambda\Delta t \in S$ if and only if the scheme is eigenvalue stable for this choice of λ and Δt. We then call S the *stability region* of the particular ODE scheme.

Consider now the linear ODE,

$$\dot{\boldsymbol{u}}(t) = \boldsymbol{\mathcal{A}}\boldsymbol{u}(t) \,, \tag{3.21}$$

where we assume that the matrix $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N\times N}$ is diagonalizable, $\boldsymbol{\mathcal{A}} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1}$ with $\boldsymbol{D} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$. Setting $\boldsymbol{y}(t) = \boldsymbol{V}^{-1}\boldsymbol{u}(t)$, the above system becomes equivalent to

$$\dot{\boldsymbol{y}}(t) = \boldsymbol{D}\boldsymbol{y}(t) \quad \text{or} \quad \dot{\boldsymbol{y}}(k)(t) = \lambda_k \boldsymbol{y}(k)(t) \quad \text{for } k = 1, 2, \ldots, N\,.$$

The reason for studying the case $f(u,t) = \lambda u$ is now clear. If it is possible to choose Δt such that $\lambda_k \Delta t$ lies in the stability region for all k, then we are sure that the approximate solution $\boldsymbol{y}^n$ does not "blow up". In turn, $\boldsymbol{u}^n = \boldsymbol{V}\boldsymbol{y}^n$ will stay bounded. Since the stability thus comes down to Δt and the eigenvalues of $\boldsymbol{\mathcal{A}}$, we see the reason for the term *eigenvalue stability*.

3.2.2 The Explicit Midpoint Rule

The explicit midpoint rule has the following appearance,

$$\frac{u^{n+1} - u^{n-1}}{2\Delta t} = f(u^n, t^n)\,.$$

It is explicit since u^{n+1} can be isolated, $u^{n+1} = 2\Delta t f(u^n, t^n) + u^{n-1}$, without having further information about $f(u,t)$. It is furthermore a two-step rule since u^{n+1} depends on the values of both u^n and u^{n-1}.

3.2.2.1 First Order Equations

To investigate eigenvalue stability we set $f(u,t) = \lambda u$ and get

$$\begin{bmatrix} u^n \\ u^{n+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 2\Delta t\lambda \end{bmatrix} \begin{bmatrix} u^{n-1} \\ u^n \end{bmatrix},$$

when formulating it as a one-step scheme. An eigenvalue σ of the one-step-forward matrix is seen to satisfy the equation

$$\sigma^2 - 2\Delta\lambda\sigma - 1 = 0 \quad \Leftrightarrow \quad \sigma - \frac{1}{\sigma} = 2\Delta\lambda\,.$$

This shows that if an eigenvalue has $|\sigma| < 1$ then one also exists with $|\sigma| > 1$. Thus for eigenvalue stability we must have $\sigma = e^{i\theta}$ for $\theta \in \mathbb{R}$. We now get

$$e^{i\theta} - e^{-i\theta} = 2i\sin(\theta) = 2\Delta t\lambda \quad \Leftrightarrow \quad \sin(\theta) = -i\Delta t\lambda\,,$$

which shows that λ must be purely imaginary and $|\Delta t\lambda| \le 1$. This makes the explicit midpoint rule especially suited for hyperbolic systems.

3.2.2.2 Second Order Equations

We turn to a second order system $\boldsymbol{C}\ddot{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{u}$, where $\boldsymbol{C} \in \mathbb{R}^{N\times N}$ and $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ are both symmetric, and positive and negative definite, respectively. We rewrite into a first order system as in (3.14) and introduce it to the explicit midpoint scheme:

$$\begin{bmatrix} \boldsymbol{u}^{n+1/2} \\ \boldsymbol{v}^{n+1/2} \end{bmatrix} - \begin{bmatrix} \boldsymbol{u}^{n-1/2} \\ \boldsymbol{v}^{n-1/2} \end{bmatrix} = \Delta t \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{C}^{-1}\boldsymbol{A} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}^n \\ \boldsymbol{v}^n \end{bmatrix}.$$

The reason for using half time steps become clear since the relations

$$\boldsymbol{u}^{n+1/2} - \boldsymbol{u}^{n-1/2} = \Delta t \ \boldsymbol{v}^n,$$
$$\boldsymbol{v}^{n+1/2} - \boldsymbol{v}^{n-1/2} = \Delta t \ \boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{u}^n \ ,$$

can be appropriately combined into the well-known

$$\boldsymbol{C}\frac{\boldsymbol{u}^{n+1} - 2\boldsymbol{u}^n + \boldsymbol{u}^{n-1}}{\Delta t^2} = \boldsymbol{A}\boldsymbol{u}^n \ . \tag{3.22}$$

Let us find an expression for $\boldsymbol{u}^n$ in terms of $\boldsymbol{C}^{-1}\boldsymbol{A}$'s eigenvectors,

$$\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{w}_k = \lambda_k \boldsymbol{w}_k = -\mu_k^2 \boldsymbol{w}_k \ ,$$

$k = 1, 2, \ldots, N$. We reformulate (3.22) into

$$\begin{bmatrix} \boldsymbol{u}^n \\ \boldsymbol{u}^{n+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ -\boldsymbol{I} & 2\boldsymbol{I} + \Delta t^2 \boldsymbol{C}^{-1}\boldsymbol{A} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}^{n-1} \\ \boldsymbol{u}^n \end{bmatrix} \ ,$$

and consider therefore the eigenvalue problem,

$$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ -\boldsymbol{I} & 2\boldsymbol{I} + \Delta t^2 \boldsymbol{C}^{-1}\boldsymbol{A} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_k \\ \sigma_k \boldsymbol{w}_k \end{bmatrix} = \sigma_k \begin{bmatrix} \boldsymbol{w}_k \\ \sigma_k \boldsymbol{w}_k \end{bmatrix} \ .$$

We get

$$(2 - \Delta t^2 \mu_k^2)\sigma_k - 1 = \sigma_k^2 \quad \Leftrightarrow \quad \sigma_k + \frac{1}{\sigma_k} = 2 - \Delta t^2 \mu_k^2 \ ,$$

and again we must have $\sigma_k = e^{i\theta_k}$, $\theta_k \in \mathbb{R}$. This implies that

$$\cos(\theta_k) = 1 - \tfrac{1}{2}\Delta t^2 \mu_k^2, \quad k = 1, 2, \ldots, N \ . \tag{3.23}$$

Thus the eigenvalue stability criterion is $\Delta t \leq 2/\mu_k$ for $k = 1, 2, \ldots, N$ and real solutions must have the appearance

$$\boldsymbol{u}^n = \sum_{k=1}^{N} \left[\tilde{a}_k \cos(n\theta_k) + \tilde{b}_k \sin(n\theta_k) \right] \boldsymbol{w}_k \ , \tag{3.24}$$

where the coefficients $\langle \tilde{a}_k \rangle_{k=1}^N$ and $\langle \tilde{b}_k \rangle_{k=1}^N$ are to be determined. We do that from the following initial conditions and approximation to $\dot{\boldsymbol{u}}(0)$ (compare to (3.13) and (3.17)):

$$\boldsymbol{u}^0 = \sum_{k=1}^{N} a_k \boldsymbol{w}_k, \qquad \frac{\boldsymbol{u}^1 - \boldsymbol{u}^{-1}}{2\Delta t} = \sum_{k=1}^{N} b_k \boldsymbol{w}_k \ . \tag{3.25}$$

Inserting (3.24) into these conditions we get

$$\tilde{a}_k = a_k \qquad \text{and} \qquad \tilde{b}_k = \frac{\Delta t}{\sin(\theta_k)} b_k = \frac{1}{\mu_k \sqrt{1 + \frac{1}{4}\Delta t^2 \mu_k^2}} b_k \ ,$$

where the relation (3.23) has been used. To summarize, the complete solution to (3.22) with initial conditions (3.25) is given by

$$\boldsymbol{u}^n = \sum_{k=1}^{N} \left[a_k \cos(n\theta_k) + \frac{1}{\mu_k \sqrt{1 + \frac{1}{4}\Delta t^2 \mu_k^2}} b_k \sin(n\theta_k) \right] \boldsymbol{w}_k \,, \tag{3.26}$$

where each θ_k must satisfy (3.23).

3.2.2.3 Energy Norm for Second Order Equations

To have an energy measure with properties similar to that of the semi-discrete system (3.19) (and, in turn, to that of a continuous system), we define

$$\begin{aligned} E^n &= \tfrac{1}{2} \left[\left\langle \frac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^n}{\Delta t}, \boldsymbol{C} \frac{\boldsymbol{u}^n - \boldsymbol{u}^{n-1}}{\Delta t} \right\rangle - \langle \boldsymbol{u}^n, \boldsymbol{A}\boldsymbol{u}^n \rangle \right] \\ &= \frac{1}{2\Delta t^2} \Big[\langle \boldsymbol{u}^n, \boldsymbol{C}\boldsymbol{u}^n \rangle - \langle \boldsymbol{u}^{n+1}, \boldsymbol{C}\boldsymbol{u}^{n-1} \rangle \Big] \,. \end{aligned} \tag{3.27}$$

Consider now

$$\begin{aligned} 2\Delta t^2 (E^{n+1} - E^n) =& \Big(\langle \boldsymbol{u}^{n+1}, \boldsymbol{C}\boldsymbol{u}^{n+1} \rangle - \langle \boldsymbol{u}^{n+2}, \boldsymbol{C}\boldsymbol{u}^n \rangle \Big) \\ & - \Big(\langle \boldsymbol{u}^n, \boldsymbol{C}\boldsymbol{u}^n \rangle - \langle \boldsymbol{u}^{n+1}, \boldsymbol{C}\boldsymbol{u}^{n-1} \rangle \Big) \\ =& \Big(\langle \boldsymbol{u}^{n+1}, \boldsymbol{C}\boldsymbol{u}^{n+1} \rangle - 2\langle \boldsymbol{u}^{n+1}, \boldsymbol{C}\boldsymbol{u}^n \rangle + \langle \boldsymbol{u}^{n+1}, \boldsymbol{C}\boldsymbol{u}^{n-1} \rangle \Big) \\ & - \Big(\langle \boldsymbol{u}^n, \boldsymbol{C}\boldsymbol{u}^{n+2} \rangle - 2\langle \boldsymbol{u}^n, \boldsymbol{C}\boldsymbol{u}^{n+1} \rangle + \langle \boldsymbol{u}^n, \boldsymbol{C}\boldsymbol{u}^n \rangle \Big) \\ =& \langle \boldsymbol{u}^{n+1}, \Delta t^2 \boldsymbol{A}\boldsymbol{u}^n \rangle - \langle \boldsymbol{u}^n, \Delta t^2 \boldsymbol{A}\boldsymbol{u}^{n+1} \rangle = 0 \,, \end{aligned}$$

which shows that $E^n = E^0$ for all $n = 0, 1, 2, \ldots$.

An energy expression similar to that in (3.27) has previously been formulated in *Negreanu and Zuazua (2003)*. In that paper, they consider only the second order centered difference scheme for time and space discretization (corresponding to the midpoint rule with $\boldsymbol{C}_0$ and $\boldsymbol{A}$ from (3.5)). Their approach, however, does not use matrix notation and the above treatment is thus simpler and more general.

But can E^0 act as a norm? Let us insert the expression for the solution (3.26) into (3.27) for $n = 0$ and we get

$$E^0 = \frac{1}{2} \left(\sum_{k=1}^{N} \Big[(1 - \tfrac{1}{4}\Delta t^2 \mu_k^2) \mu_k^2 a_k^2 + b_k^2 \Big] \langle \boldsymbol{w}_k, \boldsymbol{C}\boldsymbol{w}_k \rangle \right) \,,$$

using Theorem 3.1.1 to eliminate all "mixed" inner products. The expression clearly shows that E^0 is a norm if and only if the condition $\Delta t < 2/\mu_k$ is satisfied for all k. Note that this is similar to the stability criterion. A norm corresponding to the energy can be written

$$\left\| \begin{bmatrix} \boldsymbol{u}^0 \\ \overline{\boldsymbol{u}}^0 \end{bmatrix} \right\|_{\widetilde{\boldsymbol{Q}}}^2 = \begin{bmatrix} \boldsymbol{u}^0 \\ \overline{\boldsymbol{u}}^0 \end{bmatrix}^T \widetilde{\boldsymbol{Q}} \begin{bmatrix} \boldsymbol{u}^0 \\ \overline{\boldsymbol{u}}^0 \end{bmatrix}, \qquad \text{where} \quad \widetilde{\boldsymbol{Q}} = \begin{bmatrix} -\boldsymbol{A}(\boldsymbol{I} + \frac{1}{4}\Delta t^2 \boldsymbol{C}^{-1}\boldsymbol{A}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix} .$$

3.2.3 The Trapezoid Rule

We now consider the trapezoid rule, which has the general formulation,

$$\frac{u^{n+1}-u^n}{\Delta t} = \frac{f(u^{n+1},t^{n+1})+f(u^n,t^n)}{2},$$

an implicit one-step scheme. This scheme has some nice properties when it comes to stability and, as we shall see later, to controllability.

3.2.3.1 First Order Equations

Let us again start out with the simple case $f(u,t)=\lambda u$. We get

$$u^{n+1}-u^n = \tfrac{1}{2}\Delta t\lambda(u^{n+1}+u^n) \quad \Leftrightarrow \quad u^{n+1} = \left(\frac{1+\frac{1}{2}\Delta t\lambda}{1-\frac{1}{2}\Delta t\lambda}\right)u^n .$$

Setting $\alpha=\Delta t\lambda$ we have

$$u^n = \left(\frac{2+\alpha}{2-\alpha}\right)^n u^0 , \tag{3.28}$$

so finding the stability region is easy,

$$\left|\frac{2+\alpha}{2-\alpha}\right|^2 = \frac{4+2\alpha+2\overline{\alpha}+|\alpha|^2}{4-2\alpha-2\overline{\alpha}+|\alpha|^2} \le 1 \quad \Leftrightarrow \quad \mathrm{Re}(\alpha)\le 0 , \tag{3.29}$$

where $\overline{\alpha}$ denotes the complex conjugate of α.

Considering now the case $f(\boldsymbol{u},t)=\boldsymbol{A}\boldsymbol{u}$, we get

$$(\boldsymbol{I}-\tfrac{1}{2}\Delta t\boldsymbol{A})\boldsymbol{u}^{n+1} = (\boldsymbol{I}+\tfrac{1}{2}\Delta t\boldsymbol{A})\boldsymbol{u}^n ,$$

from which the implicitness is obvious. Using the eigenvalue information (3.12), the solution is easily written using (3.28),

$$\boldsymbol{u}^n = \sum_{k=1}^{N} a_k\rho_k^n\boldsymbol{w}_k ,$$

for $n=1,2,\ldots$, where

$$\boldsymbol{u}^0 = \sum_{k=1}^{N} a_k\boldsymbol{w}_k, \qquad \text{and} \qquad \rho_k = \frac{2+\Delta t\lambda_k}{2-\Delta t\lambda_k} .$$

3.2.3.2 Second Order Equations

We turn again to a second order system $\boldsymbol{C}\ddot{\boldsymbol{u}}=\boldsymbol{A}\boldsymbol{u}$. The trapezoid scheme applied to this system becomes

$$\begin{bmatrix}\boldsymbol{u}^{n+1}\\ \overline{\boldsymbol{u}}^{n+1}\end{bmatrix} - \begin{bmatrix}\boldsymbol{u}^{n}\\ \overline{\boldsymbol{u}}^{n}\end{bmatrix} = \tfrac{1}{2}\Delta t\boldsymbol{S}\left(\begin{bmatrix}\boldsymbol{u}^{n+1}\\ \overline{\boldsymbol{u}}^{n+1}\end{bmatrix} + \begin{bmatrix}\boldsymbol{u}^{n}\\ \overline{\boldsymbol{u}}^{n}\end{bmatrix}\right), \quad \boldsymbol{S} = \begin{bmatrix}\boldsymbol{0} & \boldsymbol{I}\\ \boldsymbol{C}^{-1}\boldsymbol{A} & \boldsymbol{0}\end{bmatrix} . \tag{3.30}$$

We let the initial conditions be as in (3.17), also written as

$$\begin{bmatrix} \boldsymbol{u}^0 \\ \overline{\boldsymbol{u}}^0 \end{bmatrix} = \sum_{1 \le |k| \le N} c_k \boldsymbol{z}_k, \qquad \text{with } c_k = \tfrac{1}{2}(a_k - ib_k/\mu_k),\ c_{-k} = \overline{c}_k,\ 1 \le k \le N\,,$$

using the eigenvectors of the matrix $\boldsymbol{S}$, see (3.15). We get

$$\rho_k = \frac{2 + \Delta t \sigma_k}{2 - \Delta t \sigma_k}, \quad 1 \le |k| \le N\,,$$

which implies that

$$\rho_k = \frac{2 + i\Delta t \mu_k}{2 - i\Delta t \mu_k} = \frac{4 - \Delta t^2 \mu_k^2}{4 + \Delta t^2 \mu_k^2} + i \frac{4 \Delta t \mu_k}{4 + \Delta t^2 \mu_k^2}, \qquad \rho_{-k} = \overline{\rho}_k\,,$$

for $k = 1, 2, \ldots, N$. We observe that $|\rho_k| = 1$ and let $\theta_k \in (-\pi, \pi)$ be chosen such that $e^{i\theta_k} = \rho_k$. Note that $\theta_{-k} = -\theta_k$. The full solution is now

$$\begin{bmatrix} \boldsymbol{u}^n \\ \overline{\boldsymbol{u}}^n \end{bmatrix} = \sum_{1 \le |k| \le N} c_k \rho_k^n \boldsymbol{z}_k = \sum_{1 \le |k| \le N} c_k e^{i\theta_k n} \boldsymbol{z}_k\,,$$

or written out as in (3.18),

$$\begin{aligned} \boldsymbol{u}^n &= \sum_{k=1}^{N} \big[a_k \cos(\theta_k n) + b_k/\mu_k \sin(\theta_k n) \big] \boldsymbol{w}_k, \\ \overline{\boldsymbol{u}}^n &= \sum_{k=1}^{N} \big[-a_k \mu_k \sin(\theta_k n) + b_k \cos(\theta_k n) \big] \boldsymbol{w}_k\,. \end{aligned} \tag{3.31}$$

The trapezoid scheme for second order systems can be rewritten in an interesting way. Consider the relation from (3.30), written for two neighboring values of n:

$$\begin{aligned} \boldsymbol{u}^{n+1} - \boldsymbol{u}^n &= \tfrac{1}{2}\Delta t(\overline{\boldsymbol{u}}^{n+1} + \overline{\boldsymbol{u}}^n), & \boldsymbol{u}^n - \boldsymbol{u}^{n-1} &= \tfrac{1}{2}\Delta t(\overline{\boldsymbol{u}}^n + \overline{\boldsymbol{u}}^{n-1}), \\ \overline{\boldsymbol{u}}^{n+1} - \overline{\boldsymbol{u}}^n &= \tfrac{1}{2}\Delta t \boldsymbol{C}^{-1}\boldsymbol{A}(\boldsymbol{u}^{n+1} + \boldsymbol{u}^n), & \overline{\boldsymbol{u}}^n - \overline{\boldsymbol{u}}^{n-1} &= \tfrac{1}{2}\Delta t \boldsymbol{C}^{-1}\boldsymbol{A}(\boldsymbol{u}^n + \boldsymbol{u}^{n-1}). \end{aligned}$$

Combining these relations we get

$$\boldsymbol{u}^{n+1} - 2\boldsymbol{u}^n + \boldsymbol{u}^{n-1} = \tfrac{1}{2}\Delta t(\overline{\boldsymbol{u}}^{n+1} - \overline{\boldsymbol{u}}^{n-1}) = \tfrac{1}{4}\Delta t^2 \boldsymbol{C}^{-1}\boldsymbol{A}(\boldsymbol{u}^{n+1} + 2\boldsymbol{u}^n + \boldsymbol{u}^{n-1})\,,$$

or equivalently,

$$\boldsymbol{C}\frac{\boldsymbol{u}^{n+1} - 2\boldsymbol{u}^n + \boldsymbol{u}^{n-1}}{\Delta t^2} = \boldsymbol{A}\frac{\boldsymbol{u}^{n+1} + 2\boldsymbol{u}^n + \boldsymbol{u}^{n-1}}{4}\,. \tag{3.32}$$

Note the similarity to the box method space discretization, see (3.4) with $\alpha = 1/4$.

3.2.3.3 Energy Norm for Second Order Equations

As far as the author knows, a suitable energy norm has not previously been studied in the literature (at least not in the context of control theory).

We define the energy of the system (3.30) as

$$E^n = \tfrac{1}{2}\big(\langle \overline{\boldsymbol{u}}^n, \boldsymbol{C}\overline{\boldsymbol{u}}^n\rangle - \langle \boldsymbol{u}^n, \boldsymbol{A}\boldsymbol{u}^n\rangle\big)\,.$$

Note how well this energy expression corresponds to that of the semi-discrete system, see (3.19).

The energy is constant in time, which is seen from

$$\begin{aligned}2(E^{n+1}-E^n) =& \langle \overline{\boldsymbol{u}}^{n+1}, \boldsymbol{C}\overline{\boldsymbol{u}}^{n+1}\rangle - \langle \overline{\boldsymbol{u}}^n, \boldsymbol{C}\overline{\boldsymbol{u}}^n\rangle \\ &- \langle \boldsymbol{u}^{n+1}, \boldsymbol{A}\boldsymbol{u}^{n+1}\rangle + \langle \boldsymbol{u}^n, \boldsymbol{A}\boldsymbol{u}^n\rangle \\ =& \langle \overline{\boldsymbol{u}}^{n+1}+\overline{\boldsymbol{u}}^n, \boldsymbol{C}(\overline{\boldsymbol{u}}^{n+1}-\overline{\boldsymbol{u}}^n)\rangle - \langle \boldsymbol{u}^{n+1}-\boldsymbol{u}^n, \boldsymbol{A}(\boldsymbol{u}^{n+1}+\boldsymbol{u}^n)\rangle \\ =& \langle \overline{\boldsymbol{u}}^{n+1}+\overline{\boldsymbol{u}}^n, \tfrac{1}{2}\Delta t\boldsymbol{A}(\boldsymbol{u}^{n+1}+\boldsymbol{u}^n)\rangle \\ &- \langle \tfrac{1}{2}\Delta t(\overline{\boldsymbol{u}}^{n+1}+\overline{\boldsymbol{u}}^n), \boldsymbol{A}(\boldsymbol{u}^{n+1}+\boldsymbol{u}^n)\rangle = 0\,,\end{aligned}$$

so $E^n = E^0$ for all n.

We finally compute the energy when the initial conditions are given in terms of eigenvalues and eigenvectors,

$$\begin{aligned}E^0 &= \tfrac{1}{2}\big(\langle \overline{\boldsymbol{u}}^0, \boldsymbol{C}\overline{\boldsymbol{u}}^0\rangle - \langle \boldsymbol{u}^0, \boldsymbol{A}\boldsymbol{u}^0\rangle\big) \\ &= \tfrac{1}{2}\left(\sum_{k=1}^N b_k^2\langle \boldsymbol{w}_k, \boldsymbol{C}\boldsymbol{w}_k\rangle - \sum_{k=1}^N a_k^2\langle \boldsymbol{w}_k, \boldsymbol{A}\boldsymbol{w}_k\rangle\right) \\ &= \tfrac{1}{2}\sum_{k=1}^N (\mu_k^2 a_k^2 + b_k^2)\langle \boldsymbol{w}_k, \boldsymbol{C}\boldsymbol{w}_k\rangle\,.\end{aligned} \tag{3.33}$$

This shows that the energy induces a norm,

$$\left\|\begin{bmatrix}\boldsymbol{u}^0\\ \boldsymbol{v}^0\end{bmatrix}\right\|_{\widetilde{\boldsymbol{Q}}}^2 = \begin{bmatrix}\boldsymbol{u}^0\\ \boldsymbol{v}^0\end{bmatrix}^T \widetilde{\boldsymbol{Q}} \begin{bmatrix}\boldsymbol{u}^0\\ \boldsymbol{v}^0\end{bmatrix}, \qquad \text{where} \quad \widetilde{\boldsymbol{Q}} = \begin{bmatrix}-\boldsymbol{A} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}\end{bmatrix}\,.$$

Using an analogous procedure, it is straightforward to show that another norm is also constant in time for second order systems when using the trapezoid scheme,

$$\left\|\begin{bmatrix}\boldsymbol{u}^0\\ \boldsymbol{v}^0\end{bmatrix}\right\|_{\boldsymbol{Q}'}^2 = \begin{bmatrix}\boldsymbol{u}^0\\ \boldsymbol{v}^0\end{bmatrix}^T \boldsymbol{Q}' \begin{bmatrix}\boldsymbol{u}^0\\ \boldsymbol{v}^0\end{bmatrix}, \qquad \text{where} \quad \boldsymbol{Q}' = \begin{bmatrix}\boldsymbol{C} & \boldsymbol{0}\\ \boldsymbol{0} & -\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{C}\end{bmatrix}\,.$$

For the wave equation with homogeneous Dirichlet boundary conditions, where $\boldsymbol{C}^{-1}\boldsymbol{A}$ approximates the Laplacian, Δ, the $\|\cdot\|_{\widetilde{\boldsymbol{Q}}}$-norm approximates the $H_0^1(\Omega)\times L^2(\Omega)$-norm. Similarly, the $\|\cdot\|_{\boldsymbol{Q}'}$-norm approximates the $L^2(\Omega)\times H^{-1}(\Omega)$-norm. The continuous wave equation is well posed in both norms (the discrete $\boldsymbol{Q}'$-norm corresponds to the continuous H'-norm of Chapter 2, in which the control system is well posed).

3.3 Convergence of PDEs

To solve PDEs approximately, one can apply the following *method of lines* approach: Discretize in space to obtain an Nth order ODE, then use an ODE solver to integrate in time, choosing Δt small enough for eigenvalue stability. Now simultaneously increasing N and appropriately decreasing Δt, we obtain better and better approximations to the continuous PDE solution—or do we? The procedure can actually fail.

We need to make the concept of convergence more precise, and to that end we need a stronger form of stability. Let the PDE we wish to approximate have the generic appearance,

$$\begin{cases} \dfrac{du(t)}{dt} = Au(t) & \text{in } (0,T)\times\Omega\,, \\ u(0) = u^0 & \text{in } \Omega\,, \end{cases} \tag{3.34}$$

which is assumed well posed in the sense that the solution lies in the Hilbert space H, $u(t) \in H$, for $0 \leq t < T$. Note that boundary conditions are assumed built into the differential operator A and the Hilbert space H.

We will use the following abstract formulation for a PDE discretized in both space and time,

$$\boldsymbol{u}_N^{n+1} = \boldsymbol{S}(N)\boldsymbol{u}_N^n\,,$$

where the subscripts emphasize the dimension of the space discretization. We assume that an appropriate time step Δt is built into the operator $\boldsymbol{S}(N)$. We thus have

$$\boldsymbol{u}_N^n = \boldsymbol{S}(N)^n\boldsymbol{u}_N^0\,. \tag{3.35}$$

We need to be able to set the initial condition $\boldsymbol{u}_N^0$ for the discrete system from the continuous function u^0. To that end we introduce an operator $R_N : H \mapsto \mathbb{R}^N$ that does the job,

$$\boldsymbol{u}_N^0 = R_N u^0\,.$$

Conversely, to speak of convergence, we must be able to compare $\boldsymbol{u}_N^n$ to $u(n\Delta t)$. We introduce $E_N : \mathbb{R}^N \mapsto H$ that interpolates a vector in $\mathbb{R}^N$ of the discrete system to the Hilbert space H. We require that $R_N E_N = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity in $\mathbb{R}^N$, and we set $S(N) = E_N\boldsymbol{S}(N)R_N$. We can now formulate (3.35) as

$$u_N^n = S(N)^n u^0\,,$$

where $u_N^n \in H$ for all n, N and we see that $\boldsymbol{u}_N^n = R_N u_N^n$. We are now ready to define convergence.

Definition 3.3.1. *Let $N_1, N_2, \ldots$ be a sequence of natural numbers. Let corresponding time steps $\Delta t_j > 0$ and $n_j \in \mathbb{N}$ be chosen such that $\Delta t_j \to 0$ and $\Delta t_j n_j \to t$ where $0 \leq t \leq T$. A PDE discretization is now said to be* convergent *if*

$$\|S(N_j)^{n_j}u(0) - u(t)\|_H \to 0, \quad \text{as } j \to \infty\,,$$

for all solutions $u(t)$ to (3.34).

It turns out that convergence is intimately tied to two other concepts: stability and consistency. The following type of stability makes sure that the numerical solution can not "blow up".

Definition 3.3.2. *A PDE discretization is* stable *if a constant $C > 0$ exists such that*

$$\left\| S(N)^n u^0 \right\|_H \leq C \,,$$

for all n where $0 \leq n\Delta t \leq T$ and for all N (recall that Δt depends on N).

Stability, however, is *independent* of the continuous system. We need to make sure that the discrete scheme is actually a discretization of the right equation. This is the subject of consistency.

Definition 3.3.3. *A PDE discretization is* consistent *with the differential equation (3.34) if*

$$\left\| \left(\frac{S(N) - I}{\Delta t} - A \right) u(t) \right\|_H \to 0, \quad \textit{as } N \to \infty \,,$$

for every solution $u(t)$ of the system (3.34) with u^0 belonging to a dense subset of H.

The condition of this definition can be rewritten into something that is easier to handle in practice. Assume

$$\left\| S(N)u(t) - u(t + \Delta t) \right\|_H = \mathcal{O}(\Delta t^{p+1}), \quad \text{as } N \to \infty \,,$$

for any solution $u(t)$ of the system (3.34) with u^0 in a *dense subset* of H and where $0 \leq t < T$. The discretization is now consistent if $p > 0$. The number p is called the order of accuracy.

We are now ready for the main theorem of this section, a classical theorem of great importance.

Theorem 3.3.1 (Lax Equivalence Theorem). *Let a consistent discretization be given of a well-posed linear initial-value system of the type (3.34). The discretization scheme is now convergent if and only if it is stable.*

The proof can be found in *Lax and Richtmyer (1956)* and *Richtmyer and Morton (1967)*, which both contain further details on convergence of discretizations. See also *Trefethen (1996)*.

Note how it, in the light of the present section, does not make sense just to talk of whether a discretization scheme $\boldsymbol{S}_N$, in itself, is convergent or not. One also has to specify how to set the initial conditions of the discrete system (here done by R_N), how to put a discrete solution in the same space as the continuous system (here done by E_N) and finally by specifying in which norm the solution should converge (here $\| \cdot \|_H$). An example of rigorously showing convergence from consistency and stability for a particular discretization of the one dimensional wave equation can be found in Section 7.4.2.

3.4 Group Velocity for Hyperbolic Systems

This section will provide a short introduction to the concept of group velocity, mostly related to the discretizations that we will deal with in the present thesis. The presentation is primarily based on the paper *Trefethen (1982)*. We will, however, present some new insight by considering the general α-discretizations. We will furthermore derive a space–time discretization of the two-dimensional wave equation, which has never before been studied in the context of control.

Group velocity turns out to explain many things related to controllability. The explanations, however, turn out to be more intuitive than actual proofs. Some attempts have been made, though, to make rigorous proofs using the ideas of group velocity, see *Maciá (2003)*.

Let us start out in one space dimension. The central idea is to consider solutions of the form

$$u(t,x) = e^{i(\omega t - \xi x)} \,, \tag{3.36}$$

where ω is denoted the *frequency* and ξ the *wave number*. Inserting such a solution into the PDE in question leads to a *dispersion relation*,

$$\omega = \omega(\xi) \,,$$

which shows the *necessary relation* between frequencies and wave numbers. For instance, consider the wave equation

$$u_{tt} = u_{xx} \,, \tag{3.37}$$

which, when inserting (3.36), leads to $\omega^2 = \xi^2$. The quantity

$$c(\xi) = \frac{\omega(\xi)}{\xi} \,,$$

is called the *phase speed*, which is the speed with which the solution (3.36) travels to the right. It is, however, the *group velocity*,

$$C(\xi) = \frac{d\omega(\xi)}{d\xi} \,,$$

that dictates the speed with which wave packets of dominating wave number ξ travels.

Let us now consider a family of semi-discretization of the wave equation (3.37),

$$\alpha \ddot{\boldsymbol{u}}_{j+1} + (1-2\alpha)\ddot{\boldsymbol{u}}_j + \alpha \ddot{\boldsymbol{u}}_{j-1} = \frac{\boldsymbol{u}_{j+1} - 2\boldsymbol{u}_j + \boldsymbol{u}_{j-1}}{h^2} \,, \tag{3.38}$$

using the α-discretization of the Laplacian introduced in (3.4). We now insert (3.36) with $x = jh$ and get, after some manipulation,

$$\omega^2 = \frac{4\sin^2(\frac{1}{2}\xi h)}{h^2(1 - 4\alpha \sin^2(\frac{1}{2}\xi h))} = \xi^2 \left(\frac{\sin(\frac{1}{2}\xi h)}{\frac{1}{2}\xi h}\right)^2 \frac{1}{1 - 4\alpha\sin^2(\frac{1}{2}\xi h)} \tag{3.39}$$

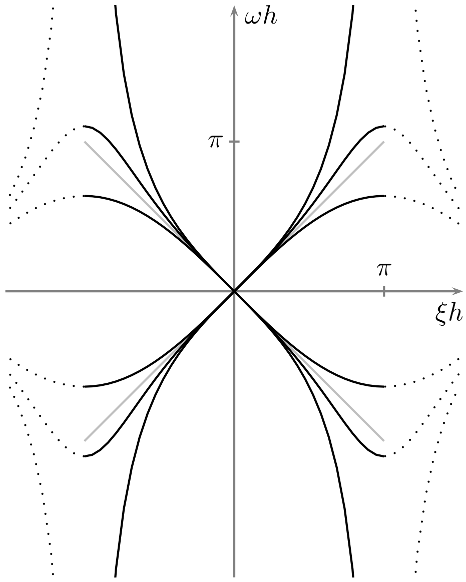


Figure 3.5: Illustration of the dispersion relation (3.39) for the semi-discretization shown in (3.38). The solid, black lines correspond, respectively, to $\alpha = 1/4$, $\alpha = 1/6$ and $\alpha = 0$ from the top (ignoring the mirror image below the ξh-axis). The dotted lines indicate the periodic nature of the dispersion relation, $\omega(\xi h) = \omega(\xi h + 2\pi)$.

(note the similarity to the expression (3.6) for the eigenvalues of the approximate Laplacian).

An illustration of the dispersion relation above can be seen in Figure 3.5. Around $\xi h = 0$ the group velocity (the slope of the curve) is close to one as required by consistency. However, around $\xi h = \pm\pi$ for $\alpha < 1/4$, the group velocity is close to zero. This means that such high frequency waves hardly move in the discrete media. But what do the waves look like when $\xi h \simeq \pi$? For $j \in \mathbb{Z}$ and $\epsilon \in \mathbb{R}$ we get

$$\cos((\pi - \epsilon)j) = \cos(\pi j)\cos(\epsilon j) + \sin(\pi j)\sin(\epsilon j) = (-1)^j \cos(\epsilon j)\,,$$

implying that $\cos((\pi - \epsilon)j) \simeq (-1)^j$ for $\epsilon j \ll 1$.

Let us move on to discretizing in time also. We start out with the midpoint scheme for the time discretization. Again we use the α-discretization in space, see (3.4) and (3.5),

$$\boldsymbol{C}_\alpha \frac{\boldsymbol{u}^{n+1} - 2\boldsymbol{u}^n + \boldsymbol{u}^{n-1}}{\Delta t^2} = \boldsymbol{A}\boldsymbol{u}^n\,. \tag{3.40}$$

Inserting a solution of the type

$$\boldsymbol{u}_j^n = e^{i(\omega n \Delta t - \xi j h)}\,, \tag{3.41}$$

we arrive at the relation

$$\sin^2(\tfrac{1}{2}\omega\Delta t) = \frac{\eta^2 \sin^2(\frac{1}{2}\xi h)}{1 - 4\alpha\sin^2(\frac{1}{2}\xi h)}, \quad \omega h = \pm\frac{2}{\eta}\arcsin\left(\frac{\eta\sin(\frac{1}{2}\xi h)}{\sqrt{1 - 4\alpha\sin^2(\frac{1}{2}\xi h)}}\right)\,. \tag{3.42}$$

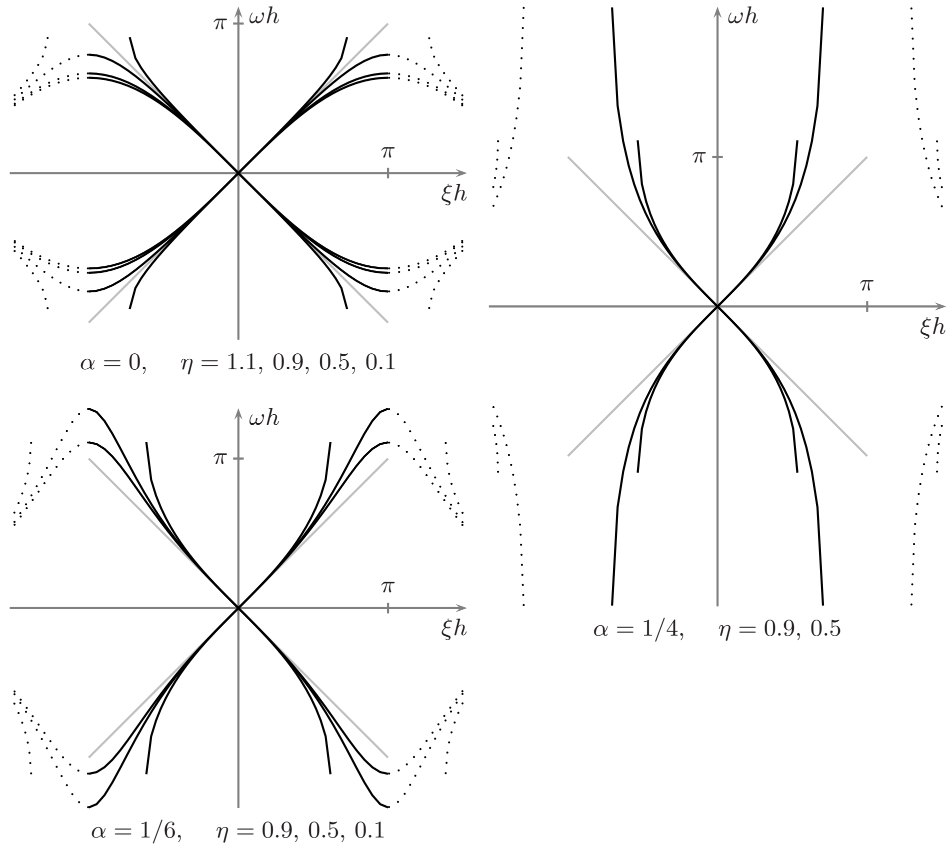


Figure 3.6: Dispersion relations of the 1D wave equation using the midpoint scheme for time discretization and α-discretization for the space discretization. Under each plot is shown the value of α together with the values of $\eta = \Delta t/h$ used, corresponding to the solid, black lines, counting from above. No curve has been cut in the ωh-direction, and missing parts indicate that the corresponding ωh has a non-zero imaginary part.

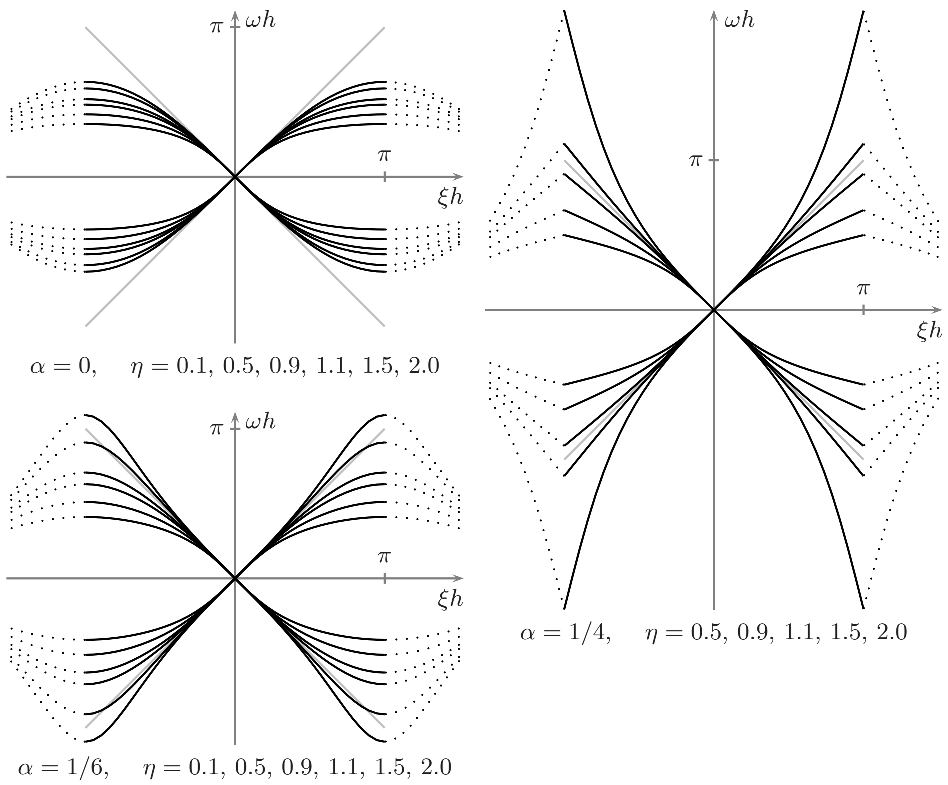


Figure 3.7: Dispersion relations of the 1D wave equation using the trapezoid scheme for time discretization and α-discretization for the space discretization. Under each plot is shown the value of α together with the values of $\eta = \Delta t/h$ used, corresponding to the solid, black lines, counting from above. No curve has been cut in the ωh-direction.

This relation is shown for the values $\alpha = 0,\ 1/6,\ 1/4$ in Figure 3.6. Note that these exhibit the same deficiencies for $\xi h \simeq \pi$ as for the semi-discrete case.

We now turn to the trapezoid method for time discretization, and we will use the formulation from (3.32),

$$\boldsymbol{C}_\alpha \frac{\boldsymbol{u}^{n+1} - 2\boldsymbol{u}^n + \boldsymbol{u}^{n-1}}{\Delta t^2} = \boldsymbol{A} \frac{\boldsymbol{u}^{n+1} + 2\boldsymbol{u}^n + \boldsymbol{u}^{n-1}}{4}\,.$$

We insert the (local) solution (3.39) once again and obtain

$$\tan^2(\tfrac{1}{2}\omega\Delta t) = \frac{\eta^2 \sin^2(\frac{1}{2}\xi h)}{1 - 4\alpha \sin^2(\frac{1}{2}\xi h)}, \quad \omega h = \pm\frac{2}{\eta} \arctan\left(\frac{\eta \sin(\frac{1}{2}\xi h)}{\sqrt{1 - 4\alpha\sin^2(\frac{1}{2}\xi h)}} \right).$$

This relation is shown for the values $\alpha = 0,\ 1/6,\ 1/4$ in Figure 3.7. Note in the figure the dispersion relation for the very important case of $\alpha = 1/4$,

$$\tan^2(\tfrac{1}{2}\omega\Delta t) = \eta^2 \tan^2(\tfrac{1}{2}\xi h)\,. \tag{3.43}$$

This case has the very special feature that the group velocity is not $\simeq 0$ for $\xi h \simeq \pm\pi$, or anywhere else for that matter. This creates hope that the scheme will do well when it comes to control. We shall study the scheme in Section 7.4.

Let us consider an example to illustrate just how exact the previous concepts can predict wave propagation in a discrete media. Consider the function

$$f(x) = \begin{cases} \sin(\xi x)\exp\Big(\frac{4}{(2x+1)(2x-1)}\Big), & -\frac{1}{2} < x < \frac{1}{2}\,, \\ 0, & \text{otherwise}\,, \end{cases}$$

which is a sine wave with wave number ξ, multiplied point-wise with a smooth pulse function, supported in $(-\frac{1}{2}, \frac{1}{2})$. It is now clear that $u(t,x) = f(x-t)$ is a solution to the one-dimensional wave equation, $u_{tt} = u_{xx}$, on the real line $\mathbb{R}$. For the discretization, we use the second order centered difference formula in both space and time, that is, we use (3.40) with $\alpha = 0$. In turn, the dispersion relation is seen from (3.42), leading to the group velocity

$$\omega'(\xi) = \pm\frac{\cos(\frac{1}{2}\xi h)}{\sqrt{1-\eta^2\sin^2(\frac{1}{2}\xi h)}}$$

(easiest obtained by applying implicit differentiation to the left-most expression in (3.42)). In our example we have $h = 1/300$, $\eta = 0.1$ and $\Delta t = \eta h$. We set $\xi = 100$ as the wave number for the function f, and by using initial conditions $u(0,x) = f(x)$, $u_t(0,x) = -f'(x)$ we get the solution $u(t,x) = f(x-t)$, that is, the function f traveling right at speed one. Inserting these numbers, we get the group velocity $\omega'(\xi) \simeq \pm 0.7860$. This means that in time one, the true solution has travelled one unit to the right whereas a discrete wave with wave number $\xi = 100$ should travel the distance 0.7860 either left or right. By looking at Figure 3.8, we see that this is highly accurate.

3.4.1 Group Velocity in 2D

In two (or more) dimensions we simply insert waves of the type

$$u(t,\boldsymbol{x}) = e^{i(\omega t - \boldsymbol{\xi}\cdot\boldsymbol{x})}\,,$$

with $\boldsymbol{\xi} \in \mathbb{R}^d$ and thereby obtaining a dispersion relation of the type

$$\omega = \omega(\boldsymbol{\xi})\,.$$

The group velocity now becomes a vector field,

$$\boldsymbol{C} = \nabla\omega\,.$$

The length and direction of the vector $\boldsymbol{C}(\boldsymbol{\xi})$ reveals the speed and the direction of the corresponding wave, respectively.

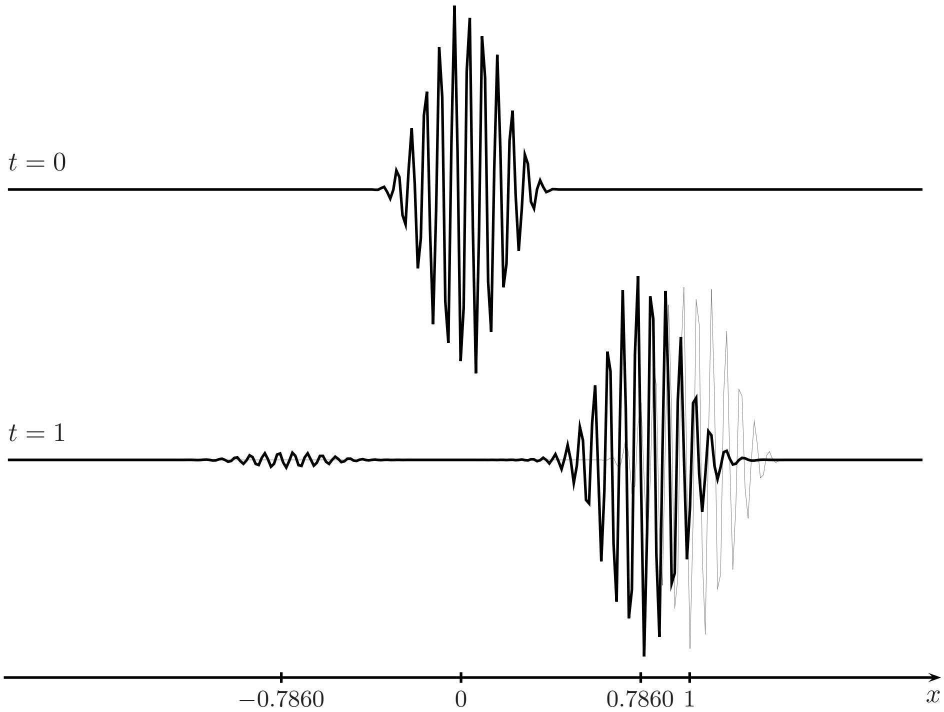


Figure 3.8: Propagation of a wave in a discrete medium with grid size $h = 1/300$. The top-most plot shows the initial condition, a dampened $\sin(100x)$ wave. The initial velocity is chosen such that the true solution should travel to the right at speed 1. This is illustrated in gray in the bottom plot. However, because of discretization effects, solving the wave equation using a finite difference scheme, group velocity calculations predict that this particular wave should propagate at speed ± 0.7860. This is seen to be highly accurate.

Let us consider the wave equation in two dimensions,

$$u_{tt} = u_{xx} + u_{yy}\ , \tag{3.44}$$

which has the simple dispersion relation $\omega^2 = \xi_1^2 + \xi_2^2$, using $\boldsymbol{\xi} = (\xi_1, \xi_2)$.

Consider as an example the following finite difference discretization of the two-dimensional wave equation,

$$\frac{\boldsymbol{u}_{j,k}^{n+1} - 2\boldsymbol{u}_{j,k}^{n} + \boldsymbol{u}_{j,k}^{n-1}}{\Delta t^2} = \frac{\boldsymbol{u}_{j+1,k}^{n} - 2\boldsymbol{u}_{j,k}^{n} + \boldsymbol{u}_{j-1,k}^{n}}{h^2} + \frac{\boldsymbol{u}_{j,k+1}^{n} - 2\boldsymbol{u}_{j,k}^{n} + \boldsymbol{u}_{j,k-1}^{n}}{h^2}\ . \tag{3.45}$$

Inserting

$$\boldsymbol{u}_{j,k}^{n} = e^{i(\omega\Delta t n - \xi_1 jh - \xi_2 kh)}\ , \tag{3.46}$$

we arrive at

$$\sin^2(\tfrac{1}{2}\omega\Delta t) = \eta^2\big[\sin^2(\tfrac{1}{2}\xi_1 h) + \sin^2(\tfrac{1}{2}\xi_2 h)\big]\ . \tag{3.47}$$

The dispersion relation for $\eta = \Delta t/h = 1/\sqrt{2}$ can be seen in Figure 3.9. The contour plot here is very informative. It shows level curves in the (ξ_1, ξ_2)-plane for constant values of $\omega h\eta = \pi/20,\ 2\pi/20,\ \ldots,\ 19\pi/20$. The distance between two adjacent curves indicates the speed of a wave with a corresponding (ξ_1, ξ_2) wave number. For comparison, the distance between the curves around $(\xi_1, \xi_2) = (0, 0)$ corresponds to waves with the true unit wave speed of the underlying system. The level curves provide further useful information: The direction of a wave is perpendicular to a corresponding level curve. We know from the continuous system that the dispersion relation is $\omega^2 = \xi_1^2 + \xi_2^2$, meaning that the direction of a (ξ_1, ξ_2) wave is the same as a $(0, 0) \to (\xi_1, \xi_2)$ vector. For our discrete system, as seen in the figure, this fits well around $(\xi_1, \xi_2) = (0, 0)$ and around the lines $\xi_1 = \pm\xi_2$. But other places, for instance around $(\xi_1, \xi_2) = (\pi, \pi/5)$, the direction is almost *perpendicular* to the true direction!

Does a discretization of the two-dimensional wave equation (3.44) exist that has the same advantageous properties, when it comes to the dispersion relation, as (3.43) for the one-dimensional case?

Let us start out with the dispersion relation that we would like. A relation like

$$\tan^2(\tfrac{1}{2}\omega\Delta t) = \eta^2\big[\tan^2(\tfrac{1}{2}\xi_1 h) + \tan^2(\tfrac{1}{2}\xi_2 h)\big]\ , \tag{3.48}$$

would be very nice, see Figure 3.10. All waves travel with speed one or a little more (remember that when it comes to control, it is waves with speed less than one that are problematic). The direction of the waves are reasonable, although some inaccuracies occur (as usual) when $\xi_1 \simeq \pm\pi$ and/or $\xi_2 \simeq \pm\pi$.

We now make the following rewrites of (3.48),

$$\begin{aligned}\sin^2(\tfrac{1}{2}\omega\Delta t)\cos^2(\tfrac{1}{2}\xi_1 h)\cos^2(\tfrac{1}{2}\xi_2 h) =& \eta^2 \sin^2(\tfrac{1}{2}\xi_1 h)\cos^2(\tfrac{1}{2}\xi_2 h)\cos^2(\tfrac{1}{2}\omega\Delta t)\\ &+ \eta^2 \sin^2(\tfrac{1}{2}\xi_2 h)\cos^2(\tfrac{1}{2}\xi_1 h)\cos^2(\tfrac{1}{2}\omega\Delta t) \quad \Leftrightarrow\end{aligned}$$

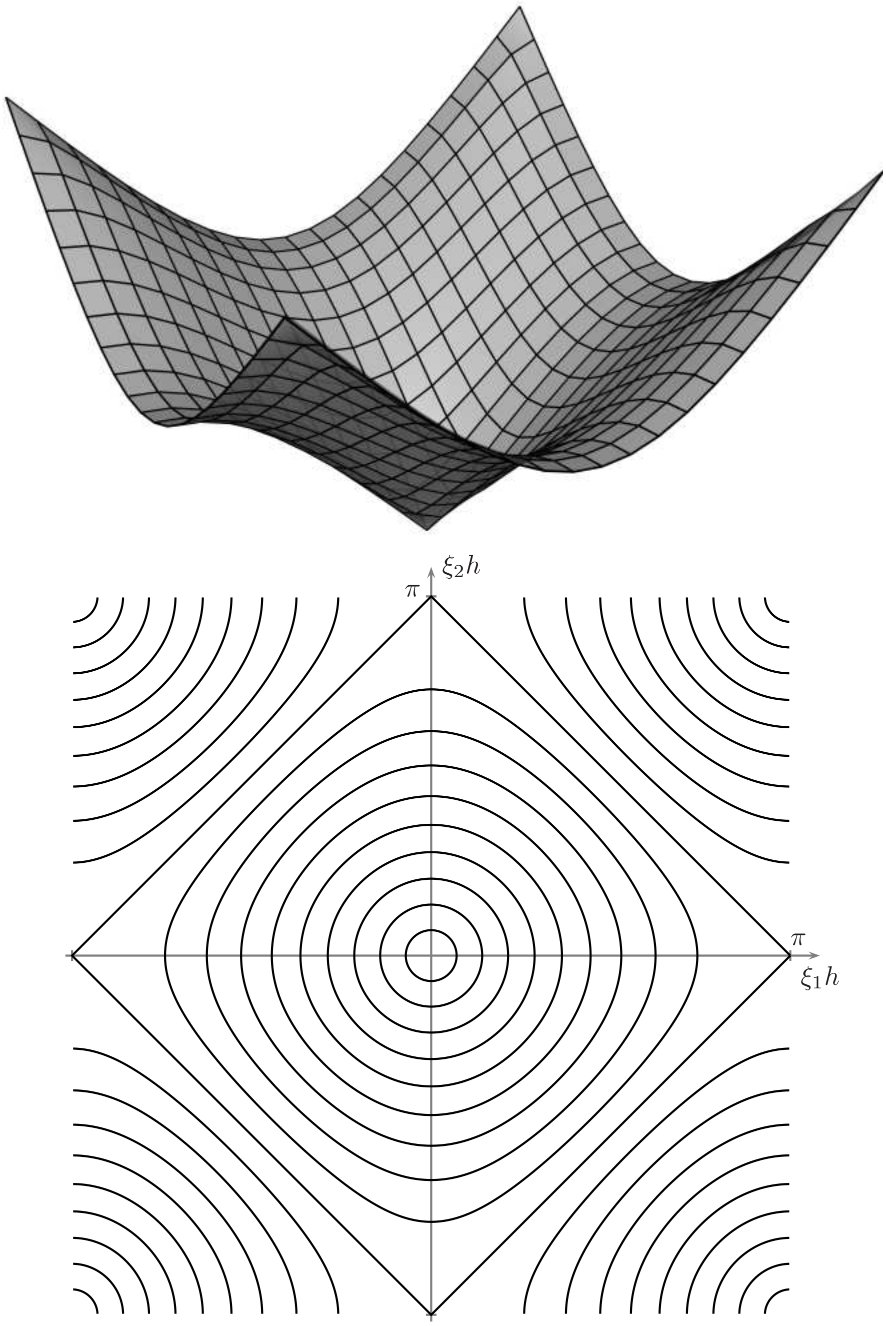


Figure 3.9: Illustrations of the 2D dispersion relation (3.47) with $\eta = 1/\sqrt{2}$. In the contour plot, the curves correspond to $\omega h \eta = \pi/20$, $2\pi/20$, $\ldots$, $19\pi/20$.

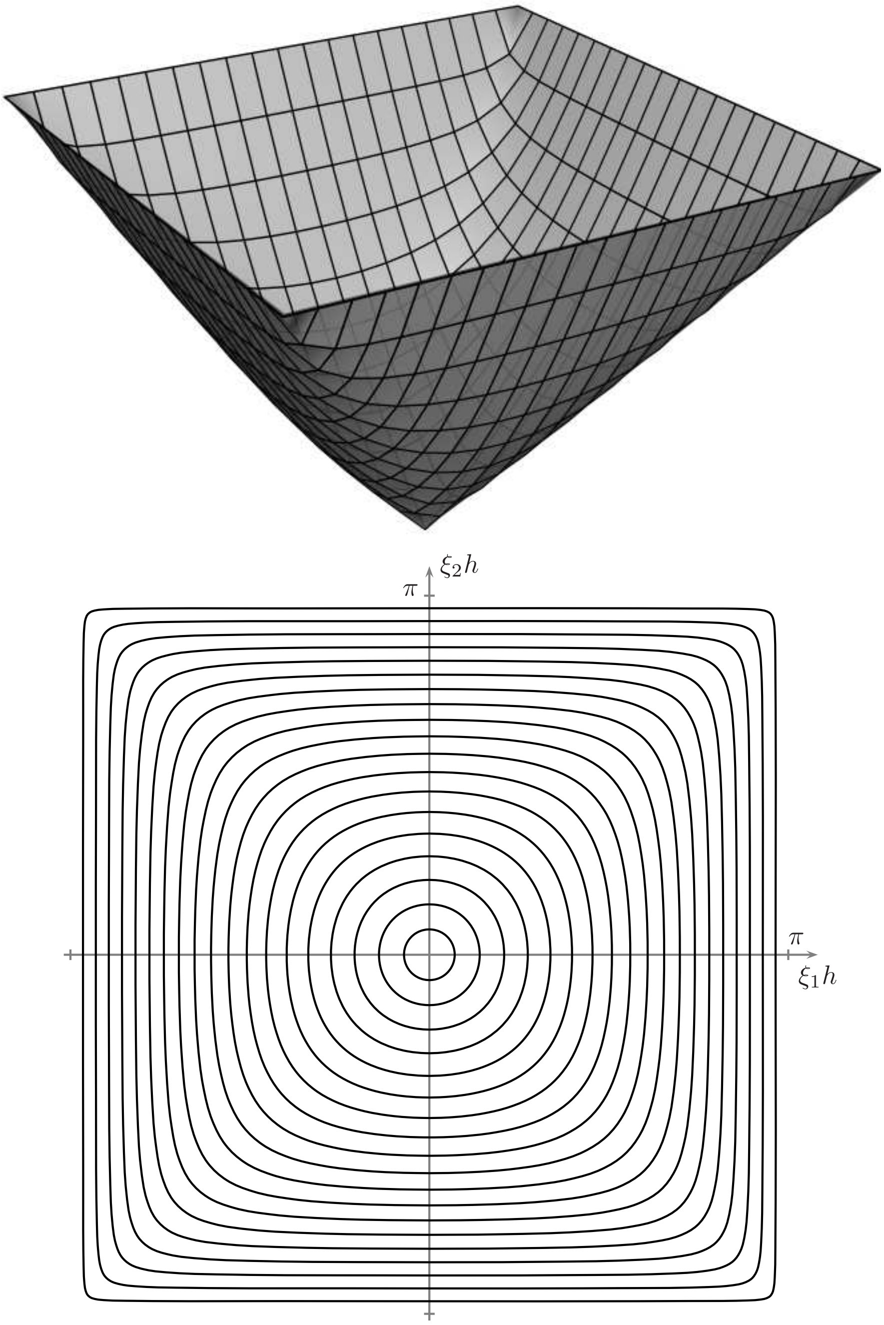


Figure 3.10: Illustrations of the 2D dispersion relation (3.48) with $\eta = 1/\sqrt{2}$. In the contour plot, the curves correspond to $\omega h \eta = \pi/20,\ 2\pi/20,\ \ldots,\ 19\pi/20$.

$$
\begin{aligned}
&(1-\cos(\omega\Delta t))(1+\cos(\xi_1 h))(1+\cos(\xi_2 h)) \\
&\qquad = \eta^2(1-\cos(\xi_1 h))(1+\cos(\xi_2 h))(1+\cos(\omega\Delta t)) \\
&\qquad\quad + \eta^2(1-\cos(\xi_2 h))(1+\cos(\xi_1 h))(1+\cos(\omega\Delta t)) \quad \Leftrightarrow \\
&(e^{i\omega\Delta t}-2+e^{-i\omega\Delta t})(e^{i\xi_1 h}+2+e^{-i\xi_1 h})(e^{i\xi_2 h}+2+e^{-i\xi_2 h}) \\
&\qquad = \eta^2(e^{i\xi_1 h}-2+e^{-i\xi_1 h})(e^{i\xi_2 h}+2+e^{-i\xi_2 h})(e^{i}+2+e^{-i\omega\Delta t}) \\
&\qquad\quad + \eta^2(e^{i\xi_1 h}-2+e^{-i\xi_1 h})(e^{i\xi_2 h}+2+e^{-i\xi_2 h})(e^{i}+2+e^{-i\omega\Delta t})\,.
\end{aligned}
$$

Now multiplying out the parenthesis and multiplying both sides of the equation by $e^{i(\omega n\Delta t-\xi_1 jh-\xi_2 kh)}$, we can use (3.46) to identify each term (the author recommends going meticulously through these calculations if the reader feels unusually bored—a total of 81 terms should appear). We now arrive at the scheme:

$$
\begin{aligned}
&\frac{1}{16}\left[\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}^{n+1} - 2\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}^{n} + \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}^{n-1}\right] \\
&\qquad = \frac{\Delta t^2}{8h^2}\left[\begin{pmatrix} 1 & & 1 \\ & -4 & \\ 1 & & 1 \end{pmatrix}^{n+1} + 2\begin{pmatrix} 1 & & 1 \\ & -4 & \\ 1 & & 1 \end{pmatrix}^{n} + \begin{pmatrix} 1 & & 1 \\ & -4 & \\ 1 & & 1 \end{pmatrix}^{n-1}\right].
\end{aligned}
\tag{3.49}
$$

The 9-point (and 5-point) computational molecules/stencils represent the spacial discretizations (see *Iserles, 1996*, Section 7.2) while the superscripts represent the time steps. To be more specific, the stencil

$$
\begin{pmatrix} w_1 & w_2 & w_3 \\ w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 \end{pmatrix}^{n},
$$

is short for

$$
\begin{aligned}
&w_1\boldsymbol{u}^n_{j-1,k-1} + w_2\boldsymbol{u}^n_{j-1,k} + w_3\boldsymbol{u}^n_{j-1,k+1} + w_4\boldsymbol{u}^n_{j,k-1} + w_5\boldsymbol{u}^n_{j,k} \\
&\quad + w_6\boldsymbol{u}^n_{j,k+1} + w_7\boldsymbol{u}^n_{j+1,k-1} + w_8\boldsymbol{u}^n_{j+1,k} + w_9\boldsymbol{u}^n_{j+1,k+1}\,.
\end{aligned}
$$

But is the scheme (3.49) a discretization of the two-dimensional wave equation at all? Yes, luckily it is. Consider first the time discretization. We immediately recognize it as the trapezoid method as it is shown in (3.32). For the space discretization we observe that the stencil

$$
\frac{1}{16}\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}
$$

simply is a weighted average that, for sufficiently smooth functions, converge towards the value of the center. Left is the stencil

$$\frac{1}{2h^2}\begin{pmatrix} 1 & & 1 \\ & -4 & \\ 1 & & 1 \end{pmatrix},$$

which, by using Taylor series on sufficiently smooth functions, is seen to approximate $u_{xx} + u_{yy}$ to second order.

An implementation of the two-dimensional scheme (3.49) used for exact controllability, together with numerical results, can be found in Chapter 9.

CHAPTER 4

Boundary Control of Discrete Systems

How to apply general Principles to particular Cases.
— EPICTETUS (50–138)

This chapter will transfer many of the relations and concepts from Chapter 2 into a discrete setting. For instance, inner products, boundary conditions and boundary integrals will be expressed using matrices and sums.

For PDE systems the time available for control can be essential. For discrete systems, the minimal control time can always be arbitrarily short. However, as we discretize more and more accurately, the time available for control can turn out to be important again. The essential condition for correct behavior "in the limit" turns out to be the existence of a uniform constant in the corresponding discrete observability inequalities.

4.1 General Description

This section will go through most of the concepts of Chapter 2, but adapted to a discrete setting. The literature's treatment of discrete control system, especially when considering convergence of controls, has been restricted to specific systems, treating each system seperately. Just as the second chapter presented a new, unified approach to controllability, so will this section.

We will first consider semi-discrete systems and show necessary and sufficient conditions for when such a system is controllable. We then continue to consider fully discrete systems, in particular using the midpoint and trapezoid schemes.

For both semi-discretizations and full discretizations, we will examine important HUM relations and, in turn, how to compute HUM controls.

4.1.1 Semi-discretization

When it comes to convergence of controls, semi-discrete control systems have, by far, gotten the most attention in the literature. Primarily the wave equation has been studied (see, e.g., *Zuazua, 1999*, *Infante and Zuazua, 1999* or *Micu, 2002*), although also the heat equation *(Zuazua, 2003)* and the beam equation *(León and Zuazua, 2002)* have been considered.

We will consider a discrete control system with the following generic appearance,

$$\begin{cases} \boldsymbol{\mathcal{C}}\dot{\boldsymbol{u}}(t) = \boldsymbol{\mathcal{A}}\boldsymbol{u}(t) + \boldsymbol{\mathcal{B}}\boldsymbol{k}(t), \quad 0 \le t \le T\,, \\ \boldsymbol{u}(0) = \boldsymbol{u}^0\,, \end{cases} \tag{4.1}$$

where $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{C}} \in \mathbb{R}^{N\times N}$, $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{N\times m}$, the solution $\boldsymbol{u} \in C([0,T],\mathbb{R}^N)$ and the control $\boldsymbol{k} \in C([0,T],\mathbb{R}^m)$. The matrix $\boldsymbol{\mathcal{C}}$ must be symmetric and positive definite, and although it could be left out, it will prove quite useful. The boundary conditions must be implicitly built into $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$. Note how both the wave equation and the heat equation easily fit into the above formulation.

Analogous with the PDE case, we introduce an adjoint system,

$$\begin{cases} \boldsymbol{\mathcal{C}}\dot{\boldsymbol{v}}(t) = -\widetilde{\boldsymbol{\mathcal{A}}}\boldsymbol{v}(t), \quad 0 \le t \le T\,, \\ \boldsymbol{v}(T) = \boldsymbol{v}^0\,, \end{cases} \tag{4.2}$$

where $\widetilde{\boldsymbol{\mathcal{A}}} \in \mathbb{R}^{N\times N}$ and the solution $\boldsymbol{v} \in C([0,T],\mathbb{R}^N)$. Note that the $\boldsymbol{\mathcal{C}}$-matrix is the same as for the control system.

A duality pairing between solutions of the control and adjoint system is established using the bilinear form $\{\cdot,\cdot\}$,

$$\{\boldsymbol{u},\boldsymbol{v}\} = \langle \boldsymbol{u}, \boldsymbol{M}\boldsymbol{v}\rangle_{\boldsymbol{\mathcal{C}}}, \quad \text{for all } \boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^N\,,$$

where $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ is a regular matrix. The form $\langle\cdot,\cdot\rangle_{\boldsymbol{\mathcal{C}}}$ is a generalized inner product on $\mathbb{R}^N$,

$$\langle \boldsymbol{u},\boldsymbol{v}\rangle_{\boldsymbol{\mathcal{C}}} = \langle \boldsymbol{u}, \boldsymbol{\mathcal{C}}\boldsymbol{v}\rangle_{\mathbb{R}^N} = \boldsymbol{u}^T\boldsymbol{\mathcal{C}}\boldsymbol{v}, \quad \text{for all } \boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^N\,,$$

where $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{N\times N}$ is a symmetric and positive definite matrix. Note that $\boldsymbol{\mathcal{C}}$ must be the same matrix that appears in the control and adjoint system above.

We will furthermore assume that $\boldsymbol{\mathcal{C}}$ and $\boldsymbol{M}$ commute, $\boldsymbol{\mathcal{C}}\boldsymbol{M} = \boldsymbol{M}\boldsymbol{\mathcal{C}}$, which implies

$$\{\boldsymbol{u},\boldsymbol{v}\} = \langle \boldsymbol{M}^T\boldsymbol{u}, \boldsymbol{v}\rangle_{\boldsymbol{\mathcal{C}}} = \langle \boldsymbol{u}, \boldsymbol{M}\boldsymbol{v}\rangle_{\boldsymbol{\mathcal{C}}}\,,$$

just as in the Hilbert space case (compare to the continuous duality pairing (2.3), page 11).

The operators $\boldsymbol{\mathcal{C}}^{-1}\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{C}}^{-1}\widetilde{\boldsymbol{\mathcal{A}}}$ must be adjoint with respect to the duality pairing, $\{\cdot,\cdot\}$, or more precisely,

$$\{\boldsymbol{\mathcal{C}}^{-1}\boldsymbol{\mathcal{A}}\boldsymbol{u}, \boldsymbol{v}\} = \{\boldsymbol{u}, \boldsymbol{\mathcal{C}}^{-1}\widetilde{\boldsymbol{\mathcal{A}}}\boldsymbol{v}\}, \quad \text{for all } \boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^N\,. \tag{4.3}$$

Using the definitions above we see that the relation between $\boldsymbol{\mathcal{A}}$ and $\widetilde{\boldsymbol{\mathcal{A}}}$ must be

$$\boldsymbol{\mathcal{C}}^{-1}\boldsymbol{\mathcal{A}}^T\boldsymbol{M} = \boldsymbol{M}\boldsymbol{\mathcal{C}}^{-1}\widetilde{\boldsymbol{\mathcal{A}}} \quad \Leftrightarrow \quad \boldsymbol{\mathcal{A}}^T\boldsymbol{M} = \boldsymbol{M}\widetilde{\boldsymbol{\mathcal{A}}} \tag{4.4}$$

(the equivalence holds because $\boldsymbol{\mathcal{C}}$ and $\boldsymbol{M}$ commute).

The norms with which we measure the size of solutions is not so important on a fixed discretization level, since all norms are equivalent in finite dimensional spaces. However, since we are interested in letting the discretization level $N \to \infty$, the norms are very important. They must correspond to the continuous norms in an appropriate way.

We let solutions of the adjoint system be measured by the norm

$$\|\boldsymbol{v}\|^2_{\widetilde{\boldsymbol{Q}}} = \langle \boldsymbol{v}, \boldsymbol{v}\rangle_{\widetilde{\boldsymbol{Q}}}, \quad \langle \boldsymbol{v}, \overline{\boldsymbol{v}}\rangle_{\widetilde{\boldsymbol{Q}}} = \boldsymbol{v}^T\widetilde{\boldsymbol{Q}}\overline{\boldsymbol{v}}\,,$$

where $\widetilde{\boldsymbol{Q}}$ is symmetric and positive definite. This norm corresponds to the norm of the Hilbert space $\widetilde{H}$ of Chapter 2. We deduce the norm of the corresponding dual space by setting $\widetilde{\boldsymbol{Q}} = \boldsymbol{R}^T\boldsymbol{R}$ and then we get

$$\begin{aligned}
\|\boldsymbol{u}\|^2_{\widetilde{\boldsymbol{Q}}'} &= \sup_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{\left|\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\boldsymbol{\mathcal{C}}}\right|^2}{\|\boldsymbol{v}\|^2_{\widetilde{\boldsymbol{Q}}}} = \sup_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{\left(\boldsymbol{u}^T\boldsymbol{\mathcal{C}}\boldsymbol{v}\right)^2}{\boldsymbol{v}^T\widetilde{\boldsymbol{Q}}\boldsymbol{v}} = \sup_{\boldsymbol{w}\neq\boldsymbol{0}} \frac{\left(\boldsymbol{u}^T\boldsymbol{\mathcal{C}}\boldsymbol{R}^{-1}\boldsymbol{w}\right)^2}{\boldsymbol{w}^T\boldsymbol{w}} \\
&= \frac{\left(\boldsymbol{u}^T\boldsymbol{\mathcal{C}}\boldsymbol{R}^{-1}\boldsymbol{R}^{-T}\boldsymbol{\mathcal{C}}\boldsymbol{u}\right)^2}{\boldsymbol{u}^T\boldsymbol{\mathcal{C}}\boldsymbol{R}^{-1}\boldsymbol{R}^{-T}\boldsymbol{\mathcal{C}}\boldsymbol{u}} = \boldsymbol{u}^T\boldsymbol{\mathcal{C}}\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\mathcal{C}}\boldsymbol{u}\,,
\end{aligned}$$

showing that $\widetilde{\boldsymbol{Q}}' = \boldsymbol{\mathcal{C}}\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\mathcal{C}}$. To obtain the "no-tilde norms" we use the relation (2.2) from Chapter 2, and the comments that follow, such that

$$\|\boldsymbol{v}\|_{\widetilde{\boldsymbol{Q}}} = \|\boldsymbol{M}\boldsymbol{v}\|_{\boldsymbol{Q}} \quad \text{and} \quad \|\boldsymbol{v}\|_{\widetilde{\boldsymbol{Q}}'} = \|\boldsymbol{M}^{-T}\boldsymbol{v}\|_{\boldsymbol{Q}'}\,, \tag{4.5}$$

implying that

$$\boldsymbol{Q} = \boldsymbol{M}^{-T}\widetilde{\boldsymbol{Q}}\boldsymbol{M}^{-1} \quad \text{and} \quad \boldsymbol{Q}' = \boldsymbol{M}\boldsymbol{\mathcal{C}}\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\mathcal{C}}\boldsymbol{M}^T\,. \tag{4.6}$$

One of the most essential relations of Chapter 2 was (2.8) on page 12. The following relation is the semi-discrete equivalent of that,

$$\begin{aligned}
\big[\{\boldsymbol{u}, \boldsymbol{v}\}\big]_0^T &= \int_0^T \big(\{\dot{\boldsymbol{u}}, \boldsymbol{v}\} + \{\boldsymbol{u}, \dot{\boldsymbol{v}}\}\big)dt \\
&= \int_0^T \big(\{\boldsymbol{\mathcal{C}}^{-1}(\boldsymbol{\mathcal{A}}\boldsymbol{u} + \boldsymbol{\mathcal{B}}\boldsymbol{k}), \boldsymbol{v}\} - \{\boldsymbol{u}, \boldsymbol{\mathcal{C}}^{-1}\widetilde{\boldsymbol{\mathcal{A}}}\boldsymbol{v}\}\big)dt \\
&= \int_0^T \{\boldsymbol{\mathcal{C}}^{-1}\boldsymbol{\mathcal{B}}\boldsymbol{k}, \boldsymbol{v}\}dt = \int_0^T \langle \boldsymbol{\mathcal{B}}\boldsymbol{k}, \boldsymbol{M}\boldsymbol{v}\rangle_{\mathbb{R}^N}\,dt = \int_0^T \langle \boldsymbol{k}, \boldsymbol{\mathcal{B}}^T\boldsymbol{M}\boldsymbol{v}\rangle_{\mathbb{R}^m}\,dt\,,
\end{aligned} \tag{4.7}$$

valid for all solutions $\boldsymbol{u}(t)$ and $\boldsymbol{v}(t)$ of the systems (4.1) and (4.2) respectively.

Analogous with Section 2.1.3 we will also introduce two mappings $\boldsymbol{L}_T^h$ and $\boldsymbol{L}_T^{h,*}$. If we apply no control to the control system (4.1) we set

$$\boldsymbol{L}_T^h \boldsymbol{u}^0 = \boldsymbol{u}(T)\,,$$

where $\boldsymbol{u}(t)$ is a solution to (4.1) with initial condition $\boldsymbol{u}(0) = \boldsymbol{u}^0$ and control $\boldsymbol{k}(t) = \boldsymbol{0}$. Likewise for the adjoint system,

$$\boldsymbol{L}_T^{h,*} \boldsymbol{v}^0 = \boldsymbol{v}(0)\,,$$

where $\boldsymbol{v}(t)$ is a solution to (4.2) with initial condition $\boldsymbol{v}(T) = \boldsymbol{v}^0$. As in Chapter 2, the two mappings are adjoint in the sense

$$\{\boldsymbol{L}_T^h \boldsymbol{u}^0, \boldsymbol{v}^0\} = \{\boldsymbol{u}^0, \boldsymbol{L}_T^{h,*} \boldsymbol{v}^0\}\,,$$

for all $\boldsymbol{u}^0, \boldsymbol{v}^0 \in \mathbb{R}^N$, a consequence of relation (4.7).

We now wish to find the HUM controllability operator for the semi-discrete case. Consider the systems (4.1) and (4.2) and by setting $\boldsymbol{u}(0) = \boldsymbol{u}^0 = \boldsymbol{0}$ and $\boldsymbol{k}(t) = \boldsymbol{\mathcal{B}}^T \boldsymbol{M}\boldsymbol{v}(t)$, we can define $\boldsymbol{\Lambda}_T^h \boldsymbol{v}^0 = \boldsymbol{M}^T \boldsymbol{u}(T)$. This corresponds to the definition $\Lambda_T = G_T^* G_T$ for the general systems of Chapter 2. If we now let $\boldsymbol{w}$ be a solution to the adjoint system (4.2) with initial condition $\boldsymbol{w}^0$ we get from (4.7):

$$\langle \boldsymbol{\Lambda}_T^h \boldsymbol{v}^0, \boldsymbol{w}^0\rangle_{\boldsymbol{\mathcal{C}}} = \langle \boldsymbol{u}(T), \boldsymbol{M}\boldsymbol{w}^0\rangle_{\boldsymbol{\mathcal{C}}} = \big[\{\boldsymbol{u}, \boldsymbol{w}\}\big]_0^T = \int_0^T \langle \boldsymbol{\mathcal{B}}^T \boldsymbol{M}\boldsymbol{v}, \boldsymbol{\mathcal{B}}^T \boldsymbol{M}\boldsymbol{w}\rangle_{\mathbb{R}^m}\, dt\,. \tag{4.8}$$

This clearly shows that $\boldsymbol{\Lambda}_T^h$ is symmetric with respect to the inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{\mathcal{C}}}$ and is positive semi-definite.

As in the continuous case, we introduce

$$\gamma_T(\boldsymbol{u}^0, \boldsymbol{w}^0) = \langle \boldsymbol{\Lambda}_T^h \boldsymbol{u}^0, \boldsymbol{w}^0\rangle_{\boldsymbol{\mathcal{C}}} = \int_0^T \langle \boldsymbol{\mathcal{B}}^T \boldsymbol{M}\boldsymbol{v}, \boldsymbol{\mathcal{B}}^T \boldsymbol{M}\boldsymbol{w}\rangle_{\mathbb{R}^m}\, dt\,.$$

Naturally, this bilinear form is symmetric and positive semi-definite. We now have the first result concerning controllability of a semi-discrete system.

Theorem 4.1.1. *The semi-discrete control system (4.1) is controllable at time T if and only if γ_T is positive definite, that is, if and only if*

$$0 < \gamma_T(\boldsymbol{v}^0, \boldsymbol{v}^0) = \int_0^T \big\|\boldsymbol{\mathcal{B}}^T \boldsymbol{M}\boldsymbol{v}(t)\big\|_{\mathbb{R}^m}^2 dt, \quad \textit{for all } \boldsymbol{v}^0 \in \mathbb{R}^N \setminus \{0\}\,, \tag{4.9}$$

where $\boldsymbol{v}(t)$ is the solution of the adjoint semi-discrete system (4.2) with initial condition $\boldsymbol{v}^0$.

Proof. Assume that (4.9) holds for some T. Let $\boldsymbol{u}^0, \boldsymbol{u}^1 \in \mathbb{R}^N$ be given. We will now show how to find a control $\boldsymbol{k}$ that drives the control system from state $\boldsymbol{u}^0$ to the state $\boldsymbol{u}^1$.

We start by solving the linear equation

$$\boldsymbol{\Lambda}_T^h \boldsymbol{w}^0 = \boldsymbol{M}^T(\boldsymbol{u}^1 - \boldsymbol{L}_T^h \boldsymbol{u}^0) \tag{4.10}$$

for $\boldsymbol{w}^0$. This is possible since the assumed positivity of γ_T implies that $\boldsymbol{\Lambda}_T^h$ is positive definite and thus invertible. The above relation leads to

$$\begin{aligned}\langle \boldsymbol{\Lambda}_T^h \boldsymbol{w}^0, \boldsymbol{v}^0\rangle_{\boldsymbol{\mathcal{C}}} = \langle \boldsymbol{M}^T(\boldsymbol{u}^1 - \boldsymbol{L}_T^h \boldsymbol{u}^0), \boldsymbol{v}^0\rangle_{\boldsymbol{\mathcal{C}}} \qquad \Leftrightarrow \\ \gamma_T(\boldsymbol{w}^0, \boldsymbol{v}^0) = \{\boldsymbol{u}^1 - \boldsymbol{L}_T^h \boldsymbol{u}^0, \boldsymbol{v}^0\} = \{\boldsymbol{u}^1, \boldsymbol{v}^0\} - \{\boldsymbol{u}^0, \boldsymbol{L}_T^{h,*} \boldsymbol{v}^0\}\end{aligned}$$

for all $\boldsymbol{v}^0 \in \mathbb{R}^N$. This last expression shows, cf. (4.7), that the control $\boldsymbol{k}(t) = \boldsymbol{\mathcal{B}}^T \boldsymbol{M}\boldsymbol{w}(t)$, where $\boldsymbol{w}(t)$ is a solution to the adjoint system with initial condition $\boldsymbol{w}^0$, drives the control system from $\boldsymbol{u}^0$ to $\boldsymbol{u}^1$. Note that we have thus found a HUM control.

Consider now the case where a linear operator $K_T : \mathbb{R}^N \mapsto L^2((0,T);\mathbb{R}^m)$ exists that can drive the control system from $\mathbf{0}$ to any final state $\boldsymbol{u}$ (this is sufficient to consider, see the comment in the beginning of Section 2.4). From relation (4.7) we now have

$$\int_0^T \langle K_T\boldsymbol{u}, \boldsymbol{\mathcal{B}}^T\boldsymbol{M}\boldsymbol{v}\rangle_{\mathbb{R}^m}\, dt = \{\boldsymbol{u}, \boldsymbol{v}^0\}, \quad \text{for all } \boldsymbol{v}^0 \in \mathbb{R}^N$$

We get for an arbitrary $\boldsymbol{v}^0 \in \mathbb{R}^N$,

$$\begin{aligned}\|\boldsymbol{M}\boldsymbol{v}^0\| &= \max_{\boldsymbol{u}\in\mathbb{R}^N\setminus\{\mathbf{0}\}} \frac{\big|\langle \boldsymbol{u}, \boldsymbol{M}\boldsymbol{v}^0\rangle\big|}{\|\boldsymbol{u}\|} = \max_{\boldsymbol{u}\in\mathbb{R}^N\setminus\{\mathbf{0}\}} \frac{\big|\langle K_T\boldsymbol{u}, \boldsymbol{\mathcal{B}}^T\boldsymbol{M}\boldsymbol{v}\rangle_{L^2((0,T),\mathbb{R}^m)}\big|}{\|\boldsymbol{u}\|} \\ &\le \max_{\boldsymbol{u}\in\mathbb{R}^N\setminus\{\mathbf{0}\}} \frac{\|K_T\boldsymbol{u}\|_{L^2((0,T),\mathbb{R}^m)}}{\|\boldsymbol{u}\|} \|\boldsymbol{\mathcal{B}}^T\boldsymbol{M}\boldsymbol{v}\|_{L^2((0,T),\mathbb{R}^m)} \\ &= \|K_T\|\gamma_T(\boldsymbol{v}^0,\boldsymbol{v}^0)^{1/2},\end{aligned}$$

which implies statement (4.9). □

Note how the condition $0 < \gamma_T(\boldsymbol{v}^0, \boldsymbol{v}^0)$ of the theorem corresponds to the unique continuation property mentioned in Section 2.1.3. In the continuous setting this property only implied approximate controllability, whereas it here implies exact controllability. This is a consequence of the finite space dimension.

Let us now introduce the matrix

$$\boldsymbol{R} = \begin{bmatrix}\boldsymbol{\mathcal{B}} & \boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{B}} & \dots & \boldsymbol{\mathcal{A}}^{N-1}\boldsymbol{\mathcal{B}}\end{bmatrix} \in \mathbb{R}^{N\times Nm}.$$

Some of the results that follow, concerning controllability in relation to properties of $\boldsymbol{R}$ or eigenvalue properties of $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$, are classical results, see, for instance, *Kalman (1963)*, *Russell (1978)* or *Sontag (1990)*. First we will show that the traditional rank condition of $\boldsymbol{R}$ is equivalent to the condition of Theorem 4.1.1. Note how this theorem ties together a classical result with our new approach.

Theorem 4.1.2. *The controllability condition of Theorem 4.1.1,*

$$0 < \gamma_T(\boldsymbol{v}^0, \boldsymbol{v}^0), \quad \textit{for all } \boldsymbol{v}^0 \in \mathbb{R}^N \setminus \{0\}\,,$$

for some $T > 0$, is equivalent to the condition

$$\operatorname{rank} \boldsymbol{R} = N\,.$$

Proof. We will show that the *negation* of the statements are equivalent. Assume therefore that a $\boldsymbol{v}^0 \neq 0$ exists such that

$$\int_0^T \left\|\boldsymbol{\mathcal{B}}^T \boldsymbol{M}\boldsymbol{v}(t)\right\|^2_{\mathbb{R}^m} dt = 0 \tag{4.11}$$

By inserting that $\boldsymbol{v}(t) = e^{\widetilde{\boldsymbol{\mathcal{A}}}(T-t)}\boldsymbol{v}^0$ we get

$$\left\|\boldsymbol{\mathcal{B}}^T \boldsymbol{M} e^{\widetilde{\boldsymbol{\mathcal{A}}}(T-t)}\boldsymbol{v}^0\right\|_{\mathbb{R}^m} = 0, \qquad 0 \le t \le T\,.$$

Using the relation between $\boldsymbol{\mathcal{A}}$ and $\widetilde{\boldsymbol{\mathcal{A}}}$ from (4.4) and the Taylor expansion of e^x we get

$$\left\|\boldsymbol{\mathcal{B}}^T e^{\boldsymbol{\mathcal{A}}^T t}\boldsymbol{M}\boldsymbol{v}^0\right\|_{\mathbb{R}^m} = 0, \qquad 0 \le t \le T\,.$$

Setting $\boldsymbol{z} = \boldsymbol{M}\boldsymbol{v}^0$ we see that $\boldsymbol{z} \neq 0$ and

$$\boldsymbol{z}^T e^{\boldsymbol{\mathcal{A}}t}\boldsymbol{\mathcal{B}} = \boldsymbol{0}\,, \tag{4.12}$$

for $0 \le t \le T$. The analyticity of the exponential implies now

$$\boldsymbol{z}^T \boldsymbol{\mathcal{A}}^k \boldsymbol{\mathcal{B}} = \boldsymbol{0}, \quad \text{for integer } k \ge 0\,, \tag{4.13}$$

which clearly implies

$$\operatorname{rank} \boldsymbol{R} < N\,. \tag{4.14}$$

Let us now see that these implications can be reversed. Assume therefore the rank condition (4.14) above holds. This means that a $\boldsymbol{z} \neq 0$ exists such that

$$\boldsymbol{z}^T \boldsymbol{A}^k \boldsymbol{\mathcal{B}} = \boldsymbol{0}, \quad \text{for } k = 0, 1, \ldots, N-1\,. \tag{4.15}$$

Since $\boldsymbol{\mathcal{A}}$ is a root in its own characteristic polynomial (the Cayley–Hamilton theorem), it is seen that $\boldsymbol{\mathcal{A}}^n$ can be written as a linear combination of $\boldsymbol{I}$, $\boldsymbol{\mathcal{A}}$, …, $\boldsymbol{\mathcal{A}}^{N-1}$. Using this result, one can easily show, by induction on k, that (4.15) implies (4.13). This implies, in turn, that (4.12) holds for *all* $t \in \mathbb{R}$, and in particular for $0 \le t \le T$. By choosing $\boldsymbol{v}^0 = \boldsymbol{M}^{-1}\boldsymbol{z}$ we easily arrive at the wanted equality (4.11). ☐

Testing the rank condition of $\boldsymbol{R}$ is fairly simple in the case where $\boldsymbol{\mathcal{A}}$ is diagonalizable. This is the subject of the following, well-known, theorem.

Theorem 4.1.3. *Let a control system with $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N\times N}$ and $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{N\times m}$ be given. Let furthermore $\boldsymbol{\mathcal{A}}$ be diagonalizable, $\boldsymbol{\mathcal{A}} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1}$, such that*

$$\boldsymbol{D} = \begin{pmatrix} \lambda_1 \boldsymbol{I}_{p_1} & & & \\ & \lambda_2 \boldsymbol{I}_{p_2} & & \\ & & \ddots & \\ & & & \lambda_d \boldsymbol{I}_{p_d} \end{pmatrix}, \qquad \boldsymbol{V}^{-1}\boldsymbol{\mathcal{B}} = \begin{pmatrix} \boldsymbol{\mathcal{B}}_1 \\ \boldsymbol{\mathcal{B}}_2 \\ \vdots \\ \boldsymbol{\mathcal{B}}_d \end{pmatrix},$$

where $\boldsymbol{I}_p$ is an identity matrix of size $p \times p$, the quantity d is the number of distinct eigenvalues, $p_1 + p_2 + \cdots + p_d = N$, and the structure of $\boldsymbol{V}^{-1}\boldsymbol{\mathcal{B}}$ corresponds to that of $\boldsymbol{D}$.

The semi-discrete control system (4.1) is now controllable if and only if the rows of each $\boldsymbol{\mathcal{B}}_k$ are linearly independent for $k = 1, 2, \ldots, d$.

Proof. Observe from Theorem 4.1.2 that the control system is *not* controllable if and only if a $\boldsymbol{z} \neq 0$ exists such that

$$\boldsymbol{z}^T \boldsymbol{\mathcal{A}}^k \boldsymbol{\mathcal{B}} = 0, \quad \text{for } k = 0, 1, \ldots, N-1\,.$$

Using the eigenvalue decomposition, we get the equivalent

$$(\boldsymbol{V}^T \boldsymbol{z})^T \boldsymbol{D}^k (\boldsymbol{V}^{-1}\boldsymbol{\mathcal{B}}) = \boldsymbol{0}, \quad \text{for } k = 0, 1, \ldots, N-1\,. \tag{4.16}$$

We set $(\boldsymbol{V}^T \boldsymbol{z})^T = [\boldsymbol{w}_1^T \ \boldsymbol{w}_2^T \ \cdots \ \boldsymbol{w}_d^T]$, corresponding to the structure of $\boldsymbol{D}$, and see that (4.16) is equivalent to

$$\lambda_1^k \boldsymbol{w}_1^T \boldsymbol{\mathcal{B}}_1 + \lambda_2^k \boldsymbol{w}_2^T \boldsymbol{\mathcal{B}}_2 + \cdots + \lambda_d^k \boldsymbol{w}_d^T \boldsymbol{\mathcal{B}}_d = \boldsymbol{0}, \quad \text{for } k = 0, 1, \ldots, N-1\,,$$

which, in turn, can be written

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_d \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{N-1} & \lambda_2^{N-1} & \cdots & \lambda_d^{N-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_1^T \boldsymbol{\mathcal{B}}_1 \\ \boldsymbol{w}_2^T \boldsymbol{\mathcal{B}}_2 \\ \vdots \\ \boldsymbol{w}_d^T \boldsymbol{\mathcal{B}}_d \end{pmatrix} = \boldsymbol{0}\,.$$

The left-hand Vandermonde matrix is regular since the numbers $\lambda_1, \lambda_2, \ldots, \lambda_d$ are mutually distinct (see *Golub and Van Loan, 1996*, page 184) and the above equation is thus satisfied if and only if

$$\boldsymbol{w}_1^T \boldsymbol{\mathcal{B}}_1 = \boldsymbol{w}_2^T \boldsymbol{\mathcal{B}}_2 = \cdots = \boldsymbol{w}_d^T \boldsymbol{\mathcal{B}}_d = \boldsymbol{0}\,.$$

Since $\boldsymbol{V}^T \boldsymbol{z} = [\boldsymbol{w}_1^T \ \boldsymbol{w}_2^T \ \cdots \ \boldsymbol{w}_d^T]^T$ where $\boldsymbol{V}$ is regular, the result follows. □

The case where $\boldsymbol{\mathcal{B}}$ has only one column deserves special attention. Theorem 4.1.3 reduces in this case to the following.

Corollary 4.1.1. *Let $\boldsymbol{\mathcal{A}}$ be diagonalizable and $\boldsymbol{\mathcal{B}}$ a column vector. The semi-discrete control system (4.1) is controllable if and only if $\boldsymbol{\mathcal{A}}$ has no multiple eigenvalues and the vector $\boldsymbol{V}^{-1}\boldsymbol{\mathcal{B}}$ contains no zeroes.*

The controllability condition of Theorem 4.1.3 is quite interesting. It states that if $\boldsymbol{\mathcal{A}}$ has an eigenvalue with multiplicity *larger* than the number of columns in $\boldsymbol{\mathcal{B}}$, then controllability is impossible, *no matter how* you choose the entries of the control matrix $\boldsymbol{\mathcal{B}}$.

Another way controllability will fail is if $\boldsymbol{V}^{-1}\boldsymbol{\mathcal{B}}$ contains a zero-valued row. Note that this can always happen, as long as $\boldsymbol{\mathcal{A}}$ is diagonalizable. The reason for the failure becomes clear if we premultiply the control system with $\boldsymbol{V}^{-1}$ and obtain

$$\boldsymbol{V}^{-1}\dot{\boldsymbol{u}}(t) = \boldsymbol{V}^{-1}\boldsymbol{\mathcal{A}}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{u}(t) + \boldsymbol{V}^{-1}\boldsymbol{\mathcal{B}}\boldsymbol{k}(t)\,.$$

Setting now $\boldsymbol{\tau}(t) = \boldsymbol{V}^{-1}\boldsymbol{u}(t)$, we see the control system in an eigenvector basis,

$$\dot{\boldsymbol{\tau}}(t) = \boldsymbol{D}\boldsymbol{\tau}(t) + (\boldsymbol{V}^{-1}\boldsymbol{\mathcal{B}})\boldsymbol{k}(t)\,.$$

All rows of this ODE have been decoupled and it is now clear that if the matrix $\boldsymbol{V}^{-1}\boldsymbol{\mathcal{B}}$ contains a zero-valued row, then it is impossible to control the corresponding element of $\boldsymbol{\tau}(t)$, and thereby, the corresponding eigenmode.

4.1.1.1 HUM for Hyperbolic Semi-Discrete Systems

We consider the case of a hyperbolic control system,

$$\begin{cases} \boldsymbol{C}\ddot{\boldsymbol{u}}(t) = \boldsymbol{A}\boldsymbol{u}(t) + \boldsymbol{B}\boldsymbol{k}(t), \quad 0 \le t \le T\,, \\ \boldsymbol{u}(0) = \boldsymbol{u}^0, \quad \dot{\boldsymbol{u}}(0) = \overline{\boldsymbol{u}}^0\,, \end{cases} \tag{4.17}$$

where $\boldsymbol{C}$ and $\boldsymbol{A}$ are order N matrices that are symmetric, and positive and negative definite, respectively. The adjoint system is of the form

$$\begin{cases} \boldsymbol{C}\ddot{\boldsymbol{v}}(t) = \boldsymbol{A}\boldsymbol{v}(t), \quad 0 \le t \le T\,, \\ \boldsymbol{v}(T) = \boldsymbol{v}^0, \quad \boldsymbol{v}(T) = \overline{\boldsymbol{v}}^0\,. \end{cases} \tag{4.18}$$

We can easily apply the general results obtained in Section 4.1.1, when we observe that the above systems are equivalent to the first order systems (4.1) and (4.2) with

$$\boldsymbol{\mathcal{A}} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{C} \\ \boldsymbol{A} & \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{\mathcal{C}} = \begin{bmatrix} \boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix}, \quad \widetilde{\boldsymbol{\mathcal{A}}} = \begin{bmatrix} \boldsymbol{0} & -\boldsymbol{C} \\ -\boldsymbol{A} & \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{\mathcal{B}} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{B} \end{bmatrix}. \tag{4.19}$$

All we need to determine is the matrix $\boldsymbol{M}$ such that $\boldsymbol{\mathcal{C}}^{-1}\boldsymbol{\mathcal{A}}^T\boldsymbol{M} = \boldsymbol{M}\boldsymbol{\mathcal{C}}^{-1}\widetilde{\boldsymbol{\mathcal{A}}}$. This is clearly fulfilled with

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{0} & -\boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}. \tag{4.20}$$

We can now easily formulate relation (4.7) for this case,

$$\big[\{(\boldsymbol{u}, \overline{\boldsymbol{u}}), (\boldsymbol{v}, \overline{\boldsymbol{v}})\}\big]_0^T = \int_0^T \langle \boldsymbol{k}, \boldsymbol{B}^T\boldsymbol{v}\rangle_{\mathbb{R}^m}\, dt\,,$$

since

$$\boldsymbol{\mathcal{B}}^T \boldsymbol{M} \begin{bmatrix} \boldsymbol{v} \\ \overline{\boldsymbol{v}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{B}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{0} & -\boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{v} \\ \overline{\boldsymbol{v}} \end{bmatrix} = \boldsymbol{B}^T \boldsymbol{v} \,.$$

The controllability operator is easily defined using the same procedure as for the general first order system. When we solve the adjoint system (4.18) followed by solving the control system (4.17) with $(\boldsymbol{u}^0, \overline{\boldsymbol{u}}^0) = (\boldsymbol{0}, \boldsymbol{0})$ and $\boldsymbol{k} = \boldsymbol{B}^T \boldsymbol{v}$, we can define $\boldsymbol{\Lambda}_T^h$ as

$$\boldsymbol{\Lambda}_T^h \begin{pmatrix} \boldsymbol{v}^0 \\ \overline{\boldsymbol{v}}^0 \end{pmatrix} = \boldsymbol{M}^T \begin{pmatrix} \boldsymbol{u}(T) \\ \overline{\boldsymbol{u}}(T) \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{u}}(T) \\ -\boldsymbol{u}(T) \end{pmatrix} ,$$

leading to the relation

$$\begin{aligned} \left\langle \boldsymbol{\Lambda}_T^h(\boldsymbol{v}^0, \overline{\boldsymbol{v}}^0), (\boldsymbol{w}^0, \overline{\boldsymbol{w}}^0) \right\rangle_{\boldsymbol{\mathcal{C}}} &= \left\langle \boldsymbol{M}^T (\boldsymbol{u}(T), \overline{\boldsymbol{u}}(T))^T, (\boldsymbol{w}^0, \overline{\boldsymbol{w}}^0)^T \right\rangle_{\boldsymbol{\mathcal{C}}} \\ &= \Big[\{(\boldsymbol{u}, \overline{\boldsymbol{u}}), (\boldsymbol{w}, \overline{\boldsymbol{w}})\} \Big]_0^T = \int_0^T \langle \boldsymbol{B}^T \boldsymbol{v}, \boldsymbol{B}^T \boldsymbol{w} \rangle_{\mathbb{R}^m} \, dt \,. \end{aligned}$$

Note how our general approach reveals, that what we in Chapter 2 called the complementary boundary operator, *must have the form* $\boldsymbol{B}^T$. So given an actual continuous control system with corresponding semi-discretization, one should make sure that $\boldsymbol{B}^T$ is a consistent discretization of the complementary boundary operator $\mathcal{C}$.

4.1.2 Full Discretization

We now turn our attention to fully discrete control systems. It is interesting to note that when it comes to choosing time discretization schemes for controllability problems, the explicit midpoint rule has almost exclusively been used in the literature. The papers *Glowinski, Li, and Lions (1990)*, *Glowinski (1992b)*, *Glowinski and Lions (1995)*, *Asch and Lebeau (1998)* and *Negreanu and Zuazua (2003)* are examples of this. See *Eljendy (1992)* for a so-called discrete-time Galerkin approximation, which, as far as the author knows, is the only paper that does *not* use the explicit midpoint rule for time discretization (it is not clear, though, exactly why that scheme was chosen).

We will initially consider time discretization in great generality. When we study some schemes in more detail, however, we will consider only the explicit midpoint rule and the trapezoid rule.

Let us first consider a fully discrete control system of the following general form,

$$\boldsymbol{u}^{n+1} = \boldsymbol{G}\boldsymbol{u}^n + \boldsymbol{F}\boldsymbol{k}^n \,, \tag{4.21}$$

where $\boldsymbol{u}^0 \in \mathbb{R}^N$ is given and $\boldsymbol{G} \in \mathbb{R}^{N \times N}$ and $\boldsymbol{F} \in \mathbb{R}^{N \times m}$. The nth iterate can be written as

$$\boldsymbol{u}^n = \boldsymbol{G}^n \boldsymbol{u}^0 + \sum_{k=0}^{n-1} \boldsymbol{G}^k \boldsymbol{F} \boldsymbol{k}^{n-1-k}, \quad n \geq 0 \,, \tag{4.22}$$

which is easily seen by induction. Let us introduce the notation

$$\boldsymbol{R}_k = \begin{bmatrix}\boldsymbol{F} & \boldsymbol{GF} & \cdots & \boldsymbol{G}^{k-1}\boldsymbol{F}\end{bmatrix} \in \mathbb{R}^{N\times km} .$$

Observe that we will always have rank $\boldsymbol{R}_k$ = rank $\boldsymbol{R}_N$ for $k \geq N$ since powers $\boldsymbol{G}^k$ with $k \geq N$ can always be expressed as a linear combination of $\boldsymbol{I}$, $\boldsymbol{G}$, …, $\boldsymbol{G}^{N-1}$. This follows from the fact that any matrix is a root in its own characteristic polynomial.

We now have the following result concerning controllability of a fully discrete system.

Theorem 4.1.4. *The (fully) discrete control system (4.21) is controllable at iteration n if and only if* rank $\boldsymbol{R}_N = N$ *and $n \geq n_0$, where n_0 is the smallest integer such that* rank $\boldsymbol{R}_{n_0} = N$.

Proof. Consider the following rewrite of (4.22):

$$\sum_{k=0}^{n-1} \boldsymbol{G}^k \boldsymbol{F}\boldsymbol{k}^{n-1-k} = \boldsymbol{R}_n \begin{bmatrix}\boldsymbol{k}^{n-1}\\ \vdots \\ \boldsymbol{k}^0\end{bmatrix} = \boldsymbol{u}^n - \boldsymbol{G}^n\boldsymbol{u}^0 .$$

From this we see that the control system is controllable at iteration n if and only if rank$\boldsymbol{R}_n = N$. It now follows from the definition of n_0 and the fact that rank$\boldsymbol{R}_{n_1} \leq$ rank $\boldsymbol{R}_{n_2}$ for $n_1 \leq n_2$, that rank $\boldsymbol{R}_n = N$ for $n \geq n_0$. □

Assume that rank $\boldsymbol{R}_N = N$. Now in order for $\boldsymbol{R}_{n_0}$ to have rank N, it must have at least N columns, so $m \cdot n_0 \geq N$, where m is the number of columns of $\boldsymbol{F}$. This implies

$$\left\lceil \frac{N}{m} \right\rceil \leq n_0 \leq N .$$

Note how $\boldsymbol{R}_N$ is analogous to the matrix $\boldsymbol{R}$ of the previous section. This means that Theorem 4.1.3 and Corollary 4.1.1 can be used for showing the rank condition rank $\boldsymbol{R}_N = N$.

We now introduce the adjoint system,

$$\boldsymbol{v}^{n-1} = \widetilde{\boldsymbol{G}}\boldsymbol{v}^n , \tag{4.23}$$

where $\boldsymbol{v}^M \in \mathbb{R}^N$ is given. The matrix $\widetilde{\boldsymbol{G}}$ is the dual of $\boldsymbol{G}$ with respect to the duality pairing $\{\cdot,\cdot\}$, that is, $\{\boldsymbol{Gu},\boldsymbol{v}\} = \{\boldsymbol{u},\widetilde{\boldsymbol{G}}\boldsymbol{v}\}$ for all $\boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^N$. We can now derive a relation similar to (2.7) for PDEs and similar to (4.7) for semi-discrete systems,

$$\begin{aligned}
\{\boldsymbol{u}^M,\boldsymbol{v}^M\} - \{\boldsymbol{u}^0,\boldsymbol{v}^0\} &= \sum_{n=0}^{M-1}\big(\{\boldsymbol{u}^{n+1},\boldsymbol{v}^{n+1}\} - \{\boldsymbol{u}^n,\boldsymbol{v}^n\}\big)\\
&= \sum_{n=0}^{M-1}\big(\{\boldsymbol{Gu}^n + \boldsymbol{Fk}^n,\boldsymbol{v}^{n+1}\} - \{\boldsymbol{u}^n,\boldsymbol{G}^*\boldsymbol{v}^{n+1}\}\big)\\
&= \sum_{n=0}^{M-1}\{\boldsymbol{Fk}^n,\boldsymbol{v}^{n+1}\} = \sum_{n=0}^{M-1}\langle\boldsymbol{k}^n,\boldsymbol{F}^T\boldsymbol{\mathcal{C}}\boldsymbol{M}\boldsymbol{v}^{n+1}\rangle ,
\end{aligned}$$

which holds for all solution of (4.21) and (4.23). So a HUM control must be of the form $\boldsymbol{k}^n = \boldsymbol{F}^T \boldsymbol{\mathcal{C}} \boldsymbol{M} \boldsymbol{v}^{n+1}$ for $n = 0, 1, \dots, M-1$.

It would be possible, for a given time discretization, to compute $\boldsymbol{G}$, $\boldsymbol{F}$ and $\widetilde{\boldsymbol{G}}$, and then to use the relation above to derive the observability inequality. In practice, however, it is much easier to consider each time discretization separately. The next two sections will show how, for the midpoint and trapezoid rules.

4.1.2.1 The Midpoint Rule

Recall from Section 3.2.2 that the midpoint rule only makes sense for hyperbolic systems (a necessary condition for the stability of this scheme was that the first order system matrix must have purely imaginary eigenvalues). The control system is thus of the form

$$\left\{ \begin{aligned} \boldsymbol{C}\frac{\boldsymbol{u}^{n+1} - 2\boldsymbol{u}^n + \boldsymbol{u}^{n-1}}{\Delta t^2} &= \boldsymbol{A}\boldsymbol{u}^n + \boldsymbol{B}\boldsymbol{k}^n \,, \\ \boldsymbol{u}^0 \text{ given,} \quad \frac{\boldsymbol{u}^1 - \boldsymbol{u}^{-1}}{2\Delta t} &= \overline{\boldsymbol{u}}^0 \,. \end{aligned} \right.$$

Similarly, the adjoint system becomes

$$\left\{ \begin{aligned} \boldsymbol{C}\frac{\boldsymbol{v}^{n+1} - 2\boldsymbol{v}^n + \boldsymbol{v}^{n-1}}{\Delta t^2} &= \boldsymbol{A}\boldsymbol{v}^n \,, \\ \boldsymbol{v}^M \text{ given,} \quad \frac{\boldsymbol{v}^{M+1} - \boldsymbol{v}^{M-1}}{2\Delta t} &= \overline{\boldsymbol{v}}^M \,. \end{aligned} \right.$$

Let now $\boldsymbol{M}$ be as for the previous hyperbolic systems, see (4.20), and we have,

$$\begin{aligned} &\left\{ \begin{pmatrix} \boldsymbol{u}^M \\ \overline{\boldsymbol{u}}^M \end{pmatrix}, \begin{pmatrix} \boldsymbol{v}^M \\ \overline{\boldsymbol{v}}^M \end{pmatrix} \right\} - \left\{ \begin{pmatrix} \boldsymbol{u}^0 \\ \overline{\boldsymbol{u}}^0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{v}^0 \\ \overline{\boldsymbol{v}}^0 \end{pmatrix} \right\} \\ &\quad = \left\langle \frac{\boldsymbol{u}^{M+1} - \boldsymbol{u}^{M-1}}{2\Delta t}, \boldsymbol{C}\boldsymbol{v}^M \right\rangle - \left\langle \boldsymbol{u}^M, \boldsymbol{C}\frac{\boldsymbol{v}^{M+1} - \boldsymbol{v}^{M-1}}{2\Delta t} \right\rangle \\ &\qquad - \left\langle \frac{\boldsymbol{u}^1 - \boldsymbol{u}^{-1}}{2\Delta t}, \boldsymbol{C}\boldsymbol{v}^0 \right\rangle + \left\langle \boldsymbol{u}^0, \boldsymbol{C}\frac{\boldsymbol{v}^1 - \boldsymbol{v}^{-1}}{2\Delta t} \right\rangle \\ &\quad = \Delta t \sum_{n=0}^{M}{}' \left(\left\langle \frac{\boldsymbol{u}^{n+1} - 2\boldsymbol{u}^n + \boldsymbol{u}^{n-1}}{\Delta t^2}, \boldsymbol{C}\boldsymbol{v}^n \right\rangle - \left\langle \frac{\boldsymbol{v}^{n+1} - 2\boldsymbol{v}^n + \boldsymbol{v}^{n-1}}{\Delta t^2}, \boldsymbol{C}\boldsymbol{u}^n \right\rangle \right) \\ &\quad = \Delta t \sum_{n=0}^{M}{}' \Big(\langle \boldsymbol{A}\boldsymbol{u}^n + \boldsymbol{B}\boldsymbol{k}^n, \boldsymbol{v}^n \rangle - \langle \boldsymbol{A}\boldsymbol{v}^n, \boldsymbol{u}^n \rangle \Big) \\ &\quad = \Delta t \sum_{n=0}^{M}{}' \left\langle \boldsymbol{k}^n, \boldsymbol{B}^T \boldsymbol{v}^n \right\rangle \,. \end{aligned}$$

The second equality makes use of a discrete version of the identity

$$\int_0^T (f''g - fg'')dt = \big[f'g\big]_0^T - \big[fg'\big]_0^T \,,$$

which can be found, including a derivation, in Detail 4, page 183. The primed summation sign $\sum'$ means that the first and last term should be weighed with 1/2, while the intervening terms should be weighed with 1 as usual.

With $(\boldsymbol{u}^0, \overline{\boldsymbol{u}}^0) = (\boldsymbol{0}, \boldsymbol{0})$ and $\boldsymbol{k}^n = \boldsymbol{B}^T \boldsymbol{v}^n$ we define

$$\boldsymbol{\Lambda}_M^{\Delta t} \begin{pmatrix} \boldsymbol{v}^M \\ \overline{\boldsymbol{v}}^M \end{pmatrix} = \boldsymbol{M}^T \begin{pmatrix} \boldsymbol{u}^M \\ (\boldsymbol{u}^{M+1} - \boldsymbol{u}^{M-1})/(2\Delta t) \end{pmatrix} ,$$

leading to the important equality

$$\left\langle \boldsymbol{\Lambda}_M^{\Delta t} \begin{pmatrix} \boldsymbol{v}^M \\ \overline{\boldsymbol{v}}^M \end{pmatrix}, \begin{pmatrix} \boldsymbol{w}^M \\ \overline{\boldsymbol{w}}^M \end{pmatrix} \right\rangle_{\boldsymbol{\mathcal{C}}} = \Delta t \sum_{n=0}^{M}{}' \left\langle \boldsymbol{B}^T \boldsymbol{v}^n, \boldsymbol{B}^T \boldsymbol{w}^n \right\rangle . \tag{4.24}$$

This relation corresponds to that obtained in *Glowinski, Li, and Lions (1990)*. The above relation is, however, more general (in the cited paper, the space discretization is a simple 2D finite element discretization of the Laplacian where $\boldsymbol{\mathcal{C}}$, had they used that notation, is a diagonal matrix).

4.1.2.2 The Trapezoid Rule

As mentioned in the beginning of this chapter, the trapezoid rule has never been used in the context of discrete control systems. We will now derive the necessary relations.

When applying the trapezoid rule to the semi-discrete control system (4.1) we get

$$\boldsymbol{\mathcal{C}} \frac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^n}{\Delta t} = \boldsymbol{\mathcal{A}} \, \frac{\boldsymbol{u}^{n+1} + \boldsymbol{u}^n}{2} + \boldsymbol{\mathcal{B}} \, \frac{\boldsymbol{k}^{n+1} + \boldsymbol{k}^n}{2} , \tag{4.25}$$

where the initial condition is given as $\boldsymbol{u}^0$. Similarly, the semi-discrete adjoint system (4.2) turns in to

$$\boldsymbol{\mathcal{C}} \frac{\boldsymbol{v}^{n+1} - \boldsymbol{v}^n}{\Delta t} = -\widetilde{\boldsymbol{\mathcal{A}}} \, \frac{\boldsymbol{v}^{n+1} + \boldsymbol{v}^n}{2} , \tag{4.26}$$

where the initial condition is represented by $\boldsymbol{v}^M$ (the adjoint system is solved backwards in time).

The relation (4.7) now gets the appearance

$$\begin{aligned}
\{\boldsymbol{u}^M, \boldsymbol{v}^M\} - \{\boldsymbol{u}^0, \boldsymbol{v}^0\} &= \sum_{n=0}^{M-1} \Big(\{\boldsymbol{u}^{n+1}, \boldsymbol{v}^{n+1}\} - \{\boldsymbol{u}^n, \boldsymbol{v}^n\} \Big) \\
&= \Delta t \sum_{n=0}^{M-1} \left(\left\{ \frac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^n}{\Delta t}, \frac{\boldsymbol{v}^{n+1} + \boldsymbol{v}^n}{2} \right\} + \left\{ \frac{\boldsymbol{u}^{n+1} + \boldsymbol{u}^n}{2}, \frac{\boldsymbol{v}^{n+1} - \boldsymbol{v}^n}{\Delta t} \right\} \right) \\
&= \Delta t \sum_{n=0}^{M-1} \left(\left\{ \boldsymbol{\mathcal{C}}^{-1} \boldsymbol{\mathcal{A}} \, \frac{\boldsymbol{u}^{n+1} + \boldsymbol{u}^n}{2} + \boldsymbol{\mathcal{C}}^{-1} \boldsymbol{\mathcal{B}} \, \frac{\boldsymbol{k}^{n+1} + \boldsymbol{k}^n}{2}, \frac{\boldsymbol{v}^{n+1} + \boldsymbol{v}^n}{2} \right\} \right. \\
&\qquad \left. - \left\{ \frac{\boldsymbol{u}^{n+1} + \boldsymbol{u}^n}{2}, \boldsymbol{\mathcal{C}}^{-1} \widetilde{\boldsymbol{\mathcal{A}}} \, \frac{\boldsymbol{v}^{n+1} + \boldsymbol{v}^n}{2} \right\} \right)
\end{aligned}$$

$$= \Delta t \sum_{n=0}^{M-1} \left\langle \frac{\boldsymbol{k}^{n+1} + \boldsymbol{k}^n}{2}, \boldsymbol{\mathcal{B}}^T \boldsymbol{M} \frac{\boldsymbol{v}^{n+1} + \boldsymbol{v}^n}{2} \right\rangle .$$

When it comes to HUM, we can proceed similar to earlier and solve (4.25) and (4.26) with $\boldsymbol{u}^0 = \boldsymbol{0}$ and $\boldsymbol{k}^n = \boldsymbol{\mathcal{B}}^T \boldsymbol{M} \boldsymbol{v}^n$, thereby defining

$$\boldsymbol{\Lambda}_M^{\Delta t} \boldsymbol{v}^M = \boldsymbol{M}^T \boldsymbol{u}^M .$$

Inserting this into the relation above we get

$$\begin{aligned} \left\langle \boldsymbol{\Lambda}_M^{\Delta t} \boldsymbol{v}^M, \boldsymbol{w}^M \right\rangle_{\boldsymbol{\mathcal{C}}} &= \left\langle \boldsymbol{M}^T \boldsymbol{u}^M, \boldsymbol{w}^M \right\rangle_{\boldsymbol{\mathcal{C}}} = \{\boldsymbol{u}^M, \boldsymbol{w}^M\} - \{\boldsymbol{u}^0, \boldsymbol{w}^0\} \\ &= \Delta t \sum_{n=0}^{M-1} \left\langle \boldsymbol{\mathcal{B}}^T \boldsymbol{M} \frac{\boldsymbol{v}^{n+1} + \boldsymbol{v}^n}{2}, \boldsymbol{\mathcal{B}}^T \boldsymbol{M} \frac{\boldsymbol{w}^{n+1} + \boldsymbol{w}^n}{2} \right\rangle . \end{aligned} \tag{4.27}$$

When we consider hyperbolic systems as in (4.17) and (4.18) we can set $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{B}}$, $\boldsymbol{\mathcal{C}}$, $\widetilde{\boldsymbol{\mathcal{A}}}$, $\boldsymbol{M}$, as in (4.19) and (4.20), and replace $\boldsymbol{u}^n$ and $\boldsymbol{v}^n$ by $(\boldsymbol{u}^n, \overline{\boldsymbol{u}}^n)$ and $(\boldsymbol{v}^n, \overline{\boldsymbol{v}}^n)$, respectively. With $(\boldsymbol{u}^0, \overline{\boldsymbol{u}}^0) = (\boldsymbol{0}, \boldsymbol{0})$ and $\boldsymbol{k}^n = \boldsymbol{B}^T \boldsymbol{v}^n$ we set

$$\boldsymbol{\Lambda}_M^{\Delta t} \begin{pmatrix} \boldsymbol{v}^M \\ \overline{\boldsymbol{v}}^M \end{pmatrix} = \boldsymbol{M}^T \begin{pmatrix} \boldsymbol{u}^M \\ \overline{\boldsymbol{u}}^M \end{pmatrix} ,$$

leading to

$$\left\langle \boldsymbol{\Lambda}_M^{\Delta t} \begin{pmatrix} \boldsymbol{v}^M \\ \overline{\boldsymbol{v}}^M \end{pmatrix}, \begin{pmatrix} \boldsymbol{w}^M \\ \overline{\boldsymbol{w}}^M \end{pmatrix} \right\rangle_{\boldsymbol{\mathcal{C}}} = \Delta t \sum_{n=0}^{M-1} \left\langle \boldsymbol{B}^T \frac{\boldsymbol{v}^{n+1} + \boldsymbol{v}^n}{2}, \boldsymbol{B}^T \frac{\boldsymbol{w}^{n+1} + \boldsymbol{w}^n}{2} \right\rangle . \tag{4.28}$$

4.2 Uniform Observability

It's all very well in practice,
but it will never work in theory.
— French management saying

A necessary condition for having controllability is that the controllability operator is positive and thereby invertible. This is true whether we speak of continuous, semi-discrete or fully discrete systems. But for the two latter cases, what happens when the discretization level, as measured by the space dimension N, goes to infinity? Do the computed controls converge?

Consider the case of exact controllability for a fully discrete system. Assume now that constants $C_1, C_2 > 0$ exist such that

$$C_1 \|\boldsymbol{v}\|_{\widetilde{\boldsymbol{Q}}}^2 \le \langle \boldsymbol{\Lambda}_M^{\Delta t} \boldsymbol{v}, \boldsymbol{v} \rangle_{\boldsymbol{\mathcal{C}}} \le C_2 \|\boldsymbol{v}\|_{\widetilde{\boldsymbol{Q}}}^2 , \tag{4.29}$$

holds for all $\boldsymbol{v} \in \mathbb{R}^N$ on all discretization levels, $N \in \mathbb{N}$ (all quantities except C_1 and C_2 in this relation should be indexed with N, but we will omit such indices for simpler notation). When the constants C_1 and C_2 in this way do not depend on N, we call it *uniform observability.*

The above inequality can equivalently be written as

$$C_1 \boldsymbol{v}^T \widetilde{\boldsymbol{Q}} \boldsymbol{v} \le \boldsymbol{v}^T \boldsymbol{\mathcal{C}} \boldsymbol{\Lambda}_M^{\Delta t} \boldsymbol{v} \le C_2 \boldsymbol{v}^T \widetilde{\boldsymbol{Q}} \boldsymbol{v} \,,$$

for all $\boldsymbol{v} \in \mathbb{R}^N$ and all $N \in \mathbb{N}$. Since $\widetilde{\boldsymbol{Q}}$ is required to be symmetric and positive definite, it has a Cholesky factorization $\widetilde{\boldsymbol{Q}} = \boldsymbol{R}^T \boldsymbol{R}$, where $\boldsymbol{R}$ is an upper triangular matrix (see *Golub and Van Loan, 1996*, Theorem 4.2.5). Setting $\boldsymbol{w} = \boldsymbol{R}\boldsymbol{v}$, we get

$$C_1 \boldsymbol{w}^T \boldsymbol{w} \le \boldsymbol{w}^T \boldsymbol{R}^{-T} \boldsymbol{\mathcal{C}} \boldsymbol{\Lambda}_M^{\Delta t} \boldsymbol{R}^{-1} \boldsymbol{w} \le C_2 \boldsymbol{w}^T \boldsymbol{w} \,.$$

This shows that the eigenvalues of the symmetric matrix $\boldsymbol{R}^{-T} \boldsymbol{\mathcal{C}} \boldsymbol{\Lambda}_M^{\Delta t} \boldsymbol{R}^{-1}$ all lie between C_1 and C_2 for all N. From the relation we can also derive the following inequality for the inverse,

$$\frac{1}{C_2} \boldsymbol{w}^T \boldsymbol{w} \le \boldsymbol{w}^T \boldsymbol{R} \big(\boldsymbol{\Lambda}_M^{\Delta t}\big)^{-1} \boldsymbol{\mathcal{C}}^{-1} \boldsymbol{R}^T \boldsymbol{w} \le \frac{1}{C_1} \boldsymbol{w}^T \boldsymbol{w} \,.$$

Making the replacement $\boldsymbol{w} = \boldsymbol{R}^{-T} \boldsymbol{\mathcal{C}} \boldsymbol{u}$ this inequality is seen to be equivalent to

$$\frac{1}{C_2} \|\boldsymbol{u}\|^2_{\widetilde{\boldsymbol{Q}}'} \le \Big\langle \big(\boldsymbol{\Lambda}_M^{\Delta t}\big)^{-1} \boldsymbol{u}, \boldsymbol{u} \Big\rangle_{\boldsymbol{\mathcal{C}}} \le \frac{1}{C_1} \|\boldsymbol{u}\|^2_{\widetilde{\boldsymbol{Q}}'} \,,$$

for all $\boldsymbol{u} \in \mathbb{R}^N$ and all $N \in \mathbb{N}$ (see the definition of the discrete $\widetilde{\boldsymbol{Q}}'$-norm in (4.5)).

Since the condition number of a symmetric and positive definite matrix is the ratio of the largest eigenvalue to the smallest, we also see that *the condition number of the matrix $\boldsymbol{R}^{-T} \boldsymbol{\mathcal{C}} \boldsymbol{\Lambda}_M^{\Delta t} \boldsymbol{R}^{-1}$ is bounded by C_2/C_1, uniformly in N.*

Assume now that we wish to compute controls for an exact controllability problems, given some initial and final conditions, see Equation (4.10). Let a sequence of vectors $\boldsymbol{y}_N \in \mathbb{R}^N$ be given such that $\boldsymbol{y}_N$ converges to a limit vector $y \in H'$. Exactly how the convergence occurs is not important for now, but we assume that $\|\boldsymbol{y}_N\|_{\boldsymbol{Q}'} < C_y$ for all N and some $C_y > 0$ that does not depend on N. We now solve, for increasing values of N,

$$\boldsymbol{\Lambda}_M^{\Delta t} \boldsymbol{v}_N = \boldsymbol{M}^T \boldsymbol{y}_N \,, \tag{4.30}$$

an equation which can be written equivalently as

$$\big(\boldsymbol{R}^{-T} \boldsymbol{\mathcal{C}} \boldsymbol{\Lambda}_M^{\Delta t} \boldsymbol{R}^{-1}\big)(\boldsymbol{R}\boldsymbol{v}_N) = \boldsymbol{R}^{-T} \boldsymbol{\mathcal{C}} \boldsymbol{M}^T \boldsymbol{y}_N \,.$$

We now have the important bound,

$$\begin{aligned} \|\boldsymbol{v}_N\|_{\widetilde{\boldsymbol{Q}}} = \|\boldsymbol{R}\boldsymbol{v}_N\| &\le \big\|\boldsymbol{R}\big(\boldsymbol{\Lambda}_M^{\Delta t}\big)^{-1} \boldsymbol{\mathcal{C}}^{-1} \boldsymbol{R}^T\big\| \big\|\boldsymbol{R}^{-T} \boldsymbol{\mathcal{C}} \boldsymbol{M}^T \boldsymbol{y}_N\big\| \\ &\le \frac{1}{C_1} \|\boldsymbol{y}_N\|_{\boldsymbol{Q}'} \le \frac{C_y}{C_1} \,, \end{aligned}$$

cf. the norms introduced earlier, see (4.6). So the solutions to Equation (4.30) will be uniformly bounded in the $\|\cdot\|_{\widetilde{\boldsymbol{Q}}}$-norm as $N \to \infty$. Recall, though, that the actual control, for a given N, is computed via the adjoint system with $\boldsymbol{v}_N$ as initial condition. But the norm of the control is given by the inner product $\langle \boldsymbol{\Lambda}_M^{\Delta t}\boldsymbol{v}, \boldsymbol{v}\rangle_{\boldsymbol{\mathcal{C}}}$ (see (4.8) and (4.24) for a semi-discrete and fully discrete example, respectively). We thus have

$$\langle \boldsymbol{\Lambda}_M^{\Delta t}\boldsymbol{v}_N, \boldsymbol{v}_N\rangle_{\boldsymbol{\mathcal{C}}} \leq C_2\|\boldsymbol{v}_N\|_{\boldsymbol{Q}}^2 \leq C_y^2\frac{C_2}{C_1^2}\,, \tag{4.31}$$

showing that the corresponding controls stay uniformly bounded.

To finish the proof that uniform observability leads to convergent controls, we need to show that applying the obtained controls to the discrete control system is *consistent* with the continuous control system. This should follow if the discretization of the adjoint system and the control system have been proved convergent. In a rigorous proof, though, it would be required.

Assume, on the other hand, that uniform observability does *not* hold. This means that $C_1 \to 0$ and/or $C_2 \to \infty$ as $N \to \infty$ in the double inequality (4.29). The boundedness of controls can thus not be guaranteed, easily seen from (4.31). Furthermore, the condition number C_2/C_1 of the matrix $\boldsymbol{R}^{-T}\boldsymbol{\mathcal{C}}\boldsymbol{\Lambda}_M^{\Delta t}\boldsymbol{R}^{-1}$ will tend to infinity. This means that in a practical implementation, computing a control will become increasingly liable to rounding errors (see also Section 9.5).

The fact that uniform observability is necessary in order to have convergent controls was first observed in *Infante and Zuazua (1998)* (see *Infante and Zuazua (1999)* for an improved version of this paper). It has since then been commented upon in, e.g., *Micu (2002)*, *Zuazua (2003)*, *Negreanu and Zuazua (2004a)*, and other publications by these authors. Only in *León and Zuazua (2002)* has a rigorous proof been given for the convergence of controls (for the case of the one-dimensional beam equation, $u_{tt} = u_{xxxx}$).

The derivations of this section show, in great generality, the implications of having uniform observability, and how to show convergence of controls for a concrete control system.

4.2.1 Hyperbolic Systems

This and the following two sections will present theorems that have been used in the literature to show observability inequalities for continuous systems and to show uniform observability inequalities for (semi-)discrete systems. How to apply these results in practice will be postponed until later chapters.

The first theorem is a classical result in the area of non-harmonic Fourier series. It was first published in *Ingham (1936)*, see also *Young (2001)*. (We here use ℓ^2 to denote the set of infinite sequences $\langle a_k\rangle$ of real or complex numbers that are square summable, $\sum |a_k|^2 < \infty$.)

Theorem 4.2.1 (Ingham). *Let $\langle \mu_k\rangle_{k\in\mathbb{Z}}$ be a sequence of real numbers for which*

$$\mu_{k+1} - \mu_k \geq \gamma, \quad \forall k \in \mathbb{Z}\,,$$

for some $\gamma > 0$. *Then for any* $T > 2\pi/\gamma$ *there exist constants* $C_1, C_2 > 0$, *both depending* only *on* T *and* γ, *such that*

$$C_1 \sum_{k\in\mathbb{Z}} |c_k|^2 \le \int_0^T \left| \sum_{k\in\mathbb{Z}} c_k e^{i\mu_k t} \right|^2 dt \le C_2 \sum_{k\in\mathbb{Z}} |c_k|^2 \,,$$

for all sequences of complex numbers $\langle c_k \rangle_{k\in\mathbb{Z}} \in \ell^2$.

The next theorem is very similar. However, whereas Ingham's Theorem demanded real numbers $\langle \mu_k \rangle_{k\in\mathbb{Z}}$ with a uniform gap, the following theorem allows each μ_k to be complex valued, as long as they lie appropriately close to uniformly distributed points on the imaginary axis. It is Kadec' classical 1/4-theorem, see *Kadec (1964)*.

Theorem 4.2.2 (Kadec). *Let* $\langle \mu_k \rangle_{k\in\mathbb{Z}}$ *be a sequence of complex numbers for which*

$$\sup_{k\in\mathbb{Z}} \left| \frac{\mu_k}{\gamma} - k \right| < \frac{1}{4} \,,$$

for some $\gamma > 0$. *Then for any* $T \ge 2\pi/\gamma$ *there exist constants* $C_1, C_2 > 0$, *both depending* only *on* T *and* γ, *such that*

$$C_1 \sum_{k\in\mathbb{Z}} |c_k|^2 \le \int_0^T \left| \sum_{k\in\mathbb{Z}} c_k e^{i\mu_k t} \right|^2 dt \le C_2 \sum_{k\in\mathbb{Z}} |c_k|^2 \,,$$

for all sequences of complex numbers $\langle c_k \rangle_{k\in\mathbb{Z}} \in \ell^2$.

Note how the previous two theorems actually indicate when the functions $\{e^{i\mu_k t} \mid k \in \mathbb{Z}\}$ constitute a Riesz basis on the interval $(0, T)$.

For the most general results, complex exponentials were used in the above theorems. Let us see how real coefficients with sine and cosine functions can be written in such a form. Let two sequences $\langle a_k \rangle_{k\in\mathbb{N}} \in \ell^2$ and $\langle b_k \rangle_{k\in\mathbb{N}} \in \ell^2$ be given and let $\langle \eta_k \rangle_{k\in\mathbb{N}}$ be a sequence of real numbers. Then by setting

$$\begin{aligned} c_k &= \tfrac{1}{2}(a_k - ib_k)\,, &\quad \mu_k &= \eta_k \,, \\ c_0 &= 0\,, && \\ c_{-k} &= \tfrac{1}{2}(a_k + ib_k)\,, & \mu_{-k} &= -\eta_k \,, \end{aligned}$$

for $k \in \mathbb{N}$, we have

$$\sum_{k\in\mathbb{N}} (a_k \cos(t\eta_k) + b_k \sin(t\eta_k)) = \sum_{k\in\mathbb{Z}} c_k e^{i\mu_k t}, \quad \forall t \in \mathbb{R} \,.$$

We furthermore have the following relation between the ℓ^2-norms,

$$\sum_{k\in\mathbb{Z}} (|a_k|^2 + |b_k|^2) = 2 \sum_{k\in\mathbb{Z}} |c_k|^2 \,.$$

4.2.2 Parabolic Systems

Just as the two previous theorems are relevant in the context of hyperbolic systems, the following theorem is relevant for parabolic systems. It was first published in *Fattorini and Russell (1974)*, see also *López and Zuazua (2002)*.

We start with some notation. Given $\xi > 0$ and a decreasing function $M : (0,\infty) \mapsto \mathbb{N}$ such that $M(\delta) \to \infty$ as $\delta \to 0$, we introduce the class $\mathcal{P}(\xi, M)$ of increasing sequences of positive real numbers $\langle \mu_j \rangle_{j\in\mathbb{N}}$ such that

$$\mu_{j+1} - \mu_j \geq \xi > 0, \qquad \forall j \in \mathbb{N}\,, \tag{4.32}$$

$$\sum_{k\geq M(\delta)} \mu_k^{-1} \leq \delta, \qquad \forall \delta > 0\,. \tag{4.33}$$

The following now holds.

Theorem 4.2.3. *Given a class $\mathcal{P}(\xi, M)$ of sequences and $T > 0$ there exists a constant $C > 0$ (which depends on ξ, M and T) such that*

$$\int_0^T \left| \sum_{k=1}^{\infty} a_k e^{-\mu_k t} \right|^2 dt \geq \frac{C}{\sum_{k=1}^{\infty} \mu_k^{-1}} \sum_{k=1}^{\infty} \frac{e^{-2\mu_k T}}{\mu_k} |a_k|^2$$

for all $\langle \mu_j \rangle_{j\in\mathbb{N}} \in \mathcal{P}(\xi, N)$ and all bounded sequences of real numbers $\langle a_k \rangle_{k\in\mathbb{N}}$.

4.2.3 Time Discrete Version of Ingham's Theorem

The three theorems we have seen by now can only be applied to PDEs or semi-discrete systems because of the $\int_0^T$ time integrals. We will in this section prove a time-discrete version of Ingham's Theorem. The proof builds upon a similar result in *Negreanu and Zuazua (2004b)*, which, in turn, builds upon the original proof in *Ingham (1936)*, see also *Young (2001)*. The assumptions of the present theorem have been improved, however, and no unknown constants appear.

Let $g(x) = \sin(\frac{1}{2}x)\chi_{[0,2\pi]}(x)$ with corresponding Fourier transform $\hat{g}(\xi)$,

$$\hat{g}(\xi) = \int_{\mathbb{R}} g(x) e^{-i\xi x} dx = \int_0^{2\pi} \sin(\tfrac{1}{2}x) e^{-i\xi x} dx = \frac{2 + 2e^{-2i\pi\xi}}{1 - 4\xi^2}\,. \tag{4.34}$$

When sampling the function $g(x)$ onto the grid $h\mathbb{Z}$, we obtain the discrete Fourier transform $\hat{g}_h$ as

$$\hat{g}_h(\xi) = h \sum_{n\in\mathbb{Z}} g(nh) e^{-i\xi nh} = h \sum_{n=0}^{\lfloor 2\pi/h \rfloor} \sin(\tfrac{1}{2}nh) e^{-i\xi nh}\,. \tag{4.35}$$

We are now interested in how well $\hat{g}_h(\xi)$ approximates $\hat{g}(\xi)$ as $h \to 0$. To this end the *Poisson summation formula* proves very useful,

$$\hat{g}_h(\xi) = \sum_{j\in\mathbb{Z}} \hat{g}(\xi + 2\pi j/h)\,, \tag{4.36}$$

see *Henrici (1977)*, Theorem 10.6e, or *Trefethen (1996)*, Theorem 2.7. Using now that $|\hat{g}(\xi)| \le 2/\xi^2$ for all $\xi \in \mathbb{R}$ (see Detail 5, page 184) we get

$$\begin{aligned} \big|\hat{g}_h(\xi) - \hat{g}(\xi)\big| &= \left|\sum_{j=1}^{\infty}\big(\hat{g}(\xi + 2j\pi/h) + \hat{g}(\xi - 2j\pi/h)\big)\right| \\ &\le \sum_{j=1,3,5,\ldots} \big|\hat{g}(-j\pi/h) + \hat{g}(j\pi/h)\big| \le \frac{4h^2}{\pi^2} \sum_{j=1,3,5,\ldots} j^{-2} = \tfrac{1}{2}h^2\,, \end{aligned} \tag{4.37}$$

for $-\frac{\pi}{h} \le \xi \le \frac{\pi}{h}$. Note that no unknown constants appear in this bound. We now have the following lemma that concerns an inverse inequality for a time interval of length 2π.

Lemma 4.2.1. *Let $h > 0$, $\gamma > 1$, $N \in \mathbb{N}$ and $\lambda_{-N}, \lambda_{-N+1}, \ldots, \lambda_N \in \mathbb{R}$ be such that*

$$\begin{aligned} &\lambda_{k+1} - \lambda_k \ge \gamma, && \text{for } k = -N, \ldots, N-1, \\ &\lambda_N - \lambda_{-N} \le 2\pi/h - \gamma\,. \end{aligned} \tag{4.38}$$

Then for all complex sequences $\langle c_k \rangle_{k=-N}^{N}$ we have

$$C_1(\gamma, h, N) \sum_{k=-N}^{N} |c_k|^2 \le h \sum_{n=0}^{\lfloor 2\pi/h \rfloor} \left| \sum_{k=-N}^{N} c_k e^{i\lambda_k nh} \right|^2 , \tag{4.39}$$

where $C_1(\gamma, h, N) = 4\left(1 - \dfrac{1}{\gamma^2}\right) - \frac{1}{2}(1 + 2N)h^2$.

Proof. Observe that

$$\begin{aligned} h \sum_{n=0}^{\lfloor 2\pi/h \rfloor} \left| \sum_{k=-N}^{N} c_k e^{i\lambda_k nh} \right|^2 &\ge h \sum_{n=0}^{\lfloor 2\pi/h \rfloor} \sin(\tfrac{1}{2}nh) \left| \sum_{k=-N}^{N} c_k e^{i\lambda_k nh} \right|^2 \\ &= \sum_{k=-N}^{N} \sum_{l=-N}^{N} c_k \overline{c}_l \; \hat{g}_h(\lambda_l - \lambda_k) \\ &\ge \hat{g}_h(0) \sum_{k=-N}^{N} |c_k|^2 - \sum_{k=-N}^{N} \sum_{\substack{l=-N \\ l \ne k}}^{N} |c_k \overline{c}_l|\; |\hat{g}_h(\lambda_l - \lambda_k)| \\ &\ge \hat{g}_h(0) \sum_{k=-N}^{N} |c_k|^2 - \sum_{k=-N}^{N} |c_k|^2 \sum_{\substack{l=-N \\ l \ne k}}^{N} |\hat{g}_h(\lambda_l - \lambda_k)|\,, \end{aligned} \tag{4.40}$$

where we use that $2|c_k\overline{c}_l| \le |c_k|^2 + |c_l|^2$ and that $\hat{g}_h$ is an even function. We now have

$$\sum_{\substack{l=-N\\ l\neq k}}^{N} |\hat{g}_h(\lambda_l - \lambda_k)| = \sum_{\substack{l=-N, l\neq k\\ |\lambda_l-\lambda_k|\leq\frac{\pi}{h}}}^{N} |\hat{g}_h(|\lambda_l - \lambda_k|)| + \sum_{\substack{l=-N, l\neq k\\ |\lambda_l-\lambda_k|>\frac{\pi}{h}}}^{N} |\hat{g}_h(\tfrac{2\pi}{h} - |\lambda_l - \lambda_k|)|$$

$$\leq \sum_{\substack{l=-N, l\neq k\\ |\lambda_l-\lambda_k|\leq\frac{\pi}{h}}}^{N} |\hat{g}(|\lambda_l - \lambda_k|)| + \sum_{\substack{l=-N, l\neq k\\ |\lambda_l-\lambda_k|>\frac{\pi}{h}}}^{N} |\hat{g}(\tfrac{2\pi}{h} - |\lambda_l - \lambda_k|)| + Nh^2\,,$$

where we have used the bound (4.37). We now deduce from (4.38) that

$$\begin{aligned} |\lambda_l - \lambda_k| &\geq \gamma|l-k|,\\ \tfrac{2\pi}{h} - |\lambda_l - \lambda_k| &\geq \gamma(2N+1-|l-k|)\,, \end{aligned}$$

for all $k, l \in \{-N, -N+1, \dots, N\}$. This yields

$$\begin{aligned} &\sum_{\substack{l=-N, l\neq k\\ |\lambda_l-\lambda_k|\leq\frac{\pi}{h}}}^{N} |\hat{g}(|\lambda_l - \lambda_k|)| + \sum_{\substack{l=-N, l\neq k\\ |\lambda_l-\lambda_k|>\frac{\pi}{h}}}^{N} |\hat{g}(\tfrac{2\pi}{h} - |\lambda_l - \lambda_k|)| \\ &\qquad\leq \sum_{\substack{l=-N, l\neq k\\ |\lambda_l-\lambda_k|\leq\frac{\pi}{h}}}^{N} \frac{4}{4|\lambda_l-\lambda_k|^2-1} + \sum_{\substack{l=-N, l\neq k\\ |\lambda_l-\lambda_k|>\frac{\pi}{h}}}^{N} \frac{4}{4(\frac{2\pi}{h} - |\lambda_l-\lambda_k|)^2-1} \\ &\qquad\leq \sum_{\substack{l=-N, l\neq k\\ |\lambda_l-\lambda_k|\leq\frac{\pi}{h}}}^{N} \frac{4}{4\gamma^2|l-k|^2-1} + \sum_{\substack{l=-N, l\neq k\\ |\lambda_l-\lambda_k|>\frac{\pi}{h}}}^{N} \frac{4}{4\gamma^2\big(2N+1-|l-k|\big)^2-1} \\ &\qquad\leq \sum_{l=-N, l\neq k}^{N} \frac{4}{4\gamma^2|l-k|^2-1} + \sum_{l=-N, l\neq k}^{N} \frac{4}{4\gamma^2\big(2N+1-|l-k|\big)^2-1} \\ &\qquad= \sum_{r=1}^{2N} \frac{8}{4\gamma^2r^2-1} < \frac{8}{\gamma^2}\sum_{r=1}^{\infty}\frac{1}{4r^2-1} = \frac{4}{\gamma^2}\sum_{r=1}^{\infty}\left(\frac{1}{2r-1} - \frac{1}{2r+1}\right) = \frac{4}{\gamma^2}\,. \end{aligned}$$

Inserting the bounds into (4.40) we get

$$\begin{aligned} &h\sum_{n=0}^{\lfloor 2\pi/h\rfloor}\left|\sum_{k=-N}^{N} c_k e^{i\lambda_k nh}\right|^2 \\ &\qquad\geq (\hat{g}(0) - \tfrac{1}{2}h^2)\sum_{k=-N}^{N}|c_k|^2 - (4/\gamma^2 + Nh^2)\sum_{k=-N}^{N}|c_k|^2 \\ &\qquad= \left[4\left(1 - \frac{1}{\gamma^2}\right) - \tfrac{1}{2}(2N+1)h^2\right]\sum_{k=-N}^{N}|c_k|^2\,. \end{aligned}$$

$\square$

We then consider the direct inequality, still for a time interval of length 2π.

Lemma 4.2.2. *Under the conditions of Lemma 4.2.1 we have for all complex sequences* $\langle c_k\rangle_{k=-N}^{N}$,

$$h \sum_{n=0}^{\lfloor 2\pi/h\rfloor} \left| \sum_{k=-N}^{N} c_k e^{i\lambda_k nh} \right|^2 \leq C_2(\gamma, h, N) \sum_{k=-N}^{N} |c_k|^2 , \tag{4.41}$$

where $C_2(\gamma, h, N) = 16\left(1 - \dfrac{1}{9\gamma^2}\right) - \frac{1}{18}(1+2N)h^2$.

Proof. Assume $h \leq 1$ and we get

$$h \sum_{n=0}^{\lfloor 2\pi/h\rfloor} \left| \sum_{k=-N}^{N} c_k e^{i\lambda_k nh} \right|^2 \leq \tfrac{4}{3} h \sum_{n=0}^{\lfloor 6\pi/h\rfloor} \sin(\tfrac{1}{6}nh) \left| \sum_{k=-N}^{N} c_k e^{i\lambda_k (n-M)h} \right|^2 = I ,$$

since $1 \leq \frac{4}{3}\sin(\frac{1}{6}nh)$ for $\lfloor 2\pi/h\rfloor \leq n \leq 2\lfloor 2\pi/h\rfloor$. We set $h' = h/3$ and $c'_k = c_k e^{-i\lambda_k Mh}$ and get

$$I = 4h' \sum_{n=0}^{\lfloor 2\pi/h'\rfloor} \sin(\tfrac{1}{2}nh') \left| \sum_{k=-N}^{N} c'_k e^{i3\lambda_k nh'} \right|^2 = 4 \sum_{k,l=-N}^{N} c_k \overline{c}_l \; \hat{g}_{h'}(3(\lambda_l - \lambda_k)) .$$

Using now the same techniques as in Lemma 4.2.1, but using h' for h and 3γ for γ, we get the desired expression for C_2. Finally we observe that $|c'_k| = |c_k|$. ◻

We finally collect the results of the two lemmas and generalize to time intervals of any length.

Theorem 4.2.4. *Let time $T > 0$, time step $\Delta t > 0$ and $M \in \mathbb{N}$ be given such that $T = M\Delta t$. If the real numbers $\lambda_{-N}, \lambda_{-N+1}, \ldots, \lambda_N \in \mathbb{R}$ satisfy*

$$\lambda_{k+1} - \lambda_k \geq \gamma, \qquad \text{for } k = -N, \ldots, N-1, \tag{4.42}$$

$$\lambda_N - \lambda_{-N} \leq \frac{2\pi}{\Delta t} - \gamma , \tag{4.43}$$

where $\gamma > \dfrac{2\pi}{T}$, *then for all complex sequences* $\langle c_k\rangle_{k=-N}^{N}$ *we have*

$$C_1(T,\gamma,N,M) \sum_{k=-N}^{N} |c_k|^2 \leq \Delta t \sum_{n=0}^{M} \left| \sum_{k=-N}^{N} c_k e^{i\lambda_k n\Delta t} \right|^2 \leq C_2(T,\gamma,N,M) \sum_{k=-N}^{N} |c_k|^2 , \tag{4.44}$$

where

$$\begin{aligned} C_1(T,\gamma,M,N) &= \frac{2T}{\pi}\left(1 - \frac{4\pi^2}{T^2\gamma^2}\right) - \frac{4\pi^2(1+2N)}{2M^2}, \\ C_2(T,\gamma,M,N) &= \frac{8T}{\pi}\left(1 + \frac{4\pi^2}{9T^2\gamma^2}\right) + \frac{4\pi^2(1+2N)}{18M^2} . \end{aligned} \tag{4.45}$$

Proof. Let $\Delta t' = \frac{2\pi}{T}\Delta t$ such that $M\Delta t' = 2\pi$. We get

$$\Delta t \sum_{n=0}^{M} \left| \sum_{k=-N}^{N} c_k e^{i\lambda_k n \Delta t} \right|^2 = \tfrac{T}{2\pi}\Delta t' \sum_{n=0}^{2\pi/\Delta t'} \left| \sum_{k=-N}^{N} c_k e^{i\frac{T}{2\pi}\lambda_k n \Delta t'} \right|^2 .$$

We observe that

$$\begin{aligned} \lambda_{k+1} - \lambda_k \geq \gamma > \frac{2\pi}{T} \quad &\Leftrightarrow \quad \frac{T}{2\pi}(\lambda_{k+1} - \lambda_k) \geq \frac{T}{2\pi}\gamma = \gamma' > 1, \\ \lambda_N - \lambda_{-N} \leq \frac{2\pi}{\Delta t} - \gamma \quad &\Leftrightarrow \quad \frac{T}{2\pi}(\lambda_N - \lambda_{-N}) \leq \frac{2\pi}{\frac{2\pi}{T}\Delta t} - \frac{T}{2\pi}\gamma = \frac{2\pi}{\Delta t'} - \gamma' , \end{aligned}$$

so Lemmas 4.2.1 and 4.2.2 can be applied. The result is obtained by insertion into the bounds of the lemmas. □

The condition (4.43) was in *Negreanu and Zuazua (2004b)* replaced by

$$|\lambda_k - \lambda_l| \leq \frac{2\pi - (\Delta t)^p}{\Delta t} ,$$

for all $k, l \in \mathbb{Z}$ and some $0 \leq p \leq 1/2$, with which the authors could complete a similar theorem. Using the condition (4.43), however, is quite natural and follows from periodic nature of the function $\hat{g}_h(\xi)$.

Chapter 5

Properties of the Controllability Operator

This chapter collects together
some basic mathematical properties
— Jeffrey H. Kingston (1990)

The controllability operator Λ_T is an essential operator when it comes to controllability and HUM. Recall that Λ_T is always linear, symmetric and positive semi-definite. Its invertibility is reflected by the observability inequalities of Chapter 2.

This chapter treats two topics. One is how two compute a (possibly infinite) matrix representation of Λ_T with respect to some appropriate bases. This is possible to do analytically for some special cases such as the one dimensional wave equation and the one dimensional heat equation. When Λ_T is finite dimensional, its matrix representation can be quite useful for computing HUM controls in practice.

The second topic is asymptotic properties of the controllability operator, or more specifically, whether Λ_T converges to some limit operator as $T \to \infty$. This will be studied for the heat equation and wave equation, and it turns out that the controllability operator for the wave equation has a very simple limit operator for some interesting domains.

5.1 Computing the Controllability Operator

Recall from Section 2.5 that the controllability operator is an operator between a Hilbert space $\widetilde{H}$ and its adjoint, $\Lambda_T : \widetilde{H} \mapsto \widetilde{H}'$.

Assume we are given a basis $\langle e_j \rangle$ for $\widetilde{H}$ and another $\langle e'_i \rangle$ for $\widetilde{H}'$. We would like to compute a (possibly infinite) matrix $\mathbf{\Lambda}_T$ such that

$$\boldsymbol{y} = \mathbf{\Lambda}_T \boldsymbol{v} \quad \Leftrightarrow \quad y = \Lambda_T v$$

for every instance of

$$y = \sum_i \boldsymbol{y}_i e'_i \quad \text{and} \quad v = \sum_j \boldsymbol{v}_j e_j \,.$$

We will assume that the bases $\langle e_j \rangle$ and $\langle e'_i \rangle$ have the following orthogonality property,

$$\langle e'_i, e_j \rangle_{\widetilde{H}' \times \widetilde{H}} = \delta_{ij} \,.$$

We now get

$$\Lambda_T v = y \quad \Leftrightarrow \quad \sum_j (\Lambda_T e_j)\boldsymbol{v}_j = \sum_i e'_i \boldsymbol{y}_i \quad \Leftrightarrow \quad \sum_j \langle \Lambda_T e_j, e_i \rangle \boldsymbol{v}_j = \boldsymbol{y}_i \,.$$

This clearly shows that the (i,j)th entry of the matrix $\boldsymbol{\Lambda}_T$ is the number $\langle \Lambda_T e_j, e_i \rangle$. Two methods now suggest themselves.

The direct method. This method relies on computing directly $y = \Lambda_T e_j$ and then $\langle y, e_i \rangle$, thus determining one column of $\boldsymbol{\Lambda}_T$ for each application of Λ_T. Let us recall how an arbitrary vector $v^0 \in \widetilde{H}$ is mapped by Λ_T. We initially solve the adjoint system,

$$\begin{cases} v_t = -\widetilde{\mathcal{A}} v & \text{in } Q \;, \\ \mathcal{B} v = 0 & \text{in } \Sigma \;, \\ v(T) = v^0 & \text{in } \Omega \;, \end{cases} \tag{5.1}$$

followed by

$$\begin{cases} u_t = \mathcal{A} u & \text{in } Q \;, \\ \mathcal{B} u = \begin{cases} \mathcal{C} v \\ 0 \end{cases} & \begin{matrix} \text{in } \Sigma_0 \;, \\ \text{in } \Sigma \setminus \Sigma_0 \;, \end{matrix} \\ u(0) = 0 & \text{in } \Omega \;, \end{cases} \tag{5.2}$$

and then finally setting

$$\Lambda_T(v^0) = \mathcal{M}^T u(T) \,.$$

The inner product method. The following expression from Chapter 2 provides exactly what we need,

$$\left\langle \Lambda_T v^0, w^0 \right\rangle = \gamma_T(v^0, w^0) = \int_0^T \int_{\Gamma_0} \mathcal{C} v \, \mathcal{C} w \, d\Gamma dt \,, \tag{5.3}$$

where $v(t)$ and $w(t)$ are solutions of adjoint system (5.1) with initial conditions v^0 and w^0, respectively. If we now set $v^0 = e_j$ and $w^0 = e_i$ and alternate the indices, we compute the $\boldsymbol{\Lambda}_T$ matrix indices.

The inner product method has some obvious advantages:

- It is not necessary to compute *any* solutions to the control system (5.2).

- We can obtain $\mathbf{\Lambda}_{T_2}$ from $\mathbf{\Lambda}_{T_1}$ by using

$$\left\langle \Lambda_{T_2} v^0, w^0 \right\rangle = \left\langle \Lambda_{T_1} v^0, w^0 \right\rangle + \int_{T_1}^{T_2} \int_{\Gamma_0} \mathcal{C}v\, \mathcal{C}w \; d\Gamma\, dt \,.$$

- Since $\langle \Lambda_T e_j, e_i \rangle = \langle \Lambda_T e_i, e_j \rangle$, we only need to compute half the entries.

The direct method has been used in different contexts in the literature. In *Glowinski, Li, and Lions (1990)*, the authors used the direct method to compute the controllability operator analytically for the wave equation on the 2D domain $\Omega = (0,1) \times (0,1)$ (they based their calculations on $\{\sin(i\pi x)\} \times \{\sin(j\pi y)\}$ bases, and restricted the control time T to the cases $T = (n + 3/4)/\sqrt{2}$, $n = 0, 1, \ldots$, for easier computations). In *Glowinski, Li, and Lions (1990)*, *Asch and Lebeau (1998)* and *Negreanu and Zuazua (2003)*, the authors use the direct method in a discrete setting. In each of these publications they solve controllability problems using the Conjugate Gradients algorithm (see Section 2.7.1), and the direct method is used for computing the map $v \mapsto \Lambda_T v$.

The idea behind the inner product method is quite simple, but the method has not been described before in the literature (in *Eljendy (1992)*, though, similar ideas are used for solving exact controllability problems through optimization). The inner product method has some very appealing properties, some of which we will return to in Chapter 9 concerning implementations.

Chapters 6 and 7 will provide examples of both methods for the heat equation and wave equation, respectively, where analytical representations of the controllability operators will be calculated.

5.1.1 Special Considerations for Discretizations

Both methods apply easily to both semi-discretizations and full discretizations.

To use the direct method, one does not even have to think about bases. By successively mapping the columns of an appropriately dimensioned identity matrix, the corresponding columns of $\mathbf{\Lambda}_T$ will be computed.

When using the inner product method, the relation (5.3) can obviously not be used. The discrete equivalences in Chapter 4 must be used instead, see the relations (4.8), (4.24), (4.27) and (4.28). Note, however, that we in these cases compute

$$\left\langle \mathbf{\Lambda}_T^h \boldsymbol{v}, \boldsymbol{w} \right\rangle_{\boldsymbol{\mathcal{C}}} \quad \text{or} \quad \left\langle \mathbf{\Lambda}_M^{\Delta t} \boldsymbol{v}, \boldsymbol{w} \right\rangle_{\boldsymbol{\mathcal{C}}} \;,$$

so, noting the inner product used, we actually compute the entries of $\boldsymbol{\mathcal{C}}\mathbf{\Lambda}_T^h$ or $\boldsymbol{\mathcal{C}}\mathbf{\Lambda}_M^{\Delta t}$.

5.2 Asymptotic Properties of the Controllability Operator

We must act as if we had eternity before us.
— Umberto Eco (The Name of the Rose, 1980)

This section will investigate whether the controllability operator Λ_T has a limit operator as $T \to \infty$. Such results for the wave equation have previously appeared in *Glowinski, Li, and Lions (1990)*, *Glowinski and Lions (1995)*, *Bensoussan (1990)* and Bensoussan (1992). Actually, the two latter references studied general skew-symmetric operators and only the same two references contained proofs.

We will similarly consider the wave equation, including detailed proofs. We furthermore prove that also for the heat equation does a limit controllability operator exist. A general result for the abstract formulation of Chapter 2 is not known.

Our approach in both cases is intimately tied to the eigenvectors of the (negative) Laplacian, and we will assume that we have an orthonormal basis of eigenvectors in $L^2(\Omega)$,

$$\begin{cases} -\Delta w_k = \lambda_k w_k, & \text{in } \Omega\,, \\ \mathcal{B} w_k = 0, & \text{on } \Gamma\,, \\ \langle w_k, w_l \rangle_{L^2(\Omega)} = \delta_{kl}\,, \end{cases} \tag{5.4}$$

for all $k, l \in \mathbb{N}$, and where all eigenvalues are distinct, $0 < \lambda_1 < \lambda_2 < \cdots$, and $\lambda_k \to \infty$ for $k \to \infty$. For shorter notation we will use $\mu_k^2 = \lambda_k$.

We introduce a Hilbert space $\widetilde{H}_1$ by defining its inner product in terms of the eigenvectors,

$$\langle w_k, w_l \rangle_{\widetilde{H}_1} = \delta_{kl} \lambda_k \quad \text{for all } k, l \in \mathbb{N}\,. \tag{5.5}$$

This space is equivalent to $H_0^1(\Omega)$ when we deal with Dirichlet boundary conditions, $\mathcal{B} = I$.

Another important assumption deals with the complementary boundary operator $\mathcal{C}$ applied to the eigenvectors, namely the bound

$$\int_{\Gamma_0} |\mathcal{C} w_k|^2 d\Gamma \le K \lambda_k \quad \text{for all } k \in \mathbb{N}\,, \tag{5.6}$$

for some constant $K > 0$. This relation is not trivial. For the case $\mathcal{B} = I$, where the domain Ω has certain properties, it can be derived. Furthermore, in the case of the wave equation it is a consequence of the well-posedness of the control system. How to derive (5.6) in both cases will be shown later.

5.2.1 The Heat Equation

The adjoint system for the heat equation is

$$\begin{cases} v_t = -\Delta v & \text{in } Q \ , \\ \mathcal{B}v = 0 & \text{in } \Sigma \ , \\ v(T) = v^0 & \text{in } \Omega \ , \end{cases}$$

with $v^0 \in \widetilde{H} = \widetilde{H}_1$ and thus $v(t) \in \widetilde{H}$. Given initial conditions

$$v^0 = \sum_{k=1}^{\infty} a_k w_k \ ,$$

with $\langle \mu_k a_k \rangle_{k=1}^{\infty} \in \ell^2$, the heat equation has the solution,

$$v(t) = \sum_{k=1}^{\infty} a_k e^{-\lambda_k (T-t)} w_k \ .$$

Let us similarly consider a solution $\tilde{v}(t)$, corresponding to the coefficients $\langle \tilde{a}_k \rangle_{k=1}^{\infty}$ for which $\langle \mu_k \tilde{a}_k \rangle_{k=1}^{\infty} \in \ell^2$.

Using the equality (5.3), we now have

$$\begin{aligned} \langle \Lambda_T v^0, \tilde{v}^0 \rangle &= \int_0^T \int_{\Gamma_0} \mathcal{C}v \, \mathcal{C}\tilde{v} \; d\Gamma dt \\ &= \int_0^T \int_{\Gamma_0} \left(\sum_{j=1}^{\infty} a_j e^{-\lambda_j (T-t)} \mathcal{C}w_j \right) \left(\sum_{k=1}^{\infty} \tilde{a}_k e^{-\lambda_k (T-t)} \mathcal{C}w_k \right) d\Gamma dt \\ &= \int_0^T \int_{\Gamma_0} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_j \tilde{a}_k e^{-(\lambda_j + \lambda_k) t} \, \mathcal{C}w_j \, \mathcal{C}w_k \; d\Gamma dt \\ &= \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_j \tilde{a}_k \frac{1}{\lambda_j + \lambda_k} \left(1 - e^{-(\lambda_j + \lambda_k) T} \right) \int_{\Gamma_0} \mathcal{C}w_j \, \mathcal{C}w_k \; d\Gamma \ . \end{aligned}$$

Let us define what turns out to be the limit operator,

$$\langle \Pi_\infty v^0, \tilde{v}^0 \rangle = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_j \tilde{a}_k \frac{1}{\lambda_j + \lambda_k} \int_{\Gamma_0} \mathcal{C}w_j \, \mathcal{C}w_k \; d\Gamma \ .$$

This operator is bounded since

$$\begin{aligned} |\langle \Pi_\infty v^0, \tilde{v}^0 \rangle| &\le \frac{1}{2\lambda_1} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} |a_j| |\tilde{a}_k| \left| \int_{\Gamma_0} \mathcal{C}w_j \, \mathcal{C}w_k \; d\Gamma \right| \\ &\le \frac{K}{2\lambda_1} \left(\sum_{j=1}^{\infty} \lambda_j |a_j|^2 \right)^{1/2} \left(\sum_{k=1}^{\infty} \lambda_k |\tilde{a}_k|^2 \right)^{1/2} \\ &= \frac{K}{2\lambda_1} \|v^0\|_{\widetilde{H}}^2 \|\tilde{v}^0\|_{\widetilde{H}}^2 \ , \end{aligned}$$

using the assumption (5.6), and thus well defined. We can now prove the following theorem.

Theorem 5.2.1. *Under the assumptions (5.4), (5.5) and (5.6) we have*

$$\Lambda_T \to \Pi_\infty \quad \text{in the operator norm as } T \to \infty \,. \tag{5.7}$$

Proof. We have that

$$\|\Lambda_T - \Pi_\infty\| = \sup_{v^0, \tilde{v}^0 \in \widetilde{H} \setminus \{0\}} \frac{\left| \left\langle (\Lambda_T - \Pi_\infty) v^0, \tilde{v}^0 \right\rangle_{\widetilde{H}' \times \widetilde{H}} \right|}{\|v^0\|_{\widetilde{H}} \|\tilde{v}^0\|_{\widetilde{H}}} \,,$$

where

$$\begin{aligned}\left| \langle \Pi_\infty v^0, \tilde{v}^0 \rangle - \langle \Lambda_T v^0, \tilde{v}^0 \rangle \right| &= \left| \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_j \tilde{a}_k \frac{e^{-(\lambda_j + \lambda_k)T}}{\lambda_j + \lambda_k} \int_{\Gamma_0} \mathcal{C} w_j \, \mathcal{C} w_k \, d\Gamma \right| \\ &\le \frac{K e^{-2T\lambda_1}}{2\lambda_1} \|v^0\|_{\widetilde{H}}^2 \|\tilde{v}^0\|_{\widetilde{H}}^2 \,.\end{aligned}$$

□

Note how the fact that all eigenvalues λ_k were real and strictly greater than zero, was exactly what made the preceeding result possible. Note also that the convergence in (5.7) was in the operator norm, which is quite strong.

Assume that we are given a specific null-controllability problem for the heat equation. If now the limit operator Π_∞ is invertible for these particular data, we can then compute a control function, to which controls, corresponding to increasing values of T, will converge. It would similarly imply that the norm of the computed controls will converge to a certain level as $T \to \infty$.

5.2.2 The Wave Equation

We now focus on the controllability operator for the wave equation. Recall that in this case the adjoint system has the form

$$\begin{cases} v_{tt} = \Delta v & \text{in } Q \,, \\ \mathcal{B} v = 0 & \text{in } \Sigma \,, \\ v(T) = y^0, \quad v_t(T) = \overline{y}^0 & \text{in } \Omega \,, \end{cases} \tag{5.8}$$

for some fixed control time $T > 0$. Note that the boundary operator $\mathcal{B}$, as introduced in Chapter 2, was defined on the "first order state", which here is $(v(t), v_t(t))$. For easier notation, $\mathcal{B}$ is in this case defined on $v(t)$ only. This reduces the generality slightly, but the above formulation still includes the most important cases. Likewise, the complementary boundary operator $\mathcal{C}$ is defined only on $v(t)$.

We now set $\widetilde{H} = \widetilde{H}_1 \times \widetilde{H}_2$ with $\widetilde{H}_2 = L^2(\Omega)$, where $\widetilde{H}$ is a space in which the wave equation (5.8) is well posed (recall that $\langle w_k, w_l \rangle_{\widetilde{H}_2} = \delta_{kl}$). We have

$$\|v^0\|^2_{\widetilde{H}} = \|(y^0, \overline{y}^0)\|^2_{\widetilde{H}} = \|y^0\|^2_{\widetilde{H}_1} + \|\overline{y}^0\|^2_{\widetilde{H}_2} \qquad \text{for all } v^0 = (y^0, \overline{y}^0) \in \widetilde{H}\,.$$

Given initial conditions of the system (5.8),

$$y^0 = \sum_{k=1}^{\infty} a_k w_k, \qquad \overline{y}^0 = \sum_{k=1}^{\infty} b_k w_k\,,$$

where $\langle \mu_k a_k \rangle, \langle b_k \rangle \in \ell^2$, one can easily verify that

$$\big\|(u(y), u_t(t))\big\|^2_{\widetilde{H}} = \sum_{k=1}^{\infty} (\mu_k^2 a_k^2 + b_k^2)\,,$$

for all $t \in \mathbb{R}$. Although we operate in real Hilbert spaces, it will prove convenient to write the solution of the system (5.8) using complex exponentials. Let the initial conditions $v^0 = (y^0, \overline{y}^0)$ be represented by the complex sequence $\langle c_k \rangle_{k=1}^{\infty}$ where we set $c_k = \frac{1}{2}(a_k - i b_k/\mu_k)$, $k \in \mathbb{N}$. From the assumptions above we see that this implies $\langle \mu_k c_k \rangle \in \ell^2$. The solution corresponding to v^0 can now be written

$$v(t,x) = \sum_{k=1}^{\infty} \big(c_k e^{i\mu_k t} + \overline{c_k} e^{-i\mu_k t}\big) w_k(x) = 2 \sum_{k=1}^{\infty} \mathrm{Re}\big(c_k e^{i\mu_k t}\big) w_k(x)\,. \tag{5.9}$$

In a similar way we will let the initial conditions $\tilde{v}^0$ be represented by the complex sequence $\langle \tilde{c}_k \rangle_{k=1}^{\infty}$.

Let us finally introduce the projection operator $P_N : \widetilde{H} \mapsto \widetilde{H}$ by

$$P_N \left(\sum_{k=1}^{\infty} a_k w_k, \sum_{k=1}^{\infty} b_k w_k \right) = \left(\sum_{k=1}^{N} a_k w_k, \sum_{k=1}^{N} b_k w_k \right)\,,$$

for all $N \in \mathbb{N}$ and sequences $\langle a_k \rangle_{k=1}^{\infty}$, $\langle b_k \rangle_{k=1}^{\infty}$ for which $\langle \mu_k a_k \rangle, \langle b_k \rangle \in \ell^2$. So P_N does a simple spectral truncation of the initial conditions, leading to a solution as in (5.9) but with the upper limit of the sums replaced by N.

Recall that we in the general setting of Chapter 2 assume the bound,

$$\int_0^T \int_{\Gamma_0} |\mathcal{C}v|^2 d\Gamma\, dt \le K(T) \|v^0\|^2_{\widetilde{H}}\,, \tag{5.10}$$

where $v(t)$ is a solution of the adjoint system with initial data v^0 (see Equation (2.6), page 12). Consider now the initial conditions $(y^0, \overline{y}^0) = (w_k, 0)$, which leads to the solution $v(t) = \cos(\mu_k t) w_k$. Inserted into the bound (5.10) we get

$$\begin{aligned}\int_0^T \int_{\Gamma_0} |\mathcal{C}v|^2 d\Gamma\, dt &= \int_0^T \int_{\Gamma_0} \big|\cos(\mu_k t) \mathcal{C} w_k\big|^2 d\Gamma\, dt = \int_0^T \cos^2(\mu_k t)\, dt \int_{\Gamma_0} |\mathcal{C} w_k|^2 d\Gamma \\ &= \big(\tfrac{1}{2}T + \tfrac{1}{4\mu_k} \sin(2\mu_k T)\big) \int_{\Gamma_0} |\mathcal{C} w_k|^2 d\Gamma \le K(T) \|w_k\|^2_{\widetilde{H}_1}\,.\end{aligned}$$

If we now choose T sufficiently large, we see that

$$\int_{\Gamma_0} |\mathcal{C}w_k|^2 d\Gamma \leq K\|w_k\|^2_{\widetilde{H}_1} = K\lambda_k\,, \tag{5.11}$$

for all $k \in \mathbb{N}$. So for the wave equation, the relation (5.10) of a well-posed adjoint system implies (5.11).

Let us turn to the operators that we are interested in. First the controllability operator, which can be defined as

$$\langle \Lambda_T v^0, \tilde{v}^0 \rangle = \int_0^T \int_{\Gamma_0} \mathcal{C}v\, \mathcal{C}\tilde{v}\; d\Gamma dt\,,$$

for all $v^0, \tilde{v}^0 \in \widetilde{H}$. Using (5.10) we see that this operator is bounded,

$$\begin{aligned} \big|\langle \Lambda_T v^0, \tilde{v}^0 \rangle\big| &= \big|\langle \mathcal{C}v, \mathcal{C}\tilde{v} \rangle_{L^2(\Sigma_0)}\big| \leq \|\mathcal{C}v\|_{L^2(\Sigma_0)} \|\mathcal{C}\tilde{v}\|_{L^2(\Sigma_0)} \\ &\leq K(T)\|v^0\|_{\widetilde{H}} \|\tilde{v}^0\|_{\widetilde{H}}\,, \end{aligned}$$

for $v^0, \tilde{v}^0 \in \widetilde{H}$. As we shall see, Λ_T/T is the interesting operator concerning convergence for $T \to \infty$. For convenience, we approach it through a bilinear form,

$$\pi_T(v^0, \tilde{v}^0) = \frac{1}{T} \langle \Lambda_T v^0, \tilde{v}^0 \rangle_{\widetilde{H}' \times \widetilde{H}}\,.$$

In the limit (in a sense that will be made precise), the form π_T turns out to approach the form π,

$$\pi(v^0, \tilde{v}^0) = 2\sum_{j=1}^{\infty} \mathrm{Re}(\overline{c_j}\tilde{c}_j) \int_{\Gamma_0} |\mathcal{C}w_j|^2\, d\Gamma\,.$$

That π is bounded follows from (5.11) and the fact that $\langle \mu_k c_k \rangle, \langle \mu_k \tilde{c}_k \rangle \in \ell^2$. Let the operator $\Pi : \widetilde{H} \mapsto \widetilde{H}'$ be defined by the relation

$$\pi(v^0, \tilde{v}^0) = \langle \Pi v^0, \tilde{v}^0 \rangle_{\widetilde{H}' \times \widetilde{H}} \quad \text{for all } v^0, \tilde{v}^0 \in \widetilde{H}\,.$$

This will be the limit operator for Λ_T/T as $T \to \infty$.

We are now ready to prove the convergence. We break up the proof in smaller parts and start out with three lemmas.

Lemma 5.2.1. *Assume that $K(T)/T \leq K$ for T large enough. Then*

$$\begin{aligned} &\big|\pi_T(v^0, \tilde{v}^0) - \pi_T(P_N v^0, P_N \tilde{v}^0)\big| \\ &\qquad \leq K\Big(\|v^0 - P_N v^0\|_{\widetilde{H}} \|\tilde{v}^0\|_{\widetilde{H}} + \|\tilde{v}^0 - P_N \tilde{v}^0\|_{\widetilde{H}} \|v^0\|_{\widetilde{H}}\Big)\,, \end{aligned}$$

for all $v^0, \tilde{v}^0 \in \widetilde{H}$.

Proof. Note that

$$\pi_T(v^0,\tilde{v}^0) - \pi_T(P_N v^0, P_N\tilde{v}^0) = \pi_T(v^0 - P_N v^0, \tilde{v}^0) + \pi_T(P_N v^0, \tilde{v}^0 - P_N\tilde{v}^0)\,,$$

for $v^0, \tilde{v}^0 \in \widetilde{H}$. We get

$$\begin{aligned}\left|\pi_T(v^0 - P_N v^0, \tilde{v}^0)\right| &= \frac{1}{T}\left|\int_0^T \int_{\Gamma_0} \mathcal{C}(v - P_N v)\, \mathcal{C}\tilde{v}\; d\mathit{\Gamma} dt\right| \\ &\le K\|v^0 - P_N v^0\|_{\widetilde{H}} \|\tilde{v}^0\|_{\widetilde{H}}\,,\end{aligned}$$

for T large enough, and where $v(t)$ is a solution corresponding to the initial conditions $v^0 - P_N v^0$. Similarly,

$$\left|\pi_T(P_N v^0, \tilde{v}^0 - P_N\tilde{v}^0)\right| \le K\|v^0\|_{\widetilde{H}} \|\tilde{v}^0 - P_N\tilde{v}^0\|_{\widetilde{H}}\,.$$

□

The requirement "$K(T)/T \le K$ for T large enough" is resonable. Every bound of the type (5.10) for the wave equation known to the author, is of the form $K(T) = c_1 T + c_2$ for some $c_1, c_2 > 0$. Such bounds are typically proved using *multipliers*, see Section 7.1 for a derivation for the one-dimensional wave equation, and see Section 7.6 for further information on multipliers.

Lemma 5.2.2. *The following bound holds,*

$$\left|\pi_T(P_N v^0, P_N\tilde{v}^0) - \pi(P_N v^0, P_N\tilde{v}^0)\right| \le \frac{\delta_N}{T}\|v^0\|_{\widetilde{H}} \|\tilde{v}^0\|_{\widetilde{H}}\,,$$

for all $v^0, \tilde{v}^0 \in \widetilde{H}$, and where δ_N is a non-decreasing function of N.

Proof. Let $v^0, \tilde{v}^0 \in \widetilde{H}$ and we have

$$\begin{aligned}&\pi_T(P_N v^0, P_N\tilde{v}^0) - \pi(P_N v^0, P_N\tilde{v}^0) \\ &= \frac{1}{T}\int_0^T\int_{\Gamma_0}\left(\sum_{j=1}^N (c_j e^{i\mu_j t} + \overline{c_j} e^{-i\mu_j t})\mathcal{C}w_j\right)\left(\sum_{k=1}^N (\tilde{c}_k e^{i\mu_k t} + \overline{\tilde{c}_k} e^{-i\mu_k t})\mathcal{C}w_k\right) d\mathit{\Gamma} dt \\ &\qquad - 2\sum_{j=1}^{\infty} \mathrm{Re}(\overline{c_j}\tilde{c}_j)\int_{\Gamma_0} |\mathcal{C}w_j|^2\, d\mathit{\Gamma} \\ &= \frac{1}{T}\int_0^T\int_{\Gamma_0}\Bigg[\sum_{\substack{j,k=1\\ j\ne k}}^N c_j\overline{\tilde{c}_k} e^{i(\mu_j-\mu_k)t}\mathcal{C}w_j\mathcal{C}w_k + \sum_{\substack{j,k=1\\ j\ne k}}^N \overline{c_j}\tilde{c}_k e^{i(\mu_k-\mu_j)t}\mathcal{C}w_j\mathcal{C}w_k \\ &\qquad\qquad + \sum_{j,k=1}^N c_j\tilde{c}_k e^{i(\mu_j+\mu_k)t}\mathcal{C}w_j\mathcal{C}w_k + \sum_{j,k=1}^N \overline{c_j}\overline{\tilde{c}_k} e^{-i(\mu_j+\mu_k)t}\mathcal{C}w_j\mathcal{C}w_k\Bigg] d\mathit{\Gamma} dt\,.\end{aligned}$$

Focusing on the first sum of the last expression, we get

$$\int_0^T \int_{\Gamma_0} \sum_{\substack{j,k=1 \\ j\neq k}}^{N} c_j \tilde{c}_k e^{i(\mu_j-\mu_k)t} \mathcal{C}w_j \mathcal{C}w_k d\Gamma dt$$
$$= \sum_{\substack{j,k=1 \\ j\neq k}}^{N} c_j \tilde{c}_k \int_0^T e^{i(\mu_j-\mu_k)t} dt \int_{\Gamma_0} \mathcal{C}w_j \mathcal{C}w_k d\Gamma = I_1 \,.$$

We obtain

$$\left| \int_0^T e^{i(\mu_j-\mu_k)t} dt \right| = \frac{1}{\mu_j - \mu_k} \left| e^{i(\mu_j-\mu_k)T} - 1 \right| \leq \frac{2}{\mu_j - \mu_k} \,.$$

Setting $\delta_N^1 = \max\{2/(\mu_j - \mu_k) \mid j,k = 1,2,\ldots,N,\ j \neq k\}$ and using (5.11) we get

$$|I_1| \leq \delta_N^1 K \|P_N v^0\|_{\widetilde{H}} \|P_N \tilde{v}^0\|_{\widetilde{H}} \leq \delta_N^1 K \|v^0\|_{\widetilde{H}} \|\tilde{v}^0\|_{\widetilde{H}} \,.$$

Analogous bounds can be obtained for the other three sums, leading to the desired result. □

Lemma 5.2.3. *We have*

$$\left|\pi(P_N v^0, P_N \tilde{v}^0) - \pi(v^0, \tilde{v}^0)\right| \leq \|v^0 - P_N v^0\|_{\widetilde{H}} \|\tilde{v}^0 - P_N \tilde{v}^0\|_{\widetilde{H}} \,,$$

for all $v^0, \tilde{v}^0 \in \widetilde{H}$.

Proof. Since π is bounded and

$$\pi(v^0, \tilde{v}^0) - \pi(P_N v^0, P_N \tilde{v}^0) = 2 \sum_{j=N+1}^{\infty} \mathrm{Re}(\overline{c_j} \tilde{c}_j) \int_{\Gamma_0} |\mathcal{C}w_j|^2 \, d\Gamma \,,$$

the result follows. □

We are now ready for the main theorem.

Theorem 5.2.2. *Assume that* $K(T)/T$ *is bounded for* T *large enough. Then*

$$\pi_T(v^0, \tilde{v}^0) \to \pi(v^0, \tilde{v}^0) \quad \textit{as } T \to \infty \,,$$

for all $v^0, \tilde{v}^0 \in \widetilde{H}$, *or equivalently*

$$\tfrac{1}{T} \Lambda_T v^0 \rightharpoonup \Pi v^0 \quad \textit{weakly in } \widetilde{H}' \textit{ as } T \to \infty \,, \tag{5.12}$$

for all $v^0 \in \widetilde{H}$.

Proof. Let $v^0, \tilde{v}^0 \in \widetilde{H}$. From Lemmas 5.2.1, 5.2.2 and 5.2.3 we get

$$\begin{aligned}
\left|\pi_T(v^0,\tilde{v}^0) - \pi(v^0,\tilde{v}^0)\right| &\le \left|\pi_T(v^0,\tilde{v}^0) - \pi_T(P_N v^0, P_N\tilde{v}^0)\right| \\
&\quad + \left|\pi_T(P_N v^0, P_N\tilde{v}^0) - \pi(P_N v^0, P_N\tilde{v}^0)\right| \\
&\quad + \left|\pi(P_N v^0, P_N\tilde{v}^0) - \pi(v^0,\tilde{v}^0)\right| \\
&\le K\Big(\|v^0 - P_N v^0\|_{\widetilde{H}}\|\tilde{v}^0\|_{\widetilde{H}} + \|v^0\|_{\widetilde{H}}\|\tilde{v}^0 - P_N\tilde{v}^0\|_{\widetilde{H}}\Big) \\
&\quad + \frac{\delta_N}{T}\|v^0\|_{\widetilde{H}}\|\tilde{v}^0\|_{\widetilde{H}} + \|v^0 - P_N v^0\|_{\widetilde{H}}\|\tilde{v}^0 - P_N\tilde{v}^0\|_{\widetilde{H}} \\
&\le I_N^1 + I_N^2/T + I_N^3\,.
\end{aligned}$$

Let $\epsilon > 0$ be an arbitrary (small) number. Choose now N so large that $I_N^1 \le \epsilon/3$ and $I_N^3 \le \epsilon/3$. Fixing this N, we choose T such that $I_N^2/T \le \epsilon/3$. This means that

$$\left|\pi_T(v^0,\tilde{v}^0) - \pi(v^0,\tilde{v}^0)\right| \le \epsilon\,.$$

□

Note that the limit of the above theorem suggests that the norm of controls, for increasing values of T, will be approximately proportional to $1/T$.

But what kind of operator is Π? Does it have a simple, closed form? When the eigenvectors of the Laplace operator satisfy a special relation, it actually has a very simple form.

Theorem 5.2.3. *Given the relation*

$$\int_{\Gamma_0} |\mathcal{C}w_k|^2 d\Gamma = K\lambda_k\,, \tag{5.13}$$

for all $k \in \mathbb{N}$, we have

$$\Pi = \tfrac{K}{2}\begin{bmatrix} -\Delta & 0 \\ 0 & I \end{bmatrix}.$$

Proof. Let $v^0 = (y^0, \overline{y}^0) = (\sum_{k=1}^\infty a_k w_k, \sum_{k=1}^\infty b_k w_k)$ and $c_k = \frac{1}{2}(a_k - i b_k/\mu_k)$. Let analogous relations hold for $\tilde{v}^0$, $\tilde{a}_k$, $\tilde{b}_k$ and $\tilde{c}_k$. Then

$$\begin{aligned}
\langle \Pi v^0, \tilde{v}^0\rangle &= 2\sum_{k=1}^\infty \mathrm{Re}(\overline{c_k}\tilde{c}_k)\int_{\Gamma_0} |\mathcal{C}w_k|^2\, d\Gamma = 2K\sum_{k=1}^\infty \mathrm{Re}(\overline{c_k}\tilde{c}_k)\mu_k^2 \\
&= \tfrac{K}{2}\sum_{k=1}^\infty \big(a_k\tilde{a}_k + b_k\tilde{b}_k/\mu_k^2\big)\mu_k^2 \\
&= \tfrac{K}{2}\left\langle \left(\sum_{k=1}^\infty \mu_k^2 a_k w_k, \sum_{k=1}^\infty b_k w_k\right), \left(\sum_{k=1}^\infty \tilde{a}_k w_k, \sum_{k=1}^\infty \tilde{b}_k w_k\right)\right\rangle_{\widetilde{H}'\times\widetilde{H}} \\
&= \tfrac{K}{2}\left\langle \begin{bmatrix} -\Delta & 0 \\ 0 & I \end{bmatrix} v^0, \tilde{v}^0\right\rangle_{\widetilde{H}'\times\widetilde{H}}.
\end{aligned}$$

□

But when does the relation (5.13) hold? The following section shows that it holds, at least, for Dirichlet control on the whole boundary of some very interesting domains.

5.2.2.1 A Special Case

Let us consider Dirichlet control on the whole boundary,

$$\Gamma_0 = \Gamma, \quad \mathcal{B} = I \quad \text{and} \quad \mathcal{C} = -\frac{\partial}{\partial n} \quad \text{on } \Gamma\,.$$

In this setting, the eigensolutions from (5.4) become

$$\begin{cases} -\Delta w_k = \lambda_k w_k\,, \\ w_k|_\Gamma = 0\,, \\ \langle w_k, w_l\rangle_{L^2(\Omega)} = \delta_{kl}\,, \end{cases}$$

for $k, l \in \mathbb{N}$. Note then the important relation, using Green's Theorem,

$$\begin{aligned} \int_\Omega \nabla w_j \cdot \nabla w_k dx &= \int_\Gamma \frac{\partial w_j}{\partial n} w_k d\mathit{\Gamma} - \int_\Omega \Delta w_j w_k dx = \lambda_j \int_\Omega w_j w_k dx \\ &= \delta_{jk}\lambda_j\,. \end{aligned} \tag{5.14}$$

We now have the following interesting theorem (the same result and a sketchy proof can be found in *Bensoussan, 1990*, page 214).

Theorem 5.2.4. *Let $\Omega \subset \mathbb{R}^d$, $\Gamma = \partial\Omega$ and $m(x) = x - x_0$, where $x_0 \in \mathbb{R}^d$. Then*

$$\int_\Gamma \left|\frac{\partial w_k}{\partial n}\right|^2 m \cdot n\, d\mathit{\Gamma} = 2\lambda_k\,.$$

Proof. Note first that since $w_k = 0$ on the boundary, the gradient $\nabla w_k(x)$ and the outward normal $n(x)$ point in the same direction for any $x \in \Gamma$. This leads to

$$\begin{aligned} 2\int_\Gamma \left|\frac{\partial w_k}{\partial n}\right|^2 m \cdot n\, d\mathit{\Gamma} &= 2\int_\Gamma \frac{\partial w_k}{\partial n} \nabla w_k \cdot m\, d\mathit{\Gamma} \\ &= 2\int_\Omega \Delta w_k \nabla w_k \cdot m\, dx + 2\int_\Omega \nabla w_k \cdot \nabla(\nabla w_k \cdot m) dx = I_1 + I_2\,. \end{aligned} \tag{5.15}$$

We now get

$$I_1 = -\lambda_k \int_\Omega m \cdot \nabla(w_k^2) dx = -\lambda_k \int_\Gamma m \cdot n\, w_k^2 d\mathit{\Gamma} + \lambda_k \int_\Omega \nabla \cdot m\, w_k^2 dx = d\lambda_k\,.$$

We rewrite I_2 as follows,

$$\begin{aligned}
I_2 &= 2\sum_{\alpha,\beta=1}^{d}\int_\Omega \frac{\partial w_k}{\partial x_\alpha}\frac{\partial}{\partial x_\alpha}\left(\frac{\partial w_k}{\partial x_\beta}m_\beta\right)dx \\
&= 2\sum_{\alpha,\beta=1}^{d}\int_\Omega \left(\frac{\partial w_k}{\partial x_\alpha}\frac{\partial^2 w_k}{\partial x_\alpha x_\beta}m_\beta + \frac{\partial w_k}{\partial x_\alpha}\frac{\partial w_k}{\partial x_\beta}\frac{\partial m_\beta}{\partial x_\alpha}\right)dx \\
&= \int_\Omega m\cdot\nabla(\nabla w_k\cdot\nabla w_k)dx + 2\int_\Omega \nabla w_k\cdot\nabla w_k dx = I_3 + 2\lambda_k\,,
\end{aligned}$$

using (5.14) and the fact that $\partial m_\alpha/\partial x_\beta = \delta_{\alpha\beta}$. Proceeding, we get

$$\begin{aligned}
I_3 &= \int_\Gamma m\cdot n\,\nabla w_k\cdot\nabla w_k d\mathit{\Gamma} - \int_\Omega \nabla\cdot m\,\nabla w_k\cdot\nabla w_k dx \\
&= \int_\Gamma \left|\frac{\partial w_k}{\partial n}\right|^2 m\cdot n\,d\mathit{\Gamma} - d\lambda_k\,.
\end{aligned}$$

Inserting the obtained expressions into (5.15), we get

$$2\int_\Gamma \left|\frac{\partial w_k}{\partial n}\right|^2 m\cdot n\,d\mathit{\Gamma} = d\lambda_k + \int_\Gamma \left|\frac{\partial w_k}{\partial n}\right|^2 m\cdot n\,d\mathit{\Gamma} - d\lambda_k + 2\lambda_k\,,$$

leading to the desired result. □

Combining Theorems 5.2.3 and 5.2.4, we easily deduce the following theorem.

Theorem 5.2.5. *Let $\Omega\subset\mathbb{R}^d$, $\Gamma=\partial\Omega$ and $m(x)=x-x_0$, where $x_0\in\mathbb{R}^d$. If*

$$m(x)\cdot n(x) = C_n, \quad \textit{for almost all } x\in\Gamma\,, \tag{5.16}$$

for a constant $C_n>0$, then

$$\Pi = \frac{1}{C_n}\begin{bmatrix}-\Delta & 0\\ 0 & I\end{bmatrix}.$$

The "almost all" in (5.16) means that the property is allowed to fail in a set of measure zero (for instance, it is ok if the property fails in a finite number of points).

This result has previously been mentioned in *Glowinski, Li, and Lions (1990)*, page 6, and in *Glowinski and Lions (1995)*, page 257, but without proof.

5.2.2.2 Domains of Constant Normal Width

How do the domains look for which $m\cdot n = C_n$ on the whole boundary?

In one dimension, the characterization is trivial. If $\Omega=(a,b)$ choose $x^0 = (a+b)/2$ and we have $C_n=(b-a)/2$.

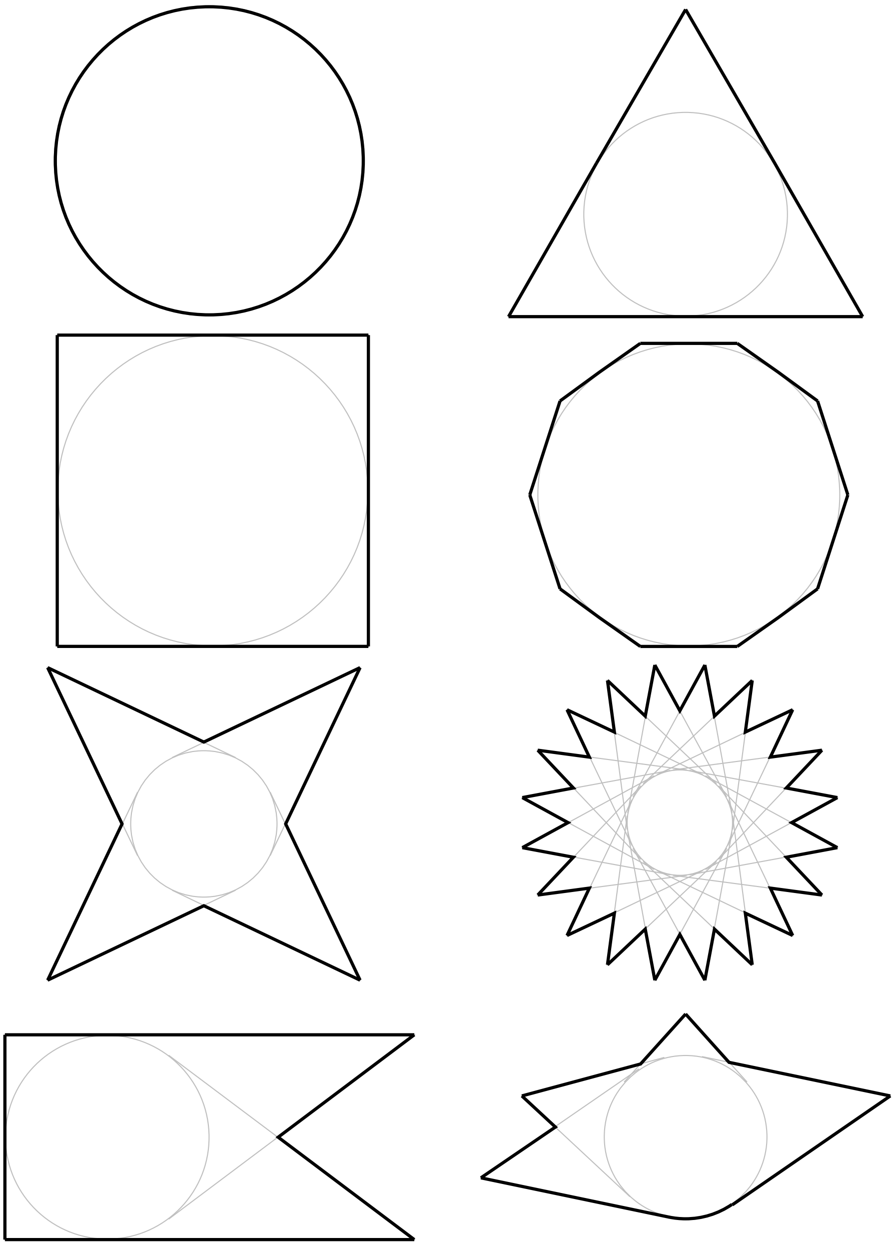


Figure 5.1: A selection of various domains having constant normal width, that is, a point $x^0 \in \mathbb{R}^2$ exists such that $(x - x^0) \cdot n(x) = C_n$ holds for almost all boundary points. For each of the shown domains, x^0 is the center of the circle and C_n is equal to the radius of that circle. (The black lines bound the actual domains, the gray lines and gray circle indicate how to easily construct such domains.)

In two dimensions, the domains having this property are much more interesting. Figure 5.1 shows a selection of such domains, including visual indications of how to construct them.

In three dimensions or more, these domains become even more versatile. We will, however, not try to give an exhaustive characterization.

CHAPTER 6

The Heat Equation

If you can't stand the heat, get out of the kitchen.
— HARRY VAUGHAN, 1952

We will in this chapter, and the following two, focus on specific control systems. Different parts of the theory of the previous chapters will be applied in these case studies.

The heat equation is a parabolic PDE that describes the distribution of temperature in some object, as it depends on time. As time goes by, the temperature profile in the object becomes very smooth (this was illustrated in the Introduction, see Figure 1.3) and only null-controllability will be possible.

We consider as control system the heat equation in the domain Ω with Dirichlet control on a part of the boundary, Γ_0,

$$\begin{cases} u_t = \Delta u & \text{in } Q \ , \\ u = \begin{cases} k \\ 0 \end{cases} & \begin{array}{l} \text{in } \Sigma_0 \ , \\ \text{in } \Sigma \setminus \Sigma_0 \ , \end{array} \\ u(0) = u^0 & \text{in } \Omega \ , \end{cases} \tag{6.1}$$

with $k \in L^2(\Sigma_0)$ and $u^0 \in H' = H^{-1}(\Omega)$ (recall that $\Sigma_0 = (0,T) \times \Gamma_0$). We shall see shortly how to show that this formulation leads to a well-posed system with solution $u(t) \in H'$.

6.1 Well-posedness

Let us set up the different maps and quantities of Chapter 2 as it applied to this case. At the same time, we must argue that the assumptions of that chapter are fulfilled.

We first introduce the adjoint system,

$$\begin{cases} v_t = -\Delta v & \text{in } Q\ , \\ v = 0 & \text{in } \Sigma\ , \\ v(T) = v^0 & \text{in } \Omega\ , \end{cases}$$

with $v(t) \in \widetilde{H} = H_0^1(\Omega)$. In the notation of Chapter 2 we have

$$\mathcal{A} = \Delta, \quad \mathcal{B} = I, \quad \widetilde{\mathcal{A}} = \Delta\,.$$

By observing

$$\langle \mathcal{A}u, v\rangle = \langle \Delta u, v\rangle_{H'\times H} = \int_\Omega \Delta u\, v\, dx = \int_\Omega u\,\Delta v\, dx = \langle u, \widetilde{\mathcal{A}}v\rangle\,,$$

when $\mathcal{A}$ is considered with homogeneous boundary conditions, we see that $\widetilde{\mathcal{A}} = \mathcal{A}^*$ and thus $\mathcal{M} = 1$. What we need as the last thing is the complementary boundary operator $\mathcal{C}$. With $u, v \in C^\infty(\overline{\Omega})$ we include boundary conditions and get

$$\begin{aligned} \{\mathcal{A}u, v\} - \{u, \widetilde{\mathcal{A}}v\} &= \int_\Omega (\Delta u v - u \Delta v) dx = \int_\Gamma \left(\frac{\partial u}{\partial n} v - u \frac{\partial v}{\partial n}\right) d\Gamma \\ &= \int_\Gamma u \left(-\frac{\partial v}{\partial n}\right) d\Gamma\,. \end{aligned} \tag{6.2}$$

This shows that $\mathcal{C} = -\partial/\partial n$, the negated normal derivative.

We show well-posedness of the adjoint system first. We introduce the energy of the adjoint system at time t as

$$E(t) = \tfrac{1}{2} \int_\Omega |\nabla v|^2 dx\,.$$

By differentiating the energy,

$$\begin{aligned} E'(t) &= \int_\Omega \nabla v \cdot \nabla \dot{v} dx = -\int_\Omega \nabla v \cdot \nabla(\Delta v) dx = \int_\Omega |\Delta v|^2 dx - \int_\Gamma \frac{\partial v}{\partial n} \Delta v\, d\Gamma \\ &= \int_\Omega |\Delta v|^2 dx\,, \end{aligned}$$

we see that $E'(t) \geq 0$ and so the energy can only *decrease* as we solve the adjoint system backwards in time,

$$0 \leq E(t) \leq E(T), \quad \text{for } 0 \leq t \leq T\,.$$

We finally observe that $E(t) < \infty$ if and only if the corresponding solution $v(t) \in H_0^1(\Omega)$, and the well-posedness of the adjoint system is shown.

To show well-posedness of the control system we first need to establish that applying the complementary boundary operator $\mathcal{C}$ to a solution $v(t)$ of the adjoint system lies in $L^2(\Sigma_0)$,

$$\|\mathcal{C}v(\cdot)\|^2_{L^2(\Sigma_0)} = \int_0^T \int_{\Gamma_0} \left|\frac{\partial v}{\partial n}\right|^2 d\Gamma dt \le K(T)\|v^0\|^2_{H_0^1(\Omega)} , \tag{6.3}$$

for all $v^0 \in H_0^1(0,1)$ with corresponding solution v of the adjoint system. This inequality is typically shown using *multipliers*, and we will prove it for the simple one dimensional case of $\Omega = (0,1)$ and $\Gamma_0 = \{1\}$. To show it for more general domains, some regularity assumptions must be made about Ω, see *Lions (1985)* and *Komornik (1994)* for further details about the multiplier method (these references discuss the wave equation, but the principles are also applicable to the heat equation).

To show (6.3) in the 1D case, we proceed by applying the multiplier $v_x x$ to the (negative) right hand side of the heat equation, $\dot{v} = -v_{xx}$,

$$\int_0^1 v_{xx}v_x x dx = \left[v_x^2 x\right]_0^1 - \int_0^1 v_x(v_{xx}x + v_x)dx = v_x^2(t,1) - \int_0^1 v_x v_{xx} x dx - \int_0^1 v_x^2 dx ,$$

leading to the bound

$$\begin{aligned}\int_0^T v_x^2(t,1)dt &\le 2\left|\int_0^T \int_0^1 v_{xx}v_x x dx dt\right| + \int_0^T\int_0^1 v_x^2 dx dt \\ &\le 2\left(\int_0^T\int_0^1 v_{xx}^2 dx dt\right)^{1/2}\left(\int_0^T\int_0^1 v_x^2 dx dt\right)^{1/2} + \int_0^T\int_0^1 v_x^2 dx dt .\end{aligned}$$

Now using

$$\begin{aligned}\int_0^T\int_0^1 v_{xx}^2 dx dt &= \int_0^T E'(t)dt = E(T) - E(0) \le E(T), \\ \int_0^T\int_0^1 v_x^2 dx dt &= 2\int_0^T E(t)dt \le 2TE(T) ,\end{aligned}$$

we obtain the desired inequality,

$$\int_0^T v_x^2(t,1)dt \le 2(\sqrt{2T} + T)E(T) . \tag{6.4}$$

Let us now repeat some arguments of Section 2.1.2 to show that the control system (6.1) is well posed. Assume that solutions actually exists to the control system for sufficiently smooth initial data u^0 and control k, dense in $H^{-1}(\Omega)$ and $L^2((0,T)\times\Gamma_0)$ respectively. Given any solution v to the adjoint system with initial data v^0 we get

$$\langle u(T), v^0\rangle - \langle u^0, v(0)\rangle = \int_0^T\int_{\Gamma_0} k\frac{\partial v}{\partial n} d\Gamma dt = \langle k, \mathcal{C}v\rangle_{L^2(\Sigma_0)} ,$$

and, in turn,

$$\|u(T)\| = \sup_{v^0\in\widetilde{H}\setminus\{0\}} \frac{|\langle u(T), v^0\rangle|}{\|v^0\|} \le \sup_{v^0\in\widetilde{H}\setminus\{0\}} \frac{|\langle u^0, v(0)\rangle| + |\langle k, \mathcal{C}v\rangle|}{\|v^0\|}$$
$$\le \sup_{v^0\in\widetilde{H}\setminus\{0\}} \frac{\|u^0\|\|v^0\| + K(T)\|k\|\|v^0\|}{\|v^0\|} = \|u^0\| + K(T)\|k\|\,.$$

This shows that if $u^0 \in H^{-1}(\Omega)$ and $k \in L^2((0,T)\times\Gamma_0)$, we will have $u(T) \in H^{-1}(\Omega)$. Note that all of the above could be carried out for any $T > 0$.

The details of well-posedness are not normally shown in such detail in the literature. We do it here to illustrate how it can be done and because it is quite important. Indeed, both the control system and adjoint system are well posed in $L^2(\Omega)$, but the bound in (6.3), with the $H_0^1(\Omega)$-norm replaced by the $L^2(\Omega)$-norm, does not hold in this case.

6.1.1 Other Types of Control Operators

Let us at this point briefly consider another type of boundary control. We set $\mathcal{B}u = \partial u/\partial n - \alpha u$ for some $\alpha \in \mathbb{R}$. This formulation includes both Neumann control and Robin control. The control system becomes

$$\begin{cases} u_t = \Delta u & \text{in } Q\ , \\ \frac{\partial u}{\partial n} - \alpha u = \begin{cases} k \\ 0 \end{cases} & \begin{array}{l}\text{in } \Sigma_0\ , \\ \text{in } \Sigma\setminus\Sigma_0\ ,\end{array} \\ u(0) = u^0 & \text{in } \Omega\ , \end{cases}$$

and the adjoint system,

$$\begin{cases} v_t = -\Delta v & \text{in } Q\ , \\ \frac{\partial v}{\partial n} - \alpha v = 0 & \text{in } \Sigma\ , \\ v(T) = v^0 & \text{in } \Omega\ , \end{cases}$$

where the boundary conditions have been replaced accordingly. To derive the complementary boundary operator, we start out as in (6.2) and get

$$\{\mathcal{A}u, v\} - \{u, \widetilde{\mathcal{A}}v\} = \int_\Omega (\Delta u v - u\Delta v)dx = \int_\Gamma \left(\frac{\partial u}{\partial n}v - u\frac{\partial v}{\partial n}\right) d\Gamma$$
$$= \int_\Gamma \left(\frac{\partial u}{\partial n}v - \alpha u v\right) d\Gamma = \int_{\Gamma_0} \left(\frac{\partial u}{\partial n} - \alpha u\right) v\, d\Gamma\,,$$

using that $\partial v/\partial n = \alpha v$. It shows that we must simply use $\mathcal{C} = I$ for the complementary boundary operator.

This was a simple illustration of how to use another form of boundary control. Of course, well-posedness would have to be shown in this particular case, before the control “machinery” of Chapter 2 could be applied. In the following we return to Dirichlet control.

6.2 Analytical Solution in 1D Using Fourier Series

Let us focus on the case $\Omega = (0,1)$ and $\Gamma_0 = \{1\}$. We wish here to represent the controllability operator Λ_T as an infinite matrix $\mathbf{\Lambda}_T$ using the basis $e_j = \sin(j\pi\cdot)$, $j \in \mathbb{N}$, in $H_0^1(0,1)$ and the same basis $e_i' = e_i$ in $H^{-1}(0,1)$. As seen in Section 5.1, this can be done in two ways. We will go through both methods.

For the direct method we must, given $T > 0$, compute

$$\Lambda_T(v^0) = u(T,\cdot)\,,$$

where $v^0 \in H_0^1(0,1)$ and u is found by first solving

$$\begin{cases} v_t(t,x) = -v_{xx}(t,x), & \text{in } (0,T)\times(0,1)\,,\\ v(t,0) = v(t,1) = 0, & \text{in } (0,T)\,,\\ v(T,x) = v^0(x), & \text{in } (0,1)\,, \end{cases} \tag{6.5}$$

followed by

$$\begin{cases} u_t(t,x) = u_{xx}(t,x), & \text{in } (0,T)\times(0,1),\\ u(t,0) = 0, \quad u(t,1) = -v_x(t,1), & \text{in } (0,T),\\ u(0,x) = 0, & \text{in } (0,1)\,. \end{cases} \tag{6.6}$$

We will now apply the Λ_T-map to an arbitrary basis vector,

$$v^0(x) = e_j(x) = \sin(j\pi x), \quad j \in \mathbb{N}\,.$$

The solution to (6.5) is clearly

$$v(t,x) = e^{-j^2\pi^2(T-t)}\sin(j\pi x)\,,$$

and thus

$$v_x(t,1) = (-1)^j j\pi e^{-j^2\pi^2(T-t)}\,. \tag{6.7}$$

We now wish to solve (6.6) given these boundary conditions. These are satisfied if we set

$$u(t,x) = (-1)^{j+1} j\pi e^{-j^2\pi^2(T-t)}x + \sum_{i=1}^{\infty} a_i(t)\sin(i\pi x)\,, \tag{6.8}$$

where the real functions $a_1(t)$, $a_2(t)$, ..., are to be determined. However, using $u_t = u_{xx}$ and the initial condition $u(0,x) = 0$ we get

$$\begin{cases} a_i'(t) + i^2\pi^2 a_i(t) = (-1)^{i+j+1}\dfrac{2i^3\pi^2}{j}e^{-j^2\pi^2(T-t)},\\ a_i(0) = (-1)^{i+j+1}\dfrac{2j}{i}e^{-j^2\pi^2 T}\,, \end{cases}$$

for each $i \in \mathbb{N}$. For fixed i, this is an ordinary differential equation with the solution

$$a_i(t) = (-1)^{i+j+1}\frac{2k}{i(i^2+j^2)}e^{-j^2\pi^2 T}\left(i^2 e^{-i^2\pi^2 t} + j^2 e^{j^2\pi^2 t}\right) \quad \text{for all } i \in \mathbb{N}\,.$$

Inserting this into (6.8) while writing x in the $\langle e_i' \rangle$ basis, we finally get

$$u(T,x) = \sum_{i=1}^{\infty} (-1)^{i+j} \frac{2ij}{i^2+j^2} \left(1 - e^{-(i^2+j^2)\pi^2 T}\right) \sin(i\pi x)\,.$$

So the (i,j)th entry of $\mathbf{\Lambda}_T$ is simply

$$\mathbf{\Lambda}_T(i,j) = (-1)^{i+j} \frac{2ij}{i^2+j^2} \left(1 - e^{-(i^2+j^2)\pi^2 T}\right).$$

For the inner product method, we let v and $\tilde{v}$ be two solutions of the adjoint system with initial conditions e_j and e_i, respectively. Note that we have $\langle e_j', e_i \rangle = 1/2\ \delta_{ij}$, but we will deal with the non-uniform scaling afterwards (although the basis $\langle e_j' \rangle$ does not appear in our calculations for the inner product method at all, it is still the basis for the range of the map). We now have, cf. (6.7),

$$\begin{aligned}\langle \Lambda_T e_j, e_i \rangle &= \int_0^T v_x(t,1)\tilde{v}_x(t,1)dt = \int_0^T (-1)^{i+j} ij\pi^2 e^{-(i^2+j^2)\pi^2 t} dt \\ &= (-1)^{i+j} \frac{ij}{i^2+j^2} \left(1 - e^{-(i^2+j^2)\pi^2 T}\right).\end{aligned}$$

To finish, we just have to take care of the scaling and we end up with,

$$\mathbf{\Lambda}_T(i,j) = \frac{\langle \Lambda_T e_j, e_i \rangle}{\langle e_i', e_i \rangle_{L^2(0,1)}} = (-1)^{i+j} \frac{2ij}{i^2+j^2} \left(1 - e^{-(i^2+j^2)\pi^2 T}\right), \tag{6.9}$$

naturally the same as for the direct method. Note how the term $e^{[\cdots]}$ will vanish as $T \to \infty$, fitting well with the result of Theorem 5.2.1, concerning the limit controllability operator for the heat equation.

An analytical representation of the controllability operator for the heat equation in one dimension, has not previously been seen in the literature. As just seen above, the representation (6.9) provides insight into the properties of the controllability operator, and it would also be useful for computing approximate controls of any accuracy (by truncating the infinite matrix into a sufficiently large, but finite, matrix).

6.3 Null-controllability in 1D

We wish to show null-controllability for the heat equation in one dimension. If one could argue that the infinite matrix of the previous section was invertible, we would be done. That seems like a difficult task, though.

We can also use Theorem 2.3.1, page 19. This involves showing an observability inequality for the adjoint system. Let initial conditions for the adjoint system be given as

$$v(T,x) = v^0(x) = \sum_{j=1}^{\infty} a_j \sin(j\pi x), \qquad 0 \le x \le 1\,,$$

where $\langle ja_j\rangle_{j=1}^{\infty} \in \ell^2$ such that $v^0 \in H_0^1(0,1)$. The solution now becomes

$$v(t,x) = \sum_{j=1}^{\infty} a_j e^{-j^2\pi^2(T-t)} \sin(j\pi x), \qquad 0 \le t \le T, \quad 0 \le x \le 1\,.$$

We have

$$\int_0^T v_x^2(t,1)dt \ge \int_{T/2}^T v_x^2(t,1)dt = \int_0^{T/2} \left(\sum_{j=1}^{\infty} (-1)^j j\pi a_j e^{-j^2\pi^2 t} \right) dt\,.$$

This brings us into a position where we can use Theorem 4.2.3 (page 77), the parabolic version of Ingham's Theorem. We get that a constant $C_p > 0$, independent of the coefficients $\langle a_j\rangle_{j=1}^{\infty}$, exists such that

$$\begin{aligned}
\int_0^T v_x^2(t,1)dt &\ge \int_0^{T/2} \left(\sum_{j=1}^{\infty} (-1)^j j\pi a_j e^{-j^2\pi^2 t} \right) dt \\
&\ge \frac{C_p}{\sum_{j=1}^{\infty} (j\pi)^{-2}} \sum_{j=1}^{\infty} \frac{e^{-j^2\pi^2 T}}{j^2\pi^2} j^2\pi^2 |a_j|^2 \\
&= 6C_p \sum_{j=1}^{\infty} e^{-2j^2\pi^2 T} j^2\pi^2 \frac{e^{j^2\pi^2 T}}{j^2\pi^2} |a_j|^2 \\
&\ge 6C_p \frac{e^{\pi^2 T}}{\pi^2} \sum_{j=1}^{\infty} e^{-2j^2\pi^2 T} j^2\pi^2 |a_j|^2 \ge \frac{6}{\pi^2} C_p \|v(0,\cdot)\|_{H_0^1(0,1)}^2\,,
\end{aligned}$$

for all $T > 0$.

This proves the null-controllability of the heat equation in one dimension. See *López and Zuazua (1998)* and *López and Zuazua (2002)* for similar null-controllability results related to the heat equation in one dimension.

6.4 Uniform Observability of a Semi-discretization

Let us consider the following (family of) semi-discretizations of the heat equation in one dimension,

$$\begin{cases} \boldsymbol{C}_\alpha \dot{\boldsymbol{u}}(t) = \boldsymbol{A}\boldsymbol{u}(t) + \boldsymbol{B}\boldsymbol{k}(t)\,, \\ \boldsymbol{u}(0) = \boldsymbol{u}^0\,, \end{cases}$$

where we use

$$\boldsymbol{C}_\alpha = h \begin{bmatrix} 1-2\alpha & \alpha & & & \\ \alpha & 1-2\alpha & \alpha & & \\ & \ddots & \ddots & \ddots & \\ & & & \alpha & 1-2\alpha \end{bmatrix},$$

for $0 \le \alpha \le 1/4$ and

$$\boldsymbol{A} = \frac{1}{h}\begin{bmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & -2 \end{bmatrix}, \quad \boldsymbol{B} = \frac{1}{h}\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

The matrix $\boldsymbol{B}$ takes care of the Dirichlet boundary conditions at the right end-point by being the natural "next column" of $\boldsymbol{A}$.

The eigenvalues λ_k^α and corresponding eigenvectors $\boldsymbol{w}_k$ of $\boldsymbol{C}_\alpha^{-1}\boldsymbol{A}$ are

$$\lambda_k^\alpha = -\frac{4\sin^2(\frac{1}{2}k\pi h)}{h^2(1-4\alpha\sin^2(\frac{1}{2}k\pi h))},$$

$$\boldsymbol{w}_k(j) = \sin(jkh\pi), \qquad j = 1, 2, \ldots, N,$$

for $k = 1, 2, \ldots, N$, as described in Section 3.1.

The adjoint system has the appearance,

$$\begin{cases} \boldsymbol{C}_\alpha \dot{\boldsymbol{v}}(t) = -\boldsymbol{A}\boldsymbol{v}(t), \\ \quad \boldsymbol{v}(T) = \boldsymbol{v}^0, \end{cases}$$

for which we will use the norm $\|\boldsymbol{v}\|_{\boldsymbol{Q}}^2 = -\langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{v}\rangle = -\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}$.

We aim to prove an observability inequality, similar to the continuous system in the previous section, which is uniform in N (the number N denotes the discretization level and is the vector length of $\boldsymbol{u}(t)$ and $\boldsymbol{v}(t)$). So we need to study solutions of the adjoint system. The initial conditions will be set as

$$\boldsymbol{v}^0 = \sum_{k=1}^{N} a_k \boldsymbol{w}_k,$$

where $\langle k a_k \rangle \in \ell^2$, leading to solutions of the form

$$\boldsymbol{v}(t) = \sum_{k=1}^{N} a_k e^{\lambda_k^\alpha (T-t)} \boldsymbol{w}_k.$$

One of the elements of the observability inequality is the norm of the solution at time $t = 0$,

$$\|\boldsymbol{v}(0)\|_{\boldsymbol{Q}}^2 = -\sum_{k=1}^{N} |a_k|^2 e^{\lambda_k^\alpha T}\langle \boldsymbol{w}_k, \boldsymbol{A}\boldsymbol{w}_k\rangle = \sum_{k=1}^{N} \frac{2\sin^2(\frac{1}{2}kh\pi)}{h^2} e^{2\lambda_k^\alpha T}|a_k|^2. \tag{6.10}$$

We furthermore have

$$
\begin{aligned}
\int_0^T \left|\boldsymbol{B}^T \boldsymbol{v}(t)\right|^2 dt &= \int_0^T \left|\frac{\boldsymbol{v}(t)(N)}{h}\right|^2 dt \\
&= \int_0^T \left|\sum_{k=1}^N \frac{(-1)^k \sin(kh\pi)}{h} e^{\lambda_k^\alpha (T-t)} a_k\right|^2 dt \qquad (6.11)\\
&\geq \int_0^{T/2} \left|\sum_{k=1}^N \frac{(-1)^k \sin(kh\pi)}{h} e^{\lambda_k^\alpha t} a_k\right|^2 dt\,.
\end{aligned}
$$

In order to apply Theorem 4.2.3, the parabolic version of Ingham's Theorem, we need to show that there is a *uniform* gap (uniform in both the index k and discretization level N) among the eigenvalues λ_k^α. First we realize that the smallest gaps occur when $\alpha = 0$,

$$
\begin{aligned}
-(\lambda_{k+1}^\alpha - \lambda_k^\alpha) &= \frac{4\sin^2(\frac{1}{2}(k+1)h\pi)}{h^2(1-4\alpha\sin^2(\frac{1}{2}(k+1)h\pi))} - \frac{4\sin^2(\frac{1}{2}kh\pi)}{h^2(1-4\alpha\sin^2(\frac{1}{2}kh\pi))} \\
&\geq \frac{-1}{1-4\alpha\sin^2(\frac{1}{2}k\pi h)}\left(\lambda_{k+1}^0 - \lambda_k^0\right) \geq -\left(\lambda_{k+1}^0 - \lambda_k^0\right).
\end{aligned}
$$

We now get

$$
\begin{aligned}
-\left(\lambda_{k+1}^0 - \lambda_k^0\right) &= \frac{4}{h^2}\left(\sin^2(\tfrac{1}{2}(k+1)h\pi) - \sin^2(\tfrac{1}{2}kh\pi)\right) = \frac{4}{h^2}(\tfrac{1}{2}\pi h)\sin(2\xi) \\
&\geq 2\pi^2\frac{\sin(\pi h)}{\pi h} \geq 2\pi^2 \frac{\sin(\pi/6)}{\pi/6} = 6\pi\,,
\end{aligned}
$$

with $\frac{1}{2}k\pi h \leq \xi \leq \frac{1}{2}(k+1)\pi h$ by the Mean Value Theorem, and for all $N \geq 5$. A uniform gap has thus been established. Since we also have

$$
-\lambda_k^\alpha \geq \frac{4}{h^2}\sin^2(\tfrac{1}{2}kh\pi) = k^2\pi^2\left(\frac{\sin(\frac{1}{2}kh\pi)}{\frac{1}{2}kh\pi}\right)^2 \geq k^2\pi^2\left(\frac{\sin(\frac{1}{2}\pi)}{\frac{1}{2}\pi}\right)^2 = 4k^2\,,
$$

we clearly satisfy the conditions in (4.32).

We now apply Theorem 4.2.3 to the last expression of (6.11) and get

$$
\int_0^T \left|\boldsymbol{B}^T\boldsymbol{v}(t)\right|^2 dt \geq \frac{C_1}{\sum_{k=1}^N (-\lambda_k^\alpha)^{-1}} \sum_{k=1}^N \frac{e^{-\lambda_k^\alpha T}\sin^2(kh\pi)}{-\lambda_k^\alpha h^2} e^{2\lambda_k^\alpha T}|a_k|^2\,, \qquad (6.12)
$$

for some positive constant C_1 which is independent of the discretization level N, the parameter α, and the coefficients $\langle a_k\rangle_{k=1}^\infty$.

We now wish to show that the factor in front of $e^{2\lambda_k^\alpha T}|a_k|^2$ in (6.12), majorizes the corresponding coefficient in (6.10). The ratio between such two factors is

$$
\frac{\sin^2(kh\pi)}{2\sin^2(\frac{1}{2}kh\pi)}\frac{e^{-\lambda_k^\alpha T}}{-\lambda_k^\alpha} = \frac{\sin^2(x\pi)}{2\sin^2(\frac{1}{2}x\pi)}\frac{e^{Tf_\alpha(x)/h^2}}{f_\alpha(x)/h^2} = r(x,h,\alpha)\,, \qquad (6.13)
$$

where $x = kh$ and

$$f_\alpha(x) = \frac{4\sin^2(\frac{1}{2}x\pi)}{1 - 4\alpha\sin^2(\frac{1}{2}x\pi)} \,.$$

For $0 < a \le x \le b < 1$, with a and b fixed, $r(x, h, \alpha)$ is clearly bounded from below as $h \to 0$. With $x = h \to 0$ we get

$$r(h, h, \alpha) \to \frac{h^2\pi^2}{\frac{1}{2}h^2\pi^2}\frac{e^{T\pi^2}}{\pi^2} = \frac{2e^{T\pi^2}}{\pi^2} \,,$$

so a uniform bound exists here also. For $x = Nh = N/(N+1)$ as $h \to 0$ we split into two cases. For $0 \le \alpha < 1/4$ we have

$$r(Nh, h, \alpha) \to \frac{h^2\pi^2}{2}\frac{e^{4T/\left((1-4\alpha)h^2\right)}}{4/\left((1-4\alpha)h^2\right)} \to \infty \,,$$

and for $\alpha = 1/4$ we have $f_\alpha(Nh) = 4/\cos^2(\frac{1}{2}Nh\pi) = 4/\sin^2(\frac{1}{2}h\pi) \to 16/(h^2\pi^2)$ and thus

$$r(Nh, h, 1/4) \to \frac{h^2\pi^2}{2}\frac{e^{16T/(h^4\pi^2)}}{16/(h^4\pi^2)} \to \infty \,.$$

All in all, the ratio $r(x, h, \alpha)$ in (6.13) is uniformly bounded away from zero for $h \le x \le Nh$ and $0 \le \alpha \le 1/4$ as $h \to 0$. So we finally conclude that the observability inequality

$$C_s\|\boldsymbol{v}(0)\|^2_{\widetilde{\boldsymbol{Q}}} \le \int_0^T \left|\boldsymbol{B}^T\boldsymbol{v}(t)\right|^2 dt \,,$$

holds for a constant $C_s > 0$ which is independent of N and the initial condition $\boldsymbol{v}^0$.

The so-called direct inequality, which is the semi-discrete analog of (6.4), is easier to show. We first bound the norm of the initial condition $\boldsymbol{v}^0$,

$$\begin{aligned}\|\boldsymbol{v}^0\|^2_{\widetilde{\boldsymbol{Q}}} &= -\sum_{k=1}^N |a_k|^2\langle \boldsymbol{w}_k, \boldsymbol{A}\boldsymbol{w}_k\rangle = 2\sum_{k=1}^N \frac{\sin^2(\frac{1}{2}kh\pi)}{h^2}|a_k|^2 \\ &= \tfrac{1}{2}\sum_{k=1}^N k^2\pi^2\left(\frac{\sin(\frac{1}{2}kh\pi)}{\frac{1}{2}kh\pi}\right)^2|a_k|^2 \ge 2\sum_{k=1}^N |ka_k|^2 \,.\end{aligned}$$

Next we have, see (6.11),

$$\begin{aligned}
\int_0^T \left|\boldsymbol{B}^T \boldsymbol{v}(t)\right|^2 dt &= \int_0^T \left| \sum_{k=1}^N \frac{(-1)^k \sin(kh\pi)}{h} e^{\lambda_k^\alpha t} a_k \right|^2 dt \\
&= \int_0^T \left| \sum_{k=1}^N \sum_{l=1}^N \frac{(-1)^{k+l} \sin(kh\pi)\sin(kh\pi)}{h^2} e^{(\lambda_k^\alpha + \lambda_l^\alpha)t} a_k a_l \right| dt \\
&\leq \sum_{k=1}^N \sum_{l=1}^N \left|\frac{\sin(kh\pi)}{kh\pi}\right| \left|\frac{\sin(lh\pi)}{kh\pi}\right| \left| \int_0^T e^{(\lambda_k^\alpha + \lambda_l^\alpha)t} dt \right| |k\pi a_k||l\pi a_l| \\
&\leq \frac{-1}{2\lambda_1^\alpha} \left(\sum_{k=1}^N |k\pi a_k|^2 \right)^{1/2} \left(\sum_{l=1}^N |l\pi a_l|^2 \right)^{1/2} = \frac{-\pi^2}{2\lambda_1^\alpha} \sum_{k=1}^N |k a_k|^2 \\
&\leq \frac{-\pi^2}{4\lambda_1^\alpha} \|\boldsymbol{v}^0\|_{\widetilde{\boldsymbol{Q}}}^2 ,
\end{aligned}$$

which proves the direct inequality.

We have now established a uniform observability inequality for the semi-discrete heat equation in one dimension, implying that computed controls will converge. A similar result has been shown in *López and Zuazua (1998)*, see also *Zuazua (2003)*.

Uniform observability for a full discretization of the heat equation has yet to be proved. Even if such a result is obtained, some practical difficulties will most likely occur when computing controls, one difficulty being the high condition number of the controllability operator. We will return to this subject in Section 9.5.2.

The literature has presented some numerical results regarding controllability of the heat equation, but exact (null-)controllability *is* hard to do. In *Carthel, Glowinski, and Lions (1994)*, the authors argue that for exact control, only very smooth functions can be reached. They then move on to *approximate* controllability, where they use a Tikhonov regularization method together with the conjugate gradient algorithm for computing the controls (they use a simple finite element discretization in space and a backward Euler time discretization). The paper *Kindermann (1999)* uses a similar approach, but focuses on the speed of convergence. In *Park and Lee (2002)*, the authors also consider approximate controllability. They do this by minimizing an appropriate functional using a conjugate gradient-like algorithm.

Chapter 7

The Wave Equation

If everything seems under control,
you're just not going fast enough.
— Mario Andretti

We now turn to the wave equation in $\Omega \subset \mathbb{R}^d$, to which we apply Dirichlet control on a subset of the boundary, $\Gamma_0 \subset \Gamma = \partial\Omega$,

$$\begin{cases} u_{tt} = \Delta u & \text{in } Q \ , \\ u = \begin{cases} k \\ 0 \end{cases} & \begin{array}{l} \text{in } \Sigma_0 \ , \\ \text{in } \Sigma \setminus \Sigma_0 \ , \end{array} \\ u(0) = u^0, \quad u_t(0) = \overline{u}^0 & \text{in } \Omega \ , \end{cases}$$

where $k \in L^2(\Sigma_0)$ and $(u^0, \overline{u}^0) \in H' = L^2(\Omega) \times H^{-1}(\Omega)$. In order for this control system to fit into the framework of Chapter 2, we must write it as a first order (in time) system. This is easily done as

$$\begin{cases} y_t = \mathcal{A}y & \text{in } Q \ , \\ \mathcal{B}y = \begin{cases} k \\ 0 \end{cases} & \begin{array}{l} \text{in } \Sigma_0 \ , \\ \text{in } \Sigma \setminus \Sigma_0 \ , \end{array} \\ y(0) = y^0 & \text{in } \Omega \ , \end{cases} \tag{7.1}$$

where

$$y(t) = \begin{bmatrix} u(t) \\ u_t(t) \end{bmatrix}, \qquad y^0 = \begin{bmatrix} u^0 \\ \overline{u}^0 \end{bmatrix}, \qquad \mathcal{A} = \begin{bmatrix} 0 & I \\ \Delta & 0 \end{bmatrix}, \qquad \mathcal{B} = \begin{bmatrix} u \\ \overline{u} \end{bmatrix} \mapsto u \, .$$

We clearly have (obtained by ignoring boundary conditions)

$$\mathcal{A}^* = \begin{bmatrix} 0 & \Delta \\ I & 0 \end{bmatrix},$$

but using the operator $-\mathcal{A}^*$ for the adjoint system leads to

$$\begin{bmatrix} v_t \\ v \end{bmatrix}_t = \begin{bmatrix} 0 & \Delta \\ I & 0 \end{bmatrix} \begin{bmatrix} v_t \\ v \end{bmatrix} ,$$

which we prefer to write as

$$\begin{bmatrix} v \\ v_t \end{bmatrix}_t = - \begin{bmatrix} 0 & -I \\ -\Delta & 0 \end{bmatrix} \begin{bmatrix} v \\ v_t \end{bmatrix} ,$$

since the initial conditions are in a more natural order (a minus-sign is placed in front of the operator since it is required from the general formulation (2.4) of the adjoint system). This means that we set

$$\widetilde{\mathcal{A}} = \begin{bmatrix} 0 & -I \\ -\Delta & 0 \end{bmatrix} ,$$

for the adjoint system,

$$\begin{cases} z_t = -\widetilde{\mathcal{A}}z & \text{in } Q \ , \\ \mathcal{B}z = 0 & \text{in } \Sigma \ , \\ z(T) = z^0 & \text{in } \Omega \ , \end{cases} \tag{7.2}$$

for some fixed control time $T > 0$.

Since $\widetilde{\mathcal{A}} \neq \mathcal{A}^*$ we need a non-trivial "conversion matrix" $\mathcal{M}$. By the requirement $\mathcal{M}\widetilde{\mathcal{A}} = \mathcal{A}^*\mathcal{M}$ we derive

$$\mathcal{M} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} .$$

(Note that any scaling of this matrix will do.) Given this matrix, we can deduce the relevant Hilbert spaces,

$$H = L^2(\Omega) \times H_0^1(\Omega) , \qquad \widetilde{H} = H_0^1(\Omega) \times L^2(\Omega) ,$$

$$H' = L^2(\Omega) \times H^{-1}(\Omega) , \qquad \widetilde{H}' = H^{-1}(\Omega) \times L^2(\Omega) ,$$

and the duality pairing $\{\cdot,\cdot\}$,

$$\left\{ \begin{bmatrix} u \\ \overline{u} \end{bmatrix}, \begin{bmatrix} v \\ \overline{v} \end{bmatrix} \right\} = \left\langle \begin{bmatrix} u \\ \overline{u} \end{bmatrix}, \mathcal{M} \begin{bmatrix} v \\ \overline{v} \end{bmatrix} \right\rangle = \left\langle \begin{bmatrix} u \\ \overline{u} \end{bmatrix}, \begin{bmatrix} -\overline{v} \\ v \end{bmatrix} \right\rangle = \langle \overline{u}, v \rangle - \langle u, \overline{v} \rangle ,$$

for all instances of $(u, \overline{u}) \in H'$ and $(v, \overline{v}) \in \widetilde{H}$. Recall that H' is the space in which the control system (7.1) is posed and $\widetilde{H}$ is the space in which the adjoint system (7.2) is posed. We shall show shortly that this leads to well-posed systems.

Left to find is the complementary boundary operator $\mathcal{C}$. We set $u, \overline{u}, v, \overline{v} \in C^\infty(\overline{\Omega})$ and get

$$\begin{aligned} \left\{ \mathcal{A} \begin{bmatrix} u \\ \overline{u} \end{bmatrix}, \begin{bmatrix} v \\ \overline{v} \end{bmatrix} \right\} - \left\{ \begin{bmatrix} u \\ \overline{u} \end{bmatrix}, \widetilde{\mathcal{A}} \begin{bmatrix} v \\ \overline{v} \end{bmatrix} \right\} &= \left\{ \begin{bmatrix} \overline{u} \\ \Delta u \end{bmatrix}, \begin{bmatrix} v \\ \overline{v} \end{bmatrix} \right\} - \left\{ \begin{bmatrix} u \\ \overline{u} \end{bmatrix}, \begin{bmatrix} -\overline{v} \\ -\Delta v \end{bmatrix} \right\} \\ &= \langle \Delta u, v \rangle - \langle \overline{u}, \overline{v} \rangle + \langle \overline{u}, \overline{v} \rangle - \langle u, \Delta v \rangle = \int_\Gamma \left(\frac{\partial u}{\partial n} v - u \frac{\partial v}{\partial n} \right) d\Gamma \\ &= \int_{\Gamma_0} u \left(-\frac{\partial v}{\partial n} \right) d\Gamma , \end{aligned}$$

when using the boundary conditions of the control and adjoint systems, showing that the complementary boundary operator is $\mathcal{C}(v,\overline{v}) = -\partial v/\partial n$. Different types of boundary control easily could be used, analogous to Section 6.1.1 for the heat equation.

7.1 Well-posedness

The adjoint system (7.2) can be written more readable as

$$\begin{cases} v_{tt} = \Delta v & \text{in } Q \ , \\ v = 0 & \text{in } \Sigma \ , \\ v(T) = v^0, \quad v_t(T) = \overline{v}^0 & \text{in } \Omega \ , \end{cases}$$

by setting $(v, v_t) = z$. We will now show the well-posedness of this system. Let us introduce the energy of the system at time t,

$$E(t) = \tfrac{1}{2}\int_\Omega \big(|\nabla v(t,x)|^2 + |v_t(t,x)|^2\big)dx\,.$$

The important observation is here that $(v(t), v_t(t)) \in \widetilde{H}$ if and only if $E(t) < \infty$. By differentiating the energy,

$$E'(t) = \int_\Omega \big(\nabla v\cdot\nabla v_t + v_t v_{tt}\big)dx = \int_\Omega \big(\nabla v\cdot\nabla v_t + v_t\Delta v\big)dx = \int_\Gamma \frac{\partial v}{\partial n} v_t\, d\varGamma = 0\,,$$

we see that the energy remains *constant* through time, $E(t) = E(T)$ for all t. This means that $(v^0, \overline{v}^0) \in \widetilde{H}$ implies $(v(t), v_t(t)) \in \widetilde{H}$ for all t.

We now wish to show that the complementary boundary operator is bounded, in the sense that a point-wise positive function $K(T)$ exists such that

$$\int_{\Sigma_0} \big|\mathcal{C}(v,\overline{v})\big|^2 d\varGamma dt = \int_0^T \int_{\Gamma_0} \left|\frac{\partial v}{\partial n}\right|^2 d\varGamma dt \le K(T)\big\|(v^0,\overline{v}^0)\big\|^2_{\widetilde{H}}\,, \tag{7.3}$$

for every solution $(v(t), v_t(t))$ of the adjoint system with initial conditions $(v^0, \overline{v}^0) \in \widetilde{H}$.

We show the above inequality for the simple case of $\Omega = (0,1)$ and $\Gamma_0 = \{1\}$. We apply the multiplier $v_x x$ to the right-hand side of $v_{tt} = v_{xx}$ and integrate over $(0,T)\times(0,1)$,

$$\int_0^T\int_0^1 v_{xx}v_x x\,dxdt = \int_0^T \big[v_x^2 x\big]_{x=0}^1 dt - \int_0^T\int_0^1 v_x v_{xx} x\,dxdt - \int_0^T\int_0^1 v_x^2\,dx\,,$$

which leads to

$$\int_0^T |v_x(t,1)|^2 dt = 2\int_0^T\int_0^1 v_{xx}v_x x\,dxdt + \int_0^T\int_0^1 v_x^2\,dxdt\,. \tag{7.4}$$

Our goal is to bound the absolute value of the right-hand side. Using $v_{tt} = v_{xx}$ on the first term of the right-hand side we get

$$\begin{aligned} 2\int_0^T \int_0^1 v_{xx} v_x x dx dt &= 2\int_0^1 \int_0^T v_{tt} v_x x dt dx \\ &= 2\int_0^1 \left[v_t v_x x\right]_{t=0}^T dx - 2\int_0^1 \int_0^T v_t v_{xt} x dt dx \end{aligned} \tag{7.5}$$

In order to rewrite the last term, we consider

$$\int_0^1 v_{xt} v_t x dx = \left[v_t^2 x\right]_{x=0}^1 - \int_0^1 v_t v_{tx} x dx - \int_0^1 v_t^2 dx\,,$$

obtaining (since $v_t(t,1) = 0$),

$$2\int_0^1 v_{xt} v_t x dx = -\int_0^1 v_t^2 dx\,.$$

Substituting this expression into (7.5), which in turn gets inserted into (7.4), we get

$$\int_0^T |v_x(t,1)|^2 dt = 2\int_0^1 \left[v_t v_x x\right]_{t=0}^T dx + \int_0^T \int_0^1 v_t^2 dx dt + \int_0^T \int_0^1 v_x^2 dx dt\,.$$

We now bound the absolute value of each term of the right-hand side,

$$\left|\int_0^1 v_t v_x x dx\right|^2 \le \int_0^1 v_t^2 dx \int_0^1 v_x^2 x dx \le 4E(t)^2 = 4E(0)^2\,,$$

$$\int_0^T \int_0^1 \left(v_t^2 + v_x^2\right) dx dt = 2\int_0^T E(t) dt = 2TE(0)\,,$$

which finally yields

$$\int_0^T |v_x(t,1)|^2 dt \le 8E(0) + 2TE(0) = 2E(0)(T+4)\,. \tag{7.6}$$

Since $E(0)$ and the $\|\cdot\|_{\widetilde{H}}$-norm are of the same order (easily seen in one dimension by writing these quantities in terms of $\sin(k\pi\cdot)$-basis coefficients), this inequality implies (7.3).

7.2 Analytical Solution in 1D Using Fourier Series

For the case $\Omega = (0,1)$, the controllability operator has the "type" $\Lambda_T : H_0^1(0,1) \times L^2(0,1) \mapsto H^{-1}(0,1) \times L^2(0,1)$. We will now find a matrix representation $\mathbf{\Lambda}_T$ for the controllability operator, using a $\sin(j\pi\cdot)$ basis for each of these spaces.

Let us start out with the simple case where $T \in 2\mathbb{N}$, and use the inner product method. With initial conditions $(v(T), v_t(T) = (e_j, 0)$ for the adjoint system we have the solution

$$v(t,x) = \cos(j\pi(T-t))e_j(x)\,,$$

and thus

$$\mathcal{C}(v(t), v_t(t)) = -v_x(t,1) = (-1)^{j+1} j\pi \cos(j\pi(T-t))\,.$$

Let similarly $\tilde{v}$ be a solution to the adjoint system with initial conditions $(e_i, 0)$. This leads to

$$\begin{aligned}\left\langle \Lambda_T \begin{pmatrix} e_j \\ 0 \end{pmatrix}, \begin{pmatrix} e_i \\ 0 \end{pmatrix} \right\rangle &= \int_0^T v_x(t,1)\tilde{v}_x(t,1)dt \\ &= (-1)^{i+j} ij\pi^2 \int_0^T \cos(i\pi t)\cos(j\pi t)dt = (-1)^{i+j} ij\pi^2 \delta_{ij} T/2\,.\end{aligned}$$

With initial conditions $(v(T), v_t(T)) = (0, e_j)$ we get

$$-v_x(t,1) = (-1)^j \sin(j\pi(T-t))\,,$$

leading to

$$\begin{aligned}\left\langle \Lambda_T \begin{pmatrix} 0 \\ e_j \end{pmatrix}, \begin{pmatrix} 0 \\ e_i \end{pmatrix} \right\rangle &= \int_0^T v_x(t,1)\tilde{v}_x(t,1)dt \\ &= (-1)^{i+j} \int_0^T \sin(i\pi t)\sin(j\pi t)dt = (-1)^{i+j} \delta_{ij} T/2\,.\end{aligned}$$

By similar calculations we finally get

$$\left\langle \Lambda_T \begin{pmatrix} e_j \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ e_i \end{pmatrix} \right\rangle = \left\langle \Lambda_T \begin{pmatrix} 0 \\ e_j \end{pmatrix}, \begin{pmatrix} e_i \\ 0 \end{pmatrix} \right\rangle = 0\,.$$

As was the case for the heat equation in Section 6.2, we also use the $\sin(j\pi\cdot)$ basis for the range of the controllability operator, which means that we must scale each inner product by a factor of two. All in all, a very simple matrix occurs for the case $T \in 2\mathbb{N}$,

$$\mathbf{\Lambda}_T = T \left[\begin{array}{ccc|ccc} 1^2\pi^2 & & & & & \\ & 2^2\pi^2 & & & & \\ & & \ddots & & & \\ \hline & & & 1 & & \\ & & & & 1 & \\ & & & & & \ddots \end{array}\right], \tag{7.7}$$

where all the blank entries denote zeroes. The matrix entries for all $T > 0$ are computed in Detail 6, page 185, using the direct method.

This analytical expression for the controllability operator has not been published before. But as mentioned in Section 5.1, an analytical representation for the controllability operator for the wave equation on $\Omega = (0,1)^2$ was calculated in *Glowinski, Li, and Lions (1990)*.

Let us compute the control function $k \in L^2(0,T)$ in the case $T \in 2\mathbb{N}$. We wish to drive the zero state at $t = 0$ to the state $(y^0, \overline{y}^0)$ at time $t = T$. Let

$$y^0 = \sum_{k=1}^{\infty} a_k \sin(k\pi\cdot), \qquad \overline{y}^0 = \sum_{k=1}^{\infty} b_k \sin(k\pi\cdot)\,,$$

where $\langle a_k \rangle \in \ell^2$ and $\langle b_k/k \rangle \in \ell^2$, ensuring that $(y^0, \overline{y}^0) \in L^2(0,1) \times H^{-1}(0,1)$. We must now solve

$$\Lambda_T \begin{pmatrix} v^0 \\ \overline{v}^0 \end{pmatrix} = \mathcal{M}^T \begin{pmatrix} y^0 \\ \overline{y}^0 \end{pmatrix} = \begin{pmatrix} \overline{y}^0 \\ -y^0 \end{pmatrix}\,,$$

which is easily done using the matrix representation in (7.7),

$$v^0 = \frac{1}{T} \sum_{k=1}^{\infty} \frac{b_k}{k^2\pi^2} \sin(k\pi\cdot), \qquad \overline{v}^0 = -\frac{1}{T} \sum_{k=1}^{\infty} a_k \sin(k\pi\cdot)\,,$$

These initial conditions lead to a solution of the form

$$v(t,x) = \frac{1}{T} \sum_{k=1}^{\infty} \left(\frac{b_k}{k^2\pi^2} \cos(k\pi t) - \frac{a_k}{k\pi} \sin(k\pi t) \right) \sin(k\pi x)\,,$$

and thus the control

$$k(t) = -v_x(t,1) = \frac{1}{T} \sum_{k=1}^{\infty} (-1)^k \left(a_k \sin(k\pi t) - \frac{b_k}{k\pi} \cos(k\pi t) \right)\,. \tag{7.8}$$

Note that since $\langle a_k \rangle, \langle b_k/k \rangle \in \ell^2$ we have $k \in L^2(0,T)$ (this is of course no surprise, since it had to be this way—it was proved in (7.6)). Note also the control's dependence on time, its $L^2(0,T)$-norm is inversely proportional to T, as predicted after Theorem 5.2.2.

The matrix representation (7.7) also proves that Λ_T is invertible for $T = 2$, and thus for all $T \geq 2$. Using, for instance, Ingham's Theorem (Theorem 4.2.1) for proving exact controllability in this 1D case, we would obtain the slightly weaker condition $T > 2$.

We can easily argue that there can be no exact controllability when $T < 2$. Recall that exact controllability is equivalent to the following condition: A constant $K_e > 0$ must exist such that the observability inequality

$$\left\| (v^0, \overline{v}^0) \right\|_{\widetilde{H}}^2 \leq K_e \int_0^T v_x^2(t,1)dt\,,$$

holds for *all* solutions v of the adjoint system with initial conditions $(v^0, \overline{v}^0)$. Assume that $T = 2 - \epsilon$ for a small $\epsilon > 0$. Consider now initial conditions $(v^0, \overline{v}^0)$ that are constructed such that the support of both v^0 and $\overline{v}^0$ is contained in the open interval $(1-\epsilon, 1)$ and such that the solution exclusively travels to the *left* (as time goes from T and back to 0). Because of the constant speed of propagation equal to one we must have $v_x(t,1) = 0$ for $0 < t < T = 2-\epsilon$. The above observability inequality can thus not hold in this case (a similar argument can be found in *Zuazua, 2003*, page 14).

7.3 Characterization of Controls for the Wave Equation in 1D

We consider again the following one dimensional control problem,

$$\begin{cases} u_{tt} = u_{xx}, & \text{in } (0,T)\times(0,1)\,, \\ u(t,0) = 0, \quad u(t,1) = k(t), & \text{in } (0,T)\,, \\ u(0,x) = u^0(x), \quad u_t(0,x) = \overline{u}^0(x), & \text{in } (0,1)\,, \end{cases} \tag{7.9}$$

with $(u^0, \overline{u}^0) \in L^2(0,1)\times H^{-1}(0,1)$ and $k \in L^2(0,T)$. The goal of this section is to characterize *all* possible controls for this control system for any $T > 0$. Recall that a HUM control exists for $T \geq 2$ and is unique in the sense that it has the smallest $L^2(0,T)$-norm (in general, a HUM control has the smallest $L^2((0,T)\times\Gamma_0)$-norm). We will here present a fairly constructive way of finding a control that is optimal in any sense.

As far as the author knows, the approach taken in this section has not been considered before. Although the method is not readily possible to generalize to more dimension, it provides some useful insight.

The D'Alembert solution formula for the wave equation will be important to us. On the real line any solution to the wave equation has the appearance

$$u(t,x) = f(x+t) + g(x-t), \quad t \geq 0,\, x \in \mathbb{R}\,,$$

where f and g are twice continuously differentiable functions, $f, g \in C^2(\mathbb{R})$. So a solution is simply the superposition of two waves, one traveling to the left and one traveling to the right. To take into consideration reflection at $t = 0$ we use odd extensions and get that

$$u(t,x) = f(x+t) - f(-x+t) + g(x-t) - g(-x-t), \quad t \geq 0,\, x \geq 0\,, \tag{7.10}$$

satisfies both

$$u_{tt}(t,x) = u_{xx}(t,x) \quad \text{and} \quad u(t,0) = 0\,.$$

For more on the D'Alembert solution formula and the method of reflection, see *Strauss (1992)*.

Combining the solution formula (7.10) with the initial conditions,

$$u(0,x) = u^0(x), \quad u_t(0,x) = \overline{u}^0(x), \quad x \geq 0\,,$$

we arrive at the following expressions

$$
\begin{aligned}
u(t,x) &= \frac{1}{2}\left[u^0(x+t) + u^0(x-t) + \int_{x-t}^{x+t} \overline{u}^0(s)ds\right], \quad \text{for } |t| < x\,, \\
u(t,x) &= \frac{1}{2}\left[u^0(t+x) + u^0(t-x) + \int_{t-x}^{t+x} \overline{u}^0(s)ds\right], \quad \text{for } 0 < x < |t|\,.
\end{aligned}
\tag{7.11}
$$

But our control system is posed on the interval $(0,1)$, and not on the half-line $(0,\infty)$. Nevertheless, the D'Alembert solution formula can be used to find a control that drives the control system to rest in time two (or less). Indeed, let initial conditions $(u^0, \overline{u}^0) \in L^2(0,1) \times H^{-1}(0,1)$ for the control system (7.9) be given. Expand now the initial conditions with zeroes to the half-line $(0,\infty)$, and solve the wave equation on the interval $(0,\infty)$ with reflection at $x = 0$. Because of the unit speed of propagation it is clear that $u(t,x) = 0$ for $0 < x < 1$ for all $t > 2$. The control is now found simply by reading off the positions at $x = 1$ in the interval $0 \le t \le 2$. Using the solution formula (7.11) we get

$$
k(t) = u(t,1) = \begin{cases} \frac{1}{2}u^0(1-t) + \frac{1}{2}\displaystyle\int_{1-t}^{1} \overline{u}^0(s)ds, & 0 \le t \le 1\,, \\ \frac{1}{2}u^0(t-1) + \frac{1}{2}\displaystyle\int_{t-1}^{1} \overline{u}^0(s)ds, & 1 < t \le 2\,. \end{cases}
\tag{7.12}
$$

This method, which we will call the SUR method, was mentioned by Professor David L. Russell in *Russell (1973)*, and later developed by Professor Walter Littman and others in, for instance, *Littman (1992)*. See Figure 7.1 for an example.

7.3.1 $0 < T < 2$

If the control time available is less than two then exact controllability is impossible, as argumented at the end of Section 7.2. But for some initial conditions, control is in fact possible.

Consider first the case $0 < T \le 1$. It is clear that for any part of the solution that travels to the left, it will take more than time T before it has reflected at $x = 0$ and travelled to $x = 1$ to be "handled" be the control. Thus the solution must travel to the right, $u(t,x) = f(x-t)$, implying the relation

$$
\overline{u}^0(x) = -\frac{d}{dt}u^0(x), \quad 0 < x < 1\,,
$$

for the initial conditions. Furthermore, we must require

$$
u^0(x) = 0 \quad \text{for } 0 < x < 1 - T\,,
$$

since this part of the solution will not have time to reach the control boundary. If these two conditions are met, the control given by (7.12) will drive the control system to rest in time T.

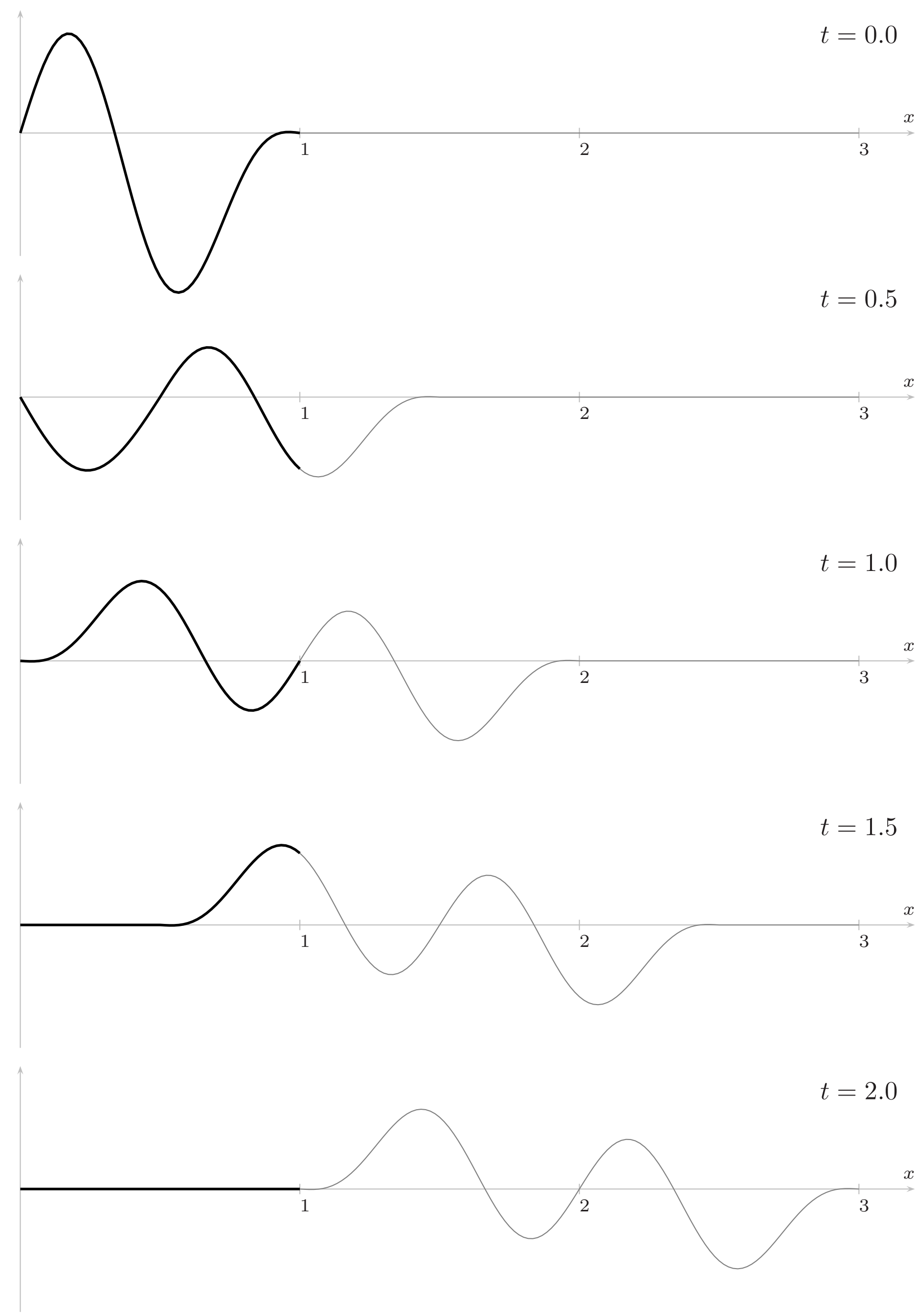


Figure 7.1: Illustration of the SUR method. The domain is expanded from $(0, 1)$ to $(0, \infty)$ and the wave equation is solved on this domain. The control is now simply read off at $x = 1$.

Consider now $1 < T < 2$. The part of the solution that at time $t = 0$ is in the interval $0 < x < T - 1$ will reach the control boundary in the time available. However, any part of the initial conditions in the interval $x \in (T-1, 1)$ has to make the solution travel to the right. We must thus require

$$\overline{u}^0(x) = -\frac{d}{dt}u^0(x), \quad T - 1 < x < 1\,,$$

of the initial conditions. Again, the control can then be computed by (7.12).

What about uniqueness for controls obtained as described? They are in fact unique, and for the simple reason that if the control did *anything other than absorb the oncoming waves*, it would take time two before that "signal" could disappear.

7.3.2 $T = 2$

By now we know of two ways of computing a control for the case $T = 2$: The HUM method, see (7.8), or the SUR method just described, see (7.12). Are they identical? No, not in general, as some examples could quickly reveal. Because of the linearity of the control system, subtracting two different controls that do "the same job" will be a *null-space control*, that is, a control that steers the control system from the zero state to the zero state. This leads to the question: What can for the case $T = 2$ be added to a control function that does not change the initial/final states? The following theorem answers this question.

Theorem 7.3.1. *A control $k \in L^2(0, 2)$ has the effect*

$$\begin{aligned} u(0, \cdot) &= 0, \quad u_t(0, \cdot) = 0\,, \\ u(2, \cdot) &= 0, \quad u_t(2, \cdot) = 0\,, \end{aligned} \tag{7.13}$$

if and only if the control function is a constant, $k(t) = k_0 \in \mathbb{R}$.

Proof. Assume that (7.13) holds. We use the D'Alembert formula to derive information about the control. Using (7.10) we get

$$\begin{aligned} u(0, x) &= f(x) - f(-x) + g(x) - g(-x) = 0\,, \\ u_t(0, x) &= f'(x) - f'(-x) - g'(x) + g'(-x) = 0\,, \\ u(2, x) &= f(x+2) - f(-x+2) + g(x-2) - g(-x-2) = 0\,, \\ u_t(2, x) &= f'(x+2) - f'(-x+2) - g'(x-2) + g'(-x-2) = 0\,, \end{aligned}$$

for $0 \le x \le 1$. Integrating the equations involving derivatives, we can combine the equations into

$$\begin{aligned} f(-x) - g(x) &= K_1\,, \qquad & f(x+2) - g(-x-2) &= K_2\,, \\ f(x) - g(-x) &= K_1\,, \qquad & f(-x+2) - g(x-2) &= K_2\,, \end{aligned}$$

for $0 \le x \le 1$, and where $K_1, K_2 \in \mathbb{R}$ are constants. We now see what happens at the boundary at $x = 1$,

$$u(t, 1) = f(1+t) - f(-1+t) + g(1-t) - g(-1-t) = K_2 - K_1\,,$$

for $0 \leq t \leq 2$. Indeed, a constant-valued control.

Assume now that $k(t) = k_0$, a constant. We now get

$$\int_0^2 k(t)v_x(t,1)dt = k_0 \int_0^2 v_x(t,1)dt = 0\,,$$

since

$$\int_0^2 \cos(j\pi(2-t))dt = \int_0^2 \sin(j\pi(2-t))dt = 0\,.$$

Using the ever useful Theorem 2.1.1, this is seen to imply (7.13). □

This means that for the case $T = 2$, a HUM control and a SUR control will be identical except for an additive constant. One can also put it this way: A HUM control is obtained from the SUR control by adjusting it by a constant in such a way that the $L^2(0,2)$-norm is the smallest possible.

7.3.3 $T > 2$

What about null-space controls for the case $T > 2$? A trivial null-space control is clearly $k(t) = 0$ for $0 < t < T$. But by Theorem 7.3.1 we can add a constant on any interval of length two, say $(t, t+2)$ for some $t \in (0, T-2)$, and we still have a null-space control. We can now do this repeatedly on different intervals of length two. For instance, for the case $T = 5$ the control

$$k(t) = \begin{cases} 2, & 0 \leq t \leq 1\,, \\ 3, & 1 < t \leq 2\,, \\ 1, & 2 < t \leq 3\,, \\ -1, & 3 < t \leq 5\,, \end{cases}$$

is a null-space control. A general result on null-space controls for the case $T > 2$ is given by the following theorem.

Theorem 7.3.2. *Let $T > 2$. The control $k \in L^2(0,T)$ has the effect*

$$\begin{aligned} u(0,\cdot) = 0, \quad u_t(0,\cdot) = 0\,, \\ u(T,\cdot) = 0, \quad u_t(T,\cdot) = 0\,, \end{aligned} \tag{7.14}$$

if and only if

$$k_0 = \sum_{\substack{t+2p \leq T \\ p=0,1,\ldots}} k(t+2p), \quad 0 \leq t < 2\,, \tag{7.15}$$

for some constant $k_0 \in \mathbb{R}$.

Proof. In order to satisfy (7.14) we see from the relation (2.7), page 12, that we must have

$$\int_0^T k(t)v_x(t,1)dt = 0\,, \tag{7.16}$$

for all solutions $v(t,x)$ to the adjoint system. Since $v_x(t,1)$ has the form

$$v_x(t,1) = \sum_{j=1}^{\infty}(-1)^j j\pi\big(a_j\cos(j\pi(T-t)) + b_j\sin(j\pi(T-t))\big), \quad \langle ja_j\rangle, \langle jb_j\rangle \in \ell^2\,, \tag{7.17}$$

we see that $v_x(\cdot,1)$ is 2π-periodic. This means that (7.16) is equivalent to

$$\int_0^2 k^\star(t)v_x(t,1)dt = 0\,, \tag{7.18}$$

for all v_x of the form (7.17) where

$$k^\star(t) = \sum_{\substack{t+2p\le T\\ p=0,1,\ldots}} k(t+2p), \quad 0\le t\le 2\,.$$

Introducing a simple phase shift we get that (7.18) is equivalent to

$$\int_0^2 k^\star(t)v_x^\star(t,1)dt = 0\,,$$

for all $v_x^\star$ of the form

$$v_x^\star(t,1) = \sum_{j=1}^{\infty}(-1)^j j\pi\big(a'_j\cos(j\pi(2-t)) + b'_j\sin(j\pi(2-t))\big), \quad \langle ja'_j\rangle, \langle jb'_j\rangle \in \ell^2\,.$$

So $k(t)$ is a null-control when the control time is T, if and only if $k^\star(t)$ is a null-control when the control time is 2. Using now the result of Theorem 7.3.1, the proof is complete. □

This result can be used in the following ways. A control that drives the control system to rest in time $T = 2$ can also be used for the case $T > 2$, simply by expanding it with zeroes on $t \in (2,T)$. We can now modify such a control by adding to it null-controls for the case T, as just described.

We can also use the result in the opposite direction. Given a control $k \in L^2(0,T)$ for some $T > 2$, we construct a null-space control $\tilde{k} \in L^2(0,T)$ for which $k(t) = \tilde{k}(t)$ on $t \in (2,T)$. Such a null-control is always possible to find, as can be seen from (7.15). Since now $k(t) - \tilde{k}(t) = 0$ for $2 < t < T$, we see that $k(t) - \tilde{k}(t)$ will be a valid control for the reduced control time 2.

7.3.4 Example of Optimal Controls in Different Norms

Figure 7.2 shows an example of four different controls that all, using the same initial conditions, drive the solution of the control system to rest. We will now describe how the four controls were computed.

The SUR control was computed simply by using formula (7.12) on a discrete time-grid. This led to a control on the interval $(0,2)$ which was then expanded with zeroes to the whole $(0,3.6)$ interval.

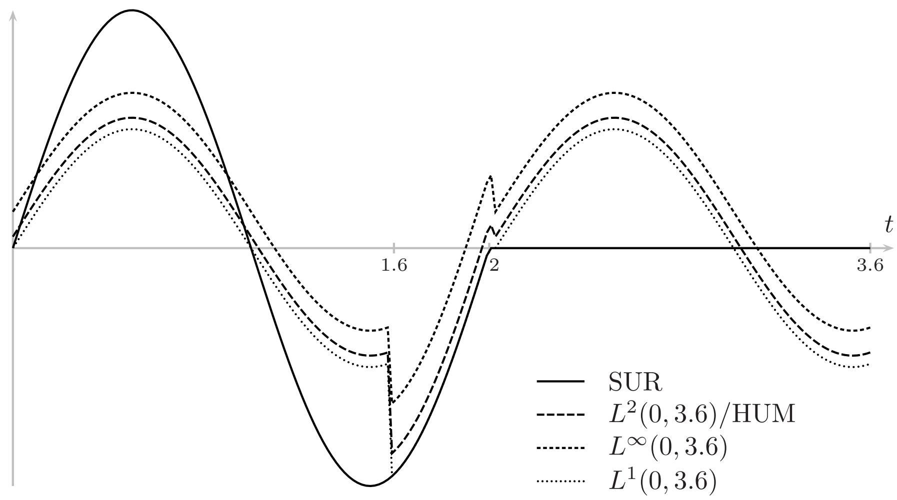


Figure 7.2: Example for the case $T = 3.6$. Using the same initial conditions, all the controls shown steer the control system to rest in time T. Apart from the control obtained by the SUR method, three controls are shown that are optimal in the sense that the $L^2(0, 3.6)$-, $L^\infty(0, 3.6)$- and $L^1(0, 3.6)$-norm of the controls are minimal (all controls are computed using numerical approximations).

Let the discrete SUR control be represented by the vector $\boldsymbol{k}_{\mathrm{SUR}}$. We now wish to modify this control by adding a null-space control such that the (discrete) $L^p(0, 3.6)$-norm is minimal for $p = 1, 2, \infty$. Recalling that adding a constant on a length-two interval does not change the effect of the control, leads to the following simple formula,

$$\boldsymbol{k}_p = \boldsymbol{k}_{\mathrm{SUR}} + \boldsymbol{N}\boldsymbol{x}^\star, \quad \boldsymbol{x}^\star = \operatorname*{argmin}_{\boldsymbol{x}} \left\| \boldsymbol{k}_{\mathrm{SUR}} + \boldsymbol{N}\boldsymbol{x} \right\|_p, \tag{7.19}$$

where $\boldsymbol{k}_p$ is the sought control and where the matrix

$$\boldsymbol{N} = \begin{bmatrix} 1 & & & \\ \vdots & 1 & & \\ 1 & \vdots & & \\ & 1 & \ddots & 1 \\ & & & \vdots \\ & & & 1 \end{bmatrix},$$

has a number of ones in each column corresponding to a time interval of length two. The norms $\| \cdot \|_p$ are the usual discrete p-norms.

The minimization in (7.19) is a standard *least squares problem* in the case $p = 2$, which can be done quite fast. The cases $p = 1, \infty$ lead to linear programming problems which are very time consuming.

7.4 A Well-behaved 1D Scheme

This section will present a full discretization of the wave equation in one dimension. It has not been studied before, and it will turn out to have some properties which are more appropriate for control than the discretization schemes previously seen in the literature.

Let us recall the continuous control system,

$$\begin{cases} u_{tt} = u_{xx}, & \text{in } (0,T)\times(0,1)\,, \\ u(t,0)=0, \quad u(t,1)=k(t), & \text{in } (0,T)\,, \\ u(0,x)=u^0(x), \quad u_t(0,x)=\overline{u}^0(x), & \text{in } (0,1)\,, \end{cases} \tag{7.20}$$

with $(u^0,\overline{u}^0)\in L^2(0,1)\times H^{-1}(0,1)$ and $k\in L^2(0,T)$. For easier notation, we will reverse the time direction when defining the adjoint system, see Section 2.5.2. This approach suffers no loss of generality since the wave equation is reversible. We have the adjoint system,

$$\begin{cases} v_{tt} = v_{xx}, & \text{in } (0,T)\times(0,1)\,, \\ v(t,0)=v(t,1)=0, & \text{in } (0,T)\,, \\ v(0,x)=v^0(x), \quad v_t(0,x)=\overline{v}^0(x), & \text{in } (0,1)\,, \end{cases} \tag{7.21}$$

with $(v^0,\overline{v}^0)\in H_0^1(0,1)\times L^2(0,1)$.

7.4.1 Discretization

We discretize the Laplacian using the box scheme (see (3.5), page 32),

$$\boldsymbol{C} = \frac{h}{4}\begin{bmatrix} 2 & 1 & & \\ 1 & 2 & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & 2 \end{bmatrix}, \quad \boldsymbol{A} = \frac{1}{h}\begin{bmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & -2 \end{bmatrix}, \quad \boldsymbol{B} = \frac{1}{h}\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

where $h = 1/(N+1)$ as usual. As for the heat equation, the matrix $\boldsymbol{B}$ takes care of the Dirichlet boundary conditions at the right end-point.

In time we use the trapezoid discretization scheme, here for the adjoint system (7.21),

$$\begin{bmatrix} \boldsymbol{v}^{n+1} \\ \overline{\boldsymbol{v}}^{n+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{v}^{n} \\ \overline{\boldsymbol{v}}^{n} \end{bmatrix} = \tfrac{1}{2}\Delta t \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{C}^{-1}\boldsymbol{A} & \boldsymbol{0} \end{bmatrix} \left(\begin{bmatrix} \boldsymbol{v}^{n+1} \\ \overline{\boldsymbol{v}}^{n+1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{v}^{n} \\ \overline{\boldsymbol{v}}^{n} \end{bmatrix} \right), \tag{7.22}$$

with $\boldsymbol{v}^0$ and $\overline{\boldsymbol{v}}^0$ given. Note how we have used the box scheme for both time and space discretization (see Equation (3.32), page 45, and the comments that surround it).

We will make extensive use of eigenvalues and eigenvectors. To that end, we recall that the eigensolutions of $\boldsymbol{C}^{-1}\boldsymbol{A}$ are

$$\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{w}_k = -\mu_k^2 \boldsymbol{w}_k \, ,$$

$$\text{with} \qquad \mu_k = \tfrac{2}{h}\tan(\tfrac{1}{2}kh\pi)\, , \quad \boldsymbol{w}_k(j) = \sin(kjh\pi)\, ,$$

for $j, k = 1, 2, \ldots, N$. With initial conditions

$$\boldsymbol{v}^0 = \sum_{k=1}^{N} a_k \boldsymbol{w}_k \quad \text{and} \quad \overline{\boldsymbol{v}}^0 = \sum_{k=1}^{N} b_k \boldsymbol{w}_k \, , \tag{7.23}$$

where $\langle k a_k\rangle, \langle b_k\rangle \in \ell^2$, the solution of (7.22) is given as

$$\boldsymbol{v}^n = \sum_{k=1}^{N} \big(a_k \cos(n\theta_k) + b_k/\mu_k \sin(n\theta_k)\big)\boldsymbol{w}_k \, ,$$

where

$$e^{i\theta_k} = \frac{2 + i\Delta t\mu_k}{2 - i\Delta t\mu_k} = \frac{1 + i\eta\tan(\frac{1}{2}kh\pi)}{1 - i\eta\tan(\frac{1}{2}kh\pi)}, \qquad \cos(\theta_k) = \frac{1 - \eta^2\tan^2(\frac{1}{2}kh\pi)}{1 + \eta^2\tan^2(\frac{1}{2}kh\pi)}\, , \tag{7.24}$$

such that $0 \le \theta_k < \pi$ for $k = 1, 2, \ldots, N$. We assume that $\eta = \Delta t/h$ is a positive constant.

From (3.33) we have the energy norm,

$$E^0 = \tfrac{1}{2}\sum_{k=1}^{N} (\mu_k^2 a_k^2 + b_k^2)\langle \boldsymbol{w}_k, \boldsymbol{C}\boldsymbol{w}_k\rangle \, ,$$

but since $\boldsymbol{C}\boldsymbol{w}_k = h\cos^2(\frac{1}{2}kh\pi)\boldsymbol{w}_k$ we get

$$\langle \boldsymbol{w}_k, \boldsymbol{C}\boldsymbol{w}_k\rangle = h\cos^2(\tfrac{1}{2}kh\pi)\langle \boldsymbol{w}_k, \boldsymbol{w}_k\rangle = \tfrac{1}{2}\cos^2(\tfrac{1}{2}kh\pi)\, ,$$

where $\langle \boldsymbol{w}_k, \boldsymbol{w}_k\rangle = 1/(2h)$, see Theorem 3.1.3, has been used. So we have

$$\begin{aligned} E^0 &= \tfrac{1}{4}\sum_{k=1}^{N}\left(\frac{4\tan^2(\frac{1}{2}kh\pi)}{h^2}a_k^2 + b_k^2\right)\cos^2(\tfrac{1}{2}kh\pi) \\ &= \tfrac{1}{4}\sum_{k=1}^{N}\left(k^2\pi^2\left(\frac{\sin(\frac{1}{2}kh\pi)}{\frac{1}{2}kh\pi}\right)^2 a_k^2 + \cos^2(\tfrac{1}{2}kh\pi)b_k^2\right) . \end{aligned} \tag{7.25}$$

7.4.2 Convergence of the Scheme

In this section we will show convergence of the discretization (7.22) used for the adjoint system. In order to do so, we will use the Lax Equivalence Theorem (see page 48), which says that consistency and stability are necessary and sufficient conditions for convergence.

First we must establish how to set the initial conditions of the discrete system, given the following initial conditions for the continuous system,

$$v^0 = \sum_{k=1}^{\infty} a_k w_k \,, \qquad \overline{v}^0 = \sum_{k=1}^{\infty} b_k w_k \,,$$

where $w_k(x) = \sin(k\pi x)$, $k \in \mathbb{N}$, and $\langle k a_k \rangle, \langle b_k \rangle \in \ell^2$. But here we simply use "spectral truncation" as used in (7.23). This is the R_N-map of Section 3.3.

We also need to be able to compare vectors from the discrete system, posed in $\mathbb{R}^N \times \mathbb{R}^N$, with vectors from the continuous system, posed in $\widetilde{H} = H_0^1(0,1) \times L^2(0,1)$. But to this end, we simply do the opposite as we did for the initial conditions: Given a spectral representation, we just replace the $\boldsymbol{w}_k$ basis vectors with the w_k basis vectors. This is the E_N-map of Section 3.3. From this point on, we will use the notation v_N^n as the vector of $\widetilde{H}$ that in this way corresponds to $\boldsymbol{v}^n$ of $\mathbb{R}^N \times \mathbb{R}^N$.

Finally, since everything is going to be represented in spectral terms, we will use the following expression for the norm in $\widetilde{H}$,

$$\left\| \left(\sum_{k=1}^{\infty} a_k w_k, \sum_{k=1}^{\infty} b_k w_k \right) \right\|_{\widetilde{H}}^2 = \tfrac{1}{2} \sum_{k=1}^{\infty} (k^2\pi^2 a_k^2 + b_k^2) \,.$$

We can now move on to show stability for the discretization of the adjoint system. In order to do so, we consider the discrete (energy) norm (7.25), which we know is constant in time for each N. If we can now show that the discrete norm converges to the continuous norm as $N \to \infty$, we are done. Let $\epsilon > 0$ be given and consider

$$\begin{aligned} |E - E^0| \le \tfrac{1}{4} & \left| \sum_{k=1}^{N_0} \left[k^2\pi^2 \left(1 - \left(\frac{\sin(\frac{1}{2}kh\pi)}{\frac{1}{2}kh\pi} \right)^2 \right) a_k^2 + (1 - \cos^2(\tfrac{1}{2}kh\pi)) b_k^2 \right] \right| \\ & + \tfrac{1}{4} \left| \sum_{k=N_0+1}^{N} \left[k^2\pi^2 \left(\frac{\sin(\frac{1}{2}kh\pi)}{\frac{1}{2}kh\pi} \right)^2 a_k^2 + \cos^2(\tfrac{1}{2}kh\pi) b_k^2 \right] \right| \\ & + \tfrac{1}{4} \left| \sum_{k=N_0+1}^{\infty} \left[k^2\pi^2 a_k^2 + b_k^2 \right] \right| = I_1 + I_2 + I_3 \,. \end{aligned}$$

Choose now N_0 such that $I_3 \le \epsilon/3$ and observe that $I_2 \le I_3$, independently of N. Fixing N_0, we can now choose N large enough so $I_1 \le \epsilon/3$. This shows that $E_N^0 \to E$ as $N \to \infty$, for every choice of initial data.

Consistency only has to be shown for a dense subset of $\widetilde{H}$ and we will assume that both v^0 and $\overline{v}^0$ are infinitely smooth. This means that the coefficients $\langle a_k \rangle$ and $\langle b_k \rangle$ decay exponentially,

$$k^2\pi^2 a_k^2 + b_k^2 \le C r^k \quad \text{for all } k \in \mathbb{N} \,,$$

for some constants $C > 0$ and $0 < r < 1$.

Loosely speaking, consistency means that going Δt forwards in time for the continuous system,

$$v(\Delta t) = \sum_{k=1}^{\infty} \Big[a_k \cos(k\pi\Delta t) + b_k \sin(k\pi\Delta t)/(k\pi) \Big] w_k \, ,$$
$$v_t(\Delta t) = \sum_{k=1}^{\infty} \Big[-a_k k\pi \sin(k\pi\Delta t) + b_k \cos(k\pi\Delta t) \Big] w_k \, ,$$

should be close to taking one step forward for the discrete system,

$$v_N^1 = \sum_{k=1}^{N} \Big[a_k \cos(\theta_k) + b_k \sin(\theta_k)/\mu_k \Big] w_k \, ,$$
$$\overline{v}_N^1 = \sum_{k=1}^{N} \Big[-a_k \mu_k \sin(\theta_k) + b_k \cos(\theta_k) \Big] w_k \, .$$

We now aim to show

$$\frac{\Big\| (v(\Delta t), v_t(\Delta t)) - (v_N^1, \overline{v}_N^1) \Big\|_{\widetilde{H}}}{\Delta t} \to 0 \, ,$$

as $N \to 0$. We get

$$\begin{aligned}
&\Big\| (v(\Delta t), v_t(\Delta t)) - (v_N^1, \overline{v}_N^1) \Big\|_{\widetilde{H}} \\
&\leq \Bigg\| \Bigg(\sum_{k=1}^{\alpha(N)} \Big[a_k(\cos(k\pi\Delta t) - \cos(\theta_k)) + b_k \left(\frac{\sin(k\pi\Delta t)}{k\pi} - \frac{\sin(\theta_k)}{\mu_k} \right) \Big] w_k, \\
&\qquad \sum_{k=1}^{\alpha(N)} \Big[-a_k(k\pi \sin(k\pi\Delta t) - \mu_k \sin(\theta_k)) + b_k(\cos(k\pi\Delta t) - \cos(\theta_k)) \Bigg) \Bigg\|_{\widetilde{H}} \\
&\quad + \Bigg\| \Bigg(\sum_{k=\alpha(N)+1}^{N} \Big[a_k \cos(\theta_k)) + b_k \frac{\sin(\theta_k)}{\mu_k} \Big] w_k, \\
&\qquad\qquad \sum_{k=\alpha(N)+1}^{N} \Big[-a_k \mu_k \sin(\theta_k) + b_k \cos(\theta_k) \Big] w_k \Bigg) \Bigg\|_{\widetilde{H}} \\
&\quad + \Bigg\| \Bigg(\sum_{k=\alpha(N)+1}^{\infty} \Big[a_k \cos(k\pi\Delta t) + b_k \frac{\sin(k\pi\Delta t)}{k\pi} \Big] w_k, \\
&\qquad\qquad \sum_{k=\alpha(N)+1}^{\infty} \Big[-a_k k\pi \sin(k\pi\Delta t) + b_k \cos(k\pi\Delta t) \Big] w_k \Bigg) \Bigg\|_{\widetilde{H}} \\
&= I_1 + I_2 + I_3 \, ,
\end{aligned}$$

where $\alpha(N) = \lfloor (N-1)^{1/3} \rfloor$ (other cut-off choices will also work).

We first show that $I_2/\Delta t \to 0$ and $I_2/\Delta t \to 0$ as $N \to \infty$. We do this by exploiting the fact that the "tail" of the series $\sum_{k=\alpha(N)+1}^{\infty}(k^2\pi^2 a_k^2 + b_k^2)$ decays exponentially as $N \to \infty$, because of the smoothness of the initial data. The calculations can be found in Detail 7, page 189.

We now turn to I_1 where the actual consistency is used. We initially rewrite the expression for I_1,

$$
\begin{aligned}
I_1^2 = \tfrac{1}{2}\sum_{k=1}^{\alpha(N)} \Bigg[& k^2\pi^2 a_k^2 \left(1 + \cos^2(\theta_k) + \frac{\mu_k^2 \sin^2(\theta_k)}{k^2\pi^2} - 2\cos(k\pi\Delta t)\cos(\theta_k)\right. \\
& \qquad \left. - 2\frac{\mu_k \sin(k\pi\Delta t)\sin(\theta_k)}{k\pi}\right) \\
& + b_k^2 \left(1 + \cos^2(\theta_k) + \frac{k^2\pi^2 \sin^2(\theta_k)}{\mu_k^2} - 2\cos(k\pi\Delta t)\cos(\theta_k)\right. \\
& \qquad \left. - 2\frac{k\pi \sin(k\pi\Delta t)\sin(\theta_k)}{\mu_k}\right) \\
& + 2k\pi a_k b_k \big(\cos(k\pi\Delta t) - \cos(\theta_k)\big)\sin(\theta_k)\left(\frac{\mu_k}{k\pi} - \frac{k\pi}{\mu_k}\right)\Bigg] \\
= \tfrac{1}{2}\sum_{k=1}^{\alpha(N)} & \Big(k^2\pi^2 a_k^2 A(kh) + b_k^2 B(kh) + 2k\pi a_k b_k F(kh)\Big),
\end{aligned}
$$

where A, B, F are functions of kh, since the expressions for $\cos(\theta_k)$, $\sin(\theta_k)$ and $k\pi/\mu_k$ can all be written in terms of kh only. Focusing on $A(y)$, we see that it is analytic for $y > 0$ and we write its Taylor expansion (easiest computed with the aid of a program for symbolic mathematics),

$$
A(kh) = \sum_{p=6}^{\infty} c_p (kh)^p ,
$$

where the real numbers c_p depend on η only. We observe that $0 < k \le (N-1)^{1/3} \Leftrightarrow 0 < kh \le (N-1)^{-2/3}$ and so we set $kh = \beta(N-1)^{-2/3}$ where $0 < \beta \le 1$. We now get

$$
\frac{A(kh)}{\Delta t^2} = \frac{(N-1)^2}{\eta^2} A\big(\beta(N-1)^{-2/3}\big) = \frac{1}{\eta^2}\sum_{p=6}^{\infty} c_p \beta^p (N-1)^{2-2p/3} ,
$$

clearly showing that $A(kh)/\Delta t^2 \to 0$ as $N \to \infty$ for all $0 < \beta \le 1$. This yields

$$
\frac{1}{\Delta t^2}\sum_{k=1}^{\alpha(N)} k^2\pi^2 a_k^2 |A(kh)| \le \max_{0 \le \beta \le 1} \frac{\big|A\big(\beta(N-1)^{-2/3}\big)\big|}{\Delta t^2} \sum_{k=1}^{\infty} k^2\pi^2 a_k^2 \to 0 ,
$$

as $N \to \infty$. The same procedure can be used for the terms involving $B(kh)$ and $F(kh)$, using in the latter case $|2k\pi a_k b_k| \le k^2\pi^2 a_k^2 + b_k^2$.

Gathering the results for I_1, I_2 and I_3 we finally get

$$\left\| \frac{(v(\Delta t), v_t(\Delta t)) - (v_N^1, \overline{v}_N^1)}{\Delta t} \right\|_{\widetilde{H}} \le \frac{I_1 + I_2 + I_3}{\Delta t} \to 0\,,$$

as $N \to \infty$, thus proving consistency.

We have now shown stability and consistency of the discretization scheme (7.22), and it then follows from the Lax Equivalence Theorem, Theorem 3.3.1, that the scheme is convergent.

7.4.3 Exact Controllability on a Fixed Level

The discrete control system can be written as

$$\begin{cases} \tilde{\boldsymbol{u}}^{n+1} = \boldsymbol{G}\tilde{\boldsymbol{u}}^n + \boldsymbol{F}\tilde{\boldsymbol{k}}^n\,, \\ \tilde{\boldsymbol{u}}^0 \text{ given}\,, \end{cases} \tag{7.26}$$

where

$$\begin{aligned} \boldsymbol{G} &= \left(\boldsymbol{I} - \tfrac{\Delta t}{2}\boldsymbol{S}\right)^{-1}\left(\boldsymbol{I} + \tfrac{\Delta t}{2}\boldsymbol{S}\right), \\ \boldsymbol{F} &= \tfrac{\Delta t}{2}\left(\boldsymbol{I} - \tfrac{\Delta t}{2}\boldsymbol{S}\right)^{-1}\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{B}\end{bmatrix}, \qquad \boldsymbol{S} = \begin{bmatrix}\boldsymbol{0} & \boldsymbol{I}\\ \boldsymbol{C}^{-1}\boldsymbol{A} & \boldsymbol{0}\end{bmatrix}, \end{aligned}$$

and

$$\tilde{\boldsymbol{u}}^n = \begin{bmatrix}\boldsymbol{u}^n\\ \overline{\boldsymbol{u}}^n\end{bmatrix}, \qquad \tilde{\boldsymbol{k}}^n = \boldsymbol{k}^{n+1} + \boldsymbol{k}^n\,.$$

We wish to show that we have exact controllability for this system, provided that M, the number of time steps, is large enough. This can be done by showing that the matrix

$$\boldsymbol{R}_N = \begin{bmatrix}\boldsymbol{F} & \boldsymbol{G}\boldsymbol{F} & \cdots & \boldsymbol{G}^{N-1}\boldsymbol{F}\end{bmatrix} \in \mathbb{R}^{2N\times 2N}$$

has full rank, see Theorem 4.1.4, page 70. This, in turn, can be done using Corollary 4.1.1. Let us recall some useful information from Section 3.1.2. We set

$$\begin{aligned} \boldsymbol{W} &= \begin{bmatrix}\boldsymbol{w}_1 & \boldsymbol{w}_2 & \cdots & \boldsymbol{w}_N\end{bmatrix}, \qquad \boldsymbol{D} = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_N), \\ \boldsymbol{Z} &= \begin{bmatrix}\boldsymbol{W} & \boldsymbol{W}\\ i\boldsymbol{W}\boldsymbol{D} & -i\boldsymbol{W}\boldsymbol{D}\end{bmatrix}, \end{aligned}$$

and get from Equation (3.10),

$$\boldsymbol{W}^{-1} = 2h\boldsymbol{W}, \quad \boldsymbol{Z}^{-1} = h\begin{bmatrix}\boldsymbol{W} & -i\boldsymbol{D}^{-1}\boldsymbol{W}\\ \boldsymbol{W} & i\boldsymbol{D}^{-1}\boldsymbol{W}\end{bmatrix}.$$

The matrix $\boldsymbol{Z}$ diagonalizes, by construction, $\boldsymbol{S}$, and it is seen to also diagonalize $\boldsymbol{G}$. We know from Section 3.2.3 that the eigenvalues of $\boldsymbol{G}$ are the $2N$ complex numbers given by

$$\frac{2+i\Delta t\mu_k}{2-i\Delta t\mu_k} \quad \text{and} \quad \frac{2-i\Delta t\mu_k}{2+i\Delta t\mu_k}, \quad \text{for } k=1,2,\ldots,N\,.$$

These are distributed on the unit circle in the complex plane, and are all distinct, since the numbers $\mu_1,\ \mu_2,\ \ldots,\ \mu_N$ are. This was the first condition of Corollary 4.1.1.

The next is checking whether the vector

$$\boldsymbol{f} = \boldsymbol{Z}^{-1}\boldsymbol{F}$$

contains any zeroes. We rewrite as follows,

$$\begin{aligned}
\left(\boldsymbol{I} - \tfrac{\Delta t}{2}\boldsymbol{S}\right)\boldsymbol{Z}\boldsymbol{f} &= \tfrac{\Delta t}{2}\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{B}\end{bmatrix},\\
\boldsymbol{Z}\left(\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix} - \tfrac{\Delta t}{2}\begin{bmatrix}i\boldsymbol{D} & \boldsymbol{0}\\ \boldsymbol{0} & -i\boldsymbol{D}\end{bmatrix}\right)\boldsymbol{f} &= \tfrac{\Delta t}{2}\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{B}\end{bmatrix},\\
\tfrac{2}{\Delta t}\left(\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix} - \tfrac{\Delta t}{2}\begin{bmatrix}i\boldsymbol{D} & \boldsymbol{0}\\ \boldsymbol{0} & -i\boldsymbol{D}\end{bmatrix}\right)\boldsymbol{f} &= \boldsymbol{Z}^{-1}\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{B}\end{bmatrix}.
\end{aligned}$$

The matrix on the left-hand side that pre-multiplies $\boldsymbol{f}$, is seen to be a diagonal matrix with non-zero entries on the diagonal. On the right-hand side is a vector with non-zero entries, since it is simply a scaled version of the last column of $\boldsymbol{Z}^{-1}$. We can now use these results to combine Corollary 4.1.1 and Theorem 4.1.4 into the following theorem.

Theorem 7.4.1. *The control system (7.26) is controllable, is the sense that any initial state* $(\boldsymbol{u}^0, \overline{\boldsymbol{u}}^0)$ *can be driven to any final state* $(\boldsymbol{u}^M, \overline{\boldsymbol{u}}^M)$*, if and only if* $M \geq 2N$.

Since we use a time step of Δt, we see that the time available for control must be

$$T \geq 2N\Delta t = 2\eta\frac{N}{N+1}\,,$$

where $\eta = \Delta t/h$. Recall that $T \geq 2$ is the requirement for the continuous system.

Note that this theorem addresses a single discretization level. But what happens when $\Delta t, h \to 0$? Does the controls obtained converge to the continuous one? This is the subject of the following section.

7.4.4 Striving Towards Uniform Observability

The subject of this section is the proof of the following theorem. The assumptions of the theorem will be assumed true for the remainder of the section.

Theorem 7.4.2. *Let $T > 2$, and let $\langle a_k\rangle_{k=1}^{\infty}$ and $\langle b_k\rangle_{k=1}^{\infty}$ be real sequences such that $\langle ka_k\rangle, \langle b_k\rangle \in \ell^2$. For each value of $N \in \mathbb{N}$ we set*

$$h = 1/(N+1), \quad M = \lceil T/(\eta h)\rceil, \quad \Delta t = T/M\,,$$

where $\eta < 1$ is a constant ($\Delta t \simeq \eta h$), and we have

$$\boldsymbol{v}^0 = \sum_{k=1}^{N} a_k \boldsymbol{w}_k, \qquad \overline{\boldsymbol{v}}^0 = \sum_{k=1}^{N} b_k \boldsymbol{w}_k$$

(note that all quantities depend on N, except for $\langle a_k\rangle_{k=1}^{\infty}$ and $\langle b_k\rangle_{k=1}^{\infty}$, but we will omit the extra subscripts for simpler notation). We now have for every value of N,

$$C_1 \left\| \begin{bmatrix} \boldsymbol{v}^0 \\ \overline{\boldsymbol{v}}^0 \end{bmatrix} \right\|_{\widetilde{\boldsymbol{Q}}}^2 \le \Delta t \sum_{n=0}^{M-1} \left| \boldsymbol{B}^T \frac{\boldsymbol{v}^{n+1} + \boldsymbol{v}^n}{2} \right|^2 \le C_2 \left\| \begin{bmatrix} \boldsymbol{v}^0 \\ \overline{\boldsymbol{v}}^0 \end{bmatrix} \right\|_{\widetilde{\boldsymbol{Q}}}^2 , \tag{7.27}$$

where $(\boldsymbol{v}^n, \overline{\boldsymbol{v}}^n)$ is a solution of the adjoint system (7.22).

We recall the expression from (3.31), page 45, for a solution of the adjoint system in terms of eigenvalues and eigenvectors, and get

$$\begin{aligned}
(\boldsymbol{v}^{n+1} + \boldsymbol{v}^n)/2 &= \tfrac{1}{2} \sum_{k=1}^{N} \big[a_k(\cos(n\theta_k) + \cos((n+1)\theta_k)) \\
&\qquad\qquad + b_k/\mu_k(\sin(n\theta_k) + \sin((n+1)\theta_k))\big] \boldsymbol{w}_k \\
&= \sum_{k=1}^{N} \big[\tilde{a}_k \cos(n\theta_k) + \tilde{b}_k \sin(n\theta_k)\big] \boldsymbol{w}_k\,,
\end{aligned}$$

where

$$\begin{aligned}
\tilde{a}_k &= \tfrac{1}{2}\big(a_k(\cos(\theta_k) + 1) + b_k/\mu_k \sin(\theta_k)\big), \\
\tilde{b}_k &= \tfrac{1}{2}\big(b_k/\mu_k(\cos(\theta_k) + 1) - a_k \sin(\theta_k)\big)\,,
\end{aligned}$$

using the addition formulas for sine and cosine. Note that

$$\tilde{a}_k^2 + \tilde{b}_k^2 = \cos^2(\tfrac{1}{2}\theta_k)(a_k^2 + b_k^2/\mu_k^2)\,. \tag{7.28}$$

Combining now that $\boldsymbol{B}^T \boldsymbol{v}^n = \boldsymbol{v}^n(N)/h$ and $\boldsymbol{w}_k(N) = \sin(Nkh\pi) = (-1)^k \sin(kh\pi)$ we get

$$\begin{aligned}
\boldsymbol{B}^T \frac{\boldsymbol{v}^{n+1} + \boldsymbol{v}^n}{2} &= \sum_{k=1}^{N} \frac{(-1)^k \sin(kh\pi)}{h} \big[\tilde{a}_k \cos(n\theta_k) + \tilde{b}_k \sin(n\theta_k)\big] \\
&= \sum_{1\le|k|\le N} c_k e^{i\tau_k n \Delta t}\,,
\end{aligned}$$

with

$$
\begin{aligned}
c_k &= \frac{(-1)^k \sin(kh\pi)}{2h}(\tilde{a}_k - i\tilde{b}_k), & c_{-k} &= \overline{c_k}, \\
\tau_k &= \theta_k/\Delta t, & \tau_{-k} &= -\tau_k \,,
\end{aligned}
$$

for $k = 1, 2, \ldots, N$. Before being able to apply Theorem 4.2.4, we have to show some properties of τ_1, τ_2, ..., τ_N, for each N. Note that this is the *most essential* part of the proof, since this is where all other discretization schemes fail. (As far as the author knows, that is. Exceptions are some schemes with $h = \Delta t$, see *Negreanu and Zuazua (2003)*).

Lemma 7.4.1. *We have for all N,*

$$
\begin{aligned}
\tau_1 &\geq \pi \,, \\
\tau_{k+1} - \tau_k &\geq \pi \,, \qquad\qquad \textit{for all } k = 1, 2, \ldots, N-1 \,, \\
\tau_N &\leq \pi/\Delta t - \pi \,.
\end{aligned}
$$

Proof. We set

$$
t(y) = \arccos\left(\frac{\cos^2(\frac{1}{2}y\pi) - \eta^2 \sin^2(\frac{1}{2}y\pi)}{\cos^2(\frac{1}{2}y\pi) + \eta^2 \sin^2(\frac{1}{2}y\pi)} \right) ,
$$

and note that $t(0) = 0$, $t(kh) = \theta_k$ for $k = 1, 2, \ldots, N$ and $t(1) = \pi$. By using a Taylor expansion it can be shown that

$$
t(y) = \eta y\pi + \mathcal{O}(y^3) \quad \Rightarrow \quad t(kh)/\Delta t = k\pi + \mathcal{O}(k^3h^2) \,,
$$

which shows that the slope of $t(kh)/\Delta t$ at $k = 0$ is equal to π for all h. Furthermore we have

$$
t''(y) = \frac{2(1-\eta^2)\pi^2\eta \tan(\frac{1}{2}y\pi)}{\big(1 + \eta^2 + (1-\eta^2)\cos(\pi y)\big)\big(1 + \eta^2 \tan^2(\frac{1}{2}y\pi)\big)} ,
$$

showing that $t''(y) \geq 0$ for $y \geq 0$ since we assume that $\eta < 1$. Using these facts and that $\tau_k = t(kh)/\Delta t$, the results follow. □

An illustration of $\tau_1, \ldots, \tau_N$ for the case $N = 40$ can be seen in Figure 7.3, for different values of $\eta = \Delta t/h$.

We can now apply Theorem 4.2.4 and get

$$
C_1' \sum_{1 \leq |k| \leq N} |c_k|^2 \leq \Delta t \sum_{n=0}^{M-1} \left| \boldsymbol{B}^T \frac{\boldsymbol{v}^{n+1} + \boldsymbol{v}^n}{2} \right|^2 \leq C_2' \sum_{1 \leq |k| \leq N} |c_k|^2 \,, \tag{7.29}
$$

for some positive constants C_1' and C_2' and for all $N \geq N_0$ (the value of N_0 is chosen to make sure that $C_1(T, \gamma, M, N)$ of Equation (4.45) is positive).

We now turn to look at

$$
\sum_{1 \leq |k| \leq N} |c_k|^2 \qquad \text{and} \qquad \left\| \begin{bmatrix} \boldsymbol{v}^0 \\ \overline{\boldsymbol{v}}^0 \end{bmatrix} \right\|_{\tilde{\boldsymbol{Q}}}^2 .
$$

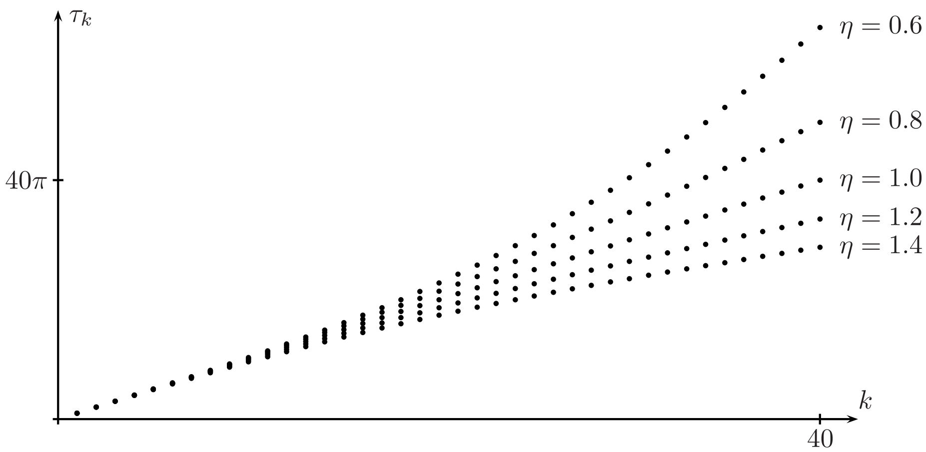


Figure 7.3: A plot of $\tau_k = \theta_k/\Delta t$ for the case $N = 40$. The slope of each curve is π in $k = 0$. Furthermore, the curves are convex for $\eta \leq 1$.

We will show that these quantities are of the same order, that is, provide upper and lower bounds that are independent of N.

First we get

$$\begin{aligned}\sum_{1\leq|k|\leq N} |c_k|^2 = 2\sum_{k=1}^{N}|c_k|^2 &= \tfrac{1}{2}\sum_{k=1}^{N}\frac{\sin^2(kh\pi)}{h^2}(\tilde{a}_k^2+\tilde{b}_k^2)\\ &= \tfrac{1}{2}\sum_{k=1}^{N}\frac{\sin^2(kh\pi)\cos^2(\frac{1}{2}\theta_k)}{h^2}(a_k^2+b_k^2/\mu_k^2),\\ &= \tfrac{1}{2}\sum_{k=1}^{N}\big(k^2\pi^2A_1^\eta(kh)a_k^2+B_1^\eta(kh)b_k^2\big)\end{aligned}$$

using (7.28) and where

$$\begin{aligned}A_1^\eta(x) &= \frac{1}{1+\eta^2\tan^2(\frac{1}{2}x\pi)}\left(\frac{\sin(x\pi)}{x\pi}\right)^2,\\ B_1^\eta(x) &= \frac{1}{1+\eta^2\tan^2(\frac{1}{2}x\pi)}\left(\frac{\sin(x\pi)}{2\tan(\frac{1}{2}x\pi)}\right)^2,\end{aligned} \tag{7.30}$$

since (7.24) leads to

$$2\cos^2(\tfrac{1}{2}\theta_k) = 1+\cos(\theta_k) = \frac{2}{1+\eta^2\tan^2(\frac{1}{2}kh\pi)}.$$

Note that we are only interested in the interval $0 < x < 1$, since $0 < kh < 1$ for all N.

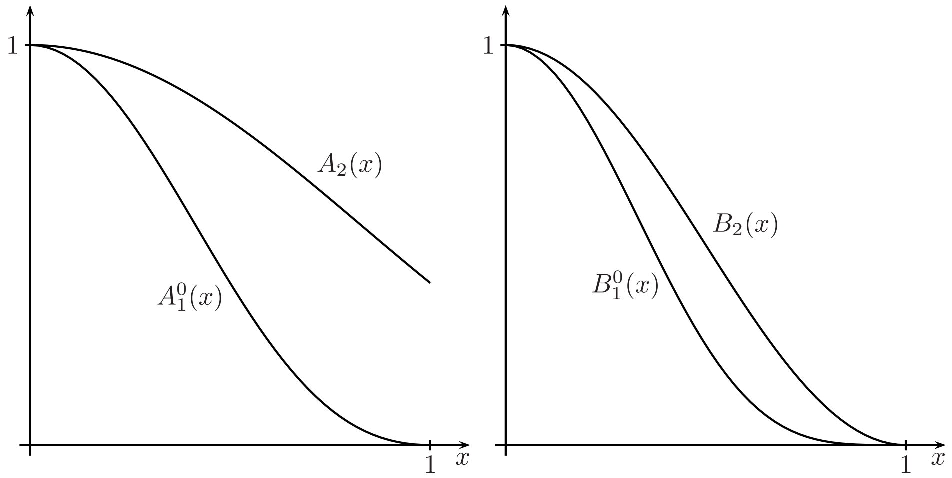


Figure 7.4: Plot of the functions A_1^0, A_2, B_1^0 and B_2 as defined in (7.30) and (7.31). They are all decreasing functions and A_2 and B_2 majorize A_1^0 and B_1^0, respectively. We here only consider the limiting case $\eta = 0$, but other values of η do not change the properties just mentioned.

For the $\widetilde{\boldsymbol{Q}}$-norm we get from (7.25),

$$\left\| \begin{bmatrix} \boldsymbol{v}^0 \\ \overline{\boldsymbol{v}}^0 \end{bmatrix} \right\|_{\widetilde{\boldsymbol{Q}}}^2 = \tfrac{1}{4} \sum_{k=1}^{N} \left(k^2\pi^2 \left(\frac{\sin(\frac{1}{2}kh\pi)}{\frac{1}{2}kh\pi} \right)^2 a_k^2 + \cos^2(\tfrac{1}{2}kh\pi) b_k^2 \right)$$
$$= \tfrac{1}{4} \sum_{k=1}^{N} \left(k^2\pi^2 A_2(kh) a_k^2 + B_2(kh) b_k^2 \right),$$

where

$$A_2(x) = \left(\frac{\sin(\frac{1}{2}x\pi)}{\frac{1}{2}x\pi} \right)^2, \qquad B_2(x) = \cos^2(\tfrac{1}{2}x\pi)\,. \tag{7.31}$$

See Figure 7.4 for an illustration of the functions A_1^0, B_1^0, A_2 and B_2. By the rewrite

$$B_1^\eta(x) = \cos^2(\tfrac{1}{2}x\pi) \frac{1}{1 + \eta^2 \tan^2(\frac{1}{2}x\pi)} \left(\frac{\sin(x\pi)}{2\sin(\frac{1}{2}x\pi)} \right)^2,$$

we clearly have $A_1^\eta(x) \le A_2(x)$ and $B_1^\eta(x) \le B_2(x)$ for $0 < x < 1$, leading to

$$\sum_{1 \le |k| \le N} |c_k|^2 \le 2 \left\| \begin{bmatrix} \boldsymbol{v}^0 \\ \overline{\boldsymbol{v}}^0 \end{bmatrix} \right\|_{\widetilde{\boldsymbol{Q}}}^2, \tag{7.32}$$

for all N. We must now establish the inverse inequality. Without loss of generality we can assume that $c_1 \neq 0$ (if $c_1 = \cdots = c_j = 0$ then (7.27) is trivially fulfilled for

$N \le j$). Naturally the quotient

$$\left\| \begin{bmatrix} \boldsymbol{v}^0 \\ \overline{\boldsymbol{v}}^0 \end{bmatrix} \right\|^2_{\widetilde{\boldsymbol{Q}}} \Bigg/ \sum_{1\le|k|\le N} |c_k|^2 \,,$$

is bounded for all $N < N_1$, where N_1 is some fixed natural number. The question is then, what happen with this quotient as $N \to \infty$? We first observe that $A_1^\eta(x)$ and $B_1^\eta(x)$ are decreasing functions in the interval $(0,1)$. We then have

$$\begin{aligned} B_1^\eta(x) &\ge B_1^\eta(\tfrac{1}{2}) = \frac{1}{4(1+\eta^2)}, \\ A_1^\eta(x) &\ge A_1^\eta(\tfrac{1}{2}) = \frac{4}{\pi^2(1+\eta^2)} \ge \frac{1}{4(1+\eta^2)} \,, \end{aligned}$$

for all $0 \le x \le \frac{1}{2}$, leading to

$$\begin{aligned} 2\frac{\left\| \begin{bmatrix} \boldsymbol{v}^0 \\ \overline{\boldsymbol{v}}^0 \end{bmatrix} \right\|^2_{\widetilde{\boldsymbol{Q}}}}{\displaystyle\sum_{1\le|k|\le N} |c_k|^2} &= \frac{\displaystyle\sum_{k=1}^{N} \big(k^2\pi^2 A_2(kh)a_k^2 + B_2(kh)b_k^2\big)}{\displaystyle\sum_{k=1}^{N} \big(k^2\pi^2 A_1^\eta(kh)a_k^2 + B_1^\eta(kh)b_k^2\big)} \\ &\le \frac{\displaystyle\sum_{k=1}^{N} \big(k^2\pi^2 a_k^2 + b_k^2\big)}{\displaystyle\frac{1}{4(1+\eta^2)} \sum_{k=1}^{\lceil N/2\rceil} \big(k^2\pi^2 a_k^2 + b_k^2\big)} \\ &\le 4(1+\eta^2) \frac{\displaystyle\sum_{k=1}^{\infty} \big(k^2\pi^2 a_k^2 + b_k^2\big)}{\displaystyle\sum_{k=1}^{\infty} \big(k^2\pi^2 a_k^2 + b_k^2\big) - \sum_{k=\lceil N/2\rceil+1}^{\infty} \big(k^2\pi^2 a_k^2 + b_k^2\big)} \\ &\to 4(1+\eta^2) \quad \text{for } N \to \infty \,. \end{aligned}$$

This shows that a constant C_1'', independent of N, exists such that

$$C_1'' \left\| \begin{bmatrix} \boldsymbol{v}^0 \\ \overline{\boldsymbol{v}}^0 \end{bmatrix} \right\|^2_{\widetilde{\boldsymbol{Q}}} \le \sum_{1\le|k|\le N} |c_k|^2 \,, \tag{7.33}$$

for all N. By combining (7.29), (7.32) and (7.33) we have now established the desired uniform inequality (7.27), but only for $N \ge N_0$. Taking now into consideration Theorem 7.4.1 of the previous section, we see that it holds for all N (by possibly adjusting the constants C_1, C_2 of the theorem).

Let us review what we have just shown. Let continuous data $(v^0, \overline{v}^0) \in \widetilde{H}$ be given,

$$v^0 = \sum_{k=1}^{\infty} a_k w_k, \qquad \overline{v}^0 = \sum_{k=1}^{\infty} b_k w_k \, ,$$

where $\langle k a_k \rangle, \langle b_k \rangle \in \ell^2$. And let discrete data,

$$\boldsymbol{v}_N^0 = \sum_{k=1}^{\infty} a_k^N \boldsymbol{w}_k, \qquad \overline{\boldsymbol{v}}_N^0 = \sum_{k=1}^{\infty} b_k^N \boldsymbol{w}_k \, ,$$

be given for every discretization level N. Theorem 7.4.2 now states that if we choose

$$a_k^N = a_k, \quad b_k^N = b_k, \quad \text{for } k = 1, 2, \ldots, N \, ,$$

for every N, the observability inequality (7.27) will hold with *uniform constants* for all N.

Does this result ensure convergence of controls as described in Section 4.2? Unfortunately not, since we do not have "true" uniform observability. The bounds in the observability inequality *depend on the sequences* $\langle a_k \rangle$ *and* $\langle b_k \rangle$.

The detailed treatment of the discretization scheme (7.22) of this section has been included for the following reasons. First, analysis of group velocity for this scheme looks very promising for controllability (see the dispersion relation (3.43) of Section 3.4). Second, this scheme is the only known fully discrete scheme with a uniform gap in the τ_k quantities (which makes it possible to apply the discrete version of Ingham's Theorem, see Lemma 7.4.1). Third, uniform observability may indeed hold, since a negative result has not yet been proved.

7.5 Other Schemes and Regularization Methods

Is there any hope for convergence of controls when using discretizations for which it is not possible to show uniform observability? Yes, if one if willing to lower the controllability requirements. Let us mention different ways of doing that.

Consider a finite difference semi-discretization of the wave equation in one dimension. This corresponds to the system $\boldsymbol{C}_0 \ddot{\boldsymbol{v}} = \boldsymbol{A}\boldsymbol{v}$, where $\boldsymbol{C}_0$ and $\boldsymbol{A}$ are defined in Section 3.1, page 31. As is hinted in Figure 3.3 of the same section, the distance/gap between the two largest square-rooted eigenvalues $\sqrt{-\lambda_k^0}$ goes to zero as $N \to \infty$. This lack of a uniform eigenvalue gap makes uniform observability impossible (see *Infante and Zuazua, 1999*, or *Zuazua, 2003*). What can be done is to only control a projection, as described in Section 2.6. By considering, at each discretization level, only the low frequency part of the spectrum, we obtain a uniform gap in the (square-rooted) eigenvalues and an observability inequality can be proved using Ingham's Theorem (see Theorem 4.2.1). One downside to this approach, of course, is that only a spectral projection is controlled. Another downside is that the minimal control time, which is two is the continuous case, depends on the projection chosen and will be *too large* (see *Infante and Zuazua, 1999*, or *Zuazua, 2003*, for the exact minimal time requirements). Note that all of the above could be said for every semi-discretization $\boldsymbol{C}_\alpha \ddot{\boldsymbol{v}} = \boldsymbol{A}\boldsymbol{v}$ with $\alpha < 1/4$.

The projection method just described can be considered a regularization method. In the field of discrete ill-posed problems this method is similar to what is called the TSDV, the Truncated Singular Value Decomposition (the TSVD method is typically used for other reasons, though). As previously noted in Section 2.7, another regularization method has been considered in the literature, namely that of using the following regularized controllability operator,

$$\Lambda_T^\alpha = \Lambda_T + \alpha \begin{bmatrix} -\Delta & 0 \\ 0 & I \end{bmatrix},$$

see *Glowinski and Li (1990)*, *Glowinski, Li, and Lions (1990)* and *Glowinski (1992b)*. It has not, however, been proved that the corresponding controls converge as the discretization grid gets finer and finer, $N \to \infty$ (and it probably does not hold).

Another method, which can also be called a regularization method, is the so-called two-grid/multigrid method. As the name suggests, the method uses two (space) grids, a coarse and a fine grid. The basic idea is a small change in the definition of the controllability operator: The input to Λ_T is posed on the coarse grid, then interpolated onto the fine grid, the usual Λ_T-mapping is done on the fine grid, and the end-result is finally restricted onto the coarse grid again. This method was first suggested in *Glowinski (1992b)*. The justification for the method is, roughly, that the high-frequency components on the fine grid are minimal since the state on the fine grid comes from a state on the coarse grid. The method has later been used in *Asch and Lebeau (1998)*, where different numerical experiments where conducted for the wave equation in two dimensions, using a finite difference discretization in both time and space, the CG algorithm and the two-grid method. It has lately been proved, in the semi-discrete case, that the two-grid method actually leads to uniform observability, but with a minimal control time which is *twice* the correct time (see *Negreanu and Zuazua, 2004a*).

A completely different approach can be chosen when uniform observability does not hold. Simply keep the discretization method, but apply restrictions to the initial conditions which can be controlled. This approach was taken in *Micu (2002)* for the finite difference semi-discretization $\boldsymbol{C}_0 \ddot{\boldsymbol{v}} = \boldsymbol{A}\boldsymbol{v}$ of the wave equation in one dimension. One of the conclusions were that if the initial conditions are analytic, then computed controls will converge to the true control as $N \to \infty$.

Let us finally mention that the authors in *Glowinski, Kinton, and Wheeler (1989)* used a so-called *mixed finite element* discretization method to solve exact controllability problems for the two-dimensional wave equation. The paper was of experimental nature and convergence of controls was not proved.

For a survey of many of the methods for obtaining convergent controls can be found in *Zuazua (2003)*. This note, currently unpublished, also contains results on the heat equation and many interesting open problems. See also *Zuazua (2004)* for similar results. The paper *Rasmussen (2003)* contains a practical comparison of the convergence of controls for many of the methods mentioned above.

7.6 Other Theoretical Results

We previously used a *multiplier technique* to show the direct inequality for the wave equation in one dimension (see inequality (7.3)). This technique can also be used to show the direct inequality in more dimensions, but also to show the inverse inequality, that is, the observability inequality for exact controllability. Some requirements of the domain $\Omega \subset \mathbb{R}^d$ and control boundary $\Gamma_0 \subset \partial\Omega$ must be met, though. For instance, that a point $x_0 \in \mathbb{R}^d$ exists such that $n(x) \cdot (x - x_0) \geq 0$ for all points on the control boundary, $x \in \Gamma_0$. The multiplier technique can also lead to upper bounds on the minimal control time. See, e.g., *Lions (1988b)*, *Komornik (1994)* or *Pedersen (2000)* for more information on the multiplier method.

Let us finish this chapter by mentioning a deep result concerning the continuous wave equation in any dimension. The proof relies on microlocal analysis and is far from the established theory of this thesis. We include the theorem here, however, because of its simple formulation and great importance.

The result was first formulated and proved in *Bardos, Lebeau, and Rauch (1992)*, and has since been referred to in many publications, for instance, *Asch and Lebeau (1998)* and *Zuazua (2003)*.

Theorem 7.6.1 (Geometric Control Condition). *Let $\Omega \subset \mathbb{R}^d$ be a class C^∞ domain with control boundary $\Gamma_0 \subset \partial\Omega$. Exact boundary controllability of the wave equation holds at time T if and only if every ray of geometric optics, propagating in Ω and reflecting on its boundary $\partial\Omega$, intersects with the control boundary Γ_0 in time less than T.*

This theorem provides a quite intuitive characterization of the domains for which exact controllability is possible. Furthermore, for most domains it is easy to see what the minimal control time is.

Note that if the whole of the boundary of a domain Ω is used as control boundary, $\Gamma_0 = \Gamma = \partial\Omega$, we *always* have controllability for large enough T. This follows from the above theorem and the fact that the domain Ω is assumed to be bounded (this statement does not generally hold if we consider the wave equation with variable coefficients, since the geometric rays may not be straight lines anymore).

Figure 7.5 shows some simple two-dimensional domains for which exact controllability is impossible, due to trapped rays that never reach the control boundary. Similarly, Figure 7.6 shows some two-dimensional domains that all fulfill the Geometric Control Condition.

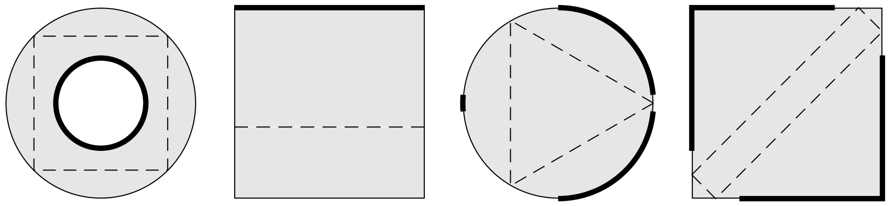


Figure 7.5: Illustration of 2D domains that do not fulfill the Geometric Control Condition. The thickly drawn part of the boundary denotes the control boundary and dashed lines show why controllability fails: they are geometric rays that never reach the control boundary.

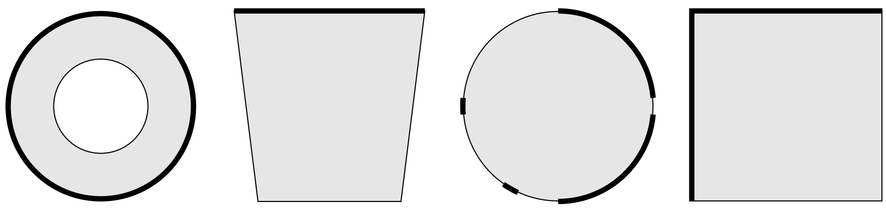


Figure 7.6: Illustration of 2D domains that do fulfill the Geometric Control Condition. The thickly drawn part of the boundary denotes the control boundary.

Chapter 8

A Linear System of Thermoelasticity

Theory can leave questions unanswered,
but practice has to come up with something
— Mason Cooley

This chapter is dedicated to studying a linear system of thermoelasticity in one dimension. The system is a coupling of a wave equation and a heat equation, and it can be thought of as an oscillating string, to which temperature is assigned in each point. The coupling terms determine how (local) oscillations affect the temperature, and vice versa. We will, however, not get into the physical derivation, or meaning, of this system, but strictly study it with respect to controllability.

Let the control time be $T > 0$, the spacial domain $\Omega = (0, 1)$, and the linear system of thermoelasticity has the following appearance,

$$\begin{cases} u_{tt} = c^2 u_{xx} - \alpha\theta_{xx} & \text{in } (0,T) \times (0,1)\,, \\ \theta_t \ = \nu\theta_{xx} - \beta u_t & \text{in } (0,T) \times (0,1)\,, \\ u(\cdot,0) = 0, \quad u(\cdot,1) = k, \quad \theta(\cdot,0) = \theta(\cdot,1) = 0 & \text{in } (0,T)\,, \\ u(0,\cdot) = u^0, \quad u_t(0,\cdot) = u^1, \quad \theta(0,\cdot) = \theta^0 & \text{in } (0,1)\,. \end{cases} \tag{8.1}$$

The constants c, ν, α, β are all assumed positive and to ensure well-posedness, we require

$$(u^0, u^1, \theta^0) \in L^2(0,1) \times H^{-1}(0,1) \times L^2(0,1) = H' \quad \text{and} \quad k \in L^2(0,T)\,,$$

which implies a solution $t \mapsto (u(t), u_t(t), \theta(t)) \in C([0,T], H')$. The above system is the control system, with the control k acting on only one of the variables, $\mathcal{B}\,(u^0, u^1, \theta^0) \mapsto u^0$, on the control boundary $\Gamma_0 = \{1\}$.

The objective is the following: Given $T > 0$ and initial data $(u^0, u^1, \theta^0) \in H'$, find $k \in L^2(0,T)$ such that

$$u(T,\cdot) = 0, \quad u_t(T,\cdot) = 0, \quad \theta(T,\cdot) = 0\,.$$

If, for fixed T, this is possible for all initial data in H', the system is *null-controllable* at time T.

This chapter is dedicated to proving that the above control system is in fact null-controllable. The proof is based on that in *Lebeau and Zuazua (1998)*, where the same system (in any dimension) was considered, but where the control was *internal*. This means that the boundary conditions were all homogeneous, but a source term was added to the first equation such as

$$u_{tt} - c^2 u_{xx} + \alpha\theta_{xx} = \chi_{\Omega_0} f\,,$$

where χ_{Ω_0} has value 1 in the control region $\Omega_0 \subset \Omega$, and 0 elsewhere.

Results similar to those we present here have also been obtained in the paper *Hansen (1994)*. In that paper, boundary null-controllability was proved for a related thermoelastic system in one dimension. See also *Zuazua (1995)*.

8.1 Well-posedness

Before deriving the adjoint system, let us first rewrite the control system (8.1) into a first order system,

$$U_t(t) = \mathcal{A}U(t)\,,$$

where

$$U(t) = \begin{pmatrix} u(t) \\ u_t(t) \\ \theta(t) \end{pmatrix} \quad \text{and} \quad \mathcal{A} = \begin{pmatrix} 0 & I & 0 \\ c^2\partial_{xx} & 0 & -\alpha\partial_{xx} \\ 0 & -\beta I & \nu\partial_{xx} \end{pmatrix}\,, \tag{8.2}$$

and the boundary and initial conditions are unchanged. We can easily figure out the adjoint operator (since $\langle u_{xx}, v\rangle = \langle u, v_{xx}\rangle$ when we consider homogeneous boundary conditions),

$$\mathcal{A}^* = \begin{pmatrix} 0 & c^2\partial_{xx} & 0 \\ I & 0 & -\beta I \\ 0 & -\alpha\partial_{xx} & \nu\partial_{xx} \end{pmatrix}\,,$$

But what system does this operator represent? Using the auxiliary variables (y, v, ψ), and introducing a minus in each equation because we want to solve the system backwards in time, we get

$$\begin{cases} y_t = -c^2 v_{xx}\,, \\ v_t = -y + \beta\psi\,, \\ \psi_t = \alpha v_{xx} - \nu\psi_{xx}\,. \end{cases}$$

What do the variables y, v, ψ represent, and in what space is this system well posed? Rewriting as follows,

$$v_{tt} = -y_t + \beta\psi_t = (c^2 + \alpha\beta)v_{xx} - \nu\beta\psi_{xx}\,,$$

we get a more convenient system,

$$\begin{cases} v_{tt} = (c^2+\alpha\beta)v_{xx} - \nu\beta\psi_{xx} & \text{in } (0,T)\times(0,1)\,, \\ \psi_t \;= -\nu\psi_{xx} + \alpha v_{xx} & \text{in } (0,T)\times(0,1)\,, \\ v(\cdot,0) = v(\cdot,1) = \psi(\cdot,0) = \psi(\cdot,1) = 0 & \text{in } (0,T)\,, \\ v(T,\cdot) = v^0, \quad v_t(T,\cdot) = v^1, \quad \psi(T,\cdot) = \psi^0 & \text{in } (0,1)\,, \end{cases} \tag{8.3}$$

which will be our *adjoint system*. Note how it, like the control system, is a coupling of a wave equation and a heat equation.

It will also be convenient to have the adjoint system in first order form,

$$V_t(t) = -\widetilde{\mathcal{A}}V(t)\,,$$

where

$$V(t) = \begin{pmatrix} v(t) \\ v_t(t) \\ \psi(t) \end{pmatrix} \quad \text{and} \quad \widetilde{\mathcal{A}} = \begin{pmatrix} 0 & -I & 0 \\ -(c^2+\alpha\beta)\partial_{xx} & 0 & \nu\beta\partial_{xx} \\ -\alpha\partial_{xx} & 0 & \nu\partial_{xx} \end{pmatrix}.$$

Note that

$$\widetilde{\mathcal{A}} = \mathcal{M}^{-1}\mathcal{A}^*\mathcal{M}\,, \quad \text{with} \quad \mathcal{M} = \begin{pmatrix} 0 & -1 & \beta \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where $\mathcal{M}$ can be used to "change back" to the variables (y, v, ψ), which were originally used for setting up the adjoint system.

We can now introduce the Hilbert space $\widetilde{H}$ with the norm

$$\|V\|^2_{\widetilde{H}} = \|\mathcal{M}V\|^2_H = \|\beta\psi^0 - v^1\|^2_{L^2(0,1)} + \|v^0\|^2_{H^1_0(0,1)} + \|\psi^0\|^2_{L^2(0,1)}\,,$$

for $V = (v^0, v^1, \psi^0)$ and the duality pairing

$$\begin{aligned} \{U, V\} &= \langle U, \mathcal{M}V\rangle_{H'\times H} \\ &= \langle u^0, \beta\psi^0 - v^1\rangle_{L^2(0,1)} + \langle u^1, v^0\rangle_{H^{-1}(0,1)\times H^1_0(0,1)} + \langle \theta^0, \psi^0\rangle_{L^2(0,1)}\,, \end{aligned}$$

for every $U = (u^0, u^1, \theta^0) \in H'$ and $V = (v^0, v^1, \psi^0) \in \widetilde{H}$.

The "adjointness" is now clear from the following relation,

$$\{AU, V\} = \langle AU, \mathcal{M}V\rangle = \langle U, A^*\mathcal{M}V\rangle = \langle U, \mathcal{M}\widetilde{\mathcal{A}}V\rangle = \{U, \widetilde{\mathcal{A}}V\}\,, \tag{8.4}$$

for all $U = (u^0, u^1, \theta^0) \in H'$ and $V = (v^0, v^1, \psi^0) \in \widetilde{H}$.

Let us consider the well-posedness of the adjoint system. We define the energy of the adjoint system as

$$E(t) = \frac{1}{2}\int_0^1 \left(|\beta\psi(t,x) - v_t(t,x)|^2 + c^2|v_x(t,x)|^2 + \frac{c^2\beta}{\alpha}|\psi(t,x)|^2 \right) dx\,,$$

for $0 \le t \le T$. By differentiating this expression,

$$E'(t) = \frac{c^2\beta\nu}{\alpha}\int_0^1 |\psi_x(t,x)|^2 dx \ge 0\,,$$

one sees that $0 \le E(t) \le E(T)$ for $0 \le t \le T$.

Note how $(v^0, v^1, \psi^0) \in \widetilde{H}$ is equivalent to $E(T) < \infty$ (with E depending appropriately on the initial data (v^0, v^1, ψ^0)). This implies that the adjoint system is well posed with initial data in $\widetilde{H}$, yielding a solution $t \mapsto (v(t), v_t(t), \psi(t)) \in C([0,T]; \widetilde{H})$.

The adjoint system furthermore has the important property

$$\int_0^T |v_x(t,1)|^2 dt \le K(T)\|(v^0, v^1, \psi^0)\|^2_{\widetilde{H}}\,,$$

for any solution $(v(t), v_t(t), \psi(t))$ of the adjoint system with initial conditions $(v^0, v^1, \psi^0) \in \widetilde{H}$. This can be shown by the use of multipliers (combine the equations of the adjoint system into $v_{tt} - \beta\psi_t = c^2 v_{xx}$, apply the multiplier $v_x x$ and integrate over $(0,T) \times (0,1)$; see Detail 8, page 190, for the derivation).

We now want to derive the complementary boundary operator $\mathcal{C}$. We consider $\mathcal{A}$ with boundary conditions as for the control system (8.1), and we get for $U = (u^0, u^1, \theta^0)^T \in H$ and $V = (v^0, v^1, \psi^0)^T \in \widetilde{H}$,

$$\begin{aligned}
&\{AU, V\} - \{U, \widetilde{\mathcal{A}}V\} \\
&= \left\{ \begin{pmatrix} u^1 \\ c^2 u^0_{xx} - \alpha\theta^0_{xx} \\ \nu\theta^0_{xx} - \beta u^1 \end{pmatrix}, \begin{pmatrix} v^0 \\ v^1 \\ \psi^0 \end{pmatrix} \right\} - \left\{ \begin{pmatrix} u^0 \\ u^1 \\ \theta^0 \end{pmatrix}, \begin{pmatrix} -v^1 \\ -(c^2+\alpha\beta)v^0_{xx} + \nu\beta\psi^0_{xx} \\ \nu\psi^0_{xx} - \alpha v^0_{xx} \end{pmatrix} \right\} \\
&= c^2\Big[\langle u^0_{xx}, v^0\rangle - \langle u^0, v^0_{xx}\rangle\Big] - \alpha\Big[\langle \theta^0_{xx}, v^0\rangle - \langle \theta^0, v^0_{xx}\rangle\Big] + \nu\Big[\langle \theta^0_{xx}, \psi^0\rangle - \langle \theta^0, \psi^0_{xx}\rangle\Big] \\
&= -c^2 u^0(1) v^0_x(1)\,,
\end{aligned} \tag{8.5}$$

which shows that $\mathcal{C}\,(v^0, v^1, \psi^0) = -c^2 v^0_x(1)$.

We can now show the well-posedness of the control system (8.1) in the following way,

$$\begin{aligned}
\|U(T)\|_{H'} &= \sup_{W^0\in H} \frac{\big|\langle U(T), W^0\rangle_{H'\times H}\big|}{\|W^0\|_H} \\
&= \sup_{W^0\in H} \frac{\big|\{U(T), \mathcal{M}^{-1}W^0\}\big|}{\|\mathcal{M}^{-1}W^0\|_{\widetilde{H}}} = \sup_{V^0\in \widetilde{H}} \frac{\big|\{U(T), V^0\}\big|}{\|V^0\|_{\widetilde{H}}} \\
&\le \sup_{V^0\in \widetilde{H}} \frac{1}{\|V^0\|_{\widetilde{H}}} \left(\big|\{U(0), V(0)\}\big| + c^2 \left| \int_0^T k(t) v_x(t,1)\,dt \right| \right) \\
&\le \sup_{V^0\in \widetilde{H}} \frac{1}{\|V^0\|_{\widetilde{H}}} \Big(\|U(0)\|_{H'} \|V(0)\|_{\widetilde{H}} + c^2 \|k\|_{L^2(0,T)} \|v_x(\cdot,1)\|_{L^2(0,T)} \Big) \\
&\le C\|U(0)\|_{H'} + c^2 K(T) \|k\|_{L^2(0,T)}\,.
\end{aligned}$$

These computations are of course valid for every $T > 0$.

8.2 Spectral Properties

In this section we will study the spectral properties of the operator $\mathcal{A}$, defined in (8.2).

Since every entry of $\mathcal{A}$ is a scaling of either the Laplace operator or the identity, we can base our study on the one-dimensional Laplace eigenproblem,

$$-\partial_{xx} e_j = \omega_j^2 e_j\,,$$

on $(0,1)$ with homogeneous boundary conditions. We have the obvious solutions,

$$e_j(x) = \sin(j\pi x), \quad \omega_j = \pi j, \quad \text{for } j = 1, 2, \ldots.$$

The eigenproblem

$$\mathcal{A} U_j = \lambda_j U_j\,,$$

with $U_j = \boldsymbol{z}_j e_j$, $\boldsymbol{z}_j \in \mathbb{R}^3$, is seen to be equivalent to the eigenproblem

$$\mathcal{A}_j \boldsymbol{z}_j = \lambda_j \boldsymbol{z}_j\,,$$

where

$$\mathcal{A}_j = \begin{pmatrix} 0 & 1 & 0 \\ -c^2 w_j^2 & 0 & \alpha w_j^2 \\ 0 & -\beta & -\nu w_j^2 \end{pmatrix}. \tag{8.6}$$

We now seek the roots of the characteristic polynomial of $\mathcal{A}_j$,

$$\det(\lambda_j I - \mathcal{A}_j) = (\lambda_j^2 + c^2\omega_j^2)(\lambda_j + \nu\omega_j^2) + \alpha\beta\lambda_j\omega_j^2 = 0\,. \tag{8.7}$$

The solutions, the eigenvalues, will be split into two parts, a parabolic part and a hyperbolic part, to be treated separately.

Parabolic eigenvalues λ_j^p. Expecting eigenvalues close to those of the heat equation part of (8.1), we let $\lambda_j^p = -\nu\omega_j^2 + \epsilon_j$ and insert into (8.7),

$$\big((\epsilon_j - \nu\omega_j^2)^2 + c^2\omega_j^2\big)\,\epsilon_j + \alpha\beta\omega_j^2(\epsilon_j - \nu\omega_j^2) = 0\,, \tag{8.8}$$

leading to

$$\epsilon_j = \frac{\alpha\beta(\nu - \epsilon_j/\omega_j^2)}{(\nu - \epsilon_j/\omega_j^2)^2 + c^2/\omega_j^2} \quad\Leftrightarrow\quad X(\epsilon_j, y_j) = \epsilon_j - \frac{\alpha\beta(\nu - \epsilon_j y_j)}{(\nu - \epsilon_j y_j)^2 + c^2 y_j} = 0\,,$$

where $y_j = 1/\omega_j^2$. We see that this equation is fulfilled for $(\epsilon_j, y_j) = (\alpha\beta/\nu, 0)$ in which $\partial X(\alpha\beta/\nu, 0)/\partial\epsilon_j = 1 \neq 0$. Therefore, by the Implicit Function Theorem, in a neighborhood around this point we have $\epsilon_j = Z(y_j)$. The function X is also seen to be analytic around this point, so we have $\epsilon_j = \alpha\beta/\nu + \mathcal{O}(y_j)$. Observe now that the coefficients in (8.8) are real. This means that if $\epsilon_j = Z(y_j)$ is a non-real root of (8.8) (with $\omega_j^2 = 1/y_j$) then we must also have $\overline{\epsilon_j} = Z(y_j)$, which is a contradiction. Hence, for sufficiently small y_j, the value of $Z(y_j)$ is real, and we have

$$\lambda_j^p = -\nu\omega_j^2 + \frac{\alpha\beta}{\nu} + \mathcal{O}\left(\frac{1}{\omega_j^2}\right) \in \mathbb{R}\,.$$

Hyperbolic eigenvalues $\lambda_j^{h,\pm}$. The remaining two roots must be complex conjugates of each other, since the coefficients of the characteristic polynomial are real. Thus $\lambda_j^{h,+} = \overline{\lambda_j^{h,-}}$ and we use the convention Im $\lambda_j^{h,+} > 0$ (we use Im to refer to the imaginary part). Here we expect eigenvalues close to those of the wave equation-part of (8.1) and we set $\lambda_j^{h,+} = ic\omega_j + \eta_j$. Equation (8.7) turns into

$$(\eta_j^2 + 2i\eta_j c\omega_j)(ic\omega_j + \nu\omega_j^2 + \eta_j) + i\alpha\beta c\omega_j^3 + \eta_j\alpha\beta\omega_j^2 = 0\,,$$

leading to

$$\eta_j = \frac{-\alpha\beta(\eta_j/\omega_j + ic)}{(\eta_j/\omega_j + 2ic)(ic/\omega_j + \nu + \eta_j/\omega_j^2)}\,.$$

Applying the Implicit Function Theorem as for the parabolic case, we get the asymptotic estimate

$$\lambda_j^{h,+} = ic\omega_j - \frac{\alpha\beta}{2\nu} + \mathcal{O}\left(\frac{1}{\omega_j}\right).$$

Let now an $R > 0$ be given. We will denote the subset of $\{\lambda_j^p \mid j \in \mathbb{N}\}$, whose elements are larger than R in magnitude, the *parabolic eigenvalues*, and similarly the *hyperbolic eigenvalues* will be the subset of $\{\lambda_j^{h,\pm} \mid j \in \mathbb{N}\}$ whose elements are larger then R in magnitude. See Figure 8.1.

We will assume that R is chosen so large that no multiple eigenvalues occur among neither the parabolic nor the hyperbolic eigenvalues. This is possible, as can be seen from the asymptotic expressions for the parabolic/hyperbolic eigenvalues above. Multiple eigenvalues *can* occur, though, among the eigenvalues with

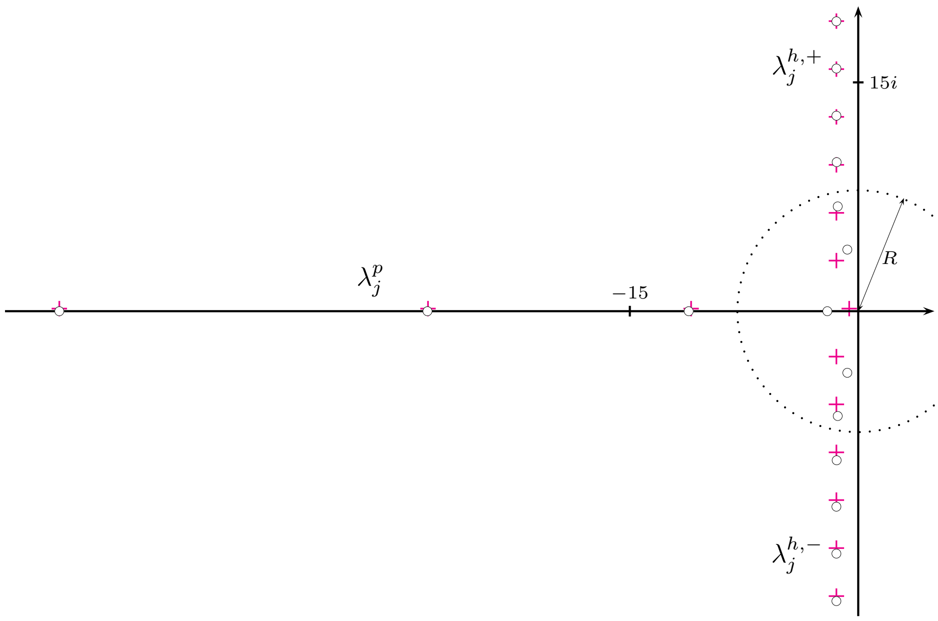
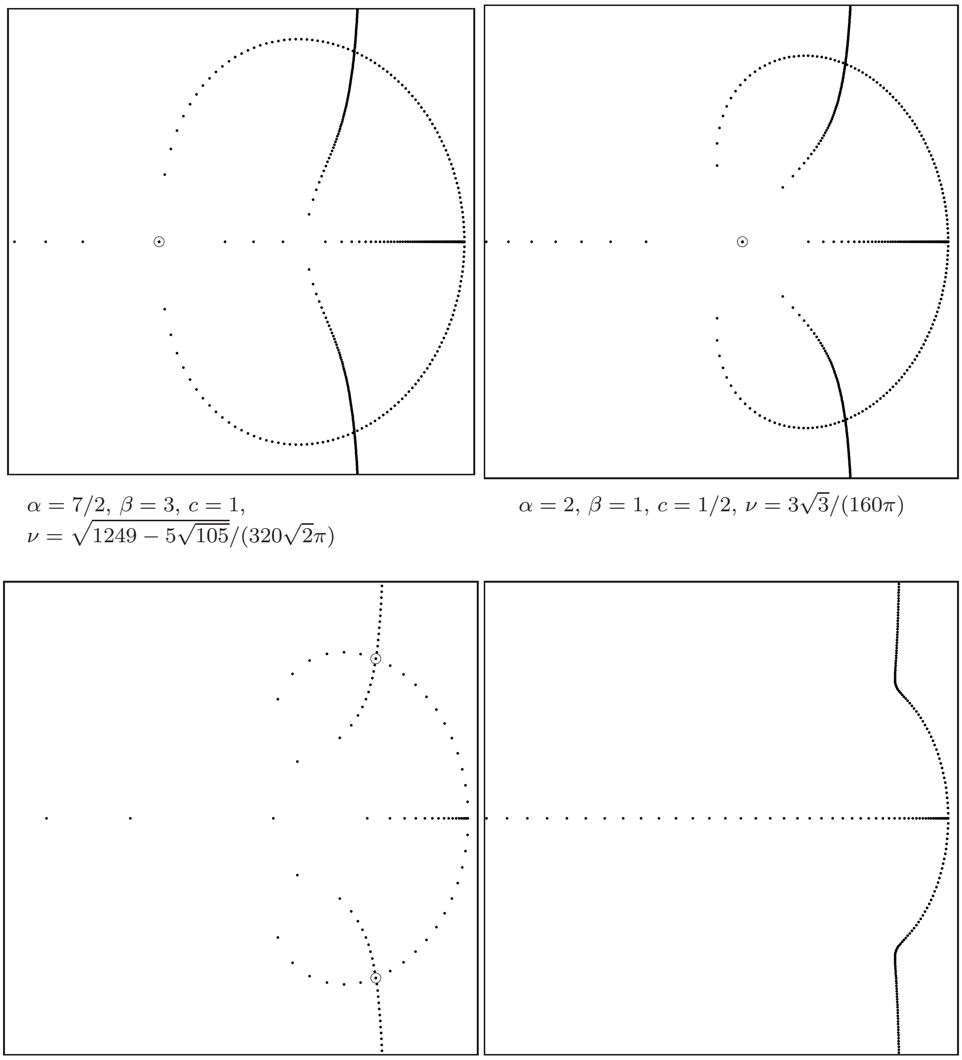


Figure 8.1: Eigenvalues of the operator $\mathcal{A}$ for the case $\alpha = \beta = c = 1$, $\nu = 0.35$ are shown with circles. The crosses show the points $-\nu\omega_j^2$ and $\pm ic\omega_j$, $j = 1, 2, \ldots$, to which the parabolic eigenvalues λ_j^p and the hyperbolic eigenvalues $\lambda_j^{h,\pm}$, respectively, approaches asymptotically.



$\alpha = 7/2$, $\beta = 3$, $c = 1$,
$\nu = \sqrt{1249 - 5\sqrt{105}}/(320\sqrt{2}\pi)$

$\alpha = 2$, $\beta = 1$, $c = 1/2$, $\nu = 3\sqrt{3}/(160\pi)$

$\alpha = 741/100$, $\beta = 1$, $c = 1$, $\nu = 29\sqrt{29}/(600\pi)$

$\alpha = 2$, $\beta = 1$, $c = 1$, $\nu = 1/50$

Figure 8.2: Eigenvalues of the operator $\mathcal{A}$ for different choices of α, β, c and ν, shown in the complex plane (the real line is centered vertically and the imaginary line is at the right-hand side for all four plots). Dots surrounded by circles indicate *multiple* eigenvalues; the two plots on the left contain eigenvalues with multiplicity two, and the top-right plot contains an eigenvalue with multiplicity three.

magnitude less than R, by appropriately choosing the parameters $c, \nu, \alpha, \beta > 0$ and $\omega_j = \pi j$. See Figure 8.2 for some examples of eigenvalues with multiplicities two and three.

We will now introduce some subspaces of H. The space $H^p \subset H$ will denote the Hilbert space spanned by the eigenvectors corresponding to the parabolic eigenvalues. Similarly, the space $H^h \subset H$ denotes the Hilbert space spanned by the eigenvectors corresponding the the hyperbolic eigenvalues. Finally, $H^0 \subset H$ denotes the "rest", that is, we have $H = H^p \oplus H^h \oplus H^0$. Note that H^0 is finite dimensional, since only a finite number of eigenvalues can have magnitude less than or equal to R.

Finally, we need some projection operators, namely the orthogonal projections $\Pi^p : H \mapsto H^p$, $\Pi^h : H \mapsto H^h$, $\Pi^0 : H \mapsto H^0$, defined in the obvious "spectral" way.

The eigenvalues of $\mathcal{A}$ and $\widetilde{\mathcal{A}}$ are identical (since $\widetilde{\mathcal{A}}_j$, created analogous to (8.6), and $\mathcal{A}_j$ have identical eigenvalues). We will define the space $\widetilde{H}^p \subset \widetilde{H}$ as the span of the eigenvectors of the adjoint system, corresponding to the parabolic eigenvalues. Let the parabolic eigenvectors be represented by $V_j^p e_j$, $V_j^p \in \mathbb{C}^3$. Now, using the fact that $\widetilde{\mathcal{A}} V_j^p e_j = \lambda_j^p V_j^p e_j$ and the asymptotic expression for λ_j^p, we get

$$V_j^p = \begin{pmatrix} 1/\omega_j^2 \\ \nu + \mathcal{O}\Big(\dfrac{1}{\omega_j^2}\Big) \\ \dfrac{\nu}{\beta} + \mathcal{O}\Big(\dfrac{1}{\omega_j^2}\Big) \end{pmatrix} . \tag{8.9}$$

The span of the eigenvectors of the adjoint system, corresponding to the hyperbolic eigenvalues, will similarly be denoted $\widetilde{H}^h \subset \widetilde{H}$. These eigenvectors $V_j^{h,\pm} e_j$ have the asymptotic representation,

$$V_j^{h,\pm} = \begin{pmatrix} 1/\omega_j \\ \mp ic + \mathcal{O}\Big(\dfrac{1}{\omega_j}\Big) \\ \mathcal{O}\Big(\dfrac{1}{\omega_j}\Big) \end{pmatrix} . \tag{8.10}$$

We will now argue that the set of eigenvectors,

$$\{V_j^p e_j\}_{|\lambda_j^p| > R} \quad \cup \quad \{V_j^{h,\pm} e_j\}_{|\lambda_j^{h,\pm}| > R} ,$$

supplemented with appropriate *generalized eigenvectors* of $\widetilde{\mathcal{A}}$, constitute a Riesz basis for $\widetilde{H}$.

A non-zero vector $f \in \widetilde{H}$ is called a generalized eigenvector of $\widetilde{\mathcal{A}}$, corresponding to some eigenvalue λ, if $(\lambda I - \widetilde{\mathcal{A}})^m f = 0$ for some positive integer m. We now have the following theorem from *Zhang and Zuazua (2003b)*, see also *Guo and Yu (2001)*.

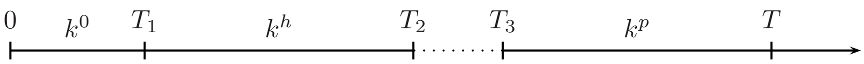


Figure 8.3: Subdivision of the time line from time 0 to T. The control is always zero in the interval (T_2, T_3), indicated with a dashed line.

Theorem 8.2.1. *Let G be a densely defined linear operator with compact resolvent in H. Let $\langle f_n\rangle_{n=1}^{\infty}$ be a Riesz basis of H. Suppose a sequence of generalized eigenvectors $\langle g_n\rangle_{n=M+1}^{\infty}$ of G satisfies*

$$\sum_{n=M+1}^{\infty} \|g_n - f_n\|^2 < \infty\,, \tag{8.11}$$

for some $M \in \mathbb{N}$. Then one can find an integer $M' \geq M$ and some generalized eigenvectors $\langle g_{n0}\rangle_{n=1}^{M'}$ of G such that

$$\{g_{n0}\}_{n=1}^{M'} \cup \{g_n\}_{n=M'+1}^{\infty}$$

forms a Riesz basis of H.

It is easily verified that the vectors

$$\begin{pmatrix} 0 \\ \nu \\ \nu/\beta \end{pmatrix} e_j, \quad \begin{pmatrix} 1/\omega_j^2 \\ -ic \\ 0 \end{pmatrix} e_j, \quad \begin{pmatrix} 1/\omega_j^2 \\ ic \\ 0 \end{pmatrix} e_j\,,$$

for $j = 1, 2, \ldots$ constitute a Riesz basis for $\widetilde{H}$, and that an appropriate subset is quadratically close to $\{V_j^p e_j\}_{|\lambda_j^p|>R} \cup \{V_j^{h,\pm} e_j\}_{|\lambda_j^{h,\pm}|>R}$ in the sense of (8.11). This means that the above theorem can be used, so that the parabolic eigenvectors (8.9) and the hyperbolic eigenvectors (8.10) of the adjoint operator $\widetilde{\mathcal{A}}$, supplemented with generalized eigenvectors of $\widetilde{\mathcal{A}}$, constitute a Riesz basis for $\widetilde{H}$.

8.3 Proving Null-controllability

As mentioned in the beginning of this chapter, our goal is to find a control $k \in L^2(0, T)$ such that the solution, corresponding to any initial state at time $t = 0$, is driven to zero at time $t = T$. To prove that this is in fact possible we will split the time line, as shown in Figure 8.3, into several parts where $0 < T_1 < T_2 < T_3 < T$. This will allow us to split the problem into smaller, easier, parts, that solve the main problem when put together.

Let us introduce notation for solutions of the control system (8.1). For example, writing

$$\mathcal{S}(T_a, U_a \mid k)(T_b)\,,$$

means the following: The initial condition is given at time T_a as U_a and the control used is k. The value of T_b dictates the time at which we wish to know the value of the solution. The control is assumed always to have support in $(0,T)$. To that end we introduce as simple "underline" notation, that extends a function with zeroes to the entire $(0,T)$ interval. For instance, if $k^h \in L^2(T_1,T_2)$ then $\underline{k}^h(t)$ is equal to $k^h(t)$ for $T_1 < t < T_2$ and zero otherwise.

We begin by proving that we can always determine a control $k^h \in L^2(T_1,T_2)$ that makes sure that the solution's projection onto the hyperbolic eigenvectors, at time $t = T$, is zero.

Theorem 8.3.1. *If $T_2 - T_1 \geq 2/c$ and R big enough, then a bounded, linear operator $K^h : H \times L^2(T_3,T) \mapsto L^2(T_1,T_2)$ exists such that*

$$\Pi^h \mathcal{S}\big(T_1, U_1 \mid \underline{K}^h(U_1,k^p) + \underline{k}^p\big)(T) = 0\,,$$

for all $U_1 \in H$ and $k^p \in L^2(T_3,T)$.

Proof. Let us first introduce the semigroup $e^{\mathcal{A}^h t} : H^h \mapsto H^h$ as the restriction of $e^{\mathcal{A}t}$ to H^h. Note that since the magnitudes of the real part of the hyperbolic eigenvalues are bounded, $e^{\mathcal{A}^h t}$ is well-defined for all $t \in \mathbb{R}$.

Let now $k^h \in L^2(T_1,T_2)$, $k^p \in L^2(T_3,T)$, $U_1 \in H$ and consider the rewrite

$$\begin{aligned}
0 &= \Pi^h \mathcal{S}\big(T_1, U_1 \mid \underline{k}^h + \underline{k}^p\big)(T) \\
&= \Pi^h \mathcal{S}\big(T_1, 0 \mid \underline{k}^h\big)(T) + \Pi^h \mathcal{S}\big(T_1, U_1 \mid \underline{k}^p\big)(T) \\
&= e^{\mathcal{A}^h(T-T_2)} \Pi^h \mathcal{S}\big(T_1, 0 \mid \underline{k}^h\big)(T_2) + \Pi^h \mathcal{S}\big(T_1, U_1 \mid \underline{k}^p\big)(T) \quad \Leftrightarrow
\end{aligned}$$

$$\Pi^h \mathcal{S}\big(T_1, 0 \mid \underline{k}^h\big)(T_2) = -e^{-\mathcal{A}^h(T-T_2)} \Pi^h \mathcal{S}\big(T_1, U_1 \mid \underline{k}^p\big)(T) = U^h\,.$$

Since $U^h \in H^h$ for every choice of $k^p \in L^2(T_3,T)$ and $U_1 \in H$, we see that the claim of the theorem is proved if we can show the following: A control $k^h \in L^2(T_1,T_2)$ exists that drives the zero state at $t = T_1$ to a state U at $t = T_2$ for which $\Pi^h U = U^h$ for *any* $U^h \in H^h$.

This *exact controllability problem* is known (see Theorem 2.6.3, page 27) to be equivalent to the following observability inequality: A constant $C_h > 0$ must exist such that

$$\big\|V^0\big\|^2_{\widetilde{H}} \leq C_h \int_{T_1}^{T_2} |v_x(t,1)|^2 dt\,, \tag{8.12}$$

for all solutions (v,ψ) of the adjoint system with initial condition $V^0 \in \widetilde{H}^h$ (at time $t = T_2$).

We first consider the left-hand side of the observability inequality (8.12). Since $V^0 \in \widetilde{H}^h$ we can write it as

$$V^0 = \sum_{|\lambda_j^{h,\pm}|>R} \left(a_j^+ V_j^{h,+} + a_j^- V_j^{h,-}\right) e_j\,.$$

Because of the Riesz basis property of the vectors $V_j^{h,\pm} e_j$, we immediately have

$$\left\|V^0\right\|_{\widetilde{H}}^2 \leq C \sum_{|\lambda_j^{h,\pm}|>R} \left(|a_j^+|^2 + |a_j^-|^2\right).$$

We now turn to the right hand-side of (8.12). The first component of the full solution, v, has the appearance

$$v(t,x) = \sum_{|\lambda_j^{h,\pm}|>R} \left(a_j^+/\omega_j e^{\lambda_j^{h,+}(T_2-t)} + a_j^-/\omega_j e^{\lambda_j^{h,-}(T_2-t)}\right) \sin(j\pi x),$$

and thus

$$\begin{aligned}
&\int_{T_1}^{T_2} |v_x(t,1)|^2 dt \\
&= \int_0^{T_2-T_1} \left| \sum_{|\lambda_j^{h,\pm}|>R} (-1)^j e^{-\frac{\alpha\beta}{2\nu}t} j\pi \left(\frac{a_j^+}{\omega_j} e^{(ic\pi j+\epsilon_j)t} + \frac{a_j^-}{\omega_j} e^{(-ic\pi j+\overline{\epsilon_j})t} \right) \right|^2 dt \\
&\geq e^{-\frac{\alpha\beta}{\nu}(T_2-T_1)} \sum_{|\lambda_j^{h,\pm}|>R} \left(\left|(-1)^j \frac{j\pi}{\omega_j} a_j^+\right|^2 + \left|(-1)^j \frac{j\pi}{\omega_j} a_j^-\right|^2 \right) \\
&= C' \sum_{|\lambda_j^{h,\pm}|>R} \left(|a_j^+|^2 + |a_j^-|^2\right)
\end{aligned}$$

for an appropriate constant C'. Here, Kadec' Theorem (Theorem 4.2.2, page 76) has been used under the assumption that $T_2 - T_1 \geq 2/c$ and that R is big enough such that each $\epsilon_j = O(1/j)$ is appropriate small.

The observability inequality (8.12), and thereby the exact controllability problem, has now been proved. □

Notice how the exact controllability condition $T_2 - T_1 \geq 2/c$ is the same as it would be for the "decoupled" wave equation $u_{tt} = c^2 u_{xx}$ of the control system (8.1).

We now move on to proving a theorem similar to the previous one. We show that a control $k^p \in L^2(T_3, T)$ exists that makes sure that the solution's projection onto the parabolic eigenvectors, at time $t = T$, is zero.

Theorem 8.3.2. *A bounded, linear operator $K^p : H \times L^2(T_1,T_2) \mapsto L^2(T_3,T)$ exists such that*

$$\Pi^p \mathcal{S}\big(T_1, U_1 \mid \underline{k}^h + \underline{K}^p(U_1,k^h)\big)(T) = 0,$$

for all $U_1 \in H$ and $k^h \in L^2(T_1,T_2)$.

Proof. Let $k^h \in L^2(T_1,T_2)$, $k^p \in L^2(T_3,T)$ and $U_1 \in H$. Observe that

$$\Pi^p \mathcal{S}\big(T_1, U_1 \mid \underline{k}^h + \underline{k}^p\big)(T) = \Pi^p \mathcal{S}\big(T_3, U_3 \mid \underline{k}^p\big)(T) = 0,$$

where
$$U_3 = \mathcal{S}\big(T_1, U_1 \mid \underline{k}^h\big)(T_3)\,.$$
This means that we must find a control $k^p \in L^2(T_3, T)$ that steers the state U_3 at time T_3 to zero at time T. This is a *null controllability problem* and the existence of such a control can be proved by showing the following observability inequality: A constant $C_p > 0$ must exist such that

$$\big\|e^{\widetilde{\mathcal{A}}(T-T_3)}V^0\big\|_{\widetilde{H}}^2 \le C_p \int_{T_3}^T |v_x(t,1)|^2 dt\,, \tag{8.13}$$

for all solutions (v, ψ) of the adjoint system with initial condition $V^0 \in \widetilde{H}^p$ (at time $t = T$), see Theorem 2.6.2, page 27.

Let now
$$V^0 = \sum_{|\lambda_j^p|>R} b_j V_j^p e_j$$

where $\langle b_j\rangle \in \ell^2$, for which the full solution becomes

$$V(t) = \sum_{|\lambda_j^p|>R} e^{\lambda_j^p(T-t)} b_j V_j^p e_j.$$

For the left-hand side of (8.13), we immediately get from the Riesz basis property,

$$C' \sum_{|\lambda_j^p|>R} e^{2\lambda_j^p(T-T_3)}|b_j|^2 \ge \big\|e^{\widetilde{\mathcal{A}}(T-T_3)}V^0\big\|_{\widetilde{H}}^2\,, \tag{8.14}$$

for an appropriate constant C'. The right hand-side of (8.13) can be bounded from below by applying Theorem 4.2.3, the parabolic version of Ingham's Theorem:

$$\begin{aligned}
\int_{T_3}^T \big|v_x(t,1)\big|^2 dt &= \int_{T_3}^T \Bigg| \sum_{|\lambda_j^p|>R} \frac{(-1)^j \pi j}{\omega_j^2} e^{\lambda_j^p(T-t)} b_j \Bigg|^2 dt \\
&\ge \int_0^{(T-T_3)/2} \Bigg| \sum_{|\lambda_j^p|>R} e^{\lambda_j^p s} \frac{(-1)^j}{\pi j} b_j \Bigg|^2 ds \\
&\ge C'' \sum_{|\lambda_j^p|>R} \frac{e^{\lambda_j^p(T-T_3)}}{-\lambda_j^p} \left|\frac{(-1)^j}{\pi j} b_j\right|^2 \\
&= C'' \sum_{|\lambda_j^p|>R} \frac{e^{-\lambda_j^p(T-T_3)}}{-\pi^2 j^2 \lambda_j^p} e^{2\lambda_j^p(T-T_3)} |b_j|^2\,.
\end{aligned} \tag{8.15}$$

Since obviously $e^{-\lambda_j^p(T-T_3)}/(-j^2\lambda_j^p) \to \infty$ as $j \to \infty$, the inequalities (8.14) and (8.15) can be combined into the observability inequality (8.13) for an appropriate constant $C_p > 0$. $\square$

Let us review the steps needed for actually computing the control k^p of the above theorem. Let a state $U_1 \in H'$ at time $t = T_1$ be given. Perform now the following steps.

1. Set $U_2 = \mathcal{S}\big(T_1, U_1 \mid k^h\big)(T_2)$.

2. Set $U_3 = e^{\mathcal{A}(T_3-T_2)}U_2$.

3. Minimize the functional

$$J_0(V^0) = \tfrac{1}{2}c^4 \int_{T_3}^{T} |v_x(t,1)|^2 dt + \{U_3, e^{\widetilde{\mathcal{A}}(T-T_3)}V^0\}\,, \tag{8.16}$$

   over $\widetilde{H}^p$ and let the minimizer be denoted V^* (compare to the functional (2.18) on page 19).

4. The wanted control is $k^p(t) = -c^2 v_x(t,1)$, where $v(t,x)$ is the first component of the solution to the adjoint system with initial condition V^*.

Important to note is that in step 3 only the H^p-part of U_3 matters since $V^0 \in \widetilde{H}^p$. This means that we can replace step 2 by

2'. Set $U_3 = e^{\mathcal{A}^p(T_3-T_2)}\mathit{\Pi}^p U_2$,

where $e^{\mathcal{A}^p t} : H^p \mapsto H^p$ is the restriction of $e^{\mathcal{A}t}$ to H^p, and the computed control will be exactly the same. The mapping in step 2' is *compact* because of the strong damping of the operator $e^{\mathcal{A}^p(T_3-T_2)}$. A more rigorous argument for this compactness can be made by first determining $\langle b_j \rangle \in \ell^2$ such that

$$\sum_{|\lambda_j^p|>R} b_j U_j^p e_j = \mathit{\Pi}^p U_2\,,$$

where $U_j^p e_j$ is the parabolic eigenvector associated with λ_j^p. We now imbed $\langle b_j \rangle \in \ell^2$ into $\ell^{2,p}$ for some $p > 0$, where $\|\langle b_j \rangle\|_{\ell^{2,p}} = \|\langle b_j/j^p \rangle\|_{\ell^2}$. This is a *compact embedding*, see Detail 9 on page 193. Setting then

$$U_3 = \sum_{|\lambda_j^p|>R} e^{\lambda_j^p(T_3-T_2)} b_j U_j^p e_j\,,$$

for $\langle b_j \rangle \in \ell^{2,p}$ will still make $U_3 \in H^p$, because of the exponential damping of the eigenvalue coefficients.

Since the composition of the steps above, where one is compact, is the map $K^p(0,\cdot)$, we see that it is itself compact. This property is important in the proof of the following theorem. The result concludes, with some restrictions, that *both* the parabolic and hyperbolic part of a solution can driven to zero at $t = T$.

Theorem 8.3.3. *Under the assumptions of Theorems 8.3.1 and 8.3.2, a subspace $\mathcal{V} \subset H'$ of finite codimension and a bounded, linear operator*

$$K_{\mathcal{V}} : \mathcal{V} \mapsto L^2(T_1, T_2) \times L^2(T_3, T) \,,$$

exists such that

$$(\Pi^p + \Pi^h)\mathcal{S}(T_1, U_1 \mid \underline{k}^h + \underline{k}^p)(T) = 0 \quad \textit{with } (k^h, k^p) = K_{\mathcal{V}}(U_1) \,,$$

for all $U_1 \in \mathcal{V}$.

Proof. Observe that the statement

$$(\Pi^p + \Pi^h)\mathcal{S}(T_1, U_1 \mid \underline{k}^h + \underline{k}^p)(T) = 0 \,,$$

is equivalent to (using the maps of the previous two theorems)

$$\begin{aligned} k^h &= K^h(U_1, k^p) = A_1(U_1) + A_2(k^p) \quad \text{and} \\ k^p &= K^p(U_1, k^h) = B_1(U_1) + B_2(k^h) \,, \end{aligned} \tag{8.17}$$

where A_1, A_2, B_1 and B_2 are trivially defined because of the linearity of K^h and K^p. Combining the above equations yields

$$\begin{aligned} k^p = B_1 U_1 + B_2(A_1 U_1 + A_2 k^p) \quad &\Leftrightarrow \\ C U_1 = (I - B_2 A_2) k^p, & \end{aligned} \tag{8.18}$$

with $C = B_1 + B_2 A_1$. Solving this equation is thus equivalent to solving (8.17).

Since $B_2 = K^p(0, \cdot)$ is compact, we have by Fredholm's alternative: There are a *finite* number of continuous maps $l_1, l_2, \ldots, l_L \in (L^2(T_3, T))'$ such that Equation (8.18) has a solution $k^p \in L^2(T_3, T)$ if and only if $U_1 \in \mathcal{V}$ where

$$\mathcal{V} = \{v \in H' \mid l_j(C(v)) = 0 \text{ for } j = 1, \ldots, L\} \,.$$

□

We must now show that a control $k^0 \in L^2(0, T_1)$ can be found, such that any state $U_0 \in H'$ at time $t = 0$ can be driven to a state at time $t = T_1$ that lies in the subspace $\mathcal{V}$.

Before doing that, however, we must argue that a basis for $\mathcal{V}^\perp$ can be chosen among finitely many eigenvectors of $\widetilde{\mathcal{A}}$. Let us first introduce the notation $\mathcal{V}_T$ to emphasize this set's dependence on the control time T, and we have

$$\mathcal{V}_T^\perp = \left\{ V \in \widetilde{H} \mid \{U, V\} = 0 \text{ for all } U \in \mathcal{V}_T \right\} .$$

Observe now that the (finite) dimension of $\mathcal{V}_T^\perp$ is non-increasing. This is because if a control exists for a given time T, in the sense of Theorem 8.3.3, then a control also exists for $T+\epsilon$, $\epsilon>0$, since the same control can be used in the interval $(0,T)$ and then simply use zero control in $(T,T+\epsilon)$. We can then assume that

$$\mathcal{V}_{T'}^{\perp}=\mathcal{V}_T^{\perp}\,,\quad T'\in(T,T+\epsilon) \tag{8.19}$$

for some $\epsilon>0$ (it may be necessary to adjust T slightly downwards, but this is always possible).

We shall now show that $\mathcal{V}_T^{\perp}=e^{\widetilde{\mathcal{A}}t}\mathcal{V}_T^{\perp}$ for all $t>0$. This follows if we can show $\mathcal{V}_T^{\perp}=e^{\widetilde{\mathcal{A}}\delta}\mathcal{V}_T^{\perp}$ for some $0<\delta<\epsilon$. Assume therefore $U\in(e^{\widetilde{\mathcal{A}}\delta}\mathcal{V}_T^{\perp})^{\perp}$. We get

$$\begin{aligned}\{U,e^{\widetilde{\mathcal{A}}\delta}V\}=\{e^{\mathcal{A}\delta}U,V\}=0\quad\text{for all } V\in\mathcal{V}_T^{\perp}\qquad&\Rightarrow\\ e^{\mathcal{A}\delta}U\in\mathcal{V}_T\quad\Rightarrow\quad U\in\mathcal{V}_{T+\delta}\quad\Rightarrow\quad U\in\mathcal{V}_T\,,&\end{aligned}$$

using (8.19). This implies

$$(e^{\widetilde{\mathcal{A}}\delta}\mathcal{V}_T^{\perp})^{\perp}\subset\mathcal{V}_T\quad\Rightarrow\quad\mathcal{V}_T^{\perp}\subset e^{\widetilde{\mathcal{A}}\delta}\mathcal{V}_T^{\perp}\,.$$

But since $e^{\widetilde{\mathcal{A}}\delta}$ is a linear operator we also know that $\dim(e^{\widetilde{\mathcal{A}}\delta}\mathcal{V}_T^{\perp})\le\dim(\mathcal{V}_T^{\perp})$. We thus have $\mathcal{V}_T^{\perp}=e^{\widetilde{\mathcal{A}}\delta}\mathcal{V}_T^{\perp}$.

Let now $Z(t)$ be the restriction of the semigroup $e^{\widetilde{\mathcal{A}}t}$ to $\mathcal{V}^{\perp}$. We set $Z(t)=e^{Bt}$, where $B:\mathcal{V}^{\perp}\mapsto\mathcal{V}^{\perp}$ is the finite dimensional linear operator such that $\widetilde{\mathcal{A}}V=BV$ for all $V\in\mathcal{V}^{\perp}$. Assume now that $V\in\mathcal{V}^{\perp}$ is a generalized eigenvector of B, that is, we have $(B-\lambda I)^mV=0$ for some integer m. We then get

$$0=(B-\lambda I)^m=\sum_{i=0}^{m}(-\lambda)^iB^{m-i}V=\sum_{i=0}^{m}(-\lambda)^i\widetilde{\mathcal{A}}^{m-i}V=(\widetilde{\mathcal{A}}-\lambda I)^m\,,$$

showing that V is also a generalized eigenvector of $\widetilde{\mathcal{A}}$. This finally means that a finite number of generalized eigenvectors of $\widetilde{\mathcal{A}}$ span $\mathcal{V}^{\perp}$,

$$\mathcal{V}^{\perp}=\widetilde{H}_M^0=\left\{\sum_{j=1}^{M}\big(a_j^+V_j^{h,+}+a_j^-V_j^{h,-}+b_jV_j^p\big)e_j\;\middle|\;a_j^+,a_j^-,b_j\in\mathbb{R},\ j=1,\dots,M\right\},$$

for some $M\in\mathbb{N}$.

We can now return to the problem of whether a control $k^0\in L^2(0,T_1)$ can be found, such that any state $U_0\in H'$ at time $t=0$ can be driven to a state at time $t=T_1$ that lies in the subspace $\mathcal{V}$. This is an exact control problem for a projection, seen from Theorem 2.6.3 to be equivalent to the observability inequality,

$$\|V^0\|_{\widetilde{H}}^2\le C_0\int_0^{T_1}|v_x(t,1)|^2dt\,,$$

for some constant $C_0>0$ and for all $V^0\in\widetilde{H}_M^0$. But since the vector space $\widetilde{H}_M^0$ is of *finite dimension*, this inequality is proved for any $T_1>0$ by the following theorem. (The theorem is basically the same as Corollary 4.1.1, page 67, but the following version is written using the notation of the present chapter).

Theorem 8.3.4. *Let $M \in \mathbb{N}$ be given. If among $\lambda_1^{h,+}$, $\lambda_1^{h,-}$, λ_1^p, $\lambda_2^{h,+}$, ..., λ_M^p there are no multiple eigenvalues, then*

$$\int_0^T |v_x(t,1)|^2 dt = 0 \quad \Rightarrow \quad V^0 = 0\,,$$

for all $V^0 \in \widetilde{H}_M^0$, where $v(t,x)$ is the first component of the solution to the adjoint system with initial condition V^0.

Proof. Let

$$V^0 = \sum_{j=1}^{M} \big(a_j^+ V_j^{h,+} + a_j^- V_j^{h,-} + b_j V_j^p\big) e_j\,,$$

so the full solution becomes

$$V(t) = \sum_{j=1}^{M} \big(a_j^+ e^{\lambda_j^{h,+}(T-t)} V_j^{h,+} + a_j^- e^{\lambda_j^{h,-}(T-t)} V_j^{h,-} + b_j e^{\lambda_j^p(T-t)} V_j^p\big) e_j\,.$$

Assume now that the eigenvectors are normalized such that (scaling does not matter because of the finite dimension)

$$\begin{aligned} v_x(t,1) &= \sum_{j=1}^{M} \big(a_j^+ e^{\lambda_j^{h,+}(T-t)} + a_j^- e^{\lambda_j^{h,-}(T-t)} + b_j e^{\lambda_j^p(T-t)}\big) \\ &= \boldsymbol{b}^T e^{\boldsymbol{L}(T-t)} \boldsymbol{z}\,, \end{aligned}$$

where $\boldsymbol{b}^T = [1,1,\dots,1] \in \mathbb{R}^{3M}$ and

$$\begin{aligned} \boldsymbol{L} &= \operatorname{diag}(l_1, l_2, l_3, l_4, ..., l_{3M}) = \operatorname{diag}(\lambda_1^{h,+}, \lambda_1^{h,-}, \lambda_1^p, \lambda_2^{h,+}, ..., \lambda_M^p)\,, \\ \boldsymbol{z}^T &= [z_1, z_2, z_3, z_4, \dots, z_{3M}] = [a_1^+, a_1^-, b_1, a_2^+, \dots, b_M]\,. \end{aligned}$$

Now observe that

$$\int_0^T |v_x(t,1)|^2 dt = 0 \quad \Leftrightarrow \quad \boldsymbol{z}^T e^{\boldsymbol{L}t} \boldsymbol{b} = 0 \quad \text{for } 0 \le t \le T\,.$$

This last expression implies, as seen by repeated differentiation, that

$$\boldsymbol{z}^T \boldsymbol{L}^k \boldsymbol{b} = 0, \quad \text{for } k = 0, 1, \dots, 3M-1\,,$$

equivalent to

$$\boldsymbol{z}^T \begin{bmatrix} 1 & l_1 & \cdots & l_1^{3M-1} \\ 1 & l_2 & \cdots & l_2^{3M-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & l_{3M} & \cdots & l_{3M}^{3M-1} \end{bmatrix} = \boldsymbol{0}^T\,.$$

Since the eigenvalues are all distinct, this Vandermonde matrix is regular and the equation is thus only satisfied for $\boldsymbol{z} = 0$, implying $V^0 = 0$. $\square$

Recall that multiple eigenvalues *can* occur, see Figure 8.2, so null-controllability will fail in these cases.

Combining the existence of a control $k^0 \in L^2(0, T_1)$ (that steers any state $U_0 \in H'$ at $t = 0$ to a state at time $t = T_1$ in the subspace $\mathcal{V}$) and Theorem 8.3.3, we have now shown that we can drive any state $U_0 \in H'$ at $t = 0$ to a state $U \in H'$ at $t = T$ for which $(\Pi^p + \Pi^h)U = 0$.

Since the interval $(0, T_1)$ can be arbitrarily short, it is possible to "squeeze in" another control, prior to k^0. This control, given k^0, k^h and k^p can be computed such that $\Pi^0 U(T) = 0$, where U is the solution to the control system. Since this is a finite dimensional control problem we have (a) it can be done under the conditions of Theorem 8.3.4, (b) it can be done arbitrarily fast and (c) the map computing the control is of finite dimension and thus *compact*. This means that the arguments of Theorem 8.3.3 can be used again to conclude: A subspace $\mathcal{V}' \subset H'$ can be found such that any state $U_0 \in \mathcal{V}'$ at $t = 0$ can be driven to zero at $t = T$. As previously done for $\mathcal{V}$, we can steer any $U_0 \in H'$ into the space $\mathcal{V}'$. We can thus finally formulate the main theorem of this chapter.

Theorem 8.3.5. *Assume that the operator $\mathcal{A}$ has no multiple eigenvalues and that $T > 2/c$. Then a bounded, linear operator*

$$K : H' \mapsto L^2(0, T)$$

exists such that

$$\mathcal{S}\big(0, U^0 \mid K(U^0)\big)(T) = 0\,,$$

for all $U^0 \in H'$.

Recall that the control system (8.1), which we have considered in this chapter, could be considered a coupling of a wave equation and a heat equation. The heat equation-part made only null-controllability possible, the wave equation-part introduced a limit to how fast the control could be done, and the single-pointed boundary control required no multiple eigenvalues. The coupling demanded all three.

Our result is similar to that obtained in *Lebeau and Zuazua (1998)* regarding the same system, but with internal control. There is one interesting difference though, in that multiple eigenvalues did not prohibit controllability in their case.

C h a p t e r 9

Implementing HUM

Now comes the nitty-gritty.
— Donald E. Knuth

The Hilbert Uniqueness Method provides more than results concerning existence and uniqueness of controls—it also tells us how to *construct* the controls.

This chapter deals with an implementation of HUM for the wave equation in two dimensions. When it comes to such implementions of HUM in practice with focus on, for instance, efficiency, accuracy and memory usage, only a few aspects have been addressed in the literature. For example, *Glowinski, Li, and Lions (1990)* and *Asch and Lebeau (1998)* contain some comments.

It should be noted that although we consider a specific discretization of a specific equation, it would not be difficult to use the same procedure in other cases as well.

9.1 The Discretization

This section will describe the discretization of the 2D wave equation as mentioned in Section 3.4.1. Recall that it (locally) discretizes according to the scheme,

$$\frac{1}{16}\left[\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}^{n+1} - 2\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}^{n} + \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}^{n-1}\right]$$

$$= \frac{\Delta t^2}{8h^2}\left[\begin{pmatrix} 1 & & 1 \\ & -4 & \\ 1 & & 1 \end{pmatrix}^{n+1} + 2\begin{pmatrix} 1 & & 1 \\ & -4 & \\ 1 & & 1 \end{pmatrix}^{n} + \begin{pmatrix} 1 & & 1 \\ & -4 & \\ 1 & & 1 \end{pmatrix}^{n-1}\right].$$

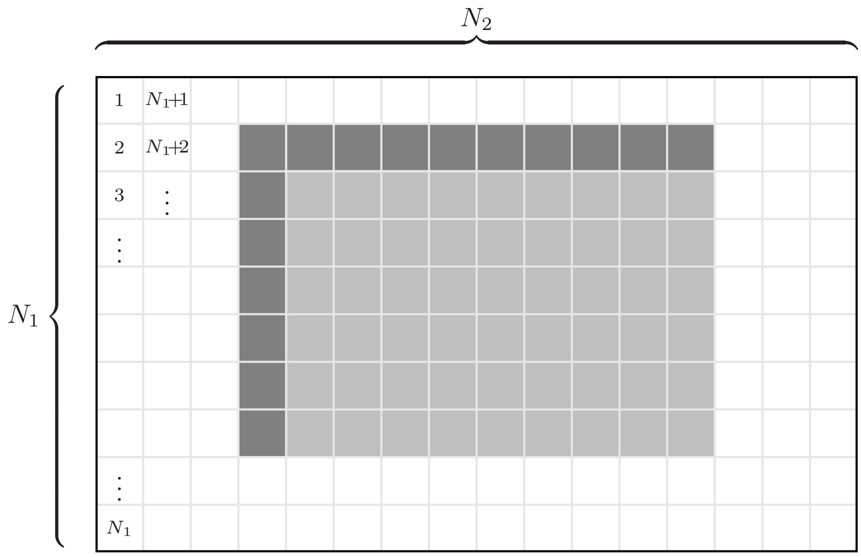


Figure 9.1: Setting up the geometry inside an $N_1 \times N_2$ rectangular matrix. Each element/point is assigned a sequential number in the range $1, 2, \ldots, N_1 \cdot N_2$, as indicated in the left part of the grid. The inner points are shown in light gray and the dark gray points represent the control part of the boundary. The rest of the points represent the non-controlled part of the boundary, kept at zero.

As also mentioned in Section 3.4.1, the time discretization corresponds to the trapezoid rule and the adjoint system can thus be written,

$$\begin{bmatrix} \boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix} \left(\begin{bmatrix} \boldsymbol{v}^{n+1} \\ \overline{\boldsymbol{v}}^{n+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{v}^{n} \\ \overline{\boldsymbol{v}}^{n} \end{bmatrix} \right) = \frac{\Delta t}{2} \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{A} & \boldsymbol{0} \end{bmatrix} \left(\begin{bmatrix} \boldsymbol{v}^{n+1} \\ \overline{\boldsymbol{v}}^{n+1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{v}^{n} \\ \overline{\boldsymbol{v}}^{n} \end{bmatrix} \right), \tag{9.1}$$

where $\overline{\boldsymbol{v}}^n$ approximates the time derivative.

We now have to set up the matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ according to the geometry. The geometry will be described using the following quantities:

- $N_1 \times N_2$: Dimension of the two-dimensional array, in which the geometry is set.
- N_0: Number of inner points.
- N_b: Number of boundary control points.
- $\boldsymbol{I}_0 \in \mathbb{N}^{N_0}$: Vector of indices of inner points.
- $\boldsymbol{I}_b \in \mathbb{N}^{N_b}$: Vector of indices of boundary control points.

These quantities are illustrated in Figure 9.1. Of course, the grid size h, time step Δt and number of time steps M need also be set. Note that the geometry need not be a rectangular shape.

Let us now introduce the matrix $\boldsymbol{A}_G \in \mathbb{R}^{N_1 \cdot N_2 \times N_1 \cdot N_2}$ as a representative for the computational stencil,

$$\frac{1}{2h^2}\begin{pmatrix} 1 & & 1 \\ & -4 & \\ 1 & & 1 \end{pmatrix}.$$

More formally, this means that

$$\boldsymbol{y} = \boldsymbol{A}_G\boldsymbol{x} \;\Leftrightarrow\; \boldsymbol{Y}_{i,j} = \frac{\boldsymbol{X}_{i-1,j-1} + \boldsymbol{X}_{i-1,j+1} + \boldsymbol{X}_{i+1,j-1} + \boldsymbol{X}_{i+1,j+1} - 4\boldsymbol{X}_{i,j}}{2h^2},$$

(references outside the $\boldsymbol{X}$-matrix should be set to zero) must hold for every instance of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{N_1 \cdot N_2}$ and $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{N_1 \times N_2}$ for which

$$\boldsymbol{X}_{i,j} = \boldsymbol{x}_{N_1(j-1)+i} \quad \text{and} \quad \boldsymbol{Y}_{i,j} = \boldsymbol{y}_{N_1(j-1)+i}, \quad i = 1, \ldots, N_1,\; j = 1, \ldots, N_2\,,$$

so $\boldsymbol{x}$ and $\boldsymbol{y}$ are column-stacked versions of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

We can now extract the relevant entries into the actual system matrix,

$$\boldsymbol{A}(i,j) = \boldsymbol{A}_G(\boldsymbol{I}_0(i), \boldsymbol{I}_0(j)), \quad \text{for all } i, j = 1, 2, \ldots, N_0,$$

or `A=AG(I0,I0)` in MATLAB notation.

Similarly, let $\boldsymbol{C}_G$ represent the stencil

$$\frac{1}{2h^2}\begin{pmatrix} 1 & & 1 \\ & -4 & \\ 1 & & 1 \end{pmatrix},$$

and we extract the relevant entries to create $\boldsymbol{C} \in \mathbb{R}^{N_0 \times N_0}$.

The way that these matrices have been created means that the entries of vectors $\boldsymbol{v}^n$ and $\overline{\boldsymbol{v}}^n$ in the scheme (9.1) represent the inner points only. The value of $\boldsymbol{v}^n(i)$, respectively $\overline{\boldsymbol{v}}^n(i)$, corresponds to the position, respectively velocity, of the grid point with index $\boldsymbol{I}_0(i)$.

Before setting up the control system, we need to incorporate the boundary conditions. But this is easily done using the already created matrix $\boldsymbol{A}_G$. We let the matrix $\boldsymbol{B} \in \mathbb{R}^{N_0 \times N_b}$ have the entries,

$$\boldsymbol{B}(i,j) = \boldsymbol{A}_G(\boldsymbol{I}_0(i), \boldsymbol{I}_b(j)), \quad \text{for } i = 1, 2, \ldots, N_0 \text{ and } j = 1, 2, \ldots, N_b\,.$$

We can now formulate the discrete control system,

$$\begin{aligned} &\begin{bmatrix} \boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix} \left(\begin{bmatrix} \boldsymbol{u}^{n+1} \\ \overline{\boldsymbol{u}}^{n+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{u}^{n} \\ \overline{\boldsymbol{u}}^{n} \end{bmatrix} \right) \\ &\qquad = \frac{\Delta t}{2} \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{A} & \boldsymbol{0} \end{bmatrix} \left(\begin{bmatrix} \boldsymbol{u}^{n+1} \\ \overline{\boldsymbol{u}}^{n+1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}^{n} \\ \overline{\boldsymbol{u}}^{n} \end{bmatrix} \right) + \frac{\Delta t}{2} \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{B} \end{bmatrix} \left(\boldsymbol{k}^{n+1} + \boldsymbol{k}^{n} \right), \end{aligned} \tag{9.2}$$

where $\boldsymbol{k}^n \in \mathbb{R}^{N_b}$ supplies the boundary conditions at time $n\Delta t$.

9.2 Computing the Controllability Operator

As explained in Section 5.1, there are two obvious ways to compute the controllability operator, the *direct method* and the *inner product method*. Both methods will be considered in the following two sections. Note that we implement the time reversed version of HUM, as described in Section 2.5.2, which is possible since we deal with the reversible wave equation. This means that we wish to drive a certain state $(\boldsymbol{y}^0, \overline{\boldsymbol{y}}^0)$ to the null state $(\mathbf{0}, \mathbf{0})$.

9.2.1 The Direct Method

Let the discrete (reversed) controllability operator be denoted $\widetilde{\boldsymbol{\Lambda}}_M$, the subscript M representing the number of time steps used. Mapping a given vector $(\boldsymbol{v}^0, \overline{\boldsymbol{v}}^0)$ is done by computing

$$\begin{aligned} \begin{bmatrix} \boldsymbol{v}^{n+1} \\ \overline{\boldsymbol{v}}^{n+1} \end{bmatrix} &= \boldsymbol{G}_1 \begin{bmatrix} \boldsymbol{v}^{n} \\ \overline{\boldsymbol{v}}^{n} \end{bmatrix}, \qquad \begin{bmatrix} \boldsymbol{v}^{0} \\ \overline{\boldsymbol{v}}^{0} \end{bmatrix} \text{ given,} \\ \boldsymbol{G}_1 &= \left(\begin{bmatrix} \boldsymbol{C} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C} \end{bmatrix} - \frac{\Delta t}{2} \begin{bmatrix} \mathbf{0} & \boldsymbol{I} \\ \boldsymbol{A} & \mathbf{0} \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} \boldsymbol{C} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C} \end{bmatrix} + \frac{\Delta t}{2} \begin{bmatrix} \mathbf{0} & \boldsymbol{I} \\ \boldsymbol{A} & \mathbf{0} \end{bmatrix} \right), \end{aligned} \tag{9.3}$$

for $n = 0, 1, \ldots, M-1$. We now set

$$\boldsymbol{k}^n = -\boldsymbol{B}^T \boldsymbol{v}^n, \tag{9.4}$$

for $n = 0, 1, \ldots, M$ (the minus sign appears because we deal with the reversed controllability operator). We then solve the control system,

$$\begin{aligned} \begin{bmatrix} \boldsymbol{u}^{n} \\ \overline{\boldsymbol{u}}^{n} \end{bmatrix} &= \boldsymbol{G}_2 \begin{bmatrix} \boldsymbol{u}^{n+1} \\ \overline{\boldsymbol{u}}^{n+1} \end{bmatrix} + \boldsymbol{F}(\boldsymbol{k}^{n+1} + \boldsymbol{k}^n), \qquad \begin{bmatrix} \boldsymbol{u}^{M} \\ \overline{\boldsymbol{u}}^{M} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \\ \boldsymbol{G}_2 &= \left(\begin{bmatrix} \boldsymbol{C} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C} \end{bmatrix} + \frac{\Delta t}{2} \begin{bmatrix} \mathbf{0} & \boldsymbol{I} \\ \boldsymbol{A} & \mathbf{0} \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} \boldsymbol{C} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C} \end{bmatrix} - \frac{\Delta t}{2} \begin{bmatrix} \mathbf{0} & \boldsymbol{I} \\ \boldsymbol{A} & \mathbf{0} \end{bmatrix} \right), \\ \boldsymbol{F} &= -\frac{\Delta t}{2} \left(\begin{bmatrix} \boldsymbol{C} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C} \end{bmatrix} + \frac{\Delta t}{2} \begin{bmatrix} \mathbf{0} & \boldsymbol{I} \\ \boldsymbol{A} & \mathbf{0} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{0} \\ \boldsymbol{B} \end{bmatrix}, \end{aligned}$$

for $n = M-1, M-2, \ldots, 0$. Finally we have

$$\widetilde{\boldsymbol{\Lambda}}_M \begin{bmatrix} \boldsymbol{v}^{0} \\ \overline{\boldsymbol{v}}^{0} \end{bmatrix} = \begin{bmatrix} \overline{\boldsymbol{u}}^{0} \\ -\boldsymbol{u}^{0} \end{bmatrix}.$$

Now, by repeatedly applying this map to the columns of a $2N_0 \times 2N_0$ identity matrix, we obtain the columns of the matrix $\widetilde{\boldsymbol{\Lambda}}_M$.

Note that instead of mapping just one column at a time, one could easily map a whole matrix. This way, we could compute the following,

$$\begin{aligned} \begin{bmatrix} \boldsymbol{V}^0 \\ \overline{\boldsymbol{V}}^0 \end{bmatrix} &= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}, \\ \begin{bmatrix} \boldsymbol{V}^{n+1} \\ \overline{\boldsymbol{V}}^{n+1} \end{bmatrix} &= \boldsymbol{G}_1 \begin{bmatrix} \boldsymbol{V}^n \\ \overline{\boldsymbol{V}}^n \end{bmatrix}, \qquad n = 0, 1, \ldots, M-1\,, \end{aligned} \tag{9.5}$$

followed by

$$\boldsymbol{K}^n = -\boldsymbol{B}^T \boldsymbol{V}^n, \qquad n = 0, 1, \ldots, M\,, \tag{9.6}$$

and

$$\begin{aligned} \begin{bmatrix} \boldsymbol{U}^M \\ \overline{\boldsymbol{U}}^M \end{bmatrix} &= \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \\ \begin{bmatrix} \boldsymbol{U}^n \\ \overline{\boldsymbol{U}}^n \end{bmatrix} &= \boldsymbol{G}_2 \begin{bmatrix} \boldsymbol{U}^{n+1} \\ \overline{\boldsymbol{U}}^{n+1} \end{bmatrix} + \boldsymbol{F}(\boldsymbol{K}^{n+1} + \boldsymbol{K}^n), \qquad n = M-1, M-2, \ldots, 0\,, \end{aligned} \tag{9.7}$$

and we would then have directly,

$$\widetilde{\boldsymbol{\Lambda}}_M = \begin{bmatrix} \overline{\boldsymbol{U}}^0 \\ -\boldsymbol{U}^0 \end{bmatrix}.$$

To actually compute the control for driving $(\boldsymbol{y}^0, \overline{\boldsymbol{y}}^0)$ to the null state, we must solve (if possible)

$$\widetilde{\boldsymbol{\Lambda}}_M \begin{bmatrix} \boldsymbol{z}^0 \\ \overline{\boldsymbol{z}}^0 \end{bmatrix} = \begin{bmatrix} \overline{\boldsymbol{y}}^0 \\ -\boldsymbol{y}^0 \end{bmatrix}.$$

We then solve the adjoint system (9.3) with $(\boldsymbol{v}^0, \overline{\boldsymbol{v}}^0) = (\boldsymbol{z}^0, \overline{\boldsymbol{z}}^0)$ and the control is finally given by (9.4).

9.2.2 The Inner Product Method

Recall from Equation 4.28, page 73, the important relation,

$$\left\langle \widetilde{\boldsymbol{\Lambda}}_M \begin{pmatrix} \boldsymbol{v}^0 \\ \overline{\boldsymbol{v}}^0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{w}^0 \\ \overline{\boldsymbol{w}}^0 \end{pmatrix} \right\rangle_{\boldsymbol{\mathcal{C}}} = \frac{\Delta t}{4} \sum_{n=0}^{M-1} \left\langle \boldsymbol{B}^T(\boldsymbol{v}^{n+1} + \boldsymbol{v}^n), \boldsymbol{B}^T(\boldsymbol{w}^{n+1} + \boldsymbol{w}^n) \right\rangle,$$

where

$$\boldsymbol{\mathcal{C}} = \begin{bmatrix} \boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix},$$

and where $(\boldsymbol{v}^n, \overline{\boldsymbol{v}}^n)$ and $(\boldsymbol{w}^n, \overline{\boldsymbol{w}}^n)$ each are solutions of the adjoint system (9.3).

Let us immediately adopt the multi-column/matrix approach of the previous section. Let $(\boldsymbol{V}^n, \overline{\boldsymbol{V}}^n)$, $n = 0, 1, \ldots, M$, consist of $2N_0$ simultaneous solutions of the adjoint system, as described by (9.5). We now have

$$\begin{aligned}
\mathcal{C}\widetilde{\boldsymbol{\Lambda}}_M &= \left\langle \widetilde{\boldsymbol{\Lambda}}_M \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \right\rangle_{\mathcal{C}} = \left\langle \widetilde{\boldsymbol{\Lambda}}_M \begin{bmatrix} \boldsymbol{V}^0 \\ \overline{\boldsymbol{V}}^0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{V}^0 \\ \overline{\boldsymbol{V}}^0 \end{bmatrix} \right\rangle_{\mathcal{C}} \\
&= \frac{\Delta t}{4} \sum_{n=0}^{M-1} \left\langle \boldsymbol{B}^T(\boldsymbol{V}^{n+1} + \boldsymbol{V}^n), \boldsymbol{B}^T(\boldsymbol{V}^{n+1} + \boldsymbol{V}^n) \right\rangle \\
&= \frac{\Delta t}{4} \sum_{n=0}^{M-1} \left(\boldsymbol{K}^{n+1} + \boldsymbol{K}^n\right)^T \left(\boldsymbol{K}^{n+1} + \boldsymbol{K}^n\right),
\end{aligned}$$

where $\boldsymbol{K}^n$, $n = 0, 1, \ldots, M$, is computed as in (9.6).

As mentioned in Section 5.1, the inner product method has some appealing properties. Let us recall them in this particular setting.

- We *never* have to solve the control system (9.7), that is, we never have to deal with solving the wave equation with inhomogeneous boundary conditions.

- It is easy to update $\mathcal{C}\widetilde{\boldsymbol{\Lambda}}_M$ from one time step to the next:

$$\mathcal{C}\widetilde{\boldsymbol{\Lambda}}_M = \mathcal{C}\widetilde{\boldsymbol{\Lambda}}_{M-1} + \frac{\Delta t}{4}\left(\boldsymbol{K}^M + \boldsymbol{K}^{M-1}\right)^T \left(\boldsymbol{K}^M + \boldsymbol{K}^{M-1}\right). \tag{9.8}$$

  This also means that the controllability operator can be updated along with solving the adjoint system.

- The matrix $\mathcal{C}\widetilde{\boldsymbol{\Lambda}}_M$ is symmetric by construction, easily seen from the update formula (9.8). This means that it is only necessary to update half the entries of $\mathcal{C}\widetilde{\boldsymbol{\Lambda}}_M$, e.g., the upper or lower triangle.

9.3 Flop Count and Memory Usage

Flop is short for *floating point operation*, and refers to when a computer performs one of the four basic operations, $+$, $-$, $\times$ or $/$, using floating point arithmetic. To get an idea of the time usage of computing the discrete controllability operator, we will count flops and ignore everything else. Since flops are the dominating ingredient of the algorithms described, this will give a reasonable idea of how the time increases as a function of the key quantities. We will furthermore only count the flops of the *main loop* of the algorithms, and ignore, for instance, initialization.

When it comes to memory usage, we will only consider the (large) matrices that are used. The data type for floating points will be what the programming language `C`, and its descendants, call `double`. We assume each `double` takes up 8 bytes.

A size summary of the key matrices that occur in the two methods, can be seen in Table 9.1.

$\mathcal{C}$	$2N_0{\times}2N_0$	$\boldsymbol{G}_1$, $\boldsymbol{G}_2$	$2N_0{\times}2N_0$
$\widetilde{\boldsymbol{\Lambda}}_M$	$2N_0{\times}2N_0$	$\boldsymbol{F}$	$2N_0{\times}N_b$
$\boldsymbol{V}^n$, $\overline{\boldsymbol{V}}^n$	$N_0{\times}2N_0$	$\boldsymbol{B}$	$N_0{\times}N_b$
$\boldsymbol{U}^n$, $\overline{\boldsymbol{U}}^n$	$N_0{\times}2N_0$	$\boldsymbol{K}^n$	$N_b{\times}2N_0$

Table 9.1: An overview of key matrices and their sizes.

9.3.1 The Direct Method

We assume that the matrix-matrix multiply operations $C \leftarrow AB$ and $C \leftarrow C+AB$, where $A \in \mathbb{R}^{m\times k}$, $B \in \mathbb{R}^{k\times n}$, each take a total of $2mkn$ flops (this is commonly called a level 3 BLAS operation, performed by the BLAS routine DGEMM).

Let us consider the steps needed to compute the controllability operator according to (9.5), (9.6) and (9.7):

- Compute $\boldsymbol{V}^n$ and $\overline{\boldsymbol{V}}^n$ for $n = 1, 2, \ldots, M$, by premultiplying M times with $\boldsymbol{G}_1$. This demands $M \cdot 2 \cdot 2N_0 \cdot 2N_0 \cdot 2N_0 = 16MN_0^3$ flops.
- Compute $\boldsymbol{K}^n = -\boldsymbol{B}^T\boldsymbol{V}^n$, $n = 0, 1, \ldots, M$. This leads to $(M+1) \cdot 2 \cdot N_b \cdot N_0 \cdot 2N_0 = 4(M+1)N_bN_0^2$ flops.
- Computing $\boldsymbol{U}^n$ and $\overline{\boldsymbol{U}}^n$ for $n = M-1, \ldots, 1, 0$ can be done by doing

  $$\boldsymbol{T}_1 \leftarrow \boldsymbol{K}^{n+1} + \boldsymbol{K}^n, \quad \boldsymbol{T}_2 \leftarrow \boldsymbol{F}\boldsymbol{T}_1, \quad \boldsymbol{T}_2 \leftarrow \boldsymbol{T}_2 + \boldsymbol{G}_2 \begin{bmatrix} \boldsymbol{U}^{n+1} \\ \overline{\boldsymbol{U}}^{n+1} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{U}^{n} \\ \overline{\boldsymbol{U}}^{n} \end{bmatrix} \leftarrow \boldsymbol{T}_2\,,$$

  a total of M times ($\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ are temporaries). This is done using $M(2N_bN_0 + 2 \cdot 2N_0 \cdot N_b \cdot 2N_0 + 2 \cdot 2N_0 \cdot 2N_0 \cdot 2N_0) = 2MN_0(N_b + 4N_0N_b + 8N_0^2)$ flops.

We get a grand total of

$$MN_0^2(32N_0 + 12N_b) \quad \text{flops},$$

when discarding lower order terms.

When it comes to memory usage, we need the following.

- The matrices $\boldsymbol{G}_1$, $\boldsymbol{G}_2$, $\boldsymbol{B}$ and $\boldsymbol{F}$ take up $8N_0^2 + 3N_0N_b$ doubles.
- The boundary data all needs to be stored, $\boldsymbol{K}^0$, $\boldsymbol{K}^1$, $\ldots$, $\boldsymbol{K}^M$ take up $(M+1)2N_0N_b$ doubles.
- Two instances of $(\boldsymbol{V}^n, \overline{\boldsymbol{V}}^n)$ need to be present in memory at one time. The same storage place can be used for two simultaneous instances of $(\boldsymbol{U}^n, \overline{\boldsymbol{U}}^n)$. This takes up $8N_0^2$ flops.

Adding up, we need about

$$N_0(16N_0 + 2MN_b) \quad \text{doubles},$$

for the matrix version of the direct method.

The need for storing all boundary data in $\boldsymbol{K}^n$, $n = 0, 1, \ldots, M$, is a major drawback of the direct method when it comes to memory usage. An alternative is to compute only a few columns of $\widetilde{\boldsymbol{\Lambda}}_M$ at a time. This leads to the same number of flops, but will clearly lead to more overhead (function calls, initializing, and so on) and probably also more cache misses (not exploiting the fast on-chip memory in an optimal way). However, the memory usage will be reduced considerably to

$$N_0(8N_0 + 3N_0) + c(4N_0 + MN_b) \quad \text{doubles},$$

where c is the number of columns computed simultaneously, $1 \leq c \leq 2N_0$.

Another way of reducing the memory requirements was proposed in *Glowinski, Li, and Lions (1990)*, Remark 4.1. The idea is the following. Solve the adjoint system with the sole purpose of computing $(\boldsymbol{V}^M, \overline{\boldsymbol{V}}^M)$. The adjoint system can now be solved backwards from $n = M$ to $n = 0$, while simultaneously computing the relevant values of $\boldsymbol{K}^n$ and $(\boldsymbol{U}^n, \overline{\boldsymbol{U}}^n)$. This eliminates the need of storing all $M + 1$ values of $\boldsymbol{K}^n$, and reduces the memory requirement to

$$N_0(8N_0 + 3N_b) + c(6N_0 + 2N_b) \quad \text{doubles},$$

where c again is the number of columns computed simultaneously. There is a drawback, however, in that we compute the solution to the adjoint system twice. The flop count will increase to

$$MN_0^2(48N_0 + 12N_b) \quad \text{flops}.$$

9.3.2 The Inner Product Method

The operation $C \leftarrow C + \alpha A^T A$ with $A \in \mathbb{R}^{k \times n}$ and $\alpha \in \mathbb{R}$ is called a rank-k update of a symmetric matrix in the language of the level 3 BLAS (performed by the function DSYRK). We will assume that such a computation takes kn^2 flops.

The main loops of this method require the following steps.

- Compute $\boldsymbol{V}^n$ and $\overline{\boldsymbol{V}}^n$ for $n = 1, 2, \ldots, M$, by premultiplying M times with $\boldsymbol{G}_1$. This demands $16MN_0^3$ flops.

- Compute $\boldsymbol{K}^n = -\boldsymbol{B}^T \boldsymbol{V}^n$, $n = 0, 1, \ldots, M$. This leads to $4(M + 1)N_b N_0^2$ flops.

- Doing
$$\boldsymbol{T}_1 \leftarrow \boldsymbol{K}^{n+1} + \boldsymbol{K}^n,$$
followed by the update
$$\boldsymbol{L} \leftarrow \boldsymbol{L} + \tfrac{\Delta t}{4} \boldsymbol{T}_1^T \boldsymbol{T}_1, \tag{9.9}$$
a total of M times amounts to $M(2N_0 N_b + 4N_0^2 N_b)$ flops (the matrix $\boldsymbol{L}$ represents $\boldsymbol{\mathcal{C}}\widetilde{\boldsymbol{\Lambda}}_M$).

	Flop count	Memory usage in doubles
Direct	$MN_0^2(32N_0 + 12N_b)$	$N_0(8N_0 + 3N_b) + c(4N_0 + MN_b)$
Glowinski	$MN_0^2(48N_0 + 12N_b)$	$N_0(8N_0 + 3N_b) + c(6N_0 + 2N_b)$
Inner product	$MN_0^2(16N_0 + \ \ 4N_b)$	$N_0(16N_0 + 5N_b)$

Table 9.2: Comparing flop count and memory usage for three different ways of computing the discrete controllability operator. The label Glowinski refers to the Glowinski, Li, Lions version of the direct method. The constant c represents the number of columns that are computed simultaneously, $1 \leq c \leq 2N_0$.

Adding up and discarding lower order terms we get

$$MN_0^2(16N_0 + 4N_b) \quad \text{flops},$$

for the inner product method.

For memory usage, we observe that we successively compute $\boldsymbol{\mathcal{C}}\widetilde{\boldsymbol{\Lambda}}_M$ from the previous time step $\boldsymbol{\mathcal{C}}\widetilde{\boldsymbol{\Lambda}}_{M-1}$. This means that we do not need to save any information across all time steps. We sum up as follows.

- The matrices $\boldsymbol{G}_1$, $\boldsymbol{B}$ and $\boldsymbol{L}$ take up $8N_0^2 + N_0N_b$ doubles.
- Two instances of $(\boldsymbol{V}^n, \overline{\boldsymbol{V}}^n)$ need to be present in memory at one time. This takes up $8N_0^2$ flops.
- Two instances of $\boldsymbol{K}^n$ need to be present in memory at one time. This takes up $4N_0N_b$ flops.

This amounts to a total of

$$N_0(16N_0 + 5N_b) \quad \text{doubles}.$$

9.3.3 Choosing the Best

Table 9.2 presents a summary of the flop count and memory usage for each method, including the Glowinski, Li and Lions version of the direct method. When it comes to speed, according to the flop counts, the inner product method is a clear winner. It is 2–3 times faster than any of the direct methods.

When it comes to memory usage, something can actually be gained by using a direct method. But no more than a factor 1/2 for small c. As c approaches its maximum value of $2N_0$, however, the direct methods require *more* memory than the inner product method.

Let us consider a concrete case, namely that of the future Section 9.4. The geometry is a square grid with grid length N, with control on two of the sides. We set $h = 1/(N + 1)$, $\Delta t = h/2$ and $T = 4$. To use nice integral values we set

$$N_0 = N^2, \quad N_b = 2N, \quad M = 8N\,. \tag{9.10}$$

N	10	20	40	60	80	100
Flops	$1.3 \cdot 10^9$	$1.7 \cdot 10^{11}$	$2.1 \cdot 10^{13}$	$3.6 \cdot 10^{14}$	$2.7 \cdot 10^{15}$	$1.3 \cdot 10^{16}$
Time	0.67 sec.	1.4 min.	2.9 hours	2.1 days	16 days	2.5 months
Memory	1.3Mb	20Mb	317Mb	1.6Gb	4.9Gb	12Gb

Table 9.3: Example of how much time it takes to compute the controllability operator, and how much memory it requires. The number of inner points N_0, number of boundary points N_b, and number of time steps M depends on the parameter N as shown in (9.10). To compute the time row, we assume that the processor in question can perform 2 Gflops per second.

This leads to

$$128N^7 + 64N^6 \ \text{ flops and } \ 2N^3(8N+5) \ \text{ doubles,}$$

for the inner product method. See Table 9.3 for an example of how much time it takes to compute the controllability operator in this case, and how much memory it requires.

To get an asymptotic idea of the running time and memory requirements, let us set

$$N_0 = N^d, \quad N_b = N^{d-1}, \quad M = N,$$

for some dimension parameter d and a grid-size parameter N. The reasoning behind these numbers are as follows. Let d be the space dimension we consider. The order of inner grid points will then be N^d, like a d-dimensional cube with side length N. One of the sides of such a cube will then contain N^{d-1} points. Let the grid point distance be $h = 1/N$ and $\Delta t = h$. The order of $M \simeq T/\Delta t$ will now be N.

Inserting in the flop count formula for the inner product method, we get order N^{3d+1} for the flop count and order N^{2d} for the memory usage.

ZZZZZZZZZZZZ...

— ANONYMOUS

9.3.4 Multiple Processors

Let us briefly consider how to do an implementation that computes the controllability operator in *parallel*.

The direct method can lead to *perfect* speed-up in a distributed memory computing environment. Such an implementation is also called *embarrassingly parallel*. Perfect speed-up means that the execution time is inversely proportional to the number of processors. This is typically only possible if no communication is needed during execution, or at least in the main loop. This is exactly the situation for the direct method. Each processor can be informed about which columns of

$\boldsymbol{\Lambda}_T$ to compute, and each processor can now do this, independently of the others. Note that this could also easily be done on a shared memory architecture.

Perfect speed-up is not possible for the inner product method with distributed memory. The matrices $\boldsymbol{V}^n$, $\overline{\boldsymbol{V}}^n$ and $\boldsymbol{K}^n$ can be distributed and computed independently on each processor, but the update (9.9) requires a considerable amount of communication. Assume now that the following is possible on a supercomputer with *shared* memory: Each processor has read/write access to disjoint columns of $\boldsymbol{V}^n$, $\overline{\boldsymbol{V}}^n$, $\boldsymbol{K}^n$, $\boldsymbol{L} = \boldsymbol{\mathcal{C}}\boldsymbol{\Lambda}_T$ and, furthermore, every processor has read access to the *whole* $\boldsymbol{K}^n$ matrix. Apart from synchronization before each processor's computation of $\boldsymbol{K}^n$ and before each update of $\boldsymbol{L}$, all computations can be done completely independently. This will lead to almost perfect speed-up.

9.4 Illustrations in 2D

An implementation of the inner product method was carried out in the programming language C, using the high performance libraries BLAS (Basic Linear Algebra Subprograms, see `www.netlib.org/blas`) and LAPACK (Linear Algebra PACKage, see `www.netlib.org/lapack`).

The program was run on two different computers: A PC with an Athlon 2000+ XP processor and a SunFire 15k shared memory computer with UltraSparc-IIICu 900 MHz processors (no parallelization was done). The latter is a part of the Sun High Performance Computing Systems at DTU, see `www.hpc.dtu.dk`.

Figure 9.2 visualizes a solution where a control has been applied, such that the system is driven to rest. The controllability operator was computed using the inner product method, control time $T = 3$, and the control was then computed as described at the end of Section 9.2.1. The initial conditions, to be controlled, were as indicated at $t = 0.0$ in the figure (initial velocities were identically zero).

9.5 Preconditioning

Let us consider computing controls in general. An essential ingredient is inverting the controllability operator, that is, solving a linear system of equations of the type $\boldsymbol{L}\boldsymbol{x} = \boldsymbol{y}$. In practice, rounding errors will influence the precision of the solution. How much will depend on the condition number of the system matrix,

$$\kappa(\boldsymbol{L}) = \|\boldsymbol{L}\|\|\boldsymbol{L}^{-1}\| ,$$

where we here use the discrete 2-norm for the definition. For a symmetric matrix $\boldsymbol{L}$, this corresponds to the ratio between the largest and the smallest (in magnitude) eigenvalues. The higher the condition number, the worse the accuracy of the solution due to rounding errors. The point of *preconditioning* is to solve an equivalent problem,

$$\big(\boldsymbol{P}_1\boldsymbol{L}\boldsymbol{P}_2\big)(\boldsymbol{P}_2^{-1}\boldsymbol{x}) = \boldsymbol{P}_1\boldsymbol{y} ,$$

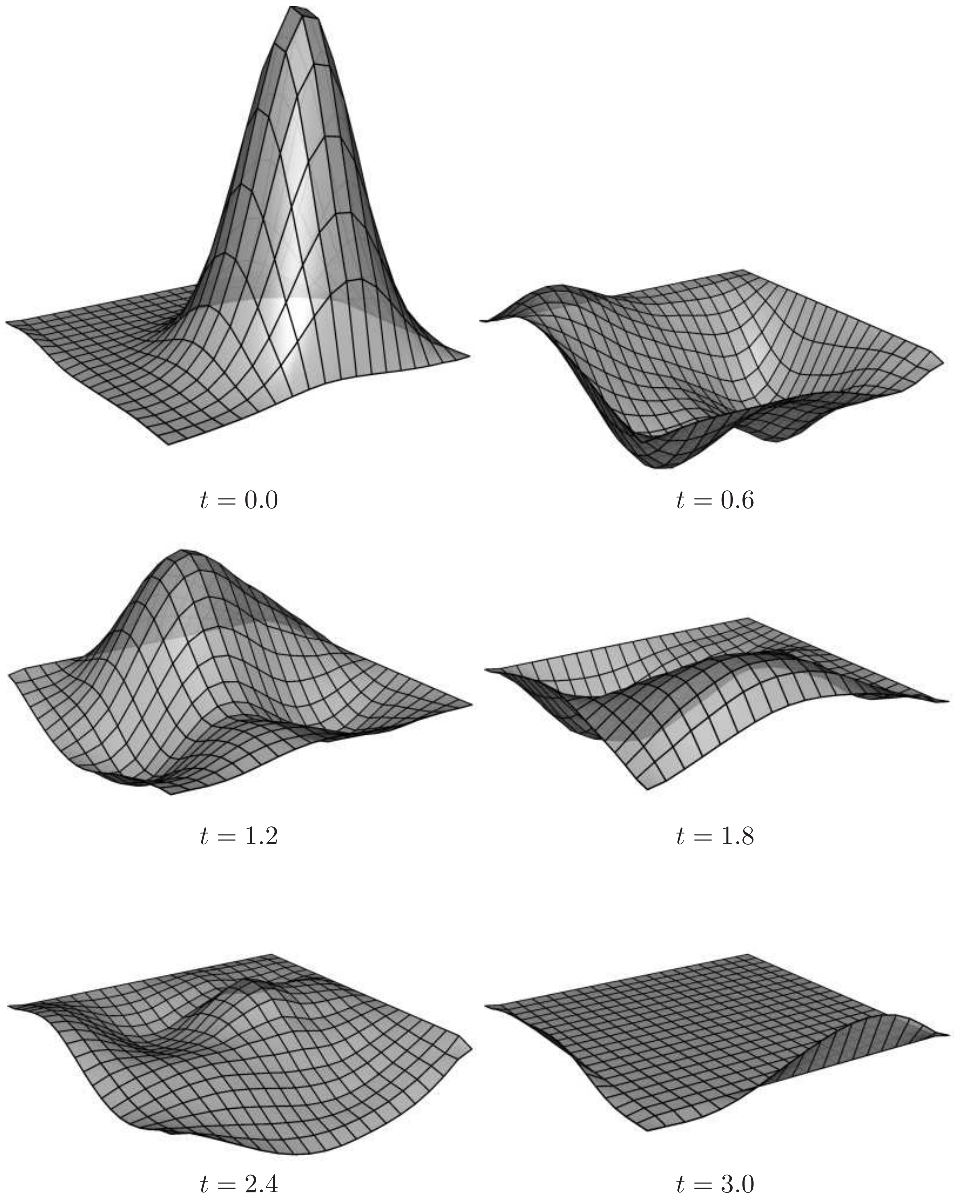


Figure 9.2: Visualization of a solution to an exact control problem of the wave equation in 2D. (The non-zero boundary at $t = 3.0$ is not an error, it would disappear in the limit, in an $L^2((0,1)^2)$ sort of way.)

where the matrix $\boldsymbol{P}_1\boldsymbol{L}\boldsymbol{P}_2$ has a smaller condition number that $\boldsymbol{L}$ itself. The big challenge, in general, is of course to find appropriate preconditioners $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$. If we have a discretization scheme with uniform observability, however, Section 4.2 provides the answer. Instead of solving

$$\boldsymbol{\Lambda}_M^{\Delta t}\boldsymbol{v} = \boldsymbol{M}^T\boldsymbol{y}\,, \tag{9.11}$$

we should use the preconditioned system,

$$\big(\boldsymbol{R}^{-T}\boldsymbol{\mathcal{C}}\boldsymbol{\Lambda}_M^{\Delta t}\boldsymbol{R}^{-1}\big)(\boldsymbol{R}\boldsymbol{v}) = \boldsymbol{R}^{-T}\boldsymbol{\mathcal{C}}\boldsymbol{M}^T\boldsymbol{y}\,, \tag{9.12}$$

where $\boldsymbol{R}^T\boldsymbol{R} = \widetilde{\boldsymbol{Q}}$ with $\boldsymbol{R}$ upper triangular. As shown in Section 4.2, uniform observability implies that the condition number of $\boldsymbol{R}^{-T}\boldsymbol{\mathcal{C}}\boldsymbol{\Lambda}_M^{\Delta t}\boldsymbol{R}^{-1}$ is bounded by a constant, independently of N. A further advantage of the above rewrite is that the matrix $\boldsymbol{R}^{-T}\boldsymbol{\mathcal{C}}\boldsymbol{\Lambda}_M^{\Delta t}\boldsymbol{R}^{-1}$ is seen to be symmetric and positive definite, an advantage when solving linear systems of equation.

9.5.1 A Preconditioner of Glowinski, Li and Lions

In the paper *Glowinski, Li, and Lions (1990)* the authors consider exact boundary controllability of the wave equation, and they solve Equation (9.11) using a conjugate gradient (CG) algorithm. This is an iterative Krylov subspace method, which works for symmetric and positive definite matrices.

It can be discussed whether or not an iterative algorithm is fitted for *exact* controllability problems. If CG performs a number of iterations that corresponds to the order the system, it is solved exactly, but stopping criteria are typically used for stopping prematurely. Indeed, in the paper mentioned, they stop the CG algorithm when an appropriate norm of the residual, $\big(\boldsymbol{M}^T\boldsymbol{y} - \boldsymbol{\Lambda}_M^{\Delta t}\boldsymbol{v}\big)$, gets below a certain threshold.

On top of this, they actually use a *preconditioned* conjugate gradient algorithm. Let us give an idea of their preconditioner by considering exact controllability for the continuous wave equation. The controllability operator maps between the following spaces,

$$\Lambda_T : \begin{pmatrix} H_0^1(\Omega) \\ L^2(\Omega) \end{pmatrix} \mapsto \begin{pmatrix} H^{-1}(\Omega) \\ L^2(\Omega) \end{pmatrix}\,.$$

This already indicates a problem in that the spaces are *different*. The preconditioner they use consists of solving an equation of the type

$$\begin{bmatrix} -\Delta & 0 \\ 0 & I \end{bmatrix} z = r,$$

internally in the CG algorithm. Solving the system $\Lambda_T x = y$ this way corresponds to solving

$$\left(\begin{bmatrix} -\Delta & 0 \\ 0 & I \end{bmatrix}^{-1/2} \Lambda_T \begin{bmatrix} -\Delta & 0 \\ 0 & I \end{bmatrix}^{-1/2}\right)\left(\begin{bmatrix} -\Delta & 0 \\ 0 & I \end{bmatrix}^{1/2} x\right) = \begin{bmatrix} -\Delta & 0 \\ 0 & I \end{bmatrix}^{-1/2} y\,,$$

see Section 10.3 of *Golub and Van Loan (1996)*. The approach makes sense for the following heuristic reason. Since $-\Delta : H_0^1(\Omega) \mapsto H^{-1}(\Omega)$, its square root will act as

$$(-\Delta)^{1/2} : H_0^1(\Omega) \mapsto L^2(\Omega) \quad \text{or} \quad (-\Delta)^{1/2} : L^2(\Omega) \mapsto H^{-1}(\Omega)\,.$$

This implies that the domain and range are

$$\begin{bmatrix} -\Delta & 0 \\ 0 & I \end{bmatrix}^{-1/2} \Lambda_T \begin{bmatrix} -\Delta & 0 \\ 0 & I \end{bmatrix}^{-1/2} : \begin{pmatrix} L^2(\Omega) \\ L^2(\Omega) \end{pmatrix} \mapsto \begin{pmatrix} L^2(\Omega) \\ L^2(\Omega) \end{pmatrix}\,,$$

so the preconditioned map is now between equivalent spaces. Using the same idea in the discrete case can now be used to reduce the condition number of the system to solve.

It should be noted that uniform observability does *not* hold for their discretization scheme. Indeed, they observe that high frequency waves lead to unwanted oscillations in the computed controls.

Although the above reasoning is quite reasonable (the authors themselves never argue why they use that particular preconditioner), it must still be considered the best solution to choose discrete norms that correspond to the continuous ones and then to use the preconditioned system given by (9.12).

9.5.2 Null-controllability and Discrete Ill-posed Problems

It's impossible to compute things which don't exist.
It's difficult to compute things which almost don't exist.
— Cleve Moler, 1997

Some null-controllability problems can present difficulties when solving them in practice. Assume for some control system that the associated observability inequality is fulfilled (see Theorem 2.3.1, page 19). Recall then that the null-controllability problem can be solved by finding w^0 such that

$$\Lambda_T w^0 = -\mathcal{M}^T L_T u^0\,. \tag{9.13}$$

Let us consider the heat equation as an example. In this case, the controllability operator $\Lambda_T : H_0^1(\Omega) \mapsto H^{-1}(\Omega)$ is *not* invertible. This is because the output of the controllability operator is the result of solving the heat equation, a highly smoothing and dampening process. The observability inequality, however, makes sure that when L_T is applied to the right-hand side, (9.13) is solvable.

When solving a discrete analog of (9.13) in the case of the heat equation, we encounter a *discrete ill-posed problem*, since it is a discretization of an ill-posed problem (it is ill-posed because of the unboundedness/discontinuity of the inverse Λ_T^{-1}, and because right-hand sides y to the equation $\Lambda_T x = y$ exist for which no solution exists). The term, discrete ill-posed problem, is most commonly used for discretizations of Fredholm integral equations of the first kind, but we will use the term here more broadly because of a similar smoothing property of Λ_T.

A discretization of Λ_T naturally inherits its smoothing property. The problem is now that when computing with finite precision, rounding errors will often blow up when applying the discrete analog of Λ_T^{-1}. It depends, though, on the method used for solving the discrete system. A good choice is probably one based on an eigenvalue decomposition, where the components of the right-hand side that correspond to the smallest eigenvalues, are ignored or dampened. This is a *regularization method* where one must be aware, of course, that only regularized solutions are obtained. It must then be showed, if possible, that the full solution is obtained in the limit.

In *Carthel, Glowinski, and Lions (1994)*, the authors consider exact controllability for the heat equation and observe that the final state has to be very smooth. They then apply regularization to compute solutions to approximate controllability problems.

See *Hansen (1998)* for an extensive treatment of discrete ill-posed problems and regularization methods.

Chapter 10

Discussion

We should know clearly before we discuss this matter;
to guess is one thing, to know clearly another.
— Aeschylus (525–456 B.C.)

Let us run through the main topics of this thesis along with open questions and suggestions for further work.

We initially established a theoretical foundation for boundary controllability of linear evolution PDEs. We showed that different types of controllability for a control system were equivalent to different types of observability inequalities for a corresponding adjoint system. The Hilbert Uniqueness Method, HUM, was presented, which is a powerful and constructive method for computing controls.

A natural abstraction of this theory would be possible, if we wrote the control system in the lines of $\dot{u}(t) = \mathcal{A}u(t) + \mathcal{B}k(t)$, where k is the control. In such a formulation, the operator $\mathcal{B}$ could represent both boundary control and internal control (see, for example, *Bensoussan, 1990*). Although constructs such as boundary integrals become less obvious/more abstract, the continuous system would actually resemble a typical semi-discretization, $\dot{\boldsymbol{u}}(t) = \boldsymbol{\mathcal{A}}\boldsymbol{u}(t) + \boldsymbol{\mathcal{B}}\boldsymbol{k}(t)$, much more.

Further generalization could also be considered by making the operator $\mathcal{A}$ space-dependent. This complicates matters considerably, however. Consider the wave equation as an example. With constant coefficients we know that waves propagate along straight lines. With variable coefficients some waves can actually propagate inside the domain and *never* reach the boundary (see *Zuazua (2003)* and references therein). This means that exact controllability can be impossible even though the domain is bounded.

Controllability problems are easily formulated for non-linear PDEs. The whole "machinery" presented in this thesis, however, relies heavily on the linearity. See the survey article *Zhang and Zuazua (2003a)*, and the references therein, on how to handle non-linear controllability problems, including a HUM-like approach.

When it came to discretizing in space, we derived fairly general statements for semi-discrete systems of the type $\dot{\boldsymbol{u}}(t) = \boldsymbol{\mathcal{A}}\boldsymbol{u}(t)+\boldsymbol{\mathcal{B}}\boldsymbol{k}(t)$. Among the most important results was that if an observability inequality holds on each discretization level with *uniform* constants, then the corresponding controls were guaranteed to converge to the true control. When it came to proving controllability properties of the discrete systems, however, we relied heavily on knowing the eigensolutions of $\boldsymbol{\mathcal{A}}$. This was possible because we considered finite difference discretizations on regular grids in one dimension. When moving to more dimensions with irregular domains and/or grids, it becomes harder, if not impossible, to have explicit eigenvalue information. Some hope may lie in the area of group velocity, since this is a local (in space) property. Whether it is possible to formulate sufficient and necessary conditions for uniform observability using analysis of group velocity is not clear. Some attempts, though still for finite difference approximations on regular grids, have been made in *Maciá (2003)*. It seems to the author, though, that uniform observability can only hold for a few, fabricated schemes and domains. In more general cases, one has to resort to filtering or other types of regularizing. Here still lies a lot of work, both in proving convergence of controls for regularized schemes and in efficient implementations.

Moving from semi-discrete systems to fully discrete systems complicates matters a little, but not much. On the other hand, it seems that a special relation must exist between the space and time discretization if any form of uniform observability is to hold. We considered the explicit midpoint rule and the trapezoid rule for time discretization. It should be possible to carry through a more general treatment of time discretizations, such as linear multi-step schemes.

The controllability operator would be worth a study in itself. But of course, its properties are intimately tied to PDEs and controllability. We focused on two aspects, namely how to compute a matrix representation for this operator and proving the existence of a limit operator as the control time went to infinity, $T \to \infty$. One way to compute a matrix representation, the direct method, was simply based on its definition. Another method, the inner product method, relied on one of its fundamental properties. Although never studied before, this latter method seemed the simplest in pen-and-paper calculations and it also proved the most efficient method to use for computer implementations.

The study of a possible limit operator for the controllability operator as $T \to \infty$ was limited to two particular equations, the wave equation and the heat equation. The existence of a limit operator for the wave equation relied on the fact that we were dealing with a skew-symmetric operator, implying special eigenvalue properties. The results would easily be extendible to general skew-symmetric system operators, see *Bensoussan (1990)* and *Bensoussan (1992)*. The limit operator for the heat equation also relied on some special eigenvalue properties. The controllability operator for the heat equation furthermore suggests that a limit operator may exist as $T \to 0$. This should be studied.

An obvious question is: Does a well-defined limit operator always exist for any well-defined controllability operator? This is not clear and is worth some study. Why is the existence of a limit operator even so interesting? For one, because it tells something about the control function's dependence of the control time T.

For instance, for the wave equation we could deduce that the norm of the control function was proportional to $1/T$ as $T \to \infty$. For the heat equation, that the norm of the control function approached a constant level as $T \to \infty$. Another possible usage of knowledge about a limit operator is that it may turn out to be a good *preconditioner* for computer implementations. This is currently unexplored.

Null-controllability using Dirichlet control was easy to show for the heat equation in one dimension. Likewise, uniform observability was straightforward to show for a semi-discretization. The nature of the heat equation suggests that similar results hold for more dimensions, and similarly for a full discretization. However, no tools or theorems, such as a time-discrete analog to the parabolic version of Ingham's Theorem, seem to be available in order to show uniform observability for a full discretization. It is suspected, though, that it holds without any surprises.

In contrast to the fact that uniform observability holds in the discrete case, is that the controls are highly unstable to compute in practice. This is due to the fact that the controllability operator has a very dampening effect, similar to discrete ill-posed problems. This means that even though the "right-hand side" is sufficiently smooth, rounding errors can have disastrous effects on the solution. This depends, of course, on the algorithm used to obtain the solution. Further study is needed in this area.

Controllability of the wave equation has, by far, gotten most attention in the literature, both in the continuous and the discrete cases. However, many things are worth more study.

The characterization of all possible controls for the one-dimensional wave equation, through the analysis of null-space controls, was quite interesting. But the analysis made use of an explicit solution formula (the D'Alembert solution formula) for the wave equation in one dimension, and such simple formulas are not available in several dimensions. However, the concept of null-space controls deserves some more study, and may lead to greater insight into controllability of the wave equation in several dimensions.

More study of discretization schemes is also needed, both time and space discretization. A previously unexplored time–space scheme was presented in Chapter 7. Uniform observability did not hold, but it had some characteristics which makes it promising in the area of control. Further analysis of this scheme is needed.

An interesting, and currently unanswered, question is: Is there any hope of having uniform observability on irregular grids, even in one dimension? From the study of group velocity properties, there is hope for uniform observability on a *regular* two-dimensional grid, cf. the scheme described in Sections 3.4.1 and 9.1.

The linear system of thermoelasticity was interesting because it provided an example of what could be done if the control system operator was fairly complicated, and only asymptotic knowledge of the eigenvalues was known. The system could be viewed as a coupling of a wave equation, for which controllability puts a condition on the control time, and a heat equation, for which only null-controllability is possible. The coupled system required both: Only null-controllability was possible and a minimum control time was required. Furthermore, no multiple eigenvalues were allowed, and this could happen.

During the controllability proof of the system of thermoelasticity, we showed that null-controllability was possible for (disjoint) *projections*. This was fairly straightforward and the difficulty of the proof lay in showing that both could be fulfilled simultaneously. It seems that it should be possible to derive general results for when such controllability results of projections can be combined. That would be a powerful result.

The Hilbert Uniqueness Method for computing controls for a discretization of the two-dimensional wave equation was implemented efficiently using high performance libraries such as BLAS and LAPACK. As the dimension increased, however, the running time increased frighteningly fast. A better running time complexity may be obtained, if a discretization scheme is used where matrix sparsity can be exploited effectively.

As mentioned earlier, when moving to several dimensions and/or irregular grids, the only hope of convergence of controls may be using an appropriate form of regularization. How should such a regularized method be implemented efficiently? This would be highly relevant as soon as the proper theoretical foundation has been established.

Let us finish with two quotes that seem appropriate at this point.

All of the sudden, Larry the Cow was in control.
And he liked it.
— The gentoo.org website

To finish a work? To finish a picture? What nonsense!
To finish it means to be through with it, to kill it,
to rid it of its soul, to give it its final blow...
— Pablo Picasso (1881–1973)

Appendix A

Details

Beware of the man who won't be bothered with details.
— William Feather (1908–1976)

Detail 1

Theorem. *Let H be a Hilbert space and consider the functional $J : H \mapsto \mathbb{R}$ defined as $J(v) = \|v\|$. Then for an arbitrary $v \neq 0$ we have,*

$$\langle \nabla J(v), w \rangle = \frac{\langle v, w \rangle}{\|v\|} \quad \text{for all } w \in H \, .$$

Proof. Fix arbitrary vectors $v, w \in H$ where $v \neq 0$. Define now

$$g(h) = \|v + hw\| \, .$$

Observe $(g^2(h))' = 2g(h)g'(h)$ which implies

$$g'(0) = \frac{(g^2(0))'}{2g(0)} \, .$$

We now get

$$(g^2(0))' = \lim_{h \to 0} \frac{\|v + hw\|^2 - \|v\|^2}{h} = \lim_{h \to 0} \frac{2h\langle v, w \rangle + h^2 \langle w, w \rangle}{h} = 2\langle v, w \rangle \, ,$$

so finally we have

$$\langle \nabla J(v), w \rangle = g'(0) = \frac{2\langle v, w \rangle}{2\|v\|} = \frac{\langle v, w \rangle}{\|v\|} \, .$$

□

Detail 2

Theorem. *Let $k \in \mathbb{Z}$. Then we have*

$$\sum_{j=1}^{N} \cos(kj\pi/(N+1)) = \begin{cases} N, & \text{for } 2(N+1) \mid k\,, \\ -1, & \text{for } 2(N+1) \nmid k,\ 2 \mid k\,, \\ 0, & \text{for } 2 \nmid k\,. \end{cases}$$

Proof. We split into four cases.

- $2(N+1) \mid k$. Straightforward, since $\cos(kj\pi/(N+1)) = 1$ for all j.
- k odd, N even.

$$\begin{aligned} &\sum_{j=1}^{N} \cos(kj\pi/(N+1)) \\ &= \sum_{j=1}^{N/2} \big(\cos(kj\pi/(N+1)) + \cos(k(N+1-j)\pi/(N+1))\big) \\ &= (1+\cos(k\pi)) \sum_{j=1}^{N/2} \cos(kj\pi/(N+1)) = 0\,. \end{aligned}$$

- k odd, N odd.

$$\begin{aligned} &\sum_{j=1}^{N} \cos(kj\pi/(N+1)) \\ &= \cos(\tfrac{1}{2}k\pi) + \sum_{j=1}^{(N-1)/2} \big(\cos(kj\pi/(N+1)) + \cos(k(N+1-j)\pi/(N+1))\big) \\ &= (1+\cos(k\pi)) \sum_{j=1}^{(N-1)/2} \cos(kj\pi/(N+1)) = 0\,. \end{aligned}$$

- k even, $2(N+1) \nmid k$. The easiest way to proceed is to consider complex exponentials,

$$\begin{aligned} &\Big(1 + e^{i\frac{1}{2}k\pi/(N+1)}\Big) \sum_{j=0}^{N} e^{ikj\pi/(N+1)} = \sum_{j=0}^{N} \Big(e^{ikj\pi/(N+1)} + e^{ik(j+\frac{1}{2})\pi/(N+1)}\Big) \\ &= \sum_{j=0}^{2N+1} e^{i\frac{1}{2}kj\pi/(N+1)} = \sum_{j=0}^{N} \Big(e^{i\frac{1}{2}kj\pi/(N+1)} + e^{i\frac{1}{2}k(j+N+1)\pi/(N+1)}\Big) \\ &= \Big(1 + e^{i\frac{1}{2}k\pi}\Big) \sum_{j=0}^{N} e^{i\frac{1}{2}kj\pi/(N+1)}\,. \end{aligned} \tag{A.1}$$

Let now $k = p \cdot 2^q$, where p is an odd integer and q is a positive integer.

If $q = 1$, we have $(1 + e^{i\frac{1}{2}k\pi}) = (1 + e^{ip\pi}) = 0$ and we get from (A.1) (the parenthesis on the left-hand side can not be zero because of the assumption $2(N+1) \nmid k$),

$$\sum_{j=0}^{N} e^{ikj\pi/(N+1)} = \sum_{j=0}^{N} \cos(kj\pi/(N+1)) + i\sum_{j=0}^{N} \sin(kj\pi/(N+1)) = 0 \quad \Rightarrow$$
$$\sum_{j=1}^{N} \cos(kj\pi/(N+1)) = -1\,.$$

For $q > 1$ we rewrite (A.1) into

$$\sum_{j=0}^{N} e^{ikj\pi/(N+1)} = \frac{1 + e^{i\frac{1}{2}k\pi}}{1 + e^{i\frac{1}{2}k\pi/(N+1)}} \sum_{j=0}^{N} e^{i\frac{1}{2}kj\pi/(N+1)}\,.$$

Since neither the numerator or the denominator of the fraction can be zero, we can repeatedly halve k, until the above case $q = 1$ can be applied (hence, we use induction).

$\square$

Detail 3

Proof of Theorem 3.1.3 (page 37). We rewrite as follows,

$$\sum_{j=1}^{N} \sin(kj\pi/(N+1)) \sin(lj\pi/(N+1))$$
$$= \tfrac{1}{2} \sum_{j=1}^{N} \cos((k-l)j\pi/(N+1)) - \tfrac{1}{2} \sum_{j=1}^{N} \cos((k+l)j\pi/(N+1)).$$

Using the theorem of Detail 2, the result follows. □

Detail 4

Let $(\alpha_0, \alpha_1, \ldots, \alpha_m) = (\frac{1}{2}, 1, 1, \ldots, 1, \frac{1}{2})$, $h \neq 0$ and two real sequences $\langle u_k \rangle_{k=-1}^{m+1}$, $\langle v_k \rangle_{k=-1}^{m+1}$ be given. We get

$$\begin{aligned}
&h \sum_{k=0}^{m} \alpha_k \left(\frac{v_{k+1} - 2v_k + v_{k-1}}{h^2} u_k - \frac{u_{k+1} - 2u_k - u_{k-1}}{h^2} v_k \right) \\
&= \frac{1}{h} \sum_{k=0}^{m} \alpha_k \left(v_{k+1}u_k - 2v_k u_k + v_{k-1}u_k - u_{k+1}v_k + 2u_k v_k - u_{k-1}v_k \right) \qquad \text{(A.2)} \\
&= \frac{1}{h} \left(\sum_{k=0}^{m} \alpha_k v_{k+1} u_k + \sum_{k=0}^{m} \alpha_k v_{k-1} u_k - \sum_{k=0}^{m} \alpha_k u_{k+1} v_k - \sum_{k=0}^{m} \alpha_k u_{k-1} v_k \right) .
\end{aligned}$$

We rewrite each sum in turn

$$\begin{aligned}
\sum_{k=0}^{m} \alpha_k v_{k+1} u_k &= \alpha_0 v_1 u_0 + \alpha_{m-1} v_m u_{m-1} + \alpha_m v_{m+1} u_m + \sum_{k=1}^{m-2} \alpha_k v_{k+1} u_k \, , \\
\sum_{k=0}^{m} \alpha_k v_{k-1} u_k &= \alpha_0 v_{-1} u_0 + \alpha_1 v_0 u_1 + \alpha_m v_{m-1} u_m + \sum_{k=1}^{m-2} \alpha_{k+1} v_k u_{k+1} \, , \\
\sum_{k=0}^{m} \alpha_k u_{k+1} v_k &= \alpha_0 u_1 v_0 + \alpha_{m-1} u_m v_{m-1} + \alpha_m u_{m+1} v_m + \sum_{k=1}^{m-2} \alpha_k u_{k+1} v_k \, , \\
\sum_{k=0}^{m} \alpha_k u_{k-1} v_k &= \alpha_0 u_{-1} v_0 + \alpha_1 u_0 v_1 + \alpha_m u_{m-1} v_m + \sum_{k=1}^{m-2} \alpha_{k+1} u_k v_{k+1} \, .
\end{aligned}$$

Using these, the final expression in Equation (A.2) can be simplified:

$$\begin{aligned}
&\frac{1}{h} \big(\alpha_0 v_1 u_0 + \alpha_{m-1} v_m u_{m-1} + \alpha_m v_{j+1} u_m + \alpha_0 v_{i-1} u_0 + \alpha_1 v_0 u_1 + \alpha_m v_{m-1} u_m \\
&\quad - \alpha_0 u_1 v_0 - \alpha_{m-1} u_m v_{m-1} - \alpha_m u_{j+1} v_m - \alpha_0 u_{i-1} v_0 - \alpha_1 u_0 v_1 - \alpha_m u_{m-1} v_m \big) \\
&= \frac{1}{h} \big((\alpha_0 - \alpha_1) v_1 u_0 + (\alpha_{m-1} - \alpha_m) v_m u_{m-1} + \alpha_m v_{j+1} u_m + \alpha_0 v_{i-1} u_0 \\
&\qquad + (\alpha_1 - \alpha_0) v_0 u_1 + (\alpha_m - \alpha_{m-1}) v_{m-1} u_m - \alpha_m u_{j+1} v_m - \alpha_0 u_{i-1} v_0 \big) \\
&= \frac{1}{h} \big(-\tfrac{1}{2} v_1 u_0 + \tfrac{1}{2} v_m u_{m-1} + \tfrac{1}{2} v_{j+1} u_m + \tfrac{1}{2} v_{i-1} u_0 \\
&\qquad + \tfrac{1}{2} v_0 u_1 - \tfrac{1}{2} v_{m-1} u_m - \tfrac{1}{2} u_{j+1} v_m - \tfrac{1}{2} u_{i-1} v_0 \big) \\
&= \frac{v_{j+1} - v_{m-1}}{2h} u_m - \frac{v_1 - v_{i-1}}{2h} u_0 - \frac{u_{j+1} - u_{m-1}}{2h} v_m + \frac{u_1 - u_{i-1}}{2h} v_0 \, .
\end{aligned}$$

The result can easily be generalized by using inner products instead of multiplication, and vectors instead of real numbers.

Detail 5

Let $|\hat{g}(\xi)| = \left|\frac{4\cos(\pi\xi)}{1-4\xi^2}\right|$. For $\xi > 1/2$ we get

$$|\hat{g}(\xi)| \le \frac{4}{4\xi^2-1} \ \le \ \frac{2}{\xi^2} \quad \Leftrightarrow \quad |\xi| \ge \frac{1}{\sqrt{2}} \simeq 0.7\,.$$

For $-1 < z < 1$ we have with $\xi = \frac{1}{2} + z$,

$$\begin{aligned} |\hat{g}(\tfrac{1}{2}+z)| &= \left|\frac{\sin(\pi z)}{z(1+z)}\right| = \frac{\pi}{1+z}\frac{\sin(\pi z)}{\pi z} \le \frac{4}{1+z} \ \le \ \frac{2}{(\frac{1}{2}+z)^2} \\ \Leftrightarrow \quad & \frac{-3-\sqrt{5}}{4} \le \tfrac{1}{2} + z \le \frac{1+\sqrt{5}}{4} \simeq 0.8\,. \end{aligned}$$

Hence, $|\hat{g}(\xi)| < 2/\xi^2$ for all $\xi \in \mathbb{R}$.

Detail 6

We wish, given $T > 0$, to calculate

$$\Lambda_T \begin{pmatrix} v^0 \\ \overline{v}^0 \end{pmatrix} = \begin{pmatrix} u_t(0,\cdot) \\ -u(0,\cdot) \end{pmatrix} ,$$

where $(v^0, \overline{v}^0) \in H_0^1(0,1) \times L^2(0,1)$ and

$$\begin{cases} v_{tt}(t,x) = v_{xx}(t,x), & \text{in } (0,T) \times (0,1)\,, \\ v(t,0) = v(t,1) = 0, & \text{in } (0,T)\,, \\ v(0,x) = v^0(x), v_t(0,x) = \overline{v}^0(x), & \text{in } (0,1)\,, \end{cases} \tag{A.3}$$

and

$$\begin{cases} u_{tt}(t,x) = u_{xx}(t,x), & \text{in } (0,T) \times (0,1)\,, \\ u(t,0) = 0, u(t,1) = v_x(t,1), & \text{in } (0,T)\,, \\ u(T,x) = u_t(T,x) = 0, & \text{in } (0,1)\,. \end{cases} \tag{A.4}$$

Since $\langle \sin(k\pi\cdot)\rangle_{k=1}^{\infty}$ constitutes an orthogonal basis for $L^2(0,1)$, we cover all $(v^0, \overline{v}^0) \in H_0^1(0,1) \times L^2(0,1)$ by looking at

$$v^0(x) = \sum_{k=1}^{\infty} v_k^0 \sin(k\pi x)\,, \quad \overline{v}^0(x) = \sum_{k=1}^{\infty} \overline{v}_k^0 \sin(k\pi x)\,,$$

with $\langle k v_k^0\rangle_{k=1}^{\infty}, \langle \overline{v}_k^0\rangle_{k=1}^{\infty} \in \ell^2$.

Because of the linearity of Λ_T, we can start by looking at what this map does to

$$v^0(x) = \sin(k\pi x), \quad \overline{v}^0(x) = 0, \quad k \in \mathbb{N}\,. \tag{A.5}$$

The solution to (A.3) is easily seen to be

$$v(t,x) = \cos(k\pi t)\sin(k\pi x)\,,$$

and for *uneven* k we have $v_x(t,1) = -k\pi\cos(k\pi t)$. The solution to (A.4) can now be written in the form

$$u(t,x) = -k\pi\cos(k\pi t)x + \sum_{n=1}^{\infty} a_n(t)\sin(n\pi x)\,, \tag{A.6}$$

so the boundary condition is guaranteed to hold. Because of $u_{tt} = u_{xx}$ we get

$$\sum_{n=1}^{\infty} \big(a_n''(t) + n^2\pi^2 a_n(t)\big)\sin(n\pi x) = -k^3\pi^3\cos(k\pi t)x\,, \tag{A.7}$$

which must hold for $t \in (0,T)$ and where the initial conditions can be derived from the fact that $u(T,x) = u_t(T,x) = 0$,

$$\begin{aligned}\sum_{n=1}^{\infty} a_n(T)\sin(n\pi x) &= k\pi\cos(k\pi T)x\,,\\ \sum_{n=1}^{\infty} a_n'(T)\sin(n\pi x) &= -k^2\pi^2\sin(k\pi T)x\,,\end{aligned} \tag{A.8}$$

for $x \in (0,1)$. By multiplying each side of the equations with $\sin(m\pi x)$ and integrating over $x = 0\ldots 1$, we transform (A.7) and (A.8) into

$$a_n''(t) + n^2\pi^2 a_n(t) = \begin{cases} \dfrac{2k^3\pi^2}{n}\cos(k\pi t), & n \geq 1 \text{ even}\,,\\ -\dfrac{2k^3\pi^2}{n}\cos(k\pi t), & n \geq 1 \text{ uneven}\,,\end{cases}$$

$$a_n(T) = \begin{cases} -\dfrac{2k}{n}\cos(k\pi T), & n \geq 1 \text{ even}\,,\\ \dfrac{2k}{n}\cos(k\pi T), & n \geq 1 \text{ uneven}\,,\end{cases}$$

$$a_n'(T) = \begin{cases} \dfrac{2k^2\pi}{n}\sin(k\pi T), & n \geq 1 \text{ even}\,,\\ -\dfrac{2k^2\pi}{n}\sin(k\pi T), & n \geq 1 \text{ uneven}\,.\end{cases}$$

This ordinary differential equation is solved straightforwardly and we get

$$a_k(t) = \tfrac{3}{2}\cos(k\pi t) + k\pi(T-t)\sin(k\pi t) + \tfrac{1}{2}\cos(k\pi(2T-t))\,,$$

for the special case $n = k$ and

$$\begin{aligned} a_n(t) = (-1)^{n+1}\frac{2k}{n^2-k^2}\Big(&n\cos(k\pi T)\cos(n\pi(T-t))\\ &+ k\sin(k\pi T)\sin(n\pi(T-t)) - \frac{k^2}{n}\cos(k\pi t)\Big)\,,\end{aligned}$$

for $n \neq k$.

The calculations for even k are identical, except for a change of sign. So for initial data given by (A.5) we insert into (A.6), use that

$$x = \sum_{n=1}^{\infty}(-1)^{n+1}\frac{2}{\pi n}\sin(n\pi x) \qquad \text{in } L^2(0,1)\,,$$

and set $t = 0$,

$$\begin{aligned}
u(0,x) =&\tfrac{1}{2}\big(\cos(2k\pi T) - 1\big)\sin(k\pi x) \\
&+ \sum_{\substack{n=1\\ n\neq k}}^{\infty} (-1)^{n+k} \frac{2k}{n^2-k^2}\Big(n\cos(k\pi T)\cos(n\pi T) \\
&\qquad\qquad + k\sin(k\pi T)\sin(n\pi T) - n\Big)\sin(n\pi x)\,, \\
u_t(0,x) =&k\pi\big(k\pi T + \tfrac{1}{2}\sin(2k\pi T)\big)\sin(k\pi x) \\
&+ \sum_{\substack{n=1\\ n\neq k}}^{\infty} (-1)^{n+k} \frac{2kn\pi}{n^2-k^2}\Big(n\cos(k\pi T)\sin(n\pi T) \\
&\qquad\qquad - k\sin(k\pi T)\cos(n\pi T)\Big)\sin(n\pi x)\,.
\end{aligned}$$

We now proceed to consider $\Lambda_T \begin{pmatrix} v^0 \\ \overline{v}^0 \end{pmatrix}$ where

$$v^0(x) = 0, \quad \overline{v}^0(x) = \sin(k\pi x), \quad k \in \mathbb{N}\,. \tag{A.9}$$

With calculations analogous to the ones above we arrive at

$$\begin{aligned}
u(0,x) =& \left(\frac{1}{2k\pi}\sin(2k\pi T) - T\right)\sin(k\pi x) \\
&+ \sum_{\substack{n=1\\ n\neq k}}^{\infty} (-1)^{n+k} \frac{2}{\pi(n^2-k^2)}\Big(n\sin(k\pi T)\cos(n\pi T) \\
&\qquad\qquad - k\cos(k\pi T)\sin(n\pi T)\Big)\sin(n\pi x)\,, \\
u_t(0,x) =&\tfrac{1}{2}\big(1 - \cos(2k\pi T)\big)\sin(k\pi x) \\
&+ \sum_{\substack{n=1\\ n\neq k}}^{\infty} (-1)^{n+k} \frac{2n}{n^2-k^2}\Big(n\sin(k\pi T)\sin(n\pi T) \\
&\qquad\qquad + k\cos(k\pi T)\cos(n\pi T) - k\Big)\sin(n\pi x)\,.
\end{aligned}$$

The controllability operator Λ_T thus has the appearance

$$\mathbf{\Lambda}_T^F = \left[\begin{array}{ccc|ccc}
\lambda_{11}^1 & \lambda_{12}^1 & \cdots & \lambda_{11}^3 & \lambda_{12}^3 & \cdots \\
\lambda_{21}^1 & \lambda_{22}^1 & \cdots & \lambda_{21}^3 & \lambda_{22}^3 & \cdots \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\
\hline
\lambda_{11}^2 & \lambda_{12}^2 & \cdots & \lambda_{11}^4 & \lambda_{12}^4 & \cdots \\
\lambda_{21}^2 & \lambda_{22}^2 & \cdots & \lambda_{21}^4 & \lambda_{22}^4 & \cdots \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots
\end{array}\right],$$

in matrix notation, using the basis $(\sin(1\pi x), \sin(2\pi x), \cdots \mid \sin(1\pi x), \sin(2\pi x), \cdots)$, where

$$
\begin{aligned}
\lambda^1_{kk} &= k\pi\big(k\pi T + \tfrac{1}{2}\sin(2k\pi T)\big)\,,\\
\lambda^1_{nk} &= (-1)^{n+k}\frac{2kn\pi}{n^2-k^2}\Big(n\cos(k\pi T)\sin(n\pi T) - k\sin(k\pi T)\cos(n\pi T)\Big)\,,\\
\lambda^2_{kk} &= \tfrac{1}{2}\big(1-\cos(2k\pi T)\big)\,,\\
\lambda^2_{nk} &= (-1)^{n+k+1}\frac{2k}{n^2-k^2}\Big(n\cos(k\pi T)\cos(n\pi T) + k\sin(k\pi T)\sin(n\pi T) - n\Big)\,,\\
\lambda^3_{kk} &= \tfrac{1}{2}\big(1-\cos(2k\pi T)\big)\,,\\
\lambda^3_{nk} &= (-1)^{n+k}\frac{2n}{n^2-k^2}\Big(n\sin(k\pi T)\sin(n\pi T) + k\cos(k\pi T)\cos(n\pi T) - k\Big)\,,\\
\lambda^4_{kk} &= T - \frac{1}{2k\pi}\sin(2k\pi T)\,,\\
\lambda^4_{nk} &= (-1)^{n+k+1}\frac{2}{\pi(n^2-k^2)}\Big(n\sin(k\pi T)\cos(n\pi T) - k\cos(k\pi T)\sin(n\pi T)\Big)\,,
\end{aligned}
$$

for all $n, k \in \mathbb{N}$ for which $n \neq k$. Notice the symmetry of this matrix, namely that

$$\lambda^1_{nk} = \lambda^1_{kn}\,, \quad \lambda^2_{nk} = \lambda^3_{kn}\,, \quad \lambda^4_{nk} = \lambda^4_{kn}\,,$$

for all $n, k \in \mathbb{N}$.

Detail 7

Consider first I_3. Using the expression for the $\widetilde{H}$-norm, we get

$$I_3^2 = \tfrac{1}{2} \sum_{k=\alpha(N)+1}^{\infty} \left[k^2\pi^2 \left(a_k \cos(k\pi\Delta t) + b_k \frac{\sin(k\pi\Delta t)}{k\pi} \right)^2 + \left(-a_k k\pi \sin(k\pi\Delta t) + b_k \cos(k\pi\Delta t)\right)^2 \right] = \tfrac{1}{2} \sum_{k=\alpha(N)+1}^{\infty} (k^2\pi^2 a_k^2 + b_k^2)\,.$$

We now have

$$\begin{aligned} \frac{I_3^2}{\Delta t^2} &= \frac{(N+1)^2}{2\eta^2} \sum_{k=\alpha(N)+1}^{\infty} (k^2\pi^2 a_k^2 + b_k^2) \le \frac{2CN^2}{\eta^2} \sum_{k=\alpha(N)+1}^{\infty} r^k \\ &= \frac{2CN^2}{\eta^2} \frac{r^{\alpha(N)+1}}{1-r} \le \frac{2C}{\eta^2(1-r)} N^2 r^{2N^{1/3}} \to 0\,, \end{aligned} \tag{A.10}$$

as $N \to \infty$.

Consider now I_2. We get

$$\begin{aligned} I_2^2 &= \tfrac{1}{2} \sum_{k=\alpha(N)+1}^{N} \left[k^2\pi^2 \left(a_k \cos(k\pi\Delta t) + b_k \frac{\sin(k\pi\Delta t)}{\mu_k} \right)^2 + \left(-a_k \mu_k \sin(k\pi\Delta t) + b_k \cos(k\pi\Delta t)\right)^2 \right] \\ &= \tfrac{1}{2} \sum_{k=\alpha(N)+1}^{N} \left[\left(k^2\pi^2 \cos^2(\theta_k) + \mu_k^2 \sin^2(\theta_k)\right) a_k^2 + \left(\frac{k^2\pi^2}{\mu_k^2} \sin^2(\theta_k) + \cos^2(\theta_k) \right) b_k^2 + 2k\pi a_k b_k \cos(\theta_k)\sin(\theta_k) \left(\frac{k\pi}{\mu_k} - \frac{\mu_k}{k\pi} \right) \right] \\ &\le C'(N+1)^2 \sum_{k=\alpha(N)+1}^{N} (k^2\pi^2 a_k^2 + b_k^2)\,, \end{aligned}$$

for some constant $C' > 0$, that only depends on η, using the bounds,

$$\begin{aligned} \frac{k\pi \sin(\theta_k)}{\mu_k} &= \frac{k\pi\eta h}{1 + \eta^2 \tan^2(\frac{1}{2}k\pi h)} \le hk\pi\eta \le \pi\eta, \\ \mu_k \sin(\theta_k) &= \frac{4\eta \tan^2(\frac{1}{2}k\pi h))}{h(1 + \eta^2 \tan^2(\frac{1}{2}k\pi h))} \le \frac{4}{\eta h} = \frac{4}{\eta}(N+1)\,. \end{aligned}$$

Proceeding as in (A.10), we get that $I_2^2/\Delta t^2 \to 0$ as $N \to \infty$.

Detail 8

We have the norm

$$\|V\|_{\widetilde{H}}^2 = \|\beta\psi^0 - v^1\|_{L^2(0,1)}^2 + \|v^0\|_{H_0^1(0,1)}^2 + \|\psi^0\|_{L^2(0,1)}^2$$

for $V = (v^0, v^1, \psi^0)^T$ and energy

$$E(t) = \frac{1}{2}\int_0^1 \left(|\beta\psi(t,x) - v_t(t,x)|^2 + c^2|v_x(t,x)|^2 + \frac{c^2\beta}{\alpha}|\psi(t,x)|^2\right) dx\,,$$

for $0 \le t \le T$. By differentiating the energy

$$E'(t) = \frac{c^2\beta\nu}{\alpha}\int_0^1 |\psi_x(t,x)|^2 dx \ge 0\,,$$

we see that $0 \le E(t) \le E(T)$ for $0 \le t \le T$. We wish to show that the inequality

$$\int_0^T |v_x(t,1)|^2 dt \le K\|(v^0, v^1, \psi^0)\|_{\widetilde{H}}^2\,,$$

holds for any solution $(v(t), v_t(t), \psi(t))$ of the adjoint system (8.3) with initial conditions $(v^0, v^1, \psi^0) \in \widetilde{H}$

First we note that the norm $\|\cdot\|_{\widetilde{H}}$ and the energy $E(t)$ are equivalent in the sense that

$$\tfrac{1}{2}\min\left\{1, c^2, \tfrac{c^2\beta}{\alpha}\right\}\left\|(v^0, v^1, \psi^0)\right\|_{\widetilde{H}}^2 \le E(t) \le \tfrac{1}{2}\max\left\{1, c^2, \tfrac{c^2\beta}{\alpha}\right\}\left\|(v^0, v^1, \psi^0)\right\|_{\widetilde{H}}^2 .$$

Assume that we are given initial data $(v^0, v^1, \psi^0) \in \widetilde{H}$. We first need a number of useful bounds. In the following, when we write f for some function f, it is short for $f(t,x)$.

$$\int_0^1 v_x^2 dx \le \frac{2}{c^2}E(t) \le \frac{2}{c^2}E(T)\,,$$

$$\int_0^T\int_0^1 v_x^2 dxdt \le \frac{2}{c^2}TE(T)\,,$$

$$\int_0^1 \left|v_t - \beta\psi\right|^2 dx \le 2E(t) \le 2E(T)\,,$$

$$\int_0^T\int_0^1 \psi_x^2 dxdt = \frac{\alpha}{c^2\beta\nu}\int_0^T E'(t)dt = \frac{\alpha}{c^2\beta\nu}\big(E(T) - E(0)\big) \le \frac{\alpha}{c^2\beta\nu}E(T)\,,$$

$$\int_0^T\int_0^1 \psi^2 dxdt = \frac{2\alpha}{c^2\beta}\int_0^T E(t)dt \le \frac{2\alpha}{c^2\beta}TE(T)\,,$$

$$
\begin{aligned}
\left(\int_0^1 v_t^2 dx\right)^{1/2} &= \big\|v_t(t,\cdot)\big\|_{L^2(0,1)} \\
&\le \big\|v_t(t,\cdot) - \beta\psi(t,\cdot)\big\|_{L^2(0,1)} + \beta\big\|\psi(t,\cdot)\big\|_{L^2(0,1)} \\
&= \left(\int_0^1 \big|v_t - \beta\psi\big|^2 dx\right)^{1/2} + \beta\left(\int_0^1 \psi^2 dx\right)^{1/2} \\
&\le \big(2E(t)\big)^{1/2} + \beta\left(\frac{2\alpha}{c^2\beta}E(t)\right)^{1/2} \le \sqrt{2}\left(1 + \frac{\sqrt{\alpha\beta}}{c}\right)\sqrt{E(T)}\,, \\
&\int_0^T\int_0^1 v_t^2 dxdt \le 2\left(1 + \frac{\sqrt{\alpha\beta}}{c}\right)^2 TE(T)\,.
\end{aligned}
$$

The two equations of the adjoint system can be combined into

$$v_{tt} - \beta\psi_t = c^2 v_{xx}\,.$$

We first apply the multiplier $v_x x$ to the right-hand side,

$$
\begin{aligned}
\int_0^1 v_{xx}v_x x dx &= \big[v_x^2 x\big]_{x=0}^1 - \int_0^1 v_x\big(v_{xx}x + v_x\big)dx \\
&= v_x^2(t,1) - \int_0^1 v_{xx}v_x x dx - \int_0^1 v_x^2 dx\,,
\end{aligned}
$$

leading to

$$\int_0^T v_x^2(t,1)dt = 2\int_0^T\int_0^1 v_{xx}v_x x dxdt - \int_0^T\int_0^1 v_x^2 dxdt\,,$$

and then

$$\int_0^T v_x^2(t,1)dt \le \frac{2}{c^2}\left|\int_0^T\int_0^1 \big(v_{tt} - \beta\psi_t\big)v_x x dxdt\right| + \frac{2}{c^2}TE(T)\,. \tag{A.11}$$

We turn to bounding the double integral. First we see that

$$\int_0^T \big(v_{tt} - \beta\psi_t\big)v_x x dt = \big[\big(v_t - \beta\psi\big)v_x x\big]_{t=0}^T - \int_0^T \big(v_t - \beta\psi\big)v_{xt}x dt\,. \tag{A.12}$$

Since, by the Cauchy-Schwartz inequality,

$$
\begin{aligned}
\left|\int_0^1 \big(v_t - \beta\psi\big)v_x x dx\right| &\le \left(\int_0^1 \big|v_t - \beta\psi\big|^2 dx\right)^{1/2}\left(\int_0^1 v_x^2 dx\right)^{1/2} \\
&\le \sqrt{2E(T)}\sqrt{\frac{2}{c^2}E(T)} = \frac{2}{c}E(T)\,,
\end{aligned}
$$

for every $t \in [0,T]$, we have

$$\left|\int_0^1 \big[\big(v_t - \beta\psi\big)v_x x\big]_{t=0}^T dx\right| \le \frac{4}{c^2}E(T)\,.$$

For the last integral of (A.12) we get

$$\left|\int_0^1\int_0^T (v_t-\beta\psi)v_{xt}xdtdx\right| \le \left|\int_0^1\int_0^T v_tv_{xt}xdtdx\right| + \beta\left|\int_0^1\int_0^T \psi v_{xt}xdtdx\right| .$$

We bound each integral of the right-hand side in turn. Since $(v_t^2x)_x = 2v_tv_{tx}x + v_t^2$ we have

$$2\int_0^1 v_tv_{xt}xdx = \int_0^1 (v_t^2x)_xdx - \int_0^1 v_t^2dx = v_t^2(t,1) - \int_0^1 v_t^2dx ,$$

leading to

$$\left|\int_0^T\int_0^1 v_tv_{xt}xdxdt\right| = \tfrac{1}{2}\left|\int_0^T\int_0^1 v_t^2dxdt\right| \le \left(1+\frac{\sqrt{\alpha\beta}}{c}\right)^2 TE(T) .$$

Next we get

$$\begin{aligned}
\left|\int_0^T\int_0^1 v_{xt}\psi xdxdt\right| &= \left|\int_0^T \left(\left[v_t\psi x\right]_{x=0}^1 - \int_0^1 v_t(\psi_xx+\psi)dx\right)dt\right| \\
&\le \left|\int_0^T\int_0^1 v_t\psi_xxdxdt\right| + \left|\int_0^T\int_0^1 v_t\psi dx\right| \\
&\le \left(\int_0^T\int_0^1 v_t^2dxdt\right)^{1/2}\left(\int_0^T\int_0^1 \psi_x^2dxdt\right)^{1/2} \\
&\quad + \left(\int_0^T\int_0^1 v_t^2dxdt\right)^{1/2}\left(\int_0^T\int_0^1 \psi^2dxdt\right)^{1/2} \\
&\le \sqrt{2T}\left(1+\frac{\sqrt{\alpha\beta}}{c}\right)\left(\frac{\sqrt{\alpha}}{c\sqrt{\beta\nu}} + \frac{\sqrt{2\alpha T}}{c\sqrt{\beta}}\right)E(T) .
\end{aligned}$$

We can finally collect all the bounds and turn (A.11) into

$$\begin{aligned}
\int_0^T v_x^2(t,1)dt &\le \frac{2}{c^2}\left[T + \frac{4}{c^2} + T\left(1+\frac{\sqrt{\alpha\beta}}{c}\right)^2\right. \\
&\qquad\left. + \frac{\sqrt{2\alpha\beta T}}{c}\left(1+\frac{\sqrt{\alpha\beta}}{c}\right)\left(\frac{1}{\sqrt{\nu}}+\sqrt{2T}\right)\right]E(T) \\
&\le \max\left\{\frac{1}{c^2}, 1, \frac{\beta}{\alpha}\right\}\left[\frac{\sqrt{2\alpha\beta T}}{c}\left(1+\frac{\sqrt{\alpha\beta}}{c}\right)\left(\frac{1}{\sqrt{\nu}}+\sqrt{2T}\right)\right. \\
&\qquad\left. + \frac{4}{c^2} + T + T\left(1+\frac{\sqrt{\alpha\beta}}{c}\right)^2\right]\left\|(v^0,v^1,\psi^0)\right\|_{\widetilde{H}}^2 ,
\end{aligned}$$

which is the type of bound we wanted to show.

Detail 9

Compact Sequence Embedding

Let ℓ^2 be the set of sequences $\langle a_k\rangle_{k=1}^{\infty}$ for which $\|\langle a_k\rangle\|_2 < \infty$, where

$$\|\langle a_k\rangle\|_2^2 = \langle\langle a_k\rangle, \langle a_k\rangle\rangle, \quad \langle\langle a_k\rangle, \langle b_k\rangle\rangle = \sum_{k=1}^{\infty} a_k \overline{b_k}\,.$$

Let similarly $\ell^{2,p}$, for $p \in \mathbb{R}$, be the set of sequences $\langle a_k\rangle_{k=1}^{\infty}$ for which $\|\langle a_k\rangle\|_{2,p} < \infty$, where

$$\|\langle a_k\rangle\|_{2,p}^2 = \|\langle a_k/k^p\rangle\|_2^2 = \sum_{k=1}^{\infty} \left|\frac{a_k}{k^p}\right|^2 .$$

Theorem. *The map* $T : \ell^2 \mapsto \ell^{2,p}$ *with* $p > 0$*, defined as* $T\big(\langle a_k\rangle\big) = \langle a_k\rangle$*, is compact.*

Proof. Consider a sequence of ℓ^2 sequences $\langle a_k^n\rangle_{k=1}^{\infty}$ that converges weakly to the sequence $\langle a_k\rangle_{k=1}^{\infty}$ in ℓ^2. This means that

$$\langle\langle a_k^n\rangle, \langle b_k\rangle\rangle \to \langle\langle a_k\rangle, \langle b_k\rangle\rangle \quad \text{for } n \to \infty\,,$$

for all sequences $\langle b_k\rangle_{k=1}^{\infty} \in \ell^2$. We will now show that this sequence converges strongly in $\ell^{2,p}$. Consider

$$\begin{aligned}\big\|\langle a_k^n\rangle - \langle a_k\rangle\big\|_{2,p}^2 &= \sum_{k=1}^{\infty} \frac{|a_k^n - a_k|^2}{k^{2p}} = \sum_{k=1}^{N_0} \frac{|a_k^n - a_k|^2}{k^{2p}} + \sum_{k=N_0+1}^{\infty} \frac{|a_k^n - a_k|^2}{k^{2p}} \\ &\le \max_{1\le k\le N_0} |a_k^n - a_k|^2 \sum_{k=1}^{N_0} \frac{1}{k^{2p}} + \frac{2}{(N_0+1)^{2p}} \left(\sum_{k=1}^{\infty} |a_k^n|^2 + \sum_{k=1}^{\infty} |a_k|^2\right) \\ &= I'_{N_0} + I''_{N_0}\,,\end{aligned}$$

where $N_0 \in \mathbb{N}$ is some constant. Consider now I''_{N_0}. The terms in the parentheses are bounded since the ℓ^2-norm of $\langle a_k\rangle$ is bounded by assumption, and $\langle a_k^n\rangle$ is uniformly bounded in ℓ^2 since

$$\|\langle a_k^n\rangle\|_2 = \sup_{\langle b_k\rangle\in\ell^2} \frac{|\langle\langle a_k^n\rangle, \langle b_k\rangle\rangle|}{\|\langle b_k\rangle\|_2} \to \sup_{\langle b_k\rangle\in\ell^2} \frac{|\langle\langle a_k\rangle, \langle b_k\rangle\rangle|}{\|\langle b_k\rangle\|_2} = \|\langle a_k\rangle\|_2\,.$$

We now choose N_0 such that $I''_{N_0} \le \epsilon/2$. Note that this choice can be made independently of n. Since I'_{N_0} consists of finitely many terms, its size can be made $I'_{N_0} \le \epsilon/2$ by choosing n large enough. □

APPENDIX B

Notation

Although we agree with the importance of the distinction,
we shall not adopt these terms.
— RENARDY AND ROGERS (1993)

We write matrices using boldface upper case letters, e.g., $\boldsymbol{A}$. Vectors are written with boldface lower case letters. A vector $\boldsymbol{v} \in \mathbb{R}^n$ should be regarded as a single-column matrix, such that $\boldsymbol{A}\boldsymbol{v}$ and $\boldsymbol{v}^T\boldsymbol{A}$ make sense when $\boldsymbol{A}$ is an appropriately sized matrix.

References to vector elements are done using parentheses, e.g., for a vector $\boldsymbol{v}$ we write $\boldsymbol{v}(i)$ to refer to the i'th element of $\boldsymbol{v}$. Similar for matrices, e.g., $\boldsymbol{A}(i,j)$.

Differentiation of a function f of one variable is written f'. Partial derivatives are written using subscripts, e.g., $u_{xx} = \partial^2 u/\partial x^2$ and $u_{tx} = \partial^2 u/\partial t \partial x$. We use dots as short-hand for differentiation with respect to time, e.g., $\dot{u} = u_t$ and $\ddot{v} = v_{tt}$. Normal derivatives are written as $\partial/\partial n$. If, e.g., $f \in H_0^1(\Omega)$ and $x \in \partial\Omega$, then $\partial f(x)/\partial n = \nabla f(x) \cdot n$, where n is a unit vector, perpendicular to $\partial\Omega$ and pointing outwards.

All Hilbert spaces are assumed to be real and separable, unless explicitly noted otherwise.

Sequences will be written using angle-brackets. For instance, we will write $\langle a_k \rangle_{k=1}^{\infty}$ as short-hand for the sequence $a_1, a_2, \ldots$ (this convention was adopted from *The Art of Computer Programming* by Donald E. Knuth).

The following two pages contain an overview of the most common notation used in this thesis.

Symbolism	Meaning	Defined
$\sum'$	Summation where the first and last terms are weighed with 1/2, the rest are weighed with 1 as normal.	Page 72
$\square'$	For a Hilbert space H, H' is the dual space, the Hilbert space of functionals $H \mapsto \mathbb{C}$.	Page 10
$\square^*$	Adjoint operator. For a linear and bounded operator $F : S_1 \mapsto S_2$, where S_1 and S_2 are Hilbert spaces, we have $F^* : S_2' \mapsto S_1'$ is the linear and bounded operator for which $\langle x, F^* y\rangle_{S_1 \times S_1'} = \langle Fx, y\rangle_{S_2 \times S_2'}$ for all $(x, y) \in S_1 \times S_2'$.	Page 10
$\square^T$	For a real matrix $\boldsymbol{X}$ we have $\boldsymbol{X}^T(i, j) = \boldsymbol{X}(j, i)$.	Page 11
$\overline{\square}$	For a set S, $\overline{S}$ is the closure of S in an appropriate topology.	Page 15
	For a complex number z, $\overline{z}$ is the complex conjugate.	Page 44
$\langle\square, \square\rangle$	For vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ we have $\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \boldsymbol{u}^T \boldsymbol{v}$.	Page 34
$\langle\square, \square\rangle_{\boldsymbol{C}}$	For vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ we have $\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\boldsymbol{C}} = \boldsymbol{u}^T \boldsymbol{C}\boldsymbol{v}$.	Eq. (3.3)
$\langle\square, \square\rangle$	For $f, g \in L^2(\Omega)$ we have $\langle f, g\rangle = \int_\Omega f(x)\overline{g(x)}dx$.	
$\langle\square, \square\rangle_{H' \times H}$	For $(f, g) \in H' \times H$ we have $\langle f, g\rangle_{H' \times H} = f(g)$.	
$\square \mid \square$	$a \mid b \Leftrightarrow a$ divides b.	Eq. (3.8)
$\square \nmid \square$	$a \nmid b \Leftrightarrow a$ does not divide b.	Page 37
$\partial\square$	$\partial\Omega$ denotes the boundary of Ω.	Sec. 2.1
δ_{ij}	Kronecker delta, δ_{ij} is equal to 1 if $i = j$ and 0 otherwise.	Page 33
Δ	The Laplace operator, the sum of second derivatives in each *space* direction.	Eq. (2.26)
Δt	Step size in the time direction.	Page 40
Γ	Boundary of Ω, $\Gamma = \partial\Omega$.	Sec. 2.1
Γ_0	Control boundary, $\Gamma_0 \subset \Gamma$.	Sec. 2.1
Λ_T	Controllability operator.	Eq. (2.23)
Ω	Open and bounded subset of $\mathbb{R}^d$, $d \in \mathbb{N}$.	Sec. 2.1
Σ	Time–boundary cylinder, $\Sigma = (0, T) \times \Gamma$.	Sec. 2.1
Σ_0	Time–control boundary cylinder, $\Sigma_0 = (0, T) \times \Gamma_0$.	Sec. 2.1
$\boldsymbol{A}$, $\boldsymbol{C}$	$\boldsymbol{C}^{-1}\boldsymbol{A}$ approximates the Laplacian, Δ.	Sec. 3.1
$\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{C}}$	Used for first order ODEs, e.g., $\boldsymbol{\mathcal{C}}\dot{\boldsymbol{v}}(t) = \boldsymbol{\mathcal{A}}\boldsymbol{v}(t)$.	Eq. (3.11)
$\boldsymbol{\mathcal{B}}$	Discrete control operator, e.g., $\boldsymbol{\mathcal{C}}\dot{\boldsymbol{v}}(t) = \boldsymbol{\mathcal{A}}\boldsymbol{v}(t) + \boldsymbol{\mathcal{B}}\boldsymbol{k}(t)$.	Eq. (4.1)

Symbolism	**Meaning**	**Defined**
$C(\square;\square)$	$C(S_1;S_2)$ is the space of continuous functions from S_1 to S_2.	
$\mathbb{C}$	The complex numbers.	
$\mathrm{diag}(\square)$	$\mathrm{diag}(x_1,\ldots,x_n)$ is an $n \times n$ diagonal matrix with entries $x_1, x_2, \ldots, x_n$ along the diagonal.	Eq. (3.16)
h	Grid size for uniformly spaced grids; in 1D we usually have $h = 1/(N+1)$.	Page 32
H	Dual space of H'.	Sec. 2.1.1
H'	Hilbert space in which solutions of a control system is well posed.	Sec. 2.1
$\widetilde{H}$	Hilbert space in which solutions of an adjoint system is well posed.	Sec. 2.1.1
$\widetilde{H}'$	The dual space of $\widetilde{H}$.	Sec. 2.1.1
$H_0^1(\square)$	$f \in H_0^1(\Omega)$ if and only if $f, f_{x_1}, \ldots, f_{x_d} \in L^2(\Omega)$ and $f(\partial\Omega) = 0$.	
$H^{-1}(\square)$	$H^{-1}(\Omega)$ is the dual space of $H_0^1(\Omega)$.	
$\mathrm{Im}\, z$	Imaginary part of the complex number z.	Page 146
$\ker \square$	For a linear operator $F : S_1 \to S_2$ we have $\ker F = \{x \in S_1 \mid F(x) = 0\}$.	
ℓ^2	Space of square summable sequences; $\langle a_k \rangle_{k=1}^{\infty} \in \ell^2$ if and only if $\sum_{k=1}^{\infty} \|a_k\|^2 < \infty$.	Page 75
$L^2(\square)$	For a set S we have $f \in L^2(S)$ if and only if $\int_S \|f(x)\|^2 dx < \infty$.	
N	Dimension of space used for space discretization.	Page 32
$\mathbb{N}$	The natural numbers, $1, 2, \ldots$.	
$\mathcal{O}(\square)$	Big-oh; $f(h) = \mathcal{O}(g(h))$ means that $\|f(h)\| \leq C\|g(h)\|$ for some $C > 0$ and $\|h\|$ sufficiently small.	
Q	Time-space cylinder, $Q = (0,T) \times \Omega$.	Sec. 2.1
$\mathbb{R}$	The real numbers.	
$\mathrm{rank}\, \square$	Matrix rank; rank $\boldsymbol{A}$ is the maximal number of linear independent rows or columns of the matrix $\boldsymbol{A}$.	
$\mathrm{Re}\, z$	Real part of the complex number z.	Page 40
T	Time available for control.	Sec. 2.1
$\mathbb{Z}$	The integers, $\ldots, -2, -1, 0, 1, 2, \ldots$.	

Bibliography

You will find it a very good practice
always to verify your references sir.
— Martin Routh

Mark Asch and Gilles Lebeau. Geometrical aspects of exact boundary controllability for the wave equation—a numerical study. *SIAM Journal of Control, Optimisation and Calculus of Variations*, 3:163–212, May 1998.

Claude Bardos, Gilles Lebeau, and Jeffrey Rauch. Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary. *SIAM J. Control and Optimization*, 30(5):1024–1065, September 1992.

Alain Bensoussan. On the general theory of exact controllability for skew symmetric operators. *Acta Appl. Math.*, 20(3):197–229, 1990.

Alain Bensoussan. Exact controllability for linear dynamic systems with skew symmetric operators. In J. P. Zolésio, editor, *Boundary Control and Boundary Variation*, volume 178 of *Lecture Notes in Control and Information Sciences*, pages 27–36. Springer-Verlag, 1992. Proceedings of IFIP WG 7.2 Conference, Sophia Antipolis, France, October 15-17, 1991.

Alain Bensoussan. An introduction to the Hilbert Uniqueness Method. In *Analysis and optimization of systems: State and frequency domain approaches for infinite-dimensional systems (Sophia-Antipolis, 1992)*, volume 185 of *Lecture Notes in Control and Information Sciences*, pages 184–198. Springer, Berlin, 1993.

Craig Carthel, Roland Glowinski, and Jacques-Louis Lions. On exact and approximate boundary controllabilities for the heat equation: A numerical approach. *Journal of Optimization Theory and Applications*, 82(3):429–484, September 1994.

James W. Daniel. *The Approximate Minimization of Functionals*. Prentice-Hall, 1971.

Ahmed Eljendy. Numerical approach to the exact controllability of hyperbolic systems. In J. P. Zolésio, editor, *Boundary Control and Boundary Variation*, volume 178 of *Lecture Notes in Control and Information Sciences*, pages 202–214. Springer-Verlag, 1992. Proceedings of IFIP WG 7.2 Conference, Sophia Antipolis, France, October 15-17, 1991.

Hector O. Fattorini and David L. Russell. Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations. *Quart. Appl. Math.*, 32:45–69, 1974.

Enrique Fernández-Cara and Enrique Zuazua. On the null controllability of the one-dimensional heat equation with BV coefficients. *Computational and Applied Mathematics*, 2(1):167–190, 2002.

Roland Glowinski. Boundary controllability problems for the wave and heat equation. In J. P. Zolésio, editor, *Boundary Control and Boundary Variation*, volume 178 of *Lecture Notes in Control and Information Sciences*, pages 221–237. Springer-Verlag, 1992a. Proceedings of IFIP WG 7.2 Conference, Sophia Antipolis, France, October 15-17, 1991.

Roland Glowinski. Ensuring well-posedness by analogy; Stokes problem and boundary control for the wave equation. *J. Comput. Phys.*, 103(2):189–221, 1992b.

Roland Glowinski, W. Kinton, and Mary 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.

Roland Glowinski and Chin Hsien Li. On the numerical implementation of the Hilbert Uniqueness Method for the exact boundary controllability of the wave equation. *C. R. Adad. Sc., Paris*, 311:135–142, 1990.

Roland Glowinski, Chin Hsien Li, and Jacques-Louis 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.*, 7:1–76, 1990.

Roland Glowinski and Jacques-Louis Lions. Exact and approximate controllability for distributed parameter systems (II). *Acta Numerica*, pages 159–333, 1995.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* The John Hopkins University Press, third edition, 1996.

Bao-Zhu Guo and Runyi Yu. The Riesz basis property of discrete operators and application to a Euler-Bernoulli beam equation with boundary linear feedback control. *IMA Journal of Mathematical Control and Information*, 18:241–251, 2001.

Per 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.

Scott W. Hansen. Boundary control of a one-dimensional linear thermoelastic rod. *SIAM J. Control and Optimization*, 32(4):1052–1074, July 1994.

Peter Henrici. *Applied and computational complex analysis. Vol. 2*. Wiley Interscience [John Wiley & Sons], New York, 1977. Special functions—integral transforms—asymptotics—continued fractions.

Juan A. Infante and Enrique Zuazua. Boundary observability for the space-discretizations of the 1-D wave equation. *C. R. Acad. Sci. Paris*, 326(6): 713–718, 1998.

Juan A. Infante and Enrique Zuazua. Boundary observability for the space semi-discretizations of the 1-D wave equation. *M2AN*, 33(2):407–438, 1999.

Albert E. Ingham. Some trigonometric inequalities with applications to the theory of series. *Mathematische Zeitschrift*, 41:367–379, 1936.

Arieh Iserles. *A first course in the numerical analysis of differential equations*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 1996.

Michail I. Kadec. The exact value of the Paley-Wiener constant. *Soviet Mathematics*, 5:559–561, 1964. Original Russian paper appeared in *Dokl. Akad. Nauk SSSR*, 155:1253–1254, 1964.

Rudolf E. Kalman. Mathematical description of linear dynamical systems. *Journal SIAM on Control*, 1(2):152–192, 1963.

Stefan Kindermann. Convergence rates of the Hilbert Uniqueness Method via Tikhonov regularization. *Journal of Optimization Theory and Applications*, 103(3):657–673, December 1999.

Vilmos Komornik. *Exact Controllability and Stabilization—The Multiplier Method*. Research in Applied Mathematics. Masson, Paris, 1994.

John E. Lagnese. The Hilbert Uniqueness Method: A retrospective. In K.-H. Hoffmann and W. Krabs, editors, *Optimal control of partial differential equations (Irsee, 1900)*, volume 149 of *Lecture Notes in Control and Information Sciences*, pages 158–181. Springer, 1991.

Peter D. Lax and Robert D. Richtmyer. Survey of the stability of linear finite difference equations. *Comm. Pure Appl. Math.*, 9:267–293, 1956. Part I: An equivalence theorem.

Gilles Lebeau and Enrique Zuazua. Null-controllability of a system of linear thermoelasticity. *Arch. Rational Mech. Anal.*, 141:297–329, 1998.

Liliana León and Enrique Zuazua. Boundary controllability of the finite-difference space semi-discretization of the beam equation. *ESIAM: Control, Optimisation and Calculus of Variations*, 8:827–862, 2002. A tribute to Jacques-Louis Lions, Tome 2.

Jacques-Louis Lions. *Optimal Control of Systems Governed by Partial Differential Equations*. Springer–Verlag, 1971.

Jacques-Louis Lions. *Control of Distributed Singular Systems*. Gauthier-Villars, 1985.

Jacques-Louis Lions. *Contrôlabilité Exacte Pertubations et Stabilisation de Systèmes Distribués, Tome 1: Contrôlabilité Exacte*, volume 8 of *Recherches en Mathématiques Appliquées*. Masson, Paris, 1988a.

Jacques-Louis Lions. Exact controllability, stabilization and pertubations for distributed systems. *SIAM Review*, 30(1):1–68, March 1988b.

Walter Littman. Remarks on boundary control for polyhedral domains and related results. In J. P. Zolésio, editor, *Boundary Control and Boundary Variation*, volume 178 of *Lecture Notes in Control and Information Sciences*, pages 272–284. Springer-Verlag, 1992. Proceedings of IFIP WG 7.2 Conference, Sophia Antipolis, France, October 15-17, 1991.

Antonio López and Enrique Zuazua. Some new results related to the null controllability of the 1-d heat equation. *Seminario EDP de la Escuela Politécnica de Paris*, VIII:1–22, 1998.

Antonio López and Enrique Zuazua. Uniform null-controllability for the one-dimensional heat equation with rapidly oscillating periodic density. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 19(5):543–580, 2002.

Fabricio Maciá. The effect of group velocity in the numerical analysis of control problems for the wave equation. In G. C. Cohen, E. Heikkola, P. Joly, and P. Neittaanmäki, editors, *Mathematical and Numerical Aspects of Wave Propagation, WAVES 2003*, pages 195–200, 2003.

Sorin Micu. Uniform boundary controllability of a semi-discrete 1-D wave equation. *Numer. Math.*, 91(4):723–768, 2002.

Sorin Micu and Enrique Zuazua. An introduction to the controllability of partial differential equations. Preprint, see `http://www.uam.es/enrique.zuazua`, 2004.

Mihaela Negreanu and Enrique Zuazua. Uniform boundary controllability of a discrete 1-D wave equation. *Systems and Control Letters*, 48(3–4):261–280, 2003.

Mihaela Negreanu and Enrique Zuazua. Convergence of a multigrid method for the controllability of a 1-d wave equation. *C. R. Acad. Sci. Paris*, 338(5): 413–418, 2004a.

Mihaela Negreanu and Enrique Zuazua. Discrete Ingham inequalities and applications. *C. R. Acad. Sci. Paris*, 338(4), 2004b.

Hung M. Park and Won J. Lee. A new numerical method for the boundary control problems of the heat conduction equation. *Int. J. Numer. Meth. Engng.*, 53: 1593–1613, 2002.

Michael Pedersen. *Functional Analysis in Applied Mathematics and Engineering.* Studies in Advanced Mathematics. Chapman & Hall, 2000.

Jan M. Rasmussen. A study of numerical methods for boundary control of the wave equation in 1D. In G. C. Cohen, E. Heikkola, P. Joly, and P. Neittaanmäki, editors, *Mathematical and Numerical Aspects of Wave Propagation, WAVES 2003*, pages 207–212, 2003.

Michael Renardy and Robert C. Rogers. *An introduction to partial differential equations*, volume 13 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 1993.

Robert D. Richtmyer and K. W. Morton. *Difference methods for initial-value problems.* Second edition. Interscience Tracts in Pure and Applied Mathematics, No. 4. Interscience Publishers John Wiley & Sons, Inc., New York-London-Sydney, 1967.

Walter Rudin. *Functional Analysis.* McGraw-Hill, 1973.

David L. Russell. A unified boundary controllability theory for hyperbolic and parabolic partial differential equations. *Studies in Appl. Math.*, 52:189–211, 1973.

David L. Russell. Controllability and stabilization theory for linear partial differential equations: Recent progress and open questions. *SIAM Review*, 20(4): 639–737, 1978.

Eduardo D. Sontag. *Mathematical control theory*, volume 6 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 1990. Deterministic finite-dimensional systems.

Walter A. Strauss. *Partial Differential Equations, An Introduction.* John Wiley & Sons, Inc., 1992.

Lloyd N. Trefethen. Group velocity in finite difference schemes. *SIAM Review*, 24 (2):113–136, April 1982.

Lloyd N. Trefethen. Finite difference and spectral methods for ordinary and partial differential equations. Unpublished text, available at http://web.comlab.ox.ac.uk/oucl/work/nick.trefethen/pdetext.html, 1996.

Lloyd N. Trefethen. *Spectral methods in MATLAB.* SIAM, 2000.

Robert Vichnevetsky and John 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.

Robert M. Young. *An introduction to nonharmonic Fourier series.* Academic Press Inc., first edition, 2001.

Xu Zhang and Enrique Zuazua. Controllability of nonlinear partial differential equations. In A. Astolfi et al., editor, *Proceedings of the Second IFAC Workshop on Lagrangian and Hamiltonian methods in Nonlinear Control*, pages 269–274, Sevilla, 2003a.

Xu Zhang and Enrique Zuazua. Unique continuation for the linearized Benjamin–Bona–Mahony equation with space-dependent potential. *Mathematische Annalen*, 325:543–582, 2003b.

Enrique Zuazua. Controllability of the linear system of thermoelasticity. *J. Math. Pures Appl.*, 74:291–315, 1995.

Enrique Zuazua. Finite-dimensional null controllability for the semilinear heat equation. *J. Math. Pures Appl. (9)*, 76(3):237–264, 1997.

Enrique Zuazua. Boundary observability for the finite-difference space semidiscretization of the 2-d wave equastion in the square. *J. Math. Pures Appl.*, 78:523–563, 1999.

Enrique Zuazua. Departamento de Matemáticas, Universidad Autónoma de Madrid, see http://www.uam.es/enrique.zuazua. Private communication, 2002a.

Enrique Zuazua. Controllability of partial differential equations and its semidiscrete approximations. *Discrete and Continuous Dynamical Systems*, 8(2): 469–513, April 2002b.

Enrique Zuazua. Propagation, observation, control and numerical approximation of waves. Universidad Autónoma de Madrid. Note obtainable from http://www.uam.es/enrique.zuazua, July 2003.

Enrique Zuazua. Optimal and approximate control of finite-difference approximation schemes for the 1-D wave equation. Preprint, see http://www.uam.es/enrique.zuazua, 2004.

Index