Gabriel Pichlbauer, BSc BSc

Numerical Analysis of Open- and Closed-Loop Control Strategies for the Heat Equation

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme:
Mathematics

submitted to

Graz University of Technology

Supervisor

Univ.-Prof. Dipl.-Math. Dr.rer.nat. Olaf Steinbach
Institute of Applied Mathematics

Co-Supervisor

Dipl.-Ing. Dr.techn. Richard Löscher
Institute of Applied Mathematics

Graz, April, 2026

Abstract

The aim of this thesis is to analyze several optimal control strategies for parabolic differential equations, specifically the heat equation. First of all, the open-loop regularization from [18], its adjoint formulation, and the interpolated formulation from [23] are compared to each other. Subsequently, the open-loop regularization from [18] is also compared to its closed-loop regularization counterpart. For the sake of simplicity, the optimal control problem for the heat equation will be reduced to a control problem constrained by ordinary differential equations by means of an eigendecomposition of the state, as it is done in [17]. Finally we present numerical results, comparing the open-loop and closed-loop strategies derived in this work.

Kurzfassung

Das Ziel dieser Arbeit ist es verschiedene optimale Kontrollstrategien für parabolische Differentialgleichungen, insbesondere der Wärmeleitgleichung, zu analysieren. Zuerst wird die Open-Loop Regularisierung aus [18], dessen adjungierte Formulierung, und die interpolierte Formulierung aus [23] miteinander verglichen. Anschließend wird die Open-Loop Regularisierung aus [18] mit der entsprechenden Closed-Loop Regularisierung verglichen. Zur Vereinfachung wird das optimale Kontrollproblem für die Wärmeleitgleichung mithilfe einer Eigenwertzerlegung zu einem durch eine gewöhnliche Differentialgleichung beschränkten Kontrollproblem reduziert, so wie in [17]. Schließlich werden numerische Resultate präsentiert, welche die in der vorliegenden Arbeit hergeleiteten Open-Loop und Closed-Loop Strategien miteinander vergleichen.

Acknowledgements

I would like to thank Prof. Dr. Steinbach for introducing me to this interesting topic and the help throughout the thesis, allowing me to gain some insight into the broad field of optimal control theory. Further, I also want to especially thank Dr. Richard Löscher, who provided me with a lot of guidance in the final stages of this work. Finally, thank you to my family and my partner Carmen for always supporting me throughout my time at the university.

Contents

Introduction

The main goal of this thesis is to analyse and compare different approaches for the distributed optimal control of tracking-type for the heat equation. A variety of results are known for parabolic optimal control with a square-integrable control, see for example [20] or [6]. In comparison, in the more recent work [18], the authors have investigated the possibility of using the weaker notion of energy regularization. Using this approach, they demonstrated that the optimal state exhibits sharper edges in discontinuous regions of the desired target state, compared to the usual square-integrable control approach. Their optimal control problem is based on the primal formulation of the heat equation from [29], but the concept of energy regularization also facilitates to use the dual formulation or the interpolated formulation discussed in this work. Though the low regularity of the state in the dual formulation makes it challenging to use this method in practice, the optimal control problem based on the interpolated formulation was already investigated in [23]. They introduced a solver on the discrete level with almost linear complexity, and showed a preferable relation between the mesh size and the regularization parameter in contrast to the primal formulation, with possible generalisations to non-linear problems. Aside from those benefits, one drawback of these works on energy regularisation is that only open-loop regularization was considered, which can be insufficient in practical applications accompanied with model uncertainties or random noise. Those uncertainties can be particularly challenging when working with non-linear equations, see for instance [17]. Thus, this work is dedicated to analysing theoretical aspects of all three optimal control problems associated with the primal, the dual, and the interpolated formulation of the heat equation. Further, we derive a closed-loop control based on the primal formulation. For the control problem based on the interpolated formulation, it is unfortunately not yet clear how to derive similar closed-loop controls, but we will describe how to incorporate non-homogeneous initial values, such that at least a form of Model Predictive Control can be realized in future works, see for example [15].

In what follows we give an overview about the contents of each chapter. Chapter 1 is devoted to the preliminaries, focussing mainly on the relevant function spaces and related theorems from functional analysis. In Chapter 2 we discuss the theoretical background of open- and closed-loop regularization techniques, with a particular emphasis on the Hamilton–Jacobi–Bellman equation and Pontryagin's Maximum Principle. In Chapter 3, the primal, the dual and the interpolated variational formulation of the heat equation are introduced, as well as a decoupling into a sequence of easier problems, similar to [17]. In Chapter 4 we analyse the three optimal control problems, focussing on well-posedness and error estimates. Further, a closed-loop regularization for the primal formulation is derived. Finally, in Chapter 5 we discuss the relations between the regularization parameter and the mesh size, and the results from the numerical experiments.

1 Preliminaries

In this section we introduce all the function spaces, the related notation, as well as the main results from functional analysis relevant for this work.

1.1 Hilbert Spaces

Recall that a real-valued Hilbert space X is a vector space over $\mathbb{R}$, equipped with an inner product $\langle \cdot, \cdot \rangle_X$, such that X is complete with respect to the norm $\|\cdot\|_X = \sqrt{\langle \cdot, \cdot \rangle_X}$. For such a Hilbert space, there holds the Cauchy–Schwarz inequality

$$\langle x, y \rangle_X \leq \|x\|_X \|y\|_X \tag{1.1}$$

for all $x, y \in X$, which we will frequently use throughout this work. The dual space of X, consisting of all linear and bounded functionals $f : X \to \mathbb{R}$, is denoted by X^*. The evaluation of a functional $f \in X^*$ in $x \in X$ is usually denoted either by $f(x)$ or $\langle f, x \rangle_{X^* \times X}$. Further, the dual space X^* is equipped with the norm

$$\|f\|_{X^*} = \sup_{0 \neq x \in X} \frac{\langle f, x \rangle_{X^* \times X}}{\|x\|_X}.$$

From the Theorem of Fréchet-Riesz, see [7, Theorem 5.5], we know that every $f \in X^*$ can be identified with an unique $x_f \in X$ via

$$f(y) = \langle x_f, y \rangle_X, \quad \forall y \in X,$$

and vice versa. The theorem also states that the mapping $J : X^* \to X, f \mapsto x_f$, which is called the Fréchet–Riesz map, satisfies $\|Jf\|_X = \|f\|_{X^*}$ for all $f \in X^*$.

For a linear operator $T : X \to Y$ with real-valued Hilbert spaces X, Y the operator norm is denoted by $\|T\|_{X \to Y}$. The following results provide powerful tools for analyzing the solvability of operator equations

$$Bx = f \quad \text{in } Y^*, \tag{1.2}$$

where $x \in X$, $f \in Y^*$ and $B : X \to Y^*$. First note that due to the Theorem of Fréchet–Riesz every linear and bounded operator $B : X \to Y^*$ can be uniquely identified with a bounded bilinear form $b : X \times Y \to \mathbb{R}$ via

$$b(x, y) = \langle Bx, y \rangle_{Y^* \times Y}, \quad \forall x \in X, \forall y \in Y,$$

and their norms agree, meaning that

$$\|B\|_{X\to Y^*} = \sup_{0\neq x\in X}\sup_{0\neq y\in Y}\frac{\langle Bx,y\rangle_{Y^*\times Y}}{\|x\|_X\|y\|_Y} = \sup_{0\neq x\in X}\sup_{0\neq y\in Y}\frac{b(x,y)}{\|x\|_X\|y\|_Y}. \tag{1.3}$$

Thus we can always interpret the operator equation (1.2) as an equivalent variational formulation of finding $x \in X$, such that

$$b(x,y) = f(y) \tag{1.4}$$

holds for all $y \in Y$. The next theorem characterizes the unique solvability of this variational formulation (1.4).

Theorem 1.1 (Banach–Nečas–Babuška, [2, Satz 4.27])**.** *Let X, Y be real-valued Hilbert spaces and $b : X \times Y \to \mathbb{R}$ be a bounded bilinear form. Then the following statements are equivalent:*

(i) *For every $f \in Y^*$ there exists a unique $x \in X$ satisfying*

$$b(x,y) = f(y)$$

for all $y \in Y$.

(ii) *There exists a constant $c_1^B > 0$ such that the so called inf-sup condition*

$$c_1^B \leq \inf_{0\neq x\in X}\sup_{0\neq y\in Y}\frac{b(x,y)}{\|x\|_X\|y\|_Y} \tag{1.5}$$

holds. Further, for every $y \in Y \setminus \{0\}$ there exists $x_y \in X$ such that $b(x_y, y) \neq 0$.

Moreover, the unique solution satisfies the stability estimate

$$\|x\|_X \leq \tfrac{1}{c_1^B}\|f\|_{Y^*}.$$

The inf-sup condition is equivalent to B having closed range and being an injective operator, whereas the latter condition in (ii) just translates to the adjoint operator of B being injective. Note that the boundedness and inf-sup stability of the bilinear form b is equivalent to the existence of constants $0 < c_1^B \leq c_2^B < \infty$ such that

$$c_1^B\|x\|_X \leq \|Bx\|_{Y^*} \leq c_2^B\|x\|_X, \qquad \forall x \in X. \tag{1.6}$$

In the case of finite-dimensional Hilbert spaces X, Y, the operator B can be interpreted as a matrix. If B satisfies (ii), it is injective, and its adjoint is also injective, which is only possible if X and Y have the same dimensions. Thus B is a square matrix, for which it is known that B is injective if and only if it is surjective. In this case there is also a simpler characterization of the inf-sup constant as the lowest singular value of a matrix.

Lemma 1.2 ([5, Proposition 3.4.5])**.** *Let X, Y be finite-dimensional real-valued Hilbert spaces, and $A : Y \to Y^*$, $D : X \to X^*$ be linear, bounded, self-adjoint and elliptic operators such that*

$$\|x\|_X^2 = \langle Dx, x\rangle_{X^*\times X}, \qquad \|y\|_Y^2 = \langle Ay, y\rangle_{Y^*\times Y}$$

holds for all $x \in X, y \in Y$. Further, let $B : X \to Y^*$ be a linear and bounded operator, and assume that

$$0 < c_1^B = \inf_{0\neq x\in X} \sup_{0\neq y\in Y} \frac{\langle Bx, y\rangle_{Y^*\times Y}}{\|x\|_Y \|y\|_Y}, \qquad c_2^B = \sup_{0\neq x\in X} \sup_{0\neq y\in Y} \frac{\langle Bx, y\rangle_{Y^*\times Y}}{\|x\|_X \|y\|_Y} < \infty.$$

Then c_1^B and c_2^B are equal to the square roots of the minimal and maximal eigenvalue from the generalised eigenvalue problem

$$B^\top A^{-1} Bx = \mu Dx$$

respectively. If $X = \mathbb{R}^n, Y = \mathbb{R}^m$, equipped with the usual Euclidean norm, the constants c_1^B and c_2^B simply represent the minimal and maximal singular value of B respectively.

Remark 1.3. *Let H, V denote real-valued Hilbert spaces, with $V \subset H$ being compactly embedded. If either $X = H$, $Y = V$ or $X = V$, $Y = H$, the result from Lemma 1.2 holds true without assuming finite dimensionality of X and Y. The proof works the same way, using the spectral theory for compact operators described in [7, Chapter 6] instead of the spectral theory for matrices. Since there are in general infinitely many eigenvalues $(\mu_k)_{k\in N_0} \subset (0, \infty)$, the constants c_1^B and c_2^B are given by the infimum and the supremum of those eigenvalues respectively, instead of the minimal and the maximal eigenvalue in the matrix case. In particular we note that the respective eigenvalues are bounded from below and from above if the operator $B : X \to Y^*$ is bijective.*

So assuming that the bilinear form $b : X \times Y \to \mathbb{R}$ related to a linear operator $B : X \to Y^*$ satisfies condition (ii) from Theorem 1.1, we know that B is invertible, and that the respective operator norm can be bounded by

$$\|x\|_X = \|B^{-1} f\|_X \leq \tfrac{1}{c_1^B} \|f\|_{Y^*}.$$

Throughout this work we will also encounter operator equations $Ax = f$ involving an operator $A : X \to X^*$, such that the related bilinear form $a : X \times X \to \mathbb{R}$ has a domain depending only on a single Hilbert space. The following results characterize solvability of operator equations involving such mappings A, and generalize the Cauchy–Schwarz inequality for inner products induced by such operators.

Theorem 1.4 (Lax–Milgram, [2, Satz 4.23])**.** *Let X be a real-valued Hilbert space and $a : X \times X \to \mathbb{R}$ be a bounded bilinear form. Further, assume that $a(\cdot, \cdot)$ is elliptic, meaning there exists some $c_1^A > 0$ such that*

$$a(x, x) \geq c_1^A \|x\|_X^2$$

holds for all $x \in X$. Then for every $f \in X^*$ there exists a unique $x \in X$, such that

$$a(x, y) = f(y)$$

holds for all $y \in X$.

Corollary 1.5 (Generalised Cauchy–Schwarz Inequality)**.** *Let $A : X \to X^*$ be a linear, bounded, self-adjoint and elliptic operator. Then $\langle \cdot, \cdot \rangle_A = \langle A\cdot, \cdot \rangle_X$ defines an inner product, and $\|\cdot\|_A = \sqrt{\langle \cdot, \cdot \rangle_A}$ defines a norm. Therefore it holds that*

$$\langle x_1, x_2 \rangle_A \leq \|x_1\|_A \|x_2\|_A, \qquad \forall x_1, x_2 \in X.$$

1.2 Sobolev Spaces

For a given $T > 0$, we denote by $L^2(0,T)$ the space of square-integrable functions on the interval $(0,T)$. For a detailed introduction into L^p-spaces consider [7, Chapter 4]. Following [31], we denote by $H^s(0,T) \subset L^2(0,T)$ the Sobolev space of fractional order $s \in [0, \infty)$. For $s = 1$ we define the subspaces

$$H^1_{0,}(0,T) = \{v \in H^1(0,T) \mid v(0) = 0\},$$
$$H^1_{,0}(0,T) = \{w \in H^1(0,T) \mid w(T) = 0\}$$

and equip them with the norm $\|v\|_{H^1_{0,}(0,T)} = \|\dot{v}\|_{L^2(0,T)}$ and $\|w\|_{H^1_{,0}(0,T)} = \|\dot{w}\|_{L^2(0,T)}$. Then we introduce the two interpolation spaces $H^s_{0,}(0,T) = [L^2(0,T), H^1_{0,}(0,T)]_s$ and $H^s_{,0}(0,T) = [L^2(0,T), H^1_{,0}(0,T)]_s$ for $s \in [0,1]$. For a detailed introduction to interpolation theory see [21], but throughout this work we will only use a characterization of those spaces via Fourier series as it is done in [31]. Denoting by $\mu_k = \frac{\pi}{2} + k\pi$, every function $v \in H^s_{0,}(0,T)$ and $w \in H^s_{,0}(0,T)$ with $s \in [0,1]$ can be written as

$$v(t) = \sum_{k=0}^{\infty} v_k \sin\left(\mu_k \tfrac{t}{T}\right), \qquad w(t) = \sum_{k=0}^{\infty} w_k \cos\left(\mu_k \tfrac{t}{T}\right). \tag{1.7}$$

In particular, every function $q \in L^2(0,T)$ can be written in both the sine and the cosine basis. Based on (1.7), the norms in $H^s_{0,}(0,T)$ and $H^s_{,0}(0,T)$ are defined by

$$\|v\|^2_{H^s_{0,}(0,T)} = \frac{1}{2T^{2s-1}} \sum_{k=0}^{\infty} \mu_k^{2s} v_k^2, \qquad \|w\|^2_{H^s_{,0}(0,T)} = \frac{1}{2T^{2s-1}} \sum_{k=0}^{\infty} \mu_k^{2s} w_k^2.$$

The respective dual spaces are denoted by $[H^s_{0,}(0,T)]^*$ and $[H^s_{,0}(0,T)]^*$, and the duality product is denoted by $\langle \cdot, \cdot \rangle_{(0,T)}$.

In the context of fractional order Sobolev spaces we also introduce the modified Hilbert Transformation $\mathcal{H}_T : H^s_{0,}(0,T) \to H^s_{,0}(0,T)$ as in [31, §3.4.2] and [24]. For $v \in H^s_{0,}(0,T)$ and $w \in H^s_{,0}(0,T)$ as in Equation (1.7), the modified Hilbert transformation and its inverse are defined by

$$\left(\mathcal{H}_T v\right)(t) = \sum_{k=0}^{\infty} v_k \cos(\mu_k \tfrac{t}{T}), \qquad \left(\mathcal{H}_T^{-1} w\right)(t) = \sum_{k=0}^{\infty} w_k \sin(\mu_k \tfrac{t}{T}). \tag{1.8}$$

Note that $\mathcal{H}_T : H^s_{0,}(0,T) \to H^s_{,0}(0,T)$ is an isometric isomorphism for all $s \in [0,1]$. For $s = \frac{1}{2}$, the modified Hilbert transformation also allows to write the inner products associated with $H^{1/2}_{0,}(0,T)$ and $H^{1/2}_{,0}(0,T)$ via

$$\langle u, v\rangle_{H^{1/2}_{0,}(0,T)} = \langle \dot{u}, \mathcal{H}_T v\rangle_{(0,T)}, \quad \langle w, z\rangle_{H^{1/2}_{,0}(0,T)} = -\langle \dot{w}, \mathcal{H}_T^{-1} z\rangle_{(0,T)}.$$

Next, we introduce the relevant Sobolev spaces for spatial problems. Let $\Omega \subset \mathbb{R}^d$ be an open subset with $d \geq 1$. Then as defined in [28], we denote by $H^1(\Omega)$ the space of all functions $v \in L^2(\Omega)$ with weak partial derivatives $\partial_{x_j} v \in L^2(\Omega)$ for all $j \in \{1, ..., d\}$, equipped with the norm

$$\|v\|_{H^1(\Omega)} = \sqrt{\|v\|^2_{L^2(\Omega)} + \|\nabla_x v\|^2_{L^2(\Omega)}}.$$

Then $H^1_0(\Omega)$ is the closure of the smooth functions with compact support, with respect to the norm $\|\cdot\|_{H^1(\Omega)}$. Note that instead of the standard norm, we will equip $H^1_0(\Omega)$ with the equivalent norm

$$\|v\|_{H^1_0(\Omega)} = \|\nabla_x v\|_{L^2(\Omega)}.$$

The dual space of $H^1_0(\Omega)$ will be denoted by $H^{-1}(\Omega)$, and the related duality product by $\langle\cdot,\cdot\rangle_\Omega$. As shown in [28, Chapter 4], the Laplace operator $-\Delta_x$ defines an isomorphism $-\Delta_x : H^1_0(\Omega) \to H^{-1}(\Omega)$. Further, it is well known that there exists an orthonormal basis $\{\phi_k\}_{k\in\mathbb{N}} \subset H^1_0(\Omega)$ of $L^2(\Omega)$ consisting of eigenfunctions of $-\Delta_x$, which is also an orthogonal basis with respect to the inner product $\langle\cdot,\cdot\rangle_{H^1_0(\Omega)}$. We denote by λ_k the respective eigenvalues, with $0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_k \overset{k\to\infty}{\longrightarrow} \infty$.

Lemma 1.6 ([25, Chapter 2] and [10, §5.7])**.** *Let $v \in H^1_0(\Omega)$, and write*

$$v(x) = \sum_{k=1}^{\infty} v_k \phi_k(x),$$

then it holds that

$$\|v\|^2_{L^2(\Omega)} = \sum_{k=1}^{\infty} v_k^2, \qquad \|v\|^2_{H^1_0(\Omega)} = \sum_{k=1}^{\infty} \lambda_k v_k^2.$$

Further, every $f \in H^{-1}(\Omega)$ can be formally written as

$$f(x) = \sum_{k=1}^{\infty} f_k \phi_k(x),$$

where $f_k = \langle f, \phi_k\rangle_\Omega$. Then it follows that

$$\langle f, v\rangle_\Omega = \sum_{k=1}^{\infty} f_k v_k, \qquad \|f\|_{H^{-1}(\Omega)} = \sum_{k=1}^{\infty} \tfrac{1}{\lambda_k} f_k^2.$$

1.3 Bochner Spaces

We denote by $Q = \Omega \times (0,T)$ the space-time cylinder associated to a domain Ω and an interval $(0,T)$, and by $L^2(Q)$ all functions which are square-integrable with respect to the domain Q. In this section, we will only cover the relevant results about Bochner spaces, which will be required throughout this work, for a more detailed introduction consider [10, §5.9.2] or [26, Chapter 10.1]. In later sections of this work we will frequently deal with functions $y \in L^2(0,T;H_0^1(\Omega))$ with a weak derivative $\partial_t y \in L^2(0,T;H^{-1}(\Omega))$. Using the results from Lemma 1.6, and interpreting $A = -\Delta_x$ as a map from $L^2(0,T;H_0^1(\Omega))$ to $L^2(0,T;H^{-1}(\Omega))$, the norms in those spaces are given as

$$\begin{aligned} \|y\|_{L^2(0,T;H_0^1(\Omega))} &= \|\nabla_x y\|_{L^2(Q)} = \sqrt{\langle Ay, y\rangle_Q} && \forall y \in L^2(0,T;H_0^1(\Omega)), \\ \|f\|_{L^2(0,T;H^{-1}(\Omega))} &= \sqrt{\langle f, A^{-1}f\rangle_Q} && \forall f \in L^2(0,T;H^{-1}(\Omega)). \end{aligned}$$

According to [26, Proposition 10.5], we know $[L^2(0,T;H_0^1(\Omega))]^* = L^2(0,T;H^{-1}(\Omega))$. The next result states that such functions $y \in L^2(0,T;H_0^1(\Omega))$ with weak derivatives in the dual space are continuous in a certain sense, and admit a well-defined trace at $t = 0$ and $t = T$.

Lemma 1.7 ([10, §5.9.2, Theorem 3])**.** *Let $y \in L^2(0,T;H_0^1(\Omega))$ be given, and assume that $\partial_t y \in L^2(0,T;H^{-1}(\Omega))$. Then it holds that $y \in C([0,T];L^2(\Omega))$, and that the map $t \mapsto \|y(t)\|^2_{L^2(\Omega)}$ is absolutely continuous with*

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\|y(t)\|^2_{L^2(\Omega)} = 2\langle \partial_t y(t), y(t)\rangle_\Omega. \tag{1.9}$$

Due to this Lemma, we know that the spaces

$$\begin{aligned} &L^2(0,T;H_0^1(\Omega)) \cap H^1_{0,}(0,T;H^{-1}(\Omega)) \\ &\qquad = \{v \in L^2(0,T;H_0^1(\Omega)) \mid \partial_t v \in L^2(0,T;H^{-1}(\Omega)),\ v(0) = 0 \text{ in } L^2(\Omega)\}, \\ &L^2(0,T;H_0^1(\Omega)) \cap H^1_{,0}(0,T;H^{-1}(\Omega)) \\ &\qquad = \{w \in L^2(0,T;H_0^1(\Omega)) \mid \partial_t w \in L^2(0,T;H^{-1}(\Omega)),\ w(T) = 0 \text{ in } L^2(\Omega)\} \end{aligned}$$

are well-defined. Another consequence of this Lemma are the following two inequalities.

Corollary 1.8. *Given the two functions $y \in L^2(0,T;H_0^1(\Omega)) \cap H^1_{0,}(0,T;H^{-1}(\Omega))$ and $z \in L^2(0,T;H_0^1(\Omega)) \cap H^1_{,0}(0,T;H^{-1}(\Omega))$, it holds that*

$$\langle \partial_t y, y\rangle_Q \geq 0, \qquad \langle \partial_t z, z\rangle_Q \leq 0.$$

Proof: Integrating over the Equation (1.9) we obtain

$$\langle \partial_t y, y\rangle_Q = \int_0^T \langle \dot{y}(t), y(t)\rangle_\Omega \,\mathrm{d}t = \tfrac{1}{2}\left(\|y(T)\|^2_{L^2(\Omega)} - \|y(0)\|^2_{L^2(\Omega)}\right).$$

Using the initial conditions immediately implies the first estimate. The second estimate follows analogously using the respective terminal condition. □

As in [31], we also introduce $H^{1/2}_{0,}(0,T;L^2(\Omega))$ as the interpolation space between $L^2(0,T;L^2(\Omega))$ and $H^1_{0,}(0,T;L^2(\Omega))$, and similarly define $H^{1/2}_{,0}(0,T;L^2(\Omega))$. Then for $u \in H^{1/2}_{0,}(0,T;L^2(\Omega))$ and $v \in H^{1/2}_{,0}(0,T;L^2(\Omega))$ we know that $\partial_t u \in [H^{1/2}_{,0}(0,T;L^2(\Omega))]^*$ and $\partial_t v \in [H^{1/2}_{0,}(0,T;L^2(\Omega))]^*$. Based on those spaces we further define the anisotropic Sobolev spaces

$$H^{1,1/2}_{0;0,}(Q) = L^2(0,T;H^1_0(\Omega)) \cap H^{1/2}_{0,}(0,T;L^2(\Omega)),$$
$$H^{1,1/2}_{0;,0}(Q) = L^2(0,T;H^1_0(\Omega)) \cap H^{1/2}_{,0}(0,T;L^2(\Omega)),$$

which are equipped with the norms

$$\|v\|^2_{H^{1,1/2}_{0;0,}(Q)} = \langle \partial_t v, \mathcal{H}_T v\rangle_Q + \|\nabla_x v\|^2_{L^2(Q)}, \quad \|w\|^2_{H^{1,1/2}_{0;0,}(Q)} = -\langle \partial_t w, \mathcal{H}_T^{-1} w\rangle_Q + \|\nabla_x w\|^2_{L^2(Q)}.$$

Next we state a quite general result about isomorphic operators between Bochner-spaces. Let $Y \subset L^2(0,T;H^1_0(\Omega))$ be a Hilbert space, with an inner product induced by an operator $D : Y \to Y^*$ in the sense that $\|\cdot\|_Y = \langle D\cdot,\cdot\rangle_Q^{1/2}$. Denoting by $\{\phi_k\}_{k=1}^\infty \subset H^1_0(\Omega)$ the orthonormal basis of $L^2(\Omega)$ from Section 1.2, we know that every $y \in Y$ can be written in the form

$$y(x,t) = \sum_{k=1}^\infty y_k(t)\phi_k(x),$$

with suitable coefficients $y_k \in L^2(0,T)$. At this point we assume that

$$\langle D(v\phi_k), w\phi_\ell\rangle_Q = 0$$

holds for all $k,\ell \in \mathbb{N}$ with $k \neq \ell$ and $v,w \in L^2(0,T)$ such that $v\phi_k, w\phi_\ell \in Y$. Then we define the spaces

$$Y_k = \{v \in L^2(0,T) \mid \langle D(v\phi_k), v\phi_k\rangle_Q < \infty\},$$

and equip those with the norm

$$\|v\|_{Y_k} = \sqrt{\langle D(v\phi_k), v\phi_k\rangle_Q}.$$

Then by definition of the norm, and our assumptions on D, it immediately follows that

$$\|y\|^2_Y = \sum_{k=1}^\infty \|y_k\|^2_{Y_k} \quad \Rightarrow \quad \|y\|_Y \geq \|y_k\|_{Y_k}.$$

Defining $D_k : Y_k \to Y_k^*$ via $\langle D_k y, z\rangle_{(0,T)} = \langle D(y\phi_k), z\phi_k\rangle_Q$, the operator D_k induces the norm $\|\cdot\|_{Y_k}$. Let us now also assume that we have another space $P \subset L^2(0,T;H^1_0(\Omega))$, with its norm being induced by an operator $A : P \to P^*$, and let P_k and $A_k : P_k \to P_k$ be constructed the same way as for the space Y. The following Lemma states that every isomorphism $B : Y \to P^*$ induces a sequence of operators $B_k : Y_k \to P_k$, which are also isomorphisms.

Lemma 1.9. *Let $B : Y \to P$ be an isomorphism, and let $0 < c_1^B \leq c_2^B < \infty$ denote the constants from Equation (1.6), related to the inf-sup stability and the boundedness of B respectively. Further, assume that*

$$\langle B(y\phi_k), p\phi_\ell \rangle_Q = 0$$

holds for all $k, \ell \in \mathbb{N}$ with $k \neq \ell$ and $y, p \in L^2(0,T)$ such that $y\phi_k \in Y$ and $p\phi_\ell \in P$. Then for all $k \in \mathbb{N}$ the operator $B_k : Y_k \to P_k^*$ defined by

$$\langle B_k y, p \rangle_{(0,T)} = \langle B(y\phi_k), (p\phi_k) \rangle_Q, \qquad \forall y \in Y_k \forall p \in P_k,$$

is isomorphic as well, and the constants $c_1^{B_k}$ and $c_2^{B_k}$ related to the inf-sup stability and the boundedness of B_k satisfy $c_1^B \leq c_1^{B_k}$ and $c_2^{B_k} \leq c_2^B$.

Proof: If B is injective, B_k has to be injective as well, so it only remains to show the boundedness and inf-sup stability. For the boundedness simply note that for all $y \in Y_k$ and $p \in P_k$ we have

$$\langle B_k y, p \rangle_{(0,T)} = \langle B(y\phi_k), (p\phi_k) \rangle_Q \leq c_2^B \|y\phi_k\|_Y \|p\phi_k\|_P = c_2^B \|y\|_{Y_k} \|p\|_{P_k}.$$

For the inf-sup stability we obtain

$$\begin{aligned} \sup_{0 \neq p \in P_k} \frac{\langle B_k y, p \rangle_{(0,T)}}{\|p\|_{P_k}} &= \sup_{0 \neq p \in P_k} \frac{\langle B(y\phi_k), (p\phi_k) \rangle_Q}{\|p\|_{P_k}} = \sup_{0 \neq p \in P} \sum_{\ell=1}^{\infty} \frac{\langle B(y\phi_k), (p_\ell \phi_\ell) \rangle_Q}{\|p_k\|_{P_k}} \\ &\geq \sup_{0 \neq p \in P} \frac{\langle B(y\phi_k), p \rangle_Q}{\|p\|_P} \geq c_1^B \|y\phi_k\|_Y = c_1^B \|y\|_{Y_k} \end{aligned}$$

□

When deriving estimates involving Bochner-spaces in later sections of this work, the following simple inequalities will also be helpful.

Lemma 1.10. *Let $a, b, c, d \in \mathbb{R}$, then it holds that*

$$(a+b)^2 \leq 2(a^2 + b^2), \tag{1.10}$$

as well as

$$ac + \alpha bd \leq \sqrt{a^2 + \alpha b^2} \sqrt{c^2 + \alpha d^2} \tag{1.11}$$

Proof: The first inequality directly follows from

$$0 \leq (a-b)^2 = a^2 - 2ab + b^2 \quad \Rightarrow \quad 2ab \leq a^2 + b^2.$$

For the second inequality we first observe that

$$\begin{aligned} (ac + \alpha bd)^2 &= a^2c^2 + 2\alpha(ad)(bc) + \alpha^2 b^2 d^2 \\ &\leq a^2c^2 + \alpha\left((ad)^2 + (bc)^2\right) + \alpha^2 b^2 d^2 = (a^2 + \alpha b^2)(c^2 + \alpha d^2). \end{aligned}$$

Taking the square root of this inequality proves the desired claim. □

2 Optimal Control Theory

In this work we will consider optimal control problems of tracking type, for which we aim to drive some state y of a dynamical system close towards a desired state y_d using an optimal control u which minimizes some related cost functional. When computing optimal control policies there is an important distinction between so called *closed-loop* control policies $u(y,t)$, which depend on both the current state and the current time, and *open-loop* control policies $u(t)$, which only depend on time. In general it is easier to to compute an open-loop control $u(t)$ for a given initial state of the system, that drives the system towards the optimal state $y(t)$. But in case that the actual state is disrupted by random perturbations not incorporated by the model, for example due to random temperature fluctuations or wind, the control $u(t)$ will yield suboptimal results at best. In cases where the desired state is not stable, like it is the case for the upright position of a pendulum for instance, the control $u(t)$ might fail to regulate the system at all. In those cases it is necessary to derive a suitable closed-loop control $u(y,t)$, which captures the optimal control response if the state is y at the time t, for all possible pairs (y,t).

In this chapter we aim to introduce the Hamilton–Jacobi–Bellman equation (HJB), which gives rise to a closed-loop control, and the Pontryagin Maximum Principle (PMP), which provides a way to compute an open-loop control. First, some fundamental principles of classical mechanics are discussed, before the relationship between Hamilton's equations and the Hamilton–Jacobi equation is established, closely following the textbooks [9, Chapter 6 and 9] and [8, Chapter 1-2]. After establishing a similar relation for the HJB equation and the PMP, we consider the prominent special case of a Linear Quadratic Regulator (LQR) as in [22]. Finally, the chapter concludes with an extension of the usual LQR setting to Hilbert spaces, following [19].

2.1 Principles of Classical Mechanics

The main purpose of this section is to introduce Hamilton's equations of motion and the Hamilton–Jacobi equation, since the solutions associated to those formalisms have a similar relation as closed-loop to open-loop controls. While solving Hamilton's equations with a fixed initial position and initial impulse provides the trajectory $\underline{x}(t)$ and the related impulse $\underline{p}(t)$ of a point mass as a function of time, the solution of the Hamilton–Jacobi equation for a fixed initial or terminal position yields a relation $\underline{p}(\underline{x},t)$, and thus characterizes the impulse of all possible trajectories.

In order to illustrate the difference between these two formalisms, we will consider the simple model problem of a point mass m within a gravitational field. Throughout this section, we denote by $\underline{x} : [0,T] \to \mathbb{R}^2, t \mapsto (x(t), z(t))^\top$ its time-dependent trajectory, where we have restricted ourselves to two spatial dimensions for the sake of simplicity.

2.1.1 Lagrange and Hamilton Formalism

Assume that we are given an initial position $\underline{x}(0) = \underline{x}_0 = (x_0, z_0)^\top \in \mathbb{R}^2$ and an initial velocity $\dot{\underline{x}}(0) = \underline{v}_0 = (v_0, w_0)^\top \in \mathbb{R}^2$. Then, in the Newtonian point of view, the particle's trajectory is solely determined by the initial conditions and Newton's equation, see for instance [8, Chapter 1.1] or [3, Chapter 1], which states that

$$m\ddot{\underline{x}}(t) = \underline{F}, \quad t \in (0, T]. \tag{2.1}$$

As a simplified model of a gravitational field we choose the constant force $\underline{F} = (0, -mg)^\top$ with $g > 0$, but in general this force may also depend on the time, the particle's position, its velocity or acceleration. Plugging this force into Newton's equation results in the system

$$\begin{aligned} m\ddot{x}(t) &= 0, & x(0) &= x_0, & \dot{x}(0) &= v_0, \\ m\ddot{z}(t) &= -mg, & z(0) &= z_0, & \dot{z}(0) &= w_0. \end{aligned} \tag{2.2}$$

Solving this system of differential equations then yields the general solution

$$\underline{x}(t) = \begin{pmatrix} x(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} x_0 + v_0 t \\ z_0 + w_0 t - \frac{g}{2} t^2 \end{pmatrix}.$$

This constant force $\underline{F}$ is part of a special class of position dependent forces $\underline{F}(\underline{x})$ called *conservative forces*, for which there exists a potential function $V : \mathbb{R}^2 \to \mathbb{R}$ such that $\underline{F}(\underline{x}) = -\nabla V(\underline{x})$. This potential function $V(\underline{x})$ is often referred to as the *potential energy* of the system. For our specific choice of $\underline{F} = (0, -mg)^\top$ the potential energy is given by $V(\underline{x}) = V(x, z) = mgz$. A related quantity is the *kinetic energy* associated with the mass m, which is defined by $T(\underline{v}) = T(v, w) = \frac{m(v^2+w^2)}{2}$. Then the *total energy* associated with the mass m at position $\underline{x}$ and with velocity $\underline{v}$ is defined as the sum

$$E(\underline{x}, \underline{v}) = T(\underline{v}) + V(\underline{x}).$$

For conservative forces, it is well known that Equation (2.1) in combination with the chain rule implies conservation of this total energy. So in our specific example, we have

$$E(\underline{x}(t), \dot{\underline{x}}(t)) = E_0 = E(\underline{x}_0, \underline{v}_0) = \frac{m(v_0^2+w_0^2)}{2} + mgz_0, \qquad \forall t \in [0, T].$$

Instead of solving Newton's equation, one can also use the Lagrangian formalism to derive the same trajectory of the point mass. A formal introduction to the Lagrangian formalism from a physicist's perspective is given in [8, Chapter 1.4] or [13, Chapter 2] for example, while books like [9, Chapter 6] or [12, Chapter 1, §4] also focus on the mathematical aspects and technical details. In contrast to Newton's description of a dynamical system using forces, the Lagrangian formalism is purely based on the concept of energy. The central object of this formalism is the *Lagrangian* $\mathcal{L}(\underline{x}, \underline{v})$ of the system, which is defined as the difference between the kinetic and the potential energy. In our specific example of a mass within a gravitational field we have

$$\mathcal{L}(\underline{x}, \underline{v}) = \mathcal{L}(x, z, v, w) = \frac{m(v^2+w^2)}{2} - mgz. \tag{2.3}$$

Then the so-called *action functional*, acting on a trajectory, is defined as

$$\mathcal{S}[\underline{x}] = \int_0^T \mathcal{L}(\underline{x}(t), \dot{\underline{x}}(t))\ \mathrm{d}t$$

Based on this action functional, the *principle of least action* states that the physically meaningful trajectory of the particle has to be a (local) minimizer of this action functional[1]. As shown in [8, Chapter 1.5] using standard techniques from calculus of variations, the trajectory leading to a stationary action functional has to satisfy the *Euler-Lagrange equations* given by

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\left(\nabla_{\underline{v}}\mathcal{L}(\underline{x}(t), \dot{\underline{x}}(t))\right) = \nabla_{\underline{x}}\mathcal{L}(\underline{x}(t), \dot{\underline{x}}(t)). \tag{2.4}$$

This system of differential equations can be solved in combination with initial conditions, in order to determine the future behaviour of a system. For our specific example we have already derived the specific form of the Lagrangian in Equation (2.3). For this Lagrangian the partial derivatives are given by

$$\begin{aligned} \tfrac{\partial\mathcal{L}}{\partial x} &= 0, & \tfrac{\partial\mathcal{L}}{\partial z} &= -mg,\\ \tfrac{\partial\mathcal{L}}{\partial v} &= mv, & \tfrac{\partial\mathcal{L}}{\partial w} &= mw. \end{aligned}$$

Therefore the Euler–Lagrange equations reduce to

$$\begin{aligned} m\ddot{x}(t) &= \tfrac{\mathrm{d}}{\mathrm{d}t}\left(\tfrac{\partial\mathcal{L}}{\partial v}(x(t), z(t), \dot{x}(t), \dot{z}(t))\right) = \tfrac{\partial\mathcal{L}}{\partial x}(x(t), z(t), \dot{x}(t), \dot{z}(t)) = 0,\\ m\ddot{z}(t) &= \tfrac{\mathrm{d}}{\mathrm{d}t}\left(\tfrac{\partial\mathcal{L}}{\partial w}(x(t), z(t), \dot{x}(t), \dot{z}(t))\right) = \tfrac{\partial\mathcal{L}}{\partial z}(x(t), z(t), \dot{x}(t), \dot{z}(t)) = -mg, \end{aligned}$$

which is exactly the same system (2.2) that resulted from Newton's equation.

A third possibility to treat the same problem is given by the so-called Hamilton formalism, which is also based on the principle of least action. Instead of the velocity $\underline{v}$, the Hamilton formalism is formulated in terms of the the impulse $\underline{p}(\underline{x}, \underline{v}) = \nabla_{\underline{v}}\mathcal{L}(\underline{x}, \underline{v})$. Assuming that this equation defining $\underline{p}(\underline{x}, \underline{v})$ can be explicitly solved for a relation $\underline{v}(\underline{x}, \underline{p})$, the so-called *Hamiltonian* of the system is defined by

$$\mathcal{H}(\underline{x}, \underline{p}) = \langle \underline{p}, \underline{v}(\underline{x}, \underline{p})\rangle - \mathcal{L}(\underline{x}, \underline{v}(\underline{x}, \underline{p})). \tag{2.5}$$

As shown in [9, Chapter 6.5], this Hamiltonian can be used to rewrite the Lagrange equations (2.4), which are in general N second order differential equations, into $2N$ first order differential equations of the form

$$\begin{aligned} \dot{\underline{x}}(t) &= \nabla_{\underline{p}}\mathcal{H}(\underline{x}(t), \underline{p}(t)),\\ \dot{\underline{p}}(t) &= -\nabla_{\underline{x}}\mathcal{H}(\underline{x}(t), \underline{p}(t)). \end{aligned} \tag{2.6}$$

[1] Note that in parts of the literature, the principle is nowadays called the *principle of stationary action*, since there are scenarios in which the particle's trajectory is only a saddle-point of the functional, see for instance [14]. But for the purely motivational purpose of this work, it is sufficient to think of the physically meaningful trajectory as the minimizer of the action functional

These are called Hamilton's equations of motion, and they are once again another equivalent reformulation of Newton's second law. In our specific example of a mass within a gravitational field the impulse relation $\underline{p} = (p, q)^\top = \nabla_{\underline{v}} \mathcal{L}$ just translates to $p = mv$ and $q = mw$. Therefore the Hamiltonian $\mathcal{H}$ is given by

$$\begin{aligned}\mathcal{H}(x, z, p, q) &= pv + qw - \mathcal{L}(x, z, v, w) \\ &= p(\tfrac{p}{m}) + q(\tfrac{q}{m}) - \left(\tfrac{m}{2}\left((\tfrac{p}{m})^2 + (\tfrac{q}{m})^2\right) - mgz\right) = \tfrac{p^2+q^2}{2m} + mgz.\end{aligned}$$

Therefore Hamilton's equations of motion (2.6) reduce to the system

$$\begin{aligned}\dot{p}(t) &= \tfrac{\partial \mathcal{H}}{\partial x}(\underline{x}(t), \underline{p}(t)) = 0, \\ \dot{q}(t) &= \tfrac{\partial \mathcal{H}}{\partial z}(\underline{x}(t), \underline{p}(t)) = mg, \\ \dot{x}(t) &= -\tfrac{\partial \mathcal{H}}{\partial p}(\underline{x}(t), \underline{p}(t)) = -\tfrac{p(t)}{m}, \\ \dot{z}(t) &= -\tfrac{\partial \mathcal{H}}{\partial p}(\underline{x}(t), \underline{p}(t)) = -\tfrac{q(t)}{m}.\end{aligned}$$

Differentiating the third and fourth equation, and plugging the first and second equation into the third and fourth equation respectively again results in the same second order system as originally obtained by Newton's equation.

2.1.2 Hamilton–Jacobi Equation

In Section 2.1.1 the principle of least action was introduced as an alternative, but equivalent formalism to Newton's second law. According to this principle of least action, we assign an action to all theoretically possible trajectories, and then try to find the trajectory with the lowest action, which consequently has to be a solution of Hamilton's equations (2.6). Let us fix a starting point $\underline{x}(0) = \underline{x}_0$ and a terminal point $\underline{x}(T) = \underline{x}_T$, and let us assume that regardless of the specific choice for $\underline{x}_0, \underline{x}_T \in \mathbb{R}^2$ and $T > 0$, there exists a unique trajectory satisfying those boundary conditions and Hamilton's equations (2.6). Then we can define

$$S(\underline{x}_T, T) = \mathcal{S}[\underline{x}] = \int_0^T \mathcal{L}(\underline{x}(t), \dot{\underline{x}}(t)) \, \mathrm{d}t$$

as the minimal value of action that can be achieved when travelling from $\underline{x}_0$ to $\underline{x}_T$ within the time T. Since the terminal condition $\underline{x}_T$ and the associated time T were chosen arbitrarily, they can be varied to obtain a function $S : \mathbb{R}^2 \times [0, \infty) \to \mathbb{R}$. From now on we will write $S(\underline{x}, t)$ instead of $S(\underline{x}_T, T)$. It turns out that this action function $S(\underline{x}, t)$ can also be characterized via the partial differential equation

$$\tfrac{\partial S}{\partial t} + \mathcal{H}(\underline{x}, \nabla_{\underline{x}} S) = 0, \tag{2.7}$$

as it is shown in [9, Chapter 9.3] for example. This partial differential equation is called the Hamilton–Jacobi equation, and its use is not just restricted to physics. As we will see in Section 2.2, this equation also plays a major role in optimal control theory. The following Lemma relates the Hamilton–Jacobi equation to Hamilton's equations.

Lemma 2.1 ([10, §3.2.5 Example 6 and §3.3.1 Theorem 2])**.** *Applying the method of characteristics to the Hamilton–Jacobi Equation* (2.7) *yields Hamilton's equations of motion* (2.6)*. Further, the Hamiltonian $\mathcal{H}$ is conserved along trajectories solving Hamilton's equations of motion.*

As discussed in [10] and [13], once the solution $S(\underline{x},t)$ of the Hamilton–Jacobi equation is known, the impulse $\underline{p}$ can be computed as $\underline{p}(\underline{x},t) = \nabla_{\underline{x}} S(\underline{x},t)$. To illustrate this impulse relation, we will again focus on our specific example of a mass m within a gravitational field, and solve the related Hamilton–Jacobi equation. Recall that we have derived the Hamiltonian

$$\mathcal{H}(\underline{q},\underline{p}) = \mathcal{H}(x,z,p,q) = \tfrac{p^2+q^2}{2m} + mgz.$$

Further, we have already shown that trajectories satisfying Hamilton's equations of motion are given by

$$\begin{pmatrix} x(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} x_0 + \frac{p_0}{m}t \\ z_0 + \frac{q_0}{m} - \frac{gt^2}{2} \end{pmatrix},$$

where the initial velocities were replaced by the initial momenta p_0 and q_0. Now assume we want to know the action it takes to travel from $(x_0,z_0)^\top$ to some point $(x',z')^\top$ within the time $t' > 0$. First of all we note that

$$\begin{pmatrix} x(t') \\ z(t') \end{pmatrix} = \begin{pmatrix} x' \\ z' \end{pmatrix} \quad \Leftrightarrow \quad \begin{pmatrix} p_0 \\ q_0 \end{pmatrix} = \begin{pmatrix} m\frac{x'-x}{t'} \\ m\left(\frac{z'-z_0}{t'} + \frac{gt'}{2}\right) \end{pmatrix}.$$

Then, using that the Hamiltonian is conserved along trajectories solving Hamilton's equations of motions, we obtain

$$\begin{aligned} \tfrac{\partial S}{\partial t}(x',z',t') &= \left(\tfrac{\partial S}{\partial t}\right)(x(t'),z(t'),t') = -\mathcal{H}(x(t'),z(t'),p(t'),q(t')) = -\mathcal{H}(x_0,z_0,p_0,q_0) \\ &= -\tfrac{p_0^2+q_0^2}{2m} - mgz_0 = -\tfrac{m}{2}\left(\tfrac{x'-x_0}{t'}\right)^2 - \tfrac{m}{2}\left(\tfrac{z'-z_0}{\tilde{t}} + \tfrac{g\tilde{t}}{2}\right)^2 - mgz_0 \\ &= -\tfrac{m}{2}\left(\left(\tfrac{x'-x_0}{t}\right)^2 + \left(\tfrac{z'-z_0}{t}\right)^2\right) - \tfrac{m(gt')^2}{8} - \tfrac{mg(z'+z_0)}{2} \end{aligned}$$

We integrate this equation with respect to t' and omit the prime when writing the coordinates in order to obtain the action function

$$S(x,z,t) = \tfrac{m}{2t}\left((x-x_0)^2 + (z-z_0)^2\right) - \tfrac{m(z+z_0)gt}{2} - \tfrac{mg^2t^3}{24}.$$

Taking the gradient of this function with respect to x and z then yields the impulse relation

$$\underline{p}(x,z,t) = \begin{pmatrix} \frac{\partial S}{\partial x} \\ \frac{\partial S}{\partial z} \end{pmatrix} = \begin{pmatrix} \frac{m(x-x_0)}{t} \\ \frac{m(z-z_0)}{t} - \frac{mgt}{2} \end{pmatrix}.$$

In Figure 2.1, the trajectory $(x(t),z(t))^\top$ is computed for $m = 1$ kg, $x_0 = 0$ m, $z_0 = 50$ m, $p_0 = 15\ \frac{\text{m·kg}}{\text{s}}$, $q_0 = 0\ \frac{\text{m·kg}}{\text{s}}$, and plotted together with the level lines

$$S(x,z,t_i) = S(x(t_i),z(t_i),t_i)$$

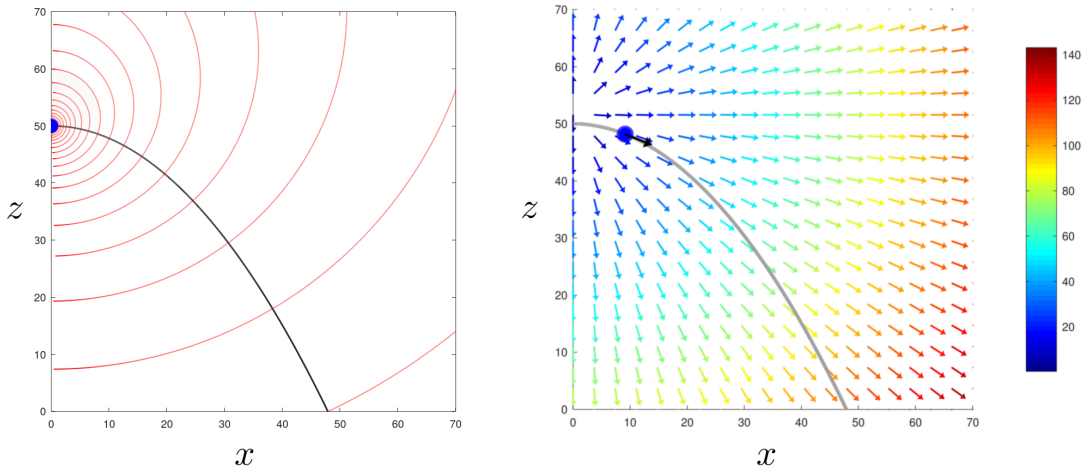

Figure 2.1: Left side: Plot of a specific trajectory together with different level lines of the action function. Right side: Plot of the same trajectory together with the impulse $\underline{p}(x,z,t) = \nabla_{\underline{x}} S(x,z,t)$ for a fixed time.

for different times $t_i \in \{\sqrt{\frac{2H}{g}} \cdot 0.8^j \mid j \in \{0,...,20\}\}$. The graph shows that the level lines are perpendicular to the trajectory. This is in alignment with the theory, since the impulse $\underline{p}(t)$ is always tangential to the curve, but also orthogonal to the level lines of $S(x,z,t)$ due to $\underline{p} = \nabla_{\underline{x}} S$. The second graph in Figure 2.1 shows the same trajectory, together with the impulse $\underline{p}(x,z,t) = \nabla_{\underline{x}} S(x,z,t)$ on a grid of points in the plane for a fixed time $t = \frac{1}{10}\sqrt{\frac{2H}{g}}$. The impulse $\underline{p}(x,z,t)$ describes the impulse the mass would have at the point $(x,z)^\top$, if it has moved from $(0,50)^\top$ to $(x,z)^\top$ within the time t.

To conclude this quick revision of concepts from classical mechanics, we want to highlight once more the difference between solving Hamilton's equations of motion and the Hamilton–Jacobi equation. As shown in the right graph in Figure 2.1, solving Hamilton's equations provides the trajectory $\underline{x}(t)$ for every point in time, as well as the change of the trajectory at every point in time in form of the impulse $p(t)$. In contrast, solving the Hamilton–Jacobi equation yields an impulse relation $\underline{p}(\underline{x},t)$, containing information about the impulse of all possible trajectories the system could take beginning from our initial condition, but does not provide a single trajectory without having to additionally solve another ordinary differential equation.

2.2 Hamilton–Jacobi–Bellman Equation

Let $T > 0$ be given. The abstract optimal control problem considered in this section is given by

$$\begin{aligned} \min_{\underline{y}\in Y, \underline{u}\in\mathcal{U}} J(\underline{y},\underline{u}) &= \int_0^T \mathcal{L}(\underline{y}(s),\underline{u}(s),s)\ \mathrm{d}s + g(\underline{y}(T)) \\ \text{subject to } \dot{\underline{y}}(t) &= \underline{f}(\underline{y}(t),\underline{u}(t),t), \quad t\in(0,T] \\ \underline{y}(0) &= \underline{y}_0. \end{aligned} \tag{2.8}$$

Here $\underline{y} : [0,T] \to \mathbb{R}^n$ represents the state of the system we would like to control, with $\underline{y}_0 \in \mathbb{R}^n$ constituting the initial state of the system, and Y denoting the state space. The function $\underline{u} : [0,T] \to \mathbb{R}^m$ allows to influence the system, with the set $\mathcal{U}$ denoting the set of all admissible controls. This set of admissible controls allows to incorporate constraints, for example to keep the control $\underline{u}(t)$ bounded. For this section we will simply assume that

$$\mathcal{U} = \{\underline{u} : [0,T] \to \mathbb{R}^m \mid \underline{u} \text{ is measurable}\},$$

following the setting used in [10, Chapter 10.3]. The dynamics are governed by the function $\underline{f} : \mathbb{R}^n \times \mathbb{R}^m \times [0,T] \to \mathbb{R}^n$, and the cost of using a specific control strategy $\underline{u} \in \mathcal{U}$ is described by the functional J. This cost functional consists of two parts, with the first part involving the integral over the function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times [0,T] \to \mathbb{R}$, which we will call the Lagrangian associated to the control problem, and the function $g : \mathbb{R}^n \to \mathbb{R}$, penalizing only the terminal state.

In what follows we will describe how to use the HJB equation to derive closed-loop and PMP to derive open-loop controls for the control problem (2.8). First of all we define the *value function* $V(\underline{y},t)$, which is the central object of the HJB equation, analogously to the role of the action function $S(\underline{x},t)$ within the Hamilton–Jacobi equation in Section 2.1.2.

Definition 2.2. *Let $\underline{y}_0 \in \mathbb{R}^n$, $t_0 \in [0,T)$ and $\underline{u} \in \mathcal{U}$ be given. Further, denote by $\underline{y}(t)$ the unique trajectory solving the initial value problem*

$$\begin{aligned} \dot{\underline{y}}(t) &= \underline{f}(\underline{y}(t),\underline{u}(t),t), \quad t\in(t_0,T], \\ \underline{y}(t_0) &= \underline{y}_0. \end{aligned}$$

Then we define the cost associated to a given control $\underline{u}$ as

$$C_{\underline{y}_0,t_0}[\underline{u}] = \int_{t_0}^T \mathcal{L}(\underline{y}(s),\underline{u}(s),s)\ \mathrm{d}s + g(\underline{y}(T)).$$

The value function $V(\underline{y}_0,\underline{t}_0)$ is then defined as the minimal cost that can be achieved, meaning that

$$V(\underline{y}_0,t_0) = \min_{\underline{u}\in\mathcal{U}} C_{\underline{y}_0,\underline{t}_0}[\underline{u}].$$

We will usually write $V(\underline{y},t)$ instead of $V(\underline{y}_0,t_0)$.

This value function $V(\underline{y}, t)$ represents the cost-to-go, if one does start at $\underline{y}(t) = \underline{y}$ instead of the fixed initial position $\underline{y}(0) = \underline{y}_0$. The following Lemma plays a key role in the derivation of the HJB equation. Roughly speaking it states that the optimal cost $V(\underline{y}, t)$, when beginning at the time t with the state y, is the same as the sum of the running costs when following the optimal trajectory for a small time $\tau > 0$ to the point $\underline{y}(t+\tau)$, and the optimal cost $V(\underline{y}(t+\tau), t+\tau)$ for the remaining time interval $(t+\tau, T]$.

Lemma 2.3 (Bellman's Optimality Principle, [10, Chapter 10.3.2])**.** *Let $V(\underline{y}, t)$ denote the value function, and let $t \in [0, T], \tau > 0$ be given such that $t + \tau \leq T$. Then it holds that*

$$V(\underline{y}, t) = \min_{\underline{u} \in \mathcal{U}} \left(\int_t^{t+\tau} \mathcal{L}(\underline{y}_{\underline{u}}(s), \underline{u}(s), s) \, \mathrm{d}s + V(\underline{y}_{\underline{u}}(t+\tau), t+\tau) \right), \tag{2.9}$$

with $\underline{y}_{\underline{u}}(s)$ denoting the solution of the initial boundary value problem

$$\begin{aligned} \dot{\underline{y}}_{\underline{u}}(s) &= \underline{f}(\underline{y}_{\underline{u}}(s), \underline{u}(s), s), \quad s \in [t, t+\tau], \\ \underline{y}_{\underline{u}}(t) &= \underline{y}. \end{aligned}$$

In order to characterize $V(\underline{y}, t)$ via a partial differential equation, we will assume that V is at least continuously differentiable twice. This is only done for the sake of simplicity, note that [10, Chapter 10.3.2] also provides a proof using the weaker notion of viscosity solutions. Now following [1, Chapter 2.2], we take the optimality principle from Equation (2.9), subtract $V(\underline{y}, t) = V(\underline{y}_{\underline{u}}(t), t)$ on both sides, and subsequently divide the equation by $\tau > 0$, resulting in

$$0 = \min_{u \in \mathcal{U}} \left(\frac{1}{\tau} \int_t^{t+\tau} \mathcal{L}(\underline{y}_{\underline{u}}(s), \underline{u}(s), s) \, \mathrm{d}s + \frac{V(\underline{y}_{\underline{u}}(t+\tau), t+\tau) - V(\underline{y}_{\underline{u}}(t), t)}{\tau} \right).$$

Now assuming that the map $t \mapsto V(\underline{y}_u(t), t)$ is sufficiently smooth, one can take the limit $\tau \to 0$ to arrive at

$$0 = \min_{\underline{u} \in \mathcal{U}} \left(\mathcal{L}(\underline{y}_{\underline{u}}(t), \underline{u}(t), t) + \left\langle \nabla_{\underline{y}} V(\underline{y}_{\underline{u}}(t), t), \underline{f}(\underline{y}_{\underline{u}}(t), \underline{u}(t), t) \right\rangle + \frac{\partial V}{\partial t}(\underline{y}_{\underline{u}}(t), t) \right).$$

Note that the minimization is formally still over all trajectories $\underline{u} \in \mathcal{U}$, but only the function value $\underline{u}(t)$ matters, such that the minimization can be reduced to finding the minimizing $\underline{u} = \underline{u}(t) \in \mathbb{R}^m$. Next we use $\underline{y}_{\underline{u}}(t) = \underline{y}$ and the fact that $\frac{\partial V}{\partial t}(\underline{y}, t)$ does not depend on $\underline{u}$ at all to conclude

$$\frac{\partial V}{\partial t}(\underline{y}, t) + \min_{u \in \mathbb{R}^m} \left(\mathcal{L}(\underline{y}, \underline{u}, t) + \left\langle \nabla_{\underline{y}} V(\underline{y}, t), \underline{f}(\underline{y}, \underline{u}, t) \right\rangle \right) = 0. \tag{2.10}$$

This is the so-called Hamilton–Jacobi–Bellman equation, in short HJB equation, and it allows to characterize the value function $V(\underline{y}, t)$ by a partial differential equation, similarly to the Hamilton–Jacobi equation (2.7) and the action function $S(\underline{x}, t)$. Additionally to the Equation (2.10), the value function satisfies the terminal boundary

condition $V(\underline{y}, T) = g(\underline{y})$. It is common in the literature to define a Hamiltonian $\mathcal{H} : \mathbb{R}^n \times \mathbb{R}^n \times [0, T] \to \mathbb{R}$ associated to the control problem via

$$\mathcal{H}(\underline{y}, \underline{p}, t) = \min_{\underline{u} \in \mathbb{R}^m} \left(\mathcal{L}(\underline{y}, \underline{u}, t) + \left\langle \underline{p}, \underline{f}(\underline{y}, \underline{u}, t) \right\rangle \right).$$

which allows to rewrite the Hamilton–Jacobi–Bellman equation into the form

$$\frac{\partial V}{\partial t} + \mathcal{H}(\underline{y}, \nabla_{\underline{y}} V(\underline{y}, t), t) = 0.$$

Note that this equation exactly resembles the Hamilton–Jacobi equation (2.7) from classical mechanics. Now in order to construct an optimal control from a solution $V(\underline{y}, t)$ of the HJB equation, we choose $u(\underline{y}, t)$ for all $\underline{y} \in \mathbb{R}^n$ and $t \in [0, T]$ such that

$$\underline{u}(\underline{y}, t) \in \operatorname*{arg\,min}_{\underline{w} \in \mathbb{R}^m} \left(\mathcal{L}(\underline{y}, \underline{w}, t) + \left\langle \nabla_{\underline{y}} V(\underline{y}, t), \underline{f}(\underline{y}, \underline{w}, t) \right\rangle \right).$$

Note that the right-hand side constitutes a set, since the minimum, and therefore also the optimal control $\underline{u}(\underline{y}, t)$, is not necessarily assumed to be unique.

As we have seen, the HJB equation can formally be written in the same form as the Hamilton–Jacobi equation. Therefore, according to Lemma 2.1, the method of characteristics imply

$$\begin{aligned}
\dot{\underline{y}}(t) &= \nabla_{\underline{p}} \mathcal{H}(\underline{y}(t), \underline{p}(t), t), \\
\dot{\underline{p}}(t) &= -\nabla_{\underline{y}} \mathcal{H}(\underline{y}(t), \underline{p}(t), t), \\
\underline{u}(t) &\in \operatorname*{arg\,min}_{\underline{w} \in \mathbb{R}^m} \left(\mathcal{L}(\underline{y}(t), \underline{w}, t) + \left\langle \underline{p}(t), \underline{f}(\underline{y}(t), \underline{w}, t) \right\rangle \right),
\end{aligned}$$

where we have introduced the adjoint state variable $\underline{p}(t)$. Note that due to the explicit time dependence of $\mathcal{H}$, the quantity $\mathcal{H}(\underline{y}(t), \underline{p}(t), t)$ is in general not time-invariant. This optimality condition is called the *Pontryagin's Maximum Principle*, in short PMP. Solving this system leads to an *open-loop* control $\underline{u}(t)$, in comparison to the *closed-loop* control $\underline{u}(\underline{y}, t)$ obtained from solving the HJB equation. Recall that the right plot in Figure 2.1 showed that when solving Hamilton's equation of motions, we obtain a single trajectory for every point within the time interval, whereas solving the Hamilton–Jacobi equation provided the direction in which the particle will move for any point in space and time, but does not produce a trajectory without having to solve another system of differential equations. Similarly, the open-loop control $\underline{u}(t)$ provides the optimal response and the related trajectory for every point in time, but only along this specific trajectory, and the closed-loop control $\underline{u}(\underline{y}, t)$ provides the optimal response at every point in space and time, but not the related optimal trajectory without having to additionally solve a system of differential equations in a post-processing step.

2.3 Linear Quadratic Regulator

This section closely follows [22, Chapter 3]. We will consider the special case of the optimal control problem (2.8), given by

$$\begin{aligned} \min_{\underline{u}\in\mathcal{U}} J(\underline{y},\underline{u}) &= \int_0^T \tfrac{1}{2}(\underline{y}(s)-\underline{y}_d(s))^\top Q(s)(\underline{y}(s)-\underline{y}_d(s)) + \tfrac{1}{2}\underline{u}(s)^\top R(s)\underline{u}(s)\,\mathrm{d}s \\ \text{subject to } \dot{\underline{y}}(t) &= A(t)\underline{y}(t)+B(t)\underline{u}(t), \quad t\in(0,T] \\ \underline{y}(0) &= \underline{y}_0. \end{aligned} \tag{2.11}$$

Here $\underline{y}(t)\in\mathbb{R}^n$ and $\underline{u}(t)\in\mathbb{R}^m$ again represent the state and the control like before, and $\underline{y}_d(t)$ represents a desired target trajectory. The dynamics are now governed by a linear system of differential equations, involving the matrix-valued maps $A:[0,T]\to\mathbb{R}^{n\times n}$ and $B:[0,T]\to\mathbb{R}^{n\times m}$. The cost functional J is based on the maps $Q:[0,T]\to\mathbb{R}^{n\times n}$ and $R:[0,T]\to\mathbb{R}^{m\times m}$, for which it is assumed that both $R(t)$ and $Q(t)$ are symmetric and positive definite for all times $t\in[0,T]$. Therefore this specific optimal control problem corresponds to the abstract control problem (2.8) with the Lagrangian

$$\mathcal{L}(\underline{y},\underline{u},\underline{t}) = \tfrac{1}{2}(\underline{y}-\underline{y}_d(t))^\top Q(t)(\underline{y}-\underline{y}_d(t)) + \tfrac{1}{2}\underline{u}^\top R(t)\underline{u}$$

and

$$\underline{f}(\underline{y},\underline{u},t) = A(t)\underline{y}+B(t)\underline{u}.$$

We will use the theory from the previous section to investigate how an optimal closed-loop control strategy for the control problem (2.11) can be derived. First note that the associated Lagrange function $L(\underline{y},\underline{u},t,\underline{p})$ of this constrained minimization problem is given by

$$\begin{aligned} L(\underline{y},\underline{u},t,\underline{p}) &= \mathcal{L}(\underline{y},\underline{u},t) + \underline{p}^\top \underline{f}(\underline{y},\underline{u},t) \\ &= \tfrac{1}{2}(\underline{y}-\underline{y}_d(t))^\top Q(t)(\underline{y}-\underline{y}_d(t)) + \tfrac{1}{2}\underline{u}^\top R(t)\underline{u} + \underline{p}^\top(A(t)\underline{y}+B(t)\underline{u}) \end{aligned}$$

and recall that the Hamiltonian associated to such a control problem is defined by

$$\mathcal{H}(\underline{y},\underline{p},t) = \min_{\underline{u}\in\mathbb{R}^m} L(\underline{y},\underline{u},t,\underline{p}).$$

Further, note that the Lagrange function $L(\underline{y},\underline{u},t,\underline{p})$ is differentiable and strictly convex with respect to $\underline{u}$, such that we can find the global minimum by means of differentiation. We just solve $\nabla_{\underline{u}} L = 0$, which yields

$$0 = R(t)\underline{u} + B(t)^\top \underline{p} \qquad \Rightarrow \qquad \underline{u} = -R(t)^{-1}B(t)^\top \underline{p}.$$

Using this result, the Hamiltonian can be written as

$$\begin{aligned} \mathcal{H}(\underline{y},\underline{p},t) &= L(\underline{y},(-R(t)^{-1}B^\top \underline{p}),t,\underline{p}) \\ &= \tfrac{1}{2}(\underline{y}-\underline{y}_d(t))^\top Q(t)(\underline{y}-\underline{y}_d(t)) + p^\top A(t)\underline{y} - \tfrac{1}{2}\underline{p}^\top B(t)R(t)^{-1}B(t)^\top \underline{p}. \end{aligned}$$

To derive a closed-loop control $\underline{u}(\underline{y},t)$ we solve the related Hamilton–Jacobi equation

$$-\tfrac{\partial V}{\partial t} = \tfrac{1}{2}(\underline{y}-\underline{y}_d)^\top Q(\underline{y}-\underline{y}_d) + (\nabla_{\underline{y}}V)^\top A\underline{y} - \tfrac{1}{2}(\nabla_{\underline{y}}V)^\top BR^{-1}B^\top(\nabla_{\underline{y}}V),$$

together with the terminal boundary condition $V(\underline{y},T)=0$. As in [22, Remark 3.6] we use an ansatz of the form

$$V(\underline{y},t) = \tfrac{1}{2}\underline{y}^\top P(t)\underline{y} + \underline{y}^\top\underline{\eta}(t) + \zeta(t),$$

where $P(t)=P(t)^\top \in \mathbb{R}^{n\times n}$. Inserting this ansatz into the HJB equation yields

$$\begin{aligned}0 &= \tfrac{1}{2}\underline{y}^\top\dot{P}(t)\underline{y} + \underline{y}^\top\dot{\underline{\eta}}(t) + \dot{\zeta}(t) + \tfrac{1}{2}\underline{y}^\top Q(t)\underline{y} - \underline{y}_d(t)^\top Q(t)\underline{y} + \tfrac{1}{2}\underline{y}_d(t)^\top Q(t)\underline{y}_d(t)\\ &\quad + \underline{y}^\top P(t)A(t)\underline{y} + \underline{\eta}(t)^\top A(t)\underline{y} - \tfrac{1}{2}\underline{y}^\top P(t)B(t)R(t)^{-1}B(t)^\top P(t)\underline{y}\\ &\quad - \underline{y}^\top P(t)B(t)R(t)^{-1}B(t)^\top\underline{\eta}(t) - \frac{1}{2}\underline{\eta}(t)^\top B(t)R(t)^{-1}B(t)^\top\underline{\eta}(t).\end{aligned}$$

Then we can use

$$\underline{y}^\top P(t)A(t)\underline{y} = \tfrac{1}{2}\underline{y}^\top P(t)A(t)\underline{y} + \tfrac{1}{2}\left(\underline{y}^\top P(t)A(t)\underline{y}\right)^\top = \tfrac{1}{2}\underline{y}^\top\left(P(t)A(t)+A(t)^\top P(t)\right)\underline{y},$$

such that comparison of the coefficients leads to the coupled system

$$\begin{aligned}0 &\overset{!}{=} \dot{P}(t) + Q(t) + P(t)A(t) + A(t)^\top P(t) - P(t)B(t)R(t)^{-1}B(t)^\top P(t),\\ 0 &\overset{!}{=} \dot{\underline{\eta}}(t) - Q(t)\underline{y}_d(t) + A(t)^\top\underline{\eta}(t) - P(t)B(t)R(t)^{-1}B(t)^\top\underline{\eta}(t),\\ 0 &\overset{!}{=} \dot{\zeta}(t) + \tfrac{1}{2}\underline{y}_d(t)^\top Q(t)\underline{y}_d(t) - \frac{1}{2}\underline{\eta}(t)^\top B(t)R(t)^{-1}B(t)^\top\underline{\eta}(t).\end{aligned}$$

To fully describe the value function, all three equations from above have to be solved. But in practical applications we are not interested in the value function $V(\underline{y},t)$, but rather in the optimal control $\underline{u}(\underline{y},t)$, which is related to the value function via

$$\underline{u}(t) = -R(t)^{-1}B(t)^\top\nabla_{\underline{y}}V(\underline{y},t) = -R(t)^{-1}B(t)^\top\left(P(t)\underline{y}+\underline{\eta}(t)\right).$$

Therefore the function $\zeta(t)$ is irrelevant if only $\underline{u}(\underline{y},t)$ is of interest, and the third differential equation can be neglected. Then it only remains to solve the coupled system

$$\begin{aligned}\dot{P}(t) &\overset{!}{=} P(t)B(t)R(t)^{-1}B(t)^\top P(t) - Q(t) - P(t)A(t) - A(t)^\top P(t), && P(T)=0,\\ \dot{\underline{\eta}}(t) &\overset{!}{=} \left(B(t)R(t)^{-1}B(t)^\top P(t) - A(t)\right)^\top\underline{\eta}(t) + Q(t)\underline{y}_d(t), && \underline{\eta}(T)=0,\end{aligned} \tag{2.12}$$

which will be done numerically in practice or even analytically if the target function is known and has a simple form. The first of these two equations is a differential equation of Riccati type, while the second is just a linear differential equation.

2.4 Tikhonov Regularization in Hilbert Spaces

In this section we follow [19], and introduce a more abstract setting for the optimal control problem (2.11) with $Q(s) = I$, $R(s) = \varrho I$ for $\varrho > 0$ and $g(y) = 0$, based on Hilbert spaces. Let H be a real-valued Hilbert space, and Y, P be continuously embedded subspaces such that $Y \subset H \subset Y^*$ and $P \subset H \subset P^*$ form Gelfand triples. Further, let $A : P \to P^*$ and $B : Y \to P^*$ be linear and bounded operators with

$$\|Ap\|_{P^*} \leq c_2^A \|p\|_P, \quad \forall y \in Y, \qquad \|By\|_{P^*} \leq c_2^B \|y\|_Y, \quad \forall x \in X.$$

Additionally we assume that A is self-adjoint and elliptic, with its ellipticity constant being denoted by $c_1^A > 0$. Then the Lemma of Lax–Milgram implies that $A : P \to P^*$ defines an isomorphism. Further, $\sqrt{\langle Ap, p\rangle_{P^*\times P}}$ and $\sqrt{\langle f, A^{-1}f\rangle_{P^*\times P}}$ define equivalent norms to $\|\cdot\|_P$ and $\|\cdot\|_{P^*}$, due to

$$c_1^A \|p\|_P^2 \leq \langle Ap, p\rangle_{P^*\times P} \leq c_2^A \|p\|_P^2, \qquad \frac{1}{c_2^A}\|f\|_{P^*}^2 \leq \left\langle f, A^{-1}f\right\rangle_{P^*\times P} \leq \frac{1}{c_1^A}\|f\|_{P^*}^2$$

for all $p \in P$ and $f \in P^*$. For B let us further assume that the adjoint operator $B^\top : P \to Y^*$ is injective, and that the inf-sup condition

$$\inf_{0\neq y\in Y} \sup_{0\neq p\in P} \frac{\langle By, p\rangle_{P^*\times P}}{\|y\|_Y \|p\|_P} \geq c_1^B > 0$$

holds. Then the Banach–Nečas–Babuška theorem implies that B defines an isomorphism as well.

Now in order to generalize the optimal control problem (2.11) for a given desired target $y_d \in H$ to Hilbert spaces, we replace the differential equation representing the system dynamics by the operator equation $By_{\text{opt}} = u$, with y_{opt} denoting the optimal state. Instead of choosing Y directly as the state space, we will write the optimal state as $y_{\text{opt}} = y + y_e$, where y_e defines a suitably chosen extension of the initial value, and $y \in Y$ is the new state variable with homogeneous initial conditions that we aim to compute. Thus the state equation can be written as $By = u + f$, with $u \in P^*$ and some $f \in P^*$ depending on y_d. Since the control u does not necessarily have to be an element of H in this setting, we will use the dual norm $\|u\|_{P^*}^2$ for the regularization. Thus, for $\varrho > 0$ the resulting optimal control problem reads

$$\begin{aligned} \min_{y\in Y, u\in P^*} \ & \tfrac{1}{2}\|y + y_e - y_d\|_H^2 + \tfrac{\varrho}{2}\left\langle u, A^{-1}u\right\rangle_{P^*\times P}, \\ \text{subject to} \quad & By = u + f \quad \text{in } P^*. \end{aligned} \tag{2.13}$$

Substituting u by $By - f$ within the cost functional, this constrained optimization problem is equivalent to the unconstrained optimization problem

$$\min_{y\in Y} \ J(y) = \tfrac{1}{2}\|y + y_e - y_d\|_H^2 + \tfrac{\varrho}{2}\left\langle By - f, A^{-1}(By - f)\right\rangle_{P^*\times P}. \tag{2.14}$$

Using that A^{-1} is self-adjoint, this functional can be rewritten as

$$J(y) = \tfrac{1}{2}\|y + y_e - y_d\|_H^2 + \tfrac{\varrho}{2}\langle B^\top A^{-1} B y, y\rangle_{Y^*\times Y} - \varrho\langle B^\top A^{-1} f, y\rangle_{Y^*\times Y} + \tfrac{\varrho}{2}\langle f, A^{-1} f\rangle_{P^*\times P}.$$

Now the optimality condition $\delta J(y; h) = 0$ for all $h \in Y$ implies

$$y + y_e - y_d + \varrho B^\top A^{-1} B y - \varrho B^\top A^{-1} f = 0 \quad \text{in } Y^*.$$

At this point we define the modified target $\tilde{y}_d = y_d - y_e + \varrho B^\top A^{-1} f$, as well as another linear and bounded operator $S : Y \to Y^*$ by $S = B^\top A^{-1} B$. Then the last equation can be rewritten into the more compact form

$$y + \varrho S y = \tilde{y}_d \quad \text{in } Y^*, \tag{2.15}$$

which is equivalent to the variational formulation of finding $y \in Y$ such that

$$\langle y, v\rangle_H + \varrho\langle Sy, v\rangle_{Y^*\times Y} = \langle \tilde{y}_d, v\rangle_H, \quad \forall v \in Y. \tag{2.16}$$

Alternatively, we can also introduce the adjoint state $p = A^{-1} B y \in P$, and write the operator equation as the equivalent system

$$\begin{aligned} &\text{(P)} && By = u + f, && \text{in } P^*, \\ &\text{(D)} && B^\top p = y + y_e - y_d, && \text{in } Y^*, \\ &\text{(G)} && p + \varrho A^{-1} u = 0, && \text{in } P, \end{aligned} \tag{2.17}$$

consisting of the primal, the dual, and the gradient equation. Note that the three equations closely resemble the optimality conditions from Pontryagin's Maximum Principle.

Lemma 2.4 ([19, Lemma 1])**.** *The operator $S : Y \to Y^*$ is a linear, bounded, self-adjoint and elliptic operator satisfying*

$$c_1^S \|y\|_Y^2 \le \langle Sy, y\rangle_{Y^*\times Y}, \quad \forall y \in Y$$

and

$$\langle Sy, z\rangle_{Y^*\times Y} \le c_2^S \|y\|_Y \|z\|_Y \quad \forall y, z \in Y,$$

with $c_1^S = c_1^A \left(\frac{c_1^B}{c_2^A}\right)^2$ and $c_2^S = \frac{(c_2^B)^2}{c_1^A}$.

These properties of S and the Lemma of Lax–Milgram immediately implies S is isomorphic, and that $\|\cdot\|_S = \sqrt{\langle S\cdot, \cdot\rangle_{Y^*\times Y}}$ is equivalent to $\|\cdot\|_Y$. Thus the operator equation (2.15), the variational formulation (2.16) and the system (2.17) are all uniquely solvable. Further, it is possible to derive error and stability estimates for this unique solution $y \in Y$ of the optimal control problem.

Lemma 2.5. *Assume that $\tilde{y}_d \in H$, and let $y \in Y$ denote the unique solution of the operator equation* (2.15)*. Then there hold the stability estimates*

$$\|y\|_H \le \|\tilde{y}_d\|_H, \qquad \|y\|_S \le \varrho^{-1/2} \|\tilde{y}_d\|_H.$$

If we further assume $\tilde{y}_d \in Y \subset H$, it holds that

$$\|y\|_S \le \|\tilde{y}_d\|_S.$$

Proof: The first two estimates are shown in [19, Lemma 2]. For the last estimate note that

$$\begin{aligned}\|\tilde{y}_d\|_S^2 &= \|\varrho Sy + y\|_S^2 = \varrho^2\|Sy\|_S^2 + 2\varrho\langle Sy, Sy\rangle_{Y^*\times Y} + \|y\|_S^2 \\ &= \varrho^2\|Sy\|_S^2 + 2\varrho\langle Sy, Sy\rangle_H + \|y\|_S^2 = \varrho^2\|Sy\|_S^2 + 2\varrho\|Sy\|_H^2 + \|y\|_S^2 \geq \|y\|_S^2. \qquad \square\end{aligned}$$

Lemma 2.6 ([19, Lemma 3])**.** *Assume that $\tilde{y}_d \in H$ and let $y \in Y$ denote the unique solution of the operator equation* (2.15)*. Then there holds the error estimate*

$$\|y - \tilde{y}_d\|_H \leq \|\tilde{y}_d\|_H.$$

If we further assume that $\tilde{y}_d \in Y$ it holds that

$$\|y - \tilde{y}_d\|_H \leq \varrho^{1/2}\|\tilde{y}_d\|_S, \qquad \|y - \tilde{y}_d\|_S \leq \|\tilde{y}_d\|_S.$$

If the target satisfies the even stronger condition $S\tilde{y}_d \in H$ it holds that

$$\|y - \tilde{y}_d\|_H \leq \varrho\|S\tilde{y}_d\|_H, \qquad \|y - \tilde{y}_d\|_S \leq \varrho^{1/2}\|S\tilde{y}_d\|_H.$$

Note that when solving the abstract optimal control problem (2.8), we are actually rather interested in estimates for $\|y + y_e - y_d\|_H$ and $\|u\|_{Y^*}$, than in estimates for $\|y - \tilde{y}_d\|_H$ and $\|y\|_H, \|y\|_S$. The following corollary relates those estimates to each other.

Corollary 2.7. *Assuming that y_d, y_e and $B^\top A^{-1} f$ are in H, the optimal control problem* (2.13) *is uniquely solvable, and $y \in Y$ is the unique minimizer if and only if y satisfies the operator equation* (2.15)*. Further, the following statements hold:*

(i) *The discrepancy error can be estimated by*

$$\|y + y_e - y_d\|_H \leq \|y_d\|_H + \|y_e\|_H + 2\varrho\|B^T A^{-1} f\|_H,$$

and the optimal control cost can be estimated by

$$\|u\|_{P^*} \leq \|f\|_{P^*} + \varrho^{-1/2}\left(\|y_d\|_H + \|y_e\|_H + \|B^T A^{-1} f\|_H\right).$$

(ii) *If further $\tilde{y}_d \in Y$ the discrepancy error can be estimated by*

$$\|y + y_e - y_d\|_H \leq \varrho^{1/2}\|\tilde{y}_d\|_S + \varrho\|B^T A^{-1} f\|_H,$$

and the optimal control cost can be estimated by

$$\|u\|_{P^*} \leq \|f\|_{P^*} + \|\tilde{y}_d\|_S$$

(iii) *If $S\tilde{y}_d \in H$ the discrepancy error can be estimated by*

$$\|y + y_e - y_d\|_H \leq \varrho\left(\|S\tilde{y}_d\|_H + \|B^\top A^{-1} f\|_H\right),$$

and the optimal control costs stay bounded with the same estimate as in (ii).

Proof: The constrained optimization problem (2.13) is equivalent to the unconstrained optimization problem (2.14). Since the quadratic functional $J(y)$ satisfies

$$\delta^2 J(y;h) = \langle Sh, h\rangle_{Y^*\times Y} \geq c_1^S \|h\|_Y^2$$

it is strictly convex. Thus the functional $J(y)$ does admit a unique minimizer, and the first order equation

$$\delta J(y;h) = 0 \quad \forall h \in Y$$

is both necessary and sufficient for a minimizer. But as we have seen, this equation is equivalent to the operator equation (2.15), which concludes the first part of the statement.

All the error estimates for $\|y + y_e - y_d\|_H$ are a direct consequence of the fact that

$$y - \tilde{y}_d = y - (y_d - y_e + \varrho B^\top A^{-1} f) = (y + y_e - y_d) - \varrho B^\top A^{-1} f,$$

the triangle inequality, the inverse triangle inequality, as well as the respective error estimates from Lemma 2.6. The control cost estimates are a consequences of Lemma 2.5 in combination with

$$\begin{aligned}\|u\|_{P^*} &= \|By - f\|_{P^*} \leq \|f\|_{P^*} + \|By\|_{P^*} \\ &= \|f\|_{P^*} + \sqrt{\langle By, A^{-1}By\rangle_{P^*\times P}} = \|f\|_{P^*} + \|y\|_S,\end{aligned}$$

which completes the proof. □

In practice, for many interesting optimal control problems the operator equation $(I + \varrho S)y = \tilde{y}_d$ will not be analytically solvable. Instead we have to rely on suitable numerical approximation schemes, for which we utilize the equivalent characterization as the variational formulation (2.16). To approximate the solution of this variational formulation, let $Y_h \subset Y$ be a finite dimensional subspace, and consider the discrete variational formulation of finding $y_h \in Y_h$ such that

$$\langle y_h, v_h\rangle_H + \langle Sy_h, v_h\rangle_{Y^*\times Y} = \langle \tilde{y}_d, v_h\rangle_H \tag{2.18}$$

holds for all $v_h \in Y_h$. Since this is a Galerkin-Bubnov type variational formulation with a conformal ansatz space, the unique solvability of the continuous problem already implies well-posedness of this discrete problem. Further, by subtracting Equation (2.18) from Equation (2.16), we obtain the Galerkin-orthogonality

$$\langle y - y_h, v_h\rangle_H + \langle S(y - y_h), v_h\rangle_{Y^*\times Y} = 0, \quad \forall v_h \in Y_h.$$

This Galerkin Orthogonality allows to derive a Cea-type estimate for the error $\|y - y_h\|_H$.

Lemma 2.8 (Cea's Lemma)**.** *Let y and y_h denote the unique solutions of the variational formulations* (2.16) *and* (2.15) *respectively. Then we have*

$$\|y - y_h\|_H \leq \inf_{v_h\in Y_h} \sqrt{\|y - v_h\|_H^2 + \varrho\|y - v_h\|_S^2}$$

Proof: Using the Galerkin-orthogonality, the generalized Cauchy-inequality from Corollary 1.5 and inequality (1.11), we conclude that

$$\begin{aligned}
\|y - y_h\|_H^2 + \varrho\|y - y_h\|_S^2 &= \langle y - y_h, y - y_h\rangle_H + \langle S(y - y_h), y - y_h\rangle_{Y^*\times Y} \\
&= \langle y - y_h, y - v_h\rangle_H + \langle S(y - y_h), y - v_h\rangle_{Y^*\times Y} \\
&\leq \|y - y_h\|_H\|y - v_h\|_H + \varrho\|y - y_h\|_S\|y - v_h\|_S \\
&\leq \sqrt{\|y - y_h\|_H^2 + \varrho\|y - y_h\|_S^2}\sqrt{\|y - v_h\|_H^2 + \varrho\|y - v_h\|_S^2} \\
\Leftrightarrow \quad \sqrt{\|y - y_h\|_H^2 + \varrho\|y - y_h\|_S^2} &\leq \sqrt{\|y - v_h\|_H^2 + \varrho\|y - v_h\|_S^2}
\end{aligned}$$

holds for any $v_h \in Y_h$. Therefore taking the infimum over $v_h \in Y_h$ completes the proof. □

Remark 2.9. *Note that this form of Cea's Lemma allows us to estimate the norm* $\|y_h + y_e - y_d\|_H$*, since the triangle inequality implies*

$$\begin{aligned}
\|y_h + y_e - y_d\|_H &\leq \|y + y_e - y_d\|_H + \|y - y_h\|_H \\
&\leq \|y + y_e - y_d\|_H + \inf_{v_h\in Y_h}\sqrt{\|y - v_h\|_H^2 + \varrho\|y - v_h\|_S^2}
\end{aligned}$$

The first term on the right-hand side of this equation does only depend on the regularity of the target, whereas the second error depends only on the approximation properties of the discrete ansatz space Y_h*.*

3 Optimal Control of the Heat Equation

Let $T > 0$ and $\Omega \subset \mathbb{R}^d$ be a bounded domain with $d \in \{1, 2, 3\}$, and further denote by $Q = \Omega \times (0, T) \subset \mathbb{R}^{(d+1)}$ the space-time cylinder. In this chapter we will consider the model problem in which the system dynamics are described by the the heat equation

$$\begin{aligned} \partial_t y(x,t) - \Delta_x y(x,t) &= u(x,t), && (x,t) \in Q, \\ y(x,0) &= 0, && x \in \Sigma_0 = \Omega \times \{0\}, \\ y(x,t) &= 0, && (x,t) \in \Sigma = \partial\Omega \times (0,T), \end{aligned} \tag{3.1}$$

with $u(x,t)$ denoting the control, which allows to influence the state $y(x,t)$. The goal is to derive a suitable control $u(x,t)$, such that the associated solution of the heat equation is close to a desired target state $y_d \in L^2(0,T)$, which can in general not be realized exactly due to insufficient regularity. How close a given state is to a target will be measured by means of an appropriate cost functional.

One popular strategy to solve this problem is to discretize the space first, in order to rewrite the heat equation into a system of first order linear differential equations of the form

$$\begin{aligned} M_h \dot{y}(t) - A_h y(t) &= u(t), && \text{for } t \in (0,T), \\ y(0) &= 0. \end{aligned}$$

Here M_h and A_h denote the mass and the stiffness matrix respectively, resulting from the spatial discretization. Then, utilizing for example a L^2-Penalty for the control, one can use the LQR-Theory described in Section 2.3 to derive suitable open-loop or closed-loop control policies. While the simplicity of this approach is appealing, one major drawback is the large size of the involved system matrices, especially for problems in three spatial dimensions with geometries requiring a fine mesh. As a remedy, there are several techniques to reduce the dimensionality of the problem while maintaining an acceptable accuracy, see for example [4] or [16], which are out of the scope of this work.

In order to simplify the analysis within this chapter, we will use the strategy described in [17] instead, and decompose the state $y(x,t)$ based on the eigenfunctions of the Laplace operator $-\Delta_x^{-1} : L^2(\Omega) \to L^2(\Omega)$, such that the considered optimal control problems can be decoupled into a sequence of independent optimal control problems, each only involving one ordinary differential equation as a constraint, instead of a partial differential equation. Note that even though this approach can also be realized numerically for domains with a very simple structure, see for instance [17], it is in general not desirable to use this decoupling for controlling the heat equation, since one would have to first numerically compute all relevant eigenfunctions of the Laplace operator $-\Delta_x$ related to the domain.

3.1 Space-Time Variational Formulations

Multiplying the heat equation (3.1) with a test function $p(x,t)$, integrating over the space-time domain Q and using integration by parts yields the weak form

$$\int_Q (\partial_t y)p + \langle \nabla_x y, \nabla_x p\rangle \,\mathrm{d}x\mathrm{d}t = \int_Q up \,\mathrm{d}x\mathrm{d}t \tag{3.2}$$

Following [18] we define the state space Y_1 and the adjoint state space P_1 via

$$Y_1 = L^2(0,T;H_0^1(\Omega)) \cap H^1_{0,}(0,T;H^{-1}(\Omega)), \qquad P_1 = L^2(0,T;H_0^1(\Omega)),$$

which are equipped with the norms

$$\|y\|_{Y_1} = \sqrt{\|\partial_t y\|^2_{L^2(0,T;H^{-1}(\Omega))} + \|\nabla_x y\|^2_{L^2(Q)}}, \qquad \|p\|_{P_1} = \|\nabla_x p\|_{L^2(Q)}.$$

Writing $A = -\Delta_x : L^2(0,T;H_0^1(\Omega)) \to L^2(0,T;H^{-1}(\Omega))$, we know $\|p\|^2_{P_1} = \langle Ap, p\rangle_Q$ holds for all $p \in P_1$, and similarly $\|f\|^2_{P_1^*} = \langle f, A^{-1}f\rangle_Q$ holds for all $f \in P_1^*$. Therefore we can define the linear, bounded, self-adjoint and elliptic operators $D_1 : Y_1 \to Y_1^*$ and $A_1 : P_1 \to P_1^*$ via

$$\langle D_1 y, z\rangle_Q = \langle \partial_t y, A^{-1}\partial_t z\rangle_Q + \langle Ay, z\rangle_Q, \qquad \langle A_1 p, q\rangle_Q = \langle Ap, q\rangle_Q,$$

in order to obtain $\|y\|^2_{Y_1} = \langle D_1 y, y\rangle_Q$ and $\|p\|^2_{P_1} = \langle A_1 p, p\rangle_Q$. Now we assume that $u \in P_1^*$ and further define the operator $B_1 : Y_1 \to P_1^*$ via

$$\langle B_1 y, p\rangle_Q = \langle \partial_t y, p\rangle_Q + \langle \nabla_x y, \nabla_x p\rangle_{L^2(Q)},$$

then the heat equation can be rewritten into the operator equation

$$B_1 y = u \quad \text{ in } P_1^*. \tag{3.3}$$

We call this formulation the primal formulation of the heat equation. The next lemma establishes unique solvability of this primal formulation.

Lemma 3.1 ([11, §3.2.1])**.** *The operator $B_1 : Y_1 \to P_1^*$ defines an isomorphism satisfying*

$$\sup_{0\neq y\in Y_1}\sup_{0\neq p\in P_1} \frac{\langle B_1 y, p\rangle_Q}{\|y\|_{Y_1}\|p\|_{P_1}} \le c_2^{B_1} \quad \textit{and} \quad \inf_{0\neq y\in Y_1}\sup_{0\neq p\in P_1} \frac{\langle B_1 y, p\rangle_Q}{\|y\|_{Y_1}\|p\|_{P_1}} \ge c_1^{B_1}$$

with $c_1^{B_1} = 1$ and $c_2^{B_1} = \sqrt{2}$.

Alternatively, we can also take the integral formulation (3.2) and integrate by parts with respect to time, resulting in the adjoint formulation, in which we aim to find $y \in Y_0$ such that

$$-\langle \partial_t p, y\rangle_Q + \langle \nabla_x y, \nabla_x p\rangle_{L^2(Q)} = \langle u, p\rangle_Q$$

holds for all $p \in P_0$, where the spaces Y_0 and P_0 are now defined as

$$Y_0 = L^2(0,T;H_0^1(\Omega)), \qquad P_0 = L^2(0,T;H_0^1(\Omega)) \cap H_{,0}^1(0,T;H^{-1}(\Omega)).$$

Similar to the primal formulation, we introduce linear, bounded, self-adjoint and elliptic operators $D_0 : Y_0 \to Y_0^*$ and $A_0 : P_0 \to P_0^*$, inducing the norms in Y_0 and P_0 via

$$\langle D_0 y, z\rangle_Q = \langle Ay, z\rangle_Q, \qquad \langle A_0 p, q\rangle_Q = \left\langle \partial_t p, A^{-1}\partial_t q\right\rangle_Q + \langle Ap, q\rangle_Q.$$

Then we can again define an operator $B_0 : Y_0 \to P_0$ by

$$\langle B_0 y, p\rangle_Q = -\langle \partial_t p, y\rangle_Q + \langle \nabla_x y, \nabla_x p\rangle_{L^2(Q)},$$

such that the heat equation corresponds to the operator equation

$$B_0 y = u \quad \text{ in } P_0^*, \tag{3.4}$$

which we refer to as the adjoint formulation. Similar to the primal formulation, we can establish well-posedness of this operator equation. The following lemma is an immediate consequence of Lemma 3.1 and the fact that $B_0 = R_T^\top B_1^\top R_T$, where R_T denotes the time flipping operator, which is an isometric isomorphism from Y_0 to P_1 and from Y_1 to P_0.

Lemma 3.2. *The operator $B_0 : Y_0 \to P_0^*$ defines an isomorphism satisfying*

$$\sup_{0\neq y\in Y_0}\sup_{0\neq p\in P_0}\frac{\langle B_0 y, p\rangle_Q}{\|y\|_{Y_0}\|p\|_{P_0}} \le c_2^{B_0} \quad \textit{and} \quad \inf_{0\neq y\in Y_0}\sup_{0\neq p\in P_0}\frac{\langle B_0 y, p\rangle_Q}{\|y\|_{Y_0}\|p\|_{P_0}} \ge c_1^{B_0}$$

with $c_1^{B_0} = 1$ and $c_2^{B_0} = \sqrt{2}$.

Note that in the primal formulation, the state space admits a higher regularity than the adjoint state space, while in the adjoint formulation the regularities are reversed. Following [23], we will also consider an interpolated formulation, in which the regularities of the state space and the adjoint state space agree. For that purpose we define $Y_{1/2}$ and $P_{1/2}$ as

$$Y_{1/2} = H_{0;0,}^{1,1/2}(Q), \qquad P_{1/2} = H_{0;,0}^{1,1/2}(Q),$$

with the corresponding norms

$$\|y\|_{Y_{1/2}} = \sqrt{\langle \partial_t y, \mathcal{H}_T y\rangle_Q + \|\nabla_x y\|_{L^2(Q)}^2}, \qquad \|p\|_{P_{1/2}} = \sqrt{-\left\langle \partial_t p, \mathcal{H}_T^{-1} p\right\rangle_Q + \|\nabla_x p\|_{L^2(Q)}^2}.$$

These norms are induced by the linear, bounded, self-adjoint and elliptic operators $D_{1/2} : Y_{1/2} \to Y_{1/2}^*$ and $A_{1/2} : P_{1/2} \to P_{1/2}^*$ respectively, with

$$\begin{aligned}\left\langle D_{1/2} y, z\right\rangle_Q &= \langle \partial_t y, \mathcal{H}_T z\rangle_Q + \langle \nabla_x y, \nabla_x z\rangle_{L^2(Q)},\\ \left\langle A_{1/2} p, q\right\rangle_Q &= -\left\langle \partial_t p, \mathcal{H}_T^{-1} q\right\rangle_Q + \langle \nabla_x p, \nabla_x q\rangle_{L^2(Q)}.\end{aligned}$$

Now let us also introduce the operator $B_{1/2} : Y_{1/2} \to P_{1/2}^*$ via

$$\left\langle B_{1/2}y, p\right\rangle_Q = \langle \partial_t y, p\rangle_Q + \langle \nabla_x y, \nabla_x p\rangle_{L^2(Q)}, \qquad \forall y \in Y_{1/2} \forall p \in P_{1/2},$$

such that the heat equation takes the form of the operator equation

$$B_{1/2}y = u \quad \text{in } P_{1/2}^*. \tag{3.5}$$

We call this formulation the interpolated formulation, and similarly to the previous formulations we can prove well-posedness.

Lemma 3.3 ([31, Theorem 3.4.19])**.** *The operator $B_{1/2} : Y_{1/2} \to P_{1/2}^*$ defines an isomorphism satisfying*

$$\sup_{0\neq y\in Y_{1/2}} \sup_{0\neq p\in P_{1/2}} \frac{\left\langle B_{1/2}y, p\right\rangle_Q}{\|y\|_{Y_{1/2}}\|p\|_{P_{1/2}}} \leq c_2^{B_{1/2}} \quad \textit{and} \quad \inf_{0\neq y\in Y_{1/2}} \sup_{0\neq p\in P_{1/2}} \frac{\left\langle B_{1/2}y, p\right\rangle_Q}{\|y\|_{Y_{1/2}}\|p\|_{P_{1/2}}} \geq c_1^{B_{1/2}}$$

with $c_1^{B_{1/2}} = \frac{1}{2}$ and $c_2^{B_{1/2}} = 1$.

3.2 Optimal Control Frameworks

Based on the three different weak formulations (3.3), (3.4) and (3.5) of the heat equation, we will establish three different optimal control problems for tracking a desired target $y_d \in L^2(Q)$. For each of those control problems, we will also utilize a decoupling as described in [17], based on the eigenfunctions $(\phi_k)_{k\in\mathbb{N}} \subset H_0^1(\Omega)$ and the eigenvalues $(\lambda_k)_{k\in\mathbb{N}}$ of the Laplace operator as described in Section 1.2. Specifically, we decompose the state y, the adjoint state p, the desired state y_d and the control u into

$$y(x,t) = \sum_{k=1}^{\infty} y_k(t)\phi_k(x), \qquad y_d(x,t) = \sum_{k=1}^{\infty} y_k^d(t)\phi_k(x),$$
$$u(x,t) = \sum_{k=1}^{\infty} u_k(t)\phi_k(x), \qquad p(x,t) = \sum_{k=1}^{\infty} p_k(t)\phi_k(x),$$

in order to obtain a sequence of decoupled problems for each of the three formulations. Note that $y^d \in L^2(Q)$ implies $y_k^d \in L^2(0,T)$ for all $k \in \mathbb{N}$.

First we consider the optimal control related to the primal formulation. To quantify what it means for y to be close to $y_d \in L^2(Q)$ we introduce the related cost functional $J : Y_1 \times P_1 \to [0,\infty)$ and the constrained minimization problem

$$\begin{aligned} &\min_{y\in Y_1, u\in P_1^*} J_1(y,u) = \tfrac{1}{2}\|y - y_d\|_{L^2(Q)}^2 + \tfrac{\varrho}{2}\|u\|_{P_1^*}^2, \\ &\text{subject to} \qquad B_1 y = u \quad \text{in } P_1^*. \end{aligned} \tag{3.6}$$

Here we have $Y_1 = L^2(0,T;H_0^1(\Omega)) \cap H_{0,}^1(0,T;H^{-1}(\Omega))$ and $P_1 = L^2(0,T;H_0^1(\Omega))$, therefore we introduce the related spaces Y_1^k and P_1^k for the coefficients y_k and p_k via

$$\begin{aligned} Y_1^k &= H_{0,}^1(0,T), \qquad \|y_k\|_{Y_1^k} = \sqrt{\tfrac{1}{\lambda_k}\|\dot{y}_k\|^2_{L^2(0,T)} + \lambda_k \|y_k\|^2_{L^2(0,T)}}, \\ P_1^k &= L^2(0,T), \qquad \|p_k\|_{P_1^k} = \sqrt{\lambda_k}\|p_k\|_{L^2(0,T)}. \end{aligned} \tag{3.7}$$

Then by means of the spectral decomposition, the optimal control problem (3.6) is equivalent to the sequence of pairwise independent optimal control problems

$$\begin{aligned} &\min_{y_k \in Y_1^k, u_k \in [P_1^k]^*} J_{1,k}(y_k,u_k) = \tfrac{1}{2}\|y_k - y_k^d\|^2_{L^2(0,T)} + \tfrac{\varrho}{2}\|u_k\|^2_{[P_1^k]^*}, \\ &\text{subject to} \qquad \langle \dot{y}_k, q_k\rangle_{L^2(0,T)} + \lambda_k \langle y_k, q_k\rangle_{L^2(0,T)} = \langle u_k, q_k\rangle_{(0,T)} \qquad \forall q_k \in P_1^k. \end{aligned} \tag{3.8}$$

Instead of the primal formulation, we can also state an optimal control problem based on the adjoint formulation. Then we aim to solve the minimization problem

$$\begin{aligned} &\min_{y \in Y_0, u \in [P_0]^*} J_0(y,u) = \tfrac{1}{2}\|y - y_d\|^2_{L^2(Q)} + \tfrac{\varrho}{2}\|u\|^2_{P_0^*}, \\ &\text{subject to} \qquad B_0 y = u \quad \text{ in } P_0^*. \end{aligned} \tag{3.9}$$

In order to obtain a sequence of decoupled control problems we introduce the spaces

$$\begin{aligned} Y_0^k &= L^2(0,T), \qquad \|y_k\|_{Y_0^k} = \sqrt{\lambda_k}\|y_k\|_{L^2(0,T)}, \\ P_0^k &= H_{,0}^1(0,T) \qquad \|p_k\|_{P_0^k} = \sqrt{\tfrac{1}{\lambda_k}\|\dot{p}_k\|^2_{L^2(0,T)} + \lambda_k \|p_k\|^2_{L^2(0,T)}}. \end{aligned} \tag{3.10}$$

Then the decoupled problems take the form

$$\begin{aligned} &\min_{y_k \in Y_0^k, u_k \in [P_0^k]^*} J_{0,k}(y_k,u_k) = \tfrac{1}{2}\|y_k - y_k^d\|^2_{L^2(0,T)} + \tfrac{\varrho}{2}\|u_k\|^2_{[P_0^k]^*}, \\ &\text{subject to} \qquad -\langle y_k, \dot{q}_k\rangle_{L^2(0,T)} + \lambda_k \langle y_k, q_k\rangle_{L^2(0,T)} = \langle u_k, q_k\rangle_{(0,T)} \qquad \forall q_k \in P_0^k. \end{aligned} \tag{3.11}$$

Finally, we can also establish a third optimal control framework, utilizing the interpolated formulation. We aim to minimize $J_{1/2} : Y_{1/2} \times P_{1/2} \to [0,\infty)$ given by

$$\begin{aligned} &\min_{y \in Y_{1/2}, u \in P_{1/2}^*} J_{1/2}(y,u) = \tfrac{1}{2}\|y - y_d\|^2_{L^2(Q)} + \tfrac{\varrho}{2}\|u\|^2_{P_{1/2}^*}, \\ &\text{subject to} \qquad B_{1/2} y = u \quad \text{ in } P_{1/2}^*. \end{aligned} \tag{3.12}$$

When using the interpolated formulation, we choose the decoupled state and adjoint state spaces $Y_{1/2}^k$ and $P_{1/2}^k$ as

$$\begin{aligned} Y_{1/2}^k &= H_{0,}^{1/2}(0,T), \qquad \|y_k\|_{Y_{1/2}^k} = \sqrt{\langle \dot{y}_k, \mathcal{H}_T y_k\rangle_{(0,T)} + \lambda_k \|y_k\|^2_{L^2(0,T)}}, \\ P_{1/2}^k &= H_{,0}^{1/2}(0,T), \qquad \|p_k\|_{P_{1/2}^k} = \sqrt{-\langle \dot{p}_k, \mathcal{H}_T^{-1} p_k\rangle_{(0,T)} + \lambda_k \|p_k\|^2_{L^2(0,T)}}, \end{aligned} \tag{3.13}$$

in order to obtain the sequence of decoupled problems

$$\begin{aligned} &\min_{y_k \in Y_{1/2}^k, u_k \in [P_{1/2}^k]^*} J_{1/2,k}(y_k,u_k) = \tfrac{1}{2}\|y_k - y_k^d\|^2_{L^2(0,T)} + \tfrac{\varrho}{2}\|u_k\|^2_{[P_{1/2}^k]^*}, \\ &\text{subject to} \qquad \langle \dot{y}_k, q_k\rangle_{(0,T)} + \lambda_k \langle y_k, q_k\rangle_{L^2(0,T)} = \langle u_k, q_k\rangle_{(0,T)} \qquad \forall q_k \in P_{1/2}^k. \end{aligned} \tag{3.14}$$

4 Optimal Control of a Single Mode

As shown in the last chapter, when working with a suitable decomposition of the state variable into eigenfunctions of the Laplace operator, it is sufficient to solve a sequence of decoupled optimal control problems. Instead of the heat equation, the state constraint is then reduced to a weak form of the initial value problem

$$\begin{aligned} \dot{y}_s(t) + \lambda y_s(t) &= u(t), \qquad t \in (0,T), \\ y_s(0) &= y_0, \end{aligned} \tag{4.1}$$

with λ representing the spatial eigenvalue associated to the considered mode. In this chapter we will analyze these decoupled problems in detail, and derive both closed loop and open loop control policies $u(t)$ in order to drive the state $y_s(t)$ towards a desired target $y_d(t)$. Since it is not obvious how to derive closed-loop policies when working with the interpolated and the adjoint formulation, we will, in contrast to the derivation within the last chapter, assume a non-vanishing initial value $y_0 \in \mathbb{R}$. By incorporating this initial value, the derived open-loop policies can at least be transformed to a closed-loop policy by means of Model Predictive Control (MPC). The basic idea of MPC is to repeatedly compute an optimal control, apply it for a short time interval, measure the new state, and then use this measured state as the initial condition to recompute the new optimal control for the next time interval, for more details see for example [15].

4.1 Primal Formulation

When working with the primal formulation, we call $y_s \in H^1(0,T)$ a weak solution of the initial value problem (4.1) if $y_s(0) = y_0$ and

$$\langle \dot{y}_s, q \rangle_{L^2(0,T)} + \lambda \langle y_s, q \rangle_{L^2(0,T)} = \langle u, q \rangle_{(0,T)}, \qquad \forall q \in L^2(0,T).$$

Let $y_e \in H^1(0,T)$ be some extension of the initial value, meaning that $y_e(0) = y_0$, then we can also write the state as $y_s = y + y_e$ with $y \in H^1_{0,}(0,T)$. With this representation of the state, we can establish another equivalent variational formulation for finding $y \in H^1_{0,}(0,T)$ such that

$$\langle \dot{y}, q \rangle_{L^2(0,T)} + \lambda \langle y, q \rangle_{L^2(0,T)} = \langle u, q \rangle_{(0,T)} - \Big(\langle \dot{y}_e, q \rangle_{L^2(0,T)} + \lambda \langle y_e, q \rangle_{L^2(0,T)} \Big) \tag{4.2}$$

holds for all $q \in L^2(0,T)$. At this point we introduce we introduce the state space $Y_1 = H^1_{0,}(0,T)$ and the adjoint state space $P_1 = L^2(0,T)$ and equip them with the norms

$$\|y\|_{Y_1} = \sqrt{\tfrac{1}{\lambda}\|\dot{y}\|^2_{L^2(0,T)} + \lambda \|y\|^2_{L^2(0,T)}}, \qquad \|p\|_{P_1} = \sqrt{\lambda}\|p\|_{L^2(0,T)},$$

which we have already defined in (3.7). These norms are induced by the linear operators $D_1 : Y_1 \to Y_1^*$ and $A_1 : P_1 \to P_1^*$ defined by

$$\begin{aligned} \langle D_1 y, z\rangle_{(0,T)} &= \tfrac{1}{\lambda}\langle \dot{y}, \dot{z}\rangle_{L^2(0,T)} + \lambda \langle y, z\rangle_{L^2(0,T)}, \\ A_1 p &= \lambda p, \end{aligned}$$

in the sense that $\|y\|_{Y_1} = \langle D_1 y, y\rangle_{(0,T)}^{1/2}$ and $\|p\|_{P_1} = \langle A_1 p, p\rangle_{(0,T)}^{1/2}$. Note that due to the Theorem of Fréchet–Riesz, we can identify both P_1 and P_1^* with $L^2(0,T)$, but the respective norms differ, since $\|p\|_{P_1} = \sqrt{\lambda}\|p\|_{L^2(0,T)}$ and $\|u\|_{P_1^*} = \frac{1}{\sqrt{\lambda}}\|u\|_{L^2(0,T)}$.

Based on the variational formulation (4.2) we define $B_1 : Y_1 \to P_1^*$ and $f_1 \in P_1^*$ by

$$\begin{aligned} B_1 y &= \dot{y} + \lambda y, \qquad \forall y \in H^1_{0,}(0,T), \\ f_1 &= -\left(\dot{y}_e + \lambda y_e\right). \end{aligned}$$

Then the variational formulation (4.2) is equivalent to the operator equation

$$B_1 y = u + f_1 \in P_1^*. \tag{4.3}$$

The following Lemma states well-posedness of this operator equation, and is a direct consequence of Lemma 3.1, Lemma 1.9 and Theorem 1.1.

Lemma 4.1. *The operator $B_1 : Y_1 \to P_1^*$ is an isomorphism that satisfies*

$$c_1^{B_1}\|y\|_{Y_1} \le \|B_1 y\|_{P_1^*} \le c_2^{B_1}\|y\|_{Y_1}, \qquad \forall y \in Y_1,$$

with $c_1^{B_1} = 1$ and $c_2^{B_1} = \sqrt{2}$. In particular, the operator equation (4.3) is uniquely solvable for every given $u \in P_1^*$.

Based on the operator equation (4.3) we can now introduce an optimal control problem similar to (3.8), which is able to incorporate a non-vanishing initial value $y_0 \in \mathbb{R}$. This new control problem takes the form

$$\begin{aligned} \min_{y\in Y_1, u\in P_1^*} \; & J_1(y,u) = \tfrac{1}{2}\|y + y_e - y_d\|^2_{L^2(0,T)} + \tfrac{\varrho}{2}\|u\|^2_{P_1^*}, \\ \text{subject to} \quad & B_1 y = u + f_1 \qquad \text{in } P_1^*. \end{aligned} \tag{4.4}$$

Following the theory from Section 2.4, we know that the optimal state y_s is given as the pointwise sum of the extension y_e and the unique solution $y \in Y_1$ of

$$y + \varrho S_1 y = \tilde{y}_d, \tag{4.5}$$

where $S_1 = B_1^\top A_1^{-1} B_1 : Y_1 \to Y_1^*$ and $\tilde{y}_d = y_d - y_e + \varrho B_1^\top A_1^{-1} f_1 \in Y_1^*$. Note that Lemma 2.4 and Lemma 4.1 already imply that S_1 is elliptic and bounded, with $c_1^{S_1} = (c_1^{B_1})^2 = 1$ and $c_2^{S_1} = (c_2^{B_1})^2 = 2$ respectively. Thus S is an isomorphism, and Equation (4.5) admits a unique solution $y \in Y_1$. As the next Lemma demonstrates, the ellipticity constant $c_1^{S_1} = 1$ is sharp, while $c_2^{S_1} = 2$ is not.

Lemma 4.2. *The operator $S_1 : Y_1 \to Y_1^*$ is elliptic with $c_1^{S_1} = 1$ and bounded with $c_2^{S_1} = 1 + \tanh(\lambda T) \leq 2$.*

Proof: According to Lemma 1.2 and Lemma 2.4, the minimal and maximal eigenvalue of the generalized eigenvalue problem

$$S_1 y = B_1^\top A_1^{-1} B_1 y \stackrel{!}{=} \mu D_1 y, \tag{4.6}$$

is equal to $c_1^{S_1}$ and $c_2^{S_1}$ respectively. Note that (4.6) is equivalent to

$$\left\langle \dot{y} + \lambda y, \tfrac{1}{\lambda}\dot{v} + v \right\rangle_{L^2(0,T)} = \tfrac{\mu}{\lambda}\langle \dot{y}, \dot{v}\rangle_{L^2(0,T)} + \mu\lambda\langle y, v\rangle_{L^2(0,T)}, \qquad \forall v \in L^2(0,T).$$

Rearranging this equation leads to

$$\left\langle \tfrac{1}{\lambda}(1-\mu)\,\dot{y} + y, \dot{v}\right\rangle_{L^2(0,T)} = \langle \lambda(\mu-1)y - \dot{y}, v\rangle_{L^2(0,T)}. \tag{4.7}$$

Hence we conclude $\frac{1}{\lambda}(1-\mu)\,\dot{y} + y \in H^1(0,T)$, further implying $y \in H^2(0,T)$. By definition of Y_1 we also know $y(0) = 0$, and using partial integration in Equation (4.7), and testing with a sequence of test functions with $v(T) = 1$ and a support that vanishes in the limit yields the terminal condition

$$(1-\mu)\dot{y}(T) + \lambda y(T) = 0.$$

What remains from Equation (4.7) after partial integration is

$$-\left\langle \tfrac{1}{\lambda}(1-\mu)\ddot{y} + \dot{y}, v\right\rangle_{L^2(0,T)} = \langle \lambda(\mu-1)y - \dot{y}, v\rangle_{L^2(0,T)}, \quad \forall v \in L^2(0,T)$$

Note that the term $-\langle \dot{y}, v\rangle_{L^2(0,T)}$ appears on both sides, thus it can be eliminated. Rearranging the rest of the equation and multiplying it with λ we conclude that y has to satisfy

$$(1-\mu)\left(\ddot{y}(t) - \lambda^2 y(t)\right) = 0$$

for almost all $t \in [0,T]$. For $\mu = 1$ this equation is automatically satisfied, and the terminal condition reduces to $y(T) = 0$, hence every $y \in H^2(0,T)$ with $y(0) = y(T) = 0$ is a solution to this eigenvalue problem with $\mu = 1$. If $\mu \neq 1$ we deduce that y is a solution of the boundary value problem

$$\begin{aligned} \ddot{y}(t) &= \lambda^2 y(t), \quad t \in (0,T), \\ y(0) &= 0, \\ (1-\mu)\dot{y}(T) + \lambda y(T) &= 0. \end{aligned}$$

The differential equation together with the initial condition $y(0) = 0$ immediately imply that y is of the form $y(t) = C \cdot \sinh(\lambda t)$ with $C \neq 0$. Then the terminal condition implies

$$0 = (1-\mu)\dot{y}(T) + \lambda y(T) = C\left((1-\mu)\lambda\cosh(\lambda T) + \lambda\sinh(\lambda T)\right),$$

thus division by $\cosh(\lambda T)$ shows that $\mu = 1 + \tanh(\lambda T) \leq 2$ is the only other eigenvalue besides 1. □

In the rest of this subsection we will prove error and control cost estimates related to this optimal state $y \in Y_1$, which are summarized in the following theorem.

Theorem 4.3. *For a given target $y_d \in L^2(0,T)$ and $y_0 \in \mathbb{R}$ we denote by $y_e(t) = y_0 e^{-\lambda t}$ the homogeneous extension of the initial value, and by $y \in Y_1$ the associated solution of the operator equation (4.5). Then it holds that*

$$\|y + y_e - y_d\|_{L^2(0,T)} \leq \|y_d\|_{L^2(0,T)} + y_0 \sqrt{\tfrac{1-e^{-2\lambda T}}{2\lambda}}$$
$$\|u\|_{Y_1^*} = \tfrac{1}{\sqrt{\lambda}} \|u\|_{L^2(0,T)} \leq \varrho^{-1/2} \left(\|y_d\|_{L^2(0,T)} + y_0 \sqrt{\tfrac{1-e^{-2\lambda T}}{2\lambda}} \right)$$

If we further assume $y_d \in H^1(0,T)$ together with the initial condition $y_d(0) = y_0$ it holds that

$$\|y + y_e - y_d\|_{L^2(0,T)} \leq \varrho^{1/2} \tfrac{1}{\sqrt{\lambda}} \|\dot{y}_d + \lambda y_d\|_{L^2(0,T)}$$
$$\|u\|_{Y_1^*} = \tfrac{1}{\sqrt{\lambda}} \|u\|_{L^2(0,T)} \leq \tfrac{1}{\sqrt{\lambda}} \|\dot{y}_d + \lambda y_d\|_{L^2(0,T)}.$$

If the target satisfies the stronger condition $y_d \in H^2(0,T)$ together with the initial condition $y_d(0) = y_0$ and the terminal condition $\dot{y}_d(T) + \lambda y_d(T) = 0$, then it holds that

$$\|y + y_e - y_d\|_{L^2(0,T)} \leq \varrho \|\lambda y_d - \tfrac{1}{\lambda} \ddot{y}_d\|_{L^2(0,T)}.$$

We will postpone the proof of this Lemma, and instead begin by introducing some necessary auxiliary results. The proof of Theorem 4.3 will be based on the abstract error and control cost estimates from Corollary 2.7, therefore we begin by finding suitable representations of $\|y\|_{S_1}$ and $\|S_1 y\|_{L^2(0,T)}$ for a given $y \in Y_1$. The norm $\|\cdot\|_{S_1}$ is given by

$$\|y\|_{S_1} = \sqrt{\langle S_1 y, y \rangle_{(0,T)}} = \left(\langle \dot{y} + \lambda y, \tfrac{1}{\lambda} \dot{y} + y \rangle_{L^2(0,T)} \right)^{1/2} = \tfrac{1}{\sqrt{\lambda}} \|\dot{y} + \lambda y\|_{L^2(0,T)},$$

and it is equivalent to the standard norm $\|\cdot\|_{Y_1}$ due to Lemma 4.2. Next, let $y \in Y_1$ and assume that $S_1 y \in L^2(0,T)$, then there exists a $q \in L^2(0,T)$ such that

$$\langle S_1 y, v \rangle_{(0,T)} = \langle \tfrac{1}{\lambda} \dot{y} + y, \dot{v} + \lambda v \rangle_{L^2(0,T)} = \langle q, v \rangle_{L^2(0,T)}$$

holds for all $v \in Y_1$. Rearranging this equation yields

$$\langle \tfrac{1}{\lambda} \dot{y} + y, \dot{v} \rangle_{L^2(0,T)} = \langle q - \dot{y} - \lambda y, v \rangle_{L^2(0,T)}. \tag{4.8}$$

We conclude that $\frac{1}{\lambda} \dot{y} + y \in H^1(0,T)$, which further implies $y \in H^2(0,T)$, together with

$$\tfrac{1}{\lambda} \ddot{y} = \lambda y - q.$$

Hence, partial integration in Equation (4.8) and testing with suitable test functions, similar to the proof in Lemma 4.2, shows that additionally to $y(0) = 0$ a terminal condition of the form $\dot{y}(T) + \lambda y(T) = 0$ has to hold. Thus $S_1 y \in L^2(0,T)$ if and only if $y \in H^2(0,T)$ with $y(0) = 0$ and $\dot{y}(T) + \lambda y(T) = 0$, and further

$$\|S_1 y\|_{L^2(0,T)} = \|q\|_{L^2(0,T)} = \|\lambda y - \tfrac{1}{\lambda} \ddot{y}\|_{L^2(0,T)}.$$

Now we are in a position to prove appropriate error and control cost estimates.

Proof of Theorem 4.3: We know that $y_e(t) = y_0 e^{-\lambda t}$ is in $L^2(0,T)$, and that $f_1 = -(\dot{y}_e + \lambda y_e) = 0 \in L^2(0,T)$. Assuming $y_d \in L^2(0,T)$, the estimates from (i) of Corollary 2.7 reduce to

$$\|y + y_e - y_d\|_{L^2(0,T)} \leq \|y_d\|_{L^2(0,T)} + \|y_e\|_{L^2(0,T)},$$
$$\|u\|_{P_1^*} \leq \varrho^{-1/2} \left(\|y_d\|_{L^2(0,T)} + \|y_e\|_{L^2(0,T)} \right).$$

Then the first two estimates claimed in the theorem follow from

$$\|y_e\|_{L^2(0,T)} = \sqrt{\int_0^T (y_0 e^{-\lambda t})^2 \, \mathrm{d}t} = y_0 \sqrt{\frac{1 - e^{-2\lambda T}}{2\lambda}}.$$

Now assume that $y_d \in H^1(0,T)$ with $y_d(0) = y_0$, then $\tilde{y}_d = y_d - y_e \in H^1_{0,}(0,T) = Y_1$, such that the estimates from (ii) of Corollary 2.7 reduce to

$$\|y + y_e - y_d\|_{L^2(0,T)} \leq \varrho^{1/2} \|\tilde{y}_d\|_{S_1},$$
$$\|u\|_{P_1^*} \leq \|\tilde{y}_d\|_{S_1}.$$

Thus the third and the fourth estimate claimed in the theorem follow from

$$\|\tilde{y}_d\|_{S_1} = \tfrac{1}{\sqrt{\lambda}} \|\dot{y}_d + \lambda y_d - (\dot{y}_e + \lambda y_e)\|_{L^2(0,T)} = \tfrac{1}{\sqrt{\lambda}} \|\dot{y}_d + \lambda y_d\|_{L^2(0,T)}.$$

Finally assume that the target satisfies $y_d \in H^2(0,T)$ together with $y_d(0) = y_0$ and $\dot{y}_d(T) + \lambda y_d(T) = 0$. Since $y_e \in H^2(0,T)$, we conclude that $\tilde{y}_d = y_d - y_e \in H^2(0,T)$ with $\tilde{y}_d(0) = 0$ and $\dot{\tilde{y}}(T) + \lambda \tilde{y}_d(T) = 0$. Thus $S_1 \tilde{y}_d \in L^2(0,T)$, and the estimate from (iii) of Corollary 2.7 reduces to

$$\|y + y_e - y_d\|_{L^2(0,T)} \leq \varrho \|S_1 \tilde{y}_d\|_{L^2(0,T)}.$$

To compute the norm on the right hand side, note that

$$\lambda y_e - \tfrac{1}{\lambda} \ddot{y}_e = -\tfrac{1}{\lambda} \left(\ddot{y}_e - \lambda^2 y_e \right) = -\tfrac{1}{\lambda} \left(\tfrac{\mathrm{d}}{\mathrm{d}t} - \lambda \right) \underbrace{(\dot{y}_e + \lambda y_e)}_{=0} = 0,$$

thus the last estimate claimed in the theorem follows from

$$\|S_1 \tilde{y}_d\|_{L^2(0,T)} = \|\lambda y_d - \tfrac{1}{\lambda} \ddot{y}_d - \left(\lambda y_e - \tfrac{1}{\lambda} \ddot{y}_e \right)\|_{L^2(0,T)} = \|\lambda y_d - \tfrac{1}{\lambda} \ddot{y}_d\|_{L^2(0,T)}. \qquad \square$$

Remark 4.4. *As Theorem 4.3 shows, the rate at which the state error converges to 0 as $\varrho \to 0$ depends on the regularity of the desired target y_d. The higher the regularity of the target, the higher the convergence rate, which is linear at best if y_d is in $H^2(0,T)$ and satisfies an appropriate initial and terminal condition. Further, if the target y_d is at least in $H^1(0,T)$, the control costs stay bounded as $\varrho \to 0$, but they tend to infinity if y_d admits a regularity lower than that. The rate at which the control costs tend to infinity increases as the regularity of the target decreases. Finally note that even though all error estimates are proven using the specific extension $y_e(t) = y_0 e^{-\lambda t}$, similar error estimates can be derived for any other chosen extension, provided the resulting modified target $\tilde{y}_d$ remains sufficiently regular.*

4.2 Adjoint Formulation

If we use the adjoint formulation instead of the primal formulation, we call $y \in L^2(0,T)$ a weak solution of the initial value problem (4.1) if

$$-\langle y, \dot{q}\rangle_{L^2(0,T)} + \lambda\langle y, q\rangle_{L^2(0,T)} = \langle u, q\rangle_{(0,T)} + y_0 q(0), \forall q \in H^1_{,0}(0,T), \tag{4.9}$$

where the term $y_0 q(0)$ on the right-hand side stems from the partial integration step. We define the respective state space $Y_0 = L^2(0,T)$ and the adjoint state space $P_0 = H^1_{,0}(0,T)$ and equip them with the norms

$$\|y\|_{Y_0} = \sqrt{\lambda}\|y\|_{L^2(0,T)}, \qquad \|p\|_{P_0} = \sqrt{\tfrac{1}{\lambda}\|\dot{p}\|^2_{L^2(0,T)} + \lambda\|p\|^2_{L^2(0,T)}},$$

which are induced by the operators $D_0 : Y_0 \to Y_0^*$ and $A_0 : P_0 \to P_0^*$ defined by

$$\begin{aligned} D_0 y &= \lambda y, \\ \langle A_0 p, q\rangle_{(0,T)} &= \tfrac{1}{\lambda}\langle \dot{p}, \dot{q}\rangle_{L^2(0,T)} + \lambda\langle p, q\rangle_{L^2(0,T)}. \end{aligned}$$

Based on the variational formulation (4.9) we define $B_0 : Y_0 \to P_0^*$ and $f_0 \in P_0^*$ by

$$\begin{aligned} \langle B_0 y, q\rangle_{(0,T)} &= -\langle y, \dot{q}\rangle_{L^2(0,T)} + \lambda\langle y, q\rangle_{L^2(0,T)}, && \forall y \in Y_0, \forall p \in P_0, \\ f_0(q) &= y_0 q(0), && \forall q \in P_0. \end{aligned}$$

Then the variational formulation (4.9) is equivalent to the operator equation

$$B_0 y = u + f_0 \in P_0^*. \tag{4.10}$$

The following Lemma is a direct consequence of Lemma 3.2, Lemma 1.9 and Theorem 1.1.

Lemma 4.5. *The operator $B_0 : Y_0 \to P_0^*$ is an isomorphism that satisfies*

$$c_1^{B_0}\|y\|_{Y_0} \le \|B_0 y\|_{P_0^*} \le c_2^{B_0}\|y\|_{Y_0}, \qquad \forall y \in Y_0,$$

with $c_1^{B_0} = 1$ and $c_2^{B_0} = \sqrt{2}$. In particular, the operator equation (4.10) is uniquely solvable for every given $u \in P_0^*$.

Based on the operator equation (4.10) and the control problem (3.8) we now consider the optimal control problem

$$\begin{aligned} \min_{y \in Y_0, u \in P_0^*} J_0(y,u) &= \frac{1}{2}\|y - y_d\|^2_{L^2(0,T)} + \frac{\varrho}{2}\|u\|^2_{P_0^*}, \\ \text{subject to} \qquad B_0 y &= u + f_0 \qquad \text{in } P_0^*. \end{aligned} \tag{4.11}$$

Following the theory from Section 2.4 we define $S_0 = B_0^\top A_0^{-1} B_0 : Y_0 \to Y_0^*$, then the optimal state $y \in Y_0$ is defined as the solution of

$$y + \varrho S_0 y = \tilde{y}_d \qquad \text{in } Y_0^*, \tag{4.12}$$

with $\tilde{y}_d = y_d + \varrho B_0^\top A_0^{-1} f_0$. The next Lemma states that S_0 is an isomorphism, thus the operator equation (4.12) is uniquely solvable.

Lemma 4.6. *The operator $S_0 : Y_0 \to Y_0^*$ is elliptic with $c_1^{S_0} = 1$ and bounded with $c_2^{S_0} = 1 + \tanh(\lambda T) \leq 2$.*

Proof: Denoting by R_T the time flipping operator, we have $B_0 = R_T^\top B_1^\top R^\top$, thus

$$\begin{aligned} c_1^{S_0} &= (c_1^{B_0})^2 = (c_1^{B_1})^2 = c_1^{S_1} = 1, \\ c_2^{S_0} &= (c_2^{B_0})^2 = (c_2^{B_1})^2 = c_2^{S_1} = 1 + \tanh(\lambda T). \end{aligned}$$ □

The rest of the subsection we will prove the following theorem which provides an estimate for the state error and the control cost.

Theorem 4.7. *Let $y_d \in L^2(0,T)$ and denote by $y \in Y_0$ the associated solution of (4.12), then it holds that*

$$\begin{aligned} \|y - y_d\|_{L^2(0,T)} &\leq \varrho \left(\|S_0 y_d\|_{L^2(0,T)} + y_0(1+2\lambda)(1+\tanh(\lambda T))\sqrt{\tfrac{\lambda(1-e^{-2\lambda T})}{2}} \right) \\ \|u\|_{P_0^*} &\leq y_0 \sqrt{\tanh(\lambda T)} + \|y_d\|_{S_0}. \end{aligned}$$

We will postpone the proof and state some auxiliary results first. In comparison to the primal formulation, it is more intricate to find a simple representation of S_0 due to the operator inversion of A_0 being non-trivial in comparison to A_1. Still, it is possible to find a very compact representation. For this purpose we choose $y \in Y_0$ and define $p = A_0^{-1} B_0 y$, then $S_0 y = B_0^\top p$ and

$$\langle A_0 p, q \rangle_{(0,T)} = \langle B_0 y, q \rangle_{(0,T)} \tag{4.13}$$

holds for all $q \in P_0$. Further, note that using the Fréchet–Riesz isomorphism, we can interpret $B_0^\top : P_0 \to Y_0^*$ as the mapping $p \to \lambda p - \dot{p}$, therefore

$$\begin{aligned} \left\langle B_0^\top p, B_0^\top q \right\rangle_{L^2(0,T)} &= \langle \lambda p - \dot{p}, \lambda q - \dot{q} \rangle_{L^2(0,T)} \\ &= \lambda^2 \langle p, q \rangle_{L^2(0,T)} + \langle \dot{p}, \dot{q} \rangle_{L^2(0,T)} - \lambda \left(\langle \dot{p}, q \rangle_{L^2(0,T)} + \langle p, \dot{q} \rangle_{L^2(0,T)} \right) \\ &= \lambda \langle A_0 p, q \rangle_{(0,T)} + \lambda p(0) q(0). \end{aligned}$$

In particular we conclude

$$\langle A_0 p, q \rangle_{(0,T)} = \frac{1}{\lambda} \left\langle B_0^\top p, B_0^\top q \right\rangle_{L^2(0,T)} - p(0) q(0). \tag{4.14}$$

Further, note that for $v_e(t) = e^{-\lambda t}$ we have $v_e \in L^2(0,T)$ and

$$\langle B_0^\top p, v_e \rangle_{L^2(0,T)} = \langle \lambda p - \dot{p}, v_e \rangle_{L^2(0,T)} = \langle p, \lambda v_e + \dot{v}_e \rangle_{L^2(0,T)} - \underbrace{p(T)}_{=0} v_e(T) + p(0) \underbrace{v_e(0)}_{=1} = p(0). \tag{4.15}$$

Combining Equations (4.13), (4.14) and (4.15) we conclude that

$$\begin{aligned}\left\langle y, B_0^\top q\right\rangle_{L^2(0,T)} &= \langle B_0 y, q\rangle_{L^2(0,T)} = \langle A_0 p, q\rangle_{(0,T)} = \frac{1}{\lambda}\left\langle B_0^\top p, B_0^\top q\right\rangle_{L^2(0,T)} - p(0)q(0)\\ &= \frac{1}{\lambda}\left\langle B_0^\top p, B_0^\top q\right\rangle_{L^2(0,T)} - \left\langle B_0^\top p, v_e\right\rangle_{L^2(0,T)}\left\langle B_0^\top q, v_e\right\rangle_{L^2(0,T)}\end{aligned}$$

holds for all $q \in P_0$. Since $B_0^\top$ can be interpreted as an isomorphism between the spaces $P_0 = H^1_{,0}(0,T)$ and $Y_0 = L^2(0,T)$, we can write $v = B_0^\top q \in L^2(0,T)$ and rewrite the last variational formulation to

$$\langle y, v\rangle_{L^2(0,T)} = \frac{1}{\lambda}\left\langle B_0^\top p, v\right\rangle_{L^2(0,T)} - \left\langle B_0^\top p, v_e\right\rangle_{L^2(0,T)}\langle v, v_e\rangle_{L^2(0,T)}, \quad \forall v \in L^2(0,T).$$

But since $S_0 y = B_0^\top p$ this means that

$$\langle S_0 y, v\rangle_{(0,T)} = \lambda\langle y, v\rangle_{L^2(0,T)} + \lambda\langle S_0 y, v_e\rangle_{L^2(0,T)}\langle v, v_e\rangle_{L^2(0,T)} \tag{4.16}$$

holds for all $y, v \in Y_0$, so it only remains to compute $\langle S_0 y, v_e\rangle_{L^2(0,T)}$. For this purpose we just plug $v = v_e$ into Equation (4.16) to obtain

$$\langle y, v_e\rangle_{L^2(0,T)} = \langle S_0 y, v_e\rangle_{(0,T)}\left(\tfrac{1}{\lambda} - \|v_e\|^2_{L^2(0,T)}\right).$$

Note that $\|v_e\|^2_{L^2(0,T)} = \int_0^T e^{-2\lambda t}\mathrm{d}t = \frac{1-e^{-2\lambda T}}{2\lambda}$, hence

$$\begin{aligned}\langle S_0 y, v_e\rangle_{(0,T)} &= \langle y, v_e\rangle_{L^2(0,T)}\left(\tfrac{1}{\lambda} - \tfrac{1-e^{-2\lambda T}}{2\lambda}\right)^{-1}\\ &= \langle y, v_e\rangle_{L^2(0,T)}\frac{2\lambda}{1+e^{-2\lambda T}} = \langle y, v_e\rangle_{L^2(0,T)}\frac{\lambda e^{\lambda T}}{\cosh(\lambda T)}.\end{aligned}$$

Thus, defining $g(t) = \sqrt{\frac{\lambda e^{\lambda T}}{\cosh(\lambda T)}}v_e(t)$ we can finally conclude that

$$\begin{aligned}\langle S_0 y, v\rangle_{(0,T)} &= \lambda\langle y, v\rangle_{L^2(0,T)} + \tfrac{\lambda^2 e^{\lambda T}}{\cosh(\lambda T)}\langle y, v_e\rangle_{L^2(0,T)}\langle v, v_e\rangle_{L^2(0,T)}\\ &= \lambda\left(\langle y, v\rangle_{L^2(0,T)} + \langle y, g\rangle_{L^2(0,T)}\langle v, g\rangle_{L^2(0,T)}\right).\end{aligned}$$

In particular, if we again use the Fréchet–Riesz isomorphism to interpret S_0 as a map $S_0 : L^2(0,T) \to L^2(0,T)$, we have found that

$$(S_0 y)(t) = \lambda\left(y(t) + \langle y, g\rangle_{L^2(0,T)}g(t)\right). \tag{4.17}$$

In order to prove the estimates from Theorem 4.7 using Corollary 2.7 we will need to compute $\|y\|_{S_0}$ and $\|S_0 y\|_{L^2(0,T)}$ for a given $y \in L^2(0,T)$. Due to the representation from Equation (4.17) we already know that

$$\|y\|_{S_0} = \sqrt{\langle S_0 y, y\rangle_{(0,T)}} = \sqrt{\lambda}\sqrt{\|y\|^2_{L^2(0,T)} + \langle y, g\rangle^2_{L^2(0,T)}}$$

and

$$\begin{aligned}\|S_0 y\|^2_{L^2(0,T)} &= \langle S_0 y, S_0 y\rangle_{L^2(0,T)} = \lambda^2 \Big\langle \Big(y + \langle y, g\rangle_{L^2(0,T)} g\Big), \Big(y + \langle y, g\rangle_{L^2(0,T)} g\Big)\Big\rangle_{L^2(0,T)} \\ &= \lambda^2 \left(\|y\|^2_{L^2(0,T)} + 2\langle y, g\rangle^2_{L^2(0,T)} + \langle y, g\rangle^2_{L^2(0,T)} \|g\|^2_{L^2(0,T)}\right) \\ &= \lambda^2 \left(\|y\|^2_{L^2(0,T)} + \langle y, g\rangle^2_{L^2(0,T)} \left(2 + \|g\|^2_{L^2(0,T)}\right)\right).\end{aligned}$$

So using that

$$\|g\|^2_{L^2(0,T)} = \frac{\lambda e^{\lambda T}}{\cosh(\lambda T)} \int_0^T e^{-2\lambda t}\, \mathrm{d}t = \frac{\lambda e^{\lambda T}}{\cosh(\lambda T)} \frac{1 - e^{-2\lambda T}}{\lambda} = 2\tanh(\lambda T),$$

we obtain the norm representation

$$\begin{aligned}\|S_0 y\|_{L^2(0,T)} &= \lambda \sqrt{\|y\|^2_{L^2(0,T)} + 2\langle y, g\rangle^2_{L^2(0,T)} \left(1 + \tanh(\lambda T)\right)} \\ &= \lambda \sqrt{\|y\|^2_{L^2(0,T)} + 2\lambda \langle y, e^{-\lambda t}\rangle^2_{L^2(0,T)} \left(1 + \tanh(\lambda T)\right)^2},\end{aligned}$$

where have expanded g using its definition and used the formula $\frac{e^x}{\cosh(x)} = (1 + \tanh(x))$ in the last step. Now we have stated all necessary auxiliary results for the proof of Theorem 4.7.

Proof of Theorem 4.7: Let as assume that $y_d \in L^2(0,T)$ and denote by $y \in Y_0$ the solution of Equation (4.12). Then Corollary 2.7 and the triangle inequality already imply

$$\begin{aligned}\|y - y_d\|_{L^2(0,T)} &\leq \varrho \left(\|S_0 y_d\|_{L^2(0,T)} + \|S_0 B_0^\top A_0^{-1} f_0\|_{L^2(0,T)} + \|B_0^\top A_0^{-1} f_0\|_{L^2(0,T)}\right) \\ \|u\|_{P_0^*} &\leq \|f_0\|_{P_0^*} + \|y_d\|_{S_0}.\end{aligned}$$

We know the explicit form of $\|y_d\|_{S_0}$ and $\|S_0 y_d\|_{L^2(0,T)}$, so it only remains to compute all the terms involving f_0. First of all let us define $p = A_0^{-1} f_0$, then $A_0 p = f_0$, which is equivalent to

$$\tfrac{1}{\lambda}\langle \dot{p}, \dot{q}\rangle_{L^2(0,T)} + \lambda \langle p, q\rangle_{L^2(0,T)} = y_0 q(0) \quad \forall q \in P_0.$$

We conclude that $p \in H^2(0,T)$ with $p(T) = 0$ and $\dot{p}(0) = -\lambda y_0$, and that

$$\ddot{p}(t) = \lambda^2 p(t).$$

Then the standard arguments for linear differential equations imply

$$p(t) = y_0 \left(\cosh(\lambda t)\tanh(\lambda T) - \sinh(\lambda t)\right).$$

Hence

$$\|f_0\|_{P_0^*} = \sqrt{\langle f_0, A_0^{-1} f_0\rangle_{(0,T)}} = \sqrt{y_0 p(0)} = y_0 \sqrt{\tanh(\lambda T)}.$$

Further, we can interpret $\phi = B_0^\top A_0^{-1} f_0$ as an element of $L^2(0,T)$, with

$$\begin{aligned}\phi(t) &= (B_0^\top A_0^{-1} f_0)(t) = (B_0^\top p)(t) = \lambda p(t) - \dot{p}(t)\\ &= \lambda y_0 \left(\cosh(\lambda t)\tanh(\lambda T) - \sinh(\lambda t) - \sinh(\lambda t)\tanh(\lambda T) + \cosh(\lambda t)\right)\\ &= \lambda y_0 \left(\cosh(\lambda t) - \sinh(\lambda t)\right)\left(1 + \tanh(\lambda T)\right) = y_0 \lambda e^{-\lambda t}(1 + \tanh(\lambda T)).\end{aligned}$$

Computing the $L^2(0,T)$-Norm of this function results in

$$\|B_0^\top A_0^{-1} f_0\|_{L^2(0,T)} = y_0\lambda(1+\tanh(\lambda T))\|e^{-\lambda t}\|_{L^2(0,T)} = y_0(1+\tanh(\lambda T))\sqrt{\tfrac{\lambda(1-e^{-2\lambda T})}{2}}.$$

Finally note that $S_0 B_0^\top A_0^{-1} f_0 = S_0\phi$ and

$$\begin{aligned}\|S_0\phi\|^2_{L^2(0,T)} &= \langle S_0\phi, S_0\phi\rangle_{L^2(0,T)} = \left\langle S_0\left(S_0^{1/2}\phi\right), S_0^{1/2}\phi\right\rangle_{(0,T)} = \|S_0^{1/2}\phi\|^2_{S_0}\\ &\le c_2^{S_0}\|S_0^{1/2}\phi\|^2_{Y_0} \le 2\lambda\|S_0^{1/2}\phi\|^2_{L^2(0,T)}\\ &= 2\lambda\left\langle S_0^{1/2}\phi, S_0^{1/2}\phi\right\rangle_{L^2(0,T)} = 2\lambda\langle S_0\phi,\phi\rangle_{(0,T)} = 2\lambda\|\phi\|^2_{S_0} \le 4\lambda^2\|\phi\|^2_{L^2(0,T)},\end{aligned}$$

which completes the proof. □

Remark 4.8. *If we compare Theorem 4.7 to Theorem 4.3, the most striking difference is that when using the adjoint formulation, it is sufficient to assume $y_d \in L^2(0,T)$ in order to get bounded control costs and a state error that converges to 0 with a linear convergence rate as $\varrho \to 0$. Moreover, using the adjoint formulation allows to set $\varrho = 0$ and still obtain a well-posed problem.*

4.3 Interpolated Formulation

Finally, when we work with the interpolated formulation, we call $y_s \in H^{1/2}(0,T)$ a weak solution of the initial value problem (4.1) if it can be written as $y_s = y + y_e$ with $y_e \in H^1(0,T)$ satisfying $y_e(0) = y_0$ and $y \in H^{1/2}_{0,}(0,T)$ such that

$$\langle \dot{y}, q\rangle_{(0,T)} + \lambda\langle y, q\rangle_{L^2(0,T)} = \langle u, q\rangle_{(0,T)} - \left(\langle \dot{y}_e, q\rangle_{L^2(0,T)} + \lambda\langle y_e, q\rangle_{L^2(0,T)}\right) \tag{4.18}$$

holds for all $q \in H^{1/2}_{,0}(0,T)$. We define the associated state space $Y_{1/2} = H^{1/2}_{0,}(0,T)$ and the adjoint state space $P_{1/2} = H^{1/2}_{,0}(0,T)$ and equip them with the norms

$$\|y\|_{Y_{1/2}} = \sqrt{\langle \dot{y}, \mathcal{H}_T y\rangle_{(0,T)} + \lambda\|y\|^2_{L^2(0,T)}}, \qquad \|p\|_{P_{1/2}} = \sqrt{-\langle \dot{p}, \mathcal{H}_T^{-1} p\rangle_{(0,T)} + \lambda\|p\|^2_{L^2(0,T)}},$$

which are itself induced by the operators $A_{1/2} : P_{1/2} \to P^*_{1/2}$ and $D_{1/2} : Y_{1/2} \to Y^*_{1/2}$ defined by

$$\left\langle A_{1/2}p, q\right\rangle_{(0,T)} = -\left\langle \dot{p}, \mathcal{H}_T^{-1} q\right\rangle_{(0,T)} + \lambda\langle p, q\rangle_{L^2(0,T)}, \tag{4.19}$$

$$\left\langle D_{1/2}y, z\right\rangle_{(0,T)} = \langle \dot{y}, \mathcal{H}_T z\rangle_{(0,T)} + \lambda\langle y, z\rangle_{L^2(0,T)}. \tag{4.20}$$

Then, based on the variational formulation (4.18), we define $B_{1/2} : Y_{1/2} \to P_{1/2}^*$ and $f_{1/2} \in P_{1/2}^*$ by

$$\begin{aligned} \langle B_{1/2} y, q \rangle_{(0,T)} &= \langle \dot{y}, q \rangle_{(0,T)} + \lambda \langle y, q \rangle_{L^2(0,T)}, \quad \forall y \in Y_{1/2}, \forall p \in P_{1/2}, \\ f_{1/2}(q) &= -\langle \dot{y}_e, q \rangle_{L^2(0,T)} - \lambda \langle y_e, q \rangle_{L^2(0,T)}, \quad \forall q \in P_{1/2}, \end{aligned} \tag{4.21}$$

such that the variational formulation can be rewritten as equivalent operator equation of the form

$$B_{1/2} y = u + f_{1/2} \in P_{1/2}^*. \tag{4.22}$$

The following Lemma is a consequence of Lemma 3.3, Lemma 1.9 and Theorem 1.1.

Lemma 4.9. *The operator $B_{1/2} : Y_{1/2} \to P_{1/2}^*$ is an isomorphism that satisfies*

$$c_1^{B_{1/2}} \|y\|_{Y_{1/2}} \leq \|B_{1/2} y\|_{P_{1/2}^*} \leq c_2^{B_{1/2}} \|y\|_{Y_{1/2}}$$

with $c_1^{B_{1/2}} = 1/2$ and $c_2^{B_{1/2}} = 1$. In particular, the operator equation (4.22) is uniquely solvable for every given $u \in P_{1/2}^*$.

The optimal control problem associated to this variational formulation is given by

$$\begin{aligned} &\min_{y \in Y_{1/2}, u \in P_{1/2}^*} J_{1/2}(y,u) = \frac{1}{2} \|y + y_e - y_d\|_{L^2(0,T)}^2 + \frac{\varrho}{2} \|u\|_{P_{1/2}^*}^2, \\ &\text{subject to} \qquad B_{1/2} y = u + f_{1/2} \qquad \text{in } P_{1/2}^*. \end{aligned} \tag{4.23}$$

Then the optimal state is given as the sum of the extension y_e and the unique solution of the operator equation

$$y + \varrho S_{1/2} y = \tilde{y}_d \quad \text{in } Y_{1/2}^*, \tag{4.24}$$

where $S_{1/2} = B_{1/2}^\top A_{1/2}^{-1} B_{1/2} : Y_{1/2} \to Y_{1/2}^*$ and $\tilde{y}_d = y_d - y_e + \varrho B_{1/2}^\top A_{1/2}^{-1} f_{1/2}$. The following Lemma is an immediate consequence of Lemma 4.9 and Lemma 2.4.

Lemma 4.10. *The operator $S_{1/2} : Y_{1/2} \to Y_{1/2}^*$ is elliptic with $c_1^{S_{1/2}} = \frac{1}{4}$ and bounded with $c_2^{S_{1/2}} = 1$.*

In the remainder of this section we will prove the state error and control cost estimates, which are summarized in the next theorem.

Theorem 4.11. *For a given $y_d \in L^2(0,T)$ and $y_0 \in \mathbb{R}$ we denote by $y_e(t) = y_0 e^{-\lambda t}$ the homogeneous extension of the initial value, and by $y \in Y_{1/2}$ the associated solution of the operator equation (4.24). Then it holds that*

$$\begin{aligned} \|y + y_e - y_d\|_{L^2(0,T)} &\leq \|y_d\|_{L^2(0,T)} + y_0 \sqrt{\tfrac{1 - e^{-2\lambda T}}{2\lambda}} \\ \|u\|_{Y_{1/2}^*} &\leq \varrho^{-1/2} \left(\|y_d\|_{L^2(0,T)} + y_0 \sqrt{\tfrac{1 - e^{-2\lambda T}}{2\lambda}} \right). \end{aligned}$$

If we further assume $(y_d - y_e) \in H^{1/2}_{0,}(0,T)$ *we obtain*

$$\|y + y_e - y_d\|_{L^2(0,T)} \leq \varrho^{1/2} \|y_d - y_e\|_{Y_{1/2}}$$
$$\|u\|_{Y^*_{1/2}} \leq \|y_d - y_e\|_{Y_{1/2}}.$$

If the target satisfies the stronger condition $y_d \in H^1(0,T)$ *with* $y_d(0) = y_0$*, then*

$$\|y + y_e - y_d\|_{L^2(0,T)} \leq \varrho \|S_{1/2}(y_d - y_e)\|_{L^2(0,T)}$$

Before we actually prove this theorem, let us state some auxiliary results, like we did in the previous sections. In order to use Corollary 2.7 we have to compute the norms $\|y\|_{S_{1/2}}$ and $\|S_{1/2}y\|_{L^2(0,T)}$. First note that we can formally write

$$S_{1/2} = B^\top_{1/2} A^{-1}_{1/2} B_{1/2} = \left(\lambda + \tfrac{\mathrm{d}}{\mathrm{d}t}\right)^\top \left(\lambda - \mathcal{H}_T \tfrac{\mathrm{d}}{\mathrm{d}t}\right)^{-1} \left(\lambda + \tfrac{\mathrm{d}}{\mathrm{d}t}\right),$$

but due to the inversion of $\lambda - \mathcal{H}_T \frac{\mathrm{d}}{\mathrm{d}t}$ involved in this definition, it is not obvious how to find a suitable explicit representation of this operator. Therefore we will only present formulas based on infinite sums, which after truncation can be used in the numerical part to approximate the system matrix with high accuracy. Firstly we can write $y, z \in Y_{1/2}$ as

$$y(t) = \sum_{k=0}^{\infty} y_k^{(s)} \sin(\mu_k \tfrac{t}{T}) = \sum_{k=0}^{\infty} y_k^{(c)} \cos(\mu_k \tfrac{t}{T}),$$
$$z(t) = \sum_{k=0}^{\infty} z_k^{(s)} \sin(\mu_k \tfrac{t}{T}) = \sum_{k=0}^{\infty} z_k^{(c)} \cos(\mu_k \tfrac{t}{T}),$$

where the infinite sums involving sine functions converge with respect to $\|\cdot\|_{H^{1/2}_{0,}}$, and the sums involving involving cosine functions merely converge with respect to $\|\cdot\|_{L^2(0,T)}$. Next, we denote by $c_{k,\ell} = \frac{2}{T}\langle \sin(\mu_\ell \tfrac{t}{T}), \cos(\mu_k \tfrac{t}{T})\rangle_{L^2(0,T)}$, then

$$(B_{1/2}y)(t) = \sum_{k=0}^{\infty} y_k^{(s)} \left(\tfrac{\mu_k}{T} \cos(\mu_k \tfrac{t}{T}) + \lambda \sin(\mu_k \tfrac{t}{T})\right)$$
$$= \sum_{k=0}^{\infty} \left(y_k^{(s)} \tfrac{\mu_k}{T} + \lambda \sum_{\ell=0}^{\infty} c_{k,\ell} y_\ell^{(s)} \right) \cos(\mu_k \tfrac{t}{T})$$

Note that $\left(\lambda - \mathcal{H}_T \frac{\mathrm{d}}{\mathrm{d}t}\right) \cos(\mu_k \frac{t}{T}) = \left(\lambda + \frac{\mu_k}{T}\right) \cos(\mu_k \frac{t}{T})$ holds for all $k \in \mathbb{N}_0$, therefore we conclude

$$(A^{-1}_{1/2} B_{1/2} y)(t) = \sum_{k=0}^{\infty} \frac{\left(y_k^{(s)} \frac{\mu_k}{T} + \lambda \sum\limits_{\ell=0}^{\infty} c_{k,\ell} y_\ell^{(s)} \right)}{\lambda + \frac{\mu_k}{T}} \cos(\mu_k \tfrac{t}{T}).$$

Then the orthogonality of the cosine basis further implies

$$\begin{aligned}
\left\langle S_{1/2}y, z\right\rangle_{(0,T)} &= \left\langle B_{1/2}z, A_{1/2}^{-1}B_{1/2}y\right\rangle_{(0,T)} \\
&= \left\langle \sum_{k=0}^{\infty}\left(z_k^{(s)}\tfrac{\mu_k}{T} + \lambda\sum_{\ell=0}^{\infty}c_{k,\ell}z_\ell^{(s)}\right)\cos(\mu_k\tfrac{t}{T}), \sum_{k=0}^{\infty}\frac{\left(y_k^{(s)}\tfrac{\mu_k}{T} + \lambda\sum\limits_{\ell=0}^{\infty}c_{k,\ell}y_\ell^{(s)}\right)}{\lambda+\tfrac{\mu_k}{T}}\cos(\mu_k\tfrac{t}{T})\right\rangle_{(0,T)} \\
&= \frac{T}{2}\sum_{k=0}^{\infty}\frac{\left(\left(y_k^{(s)}\tfrac{\mu_k}{T} + \lambda\sum\limits_{\ell=0}^{\infty}c_{k,\ell}y_\ell^{(s)}\right)\right)\left(\left(\left(z_k^{(s)}\tfrac{\mu_k}{T} + \lambda\sum\limits_{\ell=0}^{\infty}c_{k,\ell}z_\ell^{(s)}\right)\right)\right)}{\lambda+\tfrac{\mu_k}{T}} \\
&= \frac{T}{2}\sum_{k=0}^{\infty}\frac{\left(\left(y_k^{(s)}\tfrac{\mu_k}{T} + \lambda y_k^{(c)}\right)\right)\left(\left(z_k^{(s)}\tfrac{\mu_k}{T} + \lambda z_k^{(c)}\right)\right)}{\lambda+\tfrac{\mu_k}{T}}. \qquad (4.25)
\end{aligned}$$

Even though this representation is useful for the numerical part, it is neither obvious how to use it to find the sharp constants $c_1^{S_{1/2}}$ and $c_2^{S_{1/2}}$ depending on $\lambda > 0$, nor how to derive a compact representation of $\|y\|_{S_{1/2}}$ or $\|S_{1/2}y\|_{L^2(0,T)}$ like it was done for the primal and the adjoint formulation. We can still use Corollary 2.7, but it remains to characterize what it means for $y \in Y_{1/2}$ that $S_{1/2}y \in L^2(0,T)$. For this purpose we first note that $B_{1/2}\big|_{H^1_{0,}(0,T)} = B_1$, where $B_1 : H^1_{0,}(0,T) \to L^2(0,T)$ denotes the respective operator from the primal formulation. Since B_1 is an isomorphism, we know that the restricted map $B_{1/2}\big|_{H^1_{0,}(0,T)} : H^1_{0,}(0,T) \to L^2(0,T)$ is also isomorphic. Next note that $A_{1/2} = (\lambda - \mathcal{H}_T\frac{\mathrm{d}}{\mathrm{d}t})$ can be represented as

$$(A_{1/2}p)(t) = A_{1/2}\left(\sum_{k=0}^{\infty}p_k\cos(\mu_k\tfrac{t}{T})\right) = \sum_{k=0}^{\infty}p_k\left(\lambda+\tfrac{\mu_k}{T}\right)\cos(\mu_k\tfrac{t}{T}),$$

thus using the inequality (1.10) yields

$$\begin{aligned}
\|A_{1/2}p\|_{L^2(0,T)} &= \frac{T}{2}\sum_{k=0}^{\infty}p_k^2\left(\lambda+\frac{\mu_k}{T}\right)^2 \le T\lambda^2\sum_{k=0}^{\infty}p_k^2 + \frac{1}{T}\sum_{k=0}^{\infty}\mu_k^2p_k^2 \\
&= T\lambda^2\sum_{k=0}^{\infty}\tfrac{\mu_k^2}{\mu_k^2}p_k^2 + \frac{1}{T}\sum_{k=0}^{\infty}\mu_k^2p_k^2 \le 2\left(\frac{T^2\lambda^2}{\mu_0^2}+1\right)\frac{1}{2T}\left(\sum_{k=0}^{\infty}\mu_k^2p_k^2\right) \\
&\le C\|p\|_{H^1_{,0}(0,T)}.
\end{aligned}$$

We conclude that $A_{1/2}\big|_{H^1_{,0}(0,T)} : H^1_{,0}(0,T) \to L^2(0,T)$ is linear and bounded. Now let $w \in L^2(0,T)$ be given, then we can write

$$w(t) = \sum_{k=0}^{\infty}w_k\cos(\mu_k\tfrac{t}{T})$$

and subsequently define

$$p(t) = \sum_{k=0}^{\infty} w_k \left(\lambda + \tfrac{\mu_k}{T}\right)^{-1} \cos(\mu_k \tfrac{t}{T}),$$

then

$$\begin{aligned}\|p\|^2_{H^1_{,0}(0,T)} &= \frac{1}{2T}\sum_{k=0}^{\infty} \frac{\mu_k^2}{(\lambda+\frac{\mu_k}{T})^2} w_k^2 = \frac{T}{2}\sum_{k=0}^{\infty} \frac{\left(\frac{\mu_k}{T}\right)^2}{(\lambda+\frac{\mu_k}{T})^2} w_k^2 \\ &\leq \frac{T}{2}\sum_{k=0}^{\infty} \frac{\left(\lambda+\frac{\mu_k}{T}\right)^2}{(\lambda+\frac{\mu_k}{T})^2} w_k^2 = \frac{T}{2}\sum_{k=0}^{\infty} w_k^2 = \|w\|^2_{L^2(0,T)}.\end{aligned}$$

But this means that the map $A_{1/2}\big|_{H^1_{,0}(0,T)} : H^1_{,0}(0,T) \to L^2(0,T)$ admits a bounded inverse. Thus we have shown that $A_{1/2}^{-1}B_{1/2}\Big|_{H^1_{0,}(0,T)} : H^1_{0,}(0,T) \to H^1_{,0}(0,T)$ is an isomorphism.

Finally, consider $B_{1/2}^\top : H^{1/2}_{,0}(0,T) \to [H^{1/2}_{0,}(0,T)]^*$ and $p \in H^1_{,0}(0,T) \subset H^{1/2}_{,0}(0,T)$, then for all $y \in H^{1/2}_{0,}(0,T)$ it holds that

$$\left\langle B_{1/2}^\top p, y \right\rangle = \langle \dot{y}, p\rangle_{(0,T)} + \lambda\langle y, p\rangle_{L^2(0,T)} = \langle \lambda p - \dot{p}, y\rangle_{L^2(0,T)}.$$

Thus $B_{1/2}^\top\Big|_{H^1_{,0}(0,T)} p = \lambda p - \dot{p} \in L^2(0,T)$, and we know that $B_{1/2}^\top\Big|_{H^1_{,0}(0,T)}$ can be interpreted as a map from $H^1_{,0}(0,T)$ to $L^2(0,T)$. To see that it is invertible, note that $B_{1/2}^\top\Big|_{H^1_{,0}(0,T)} = R_T B_1 R_T$, where R_T denotes the time reversal operator. Since both the time reversal operator and the operator $B_1 : H^1_{0,}(0,T) \to L^2(0,T)$ are isomorphisms, we know that $B_{1/2}^\top\Big|_{H^1_{,0}(0,T)}$ is an isomorphism as well. Therefore we finally conclude that

$$S_{1/2}\big|_{H^1_{0,}(0,T)} = \left((B_{1/2})^\top (A_{1/2})^{-1} B_{1/2}\right)\Big|_{H^1_{0,}(0,T)} : H^1_{0,}(0,T) \to L^2(0,T)$$

is also isomorphic. Therefore $S_{1/2}y \in L^2(0,T)$ if and only if $y \in H^1_{0,}(0,T)$. Now we can finally prove Theorem 4.11.

Proof of Theorem 4.11: The proof is based on Corollary 2.7, and the first two estimates can be shown exactly the same way as the first two estimates in Theorem 4.3. For the next two estimates assume $y_d - y_e \in H^{1/2}_{0,}(0,T)$, which is equivalent to $\tilde{y}_d \in Y_{1/2}$, then Corollary 2.7 implies

$$\|y + y_e - y_d\|_{L^2(0,T)} \leq \varrho^{1/2}\|\tilde{y}_d\|_{S_{1/2}}, \qquad \|u\|_{Y^*_{1/2}} \leq \|\tilde{y}_d\|_{S_{1/2}}. \tag{4.26}$$

Thus the third and fourth estimate claimed in the theorem follow from (4.26) and the norm inequality $\|\tilde{y}_d\|_{S_{1/2}} \leq \|\tilde{y}_d\|_{Y_{1/2}}$. Finally, if $y_d \in H^1(0,T)$ with $y_d(0) = y_0$, then

$\tilde{y}_d = y_d - y_e \in H^1_{0,}(0,T)$, which is equivalent to $S_{1/2}\tilde{y}_d \in L^2(0,T)$, therefore Corollary 2.7 implies

$$\|y + y_e - y_d\|_{L^2(0,T)} \leq \varrho \|S_{1/2}\tilde{y}_d\|_{L^2(0,T)} = \|S_{1/2}(y_d - y_e)\|_{L^2(0,T)}. \qquad \square$$

Remark 4.12. *Note that even though the last estimate*

$$\|y + y_e - y_d\|_{L^2(0,T)} \leq \varrho \|S_{1/2}(y_d - y_e)\|_{L^2(0,T)}$$

is useful to understand the asymptotic behaviour of the state error as $\varrho \to 0$, it is neither easy to numerically evaluate the norm $\|S_{1/2}(y_d - y_e)\|_{L^2(0,T)}$ due to the complicated structure of $S_{1/2}$, nor to find an estimate from above which is easier to compute. Nevertheless those estimates reveal a similar behaviour for the control cost and state error estimates as for the primal formulation. The state error converges to 0 with a convergence rate that increases as the regularity increases, the control costs tend to infinity if the target is not sufficiently regular, and the rate at which they diverge increases as the regularity of the target decreases. But in contrast to the primal formulation, the target only has to satisfy $y_d \in H^1_{0,}(0,T)$ in order to achieve a linear convergence rate, and the control costs already stay bounded if $y_d \in H^{1/2}_{0,}(0,T)$. So by interpolating between the primal and the adjoint formulation, the regularities required to obtain the optimal estimates are also obtained by interpolation of the respective regularities required for the primal and the adjoint formulation.

4.4 Closed-Loop Regularization

As we have seen, the optimal control problem related to the primal formulation is

$$\begin{aligned} \min_{y,u} J(y,u) &= \int_0^T \tfrac{1}{2}(y(t) - y_d(t))^2 + \tfrac{\varrho}{2\lambda}u(t)^2 \,\mathrm{d}t \\ \text{subject to } \dot{y}(t) &= u(t) - \lambda y(t), \quad t \in (0,T] \\ y(0) &= y_0, \end{aligned} \tag{4.27}$$

where we have assumed $y \in H^1(0,T)$ and $u, y_d \in L^2(0,T)$. Following the definitions of Section 2.3, we have

$$\begin{aligned} Q(t) &= 1, & R(t) &= \tfrac{\varrho}{\lambda}, \\ A(t) &= -\lambda, & B(t) &= 1. \end{aligned} \tag{4.28}$$

Now in order to compute a closed-loop control policy, we plug the matrices from Equation (4.28) into the Equation (2.12) and obtain the coupled system

$$\begin{aligned} \dot{P}(t) &= \tfrac{\lambda}{\varrho}P(t)^2 - 1 + 2\lambda P(t), & P(T) &= 0, \\ \dot{\eta}(t) &= \lambda\left(\tfrac{1}{\varrho}P(t) + 1\right)\eta(t) + y_d(t), & \eta(T) &= 0. \end{aligned} \tag{4.29}$$

After solving this system for $P(t)$ and $\eta(t)$, the optimal closed-loop control $u(y,t)$ is given by

$$u(y,t) = -\tfrac{\lambda}{\varrho}\left(P(t)y + \eta(t)\right).$$

Theorem 4.13. *The optimal closed-loop control $u(y,t)$ related to (4.27) has the form*

$$u(y,t) = \lambda\left(1 - \gamma\frac{\left(\frac{\gamma+1}{\gamma-1}\right)e^{2\gamma\lambda(T-t)}-1}{\left(\frac{\gamma+1}{\gamma-1}\right)e^{2\gamma\lambda(T-t)}+1}\right)y + \frac{\lambda}{\varrho}\int_t^T \frac{\left(\frac{\gamma+1}{\gamma-1}\right)+e^{2\gamma\lambda(s-T)}}{\left(\frac{\gamma+1}{\gamma-1}\right)+e^{2\gamma\lambda(t-T)}}e^{\gamma\lambda(t-s)}y_d(s)\,\mathrm{d}s,$$

where $\gamma = \sqrt{1+\frac{1}{\lambda\varrho}} > 1$.

Proof: Since the first equation of Riccati type in (4.29) does not depend on the second differential equation, we begin by solving the Riccati equation first. For that purpose we first find a particular solution $P_p(t)$ to reduce it to a Bernoulli equation, for which there are well known transformations to obtain equivalent linear equations. Since all the coefficients are constant we make the ansatz $P_p(t) = \hat{p} \in \mathbb{R}$, which results in the quadratic equation

$$0 = \tfrac{\lambda}{\varrho}\hat{p}^2 - 1 + 2\lambda\hat{p}.$$

The two solutions to this equation are given by

$$\hat{p} = -\varrho \pm \sqrt{\varrho^2 + \tfrac{\varrho}{\lambda}} = \varrho\left(\pm\gamma - 1\right),$$

with $\gamma = \sqrt{1+\frac{1}{\lambda\varrho}}$. Without loss of generality we choose $\hat{p} = \varrho(\gamma-1)$, since we only need one particular solution. Next we plug the ansatz

$$P(t) = \hat{p} + H(t),$$

into the first differential equation in (4.29), which yields

$$\begin{aligned}
\dot{H}(t) = \dot{P}(t) &= \tfrac{\lambda}{\varrho}\left(\hat{p} + H(t)\right)^2 - 1 + 2\lambda\left(\hat{p} + H(t)\right)\\
&= \tfrac{\lambda}{\varrho}\left(\hat{p}^2 + 2\varrho\hat{p} - \tfrac{\varrho}{\lambda}\right) + \tfrac{\lambda}{\varrho}H(t)^2 + 2\lambda H(t) + \tfrac{2\lambda}{\varrho}\hat{p}H(t)\\
&= \tfrac{\lambda}{\varrho}H(t)^2 + 2\lambda\left(1 + \tfrac{\hat{p}}{\varrho}\right)H(t)\\
&= \tfrac{\lambda}{\varrho}H(t)^2 + 2\gamma\lambda H(t).
\end{aligned}$$

This is a Bernoulli equation, together with the terminal condition $H(T) = -\hat{p}$, which can be transformed to a linear equation using the ansatz

$$H(t) = \tfrac{1}{h(t)},$$

resulting in

$$-\tfrac{1}{h(t)^2}\dot{h}(t) = \tfrac{\lambda}{\varrho}\tfrac{1}{h(t)^2} + 2\gamma\lambda\tfrac{1}{h(t)}.$$

After multiplication of this equation with $h(t)^2$ we obtain

$$\dot{h}(t) + 2\gamma\lambda h(t) = -\tfrac{\lambda}{\varrho}, \qquad h(T) = -\tfrac{1}{\hat{p}}$$

Using separation of variables we can show that the solution of this differential equation is given by

$$h(t) = \left(\tfrac{1}{2\gamma\varrho} - \tfrac{1}{\hat{p}}\right) e^{2\gamma\lambda(T-t)} - \tfrac{1}{2\gamma\varrho}.$$

Therefore we have found that

$$\begin{aligned} P(t) &= \hat{p} + H(t) = \hat{p} + \tfrac{1}{h(t)} \\ &= \varrho\,(\gamma - 1) + \frac{1}{\left(\tfrac{1}{2\gamma\varrho} - \tfrac{1}{\hat{p}}\right) e^{2\gamma\lambda(T-t)} - \tfrac{1}{2\gamma\varrho}} \\ &= \varrho \left(\gamma - 1 + \frac{1}{\left(\tfrac{1}{2\gamma} - \tfrac{1}{\gamma-1}\right) e^{2\gamma\lambda(T-t)} - \tfrac{1}{2\gamma}} \right) \\ &= \varrho \left(\gamma \left(1 + \frac{2}{\left(\tfrac{1+\gamma}{1-\gamma}\right) e^{2\gamma\lambda(T-t)} - 1} \right) - 1 \right). \end{aligned}$$

Next we aim to solve the second differential equation in (4.29) for $\eta(t)$. Let us first define the auxiliary function $f(t) = \lambda(\frac{1}{\varrho}P(t) + 1)$, then the terminal value problem has the simple form

$$\dot{\eta}(t) = f(t)\eta(t) + y_d(t), \quad \eta(T) = 0.$$

Using variation of constants one can show that

$$\eta(t) = -\int_t^T \frac{\eta_h(t)}{\eta_h(s)} y_d(s)\ \mathrm{d}s,$$

is the unique solution of the terminal value problem, where $\eta_h(t)$ is any non-trivial homogeneous solution. Since $y_d(t)$ is not further specified, we cannot perform the integration, and everything that remains to be done is to compute the homogeneous solution $\eta_h(t)$ given as

$$\eta_h(t) = e^{\int f(t)\mathrm{d}t}.$$

In order to perform the integration of $f(t)$ we first note that

$$\begin{aligned} f(t) &= \lambda\left(\tfrac{1}{\varrho}P(t) + 1\right) = \lambda\gamma \left(1 + \frac{2}{\left(\tfrac{1+\gamma}{1-\gamma}\right) e^{2\gamma\lambda(T-t)} - 1} \right) \\ &= \lambda\gamma \left(1 - \frac{2}{\left(\tfrac{\gamma+1}{\gamma-1}\right) e^{2\gamma\lambda(T-t)} + 1} \right) = \lambda\gamma - \frac{2\lambda\gamma e^{2\gamma\lambda(t-T)}}{\left(\tfrac{\gamma+1}{\gamma-1}\right) + e^{2\gamma\lambda(t-T)}} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\lambda\gamma t - \ln\left(\left(\tfrac{\gamma+1}{\gamma-1}\right) + e^{2\gamma\lambda(t-T)}\right)\right). \end{aligned}$$

Therefore we conclude that

$$\eta_h(t) = e^{\left(\lambda\gamma t - \ln\left(\left(\frac{\gamma+1}{\gamma-1}\right)+e^{2\gamma\lambda(t-T)}\right)\right)} = \frac{e^{\lambda\gamma t}}{\left(\frac{\gamma+1}{\gamma-1}\right)+e^{2\gamma\lambda(t-T)}}.$$

Thus the solution $\eta(t)$ to the inhomogeneous problem takes the form

$$\eta(t) = -\int_t^T \frac{\left(\frac{\gamma+1}{\gamma-1}\right)+e^{2\gamma\lambda(s-T)}}{\left(\frac{\gamma+1}{\gamma-1}\right)+e^{2\gamma\lambda(t-T)}} e^{\gamma\lambda(t-s)} y_d(s)\ \mathrm{d}s,$$

and that the optimal control is given by

$$\begin{aligned} u(y,t) &= -\tfrac{\lambda}{\varrho}\left(P(t)y + \eta(t)\right) \\ &= \lambda\left(1 - \gamma\frac{\left(\frac{\gamma+1}{\gamma-1}\right)e^{2\gamma\lambda(T-t)}-1}{\left(\frac{\gamma+1}{\gamma-1}\right)e^{2\gamma\lambda(T-t)}+1}\right)y + \tfrac{\lambda}{\varrho}\int_t^T \frac{\left(\frac{\gamma+1}{\gamma-1}\right)+e^{2\gamma\lambda(s-T)}}{\left(\frac{\gamma+1}{\gamma-1}\right)+e^{2\gamma\lambda(t-T)}} e^{\gamma\lambda(t-s)} y_d(s)\ \mathrm{d}s, \end{aligned}$$

which was exactly the desired claim to prove. □

5 Numerical Analysis

Throughout this chapter we will always work with a uniform discretization of $(0,T)$ into $N \in \mathbb{N}$ non-overlapping intervals $I_k = (t_{k-1}, t_k)$, with $t_k = kh$ and mesh size $h = \frac{T}{N}$. Further we will denote by $S_h^0(0,T)$ the space of piecewise constant and discontinuous functions, and by $S_h^1(0,T)$ the space of piecewise linear and globally continuous functions associated to the uniform mesh. The respective basis functions are defined by

$$\psi_i(t) = \begin{cases} 1, & \text{if } t \in I_i, \\ 0, & \text{else.} \end{cases} \in S_h^0(0,T)$$

for all $i \in \{1, 2, ..., N\}$ and

$$\varphi_j(t) = \begin{cases} \frac{t-t_{j-1}}{h}, & \text{if } t \in (t_j - 1, t_j], \\ \frac{t_{j+1}-t}{h}, & \text{if } t \in (t_j, t_{j+1}), \\ 0, & \text{else.} \end{cases} \in S_h^1(0,T)$$

for all $j \in \{0, 1, ..., N\}$. All numerical experiments were conducted in *Octave 5.1.0*.

5.1 Discretization

In this section we discuss the discretizations of all relevant operators related to the primal, the adjoint and the interpolated formulation.

Primal Formulation

In order to discretize the operators D_1, A_1, B_1 and S_1 from the primal formulation, the discrete ansatz spaces are chosen conformal as $Y_{1,h} = S_h^1(0,T) \cap H_{0,}^1(0,T) \subset Y_1$ for the state space and $P_{1,h} = S_h^0(0,T) \subset P_1$ for the adjoint state space. Then by definition of D_1, the matrix $\mathbf{D}_1 = (\langle D_1 \varphi_j, \varphi_i \rangle_{(0,T)})_{1 \le i,j \le N}$ is given by

$$\mathbf{D}_1 = \left(\tfrac{1}{\lambda} \langle \dot{\varphi}_j, \dot{\varphi}_i \rangle_{L^2(0,T)} + \lambda \langle \varphi_j, \varphi_i \rangle_{L^2(0,T)} \right)_{1 \le i,j \le N} = \tfrac{1}{\lambda} \mathbf{A}^{11}_{0,;0,} + \lambda \mathbf{M}^{11}_{0,;0,},$$

where $\mathbf{A}^{11}_{0,;0,}$, $\mathbf{M}^{11}_{0,;0,}$ are just the matrices one obtains when eliminating the first row and column of the stiffness matrix

$$\mathbf{A}^{11}_h = \frac{1}{h}\begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & \ddots & \ddots & & \\ & & \ddots & \ddots & -1 & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}$$

and the mass matrix

$$\mathbf{M}^{11}_h = \frac{h}{6}\begin{pmatrix} 2 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & 1 & \ddots & \ddots & & \\ & & \ddots & \ddots & 1 & \\ & & & 1 & 4 & 1 \\ & & & & 1 & 2 \end{pmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}$$

respectively, as derived in [27]. We will frequently make use of this notation, where we write for example $\mathbf{C}_{0,;,0}$ to denote the matrix resulting from eliminating the first column and the last row of the matrix $\mathbf{C}$. In this notation, the part before the semicolon is referring to the columns and the part after the semicolon is referring to the rows, and we write the zero before / after the comma if we eliminate the first / last row or column respectively. Next, defining $\mathbf{M}^{00}$ as the mass matrix for the piecewise constant functions, it is well known that $\mathbf{M}^{00} = h \cdot \mathbf{I}$. Therefore the discretized version of A_1 is given by

$$\mathbf{A}_1 = \Big(\lambda \langle \psi_j, \psi_i \rangle_{L^2(0,T)}\Big)_{1\le i,j\le N} = \lambda \mathbf{M}^{00} = \lambda h \mathbf{I} \in \mathbb{R}^{N\times N}. \tag{5.1}$$

In order to approximate B_1 we also introduce the matrices

$$\mathbf{K}^{01} = \Big(\langle \dot{\varphi}_j, \psi_i \rangle_{L^2(0,T)}\Big)_{1\le i,j\le N} = \begin{pmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & \ddots & \ddots \\ & & & -1 \quad 1 \end{pmatrix} \in \mathbb{R}^{N\times(N+1)}$$

and

$$\mathbf{M}^{01} = \Big(\langle \varphi_j, \psi_i \rangle_{L^2(0,T)}\Big)_{1\le i,j\le N} = \frac{h}{2}\begin{pmatrix} 1 & 1 & & \\ & 1 & 1 & \\ & & \ddots & \ddots \\ & & & 1 \quad 1 \end{pmatrix} \in \mathbb{R}^{N\times(N+1)}.$$

If instead of $B_1 : H^1_{0,}(0,T) \to L^2(0,T)$ we would enlarge the domain of the operator to $H^1(0,T)$, the associated matrix approximation would have the form

$$\begin{aligned}\tilde{\mathbf{B}}_1 &= \Big(\langle B_1\varphi_j, \psi_i\rangle_{(0,T)}\Big)_{1\le i\le N,\ 0\le j\le N} = \mathbf{K}^{01} + \lambda\mathbf{M}^{01} \\ &= \begin{pmatrix} (-1+\frac{\lambda h}{2}) & (1+\frac{\lambda h}{2}) & & & \\ & (-1+\frac{\lambda h}{2}) & (1+\frac{\lambda h}{2}) & & \\ & & \ddots & \ddots & \\ & & & (-1+\frac{\lambda h}{2}) & (1+\frac{\lambda h}{2}) \end{pmatrix} \in \mathbb{R}^{N\times(N+1)}.\end{aligned}$$

Therefore the matrix approximation $\mathbf{B}_1 \in \mathbb{R}^{N\times N}$ is given by $\tilde{\mathbf{B}}_1$ without its first column, resulting in a lower triangular matrix. Next, in order to discretize S_1, first note that in Section 4.3 we have found the representation

$$\langle S_1 y, z\rangle_{(0,T)} = \tfrac{1}{\lambda}\langle \dot y, \dot z\rangle + \Big(\langle \dot y, z\rangle_{L^2(0,T)} + \langle y, \dot z\rangle_{L^2(0,T)}\Big) + \lambda\langle y, z\rangle_{L^2(0,T)}.$$

Denoting by $\mathbf{K}^{11}$ the matrix

$$\mathbf{K}^{11} = \Big(\langle \dot\varphi_j, \varphi_i\rangle_{L^2(0,T)}\Big)_{0\le i,j\le N} = \frac{1}{2}\begin{pmatrix} -1 & 1 & & & & \\ -1 & 0 & 1 & & & \\ & -1 & \ddots & \ddots & & \\ & & \ddots & \ddots & 1 & \\ & & & -1 & 0 & 1 \\ & & & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(N+1)\times(N+1)},$$

and by $\mathbf{K}^{11}_{0,;0,}$ the same matrix, but without its first row and its first column, we find that

$$\mathbf{S}_1 = \Big(\langle S_1\varphi_j, \varphi_i\rangle_{(0,T)}\Big)_{1\le i,j\le N} = \tfrac{1}{\lambda}\mathbf{A}^{11}_{0,;0,} + \Big(\mathbf{K}^{11}_{0,;0,} + \big(\mathbf{K}^{11}_{0,;0,}\big)^\top\Big) + \lambda\mathbf{M}^{11}_{0,;0,} \in \mathbb{R}^{N\times N}.$$

In this setting, where we have only looked at an optimal control problem involving an ordinary differential equation as a constraint, it was possible to determine $\mathbf{S}_1$ exactly. But if the constraint is in the form of a partial differential equation in which space and time are appearing as independent variables, this exact assembly is often not realizable. Instead, the matrix $\mathbf{S}_1$ is approximated by

$$\tilde{\mathbf{S}}_1 = \mathbf{B}_1^\top \mathbf{A}_1^{-1}\mathbf{B}_1.$$

Since $\mathbf{A}_1$ is just the identity matrix multiplied by the scalar λh, its inversion is trivial. Due to the simple and sparse structure of $\mathbf{B}_1$, one can verify that

$$\tilde{\mathbf{S}}_1 = \tfrac{1}{\lambda h}\mathbf{B}_1^\top\mathbf{B}_1 = \tfrac{1}{\lambda}\mathbf{A}^{11}_{0,;0,} + \Big(\mathbf{K}^{11}_{0,;0,} + \big(\mathbf{K}^{11}_{0,;0,}\big)^\top\Big) + \lambda\tilde{\mathbf{M}}^{11}_{0,;0,} \in \mathbb{R}^{N\times N},$$

with $\tilde{\mathbf{M}}_{0,;0,}$ being the matrix that results from eliminating the first row and column of

$$\tilde{\mathbf{M}}^{11} = \frac{h}{4}\begin{pmatrix} 1 & 1 & & & & & \\ 1 & 2 & 1 & & & & \\ & 1 & \ddots & \ddots & & & \\ & & \ddots & \ddots & 1 & & \\ & & & 1 & 2 & 1 & \\ & & & & 1 & 1 & \end{pmatrix} \in \mathbb{R}^{N\times N}. \tag{5.2}$$

Adjoint Formulation

Similar to the primal formulation, we will discretize the operators D_0, A_0, B_0 and S_0 involved in the adjoint formulation to obtain matrices $\mathbf{D}_0, \mathbf{A}_0, \mathbf{B}_0, \mathbf{S}_0$ and $\tilde{\mathbf{S}}_0$. For that purpose we choose the conformal ansatz spaces $Y_{0,h} = S_h^0(0,T)$ for the state space and $P_{0,h} = S_h^1(0,T) \cap H_{,0}^1(0,T)$ for the adjoint state space. Then from the definition of the operators it follows that

$$\mathbf{D}_0 = \lambda \mathbf{M}^{00} = \lambda h \mathbf{I},$$

and

$$\mathbf{A}_0 = \tfrac{1}{\lambda}\mathbf{A}^{11}_{,0;,0} + \lambda \mathbf{M}_{,0;,0},$$

where $\mathbf{A}^{11}_{,0;,0}$ and $\mathbf{M}^{11}_{,0;,0}$ now denote the matrices obtained from eliminating the last row and column of $\mathbf{A}^{11}$ and $\mathbf{M}^{11}$ respectively. Further, the matrix $\mathbf{B}_0$ is now obtained by taking $\tilde{\mathbf{B}}_0 = (-\mathbf{K}^{01} + \lambda \mathbf{M}^{01})^\top$ and eliminating the last row. Then, similar to $\tilde{\mathbf{S}}_1$, we define $\tilde{\mathbf{S}}_0 = \mathbf{B}_0^\top \mathbf{A}_0^{-1} \mathbf{B}_0$, but this time $\mathbf{A}_0$ is not a diagonal matrix, but a tridiagonal matrix. Therefore it is not obvious how to find a simpler representation of this matrix. In contrast, the matrix $\mathbf{S}_0$ admits a very simple representation as a rank one perturbation of a diagonal matrix. Note that using the representation from Section 4.2, Equation (4.17), we know that

$$\langle S_0 y, v\rangle_{L^2(0,T)} = \lambda \left(\langle y, v\rangle_{L^2(0,T)} + \langle y, g\rangle_{L^2(0,T)} \langle g, v\rangle_{L^2(0,T)} \right).$$

Now denoting by $\underline{g} = (g_1, ..., g_N)^\top$ the vector with elements

$$g_i = \frac{1}{\sqrt{h}} \langle g, \psi_i\rangle_{L^2(0,T)} = \sqrt{\frac{\lambda e^{\lambda T}}{h \cosh(\lambda T)}} \int_{t_{i-1}}^{t_i} e^{-\lambda t} \, \mathrm{d}t = \sqrt{\frac{e^{\lambda T}}{\lambda h \cosh(\lambda T)}} \left(e^{-\lambda t_{i-1}} - e^{-\lambda t_i} \right),$$

the matrix $\mathbf{S}_0$ takes the form

$$\mathbf{S}_0 = \lambda \left(\langle \psi_j, \psi_i\rangle_{L^2(0,T)} + \langle \psi_j, g\rangle_{L^2(0,T)} \langle g, \psi_i\rangle_{L^2(0,T)} \right)_{1\le i,j\le N} = \lambda h \left(\mathbf{I} + \underline{g}\underline{g}^\top \right).$$

Interpolated Formulation

Finally, we discuss the operator discretizations related to the interpolated formulation. In order to compute $\mathbf{S}_{1/2}$, we use an approach based on Fourier series, and write the basis functions $\varphi_j(t)$ in the form

$$\varphi_j(t) = \sum_{k=0}^{\infty} \alpha_k^{(j)} \sin\left(\mu_k \tfrac{t}{T}\right) = \sum_{k=0}^{\infty} \beta_k^{(j)} \cos\left(\mu_k \tfrac{t}{T}\right).$$

Then we know from Equation (4.25), that the system matrix admits the following representation

$$\mathbf{S}_{1/2} = \left(\tfrac{T}{2} \sum_{k=0}^{\infty} \frac{\left(\lambda T \beta_k^{(i)} + \mu_k \alpha_k^{(i)}\right)\left(\lambda T \beta_k^{(j)} + \mu_k \alpha_k^{(j)}\right)}{\lambda T + \mu_k} \right)_{1 \le i,j \le N}$$

as an infinite sum. To realize this matrix numerically, we will truncate this sum at a sufficiently large index, so it only remains to compute those coefficients. For $\alpha_k^{(j)}$ and $j \in \{1, ..., N-1\}$ we get

$$\begin{aligned}
\alpha_k^{(j)} &= \frac{2}{T} \int_0^T \varphi_j(t) \sin\left(\mu_k \tfrac{t}{T}\right) \, \mathrm{d}t \\
&= \frac{2}{T} \left(\int_{t_{j-1}}^{t_j} \left(\tfrac{t - t_{j-1}}{h}\right) \sin\left(\mu_k \tfrac{t}{T}\right) \, \mathrm{d}t + \int_{t_j}^{t_{j+1}} \left(\tfrac{t_{j+1} - t}{h}\right) \sin\left(\mu_k \tfrac{t}{T}\right) \, \mathrm{d}t \right) \\
&= \frac{2}{T} \left(\left(\frac{t - t_{j-1}}{h}\right) \frac{\left(-\cos\left(\mu_k \frac{t}{T}\right)\right)}{\frac{\mu_k}{T}} \Bigg|_{t_{j-1}}^{t_j} - \int_{t_{j-1}}^{t_j} \frac{1}{h} \frac{\left(-\cos\left(\mu_k \frac{t}{T}\right)\right)}{\frac{\mu_k}{T}} \, \mathrm{d}t \right. \\
&\quad \left. + \left(\frac{t_{j+1} - t_j}{h}\right) \frac{\left(-\cos\left(\mu_k \frac{t}{T}\right)\right)}{\frac{\mu_k}{T}} \Bigg|_{t_j}^{t_{j+1}} - \int_{t_j}^{t_{j+1}} \left(-\frac{1}{h}\right) \frac{\left(-\cos\left(\mu_k \frac{t}{T}\right)\right)}{\frac{\mu_k}{T}} \, \mathrm{d}t \right) \\
&= \frac{2}{\mu_k} \left(-\cos\left(\mu_k \tfrac{t_j}{T}\right) + \frac{1}{h} \int_{t_{j-1}}^{t_j} \cos\left(\mu_k \tfrac{t}{T}\right) \, \mathrm{d}t + \cos\left(\mu_k \tfrac{t_j}{T}\right) - \frac{1}{h} \int_{t_j}^{t_{j+1}} \cos\left(\mu_k \tfrac{t}{T}\right) \, \mathrm{d}t \right) \\
&= \frac{2T}{h\mu_k^2} \left(\sin\left(\mu_k \tfrac{t}{T}\right)\Big|_{t_{j-1}}^{t_j} - \sin\left(\mu_k \tfrac{t}{T}\right)\Big|_{t_j}^{t_{j+1}} \right) \\
&= \frac{2N}{\mu_k^2} \left(2\sin\left(\mu_k \tfrac{t_j}{T}\right) - \sin\left(\mu_k \tfrac{t_{j-1}}{T}\right) - \sin\left(\mu_k \tfrac{t_{j+1}}{T}\right) \right).
\end{aligned}$$

If $j = 0$ or $j = N$, the computation of the respective value $\alpha_k^{(j)}$ only involves one integral, instead of the two we solved earlier. Taking this changed behaviour at the boundary

into account we arrive at

$$\alpha_k^{(j)} = \begin{cases} \frac{2N}{\mu_k^2}\left(\sin\left(\mu_k\frac{t_0}{T}\right) - \sin\left(\mu_k\frac{t_1}{T}\right) + \frac{1}{N}\right), & \text{if } j = 0, \\ \frac{2N}{\mu_k^2}\left(2\sin\left(\mu_k\frac{t_j}{T}\right) - \sin\left(\mu_k\frac{t_{j-1}}{T}\right) - \sin\left(\mu_k\frac{t_{j+1}}{T}\right)\right), & \text{if } j \in \{1, ..., N-1\}, \\ \frac{2N}{\mu_k^2}\left(\sin\left(\mu_k\frac{t_N}{T}\right) - \sin\left(\mu_k\frac{t_{N-1}}{T}\right)\right), & \text{if } j = N. \end{cases}$$

Similarly, to compute $\beta_k^{(j)}$, a similar integral has to be solved, where $\sin\left(\mu_k\frac{t}{T}\right)$ is replaced by $\cos\left(\mu_k\frac{t}{T}\right)$. We skip the computation, and only state the resulting formula

$$\beta_k^{(j)} = \begin{cases} \frac{2N}{\mu_k^2}\left(\cos\left(\mu_k\frac{t_0}{T}\right) - \cos\left(\mu_k\frac{t_1}{T}\right)\right), & \text{if } j = 0, \\ \frac{2N}{\mu_k^2}\left(2\cos\left(\mu_k\frac{t_j}{T}\right) - \cos\left(\mu_k\frac{t_{j-1}}{T}\right) - \cos\left(\mu_k\frac{t_{j+1}}{T}\right)\right), & \text{if } j \in \{1, ..., N-1\}, \\ \frac{2N}{\mu_k^2}\left(\cos\left(\mu_k\frac{t_N}{T}\right) - \cos\left(\mu_k\frac{t_{N-1}}{T}\right) + \frac{(-1)^k}{N}\right), & \text{if } j = N. \end{cases}$$

Next, we discuss how to discretize $A_{1/2}, B_{1/2}, D_{1/2}$. Naively, we could choose as discrete spaces $Y_{1/2,h} = S_h^1(0,T) \cap H_{0,}^{1/2}(0,T)$ and $P_{1/2,h} = S_h^1(0,T) \cap H_{,0}^{1/2}(0,T)$. If we then define

$$\mathbf{K}^{\mathcal{H}_T} = \left(\langle\dot{\varphi}_j, \mathcal{H}_T\varphi_j\rangle_{L^2(0,T)}\right)_{0\le i,j\le N}, \qquad \mathbf{K}^{\mathcal{H}_T^{-1}} = \left(\langle\dot{\varphi}_j, \mathcal{H}_T^{-1}\varphi_i\rangle_{(0,T)}\right)_{0\le i,j\le N},$$

the Equations (4.19), (4.20) and (4.21) from Section 4.3 result in the matrix approximations

$$\begin{aligned} \mathbf{A}_{1/2} &= -\mathbf{K}_{,0;,0}^{\mathcal{H}_T^{-1}} + \lambda\mathbf{M}_{,0;,0}^{11}, \\ \mathbf{D}_{1/2} &= \mathbf{K}_{0,;0,}^{\mathcal{H}_T} + \lambda\mathbf{M}_{0,;0,}^{11}, \\ \mathbf{B}_{1/2} &= \mathbf{K}_{0,;,0}^{11} + \lambda\mathbf{M}_{0,;,0}^{11}. \end{aligned}$$

But it is important to note that as discussed in [31, §3.3], this naive approach for approximating $B_{1/2}$ yields an unstable numerical scheme, regardless of the specific choice of $\lambda > 0$ and the mesh size $h > 0$. Thus, we do not expect $\tilde{\mathbf{S}}_{1/2} = \mathbf{B}_{1/2}^\top\mathbf{A}_{1/2}^{-1}\mathbf{B}_{1/2}$ to remain elliptic as $h \to 0$. As a remedy we can enrich the approximation space $P_{1/2,h}$ by replacing it with $P_{1/2,h/2} = S_{h/2}^1(0,T) \cap H_{,0}^{1/2}(0,T)$. We denote the related matrices by $\mathbf{A}_{1/2}^{h/2} \in \mathbb{R}^{2N\times 2N}$, $\mathbf{B}_{1/2}^{h/2} \in \mathbb{R}^{2N\times N}$ and define

$$\tilde{\mathbf{S}}_{1/2}^{h/2} = \left(\mathbf{B}_{1/2}^{h/2}\right)^\top\left(\mathbf{A}_{1/2}^{h/2}\right)^{-1}\mathbf{B}_{1/2}^{h/2}.$$

In order to assemble $\mathbf{B}_{1/2}^{h/2} \in \mathbb{R}^{2N\times N}$, we can just assemble $\mathbf{B}_{1/2} \in \mathbb{R}^{2N\times 2N}$ similar to the naive approach, but on the finer mesh, and multiply $\mathbf{B}_{1/2}$ from the right-hand side with the matrix representation of the interpolation

$$I_h^{h/2} : S_h^1(0,T) \cap H_{0,}^{1/2}(0,T) \to S_{h/2}^1(0,T) \cap H_{0,}^{1/2}(0,T),$$

$$(I_h^{h/2}y_h)(\tfrac{kh}{2}) = \begin{cases} y_h(\frac{kh}{2}), & \text{if } k \text{ is even}, \\ \frac{1}{2}\left(y_h(\frac{(k+1)h}{2}) + y_h(\frac{(k-1)h}{2})\right), & \text{if } k \text{ is odd}. \end{cases},$$

This matrix representation is given by

$$\mathbf{I}_h^{h/2} = \begin{pmatrix} 0.5 & & & & & & \\ 1 & & & & & & \\ 0.5 & 0.5 & & & & & \\ & 1 & & & & & \\ & 0.5 & 0.5 & & & & \\ & & \ddots & \ddots & & & \\ & & & 0.5 & 0.5 & \\ & & & & 1 \end{pmatrix} \in \mathbb{R}^{2N\times N}.$$

Another approach ensuring stability is based on the modified Hilbert Transformation, which acts as an isomorphism $\mathcal{H}_T : Y_{1/2} \to P_{1/2}$ in the continuous setting. Thus it makes sense to consider the discrete ansatz spaces

$$Y_{1/2,h} = \operatorname{span}\{\varphi_1, ..., \varphi_N\} \subset Y_{1/2}, \qquad P_{1/2,h}^{\mathcal{H}_T} = \operatorname{span}\{\mathcal{H}_T\varphi_1, ..., \mathcal{H}_T\varphi_N\} \subset P_{1/2}.$$

Using elementary properties of the modified Hilbert-transformation, it further follows that

$$\begin{aligned} \mathbf{A}_{1/2}^{\mathcal{H}_T} &= \Big(\big\langle A_{1/2}(\mathcal{H}_T\varphi_j), (\mathcal{H}_T\varphi_i)\big\rangle_{(0,T)}\Big)_{1\le i,j\le N} \\ &= \Big(-\big\langle \tfrac{\mathrm{d}}{\mathrm{d}t}(\mathcal{H}_T\varphi_j), \mathcal{H}_T^{-1}(\mathcal{H}_T\varphi_i)\big\rangle_{(0,T)} + \langle \mathcal{H}_T\varphi_j, \mathcal{H}_T\varphi_i\rangle_{L^2(0,T)}\Big)_{1\le i,j\le N} \\ &= \Big(\langle \dot\varphi_j, \mathcal{H}_T\varphi_i\rangle_{(0,T)} + \langle \varphi_j, \varphi_i\rangle_{L^2(0,T)}\Big)_{1\le i,j\le N} \\ &= \mathbf{K}_{0,;0,}^{\mathcal{H}_T} + \lambda\mathbf{M}_{0,;0,}^{11} = \mathbf{D}_{1/2}. \end{aligned}$$

Next, let us denote by $\mathbf{M}^{\mathcal{H}_T} \in \mathbb{R}^{N\times N}$ the matrix

$$\mathbf{M}^{\mathcal{H}_T} = \Big(\langle \varphi_j, \mathcal{H}_T\varphi_i\rangle_{L^2(0,T)}\Big)_{0\le i,j\le N},$$

then $\mathbf{B}_{1/2}^{\mathcal{H}_T}$ takes the form

$$\mathbf{B}_{1/2}^{\mathcal{H}_T} = \Big(\langle \dot\varphi_j, \mathcal{H}_T\varphi_i\rangle_{(0,T)} + \langle \varphi_j, \mathcal{H}_T\varphi_i\rangle_{L^2(0,T)}\Big)_{1\le i,j\le N} = \mathbf{K}_{0,;0,}^{\mathcal{H}_T} + \lambda\mathbf{M}_{0,;0,}^{\mathcal{H}_T}.$$

Finally, we denote by $\tilde{\mathbf{S}}_{1/2}^{\mathcal{H}_T}$ the matrix

$$\tilde{\mathbf{S}}_{1/2}^{\mathcal{H}_T} = \left(\mathbf{B}_{1/2}^{\mathcal{H}_T}\right)^\top \left(\mathbf{A}_{1/2}^{\mathcal{H}_T}\right)^{-1} \mathbf{B}_{1/2}^{\mathcal{H}_T} = \left(\mathbf{B}_{1/2}^{\mathcal{H}_T}\right)^\top \mathbf{D}_{1/2}^{-1}\mathbf{B}_{1/2}^{\mathcal{H}_T}$$

5.2 Matrix Properties

The numerical realization of the matrices $\mathbf{S}$ and $\tilde{\mathbf{S}}$ can be challenging, especially in the space-time setting, regardless of the specific formulation. Thus one can also replace the $\mathbf{S}$ matrix with the spectrally equivalent $\mathbf{D}$ matrix, as proposed in [23]. For the model problem we are considering, all matrices can be realized efficiently, therefore the aim of this section is to actually analyze the spectral equivalence between the $\mathbf{S}$ matrices and the $\mathbf{D}$ matrices, which is usually not possible in the space-time setting.

Primal and Adjoint Formulation

We will analyze the following four generalized eigenvalue problems:

$$\mathbf{S}_1 \underline{y} = \mu \mathbf{D}_1 \underline{y}, \tag{5.3}$$

$$\mathbf{S}_0 \underline{y} = \mu \mathbf{D}_0 \underline{y}, \tag{5.4}$$

$$\tilde{\mathbf{S}}_1 \underline{y} = \mu \mathbf{D}_1 \underline{y}, \tag{5.5}$$

$$\tilde{\mathbf{S}}_0 \underline{y} = \mu \mathbf{D}_0 \underline{y}. \tag{5.6}$$

Before we analyze the first eigenvalue problem, we state an useful auxiliary result.

Lemma 5.1. *Let $\mathbf{S}, \mathbf{D} \in \mathbb{R}^{N \times N}$ be symmetric matrices, and assume $\mathbf{D}$ is invertible. Further, assume that there exists some $\underline{w} \in \mathbb{R}^N$ such that $\mathbf{S} = \mathbf{D} + \underline{w}\underline{w}^\top$. Then there are N linearly independent solutions to the generalized eigenvalue problem*

$$\mathbf{S}\underline{y} = \mu \mathbf{D} \underline{y}, \tag{5.7}$$

with $(N-1)$ of those eigenvectors being orthogonal to $\underline{w}$, and corresponding to the eigenvalue $\mu = 1$. The remaining eigenvector can be chosen as $\underline{y} = \mathbf{D}^{-1}\underline{w}$ and corresponds to the eigenvalue

$$\mu = 1 + \left\langle \mathbf{D}^{-1}\underline{w}, \underline{w} \right\rangle_2.$$

Proof: Using $\mathbf{S} = \mathbf{D} + \underline{w}\underline{w}^\top$, Equation (5.7) is equivalent to

$$\begin{aligned} & (\mathbf{D} + \underline{w}\underline{w}^\top)\underline{y} = \mu \mathbf{D}\underline{y} \\ \Leftrightarrow \qquad & (\mu - 1)\mathbf{D}\underline{y} = \left\langle \underline{y}, \underline{w} \right\rangle_2 \underline{w}. \end{aligned}$$

Therefore, all eigenvectors orthogonal to $\underline{w}$ correspond to $\mu = 1$, and the geometric multiplicity is $(N-1)$. If $\underline{y}$ is not orthogonal to $\underline{w}$, $\mathbf{D}\underline{y}$ and $\underline{w}$ have to be collinear. Without loss of generality, $\underline{y} = \mathbf{D}^{-1}\underline{w}$. Then it immediately follows that $\mu = 1 + \left\langle \mathbf{D}^{-1}\underline{w}, \underline{w} \right\rangle_2$. □

In Section 4.1 it was shown that the operator $S_1 : Y_1 \to Y_1^*$ is elliptic and bounded with $c^{S_1} = 1$ and $c_2^{S_1} = 1 + \tanh(\lambda T)$. Since $\mathbf{S}_1$ was the matrix resulting from a conformal discretization, we know that the eigenvalues corresponding to the generalized eigenvalue problem (5.3) are also bounded by 1 and $1 + \tanh(\lambda T)$. To compute the exact eigenvalues note that

$$\begin{aligned} \mathbf{S}_1 - \mathbf{D}_1 &= \left(\tfrac{1}{\lambda} \mathbf{A}^{11}_{0,;0,} + \left(\mathbf{K}^{11}_{0,;0,} + \left(\mathbf{K}^{11}_{0,;0,} \right)^\top \right) + \lambda \mathbf{M}^{11}_{0,;0,} \right) - \left(\mathbf{A}^{11}_{0,;0,} + \lambda \mathbf{M}^{11}_{0,;0,} \right) \\ &= \left(\mathbf{K}^{11}_{0,;0,} + \left(\mathbf{K}^{11}_{0,;0,} \right)^\top \right) = \underline{e}_N \underline{e}_N^\top, \end{aligned}$$

where $\underline{e}_N = (0, ..., 0, 1)^\top \in \mathbb{R}^N$. Therefore Lemma 5.1 implies that the minimal eigenvalue corresponding to the eigenvalue problem (5.3) is still $\mu_{\min} = 1$, and the maximal eigenvalue is given by $\mu_{\max} = 1 + \left\langle \mathbf{D}_1^{-1}\underline{e}_N, \underline{e}_N \right\rangle_2$. Denoting by $\hat{\mathbf{D}}_1$ the matrix $\mathbf{D}_1$, without its last column and its last row, it further follows from $\det(\mathbf{D}_1)\mathbf{D}_1^{-1} = \operatorname{adj}(\mathbf{D}_1)$ that

$$\mu_{\max} = 1 + \left\langle \mathbf{D}_1^{-1}\underline{e}_N, \underline{e}_N \right\rangle_2 = 1 + \frac{\det(\hat{\mathbf{D}}_1)}{\det(\mathbf{D}_1)}.$$

Recall that

$$\mathbf{D}_1 = \tfrac{1}{\lambda}\mathbf{A}^{11}_{0,;0,} + \lambda\mathbf{M}^{11}_{0,;0,}$$
$$= \begin{pmatrix} (\frac{2}{\lambda h}+\frac{2\lambda h}{3}) & (-\frac{1}{\lambda h}+\frac{\lambda h}{6}) & & & & & \\ (-\frac{1}{\lambda h}+\frac{\lambda h}{6}) & (\frac{2}{\lambda h}+\frac{2\lambda h}{3}) & (-\frac{1}{\lambda h}+\frac{\lambda h}{6}) & & & & \\ & (-\frac{1}{\lambda h}+\frac{\lambda h}{6}) & \ddots & \ddots & & & \\ & & \ddots & \ddots & (-\frac{1}{\lambda h}+\frac{\lambda h}{6}) & & \\ & & & (-\frac{1}{\lambda h}+\frac{\lambda h}{6}) & (\frac{2}{\lambda h}+\frac{2\lambda h}{3}) & (-\frac{1}{\lambda h}+\frac{\lambda h}{6}) & \\ & & & & (-\frac{1}{\lambda h}+\frac{\lambda h}{6}) & (\frac{1}{\lambda h}+\frac{\lambda h}{3}) \end{pmatrix}$$

and let us define $a = (\frac{2}{\lambda h} + \frac{2\lambda h}{3})$, $b = (-\frac{1}{\lambda h} + \frac{\lambda h}{6})$ and $c = (\frac{1}{\lambda h} + \frac{\lambda h}{3})$. Further, we introduce d_N as

$$d_N = \det \underbrace{\begin{pmatrix} a & b & & & \\ b & a & b & & \\ & b & \ddots & \ddots & \\ & & \ddots & & b \\ & & & b & a \end{pmatrix}}_{\in \mathbb{R}^{N\times N}}.$$

Then applying the Laplace expansion rule for determinants once for the last row, and then for the last column of $\mathbf{D}_1$ implies

$$\det(\mathbf{D}_1) = cd_{N-1} - b^2 d_{N-2},$$

and then by definition of d_N we also have $\det(\hat{\mathbf{D}}_1) = d_{N-1}$, therefore

$$\mu_{\max} = 1 + \frac{\det(\hat{\mathbf{D}}_1)}{\det(\mathbf{D}_1)} = 1 + \frac{d_{N-1}}{cd_{N-1} - b^2 d_{N-2}} = 1 + \frac{1}{c - b^2 \frac{d_{N-2}}{d_{N-1}}}. \tag{5.8}$$

It remains to find a suitable representation of d_N. For that purpose, note that d_N satisfies the recurrence relation $d_N = ad_{N-1} - b^2 d_{N-2}$, together with the initial values $d_0 = 1$ and $d_1 = a$. Denoting by

$$\eta_{1/2} = \frac{a}{2} \pm \sqrt{\frac{a^2}{4} - b^2}$$

the solutions of $x^2 = ax - b^2$, one can show that d_N takes the form

$$d_N = \frac{\eta_1^{N+1} - \eta_2^{N+1}}{\eta_1 - \eta_2}.$$

We conclude that the maximal eigenvalue is given by

$$\mu_{\max} = 1 + \frac{1}{c - b^2 \left(\frac{\eta_1^{N-1} - \eta_2^{N-1}}{\eta_1^N - \eta_2^N}\right)}. \tag{5.9}$$

For the analysis of the eigenvalue problem (5.4) we can also use Lemma 5.1, since we have already shown that $\mathbf{S}_0 = \lambda h \left(\mathbf{I} + \underline{g}\underline{g}^\top\right)$ and $\mathbf{D}_0 = \lambda h \mathbf{I}$, hence

$$\mathbf{S}_0 = \mathbf{D}_0 + \left(\sqrt{\lambda h}\underline{g}\right)\left(\sqrt{\lambda h}\underline{g}\right)^\top.$$

Thus, with $\underline{w} = \sqrt{\lambda h}\underline{g}$, Lemma 5.1 implies that $\mu_{\min} = 1$ and

$$\mu_{\max} = 1 + \langle \mathbf{D}_0^{-1}\underline{w}, \underline{w}\rangle = 1 + \|\underline{g}\|_2^2$$

To simplify this formula note that

$$\begin{aligned}
\|\underline{g}\|_2^2 &= \frac{e^{\lambda T}}{\lambda h \cosh(\lambda T)} \sum_{i=1}^{N} (e^{-\lambda t_{i-1}} - e^{-\lambda t_i})^2 = \frac{e^{\lambda T}}{\lambda h \cosh(\lambda T)} \sum_{i=0}^{N-1} (e^{-\lambda t_i} - e^{-\lambda(t_i+h)})^2 \\
&= \frac{e^{\lambda T}(1-e^{-\lambda h})^2}{\lambda h \cosh(\lambda T)} \sum_{i=0}^{N-1} e^{-2\lambda t_i} = \frac{e^{\lambda T}(1-e^{-\lambda h})^2}{\lambda h \cosh(\lambda T)} \sum_{i=0}^{N-1} \left(e^{-2\lambda h}\right)^i \\
&= \frac{e^{\lambda T}(1-e^{-\lambda h})^2}{\lambda h \cosh(\lambda T)} \frac{1-e^{-2\lambda N h}}{1-e^{-2\lambda h}} = \frac{e^{\lambda T}(1-e^{-\lambda h})^2(1-e^{-2\lambda T})}{\lambda h \cosh(\lambda T)(1-e^{-\lambda h})(1+e^{-\lambda h})} \\
&= \frac{(e^{\lambda T} - e^{-\lambda T})}{\lambda T \cosh(\lambda T)} \frac{(1-e^{-\lambda h})}{(1+e^{-\lambda h})} = \tanh(\lambda T) \frac{\tanh(\frac{\lambda h}{2})}{\frac{\lambda h}{2}},
\end{aligned}$$

therefore

$$\mu_{\max} = 1 + \tfrac{2}{\lambda h}\tanh(\tfrac{\lambda h}{2})\tanh(\lambda T). \tag{5.10}$$

Note that $\frac{2}{\lambda h}\tanh(\frac{\lambda h}{2}) \to 1$ as $h \to 0$, therefore we know that the maximal eigenvalue $\mu_{\max}$ actually converges towards $1 + \tanh(\lambda T)$, as we would expect.

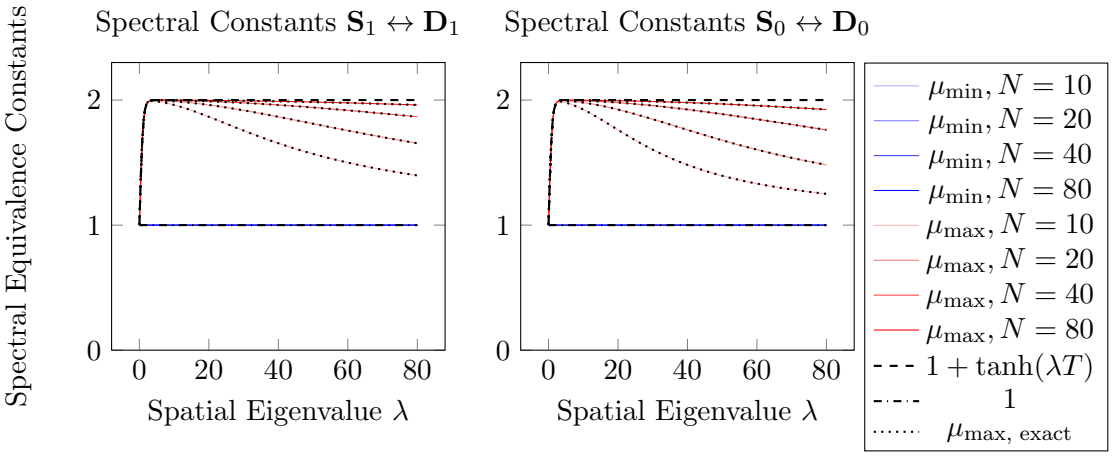

Figure 5.1: Minimal and maximal eigenvalues corresponding to the generalized eigenvalue problems (5.3) and (5.4) for $T = 1.5$.

Figure 5.1 shows the minimal and maximal eigenvalues $\mu_{\min}$ and $\mu_{\max}$ associated to the generalized eigenvalue problems (5.4) and (5.4) for different mesh sizes. Note that

the theoretical predictions of the eigenvalues and the numerical results match perfectly. Further, notice that for a fixed $\lambda > 0$, the maximal eigenvalues seem to converge towards $1 + \tanh(\lambda T)$ as $h \to 0$, which is expected for the eigenvalue problem related to primal formulation, and is proven for the eigenvalue problem related to the adjoint formulation. In contrast, for a fixed $h > 0$, the maximal eigenvalues seem to converge to 1 as $\lambda \to 0$, and as $\lambda \to \infty$. This behaviour is to be expected for the primal eigenvalue problem, since for $\lambda \to 0$, both $\mathbf{S}_1$ and $\mathbf{D}_1$ are dominated by $\mathbf{A}^{11}_{0,;0,}$, and for $\lambda \to \infty$ both matrices are dominated by $\mathbf{M}^{11}_{0,;0,}$. For the adjoint eigenvalue problem, this behaviour is an immediate consequence of Equation (5.10).

Next we want to analyze the eigenvalue problems (5.5) and (5.6). Recall that $\tilde{\mathbf{S}}_1$ is given by

$$\tilde{\mathbf{S}}_1 = \tfrac{1}{\lambda h}\mathbf{B}_1^\top \mathbf{B}_1 = \tfrac{1}{\lambda}\mathbf{A}^{11}_{0,;0,} + \left(\mathbf{K}^{11}_{0,;0,} + \left(\mathbf{K}^{11}_{0,;0,}\right)^\top\right) + \lambda \tilde{\mathbf{M}}^{11}_{0,;0,},$$

where $\tilde{\mathbf{M}}^{11}$ denotes the perturbed mass matrix defined in Equation (5.2). To interpret this matrix, let $Q_h^0 : L^2(0,T) \to S_h^0(0,T)$ denote the L^2-Projection onto the piecewise constant functions. Then it turns out that

$$\tilde{\mathbf{M}}^{11} = \left(\left\langle Q_h^0 \varphi_j, Q_h^0 \varphi_i \right\rangle_{L^2(0,T)}\right)_{0 \le i,j \le N},$$

meaning that this is the mass matrix we would obtain when working with the projection of the basis of $S_h^1(0,T)$ onto $S_h^0(0,T)$ instead of directly working with the basis of $S_h^1(0,T)$. Even though we know this compact representation of $\tilde{\mathbf{S}}_1$, it is not obvious how to analytically derive the eigenvalues for the eigenvalue problem (5.5), and the same holds for the eigenvalue problem (5.6), since we do not even have such a simple representation of $\tilde{\mathbf{S}}_0$. Still, we can at least show that the spectra related to the eigenvalue problems (5.5) and (5.6) are identical. For that purpose we denote by

$$\mathbf{R} = \begin{pmatrix} & & & 1 \\ & & 1 & \\ & \iddots & & \\ 1 & & & \end{pmatrix}$$

the discrete version of the time flipping operator, which satisfies $\mathbf{R}^\top = \mathbf{R}^{-1} = \mathbf{R}$. Next we note that both $\mathbf{B}_1$ and $\mathbf{B}_0$ are persymmetric, meaning that $\mathbf{B}_1 = \mathbf{R}\mathbf{B}_1^\top\mathbf{R}$ and $\mathbf{B}_0 = \mathbf{R}\mathbf{B}_0^\top\mathbf{R}$. Further, note that by definition of $\mathbf{B}_1$ and $\mathbf{B}_0$ we have $\mathbf{B}_1 = \mathbf{B}_0$. For the matrices $\mathbf{A}_1, \mathbf{A}_0, \mathbf{D}_1$ and $\mathbf{D}_0$ we note that $\mathbf{A}_1 = \mathbf{D}_0 = \mathbf{R}\mathbf{A}_1\mathbf{R}$, as well as $\mathbf{A}_0 = \mathbf{R}\mathbf{D}_1\mathbf{R}$.

Thus we conclude that

$$\begin{aligned}
& \mathbf{S}_0\underline{y} = \mu\mathbf{D}_0\underline{y} \\
\Leftrightarrow \quad & \mathbf{B}_0^\top\mathbf{A}_0^{-1}\mathbf{B}_0\underline{y} = \mu\mathbf{D}_0\underline{y} \\
\Leftrightarrow \quad & \mathbf{B}_1^\top\mathbf{R}\mathbf{D}_0^{-1}\mathbf{R}\mathbf{B}_1\underline{y} = \mu\mathbf{A}_1\underline{y} \\
\Leftrightarrow \quad & \mathbf{R}\mathbf{B}_1\mathbf{D}_0^{-1}\mathbf{B}_1^\top\mathbf{R}\underline{y} = \mu\mathbf{R}\mathbf{A}_1\mathbf{R}\underline{y} \\
\Leftrightarrow \quad & \mathbf{B}_1\mathbf{D}_0^{-1}\mathbf{B}_1^\top\underline{z} = \mu\mathbf{A}_1\underline{z} && (\underline{z} = \mathbf{R}\underline{y}) \\
\Leftrightarrow \quad & \mathbf{D}_0^{-1}\underline{v} = \mu\mathbf{B}_1^{-1}\mathbf{A}_1\mathbf{B}_1^{-\top}\underline{v} && (\underline{v} = \mathbf{B}_1^\top\underline{z}) \\
\Leftrightarrow \quad & \mathbf{D}_0^{-1}\underline{v} = \mu(\mathbf{B}_1^\top\mathbf{A}_1^{-1}\mathbf{B}_1)^{-1}\underline{v} \\
\Leftrightarrow \quad & \mathbf{B}_1^\top\mathbf{A}_1^{-1}\mathbf{B}_1\underline{w} = \mu\mathbf{D}_0\underline{w}, && (\underline{w} = \mathbf{D}_0^{-1}\underline{v}) \\
\Leftrightarrow \quad & \mathbf{S}_1\underline{w} = \mu\mathbf{D}_1\underline{w},
\end{aligned}$$

so the eigenvectors follow the relation $\underline{w} = \mathbf{D}_0^{-1}\mathbf{B}_1^\top\mathbf{R}\underline{y}$, and the eigenvalues agree.

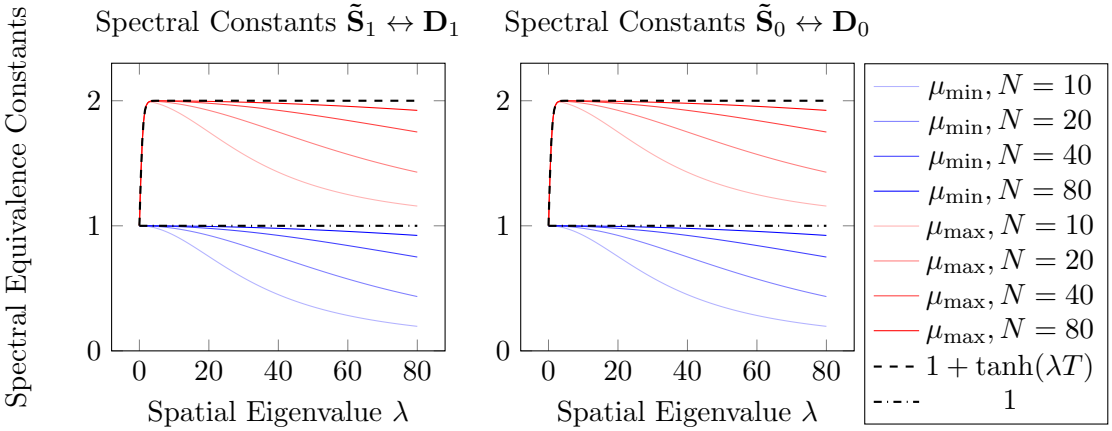

Figure 5.2: Minimal and maximal eigenvalues corresponding to the generalized eigenvalue problems (5.5) and (5.6) for $T = 1.5$.

In Figure 5.2 the minimal and maximal eigenvalues corresponding to (5.5) and (5.6) respectively are shown, which agree as we have shown before. We also observe for fixed $\lambda \geq 0$ pointwise convergence to 1 and $1 + \tanh(\lambda T)$ respectively as $h \to 0$. But in comparison to the eigenvalues depicted in Figure 5.1, the minimal eigenvalue is not bounded from below by 1. Instead, for $\lambda = 0$, the minimal eigenvalues all begin at 1, and then as $\lambda \to \infty$ they converge towards an h-dependent limit lower than 1. Due to the structure of $\tilde{\mathbf{S}}_1$ and $\mathbf{D}_1$, the minimal and maximal eigenvalues converge towards the extremal eigenvalues of the eigenvalue problem $\tilde{\mathbf{M}}^{11}_{0,;0,}\underline{y} = \mu\mathbf{M}^{11}_{0,;0,}\underline{y}$ as $\lambda \to \infty$. In fact, using the definition of the matrices $\tilde{\mathbf{M}}^{11}_{0,;0,}$ and $\mathbf{M}^{11}_{0,;0,}$, this means that

$$\lim_{\lambda\to\infty}\mu_{\min}(\lambda,h) = \inf_{0\neq z_h\in Y_{1,h}}\frac{\|Q_h^0 z_h\|^2_{L^2(0,T)}}{\|z_h\|^2_{L^2(0,T)}}, \quad \lim_{\lambda\to\infty}\mu_{\max}(\lambda,h) = \sup_{0\neq z_h\in Y_{1,h}}\frac{\|Q_h^0 z_h\|^2_{L^2(0,T)}}{\|z_h\|^2_{L^2(0,T)}}.$$

Remark 5.2. *If it is possible to realize the system matrices* $\mathbf{S}_1$ *and* $\mathbf{S}_0$ *exactly, as it is the case in our simple model problem, it is of course preferable to use them for the computation of the optimal discrete state. But if the computation is not that easy, Figure 5.2 suggests that using* $\tilde{\mathbf{S}}_1$ *and* $\tilde{\mathbf{S}}_0$ *instead also yields a stable scheme, supposed that* $h > 0$ *is chosen sufficiently small. Alternatively, one could also replace* $\mathbf{S}_1$ *and* $\mathbf{S}_0$ *by* $\mathbf{D}_1$ *and* $\mathbf{D}_0$ *respectively, since they are spectrally equivalent, as shown in Figure 5.1.*

Interpolated Formulation

We want to analyze the generalized eigenvalue problem

$$\mathbf{S}_{1/2}\underline{y} = \mu \mathbf{D}_{1/2}\underline{y}, \tag{5.11}$$

and the ones where $\mathbf{S}_{1/2}$ is replaced either by $\tilde{\mathbf{S}}_{1/2}$, $\tilde{\mathbf{S}}_{1/2}^{h/2}$ or $\tilde{\mathbf{S}}_{1/2}^{\mathcal{H}_T}$. Since it is not clear how to compute the respective eigenvalues exactly, we only inspect the numerical results.

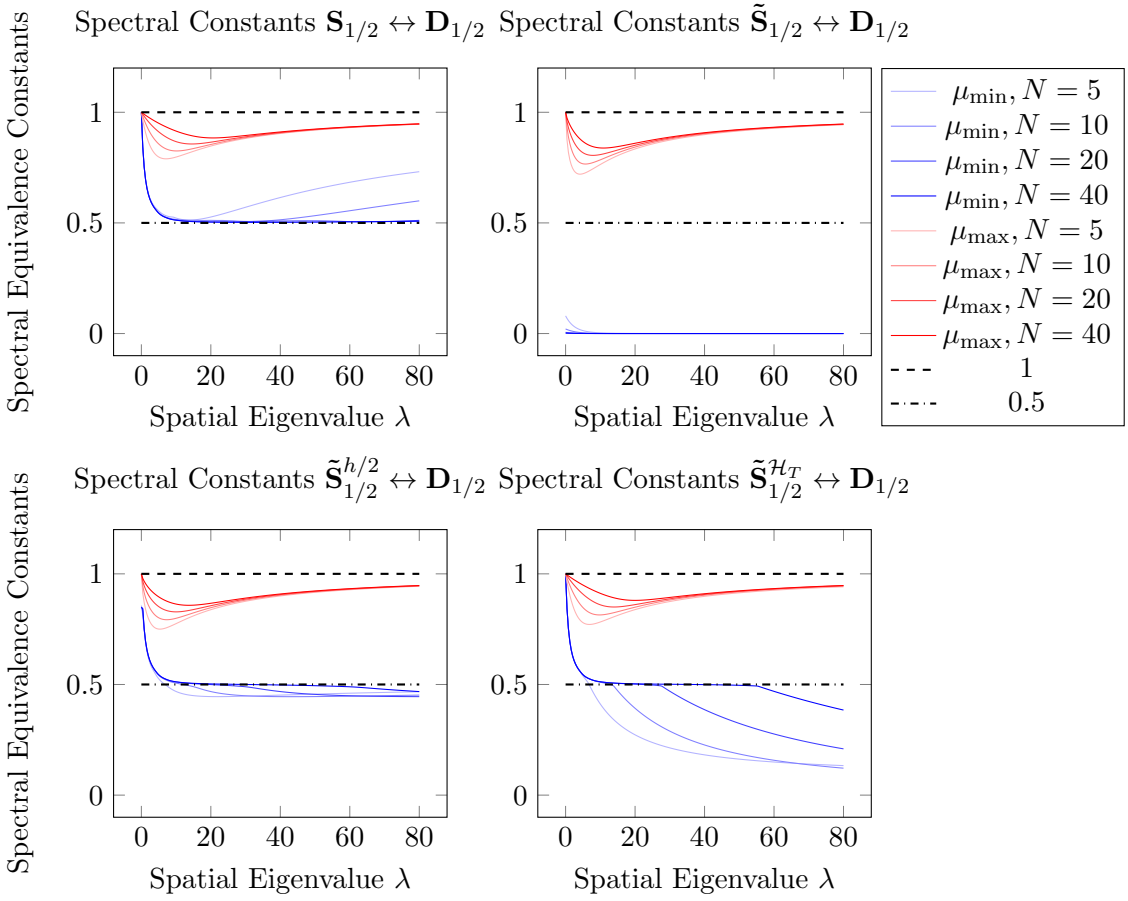

Figure 5.3: Minimal and maximal eigenvalues corresponding to the eigenvalue problem (5.11) for $T = 1$, and the ones where $\mathbf{S}_{1/2}$ is replaced by $\tilde{\mathbf{S}}_{1/2}$, $\tilde{\mathbf{S}}_{1/2}^{h/2}$ or $\tilde{\mathbf{S}}_{1/2}^{\mathcal{H}_T}$.

In Figure 5.3 the minimal and maximal eigenvalues depending on the spatial eigenvalue λ are depicted for all relevant generalized eigenvalue problems. In the upper left plot, we see that the spectral equivalence constants are bounded by $\frac{1}{2}$ and 1. In light of Lemma 2.4, this strongly suggests that the sharp inf-sup constant in Lemma 3.3 is not $\frac{1}{2}$, but rather $\frac{1}{\sqrt{2}}$. The behaviour of the largest eigenvalue is similar in all four eigenvalue problems, it is 1 for $\lambda \to 0$, then decreases slightly, before increasing again and seemingly converging towards 1 as $\lambda \to 0$. In particular, it is bounded by 1 in all four scenarios. In contrast, the minimal eigenvalue is only bounded from below by $\frac{1}{2}$ when working with the exact system matrix $\mathbf{S}_{1/2}$. As mentioned in the last section, using the matrix $\tilde{\mathbf{S}}_{1/2}$ from the naive approach results in an approximation which is not inf-sup stable, as it can be seen from the lowest eigenvalue in the respective plot. In comparison, both the matrices $\tilde{\mathbf{S}}_{1/2}^{h/2}$ and $\tilde{\mathbf{S}}_{1/2}^{\mathcal{H}_T}$ show improved lower bounds. Aside from the exact system matrix $\mathbf{S}_{1/2}$, the minimal eigenvalue is closest to 1 for the matrix $\tilde{\mathbf{S}}_{1/2}^{h/2}$.

Remark 5.3. *Since the computation of the exact matrix $\mathbf{S}_{1/2}$ is already difficult for our simple model problem, and merely possible via an approximation using Fourier series, it becomes intractable for problems involving space and time. Instead, one has to either use $\tilde{\mathbf{S}}_{1/2}^{h/2}$ or $\tilde{\mathbf{S}}_{1/2}^{\mathcal{H}_T}$, since $\tilde{\mathbf{S}}_{1/2}$ does not provide a stable approximation. Alternatively, one can also replace $\mathbf{S}_{1/2}$ by $\mathbf{D}_{1/2}$ as described in* [23].

5.3 Open-Loop Regularization

In this section we will compare the optimal states obtained by solving the primal, the adjoint, and the interpolated formulation on the discrete level. Regardless of the formulation in use, we always solve a variational formulation where we aim to find $y_h \in Y_h$ such that

$$\langle y_h, v_h \rangle_{L^2(0,T)} + \varrho \langle S y_h, v_h \rangle_{(0,T)} = \langle \tilde{y}_d, v_h \rangle_{L^2(0,T)}, \tag{5.12}$$

with $\tilde{y}_d = y_d + y_e$ denoting the sum of the desired target y_d and the homogeneous extension y_e of the initial value y_0. Depending on the formulation, we either use S_1, S_0 or $S_{1/2}$ together with the appropriate discrete ansatz spaces. Using Lemma 2.8, which states that

$$\|y_h + y_e - y_d\|_{L^2(0,T)} \leq \|y + y_e - y_d\|_{L^2(0,T)} + \inf_{v_h \in Y_h} \sqrt{\|y - v_h\|_{L^2(0,T)}^2 + \varrho \|y - v_h\|_S^2},$$

one can transfer the error estimates for y from the continuous setting to the discrete setting, but then they depend on both the regularization parameter ϱ and the mesh size h. With the aim of achieving a convergence of y_h to the target y_d with respect to h as $h \to 0$, we can derive necessary relations between the regularization parameter $\varrho > 0$ and the mesh size $h > 0$ depending on the used formulation, as it is done in [23]. Instead of proving all technical details related to these error estimates, we will only provide the key arguments, and check the plausibility of those claims numerically.

Proposition 5.4. *Let $y_h^{(1)} \in H^1_{0,}(0,T)$ be the unique solution of the discrete primal formulation. Choosing $\varrho = ch^2$, it holds for all $s \in [0,2]$ that*

$$\|y_h^{(1)} + y_e - y_d\|_{L^2(0,T)} \leq Ch^s \|y_d - y_e\|_{H^s(0,T)},$$

given that $y_d \in H^s(0,T)$, with the additional constraint $y(0) = 0$ for all $s > \frac{1}{2}$, and $\dot{y}(T) + \lambda y(T) = 0$ for $s > \frac{3}{2}$.

Proof: We will use Lemma 2.8 and Corollary 2.7 for the proof. We denote by $y^{(1)}$ the solution of the continuous problem, and note that y_e is smooth, hence $y_e \in H^s(0,T)$ for all $s > 0$. First assume that $y_d \in L^2(0,T)$, then

$$\|y_h^{(1)} - y_e + y_d\|_{L^2(0,T)} \leq \underbrace{\|y^{(1)} + y_e - y_d\|_{L^2(0,T)}}_{\leq \|y_d - y_e\|_{L^2(0,T)}} + \sqrt{\underbrace{\|y^{(1)}\|^2_{L^2(0,T)}}_{\leq \|y_d - y_e\|^2_{L^2(0,T)}} + \underbrace{\varrho \|y^{(1)}\|^2_{S_1}}_{\leq \|y_d - y_e\|^2_{L^2(0,T)}}}.$$

$$\Rightarrow \|y_h^{(1)} - y_e + y_d\|_{L^2(0,T)} \leq (1 + \sqrt{2}) \|y_d - y_e\|_{L^2(0,T)}.$$

If $y_d \in H^1(0,T)$ with $y_d(0) = y_0$ we have $\tilde{y}_d = y_d - y_e \in H^1_{0,}(0,T)$, so we can use the L^2-projection $v_h = Q_h^1 y^{(1)}$ to obtain

$$\|y_h^{(1)} - y_e + y_d\|_{L^2(0,T)} \leq \underbrace{\|y^{(1)} + y_e - y_d\|_{L^2(0,T)}}_{\leq \varrho^{1/2} \|y_d - y_e\|_{S_1}} + \sqrt{\underbrace{\|y^{(1)} - v_h\|^2_{L^2(0,T)}}_{\leq Ch^2 \|y^{(1)}\|^2_{H^1(0,T)}} + \varrho \underbrace{\|y^{(1)} - v_h\|^2_{S_1}}_{\leq C \|y^{(1)}\|^2_{H^1(0,T)}}}.$$

Since we know that $\|\cdot\|_{H^1(0,T)}$, $\|\cdot\|_{H^1_{0,}(0,T)}$ and $\|\cdot\|_{S_1}$ are all equivalent norms in the space $H^1_{0,}(0,T)$, and that $\|y^{(1)}\|_{S_1} \leq \|y_d - y_e\|_{S_1}$, we further conclude

$$\|y_h^{(1)} - y_e + y_d\|_{L^2(0,T)} \leq C \left(\varrho^{1/2} + \sqrt{h^2 + \varrho} \right) \|y_d - y_e\|_{H^1(0,T)} \leq Ch \|y_d - y_e\|_{H^1(0,T)},$$

where we have used $\varrho = ch^2$ in the last step.

Finally, assume that $y_d \in H^2(0,T)$ with $y_d(0) = 0$ and $\dot{y}_d(T) + \lambda y_d(T) = 0$, then $S_1(y_d - y_e) \in L^2(0,T)$, thus choosing $v_h = Q_h^1 y^{(1)}$ yields

$$\|y_h^{(1)} - y_e + y_d\|_{L^2(0,T)} \leq \underbrace{\|y^{(1)} + y_e - y_d\|_{L^2(0,T)}}_{\leq \varrho \|S_1(y_d - y_e)\|_{L^2(0,T)}} + \sqrt{\underbrace{\|y^{(1)} - v_h\|^2_{L^2(0,T)}}_{\leq Ch^4 \|y^{(1)}\|^2_{H^2(0,T)}} + \varrho \underbrace{\|y^{(1)} - v_h\|^2_{S_1}}_{\leq Ch^2 \|y^{(1)}\|^2_{H^2(0,T)}}}.$$

Using that $\|S_1 y^{(1)}\|_{L^2(0,T)} \leq \|S_1(y_d - y_e)\|_{L^2(0,T)}$, which can be shown in a similar fashion as the estimates in Lemma 2.6, and that $\|S_1 \cdot\|_{L^2(0,T)}$ and $\|\cdot\|_{H^2(0,T)}$ are equivalent norms, we finally obtain

$$\|y_h^{(1)} - y_e + y_d\|_{L^2(0,T)} \leq C \left(\varrho + \sqrt{h^4 + \varrho h^2} \right) \|y_d - y_e\|_{H^2(0,T)} \leq Ch^2 \|y_d - y_e\|_{H^2(0,T)},$$

where we have again used $\varrho = ch^2$ in the last step. The rest of the claim follows from interpolation of these three inequalities. □

Proposition 5.5. *Let $y_h^{(0)} \in L^2(0,T)$ be the unique solution of the discrete adjoint formulation. Choosing $\varrho = 0$, it holds for all $y_d \in H^s(0,T)$ with $s \in [0,1]$ that*

$$\|y_h^{(0)} - y_d\|_{L^2(0,T)} \leq Ch^s \|y_d\|_{H^s(0,T)}.$$

Proof: We denote by $y^{(0)}$ the related solution of the continuous problem, and will again utilize the estimates from Corollary 2.7 for the proof. If we choose $\varrho = 0$, we know that $\|y^{(0)} - y_d\|_{L^2(0,T)} = 0$, such that Cea's Lemma reduces to

$$\|y_h^{(0)} - y_d\|_{L^2(0,T)} \leq \inf_{v_h \in S_h^0(0,T)} \|y^{(0)} - v_h\|_{L^2(0,T)}.$$

Choosing the L^2-projection $v_h = Q_h^0 y^{(0)}$, and assuming that $y_d \in H^s(0,T)$ does imply $y^{(0)} \in H^s(0,T)$, and that $\|y^{(0)}\|_{H^s(0,T)}$ can be bound from above by $C\|y_d\|_{H^s(0,T)}$ we get

$$\|y_h^{(0)} - y_d\|_{L^2(0,T)} \leq \|y^{(0)} - Q_h^0 y^{(0)}\|_{L^2(0,T)} \leq Ch^s \|y^{(0)}\|_{H^s(0,T)} \leq Ch^s \|y_d\|_{H^s(0,T)}. \qquad \square$$

Proposition 5.6. *Let $y_h^{(1/2)} \in H_{0,}^{1/2}(0,T)$ be the unique solution of the discrete interpolated formulation. Choosing $\varrho = ch$, it holds for all $s \in [0,1]$ that*

$$\|y_h^{(1/2)} + y_e - y_d\|_{L^2(0,T)} \leq Ch^s \|y_d - y_e\|_{H^s(0,T)},$$

given that additionally $y_d(0) = y_0$ holds if $s > \frac{1}{2}$.

Proof: We will use make use of the abstract framework described in Corollary 2.7 once more. For $s = 0$ the estimate from the proposition follows similarly to the estimate from Proposition 5.4. Now let us assume that $(y_d - y_e) \in H_{0,}^{1/2}(0,T)$, then we can use best approximation results for the L^2-projection and Cea's Lemma as we have done in the proof of Proposition 5.4 in order to show

$$\|y_h^{(1/2)} + y_e - y_d\|_{L^2(0,T)} \leq C(\varrho^{1/2} + \sqrt{h+\varrho}) \|y_d - y_e\|_{H^{1/2}(0,T)} \leq Ch^{1/2} \|y_d - y_e\|_{H^{1/2}(0,T)},$$

where we have used $\varrho = ch$ in the last step. Similarly, if $y_d \in H^1(0,T)$ with $y_d(0) = y_0$ we know that

$$\|y_h^{(1/2)} + y_e - y_d\|_{L^2(0,T)} \leq C\left(\varrho + \sqrt{h^2 + \varrho h}\right) \|y_d - y_e\|_{H^1(0,T)} \leq Ch \|y_d - y_e\|_{H^1(0,T)},$$

if we again use the relation $\varrho = ch$. $\square$

In order to numerically test those estimates we compute the L^2-error for targets with

varying regularity. We choose $T = 5$, $\lambda = 1.5$, $y_0 = 0.5$ and consider the targets

$$y_{d,1}(t) = \begin{cases} 2, & \text{if } t < 0.5T \\ 0.5, & \text{if } 0.5T \leq t \leq 0.75T \\ 0.5 + t - 0.75T, & \text{if } 0.75T < t \end{cases} \in H^{1/2-\varepsilon}(0,T),$$

$$y_{d,2}(t) = y_0 + e^{t/4} \in C^\infty(0,T),$$

$$y_{d,3}(t) = y_0 + \sqrt{t} \in H^{1-\varepsilon}(0,T),$$

$$y_{d,4}(t) = y_0 + e^{t/4} - 1 \in C^\infty(0,T),$$

$$y_{d,5}(t) = y_0 + e^{t/4} - 1 - \frac{(\lambda(y_0 - 1) + e^{T/4}(\lambda + \frac{1}{4}))}{1 + \lambda T} t \in C^\infty(0,T).$$

Then $y_{d,1}$ has jump discontinuities, and the lowest regularity with $y_{d,1} \in H^{1/2-\varepsilon}(0,T)$ for all $\varepsilon > 0$. The target $y_{d,3}$ has a higher regularity with $\in H^{1-\varepsilon}(0,T)$ for all $\varepsilon > 0$. The remaining targets $y_{d,2}$, $y_{d,4}$ and $y_{d,5}$ are all smooth, but only $y_{d,4}$ and $y_{d,5}$ satisfy the initial condition $y_d(0) = y_0 = 0.5$. Further, $y_{d,5}$ is the only target that also satisfies the terminal condition $\dot{y}_{d,5}(T) + \lambda y_{d,5}(T) = 0$ in addition to the initial condition.

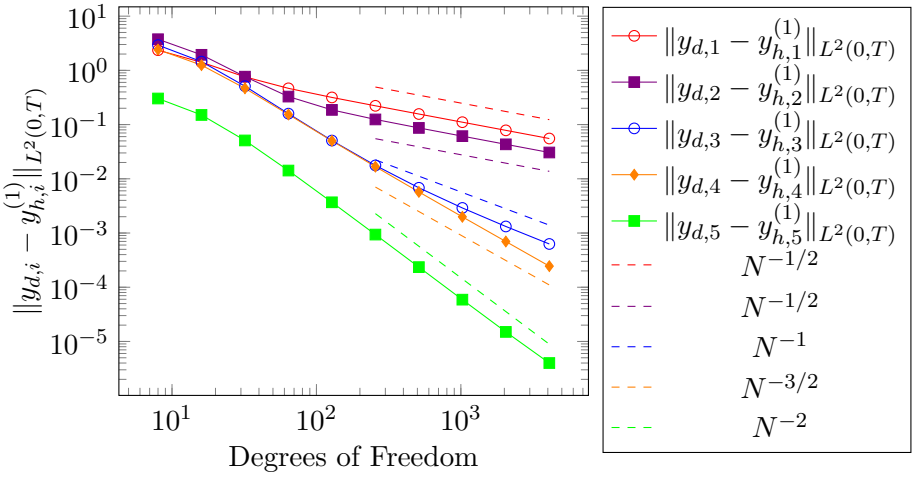

Figure 5.4: $L^2(0,T)$-Error of 5 targets for primal formulation with varying regularity, together with the anticipated error rates (dashed lines).

Figure 5.4 illustrates the $L^2(0,T)$-error related to the solution of the discrete primal problem for the different targets. As we would expect, the error related to the target $y_{d,1}$ asymptotically converges to 0 with a rate of $N^{-1/2}$ in terms of the degrees of freedom, or with a rate of $h^{1/2}$ in terms of the mesh size, due to the low regularity $y_{d,1} \in H^{1/2-\varepsilon}(0,T)$. Even though the target $y_{d,2}$ is smooth, the error asymptotically also behaves like $h^{1/2}$ due to the mismatch of the initial values. For $y_{d,3} \in H^{1-\varepsilon}(0,T)$ the initial condition is satisfied, therefore we asymptotically observe linear order of convergence. The target $y_{d,4}$ is smooth and satisfies the initial condition, but not the terminal condition specified in Proposition 5.4, therefore the error only behaves like $h^{3/2}$. Finally, the target $y_{d,5}$

is smooth, and satisfies both the initial and the terminal condition set forth in the proposition, therefore the related convergence rate is quadratic.

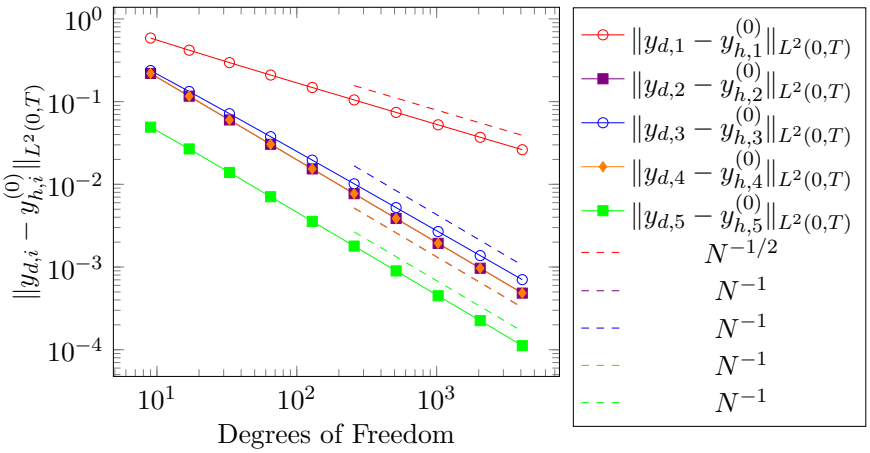

Figure 5.5: $L^2(0,T)$-Error of 5 targets for adjoint formulation with varying regularity, together with the anticipated error rates (dashed lines).

The resulting errors when using the adjoint formulation are depicted in Figure 5.5. As stated in Proposition 5.5, the error related to the target $y_{d,1} \in H^{1/2-\varepsilon}(0,T)$ asymptotically converges to 0 like $h^{1/2}$, while all other errors admit a linear order of convergence due to sufficient regularity. So in contrast to the primal formulation, the initial and the terminal condition do not matter at all, and the best order of convergence we can achieve is a linear one. The error rate related to the adjoint formulation is only superior to the rate related to the primal formulation for the second target $y_{d,2}$, which is smooth but does not satisfy the initial condition.

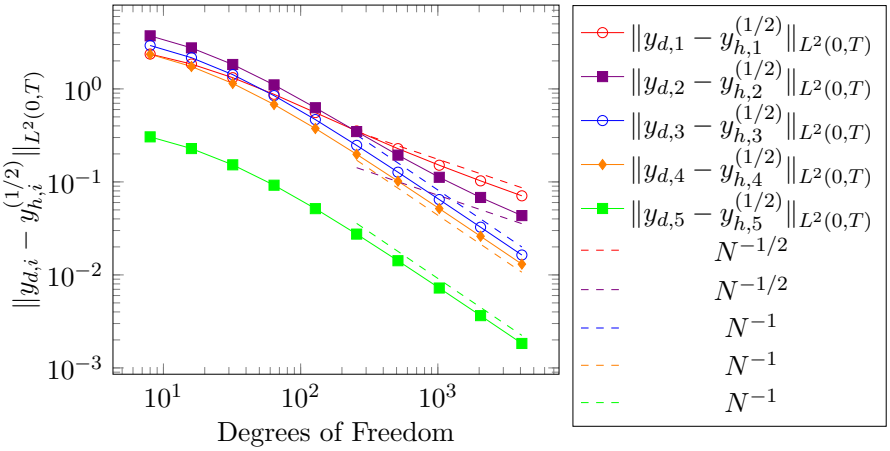

Figure 5.6: $L^2(0,T)$-Error of 5 targets for interpolated formulation with varying regularity, together with the anticipated error rates (dashed lines).

The errors obtained by using the interpolated formulation are shown in Figure 5.6. As we have also seen when working with the other formulations, the error related to the target $y_{d,1}$ converges to 0 with a rate of $h^{1/2}$. Similarly, the error related to $y_{d,2}$ asymptotically converges to 0 with the same rate $h^{1/2}$ due to the mismatch in initial conditions, even though the rates on the plot depict a slightly improved behaviour for the refinement steps shown here. All remaining targets satisfy the initial condition and are at least in $H^{1-\varepsilon}(0,T)$, therefore they all asymptotically admit a linear order of convergence, which is in agreement with the results shown in the plot. As it was the case when working with the adjoint formulation, a linear order of convergence is the best rate that can be achieved. But in contrast to the adjoint formulation, the initial condition also influences the order of convergence. Finally, note that the error rates for the first three targets $y_{d,1}$, $y_{d,2}$ and $y_{d,3}$ are the same as for the primal formulation, and for $y_{d,4}$ and $y_{d,5}$ they are the same as for the adjoint formulation.

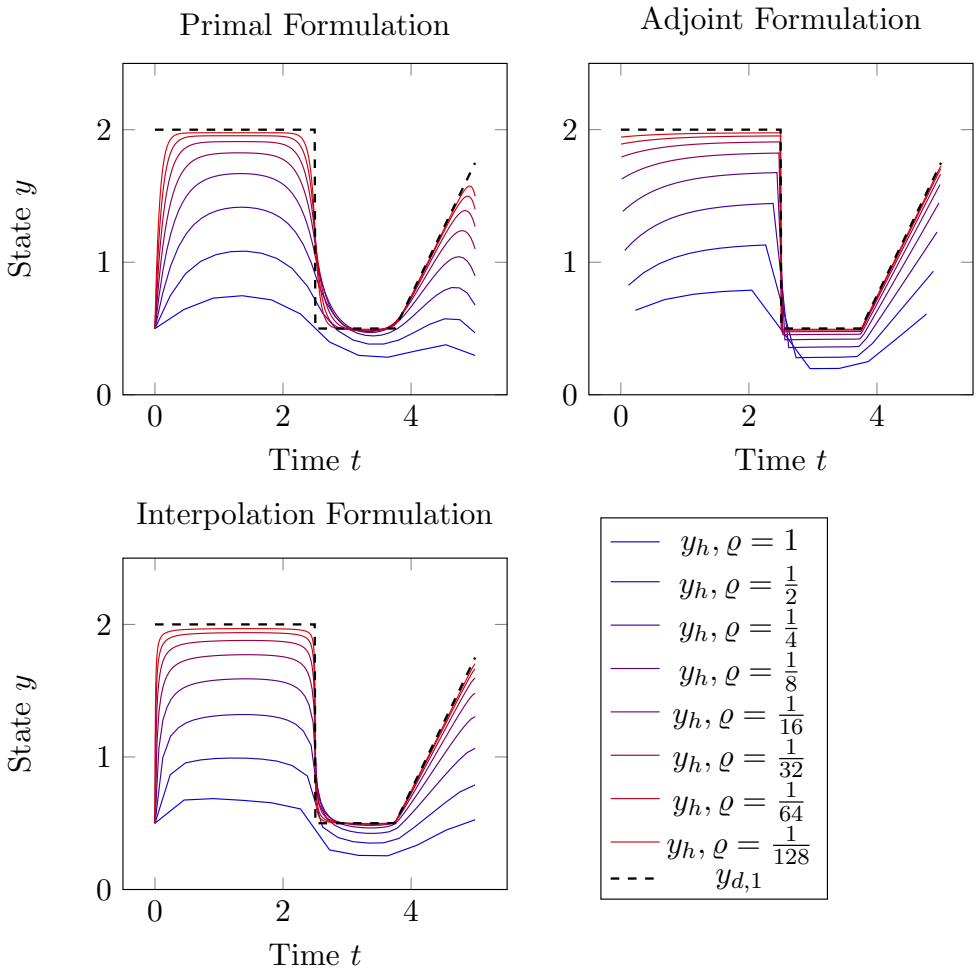

Figure 5.7: Comparison of different open-loop controls for the target $y_{d,1}$.

We conclude this chapter with a visual comparison of the optimal states obtained from each of the three formulations. Figure 5.7 shows the resulting trajectories for different choices of the regularization parameter $\varrho > 0$ and the target $y_{d,1}$. In order to make the visualizations comparable, we have also used a non-zero regularization $\varrho > 0$ for the adjoint formulation. For large values of ϱ the plots seemingly show that the trajectories associated with the primal and the interpolation formulation admit multiple kinks, but this is only an artefact from the coarse discretization. As $\varrho \to 0$ the mesh size is adapted suitably, therefore the kinks seemingly disappear. In contrast, the trajectory related to the adjoint formulation actually admits discontinuities. As we would expect, all three trajectories converge to the target trajectory pointwise in regions where y_d is continuous. Still, one can visually see that the interpolated formulation allows for steeper slopes in comparison to the primal formulation, leading to a faster reduction of the L^2-Error. Further, we can clearly see that only the trajectory associated to the primal formulation satisfies the terminal condition $\dot{y}(T) + \lambda y(T) = 0$, which translates to the condition of the optimal control converging to 0 as $t \to T$. Moreover, note that the trajectory related to the adjoint formulation does not satisfy the initial condition, since it is incorporated only in a weak sense. Even though this may seem as an advantage due to the improved error rate for mismatching initial values, it makes it difficult to apply the adjoint formulation in practice, since the related control is possibly a distribution instead of a classical function. When tracking discontinuous targets, the advantage of the interpolated formulation is that we obtain an optimal control in form of a function instead of a distribution, but the related mesh size can be chosen coarser in comparison to the primal formulation, resulting in dense system matrices of smaller size.

5.4 Comparison of Closed- and Open-Loop Regularization

Finally, we want to visualize the difference between open- and closed-loop optimal control policies. For that purpose we will only consider the primal formulation, but using MPC one could in principle also use the interpolated formulation to construct a closed-loop control policy. We choose $T = 5$, $\lambda = 0.5$, $y_0 = 0.5$, $\varrho = 0.005$ and the discontinuous, but piecewise linear target $y_{d,1}$ introduced in the last section. In order to simulate uncertainty, we also introduce a noise level $\delta > 0$ and a sequence of independent Gaussian random variables $\eta_j \sim \mathcal{N}(0, \delta^2)$, $j \in \{1, ..., N\}$. Then the respective trajectory is approximately computed in the points t_j using the explicit Euler-method

$$y_{j+1} = y_j + h(u_j - \lambda y_j + \eta_j),$$

with u_j denoting the control in the j-th step. For the open-loop control, we compute the control without any noise, by computing the optimal state y_h like described in the last section, and setting $u_h = \mathbf{B}_1 y_h + f_h$. Then we plug the control evaluated in the points t_j into the Euler-scheme to obtain the perturbed trajectory. For the closed-loop control, we just set $u_j = u(t_j, y_j)$ using the formula derived in Theorem 4.13. Note that we do not have to perform the integration analytically, and can use numerical integration schemes like Gaussian quadrature instead, since the integrand is smooth and bounded.

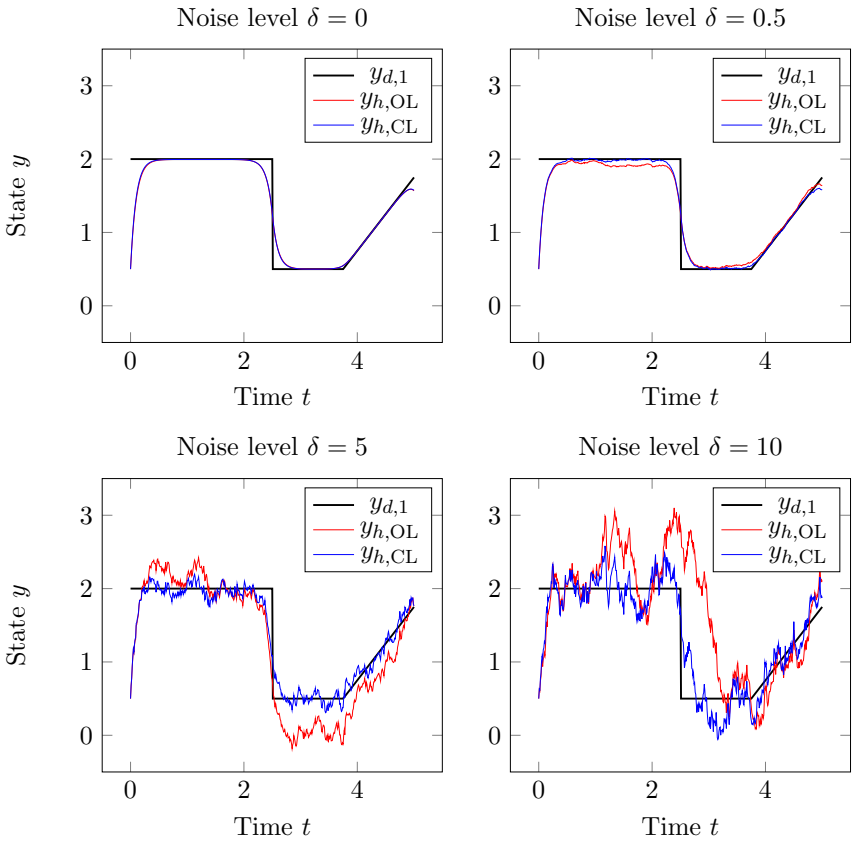

Figure 5.8: Open-Loop (OL) vs Closed-Loop (CL) regularization for varying noise levels.

Figure 5.8 shows the results of the numerical experiments, for which we have chosen $N = 500$ time intervals, and use varying noise levels $\delta \in \{0, 0.5, 5, 10\}$. For $\delta = 0$, there is only a very small difference between the trajectories obtained by the open- and closed-loop control strategies, which is an artefact from the discretization. On the continuous level, those trajectories would coincide without the presence of uncertainties. For increasing noise level we unsurprisingly observe a greater variance in both trajectories from the open- and closed-loop approach, but apparently the difference to the target trajectory is way smaller when using the closed-loop control. This is in agreement with the theory, since the closed-loop approach always applies the optimal control, even if the state does not follow the preplanned trajectory, while the open-loop approach only provides the optimal control, given that the current state agrees with the preplanned optimal trajectory. Further, note that the difference between the open- and the closed-loop control can be neglected if the noise level $\delta > 0$ is sufficiently small. But this is only possible since the dynamic system associated to the ordinary differential equation

$$\dot{y}(t) + \lambda y(t) = 0$$

with $\lambda > 0$ is asymptotically stable.

6 Conclusions

In this work we have analyzed optimal control problems related to different variational formulations of the heat equation, namely the primal, the adjoint and the interpolated formulation. Instead of directly working with the heat equation as a constraint, a spatial eigenfunction expansion was used to reduce the original control problem to a sequence of independent control problems, each only constrained by an ordinary differential equation. For these reduced problems we have derived error and control cost estimates depending on the regularization parameter, and an optimal closed-loop control for the primal formulation. On the discrete level, we have also derived error estimates using Cea's lemma, analyzed the stability of different discretizations and described the optimal relation between the mesh size and the regularization parameter for all formulations.

The numerical experiments conducted in this work indicate that the adjoint formulation is not suitable for deriving closed-loop control, since it neither allows the point evaluation of the state, nor allows an incorporation of the initial value in more than just the weak sense. In contrast, for the primal formulation both open-loop and closed-loop controls can be derived by means of solving the HJB equation and the optimality system related to PMP in an efficient way. For the interpolated formulation, it is not obvious how derive a closed-loop control based on the HJB equation, but both the theoretical and the numerical results suggest that using Model-Predictive-Control (MPC) could result in a closed-loop control suitable for discontinuous targets. Since the solution of the interpolated formulation admits sharper edges around discontinuities of the target in comparison to the solution of the primal formulation, the mesh size can be chosen larger while still maintaining a similar error as in the primal formulation. Even though this larger mesh size can potentially reduce the runtime, it is important to note that this advantage diminishes if the related system matrices are assembled in a naive way, since they are dense in contrast to the sparse matrices appearing in the primal formulation. So in order to fully exploit the larger mesh size, efficient strategies like described in [30] have to be applied. Another advantage of the interpolated formulation, which facilitates the combination with MPC, is that the related optimal states do not have to satisfy an artificial terminal boundary condition. So in contrast to the primal formulation, even without penalizing the terminal state, the related optimal control does not converge to 0 at the end of the time interval.

We finally conclude that the interpolated formulation could potentially have advantages over the primal formulation when tracking discontinuous targets, especially for non-linear problems where ordinary closed-loop strategies relying on the HJB equation are difficult to realize numerically. In contrast, for linear systems and sufficiently regular targets, it seems that the additional complexity accompanied with the interpolated formulation does not provide any advantages in comparison to the primal formulation.

Bibliography

[1] B. D. O. Anderson and J. B. Moore. *Optimal Control: Linear Quadratic Methods*. Prentice-Hall Information and System Sciences Series. Prentice-Hall, Englewood Cliffs, NJ, 1989.

[2] W. Arendt and K. Urban. *Partielle Differenzialgleichungen*. Springer Spektrum Berlin, Heidelberg, 2018.

[3] J. Awrejcewicz. *Classical mechanics*. Advances in Mechanics and Mathematics. Springer, New York, 2012. Dynamics.

[4] P. Benner and H. Faßbender. *Modellreduktion*. Springer Studium Mathematik (Master). Springer Spektrum Berlin, Heidelberg, 2024.

[5] D. Boffi, F. Brezzi, and M. Fortin. *Mixed Finite Element Methods and Applications*. Springer Series in Computational Mathematics. Springer Berlin, Heidelberg, 1 edition, 2013.

[6] A. Borzi. Multigrid methods for parabolic distributed optimal control problems. *Journal of Computational and Applied Mathematics*, 157(2):365–382, 2003.

[7] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[8] A. Deriglazov. *Classical mechanics*. Springer, Heidelberg, 2010.

[9] E. DiBenedetto. *Classical mechanics*. Cornerstones. Birkhäuser/Springer, New York, 2011.

[10] L. C. Evans. *Partial differential equations*. American Mathematical Society, 2010.

[11] M. Feuerle, R. Löscher, O. Steinbach, and K. Urban. A unified framework for the analysis, numerical approximation and model reduction of linear operator equations, Part I: Well-posedness in space and time. 2025. accepted in SINUM. `arXiv:2508.05407`.

[12] H. Fischer and H. Kaul. *Mathematik für Physiker Band 3*. Springer Spektrum Berlin, Heidelberg, 4 edition, 2017.

[13] H. Goldstein. *Classical Mechanics*. Addison-Wesley, 2 edition, 1980.

[14] C. G. Gray and E. F. Taylor. When action is not least. *American Journal of Physics*, 75(5):434–458, 2007.

[15] L. Grüne and J. Pannek. *Nonlinear Model Predictive Control*. Communications and Control Engineering. Springer Cham, 2016.

[16] M. Gubisch and S. Volkwein. Proper orthogonal decomposition for linear-quadratic optimal control. In *Model Reduction and Approximation*, chapter 1, pages 3–63. 2017.

[17] A. Kröner, K. Kunisch, and H. Zidani. Optimal feedback control of undamped wave equations by solving a HJB equation. *ESAIM: Control, Optimisation and Calculus of Variations*, 21(2):442 – 464, 2014.

[18] U. Langer, O. Steinbach, F. Tröltzsch, and H. Yang. Space-time finite element discretization of parabolic optimal control problems with energy regularization. *SIAM Journal on Numerical Analysis*, 59(2):675–695, 2021.

[19] U. Langer, O. Steinbach, and H. Yang. Robust space-time finite element methods for parabolic distributed optimal control problems with energy regularization. *Advances in Computational Mathematics*, 50(2):24, 2024.

[20] J. L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg, 1 edition, 1971.

[21] J. L. Lions and E. Magenes. *Non-Homogeneous Boundary Value Problems and Applications*. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg, 1 edition, 1972.

[22] A. Locatelli. *Optimal Control: An Introduction*. Systems & Control: Foundations & Applications. Birkhäuser, Boston, 2001.

[23] R. Löscher, M. Reichelt, and O. Steinbach. Optimal complexity solution of space-time finite element systems for state-based parabolic distributed optimal control problems. *Journal of Complexity*, 92:101976, 2026.

[24] R. Löscher, O. Steinbach, and M. Zank. On a modified hilbert transformation, the discrete inf-sup condition, and error estimates. *Computers & Mathematics with Applications*, 171:114–138, 2024.

[25] H. McLean and W. Charles. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, 2000.

[26] B. Schweizer. *Partielle Differentialgleichungen*. Masterclass. Springer Spektrum Berlin, Heidelberg, 3 edition.

[27] O. Steinbach. *Lösungsverfahren für lineare Gleichungssysteme*. Mathematik für Ingenieure und Naturwissenshaftler, Ökonomen und Landwirte. Vieweg+Teubner Verlag Wiesbaden, 2005.

[28] O. Steinbach. *Numerical Approximation Methods for Elliptic Boundary Value Problems*. Springer New York, NY, 1 edition, 2008.

[29] O. Steinbach and M. Zank. Coercive space-time finite element methods for initial boundary value problems. *Electron. Trans. Numer. Anal.*, 52:154–194, 2020.

[30] O. Steinbach and M. Zank. A note on the efficient evaluation of a modified hilbert transformation. *Journal of Numerical Mathematics*, 2020.

[31] M. Zank. *Inf-Sup Stable Space-Time Methods for Time-Dependent Partial Differential Equations*, volume 36 of *Monographic Series TU Graz / Computation in Engineering and Science*. Verlag der Technischen Universität Graz, Graz, Austria, 2020.

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitely indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present masters thesis.

Date	Signature