A Gautschi Time-stepping Approach to Optimal Control of the Wave Equation[☆]

Karl Kunisch[a], Stefan Reiterer[a]

[a]*Institute for Mathematics and Scientific Computing, University of Graz, Heinrichstrasse 36, A-8010 Graz, Austria*

Key Words— Gautschi time-stepping, optimal control, wave equation, cosine operators, finite element and

1. Introduction

This work is devoted to developing a Gautschi time stepping approach for optimal control problems associated with second order equations, including in particular the wave equation. Solving optimal control problems numerically necessitates to frequently solve the state equation and its adjoint, and hence an efficient method for the latter is indispensable. Compared to optimal control of diffusion systems, the numerical treatment of optimal control of the wave equation has received relatively little attention so far. We refer to [10, 9] where the dual weighted residual method for space-time discretization was developed, including as particular case the Crank-Nicolson discretization in time and first order finite element discretization in space. This approach has the desirable property that first discretizing the infinite dimensional optimal control problem and subsequently solving the necessary optimality conditions commutes with first setting up the necessary optimality conditions for the infinite dimensional problem and subsequently discretizing them.

In the present work the focus is put on using a Gautschi scheme for temporal discretization. It will be combined with different spatial discretizations including finite element and spectral techniques. Gautschi integrators have received a considerable amount of attention due to their desirable property that their step sizes are not restricted by the spectral properties of the underlying dynamical system. This is of particular interest for systems which allow highly oscillatory solutions. Gautschi type methods are constructed on the basis that they integrate linear systems with constant inhomogeneities exactly. We refer to [5, 4, 6, 8] and the references given there for further properties of Gautschi techniques. For Gautschi-methods, we can also show that discretizing before or after carrying out the optimization step, we obtain the same finite dimensional systems, for the class of spatial discretizations which we shall consider.

The paper is organized as follows. Section 2 contains the problem statement, first order optimality conditions and a brief recollection of cosine operators. The Gautschi time-stepping scheme in an infinite dimensional setting is presented

[☆]Both authors were supported in part by the Austrian Science Fund (FWF) under grant SFB F32 (SFB "Mathematical Optimization and Applications in Biomedical Sciences")

in Section 3. Section 4 contains it analysis for the optimal control problem, with emphasis on the inexact conjugate gradient method for its numerical realization. Numerical results, highlighting convergence rates and comparisons between different spatial discretizations are given in Section 5.

spectral methods.

2. Problem Formulation and Preliminaries

2.1. Problem Formulation

Let $V \subset H \subset V'$ be a Gelfand triple of real separable Hilbert spaces and let $T > 0$. Further let $A : V \to V'$ be a V elliptic operator, and consider for vectors $y_0 \in V$, $y_1 \in H$ and $f \in L_2(0,T;H)$ the abstract wave equation

$$\begin{aligned} \frac{\partial^2 y}{\partial t^2}(t) &= Ay(t) + f(t) \text{ for } t \in (0,T), \\ y(0) &= y_0, \\ \frac{\partial y}{\partial t} &= y_1. \end{aligned} \tag{W}$$

Definition 2.1 (Weak solution)**.** We say that $y \in L_2(0,T;V)$ is a *weak solution* of (W) iff $y_t \in L_2(0,T;H)$, $y_{tt} \in L_2(0,T;V')$,

$$\left\langle \frac{\partial^2 y}{\partial t^2}(t), \varphi \right\rangle_{V,V'} = \langle Ay(t), \varphi \rangle_{V,V'} + \langle f(t), \varphi \rangle_{V,V'} \text{ for all } \varphi \in V, \text{ and } t \in (0,T),$$

and $y(0) = y_0$, $y_t(0) = y_1$.

Existence and uniqueness of a weak solution to (W) are well understood (see e.g. [19][Chapter 29, p.436]). The solution operator $S_W : L_2(0,T;H) \times V \times H \to L_2(0,T;H)$ of the wave equation, which maps (f, y_0, y_1) to the solution y of (W), is continuous (see [19, p.437]).

For $\beta \in \mathcal{L}(L_2(0,T;H))$, $\tilde{z} \in (L^2(0,T);H)$ and $\alpha > 0$ we consider the optimal control problem

$$\begin{aligned} \min_{y,u \in L_2((0,T);H)} & \frac{1}{2} \|y - z\|^2_{L_2((0,T);H)} + \frac{\alpha}{2} \|u\|^2_{L_2((0,T);H)}, \\ \text{s. t. } & \frac{\partial^2 y}{\partial t^2} = Ay + \beta u, \\ & y(0,x) = y_0(x), \frac{\partial y}{\partial t} = y_1(x). \end{aligned} \tag{OC}$$

Define the solution operator $S : L_2(0,T;H) \to L_2(0,T;H)$ associated to the wave equation by $Su := S_W(\beta u, y_0, y_1) = y$, with y is solution to (W). We arrive at the reduced problem

$$\min_{u \in L_2(0,T;H)} \frac{1}{2} \|Su - \tilde{z}\|^2_{L_2(0,T;H)} + \frac{\alpha}{2} \|u\|^2_{L_2(0,T;H)}. \tag{1}$$

It is well known that (1) has a unique solution, see e.g. [18, p.40]. From now on we may assume without loss of generality that $y_0, y_1 = 0$, since we can express y by $y = Su = S_W(\beta u, 0, 0) + S_W(0, y_0, y_1) =: y_I + y_H$. Hence

$y - \tilde{z} = y_I - (\tilde{z} - y_H)$. Now we can replace $\tilde{z}$ in the original cost-functional by $z = \tilde{z} - y_H$ and simultaneously replace $S : L^2((0,T);H) \to L^2((0,T);H)$. by $S_W(\beta u, 0, 0)$ arriving at

$$J(u) := \frac{1}{2} \|Su - z\|^2_{L_2((0,T);H)} + \frac{\alpha}{2} \|u\|^2_{L_2((0,T);H)} . \tag{2}$$

The Gateaux derivative $J'(u)$ is given by

$$J'(u) = (S^*S + \alpha I)u - S^*z,$$

where I is the identity operator. Thus the first order necessary and sufficient optimality condition is given by the operator equation

$$\mathcal{H}u := (S^*S + \alpha I)u = S^*z. \tag{3}$$

Solving it efficiently will be in the focus of the further considerations.

2.2. Sine and Cosine Operators and the Second Order Abstract Cauchy Problem

In this subsection we briefly recall the notion of cosine and sine operators, see e.g. [1], which allow to express the solution to (W), in a manner analogous to continuous semigroups giving solutions to first order Cauchy problems. Throughout this subsection all integrals are Bochner integrals.

Definition 2.2 (Cosine functions)**.** A strongly continuous function $\mathrm{Cos} : \mathbb{R}_+ \to \mathcal{L}(H)$ is called *cosine function* if

$$\mathrm{Cos}(0) = I$$

and

$$2\,\mathrm{Cos}(t)\,\mathrm{Cos}(s) = \mathrm{Cos}(t+s) + \mathrm{Cos}(t-s) \text{ for all } t \geq s \geq 0.$$

Like the well known operator semigroups, cosine operators also have infinitesimal generators.

Proposition 2.3 (Characterization of cosine functions with Laplace transforms)**.** *Let* $\mathrm{Cos} : \mathbb{R}_+ \to \mathcal{L}(H)$ *be strongly continuous, and set* $\mathrm{abs(Cos)} := \inf\{\mathrm{Re}\,\lambda : \int_0^\infty e^{-\lambda t}\,\mathrm{Cos}(t)dt\}$ *(the abscissa of convergence). Then the following assertions are equivalent:*

(i) Cos *is a cosine function.*

(ii) One has $\mathrm{abs(Cos)} < \infty$*, and there exists* $\omega > \max\{\mathrm{abs(Cos)}, 0\}$ *and a linear operator* $A : D(A) \to H$ *such that* $(\omega^2, \infty) \subset \rho(A)$ *and*

$$\lambda R(\lambda^2, A) = \int_0^\infty e^{-\lambda t}\,\mathrm{Cos}(t)dt \text{ for } \lambda > \omega. \tag{4}$$

Proof. See [1, p.208]. □

Definition 2.4. Let Cos be a cosine function. A linear operator $A : D(A) \to H$ is called *infinitesimal generator* of the cosine function Cos, iff relation (4) is satisfied.

It is natural to extend Cos to the real line by the setting $\mathrm{Cos}(t) := \mathrm{Cos}(-t)$ for $t < 0$. We next introduce the sine operator as an integral of the cosine function.

Definition 2.5 (Sine functions)**.** Let $\mathrm{Cos} : \mathbb{R} \to \mathcal{L}(H)$ be a cosine function. The *sine function* Sin associated with Cos is defined by

$$\mathrm{Sin}(t) := \int_0^t \mathrm{Cos}(s)ds,$$

where the integral is a Bochner integral.

We can deduce some properties which are analogous to the well known trigonometric identities

Proposition 2.6 (Trigonometric Identities)**.** *Let* $\mathrm{Cos} : \mathbb{R} \to \mathcal{L}(H)$ *be a cosine function and* Sin *its associated sine function. Then the following relations hold:*

(i) $\mathrm{Sin}(-t) = -\mathrm{Sin}(t)$ *for* $t \in \mathbb{R}$

(ii) $2\,\mathrm{Sin}(t)\,\mathrm{Sin}(s) = \int_{t-s}^{t+s} \mathrm{Sin}(r)dr$

(iii) $\mathrm{Sin}(t+s) = \mathrm{Cos}(s)\,\mathrm{Sin}(t) + \mathrm{Sin}(s)\,\mathrm{Cos}(t)$ *for all* $s, t \in \mathbb{R}$.

Proof. See [1, p.209f]. □

It is important to note that for $y_0, y_1 \in H$ the generator A of Cos has to be bounded (see [1, p.213f]). Moreover, it is possible to find a suitable "phase space" $V \times H$ for the initial values.

Theorem 2.7. *Let* H *be a Banach space, and* A *an operator in* H*. Define the operator* $\mathcal{A}$ *in* $H \times H$ *by*

$$\mathcal{A} := \begin{pmatrix} 0 & I \\ A & 0 \end{pmatrix},$$

and the the norm on $H \times H$ *by* $\|(x,y)\|_{H\times H} := \|x\|_H + \|y\|_H$ *Then the following assertions are equivalent:*

(i) The operator A *generates a cosine function.*

(ii) There exists a Banach space V *such that* $D(A) \hookrightarrow V \hookrightarrow H$ *and such that the part* $\mathcal{B}$ *of* $\mathcal{A}$ *in* $V \times H$ *generates a strongly continuous semigroup.*

The Banach space V *is uniquely determined by* (ii)*. We call* $V \times H$ *the phase space associated with* A*. Moreover, one has* $\mathrm{Sin}(\cdot)y \in C(\mathbb{R}, V)$ *for all* $y \in H$*,* $\mathrm{Cos}(\cdot)x \in C^1(\mathbb{R}, H) \cap C(\mathbb{R}, V)$ *for all* $x \in V$*,* $\mathrm{Sin}(\cdot)x \in C(\mathbb{R}, D(A))$ *for all* $x \in V$*, and* $\mathcal{B}$ *generates a strongly continuous semigroup* $\mathcal{S}$ *on* $V \times H$ *given by*

$$\mathcal{S} = \begin{pmatrix} \mathrm{Cos}(t) & \mathrm{Sin}(t) \\ \mathrm{Cos}'(t) & \mathrm{Cos}(t) \end{pmatrix} = \begin{pmatrix} \mathrm{Cos}(t) & \mathrm{Sin}(t) \\ A\,\mathrm{Sin}(t) & \mathrm{Cos}(t) \end{pmatrix} \text{ for } t \in \mathbb{R},$$

and $\mathrm{Cos}'(t)x = A\,\mathrm{Sin}(t)x$ *for* $x \in V$*.*

Proof. See [1, p.212]. □

There are several other important relations between sine and cosine functions, and their generator:

- Let Cos a cosine function, Sin its associated sine function and A their generator. If $x \in D(\mathrm{A})$, then $\mathrm{Cos}(t)x$, $\mathrm{Sin}(t)x \in D(\mathrm{A})$ and $A\,\mathrm{Cos}(t)x = \mathrm{Cos}(t)\,\mathrm{A}\,x$, $\mathrm{A}\,\mathrm{Sin}(t)x = \mathrm{Sin}(t)\,\mathrm{A}\,x$, for all $t \geq 0$. I.e. the sine and cosine functions commute with their generator. (See [1, p.210f].)
- Let A be the generator of the cosine function $\mathrm{Cos}(\cdot)$ and the sine function $\mathrm{Sin}(\cdot)$. If A is self adjoint, then $\mathrm{Cos}(t)$ and $\mathrm{Sin}(t)$ are also self adjoint for all $t \geq 0$.

With the help of the sine and cosine operators we can give an explicit representation of the solution y of (W) using the well known variation of constants formula

$$y(t) := \mathrm{Cos}(t)y_0 + \mathrm{Sin}(t)y_1 + \int_0^t \mathrm{Sin}(t-s)f(s)\,ds \text{ for } t \geq 0. \tag{5}$$

In particular, the solution operator S is given by

$$(Sf)(t) = \int_0^t \mathrm{Sin}(t-s)f(s)\,ds.$$

A simple computation shows that the adjoint operator $S^* : L_2(0,T;H) \to L_2(0,T;H)$, is given by

$$(S^*g)(t) = \int_t^T \mathrm{Sin}((t-s))g(s)ds, \tag{6}$$

provided that A is self adjoint. The operator S^* is the solution operator of the equation

$$\begin{aligned} \frac{\partial^2 p}{\partial t^2} + Ap &= g, \\ p(T) &= 0, \\ p_t(T) &= 0, \end{aligned} \tag{AW}$$

Remark 2.8. Further note that p and y are related by $p(t) = y(T-t)$, where y is the solution of (W) with right hand side $f(t) = g(T-t)$, and $y_0 = y_1 = 0$.

This means that solving of the adjoint problem is equivalent to "revert" the time interval $[0,T]$ then solve the wave equation and "revert" the interval back. To be more precisely: For a given function f we denote by f^- the reverted function $f^-(t) = f(T-t)$. Then S^* could be alternatively written by

$$S^*g = (Sg^-)^-. \tag{7}$$

3. Gautschi Time-stepping

In this section we discuss the time integration of the wave equation (W) with the help of the Gautschi time stepping scheme.

Throughout we denote by $L \in \mathbb{N}$ the number of timesteps and by $\tau := \frac{T}{L}$ the stepsize.

3.1. Signals and structural properties of the Gautschi time-stepping scheme

Before we start with the definition of the Gautschi time-stepping scheme we introduce the notion of vector valued signals, to provide a proper distinction between the discrete and the continuous setting.

Definition 3.1. A mapping

$$s : \{0, \dots, L\} \to H$$

is called a (vector valued, finite) time discrete signal. We denote with

$$\mathrm{Signa}_L(H) := \{s |\, s : \{0, \dots, L\} \to H\}$$

the space of signals over $\{0, \dots, L\}$ in H. In this section let L be fixed.

Further we define the discrete scalar product over $\mathrm{Signa}_L(H)$ via the trapezoidal rule by

$$\langle f, g\rangle_{D(L)} := \tau \sum_{\ell=0}^{L} \langle f[\ell], g[\ell]\rangle_H - \frac{\tau}{2}(\langle f[0], g[0]\rangle_H + \langle f[L], g[L]\rangle_H) \tag{8}$$

To distinguish between functions (defined on a continuous domain of definition) and signals (defined on a discrete set) we use the following notation:

Notation 3.2. We write $s[\ell]$ for the ℓ-th value of the signal s in analogy to the notation which is used in signal processing.

For a function $f : [0, T] \to H$ we define the signal $f \in \mathrm{Signa}_L(H)$ by

$$f[\ell] := f(\ell\tau) \text{ for } \ell \in \{0, \dots, L\} \text{ and } \tau = \frac{T}{L}.$$

For given f the Gautschi time-stepping scheme recursively defines the signal $y \in \mathrm{Signa}_L(H)$ by

$$y[\ell + 1] = 2y[\ell] - y[\ell - 1] + 2\tau^2 \frac{\mathrm{Cos}(\tau) - I}{\tau^2 A}\,(-Ay[\ell] + f[\ell]) \text{ for } \ell \geq 1, \tag{9}$$

with initial values

$$y[0] := y_0, \tag{10}$$

and

$$y[1] := \mathrm{Cos}(\tau)y_0 + \mathrm{Sin}(\tau)y_1 + \frac{I - \mathrm{Cos}(\tau)}{A} f[0]. \tag{11}$$

Note that in case $y[\ell] \notin D(A)$ we can use the operator $(\mathrm{Cos}(\tau) - I)y[\ell]$ instead of $\frac{\mathrm{Cos}(\tau)-I}{A} Ay[\ell]$. Then the recursion is well defined for all signals on $\mathrm{Signa}_L(H)$. For more information about the Gautschi time-stepping scheme we refer to [11] and [8]. The signal y is a pointwise approximation in time of the actual solution of the wave equation (W), i.e. $y[\ell] \approx y(\ell\tau)$ for $0 \leq l \leq L$. This suggests the following definition:

Definition 3.3 (Discrete solution operator)**.** Define the (time-)discrete solution operator $S_L : \mathrm{Signa}_L(H) \to \mathrm{Signa}_L(H)$ recursively by the Gautschi time-stepping scheme with homogeneous initial values, i.e.

$$S_L(f)[\ell] := y[\ell],$$

where y is defined recursively by (9)-(11), with $y_0 = y_1 = 0$.

Definition 3.4 (Discrete Adjoint Solution Operator)**.** The discrete adjoint operator $S_L^* : \mathrm{Signa}_L(H) \to \mathrm{Signa}_L(H)$ is defined by the relation

$$\langle S_L^*(g), f\rangle_{D(L)} = \langle g, S_L(f)\rangle_{D(L)} \text{ for all } f \in \mathrm{Signa}_L(H).$$

We can expect that the discrete adjoint solution operator S_L^* will also be a backward solution of the time-stepping scheme. For this purpose we introduce the notion of reverse signals.

Definition 3.5 (Reverse signals)**.** For a signal $f \in \mathrm{Signa}_L(H)$ we define the reverse signal $f^- \in \mathrm{Signa}_L(H)$ by the relation

$$f^-[\ell] := f[L - \ell].$$

Thus the notation of a reverted signal is the continuous equivalent of the notation of a reverted function introduced in Remark 2.8.

The following result proves our conjecture concerning S_L^*. It is non trivial, since this is not true in general for other scalar products. It is also worth while to note that the proof only uses algebraic properties of the Gautschi time-stepping scheme and not approximation properties.

Theorem 3.6 (Relation between the discrete adjoint and the continuous adjoint operators)**.** *The discrete adjoint S_L^* operator with respect to the discrete scalar product $\langle\cdot,\cdot\rangle_{D(L)}$ is exactly the solution of the backward Gautschi time stepping scheme, i.e.*

$$(S_L^*(g))[n] = x[n] = (S_L(g^-))^-[n], \tag{12}$$

and $x[L+1] = 0$. Thus the relation (7) between the continuous operators carry over to the discrete operators, if we use the trapezodial rule as integration formula.

Proof. See Appendix A. □

Remark 3.7. Theorem 3.6 shows that the adjoint relationship of the discrete operators with respect to the discrete scalar products is an algebraic property. It does not rely on approximation properties of the Gautschi scheme or the trapezidoal rule. This is very important for the convergence of the Conjugate Gradient (CG) algorithm, which will be used below.

Remark 3.8. The discrete operators are singular, since if $f[j] = 0$ for $j = 0, \ldots, L$ and $f[L+1] \neq 0$, then $S_L(f) = 0$, due to the construction of the Gautschi time-stepping scheme.

3.2. Time Discretization Error of the Gautschi Time Stepping Scheme

3.2.1. Local time error

In this section we analyze the error of a single timestep of the Gautschi time-stepping scheme. Later we will use this result to provide a global error bound. In order to provide space discretization independent error bounds, we first show that in $\mathbb{R}^N$ for a symmetric and positive definite matrix $\mathfrak{A}$ the error of the time discretization does not depend on the dimension N.

We state a lemma which is proven in the thesis of Lintner [11]: Define $M_i := \max_{t\in[0,T]} \left\|f^{(i)}(t)\right\|$, where $\|\cdot\|$ denotes the Euclidean norm, and let $\ell \in \{0, \ldots, L\}$.

Lemma 3.9. *Let $\mathfrak{A}$ be a self adjoint and symmetric matrix. Then the pointwise error $e_j := y(t_j) - y[j]$ is given by the recursion*

$$e_{\ell+1} = W_\ell e_1 - \sum_{j+1}^{\ell} W_{\ell-j} d_j,$$

where

$$d_\ell = y(t_{\ell+1}) - 2y(t_\ell) + y(t_{\ell-1}) - \tau^2 G(-\mathfrak{A}y(t_\ell) + f(t_\ell)),$$

is the truncation error and

$$W_n = \sin((\ell - j + 1)\tau\sqrt{\mathfrak{A}})(\sin(\tau\sqrt{(\mathfrak{A})}))^{-1}.$$

Now we modify some results from [11] using less regularity assumptions on the right hand side than in [11]. These are needed later for the global error analysis of the Gautschi time-stepping scheme in infinite dimensional optimization.

Lemma 3.10. *For the trunctation error*

$$d_\ell = y(t_{\ell+1}) - 2y(t_\ell) + y(t_{\ell-1}) - \tau^2 G(-\sqrt{\mathfrak{A}}y(t_\ell) + f(t_\ell)),$$

we have

$$\|d_\ell\| \leq \tau^3 C M_1 \text{ if } f \in C^1(0,T;\mathbb{R}^N),$$

and

$$\|d_\ell\| \leq \tau^4 C(M_1 + M_2) \text{ if } f \in C^2(0,T;\mathbb{R}^N),$$

for a suitable constant $C > 0$ independent of N.

Proof. By (5) we have

$$d_\ell = \int_0^\tau (\tau\sqrt{\mathfrak{A}})^{-1} \sin(\tau\sqrt{\mathfrak{A}})(f(t_\ell + s) - 2f(t_\ell) + f(t_\ell - s))ds.$$

Further it holds that

$$\|d_\ell\| \leq \tau^2 \int_0^1 \left\|(\tau\sqrt{\mathfrak{A}})^{-1} \sin((1-\theta)\tau\sqrt{\mathfrak{A}})\right\| \|f(t_\ell - \theta\tau) - 2f(t_\ell) + f(t_\ell - \theta\tau)\| \, d\theta.$$

For the case $f \in C^1(0,T;R^N)$ it we have

$$\begin{aligned} \|f(t_\ell + s) - 2f(t_\ell) + f(t_\ell - s)\| &= \left\| \int_{t_\ell}^{t_\ell+s} f'(r)dr - \int_{t_\ell-s}^{t_\ell} f'(r)dr \right\| \\ &\leq 2s \max_{t\in[0,T]} \|f'(t)\|, \end{aligned} \tag{13}$$

and for $f \in C^2(0,T;R^N)$

$$\begin{aligned} &\|f(t_\ell + s) - 2f(t_\ell) + f(t_\ell - s)\| \\ &\quad = \left\| \int_{t_\ell}^{t_\ell+s} (t_\ell + s - r)(f''(r)dr) - \int_{t_\ell}^{t_\ell+s} (t_\ell + s - r)(f''(r)dr) \right\| \\ &\quad \leq s^2 \max_{t\in[0,T]} \|f''(t)\|. \end{aligned} \tag{14}$$

The estimates (13) and (14) together with the fact that $\left\|(\tau\sqrt{\mathfrak{A}})^{-1} \sin((1-\theta)\tau\sqrt{\mathfrak{A}})\right\| \leq (1-\theta)$, imply the claim. □

Proposition 3.11. *For the pointwise error e_ℓ we get*

$$\|e_\ell\| \leq \tau C(t_\ell^2 M_1) \text{ if } f \in C^1(0,T;\mathbb{R}^N),$$

or

$$\|e_\ell\| \leq \tau^2 C(t_\ell M_1 + t_\ell^2 M_2) \text{ if } f \in C^2(0,T;\mathbb{R}^N).$$

The constant C does not depend on the space discretization, i.e. on the dimension N.

Proof. We proceed as in [11, Thm. 2.5., p.28f] except we additionally apply our Lemma 3.10, to get a sharper error bound. □

Remark 3.12. We see that the error bound does not depend on the dimension N of the space discretization.

After analyzing the Gautschi time-stepping scheme in $\mathbb{R}^N$, we lift the error bounds into the infinite dimensional Hilbert space H.

Theorem 3.13. *For the Gautschi time-stepping scheme* (9)-(11) *it is true that*

$$\|y(t_\ell) - y_\ell\|_H \leq C\tau t_\ell M_1 \text{ for } f \in C^1(0,T;H),$$

or

$$\|y(t_\ell) - y_\ell\|_H \leq C\tau^2(t_\ell M_1 + t_\ell^2 M^2) \text{ if } f \in C^2(0,T;H),$$

where

$$M_i := \max_{t\in[0,T]} \left\| f^{(i)}(t) \right\|_H .$$

Proof. The proof uses a standard Galerkin approximation argument. For details we refer to [15]. □

3.2.2. Global Time Error

Since a signal is only defined on a discrete set extra considerations are required for error analysis in sense of the L_2-norm. For this purpose we have to transform a signal to an interpolating function. A natural choice is the *discrete cosine transform* of type 1 (DCT-I).

Definition 3.14 (Vector valued DCT and DFT, and Interpolating operator)**.** Let $f \in \mathrm{Signa}_L(H)$ be a vector valued signal. The DCT of f is defined by the mapping $\mathcal{C}_L : \mathrm{Signa}_L(H) \to L_2(0,T;H)$ given by

$$\mathcal{C}_L(f)(t) = \sum_{k=0}^{L} X[k] \cos\left(\frac{k\pi t}{T}\right),$$

where the signal X is defined as

$$X[k] = \frac{1}{L}\left(\frac{f[0] + (-1)^k f[L]}{2} + \sum_{j=1}^{L-1} f[k] \cos\left(\frac{kj\pi}{L}\right)\right).$$

The DFT of degree L of the signal f is the mapping $\mathcal{F}_L : \mathrm{Signa}_L(H) \to L_2(-\pi,\pi;H)$ defined by

$$\mathcal{F}_L(f)(x) = \sum_{k=1-\lceil L/2\rceil}^{\lfloor L/2\rfloor} C[k] e^{ikx},$$

where the signal C is defined by

$$C[k] = \frac{1}{L} \sum_{j=0}^{L-1} f\left(\frac{2j\pi}{L}\right) e^{-ik\frac{2\pi j}{L}}.$$

Note that we do not distinguish above between H and it's complexification. The DFT/DCT maps a signal to its trigonometric/cosine interpolation polynomial of degree L.

The interpolation operator on an equidistant grid $\mathcal{I}_L : C^0([0,T];H) \to \mathrm{Signa}_L(H)$ is defined by

$$\mathcal{I}_L(f)[n] := f[n] := f\left(\frac{nT}{L}\right).$$

The DCT of a vector valued function $f : [0,\pi] \to X$, respectively the DFT of a function $f : [-\pi,\pi] \to X$ is simply the DCT/DFT of the interpolating signal of f on the equidistant grid. I.e. we will write for the sake of simplicity

$$\mathcal{C}_L(f) = \mathcal{C}_L(\mathcal{I}_L(f)),$$

and

$$\mathcal{F}_L(f) = \mathcal{F}_L(\mathcal{I}_L(f)),$$

for continuous f.

The DCT and DFT transformations can be defined for $L_2(0,T;H)$ in analogous manner by a coordinate transform.

Remark 3.15. It is important to note that the interpolation operator $\mathcal{I}_L$ is the left inverse operator of the DCT operator $\mathcal{C}_L$, i.e. for all signals $f \in \mathrm{Signa}_L(H)$ the relation

$$\mathcal{I}_L(\mathcal{C}_L(f)) = f,$$

holds.

Remark 3.16. We will not distinguish between the operators $S_L : \mathrm{Signa}_L(H) \to \mathrm{Signa}_L(H)$ and $\mathcal{C}_L \circ S_L \circ \mathcal{I}_L : C^1([0,T];H) \to L_2(0,T;H)$ and simply write S_L for $\mathcal{C}_L \circ S_L \circ \mathcal{I}_L$ if it is clear from the context.

Further we will apply the following rule: If a signal is expected to be a continuous function then the signal will be identified by its DCT, on the contrary if it a signal should be a continuous function then we identify this function by its interpolating signal.

Definition 3.17 (Vector valued F series)**.** The formal vector valued F series of a function f defined on $[-\pi,\pi]$ is defined as

$$\sum_{k=-\infty}^{\infty} C[k]e^{ikt},$$

where the vector valued (infinite) signal $C : \mathbb{N} \to X$ is given by

$$C[k] = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)e^{-ikt}dt.$$

Alternatively one can also use its real valued representation, which is defined as in the scalar case, namely

$$F(t) := \frac{A[0]}{2} + \sum_{k=0}^{\infty} A[k]\cos(kt) + B[k]\sin(kt),$$

with the vector valued signals

$$A[k] = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos(kt)dt,\ B[k] = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin(kt)dt.$$

Remark 3.18. The DFT above is a partial sum of the F series where the coefficients are approximated by the trapezidoal rule. If we define

$$\tilde{f}(t) := \begin{cases} f(x) & \text{for } x \in [0,\pi] \\ f(-x) & \text{for } x \in [-\pi, 0] \end{cases},$$

then $\mathcal{C}_L(f) = \mathcal{F}_{2L}(\tilde{f})$. If f was continuously differentiable and we extend $\tilde{f}$ 2π-periodically to the whole real line, then the extension is periodic, continuous and piecewise continuously differentiable.

Remark 3.19. For the discrete scalar product it is true that for all trigonometric polynomials with degree less or equal to $L \in \mathbb{N}$ the relation $\frac{1}{2\pi}\langle\cdot,\cdot\rangle^2_{L_2(-\pi,\pi)} = \langle\cdot,\cdot\rangle^2_{D(L)}$ holds, since for the functions e^{ikt} for $-L/2 < k \leq L/2$ the same orthonormality relations hold in the discrete and the continuous scalar product (see [7, p.399f]). Thus for two orthogonal polynomials $f = \sum_{k=1-\lceil L/2\rceil}^{\lfloor L/2\rfloor} f[k]e^{ikx}$, and $g = \sum_{k=1-\lceil L/2\rceil}^{\lfloor L/2\rfloor} g[k]e^{ikx}$ of degree L we have the relation

$$\frac{1}{\pi}\langle f, g\rangle_{L_2(0,T;H)} = \sum_{k=1-\lceil L/2\rceil}^{\lfloor L/2\rfloor} \langle f[k], g[k]\rangle_H = \langle f, g\rangle_{D(L)}.$$

Thus the value of discrete scalar product is exact for trigonometric polynomials of degree L. Analogously the relation $\frac{1}{\pi}\langle\cdot,\cdot\rangle^2_{L_2(0,\pi)} = \langle\cdot,\cdot\rangle^2_{D(L)}$ holds for all cosine polynomials of degree L, using the relation $\mathcal{C}_L(f) = \mathcal{F}_{2L}(\tilde{f})$ from Remark 3.18.

We note also that

$$\langle f, g\rangle_{L_2(0,T;H)} = \frac{T}{\pi}\langle f, g\rangle_{L_2(0,\pi;H)} = C(T)\langle f, g\rangle_{D(L)},$$

where $C(T) = T$ is the constant which results from a coordinate transform from $[0,\pi]$ onto $[0,T]$.

Lemma 3.20 (Aliasing)**.** *For $f \in C^1([0,T];H)$ the Fourier coefficients $\hat{C}[k]$ of $\mathcal{F}_L(f)$ satisfy the identity*

$$\hat{C}[k] = \sum_{l=-\infty}^{\infty} C[k + lL],$$

for $k \in \{-N/2, \ldots, N/2\}$.

Proof.

$$\hat{C}[k] = \frac{1}{L}\sum_{j=0}^{L-1}\left(\sum_{l=-\infty}^{\infty} C[l]e^{il\frac{2\pi j}{L}}\right)e^{-ik\frac{2\pi j}{L}} = \frac{1}{L}\sum_{l=-\infty}^{\infty} C[l]\left\langle e^{ik\cdot}, e^{il\cdot}\right\rangle_{D(L)}$$
$$= \frac{1}{L}\sum_{l=-\infty}^{\infty} C[k+lL].$$

□

Corollary 3.21. *For $f \in C^1([0,T];H)$ we have*

$$f(t) - \mathcal{F}_L(f) = 2\sum_{k\in\mathbb{Z}\setminus\{-\lceil L/2\rceil-1,\dots,\lfloor L/2\rfloor\}} C[k]e^{ikt}.$$

Proof. We use Lemma 3.20. After a straightforward calculation and rearranging the sums, using absolute convergence, we get the desired result. □

Theorem 3.22. *Let $f \in C([0,T];H)$ be piecewise continuously differentiable, and let $\mathcal{C}_L(f)$ be its DFT on an equidistant grid with even number of grid points. Then it holds that*

$$\|f - \mathcal{C}_L(f)\|_{L_2} \leq \frac{c}{L}\|f\|_{H^1(0,T;H)}.$$

If f is twice continuously differentiable it is even true that

$$\|f - \mathcal{C}_L(f)\|_{L_2} \leq \frac{c}{L^2}\|f\|_{H^1(0,T;H)}.$$

Proof. We can follow the proofs of Hanke-Bourgeois [7] by using Lemma 3.20 and Bessel's inequality. □

We now have the following estimate for the solution operator.

Theorem 3.23. *Let $j \in \{1,2\}$. For $x \in C^j([0,T];H)$ there exists a constant $D(T)$ such that*

$$\|Sx - S_Lx\|_{L_2(0,T;H)} \leq \frac{D(T)\|x\|_{C^j([0,T];H)})}{L^j}.$$

Proof. We have

$$\|Sx - S_Lx\|_{L_2(0,T;H)} \leq \|Sx - \mathcal{C}_L(Sx)\|_{L_2(0,T;H)} + \|\mathcal{C}_L(Sx) - S_Lx\|_{L_2(0,T;H)}.$$

Now we apply Theorem 3.22 and Theorem 3.13, keeping in mind that

$$\|\mathcal{C}_L(Sx) - S_Lx\|_{L_2(0,T;H)} = T\|Sx - S_Lx\|_{D(L)}.$$

□

Corollary 3.24. *Let $j \in \{1,2\}$. For $x \in C^j([0,T];H)$ there exists a constant $D(T)$ such that*

$$\|S^*x - S_L^*x\|_{L_2(0,T;H)} \leq \frac{D(T)\|x\|_{C^j([0,T];H)}}{L^j}.$$

Proof. Shifting the adjoint problem (AW) like in Remark 2.8 and using the identity (12), we can apply Theorem 3.23 to show convergence. □

4. Optimization

This section is devoted to discussing the solution of (OC) on the basis of the optimality system (3). Self-adjointness $\mathcal{H} = S^*S + \alpha I$, suggests to use the conjugate gradient method, whose asymptotic convergence properties also hold in infinite dimensions (see [12]).

4.1. Convergence in time

The discrete operators S_L and S_L^* of Section 3 cannot be directly defined on $L_2(0,T;H)$, because the set $IP = \{\frac{k\pi}{LT} : k, L \in \mathbb{N}\}$ is a Lebesgue null-set. Thus two functions f, g which coincide on $[0,T] \setminus IP$ are the same functions in the Lebesgue sense, but $S_L f \neq S_L g$, if $f \neq g$ on IP. As a consequence we interpret S_L and S_L^* as operators from $C^1([0,T];H)$ into $L_2(0,T;H)$. Of course, in this sense S_L^* is not the adjoint operator of S_L.

For $B \in \mathcal{L}(C^1([0,T];H), L_2(0,T;H))$ we denote with $\|B\|_{C^1;L_2}$ the operator norm

$$\|B\|_{C^1;L_2} := \|B\|_{C^1([0,T],H);L_2(0,T;H)} := \sup_{y \in C^1, \|y\|_{L^2}=1} \|By\|_{L^2((0,T);H)}\,.$$

This norm should not be confused with the induced operator norm. Also we have to note that the normed space $(\mathcal{L}(C^1([0,T];H), L_2(0,T;H)), \|B\|_{C^1;L_2})$ is not a Banach space. Thus we cannot apply the Banach-Steinhaus Theorem, and have to take a detour in the proof of Lemma 4.1.

Lemma 4.1. *For the operators $S_L, S_L^* : C^1([0,T];H) \to L_2(0,T;H)$ with $L \in \mathbb{N}$, the sequences $\|S_L\|_{C^1;L_2}$ and $\|S_L^*\|_{C^1;L_2}$ are uniformly bounded with respect to L.*

Proof. From Theorem 3.23 it follows that for fixed $x \in C^1([0,T];H)$ we have

$$\lim_{L\to\infty} \|Sx - S_L x\|_{L_2(0,T;H)} \leq \lim_{L\to\infty} \frac{D(T)\,\|x\|_{C^1([0,T];H)}}{L} = 0,$$

and therefore

$$\lim_{L\to\infty} S_L x = Sx \text{ for all } x \in C^1([0,T];H).$$

Analogously we prove that

$$\lim_{L\to\infty} S_L^* x = S^* x \text{ for all } x \in C^1([0,T];H). \tag{15}$$

First we note that $C^1([0,T],H)$ is dense in $L_2(0,T;H)$ (See [1][L 1.3.3]), and that the cosine polynomials also are dense in L_2 in the L_2-norm. We also recall that a continuous function is identified by its interpolating signal in the discrete scalar product. Suppose now that $\limsup_{L\to\infty} \|S_L\|_{C^1;L_2} = \infty$ (or even $\|S_L\|_{C^1;L_2} = \infty$ for some $L \in \mathbb{N}$). Without loss of generality we assume that the whole sequence diverges to infinity. Then there exists a sequence $(x_n) \in C^1([0,T],H)$ such that $\|x_n\|_{L_2} = 1$ and a sequence $(M_L)_{L\in\mathbb{N}}$ such that $\|S_L x_L\|_{L_2(0,T;H)} > M_L$ and $\lim_{L\to\infty} M_L = \infty$. Since $(x_L)_{L\in\mathbb{N}}$ is bounded in $L_2([0,T],H)$, the sequence $(\mathcal{C}_L(x_L))_{L\in\mathbb{N}}$ is also bounded. This follows from the fact that the interpolation error can be estimated by the remainder of the F series times a constant factor (see Hanke-Burgeois [7] or Boyd [2, p.94]), and

therefore $\{\mathcal{C}_L(x_L) : L \in \mathbb{N}\}$ is bounded in $L_2(0,T;H)$ by Parseval's inequality. From boundedness it follows that there exists $x \in L_2(0,T;H)$ and a subsequence $(\mathcal{C}_{L_k}(x_{L_k}))_{k\in\mathbb{N}}$ which converges weakly in $L_2(0,T;H)$ to some $x \in L_2(0,T;H)$.

We also observe that for $f \in C^1([0,T];H)$, and a trigonemtric polynomial g of degree L the relation

$$C(T)\,\langle f, g\rangle_{D(L)} = C(T)\,\langle \mathcal{C}_L(f), g\rangle_{D(L)} = \langle \mathcal{C}_L(f), g\rangle_{L_2(0,T;H)} \tag{16}$$

holds. Thus for all g which are cosine polynomials of degree K we have

$$\begin{aligned}\langle S_{L_k}x_{L_k}, g\rangle_{L_2(0,T;H)} &= C(T)\,\langle S_{L_k}x_{L_k}, g\rangle_{D(L_k)} = C(T)\,\left\langle x_{L_k}, S^*_{L_k}g\right\rangle_{D(L_k)}\\ &= \left\langle \mathcal{C}_{L_k}(x_{L_k}), S^*_{L_k}g\right\rangle_{L_2(0,T;H)},\end{aligned}$$

for $L_k > K$. Now we may conclude that

$$\begin{aligned}\lim_{k\to\infty}\langle S_{L_k}x_{L_k}, g\rangle_{L_2(0,T;H)} &= \lim_{k\to\infty}\left\langle \mathcal{C}(x_{L_k})_{L_k}, S^*_{L_k}g\right\rangle_{L_2(0,T;H)} = \langle x, S^*g\rangle_{L_2(0,T;H)}\\ &= \langle Sx, g\rangle_{L_2(0,T;H)},\end{aligned}$$

using Relation (15) which implies $S^*_{L_k}g \to S^*g$, and the assumption that $\mathcal{C}_{L_k}(x_{L_k}) \rightharpoonup x$. Thus we have $\left\langle \mathcal{C}_{L_k}(x_{L_k}), S^*_{L_k}g\right\rangle \to \langle x, S^*g\rangle$ for all cosine polynomials g. Since the trigonometric polynomials are dense in $L_2(0,T;H)$, we get that $S_{L_k}x_{L_k} \rightharpoonup Sx$, which is a contradiction to the unboundedness of the sequence $(S_{L_k}x_{L_k})_{k\in\mathbb{N}}$. Boundedness of $\|S^*_L\|_{C^1;L_2}$ can be verified analogously. □

Now we (semi-)discretize the functional J defined in (2) by approximating it by the time-discrete functional J_L for $L \in \mathbb{N}$

$$J_L(u_L) := \frac{1}{2}\,\|S_Lu_L - z_L\|^2_{D(L)} + \frac{\alpha}{2}\,\|u_L\|^2_{D(L)}\,, \tag{17}$$

where $u_L \in \mathrm{Signa}_L(H)$ and $z_L = \mathcal{I}_L(z)$. Taking directional derivatives and applying Theorem 3.6 leads us to the discrete optimality condition

$$J'_L(u_L)v_L = \left.\frac{dJ_L}{dt}(u_L + tv_L)\right|_{t=0} = \langle \mathcal{H}_{\mathcal{L}}u_L - S^*_Lz_L, v_L\rangle_{D(L)} = 0,$$

for all $v_L \in \mathrm{Signa}_L(H)$ or equivalently

$$\mathcal{H}_Lu_L = S^*_Lz_L, \tag{18}$$

with the approximated Hessian

$$\mathcal{H}_L := S^*_LS_L + \alpha\,\mathcal{C}_L = \mathcal{C}_L \circ S^*_LS_L \circ \mathcal{I}_L + \alpha\,\mathcal{C}_L\,.$$

Conversely if we approximate the continuous optimality condition (3) by replacing S by S_L, and z by $z_L = \mathcal{I}_L(z)$, then we arrive again at (18). With this choice of discretization and S^*_L as in (12) of Theorem 3.6 the operator $\mathcal{H}_L$ in (18) is self-adjoint.

In order to prove $L_2(0,T;H)$ convergence of the optimization scheme we first have to estimate the error in the discrete norm.

Lemma 4.2 (Error in the discrete norm)**.** *Let u be the continuous control, u_L the solution of the discrete system $\mathcal{H}_L u_L = S_L^* z_L$ and $j \in \{1,2\}$. If u and $z \in C^j([0,T];H)$ where $z = \tilde{z} - S_H(y(0), y_t(0))$. Then the error $\|u - u_L\|_{D(L)}$ in the discrete norm is of the order $\mathcal{O}\left(\frac{1}{L^j \alpha}\right)$.*

Proof. Set $z_L := \mathcal{C}_L(\mathcal{I}(z))$, where $\mathcal{C}_L(\mathcal{I}_L(z))$ is the DCT of z defined in Definition 3.14. Recall the definitions $\mathcal{H} = S^*S + \alpha I$, and the approximated Hessian $\mathcal{H}_L := S_L^* S_L + \alpha\, \mathcal{C}_L$. First consider the error of the discrete control u_L. Recall that in the discrete norm u is identified by $\mathcal{I}_L(u)$), and that $\{\mathcal{H}_L\}$ is uniformly elliptic in the discrete scalar product with ellipticity constant α. Then it holds that

$$\alpha \|u - u_L\|_{D(L)}^2 \leq \langle \mathcal{H}_L(u - u_L), u_L - u\rangle_{D(L)} = \langle \mathcal{H}_L u - \mathcal{H}_L u_L, u - u_L\rangle_{D(L)}\,.$$

Therefore we have

$$\|u - u_L\|_{D(L)} \leq \frac{\|\mathcal{H}_L u - \mathcal{H}_L u_L\|_{D(L)}}{\alpha} = \frac{C(T)}{\alpha} \|\mathcal{H}_L u - \mathcal{H}_L u_L\|_{L_2(0,T;H)}\,. \quad (19)$$

To investigate the error further we make the following splitting:

$$\mathcal{H}_L u - \mathcal{H}_L u_L = \mathcal{H}_L u - \mathcal{H} u + \mathcal{H} u - \mathcal{H}_L u_L = (\mathcal{H}_L - \mathcal{H})u + (S^* z - S_L^* z_L).$$

The first term can be estimated by

$$\begin{aligned}
\|\mathcal{H}_L u - \mathcal{H} u\|_{L_2(0,T;H)} &= \|S_L^* S_L u - S^* S u + \alpha(\mathcal{C}_L(u) - u)\|_{L_2(0,T;H)} \\
&= \|\mathcal{C}_L \circ S_L^* \circ S_L \circ \mathcal{I}_L(u) - S^* S u + \alpha(\mathcal{C}_L(u) - u)\|_{L_2(0,T;H)} \\
&= \|\mathcal{C}_L \circ S_L^* \circ \mathcal{I}_L \circ \mathcal{C}_L \circ S_L \circ \mathcal{I}_L(u) - S^* S u + \alpha(\mathcal{C}_L(u) - u)\|_{L_2(0,T;H)} \\
&\leq \|\mathcal{C}_L \circ S_L^* \circ \mathcal{I}_L \circ \mathcal{C}_L \circ S_L \circ \mathcal{I}_L\, u - \mathcal{C}_L \circ S_L^* \circ \mathcal{I}_L(Su)\|_{L_2(0,T;H)} \\
&\quad + \|\mathcal{C}_L \circ S_L^* \circ \mathcal{I}_L(Su) - S^* S u\|_{L_2(0,T;H)} + \alpha \|\mathcal{C}_L(u) - u\|_{L_2(0,T;H)} \\
&= \|\mathcal{C}_L \circ S_L^* \circ \mathcal{I}_L(\mathcal{C}_L \circ S_L \circ \mathcal{I}_L\, u - Su)\|_{L_2(0,T;H)} \\
&\quad + \|(\mathcal{C}_L \circ S_L^* \circ \mathcal{I}_L - S^*) S u\|_{L_2(0,T;H)} + \alpha \|\mathcal{C}_L(u) - u\|_{L_2(0,T;H)} \\
&= \|S_L^*(S_L u - Su)\|_{L_2(0,T;H)} \\
&\quad + \|(S_L^* - S^*) S u\|_{L_2(0,T;H)} + \alpha \|\mathcal{C}_L(u) - u\|_{L_2(0,T;H)} \\
&\leq \|S_L^*\|_{C^1;L_2} \|(S_L u - Su)\|_{L_2(0,T;H)} \\
&\quad + \|S_L^* S u - S^* S u\|_{L_2(0,T;H)} + \alpha \|u - \mathcal{C}_L(u)\|_{L_2(0,T;H)}\,.
\end{aligned} \quad (20)$$

Now let $j \in \{1,2\}$ and $u \in C^j([0,T];H)$. From Theorem 3.23 and Lemma 4.1 we know that

$$\|S_L^*\|_{C^1;L_2} \|(S_L u - Su)\|_{L_2(0,T;H)} \leq \frac{D_1 \|u\|_{C^j([0,T];H)}}{L^j},$$

and

$$\|S_L^* S u - S^* S u\|_{L_2(0,T;H)} \leq \frac{D_2 \|Su\|_{C^j([0,T];H)}}{L^j},$$

for suitable positive constants D_1, D_2. From Corollary 3.22 we also get the estimate

$$\|u - \mathcal{C}_L(u)\|_{L_2(0,T;H)} \leq \frac{D_3 \|u\|_{H^1([0,T];H)}}{L^j}. \quad (21)$$

Thus (20) becomes

$$\|\mathcal{H}_L u - \mathcal{H} u\|_{L_2(0,T;H)} \leq \frac{D_4 \max(\|u\|_{C^j([0,T];H)}, \|Su\|_{C^j([0,T];H)})}{L^j}, \tag{22}$$

for a suitable $D_4 > 0$.

Next we estimate $S^* z - S_L^* z_L = (S^* z - S_L^* z) + S_L^*(z - z_L)$. With the help of Corollary 3.24 and Lemma 4.1 we can find suitable constants $D_5, D_6 > 0$ such that

$$\|S^* z - S_L^* z\|_{L_2(0,T;H)} \leq \frac{D_5 \|z\|_{C^j([0,T];H)}}{L^j},$$

and

$$\|S_L^*\|_{C^1;L_2} \|z - z_L\|_{L_2(0,T;H)} \leq \frac{D_6 \|z\|_{H^1([0,T];H)}}{L^j}.$$

Thus for $D_7 = \max\{D_5, D_6\}$ we have

$$\|S^* z - S_L^* z_L\| \leq \frac{D_7 \|z\|_{C^j([0,T];H)}}{L^j}. \tag{23}$$

Combining the estimates (19), (20) and (23) we finally get

$$\begin{aligned} \|u - u_L\|_{D(L)} \leq & \frac{D_4 \|u\|_{C^j([0,T];H)}}{L^j} \\ & + \frac{D_7 \|z\|_{C^j([0,T];H)}}{L^j}, \end{aligned}$$

and thus

$$\|u - u_L\|_{L_2(0,T;H)} \leq \mathcal{O}\left(\frac{1}{\alpha L^j}\right),$$

which was to be shown. □

The following theorem shows that the resulting sequence $(u_L)_{L\in\mathbb{N}}$ of solutions to the discrete optimality condition approximates the solution u of the optimality system (3) in the $L_2(0,T;H)$ sense that

$$\lim_{L\to\infty} \|u - \mathcal{C}_L(u_L)\|_{L_2(0,T;H)} = 0.$$

Theorem 4.3 (Convergence of the optimization scheme in time)**.** *Let $u, z \in C^j([0,T];H)$ for $j \in \{1,2\}$, where $z = \tilde{z} - S_H(y(0), y_t(0)) \in C^1([0,T];H)$. Then the DCTs of the sequence $(u_L)_{L\in\mathbb{N}}$ of the semi approximated solutions in time converge strongly to the solution u of* (3)*. Further the error $\|u - u_L\|_{L_2(0,T;H)}$ is of order $\mathcal{O}\left(\frac{1}{\alpha L^j}\right)$.*

Proof. We have the estimate

$$\begin{aligned} \|u - u_L\|_{L_2(0,T;H)} &\leq \|u - \mathcal{C}_L(u)\|_{L_2(0,T;H)} + \|\mathcal{C}_L(u) - u_L\|_{L_2(0,T;H)} \\ &= \|u - \mathcal{C}_L(u)\|_{L_2(0,T;H)} + \frac{1}{C(T)} \|u - u_L\|_{D(L)}. \end{aligned}$$

With Theorem 3.22 and Lemma 4.2 we can conclude the proof. □

Remark 4.4 (Notes on the influence of the parameter α on the optimization error)**.** Consider $L \in \mathbb{N}$ fixed. Since $S_L^* S_L$ is positive definite, it follows that $\alpha \leq \lambda_{\min}(\mathcal{H}_L) \leq \lambda_{\max}(\mathcal{H}_L) \leq C + \alpha$, for a constant $C \geq 0$. It follows that

$$\kappa(\mathcal{H}) \leq \frac{C+\alpha}{\alpha} = 1 + C/\alpha = \mathcal{O}(1/\alpha).$$

Therefore the number of iterations, and the error of the Conjugate Gradient scheme defined in Listing 1 strongly depend on the parameter α. Numerical experiments validate this: Dividing α by ten leads to three times more iterations to reach the stopping criterion. Denote the approximated minimum of the CG scheme with u_{CG}, and the tolerance for the residual with ε. For the error it holds that,

$$\begin{aligned}\|u_L - u_{CG}\|_{D(L)} &= \left\|\mathcal{H}_L^{-1}\mathcal{H}_L(u_L - u_{CG})\right\|_{D(L)} = \left\|\mathcal{H}_L^{-1}(S_L^* z_L - \mathcal{H}_L u_{CG})\right\|_{D(L)} \\ &\leq \left\|\mathcal{H}_L^{-1}\right\| \varepsilon \leq \frac{\varepsilon}{\alpha}.\end{aligned}$$

This can also be verified by numerical experiments.

4.2. Inexact Conjugate Gradient Method

In order to solve the operator equation (3) we deal with the abstract operators $\mathcal{H}_L$, S_L and S_L^*, which represent solution algorithms of PDE. Since these algorithms only compute numerical approximations, rather than exact solutions we have to consider an inexact matrix vector multiplication, and further the impact on the calculated solution by the Conjugate Gradient (CG) algorithm.

Considering that the results of this section apply to the CG method in general, we look at a generic linear system $Qx = b$ in a Hilbert space $\mathbb{H}$, where Q is self adjoint, bounded and $\mathbb{H}$-elliptic with ellipticity constant $c > 0$. (In our case $Q = \mathcal{H}_L$ and $\mathbb{H} = L_2(0, T; H)$)

In order to solve the linear system we execute the abstract CG algorithm. In Appendix B in Listing 1 we give a simple implementation of the algorithm.

We use the the absolute residual criterion $\|r_n\| \leq \varepsilon$ as stopping criterion. The reason why we do not use the relative criterion $\|r_n\| \leq \|r_0\| \varepsilon$ is that in our case we often start with a large starting residuum r_0. An alternative would be to use $\|r_n\| \leq \min(\|r_0\|, 1)\varepsilon$

It is well known that the the algorithm from Listing 1 converges. See for example [12].

Now we start with investigating the behavior of the inexact CG method. We denote the approximation of the operator Q at the $k-th$ iteration by $\tilde{Q}_k := Q + E_k$, where E_k is a linear disturbance of Q with $\|E_k\| < \varepsilon$ for some $\varepsilon > 0$.

We further denote by r_k, p_k, q_k, α_k the r's, p's, q's and alpha's of the k-th iteration in the CG algorithm, and additionally with ρ_k the residuum of the k-th iteration.

Let us define the disturbance vector g_k of the k-th iteration of the inexact CG algorithm by $g_k = E_k p_k$ which is the error in the inexact matrix vector product, i.e. $\tilde{Q}_k p_k = q_k = Q p_k + g_k$.

Additionally we denote by $\tilde{r}_k$ the actual residual of the inexact method defined by

$$\tilde{r}_k = b - Q\left(x_0 + \sum_{j=0}^{k} \alpha_j p_j\right).$$

We are especially interested in investigating the residual gap which is given by $\|r_k - \tilde{r}_k\|$, since it measures the accuracy which we can achieve at best.

Recall that the recurrence relations for r_k and $\tilde{r}_k$, are given by

$$r_{k+1} = r_k - \alpha_k Q p_k,$$

and

$$\tilde{r}_{k+1} = \tilde{r}_k - \alpha_k Q p_k.$$

Forming the difference yields

$$\Delta r_{k+1} := \tilde{r}_{k+1} - r_{k+1} = \Delta r_k - \alpha_k g_k,$$

and therefore

$$\Delta r_m = \sum_{k=-1}^{m} \alpha_k g_k.$$

If we consider that

$$\Delta r_0 = \tilde{r}_0 - r_0 = b - Qx_0 - b - Qx_0 - E_{-1}x_0 = E_0 x_0,$$

and set $p_{-1} := x_0$, $g_{-1} := E_{-1}p_{-1}$, $\alpha_{-1} = 1$, then it holds that

$$\|\Delta r_n\| \leq \sum_{k=-1}^{m} \|\alpha_k g_k\| \leq \sum_{k=-1}^{m} |\alpha_k| \, \|E_k\| \, \|p_k\| \, .$$

From this inequality we deduce the following theorem, using an appropriate choice for an upper bound of $\|E_k\|$:

Theorem 4.5. *For $\varepsilon > 0$, let $r_m = b - Qx_m$ be the residual of the inexact CG method, and $\tilde{r}_m$ the actual residual after m iterations. Further denote by $\tilde{Q}_k = Q + E_k$ the disturbed operator of the k-th iteration. If*

$$\|E_k\| \leq \frac{d_k \varepsilon}{D \, |\alpha_k| \, \|p_j\|},$$

where $d_k \geq 0$ and $\sum_{k=-1}^{m} d_k = D$, then the residual gap satisfies $\|\tilde{r}_m - r_m\| \leq \varepsilon$.

A reasonable choice for d_k is $1/(m+1)$, for example. But there are other choices for d_k depending on the strategy used.

The residual gap indicates which accuracy could be achieved at best. This is an important matter considering numerical convergence of the CG algorithm, since the algorithm may not converge if the matrix vector operations are done without sufficient high accuracy.

Remark 4.6. Similar results can be found in [3]. The differences between the work in [3] and this section is, that we consider the conjugate gradient methods on arbitrary Hilbert spaces instead of $\mathbb{R}^N$ and give a more practical error bound.

5. Numerical Results

5.1. Numerical tests

We describe numerical results for two test cases. In the first one, the analytical solution is known, in the second one it is not. The first problem under consideration is

$$\min j(y,u) = \frac{1}{2}\left\|y-\tilde{z}\right\|^2_{L_2(0,T;\Omega)} + \frac{\alpha}{2}\left\|u\right\|^2_{L_2(0,T;\Omega)},$$

subject to

$$\begin{aligned} \Box y &= u, \\ y(0,x_1,x_2) &= 2\sin(x_1)\sin(x_2), \\ y_t(0,x_1,x_2) &= y_0, \\ y_{|\Omega}(t,x_1,x_2) &= 0, \end{aligned}$$

where $\Box = \frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x_1^2} - \frac{\partial^2}{\partial x_2^2} + 2$ in the domain $\Omega = (0,\pi)^2$, on the time horizon $[0,T]$ with $T = \pi/2$. The desired state $\tilde{z}(t,x,y) = z_1(t)\sin(x_1)\sin(x_2)$, where

$$z_1(t) = \frac{1}{4}\cos(2\,t)\cos(t)^4 + \frac{1}{8}\,t\sin(2\,t) + \frac{1}{8}\sin(2\,t)^2 + \frac{1}{32}\sin(2\,t)\sin(4\,t) + 2\,\alpha + \frac{7}{4}\cos(2\,t),$$

and the initial value $y(0,x_1,y_1) = 2\sin(x_1)\sin(x_2)$. For these data the optimal control is given by $u(t,x,y) = (1+\cos(2t))\frac{\sin(x_1)\sin(x_2)}{2}$. While the focus is put on the performance of the temporal discretization by means of the Gautschi method, we also compare among one of following spatial discretizations

- Spectral Method with trigonometric polynomials (Fourier method),
- Spectral Method with Legendre-Chebyshev Method
- FEM with piecewise linear functions, on a rectangle grid,
- Spectral Element Method (SEM) with piecewise Legendre Polynomials For Legendre polynomials with degree 1 the SEM method is equivalent to the FEM, and for one element SEM is equivalent to the spectral method.

For evaluation of the trigonometric functions we use the approach of Hochbruck and Grimm (see [6]). This approach is a Krylov method which computes an approximation of $\mathrm{Cos}(t)v$ and $\mathrm{Sin}(t)v$ in a $K-$ dimensional Krylov space $\mathrm{span}\{v, Qv, \dots Q^{K-1}v\}$.

In the tables we use the following notation: L is the number of time-steps which is a power of 2, i.e. $L = 2^l$ for $l \in \mathbb{N}$, and by N we denote the order of approximation in space. In the case of spectral methods this is the polynomial degree, while in the case of finite and spectral elements it is the number of Ansatz functions per element, times the number of elements. The error is measured by $\|u_{NL} - u\|_{L_2(0,T;L_2(\Omega))}$, where u is the analytical solution and u_{NL} is the numerical solution. With *iter* we denote the number of required CG iterations. As stopping criterion for the CG algorithm which is used for solving, we use $\|r_k\|_H < \varepsilon$ with the residual $r_k = Qx_k - b$.

l	error	iter
4	$1.35 \cdot 10^{-03}$	14
6	$8.34 \cdot 10^{-05}$	14
8	$5.60 \cdot 10^{-06}$	14

Table 1: Fourier; $N = 4$, $K = 1$, $\varepsilon = 10^{-10}$, $\alpha = 10^{-5}$. Quadratic convergence in time.

The first results are presented for the Fourier approximation in Table 5.1. Since the exact solution is a trigonometric polynomial of degree 1 there is no space discretization error. The only errors are due to temporal discretization and the optimization algorithm (CG-method). Thus this is a good example to analyze the error of the time discretization. The numerical result show a quadratic convergence rate, not only for $\alpha = 10^{-5}$ as depicted in Table 5.1, but for a much wider range of values for α.

Next we use finite elements, with piecewise linear Ansatz functions for space discretization. Tables 2 and 3 depict the spatial and temporal convergence rated respectively. In either case we observe quadratic convergence. To suppress temporal, respectively spatial, errors the discretization for either of them is chosen to be fine for the results in Tables 2 and 3. The dimension of the Krylov space is fixed to be $K = 10$.

In Table 4 we depict the results for Gautschi timestepping with SEM space discretization with integrated Legendre polynomials on one element. This demonstrates quadratic decrease of the error with respect to time discretization, for a wide range values for the control cost α.

In Table 5 we demonstrate spectral convergence of the spatial discretization error again the the SEM space. With only a few doublings of the polynomial degree, the error decreases rapidly until it reaches the level of the temporal discretization error.

Finally we give a numerical example in $2D$ where the exact solution is *not* known. Here all the specifications are as above, except for the fact that the control only acts on the subdomain $\omega = \left(0, \frac{\pi}{2}\right)^2$. Table 6 depicts the difference between the solution at discretization level l against the numerical solution with $N = 32$, and $l = 9$. It is shown to decay quadratically as desired.

5.2. Computational complexity

In this subsection we address the computational complexity of the Gautschi timestepping with Krylov methods. First we analyze the computational complexity of solving the wave equation with these methods.

N	l	error	iter
4	6	$1.50 \cdot 10^{-01}$	5
16	6	$1.11 \cdot 10^{-02}$	5
32	6	$2.74 \cdot 10^{-03}$	5
64	6	$6.23 \cdot 10^{-04}$	5

Table 2: FEM; $K = 10$, $\varepsilon = 10^{-10}$, $\alpha = 10^{-2}$. Influence of space discretization error.

l	error	iter
2	$2.32 \cdot 10^{-03}$	3
3	$5.52 \cdot 10^{-04}$	3
4	$1.29 \cdot 10^{-04}$	3

Table 3: FEM; $N = 2^6$, $K = 10$, $\varepsilon = 10^{-10}$, $\alpha = 10^{-1}$. Influence of time discretization error.

α	l	error	iter
$1 \cdot 10^{-0}$	4	$1.40 \cdot 10^{-04}$	3
$1 \cdot 10^{-0}$	5	$3.50 \cdot 10^{-05}$	3
$1 \cdot 10^{-0}$	6	$8.74 \cdot 10^{-06}$	3
$1 \cdot 10^{-1}$	4	$7.08 \cdot 10^{-04}$	4
$1 \cdot 10^{-1}$	5	$1.77 \cdot 10^{-04}$	4
$1 \cdot 10^{-1}$	6	$4.41 \cdot 10^{-05}$	4
$1 \cdot 10^{-2}$	4	$1.40 \cdot 10^{-03}$	5
$1 \cdot 10^{-2}$	5	$3.49 \cdot 10^{-04}$	5
$1 \cdot 10^{-2}$	6	$8.70 \cdot 10^{-05}$	5
$1 \cdot 10^{-5}$	4	$2.11 \cdot 10^{-03}$	70
$1 \cdot 10^{-5}$	5	$5.25 \cdot 10^{-04}$	69
$1 \cdot 10^{-5}$	6	$1.31 \cdot 10^{-04}$	54

Table 4: SEM; $N = 2^4$, $K = 25$, $\varepsilon = 10^{-10}$.

α	N	error	ln(error)	iter
0	4	$3.17 \cdot 10^{-02}$	-3.45	4
0	8	$1.18 \cdot 10^{-05}$	-11.35	3
0	16	$3.41 \cdot 10^{-08}$	-17.19	3
1	4	$1.80 \cdot 10^{-01}$	-1.72	6
1	8	$2.16 \cdot 10^{-05}$	-10.74	5
1	16	$1.72 \cdot 10^{-07}$	-15.57	4
2	4	$7.70 \cdot 10^{-01}$	-0.26	11
2	8	$1.43 \cdot 10^{-04}$	-8.85	10
2	16	$3.40 \cdot 10^{-07}$	-14.90	6

Table 5: SEM; $K = 25$, $L = 2^{10}$, $\varepsilon = 10^{-10}$. Influence of the space discretzation.

l	iter	relative error
2	26	$1.37 \cdot 10^{+00}$
3	24	$2.90 \cdot 10^{-01}$
4	24	$6.56 \cdot 10^{-02}$
5	14	$1.58 \cdot 10^{-02}$
6	14	$3.89 \cdot 10^{-03}$
7	14	$9.26 \cdot 10^{-04}$
8	14	$1.85 \cdot 10^{-04}$

Table 6: SEM; $K = 10$, $\varepsilon = 10^{-10}$, $\alpha = 10^{-2}$. Control on partial domain. Influence of the time discretzation.

L	N	rel. err in %	time [ms]
4	4	$1.824 \cdot 10^{-1}$	147
16	16	$6.138 \cdot 10^{-3}$	517
128	16	$9.594 \cdot 10^{-5}$	3350

Table 7: Hardware AMD Dual Core with 1GHz and 3GB RAM.

Since the Gautschi timestepping is a simple timestepping method we need $\mathcal{O}(L)$ evaluations involving matrix sine and cosine operators, where L denotes again the number of timesteps. Thus we have an overall cost of $\mathcal{O}(LF(N))$, where $F(N)$ denotes the computational cost which is needed to evaluate a trigonometric matrix operator depending on the size of the spatial dicretization N. The value $F(N)$ depends on the form of the space discretization, and the algorithm which evaluates the matrix operator. The range of $F(N)$ may vary from N, if the system matrix is diagonal, to N^3, for a general matrix evaluated with conventional methods like diagonalization.

Several methods for fast evaluation of matrix functions are available. We refer to the papers by Moler [13] and [14]. A good approach for trigonometric operators with sparse input matrices, which arise in the case of FEM or SEM matrices, are Krylov methods proposed by Hockbruck, Lubich and Grimm (see e.g. [6]). Nevertheless for systems where matrix inversions and multiplications become costly (order $\mathcal{O}(N^2)$) it is advised to compute the diagonalization of the matrix, because matrix function evaluation is cheap if the diagonalization is known. Note here that optimization requires many matrix function evaluations.

Using the CG algorithm for optimization the overall computational complexity is therefore of order $\mathcal{O}(L\,F(N)\sqrt{\kappa}))$ (see e.g. [16, p.37f]), where κ denotes the condition number of H_{LN}, and H_{LN} denotes the finite dimensional approximation in space and time of the operator Hessian H.

Since $\mathbb{H} = S^*S + \alpha I$, the condition number is of order $\mathcal{O}(\frac{1}{\alpha})$. Thus only a few CG iterations are needed if α is not too small, see Tables 2 and 4.

Table 7 shows the timings for the 1 dimensional wave equation with $y_0(x) = sin(x)$, $y_1(x) = 0$, $\Omega = (0, \pi)$, with homogeneous Dirichlet boundary data with trigonometric polynomials as spatial Ansatz functions. Since for this simple case the mass and stiffness matrices are diagonal, the evaluation of the matrix functions with Krylov methods only needs a Krylov space of one dimension. Hence we have $F(N) = \mathcal{O}(N)$, and the overall complexity is $\mathcal{O}(L\,N\,\sqrt{\kappa})$. This is confirmed by Table 7, where the increase in computing time is bounded by 4 from one row to the next.

Appendices

A. Vector Valued z-Transforms and Structural Properties of the Discrete Solution Operator

In this section we give the proof of Theorem 3.6. First we look for an alternative representation of S_L.

Lemma A.1 (Implicit form of the discrete solution operator)**.** *The discrete solution operator S_L is implicitly given by*

$$y[\ell] = S_L(f)[\ell] = \sum_{j=0}^{\ell} jGf[\ell - j] - GA \sum_{j=0}^{\ell} jy[\ell - j] - Gf[0]/2.$$

Proof. First we show that for $1 \leq l \leq \ell - 1$ the relation

$$y[\ell] = (l+1)y[\ell - l] - ky[\ell - l - 1] + \sum_{j=0}^{l} jGf[\ell - j] - GA \sum_{j=0}^{k} jy[\ell - j],$$

holds. We proceed by induction with respect to k. For $k = 1$ this is simply the Gautschi time step (9). It follows by induction hypothesis that

$$\begin{aligned}
y[\ell] =&(k+1)y[\ell-l] - ly[\ell-l-1] + \sum_{j=0}^{k} jGf[\ell-j] - GA\sum_{j=0}^{l} jy[\ell-j] \\
\overset{(9)}{=}&(k+1)(2y[n-k-1] - y[n-k-2] + Gf[n-k-1] - GAy[n-k-1]) \\
&- ky[\ell-l-1] + \sum_{j=0}^{l} jGf[\ell-j] - GA\sum_{j=0}^{l} jy[\ell-j] \\
=&(2l+2)y[\ell-l-1] - (l+1)y[\ell-l-2] + (l+1)Gf[\ell-l-1] - (l+1)GAy[\ell-l-1] \\
&- ly[\ell-l-1] + \sum_{j=0}^{l} jGf[\ell-j] - GA\sum_{j=0}^{l} jy[\ell-j] \\
=&(2l+2)y[\ell-l-1] - ky[\ell-l-1] - (l+1)y[\ell-l-2] \\
&+ (l+1)Gf[\ell-l-1] + \sum_{j=0}^{l} jGf[\ell-j] - (l+1)GAy[\ell-l-1] - GA\sum_{j=0}^{l} jy[\ell-j] \\
&= (l+2)y[\ell-l-1] - (l+1)y[\ell-l-2] + \sum_{j=0}^{l+1} jGf[\ell-j] - GA\sum_{j=0}^{l+1} jy[\ell-j].
\end{aligned}$$

For $l = \ell - 1$ we get the desired result, if we take into account that $y[0] = 0$ and $y[1] = Gf[0]/2$. □

We briefly recapitulate the concept of convolution sums.

Notation A.2. In this subsection we will write $fg := \langle f, g\rangle_H$ for the inner product of $f, g \in H$.

For $f, g \in \mathrm{Signa}_L(H)$ we define the convolution sum by

$$(f * g)[\ell] := \sum_{l=0}^{\ell} f[l]g[\ell-l].$$

Definition A.3. A (formal vector valued) power series $f(z) \in \mathbb{R}[[z]]$, with

$$f(z) = \sum_{\ell=0}^{\infty} f[\ell]z^{\ell}$$

is called the z-transform, (or generating function), of the signal, where $f[\ell] = 0$ for $\ell > L + 1$.

So in fact we consider polynomials.

Notation A.4. We write $f(z)$ for the z-Transform of a signal f, in analogy to the notation which is used in signal processing. Further let f be a signal, and $f(z)$ its z-Transform. For $\ell \in \mathbb{N}$ we write

$$[z^{\ell}]f(z) := f[\ell].$$

Remark A.5. As in the scalar case we have the relation

$$[z^\ell]f(z)g(z) = \sum_{l=0}^{n} f[l]g[\ell - l] := (f * g)[\ell].$$

Thus $(f * g)$ is a signal over $\{0, \dots, L\}$ to $\mathbb{R}$.

Lemma A.6 (The z-transform of the discrete solution operator)**.** *The z-transform of $(S_L f) = y(z)$ is given by*

$$y(z) = R(z)(Gk(z)f(z) - Gk(z)f[0]/2),$$

where $k[j] = j$, and the operator $R(z)$ is given as resolvent of GA

$$R(z) := R(GA, -k(z)) = (I + GAk(z))^{-1},$$

which is well defined for z sufficiently small.

Proof. From Lemma A.1 we know that

$$y[n] := S_L(f)[n] = \sum_{j=0}^{n} jGf[n-j] - GA\sum_{j=0}^{n} jy[n-j] - Gf[0]/2.$$

We multiply this equation by z^n, and sum over all n. So we get

$$y(z) = k(z)Gf(z) - GAk(z)y(z) - k(z)f[0]/2.$$

From Remark A.5 convolutions become products under the z-transform, and thus

$$(I + k(z)GA)y(z) = k(z)Gf(z) - k(z)f[0]/2.$$

Since $k(z) = \sum_{j=0}^{L+1} jz^j = z\frac{d}{dz}\sum_{j=0}^{L+1} z^j$, we have $\lim_{z\to 0} k(z) = 0$. Since we can choose z freely, we choose it sufficiently small, such that $\|k(z)GA\| < 1$. This is possible since the operator GA is bounded. So the operator R is well defined for such z, and we have the desired identity. □

Finally we give the proof of Theorem 3.6.

Proof of Theorem 3.6. Since the operators R and G are self adjoint we can change the order of multiplication. Recall that we assumed that $y[0] = 0$. Thus

it holds that

$$\begin{aligned}
L\,\langle S_L f, g\rangle_{D(L)} &= \sum_{k=0}^{L+1} y[k]g[k] - y[L+1]g[L+1]/2 - y[0]g[0]/2 = \\
&= \sum_{k=0}^{L+1} y[k]g^-[L+1-k] - y[L+1]g^-[0]/2 = \\
&= [z^{L+1}]y(z)g^-(z) - y(z)g^-[0]/2 \\
&\overset{\text{Lemma A.6}}{=} [z^{L+1}]R(z)(Gk(z)f(z) - Gk(z)f[0]/2)g^-(z) \\
&\quad - R(z)(Gk(z)f(z) - Gk(z)f[0]/2)g^-[0]/2 \\
&= [z^{L+1}]R(z)Gk(z)f(z)g^-(z) - R(z)k(z)Gf(z)g^-[0]/2 \\
&\quad - R(z)Gk(z)f[0]/2g^-(z) + R(z)Gk(z)f[0]/2g^-[0]/2 \\
&= [z^{L+1}]f(z)R(z)(Gk(z)g^-(z) - k(z)Gg^-[0]/2) \\
&\quad - \frac{1}{2}f[0]R(z)(Gk(z)g^-(z) - Gk(z)g^-[0]/2) \\
&\overset{\text{Lemma A.6}}{=} [z^{L+1}]f(z)(S_L g^-)(z) - f[0](S_L g^-)(z)/2 \\
&\quad = [z^{L+1}]f(z)x^-(z) - f[0]x^-(z)/2 \\
&= \sum_{k=0}^{L+1} x[k]f[k] - x[0]f[0]/2 - x[L+1]f[L+1]/2 \\
&= L\,\langle f, (S_L^* g)\rangle_{D(L)}\,.
\end{aligned}$$

This was to be shown. □

B. The Conjugate Gradient Method

Listing 1: Abstract CG

```
def abstract_cg(Q,b,x0, inner_product,
                tol = 10e-5, maxiter = None,
                nr_iterations = False):
    """
    This is an implementation of the CG algorithm on general
    inner product spaces.
    INPUT::
        Q  ... elliptic linear operator on Hilbert space H
        b  ... right hand side vector in H
        x0 ... starting vector in H
        inner_product ... inner product on H
        tol ... wanted accuracy
    OUTPUT::
        x ... approximated solution of Q x = b
        k ... needed iterations if nr_iterations is set True
    """
    #Prepare input
    dim = Q.shape[1]
```

```
if maxiter is None:
    maxiter = dim*10

x = zeros_like(x0)
x += x0
r = b - Q.matvec(x0)
p = r
residuum_old = inner_product(r,r)
#Test if x0 is already a solution
if sqrt(residuum_old) < tol:
    return x, 0

#Start iterating
for k in range(maxiter):
    q = Q.matvec(p)
    alpha = residuum_old/inner_product(p,q)
    x += alpha*p
    r -= alpha*q
    residuum_new = inner_product(r,r)
    if sqrt(residuum_new) < tol: #Stopping criterion
        return x, k

    p = r + residuum_new/residuum_old*p
    residuum_old = residuum_new
else:
    print "Maxiter reached!"
    return x, (k+1)
```

C. Rational Krylov Methods for the sinc Function

Here we consider rational Krylov methods for the sinc function with $\alpha = 1$, and input vector v which does not lie in $D(A)$. For evaluating the matrix function

$$f(\tau^2 A) = \tau \operatorname{sinc}(\tau\sqrt{A})$$

with rational Krylov algorithms Grimm and Hochbruck suggested in [6] that the parameters $\alpha = 0$ or $\alpha = 1/2$ should be used, since for applications the initial value y_1 for the velocity y_t in general does not in $D(A)$.

However, it may happen that $u \notin D(A)$, but a finite dimensional approximation $\tilde{u}$ of u is in $D(A)$, and then the expression $A\tilde{u}$ is defined while Au may not be. Therefore in a discretized setting the algorithm can still be executed with $\alpha = 1$.

Theorem C.1 (Missing case $\alpha = 1$)**.** *Let A be a positive definite and self adjoint operator, and $f = \operatorname{sinc}(\sqrt{x})$. Then the following error estimate holds for the approximation $y_m^\alpha(\tau) \approx y(\tau) = F(\tau^2 A)v$, where m denotes the dimension of the Krylov space:*

$$\left\| y_m^1 - \operatorname{sinc}(\tau\sqrt{A}) \right\| \leq \frac{C(\gamma)}{m} \tau^2 \left\| Av \right\|, \text{ for } v \in D(A),$$

and $C(\gamma)$ does not depend on A or v.

Proof. Set $\psi_\alpha(x) := \frac{F(x)-F(0)}{x^\alpha}$ and $g_\gamma^\alpha := \psi_\alpha\left(\left(\frac{1}{x}-1\right)\frac{1}{\gamma}\right)$. From [6] we know that it only remains to show that $E_{m-1,1}^{m-1}$ is bounded. This can be done with the help of the modules of continuity $\omega(g_\gamma^1;\delta)$ on $[0,1]$. The modulus is defined like in [17], namely

$$\omega(F,\delta) := \sup_{x,y\in[a,b],|x-y|\le\delta} |F(x)-F(y)|\,.$$

The easiest way to calculate the modulus of continuity is to use the mean value theorem. First we show that we can extend F to a continuous function onto the boundary of $[0,1]$. For easier reading we set $z := \sqrt{\left(\frac{1}{x}-1\right)\frac{1}{\gamma}}$. Then

$$\lim_{x\to 0^+} g_\gamma^1(x) = \lim_{x\to 0^+} -\frac{\sqrt{\frac{\frac{1}{x}-1}{\gamma}} - \sin\left(\sqrt{\frac{\frac{1}{x}-1}{\gamma}}\right)}{\left(\frac{\frac{1}{x}-1}{\gamma}\right)^{1+\frac{1}{2}}} = \lim_{z\to+\infty} \frac{\sin(z)}{z^3} - z^{-2} = 0$$

and

$$\lim_{x\to 1^-} g_\gamma^1(x) = \lim_{x\to 1^-} -\frac{\sqrt{\frac{\frac{1}{x}-1}{\gamma}} - \sin\left(\sqrt{\frac{\frac{1}{x}-1}{\gamma}}\right)}{\left(\frac{\frac{1}{x}-1}{\gamma}\right)^{1+\frac{1}{2}}} = \lim_{z\to 0^+} \frac{\sin(z)-z}{z^3} = -1/6.$$

Obviously g_γ^1 is differentiable in $(0,1)$. So the mean value theorem states that $g_\gamma^1(x) - g_\gamma^1(y) = \frac{dg_\gamma^1(x)}{d\xi}(x-y)$ for all $x,y\in[a,b]$ with a value $\xi\in(x,y)$. Thus it is sufficient to investigate the derivative of g_γ^1 at 0 and 1, which is given by (calculated with Sage)

$$(g_\gamma^1(x))' = -\frac{\frac{\cos\left(\sqrt{\frac{\frac{1}{x}-1}{\gamma}}\right)}{\gamma x^2\sqrt{\frac{\frac{1}{x}-1}{\gamma}}} - \frac{1}{\gamma x^2\sqrt{\frac{\frac{1}{x}-1}{\gamma}}}}{2\left(\frac{\frac{1}{x}-1}{\gamma}\right)^{\left(\frac{3}{2}\right)}} - \frac{3\left(\sqrt{\frac{\frac{1}{x}-1}{\gamma}} - \sin\left(\sqrt{\frac{\frac{1}{x}-1}{\gamma}}\right)\right)}{2\,\gamma x^2\left(\frac{\frac{1}{x}-1}{\gamma}\right)^{\left(\frac{5}{2}\right)}}. \quad (24)$$

Since (24) is continuous in $(0,1)$ we have to only care about the boundary values again. For the case $x=1$ we use our notation of z again. Then (24) reads as follows:

$$\frac{1}{2\gamma x^2}\frac{1}{z^4}\left(3\frac{\sin(z)-z}{z} - \cos(z) + 1\right) = \frac{1}{2\gamma x^2}(-\frac{1}{30} + o(1)),\ \text{for } x\to 1$$

If we take the limit $x\to 1$ we get $-\frac{1}{60\gamma}$. For $x\to 0^+$ we see that the limit does not exist, but the function is bounded in every neighborhood of 0. If we rewrite (24) and get

$$\limsup_{x\to^- 0} g_\gamma^1(x) = \limsup_{x\to 0} \frac{\gamma}{2(1-x^2)}\left|3\frac{\sin(z(x))}{z(x)} - \cos(z(x)) - 2\right| \le \frac{3\gamma}{2},$$

since sin and cos are bounded, and $z(x)\to\infty$ for $x\to 0$. So we can conclude that $\frac{dg_\gamma^1}{dx}$ is bounded on the whole interval, and this concludes the proof. □

Due to Theorem C.1 the algorithm converges faster for $\alpha = 1$, then for $\alpha = 0$ or $\alpha = 1/2$, which was investigated in [6]. This is due to the fact that the error depends quadratically on the time stepsize τ in the case $\alpha = 1$, instead of linearly for $\alpha = 1/2$, or even constant for $\alpha = 0$. But this requires sufficient regularity of the initial data, i.e. $u \in D(A^\alpha)$.

However, in finite dimensions the choice $\alpha = 1$ is admissible. Of course, this comes at the expense of a dimension-dependent error estimate, see (25) below, where we also give a numerical example.

One could also interpret this in the following way: If we make a finite dimensional approximation of a PDE in space, the approximation $\tilde{u}$ of u has higher regularity than the initial problem, and therefore the expression $A^\alpha \tilde{u}$ makes sense, although the expression $A^\alpha u$ may not be defined. We will demonstrate here with some numerical experiments, that using $\alpha = 1$ works well in practice.

We made numerical tests in one dimension with integrated Legendre Polynomials, for the Heavyside function $H : [-1, 1] \to \mathbb{R}$, given by

$$H(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}.$$

For $A = \Delta$ it is obvious that $H \notin D(A)$. We denote with $\tau = T/L$ the time parameter, and K is the dimension of the Krylov space. The function H is approximated with help of interpolation.

We compare our calculated result with the classical evaluation methods for matrix functions of the `Scipy` package for Python. In `Scipy` the unitary diagonalization of A will be computed, i.e. $A = QDQ^*$ and the function will be evaluated by the formula $f(A) = Qf(D)Q^*$.

For comparision we choose the values for N, τ, and K, as

$$N = 2^4, 2^6, 2^8,\ \tau = 2^{-4},\ 2^{-6},\ 2^{-8} \text{ and } K = 1, 5, 10.$$

As we can see in Table 8 the error strongly depends on the size of the space discretization. We can see that it is crucial to choose a sufficiently large Krylov space and a sufficiently small stepsize τ. The results suggest, that τ should satisfy

$$\tau < 1/N,$$

such that the error is small enough, and convergence of the CG method is provided. This is reasonable since in one dimension the eigenvalues of the Laplacian Δ, are $\lambda_k = \pi^2 k^2$, thus we can expect that the largest eigenvalue of the stiffness matrix A is $\mathcal{O}(N^2)$, and therefore the error of the Krylov approximation y_K^1 is due to Theorem C.1

$$\text{error} = \left\| y_K^1 - \operatorname{sinc}(\tau\sqrt{A})v \right\|_2 \leq \frac{C(\gamma)}{K}\tau^2 \left\| Av \right\|_2 \leq \frac{\tilde{C}(\gamma)}{K}\tau^2 N^2 \left\| v \right\|_2. \tag{25}$$

Although the error depends linearly on the Krylov space dimension K, a reasonable choice of the size of the parameter K can make a significant difference as it can be seen in Table 9. We tested with two stepsizes $\tau = 2^{-8}, 2^{-10}$. It appears that a linear decay with respect to K is far too pessimistic compared to the numerical results, which suggest that the error decays super linearly or even exponentially.

N	τ	K	error
2^4	2^{-4}	10	$5.806 \cdot 10^{-15}$
2^4	2^{-6}	10	$5.263 \cdot 10^{-16}$
2^4	2^{-8}	10	$6.021 \cdot 10^{-16}$
2^4	2^{-10}	10	$3.999 \cdot 10^{-16}$
2^4	2^{-12}	10	$4.844 \cdot 10^{-16}$
2^6	2^{-4}	10	$1.066 \cdot 10^{-02}$
2^6	2^{-6}	10	$3.152 \cdot 10^{-12}$
2^6	2^{-8}	10	$4.659 \cdot 10^{-14}$
2^6	2^{-10}	10	$3.893 \cdot 10^{-15}$
2^6	2^{-12}	10	$3.713 \cdot 10^{-15}$
2^8	2^{-4}	10	$2.399 \cdot 10^{-02}$
2^8	2^{-6}	10	$1.901 \cdot 10^{-02}$
2^8	2^{-8}	10	$3.229 \cdot 10^{-03}$
2^8	2^{-10}	10	$1.360 \cdot 10^{-12}$
2^8	2^{-12}	10	$3.222 \cdot 10^{-13}$

Table 8: Relative errors of the evaluation of the function $\mathrm{sinc}(\tau\sqrt{A})$ with rational Krylov methods with parameter $\alpha = 1$.

N	τ	K	error
2^8	2^{-8}	1	$1.146 \cdot 10^{+00}$
2^8	2^{-8}	10	$3.229 \cdot 10^{-03}$
2^8	2^{-8}	20	$4.440 \cdot 10^{-13}$
2^8	2^{-8}	30	$1.733 \cdot 10^{-06}$
2^8	2^{-8}	40	$9.740 \cdot 10^{-14}$
2^8	2^{-8}	50	$1.139 \cdot 10^{-13}$
2^8	2^{-10}	1	$8.024 \cdot 10^{-01}$
2^8	2^{-10}	10	$1.360 \cdot 10^{-12}$
2^8	2^{-10}	20	$1.105 \cdot 10^{-12}$
2^8	2^{-10}	30	$9.704 \cdot 10^{-14}$
2^8	2^{-10}	40	$8.977 \cdot 10^{-14}$
2^8	2^{-10}	50	$7.832 \cdot 10^{-14}$

Table 9: Relative errors of the evaluation of the function $\mathrm{sinc}(\tau\sqrt{A})$ with rational Krylov methods with parameter $\alpha = 1$ for different K.

[1] Arendt, W., Batty, C., Hieber, M., Neubrander, F., 2011. Vector-valued Laplace transforms and Cauchy problems. Monographs in Mathematics 96.

[2] Boyd, J., 2001. Chebyshev and Fourier Spectral Methods. Dover Publications, New York.

[3] Du, X., Haber, E., Karampataki, M., Szyld, D. B., et al., 2008. Varying iteration accuracy using inexact conjugate gradients in control problems governed by pde's. Tech. rep., Tech. Report 08-06-27, Department of Mathematics, Temple University, Philadelphia.

[4] Grimm, V., 2005. A note on the Gautschi-type method for oscillatory second-order differential equations. Numerische Mathematik 102 (1), 61–66.

[5] Grimm, V., 2005. On error bounds for the gautschi-type exponential integrator applied to oscillatory second-order differential equations. Numerische Mathematik 100 (1), 71–89.
URL `http://dx.doi.org/10.1007/s00211-005-0583-8`

[6] Grimm, V., Hochbruck, M., 2008. Rational approximation to trigonometric operators. BIT Numerical Mathematics 48 (2), 215–229.

[7] Hanke-Bourgeois, M., 2006. Grundlagen der Numerischen Mathematik und des wissenschaftlichen Rechnens. Vieweg+Teubner Verlag, Wiesbaden.

[8] Hochbruck, M., Lubich, C., 1999. A Gautschi-type method for oscillatory second-order differential equations. Numerische Mathematik 83 (3), 403–426.
URL `http://dx.doi.org/10.1007/s002110050456`

[9] Kröner, A., 2011. Numerical methods for control of second order hyperbolic equations. Ph.D. thesis, Technische Universität München.

[10] Kröner, A., Kunisch, K., Vexler, B., 2011. Semismooth newton methods for optimal control of the wave equation with control constraints. SIAM Journal on Control and Optimization 49 (2), 830–858.
URL http://epubs.siam.org/doi/abs/10.1137/090766541

[11] Lintner, M., 2002. Lösung der 2d wellengleichung mittels hierarchischer matrizen. Ph.D. thesis, Technische Universität München.

[12] Luenberger, D., 1997. Optimization by Vector Space Methods. Wiley-Interscience, New York.

[13] Moler, C., Van Loan, C., 1978. Nineteen dubious ways to compute the exponential of a matrix. SIAM Review 20 (4), 801–836.

[14] Moler, C., Van Loan, C., 2003. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Review 45 (1), 3–49.

[15] Reiterer, S., 2013. Optimal Receding Horizon Control of the Wave Equation with help of Trigonometric Operators and SEM Methods. Ph.D. thesis, Universität Graz.

[16] Shewchuk, J. R., 1994. An introduction to the conjugate gradient method without the agonizing pain. Tech. rep., Pittsburgh, PA, USA.

[17] Timan, A., 1994. Theory of approximation of Functions of a Real Variable. Dover Publications, New York.

[18] Tröltzsch, F., 2009. Optimale Steuerung partieller Differentialgleichungen: Theorie, Verfahren und Anwendungen. Vieweg+Teubner, Wiesbaden.

[19] Wloka, J., 1992. Partial Differential Equations (English translation). Cambridge University Press, Cambridge.