Computer-assisted proofs of non-reachability for linear parabolic PDEs under bounded control constraints

Ivan Hasenohr, Camille Pouchol, Yannick Privat, Christophe Zhang

Computer-assisted proofs of non-reachability for linear parabolic PDEs under bounded control constraints

Ivan Hasenohr* Camille Pouchol† Yannick Privat‡ Christophe Zhang§

November 10, 2025

Abstract

Analysing reachability associated to a control system is a subtle issue, especially for infinite-dimensional dynamics, and when controls are subject to bounded constraints. We develop a computer-assisted framework for establishing non-reachability in linear parabolic PDEs governed by strongly elliptic operators, extending recent finite-dimensional techniques introduced in [26] to the PDE setting. The non-reachability of a given target is shown to be equivalent to proving that a properly defined dual functional takes negative values. Our approach combines rigorous numerics with explicit convergence estimates for discretisations of the adjoint equation, ensuring mathematically certified results with tight error bounds. We demonstrate the wide applicability of our framework on Laplacian-driven control systems, showcasing its accuracy and reliability under various types of control constraints.

Keywords: linear control systems, heat equation, bounded constraints on the control, non-reachability, minimal times of reachability, computer-assisted proof, rounding errors

AMS classification: 93B03, 49M29, 35K20, 65M22

Contents

*Université Paris Cité, FP2M, CNRS FR 2036, MAP5 UMR 8145, F-75006 Paris, France. (ivan.hasenohr@math.cnrs.fr).

†Université Paris Cité, FP2M, CNRS FR 2036, MAP5 UMR 8145, F-75006 Paris, France. (camille.pouchol@u-paris.fr).

‡Université de Lorraine, CNRS, Inria SPHINX, IECL, F-54000 Nancy, France (yannick.privat@univ-lorraine.fr).

§Université de Lorraine, CNRS, Inria SPHINX, IECL, F-54000 Nancy, France (christophe.zhang@polytechnique.org).

1 Introduction

1.1 Non-reachability issues

The context of the present work is that of constrained reachability, for linear control systems of the form

$$\begin{cases} \dot{y}(t) = Ay(t) + Bu(t), \\ y(0) = y_0. \end{cases} \tag{$\mathcal{S}$}$$

Here X, the state space, and U, the control space, are two (complex) Hilbert spaces, $y_0 \in X$, $A : \mathcal{D}(A) \subset X \to X$ is an unbounded linear operator generating a C_0 semigroup $(S_t)_{t\geq 0}$ over X, and $B : U \to X$ is a bounded linear operator. In this setting, the solution to ($\mathcal{S}$) is well-defined and satisfies $y \in \mathcal{C}^0([0,\infty); X) \cap \mathcal{C}^1((0,\infty); \mathcal{D}(A))$.

Definition 1. *Let $y_0, y_f \in X$, $T > 0$ and $\mathcal{U} \subset U$ a constraint set. We say that $y_f \in X$ is $\mathcal{U}$-reachable from y_0 in time $T > 0$ if there exists $u \in L^2(0,T;U)$ such that $u(t) \in \mathcal{U}$ for a.e. $t \in (0,T)$, and such that the solution to ($\mathcal{S}$) associated to u satisfies $y(T) = y_f$.*

Depending on the context or application, one can be interested in proving that a given target is $\mathcal{U}$-reachable at a given time T, or instead, that it is not $\mathcal{U}$-reachable at that time. Non- $\mathcal{U}$-rechability is relevant when the purpose is to ensure that the solution to ($\mathcal{S}$) cannot equal an unwanted element (or, more generally, cannot enter a given set) at time T. Our focus here is on the latter problem.

More precisely, the methodology developed in this work applies to

(i) establishing that a given target y_f is **not** $\mathcal{U}$-reachable from y_0 in time $T > 0$, in the case where

$$\mathcal{U} \text{ is convex, closed and bounded,} \tag{1}$$

(ii) in the setting of (linear) parabolic PDEs.

This paper extends the work [26], which covers the same question within the framework of (i) above, but in the finite-dimensional case. By (ii), we mean that we will deal with a large subclass of mα-accretive operators with a suitable associated discretisation method (of finite element type), see Section 2 for the detailed definitions.

Example 2. *As an example, consider the internally controlled heat equation (on the interval $(0,1)$, with Dirichlet boundary conditions)*

$$\begin{cases} \partial_t y - D\partial_{xx} y = \chi_\omega u, \\ y(t,0) = y(t,1) = 0. \\ y(0) = y_0. \end{cases}$$

with, say

$$D = 1, \quad y_0 = 0, \quad \omega = (\tfrac{1}{5}, \tfrac{2}{5}) \cup (\tfrac{4}{5}, 1), \quad \mathcal{U} = \{v \in L^2(\Omega),\ 0 \leq v \leq 1 a.e.\}. \tag{2}$$

Let $y_f \in L^2(0,1)$ and $T > 0$ be fixed; can one prove that y_f is not $\mathcal{U}$-reachable from y_0 in time $T > 0$? Or, equivalently given that $y_0 = 0$ and $0 \in \mathcal{U}$, can one prove that $T \leq T^\star$ where $T^\star$ is the minimal time $T > 0$ for which y_f is $\mathcal{U}$-reachable from $y_0 \in X$ in time $T > 0$?

1.2 State of the art

Constraint-free controllability for linear PDEs has been the focus of extensive study from the 70's, leading to numerous significant results: general conditions, as the extension of Kalman's criterion for linear PDEs, or more specific results, such as the Geometric Control Condition for the wave equation [32], the approximate controllability and exact null controllability of the heat equation [29] (see [10, 32] for lecture notes compiling most common results). In recent years, some papers have studied constrained controllability, or reachability of linear ODEs and PDEs, uncovering general obstructions to controllability [11, 20, 31], due for example to comparison principles [30], or observability-based criteria caracterising the controllability of various problems with (un-)bounded and (a-)symmetric constraints on the control [7, 8].

In this work, we focus on the reachability analysis of linear parabolic equations, typically with internal controls, a context that has been extensively studied, both in the unconstrained and constrained settings. For instance, for the reachable set of the heat equation with internal unconstrained control [14, 19, 23, 25, 27]. Constraints on the state or the control have also been considered, whether bounded or unbounded: [2, 9, 12, 34, 40].

Numerous methods exist in order to analyse reachability for finite-dimensional linear systems, including Hamilton-Jacobi type PDE formulations [15, 33], barrier functions [28, 35], and set propagation (see the comprehensive review in [1]). To the best of our knowledge, no such method has been designed in an infinite-dimensional setting, in such a way that it is possible to rigorously answer a question such as the one posed in Example 2.

Our method is based on discretising an abstract necessary and sufficient condition with explicit bounds, and using interval arithmetic. Thus, our methodology belongs to the realm of computer-assisted proofs. It is worth noting that computer-assisted proofs based on interval arithmetic for PDEs has been an active and developing area of research: see for example [4–6, 21].

1.3 Methodology and main results

1.3.1 Separation argument

Let $T > 0$ and $\mathcal{U} \subset U$ be a fixed constraint set. We let

$$E_{\mathcal{U}} := \big\{u \in L^2(0,T;U),\ \text{for a.e. } t \in (0,T), u(t) \in \mathcal{U}\big\}.$$

It follows from the linearity of ($\mathcal{S}$) that a given $y_f \in X$ is $\mathcal{U}$-reachable from y_0 in time $T > 0$ if and only if $y_f \in S_T y_0 + L_T E_{\mathcal{U}}$, where $L_T : L^2(0,T;U) \to X$ is the linear continuous operator defined by

$$\forall u \in L^2(0,T;U), \quad L_T u = \int_0^T S_{T-t} B u(t) \mathrm{d}t.$$

Clearly, if one finds a strictly separating hyperplane between $L_T E_{\mathcal{U}}$ and $\{y_f - S_T y_0\}$, i.e., if there exists $p_f \in X$ such that

$$\sup_{z \in L_T E_{\mathcal{U}}} \mathrm{Re}\langle z, p_f\rangle < \mathrm{Re}\langle y_f - S_T y_0, p_f\rangle \quad \Longleftrightarrow \quad \sup_{v \in E_{\mathcal{U}}} \mathrm{Re}\langle v, L_T^* p_f\rangle_{L^2(0,T;U)} < \mathrm{Re}\langle y_f - S_T y_0, p_f\rangle, \tag{3}$$

then y_f is not $\mathcal{U}$-reachable from y_0 in time $T>0$. Introducing the notation $\sigma_C : x \mapsto \sup_{y\in C} \mathrm{Re}\langle x, y\rangle$ for the support function of a set C, and the so-called *dual* functional

$$\forall p_f \in X, \quad J(p_f) := \sigma_{E_\mathcal{U}}(L_T^* p_f) - \mathrm{Re}\langle y_f - S_T y_0, p_f\rangle = \int_0^T \sigma_{B\mathcal{U}}(S_t^* p_f)\mathrm{d}t - \mathrm{Re}\langle y_f - S_T y_0, p_f\rangle,$$

the above condition (3) amounts to $J(p_f) < 0$.

In fact, this sufficient condition becomes an equivalence under our assumption that $\mathcal{U}$ is a convex, closed and bounded case (and hence is convex and weakly compact). A separation argument outlined in detail in [26, Proposition 2], which goes through here when applied in the weak topology of X [37, Theorem 3.4], leads to the following equivalence

$$y_f \text{ is not } \mathcal{U}\text{-reachable from } y_0 \text{ in time } T>0 \quad \Longleftrightarrow \quad \exists p_f \in X,\ J(p_f) < 0. \tag{4}$$

The condition $(\forall p_f \in X,\ J(p_f) \geq 0)$, which is equivalent to $\mathcal{U}$-reachability of y_f from y_0 in time $T>0$, is not amenable to computer-assisted proofs. Its opposite, however, is. Hence, in order to prove that y_f is not $\mathcal{U}$-reachable from y_0 in time $T>0$, we are looking for $p_f \in X$ such that $J(p_f) < 0$. We shall say that such an element p_f is a **dual certificate** of non $\mathcal{U}$-reachability (of y_f from y_0 in time $T>0$).

Remark 3. *This approach straightforwardly extends to tackling the non-reachability of a full set of targets $\mathcal{Y}_f$. By '$\mathcal{Y}_f$ is not $\mathcal{U}$-reachable from y_0 in time T', we mean that for all $y_f \in \mathcal{Y}_f$, y_f is not $\mathcal{U}$-reachable from y_0 in time T. Then, for any closed convex set $\mathcal{Y}_f \subset X$, a separation argument between $L_T E_\mathcal{U}$ and $\mathcal{Y}_f$ leads to, defining $J(p_f;\mathcal{Y}_f) := \sigma_{E_\mathcal{U}}(L_T^* p_f) + \sigma_{\mathcal{Y}_f}(-p_f) + \mathrm{Re}\langle S_T y_0, y_f\rangle$,*

$$\mathcal{Y}_f \text{ is not } \mathcal{U}\text{-reachable from } y_0 \text{ in time } T \quad \Longleftrightarrow \quad \exists p_f \in X,\ J(p_f;\mathcal{Y}_f) < 0.$$

Remark 4. *We are assuming that $\mathcal{U}$ is closed convex and bounded, and that the control operator B is bounded. As can be checked, our approach works under the slightly more general assumptions that $B\mathcal{U} \subset X$ is closed, convex and bounded, and B is an admissible operator, see [39] for a definition of admissibility.*

1.3.2 Minimal times

Let us explain how proving non-reachability at a given time T may lead to estimates for minimal times. First, let us first address the obvious fact that, in the case of bounded constraints, minimal times exist.

Lemma 5. *Let y be the solution to ($\mathcal{S}$), with $(A, \mathcal{D}(A))$ generating a C_0 semigroup, and $B\mathcal{U}$ be bounded by $M_{B\mathcal{U}} > 0$. Then for all $y_f \in X$,*

$$\|y_f - S_T y_0\| > \sup_{t\in[0,T]} \|S_t\|_{\mathcal{L}(X)} T M_{B\mathcal{U}} \quad \Longrightarrow \quad y_f \text{ is not } \mathcal{U}\text{-reachable from } y_0 \text{ in time } T.$$

Proof. For $u \in E_\mathcal{U}$, there holds

$$\begin{aligned}\|y(T) - S_T y_0\| = \|L_T u\| = \left\|\int_0^T S_{T-t} B u(t,\cdot)\mathrm{d}t\right\| &\leq \int_0^T \|S_{T-t} B u(t,\cdot)\|\mathrm{d}t \\ &\leq \int_0^T \sup_{t\in[0,T]} \|S_t\|_{\mathcal{L}(X)} \|B u(t,\cdot)\|\mathrm{d}t \leq T \sup_{t\in[0,T]} \|S_t\|_{\mathcal{L}(X)} M_{B\mathcal{U}}.\end{aligned}$$

Therefore, if $\|y_f - S_T y_0\| > \sup_{t\in[0,T]} \|S_t\|_{\mathcal{L}(X)} T M_{B\mathcal{U}}$, there exists no $u \in E_\mathcal{U}$ such that $y_f = S_T y_0 + L_T u$, which concludes the proof. □

An immediate consequence of Lemma 5 is that whenever $y_f \neq y_0$, there exists a minimal time of reachability of y_f from y_0 under the constraints given by $\mathcal{U}$, i.e.

$$T^\star(y_0, y_f) := \inf\{T \geq 0,\ \exists u \in E_{\mathcal{U}},\ y_f = S_T y_0 + L_T u\}$$

satisfies $T^\star(y_0, y_f) \in (0, +\infty]$.

In general, the set on the right-hand side of the definition of $T^\star(y_0, y_f)$ is not necessarily an interval, but it happens to be when $y_0 = 0$ or $y_f = 0$ and if there are controls $u \in \mathcal{U}$ such that $Bu = 0$. This is the content of the following lemma, see [26, Proposition 7] for a proof.

Lemma 6. *Assume that $y_0 = 0$ or $y_f = 0$, and that $\mathcal{U} \cap \ker(B) \neq \emptyset$. Then for all $T < T^\star(y_0, y_f)$, y_f is not $\mathcal{U}$-reachable from y_0 in time T, and for all $T > T^\star(y_0, y_f)$, y_f is $\mathcal{U}$-reachable from y_0 in time T.*

The above Lemma allows us to derive certified lower bounds for minimal times of reachability $T^\star(y_0, y_f)$. Indeed, if we have proved that y_f is not $\mathcal{U}$-reachable in a given time T, then $T \leq T^\star(y_0, y_f)$.

It is well known that, under the assumptions considered here, y_f is reachable at $T = T^\star(y_0, y_f)$. For a proof of this fact, based on the direct method of the calculus of variations, we refer, for instance, to [24, Lemma 2.1] or [3].

1.3.3 Control of discretisation and round-off errors

The notion of dual certificates allows us to develop a framework in which computer-assisted proofs can be established. Working on the functional J, our method provides a **numerically certified** dual certificate p_f which ensures that the target is not reachable.

In order to implement numerical methods, we must discretise the functional J. Recall that for a given $p_f \in X$, the adjoint L_T^* acts as follows: $L_T^* p_f(t) = B^* S_{T-t}^* p_f = B^* p(t)$, where p solves the *adjoint equation*

$$\begin{cases} \dot{p}(t) + A^* p(t) = 0, \\ p(T) = p_f. \end{cases} \tag{5}$$

With this notation in place, we may write

$$\begin{aligned} J(p_f) &= \int_0^T \sigma_{B\mathcal{U}}(S_{T-t}^* p_f)\mathrm{d}t - \mathrm{Re}\langle y_f, p_f\rangle_X + \mathrm{Re}\langle y_0, S_T^* p_f\rangle_X \\ &= \int_0^T \sigma_{B\mathcal{U}}(p(t))\mathrm{d}t - \mathrm{Re}\langle y_f, p_f\rangle_X + \mathrm{Re}\langle y_0, p(0)\rangle_X. \end{aligned} \tag{6}$$

In order to evaluate $J(p_f)$, not only does one need to compute a time integral and space integrals (the inner products in X), but more importantly, it is required to discretise the adjoint equation (5) both in time and in space – see Section 2.2 for details. Hence, in practice we will only be able to evaluate a discretised function J_d, defined on a finite-dimensional subspace V_h.

The next step, detailed in Section 3.2, relies on optimisation algorithms to find, if it exists, a $p_{fh} \in V_h$ for which the discretised functional J_d is negative.

Assuming such a p_{fh} is found, it provides a dual certificate for the non-reachability of y_f provided that both discretisation and round-off errors are small enough to certify the sign of the original functional at p_f.

For the first type of error, we derive in Theorem 11 a fully explicit bound using functional analytic tools, while the second type is handled using interval arithmetic[1], thanks to the Matlab library

[1]Given two real intervals $[a, b]$ and $[c, d]$, interval arithmetic defines operations so that the result is again an interval containing all possible outcomes of the corresponding real operation. For example,

$$\begin{aligned} [a, b] + [c, d] &= \{x + y : x \in [a, b],\ y \in [c, d]\} = [a + c, b + d] \\ [a, b] \cdot [c, d] &= \{xy : x \in [a, b],\ y \in [c, d]\} \quad = [\min\{ac, ad, bc, bd\},\ \max\{ac, ad, bc, bd\}]. \end{aligned}$$

Hence, the result of any basic arithmetic operation between intervals is still an interval.

IntLab [38].

In the end, we hopefully obtain a proven error bound $e(p_{fh}, p_f) \geq 0$ such that

$$J(p_f) \leq J_{\mathrm{d}}(p_{fh}) + e(p_{fh}, p_f) < 0, \tag{7}$$

which establishes that p_f is indeed a dual certificate for the non $\mathcal{U}$-reachability of y_f (from y_0, in time T).

Using this methodology, one can prove the non-reachability of targets for a large array of parabolic systems, which we develop in Section 3. Let us now give an example of a certified result obtained with our method in the context of Example 2, applied in the setting of parameters given by (2); in particular, recall that we have $y_0 = 0$, $T = 1$.

Theorem 7. *Consider $y_f = x \mapsto 0.037 \sin(\pi x)$. Then y_f is not $\mathcal{U}$-reachable from y_0 in time $T = 1$. Going further we may prove that the minimal time $T^\star$ to reach y_f from y_0 under the constraints $\mathcal{U}$, satisfies $T^\star \geq 1.12$.*

Thanks to our approach, this result is certified thanks to a dual certificate $p_f \in L^2(0,1)$ for which we prove $J(p_f; T) \in [-0.0011919, -0.0000944] < 0$, for $T = 1.12$, and where p_f can be seen in Figure 2.

One of the interests of our approach is that the above result cannot trivially be recovered using the parabolic maximum principle, since the chosen y_f is below the final state reached using the maximum allowed control $u(t,x) = 1$.

Extensions and outlook on future works. Our approach relies on a few crucial hypotheses:

- convexity and weak compactness assumptions on the constraints and target set (1) and (3) to obtain the equivalence (4)
- continuity and coercivity of $-A^*$ (see assumption (8))
- on a classical hypothesis relative to the approximation properties satisfied by the family of discretisation spaces (see assumption ($\mathcal{V}_1$)).

Let us start by highlighting a few routine generalisations. First, since finding p_f such that $J(p_f) < 0$ remains a sufficient condition for non-reachability, the method can provide dual certificates for nonconvex target sets. Similarly, the boundedness assumptions of B and $\mathcal{U}$ can be combined into the weaker assumption of boundedness of $BE_{\mathcal{U}} := \{t \mapsto Bu(t), u \in E_{\mathcal{U}}\}$. One could then consider unbounded constraints or an unbounded yet admissible operator $B \in L(U, \mathcal{D}(A^*)')$, at the cost of more complex formulae for $\sigma_{BE_{\mathcal{U}}}$ – either with a closed formula or with precise approximations.

At the expense of some generality, it might be also possible to relax certain assumptions in order to handle specific examples that do not fall within the present framework. In particular, one could try weakening the continuity-coercivity hypotheses, which are key to control discretisation errors with constants that are small with respect to the final time. This could involve restricting the search for dual certificates in a subspace where $-A^*$ is continuous and coercive, at the cost of equivalence (4), or ignoring completely these hypotheses and try to manage constants growing exponentially with T, a major hindrance for computer-assisted proofs. This extension could for instance be done using different time-discretisation schemes (see Remark 10). This would in turn make it possible to accommodate other boundary conditions, or even other classes of operators such as hypocoercive ones.

Other perspectives include major changes to the partial differential equation: an extension to non-homogeneous non-autonomous linear systems of the form $\partial_t y(t) = A(t)y(t) + B(t)u(t) + v(t)$ would require the use of resolvents to obtain (less accurate) discretisation error bounds. A completely open problem is to tackle semilinear or fully nonlinear parabolic equations, since our approach fundamentally relies on the linearity of L_T to compute $\sigma_{L_T E_{\mathcal{U}}}$. One way to recycle our approach would be to come up with inclusions of the reachable set into that of an appropriately chosen linear PDE.

Outline of the article. In Sections 2.1 and 2.2, we provide details about the functional analytic framework of the method, as well as the different hypotheses we make both on the operator A and on the discretisation methods. Proposition 9 outlines the error bounds associated to discretising the adjoint equation (5). Section 2.3 contains the main theoretical result, that is, Theorem 11 which provides precise discretisation errors incurred when approximating the dual functional J mentioned in (7) by its discretised counterpart J_{d}. Finally, Section 2.4 summarises the whole method and hypotheses.

Section 3 is devoted to exploiting these theoretical results: in Section 3.1, we give precisions as to the discretisation and interpolation methods used thereafter. Section 3.2 is dedicated to good practices on how to find p_{fh} satisfying $J_{\mathrm{d}}(p_{fh}) < 0$. Finally, in Section 3.3, we apply the methods and provide (computer-assisted) proofs of non-reachability for several 1D heat-like equations with various sets of constraints $\mathcal{U}$, operators B and target sets $\mathcal{Y}_f$.

2 Key results and proofs

2.1 Functional analytic framework

Here, we introduce some of the needed terminology at the level of an abstract operator denoted by $\mathcal{A}$. Ultimately, if A is the operator underlying the control problem, $-A^*$ will play that role.

All along this paper, X will denote a complex Hilbert space endowed with inner product $\langle\cdot,\cdot\rangle_X$ and norm $\|\cdot\|_X$. $V \subset X$ will also be a Hilbert space, densely and continuously embedded into X. Both spaces will be equipped with the norms $\|\cdot\|_X$ (associated to the inner product $\langle\cdot,\cdot\rangle_X$) and $\|\cdot\|_V$ (associated to the inner product $\langle\cdot,\cdot\rangle_V$). For ease of readability, we will drop the subscript X when dealing with $\|\cdot\|_X$ and $\langle\cdot,\cdot\rangle_X$. Let us identify X and its dual X', with the associated Gelfand triple $V \subset X \subset V'$.

We are also given $\mathcal{A} \in \mathcal{L}(V, V')$, along with its domain

$$\mathcal{D}(\mathcal{A}) = \{x \in V, \mathcal{A}x \in X\}.$$

Assume that $\mathcal{A}$ satisfies, for $0 < a_0 \leq a_1$:

$$\forall (v,w) \in \mathcal{D}(\mathcal{A}) \times V, \quad \begin{cases} |\langle \mathcal{A}v, w\rangle| \leq a_1 \|v\|_V \|w\|_V \\ \mathrm{Re}(\langle \mathcal{A}v, v\rangle) \geq a_0 \|v\|_V^2. \end{cases} \tag{8}$$

A simple division then proves that $\mathcal{A}$ satisfies the sectoriality property

$$\forall v \in \mathcal{D}(\mathcal{A}), \quad \langle \mathcal{A}v, v\rangle \in \mathcal{S}_\alpha := \big\{z \in \mathbb{C},\ z = 0 \text{ or } |\arg z| \leq \alpha\big\}, \tag{9}$$

with $\alpha = \mathrm{Arccos}(\frac{a_0}{a_1})$ satisfying $0 \leq \alpha < \frac{\pi}{2}$. Furthermore, applying Lax-Milgram's theorem to $w(z\,\mathrm{Id} - \mathcal{A})$ for a well-chosen $w \in \mathbb{C}$ provides that

$$\forall z \notin \mathcal{S}_\alpha, \quad z\,\mathrm{Id} - \mathcal{A} : \mathcal{D}(\mathcal{A}) \to X \text{ is an isomorphism.} \tag{10}$$

The combination of (9) and (10) corresponds to $(\mathcal{A}, \mathcal{D}(\mathcal{A}))$ being a so-called *$m\alpha$-accretive operator*. We shall need to use functions of such operators: if r is a rational fraction, bounded on $\mathcal{S}_\alpha$, written in the form

$$r(z) = r(\infty) + \sum_{j \in I} \frac{r_j}{(\alpha_j - z)^{m_j}},$$

with I a finite set, $\alpha_j \notin \mathcal{S}_\alpha$, then we may define $r(\mathcal{A}) \in \mathcal{L}(X)$ by

$$r(\mathcal{A}) := r(\infty)\,\mathrm{Id} + \sum_{j \in I} r_j (\alpha_j\,\mathrm{Id} - z\mathcal{A})^{-m_j}.$$

This definition then straightforwardly extends to functions f that may be written as the uniform limit of such rational fractions in $\mathcal{S}_\alpha$. For this class of functions, we have the important following estimate.

Theorem 8 ([16]). *For any function $f : \mathcal{S}_\alpha \to \mathbb{C}$ that is the uniform limit of bounded rational fractions on $\mathcal{S}_\alpha$,*

$$\|f(\mathcal{A})\|_{\mathcal{L}(X)} \leq C_\alpha \sup_{z\in\mathcal{S}_\alpha} |f(z)|,$$

where $C_\alpha \leq 2 + \frac{2}{\sqrt{3}}$.

It is commonly known that the opposite of a mα-accretive operator generates a C_0 semigroup. Furthermore, the second inequality of (8) implies that $\mathcal{A}$ is an isomorphism from V to V' by the Lax-Milgram theorem, with inverse $\mathcal{A}^{-1} : V' \to V$. Finally, we may define the adjoint $\mathcal{A}^* \in \mathcal{L}(V, V')$ of $\mathcal{A}$.

2.2 Discretisation errors

We now introduce discretisation in space. Let V_h be a finite-dimensional subspace of V, of dimension denoted M_h, associated to a discretisation parameter $h > 0$.

Let $\mathcal{A}_h : V_h \to V_h$ be defined by

$$\forall v_h, w_h \in V_h, \quad \mathcal{A}_h v_h \in V_h \quad \text{and} \quad \langle \mathcal{A}_h v_h, w_h\rangle = \langle \mathcal{A} v_h, w_h\rangle. \tag{11}$$

We will be considering standard assumptions concerning the discretisation properties associated to V_h: there exists $C_0 > 0$ such that

$$\forall f \in X, \quad \inf_{v_h\in V_h} \|\mathcal{A}^{-1} f - v_h\|_V + \inf_{v_h\in V_h} \|(\mathcal{A}^*)^{-1} f - v_h\|_V \leq C_0 h \|f\|. \tag{$\mathcal{V}_1$}$$

Given $z_0 \in X$, we are interested in approximating the unique solution $z \in \mathcal{C}^1((0,\infty); \mathcal{D}(\mathcal{A})) \cap \mathcal{C}^0([0,\infty); X)$ to

$$\begin{cases} z'(t) = -\mathcal{A} z(t) \\ z(0) = z_0. \end{cases} \tag{12}$$

Considering Euler's implicit scheme for the corresponding discrete problem (through V_h), we let $N_T \in \mathbb{N}^*$, $\Delta t := \frac{T}{N_T}$ and for $z_{h,0} \in V_h$ consider

$$\forall n \in \{0, \dots, N_T\}, \quad z_{h,n} = (\mathrm{Id} + \Delta t \mathcal{A}_h)^{-n} z_{h,0}. \tag{13}$$

The associated error can be estimated is as follows.

Proposition 9. *Assume that the couple given by $(\mathcal{A}, \mathcal{D}(\mathcal{A}))$ and V_h satisfies (8)-($\mathcal{V}_1$), and let $z_0 \in \mathcal{D}(\mathcal{A})$, $z_{h,0} \in V_h$. Then, letting $z : [0,T] \to X$ be the solution to (12) and $z_{h,n}, n \in \{0, \dots, N_T\}$ be defined by (13), we have*

$$\forall n \in \{0, \dots, N_T\}, \quad \|z(t_n) - z_{h,n}\| \leq C_\alpha \|z_0 - z_{h,0}\| + \left(C_2 h^2 + C_3 \Delta t\right) \|\mathcal{A} z_0\|.$$

where

$$\begin{cases} C_2 = C_1\left(7 + 4\ln(2)\frac{a_1}{a_0} + C_\alpha\right), \\ C_3 = \frac{a_1}{a_0} C_\alpha, \end{cases} \qquad C_1 = \begin{cases} \frac{a_1^2 C_0^2}{a_0} & \text{in general,} \\ \frac{a_1^{3/2} C_0^2}{4 a_0^{1/2}} & \text{if } (\mathcal{A}, \mathcal{D}(\mathcal{A})) \text{ is self-adjoint.} \end{cases}$$

We do not claim any originality with respect to this type of estimate; our contribution here is to derive it with explicit and optimised constants, a critical step for our approach to succeed. Its proof is postponed to Appendix B.

Remark 10. *In this article, we have only considered Euler's implicit time-discretisation scheme to approximate (12). A more general class of schemes consists in using Euler's implicit scheme to discretise*

$$\begin{cases} w'(t) = \kappa w(t) - \mathcal{A} w(t) \\ w(0) = z_0, \end{cases} \tag{14}$$

and then notice that $z(t) = e^{-\kappa t} w(t)$. This would have two main advantages: firstly, if $\mathcal{A}$ is a_0-coercive and a_1-continuous, and if V is continuously embedded in X with constant c, then $\mathcal{A} - \kappa \operatorname{Id}$ is $(a_1 + c^2\kappa)$-continuous, and if $0 < \kappa < \frac{a_0}{c^2}$ then $\mathcal{A} - \kappa \operatorname{Id}$ is $(a_1 - c^2\kappa)$-coercive. Therefore one can apply Proposition 9 to (14) and obtain discretisation errors decreasing exponentially with time.

This idea would also help extending the method to non-coercive operators: if $\mathcal{A} - \kappa \operatorname{Id}$ is coercive for $\kappa < 0$, then the same trick allows one to use Proposition 9 and obtain explicit discretisation errors – admittedly, those increase exponentially in time and would therefore be of little use even on small timescales.

Notice that one could also consider spectral methods: in the rare case where one has explicit access to eigenvalues and eigenvectors of $\mathcal{A}$, all discretisation errors related to the approximation of (12) would be avoided.

2.3 Control problem

We are now given the control problem ($\mathcal{S}$), a family of finite-dimensional subspaces V_h (of dimension denoted M_h). The important assumptions for us will concern the unbounded operator $\mathcal{A} := -A^*$, with domain $\mathcal{D}(\mathcal{A}^*)$ and the finite dimensional subspaces V_h:

- $(\mathcal{A}, \mathcal{D}(\mathcal{A}))$ satisfies (8),
- $(\mathcal{A}, \mathcal{D}(\mathcal{A}))$ and the family V_h satisfy ($\mathcal{V}_1$).

As before, we introduce the corresponding notation $\mathcal{A}_h$, defined by the relation (11).

We recall the assumption (1) that $\mathcal{U}$ is closed, convex and bounded. We let $M_{B\mathcal{U}} > 0$ be defined by

$$M_{B\mathcal{U}} := \sup_{u \in \mathcal{U}} \|Bu\|_X.$$

Note that we always have $M_{B\mathcal{U}} \leq \|B\|_{\mathcal{L}(U,X)} M_{\mathcal{U}}$, where $M_{\mathcal{U}} = \sup_{u \in \mathcal{U}} \|u\|_U$.

Starting from the equation (6) (and with the change of variable $t = T - t$ within the integral), we have

$$\forall p_f \in X, \quad J(p_f) = \int_0^T \sigma_{B\mathcal{U}}(S_t^* p_f) \mathrm{d}t - \operatorname{Re}\langle y_f, p_f \rangle_X + \operatorname{Re}\langle y_0, S_T^* p_f \rangle_X.$$

We now define the fully discretised functional $J_{\Delta t,h}$ by, for all $p_{fh} \in V_h$,

$$J_{\Delta t,h}(p_{fh}) = \Delta t \sum_{n=1}^{N_T} \sigma_{B\mathcal{U}}((\operatorname{Id} + \Delta t \mathcal{A}_h)^{-n} p_{fh}) - \operatorname{Re}\langle y_f, p_{fh} \rangle + \operatorname{Re}\left\langle y_0, (\operatorname{Id} + \Delta t \mathcal{A}_h)^{-N_T} p_{fh} \right\rangle \tag{15}$$

Then, our main result in controlling discretisation errors for the functional of interest is as follows.

Theorem 11. *Assume that $(-A^*, \mathcal{D}(A^*))$ and the family V_h satisfy (8) as well as ($\mathcal{V}_1$). Then for all $p_f \in \mathcal{D}(A^*)$, $p_{fh} \in V_h$, we have*

$$\begin{aligned} |J(p_f) - J_{\Delta t,h}(p_{fh})| \leq \left(\tfrac{1}{2} M_{B\mathcal{U}} T \Delta t + (\|y_0\| + M_{B\mathcal{U}} T)\left(C_2 h^2 + C_3 \Delta t\right)\right) \|A^* p_f\| \\ + \left((\|y_0\| + M_{B\mathcal{U}} T) C_\alpha + \|y_f\|\right) \|p_f - p_{fh}\|. \end{aligned}$$

Proof. We decompose the error in the form

$$|J(p_f) - J_{\Delta t,h}(p_{fh})| \leq \underbrace{|J(p_f) - \tilde{J}(p_f)|}_{(I)} + \underbrace{|\tilde{J}(p_f) - J_{\Delta t,h}(p_{fh})|}_{(II)},$$

where

$$\forall p_f \in X, \quad \tilde{J}(p_f) := \Delta t \sum_{n=1}^{N_T} \sigma_{B\mathcal{U}}(S_{t_n}^* p_f) - \mathrm{Re}\langle y_f, p_f\rangle + \mathrm{Re}\langle y_0, S_T^* p_f\rangle.$$

Estimate for the term (I). Recall that for a K-Lipschitz continuous function $f : [0,T] \to \mathbb{R}$, we have the estimate

$$\left| \int_0^T f(t)\mathrm{d}t - \Delta t \sum_{k=1}^{N_T} f\big(k\Delta t\big) \right| \leq \frac{1}{2} KT\Delta t.$$

Our aim is now to apply the above estimate to $f : t \mapsto \sigma_{B\mathcal{U}}(S_t^* p_f)$. Since $B\mathcal{U}$ is bounded (by $M_{B\mathcal{U}} > 0$), $\sigma_{B\mathcal{U}}$ is $M_{B\mathcal{U}}$-Lipschitz continuous, hence we have

$$|f(t) - f(s)| \leq M_{B\mathcal{U}} \|S_t^* p_f - S_s^* p_f\|. \tag{16}$$

Since $p_f \in \mathcal{D}(A^*)$, $t \mapsto S_t^* p_f$ is of class $\mathcal{C}^1$ on $[0,T]$ of derivative $t \mapsto A^* S_t^* p_f$, hence

$$|f(t) - f(s)| \leq M_{B\mathcal{U}} \sup_{t\in[0,T]} \|A^* S_t^* p_f\| \ |t-s| \leq M_{B\mathcal{U}} \|A^* p_f\| \ |t-s|, \tag{17}$$

where we used that $A^* S_t^* p_f = S_t^* A^* p_f$ as well as the bound $\|S_t^*\|_{\mathcal{L}(X)} \leq 1$ for all $t \geq 0$ (see Theorem 27). Summing up, the above Lipschitz estimate entails

$$(I) = |J(p_f) - \tilde{J}(p_f)| \leq \frac{1}{2} \Delta t M_{B\mathcal{U}} T \|A^* p_f\|$$

Estimate for the term (II). First, we write

$$\begin{aligned} |\tilde{J}(p_f) - J_{\Delta t,h}(p_{fh})| \leq \Delta t \sum_{n=1}^{N_T} \big|\sigma_{B\mathcal{U}}(S_{t_n}^* p_f) - \sigma_{B\mathcal{U}}((\mathrm{Id} + \Delta t \mathcal{A}_h)^{-n} p_{fh})\big| \\ + \big|\mathrm{Re}\left\langle y_0, S_T^* p_f - (\mathrm{Id} + \Delta t \mathcal{A}_h)^{-N_T} p_{fh}\right\rangle\big| + \|y_f\| \|p_f - p_{fh}\|. \end{aligned}$$

Using Proposition 9 with $\mathcal{A} = -A^*$,

$$\begin{aligned} \left| \mathrm{Re}\left\langle y_0, S_T^* p_f - (\mathrm{Id} + \Delta t \mathcal{A}_h)^{-N_T} p_{fh}\right\rangle \right| &\leq \|y_0\| \|S_T^* p_f - (\mathrm{Id} + \Delta t \mathcal{A}_h)^{-N_T} p_{fh}\| \\ &\leq \|y_0\| \left(C_\alpha \|p_f - p_{fh}\| + \big(C_2 h^2 + C_3 \Delta t\big) \|A^* p_f\| \right). \end{aligned}$$

The error related to the sum reads, still using Proposition 9 with $\mathcal{A} = -A^*$,

$$\begin{aligned} \Delta t \sum_{n=1}^{N_T} \big|\sigma_{B\mathcal{U}}(S_{t_n}^* p_f) - \sigma_{B\mathcal{U}}((\mathrm{Id} + \Delta t \mathcal{A}_h)^{-n} p_{fh})\big| &\leq \Delta t M_{B\mathcal{U}} \sum_{n=1}^{N_T} \|S_{t_n}^* p_f - (\mathrm{Id} + \Delta t \mathcal{A}_h)^{-n} p_{fh}\| \\ &\leq M_{B\mathcal{U}} T \left(C_\alpha \|p_f - p_{fh}\| + \big(C_2 h^2 + C_3 \Delta t\big) \|A^* p_f\| \right). \end{aligned}$$

Combining all the above estimates, we arrive at the announced result.

□

Remark 12. *Following Remark 3, if the closed convex set $\mathcal{Y}_f$ is also bounded (say by $M_{\mathcal{Y}_f}$), then the updated discretisation error bound as in Theorem 11 reads*

$$\begin{aligned} |J(p_f;\mathcal{Y}_f) - J_{\Delta t,h}(p_{fh};\mathcal{Y}_f)| \leq \big(\tfrac{1}{2} M_{B\mathcal{U}} T \Delta t + (\|y_0\| + M_{B\mathcal{U}} T)\big(C_2 h^2 + C_3 \Delta t\big)\big) \|A^* p_f\| \\ + \big((\|y_0\| + M_{B\mathcal{U}} T) C_\alpha + M_{\mathcal{Y}_f}\big) \|p_f - p_{fh}\|. \end{aligned}$$

Remark 13. *Notice that time-discretisation constants may be suboptimal in both Proposition 9 and Theorem 11. Indeed, they have been computed under the only assumption that $\mathcal{A}$ is a $m\alpha$-accretive operator, and not accounting for its coercivity. This might induce an upper-bound of the form: for all $t \geq 0$, $\|S_t^*\| \leq \gamma e^{-\varepsilon t}$, with $\gamma \geq 1, \varepsilon > 0$, which would slightly sharpen the estimates of (16)-(17). As for the discretisation errors on the adjoint equation (12), the estimates can be traced back through Proposition 9 to (32), where the supremum of the implicit Euler scheme is taken on $\mathcal{S}_\alpha$. Instead, it could be taken on the numerical range, defined as*

$$W(\mathcal{A}) = \{\langle \mathcal{A}x, x\rangle, x \in \mathcal{D}(\mathcal{A}), \|x\|_X = 1\} \subset \mathbb{C}.$$

However, although the literature about this set is extensive (see for instance [16–18] and the references therein), forfeiting the sectorial simplification might increase the value of the equivalent of the constant C_α in Theorem 8.

Remark 14. *A critical choice when it comes to applying our method is that of V_h. In view of the estimate given by Theorem 11, there are two main approaches.*

- *The first one is to rely on simple and thoroughly studied discretisation subspaces, such as the finite element method. Numerical computations and estimates of constants are made easier, but we typically do not have $p_{fh} \in \mathcal{D}(A^*)$. Hence, we must interpolate p_{fh} in some way to get $p_f \in \mathcal{D}(A^*)$, which in turn requires to appropriately bound $\|p_f - p_{fh}\|$.*
- *The second one is to choose $V_h \subset \mathcal{D}(A^*)$ (for instance, splines) so that we may simply set $p_f = p_{fh}$, thus circumventing the interpolation problem altogether. The price to pay is that the computation of mass and stiffness matrices is much more complex. Also, a larger dimension for the subspaces V_h is required to get the same discretisation parameter h, which results in increased computational costs.*

The examples studied in Section 3.3 – where A is a Dirichlet Laplacian-based operator – will be made using the first approach, with $\mathbb{P}_1$ finite elements.

In practice, there will be a natural basis $(\psi_1, \dots, \psi_{M_h})$ for V_h, yielding a mapping

$$I_h : z \in \mathbb{R}^{M_h} \mapsto \sum_{i=1}^{M_h} z_i \psi_i \in V_h.$$

Then, the operator $\mathcal{A}_h$ is equivalent to the product of matrices $\mathcal{M}^{-1}\mathcal{K}$ in the basis $(\psi_1, \dots, \psi_{M_h})$, where for all $i, j \in \{1, \dots, M_h\}$,

$$\mathcal{M}_{i,j} = \langle \psi_i, \psi_j \rangle \quad \text{and} \quad \mathcal{K}_{i,j} = \langle \mathcal{A}\psi_j, \psi_i \rangle. \tag{18}$$

In other words, $I_h(\mathcal{M}^{-1}\mathcal{K}z) = \mathcal{A}_h I_h z$ for all $z \in \mathbb{R}^{M_h}$. In the finite element setting, $\mathcal{M}$ and $\mathcal{K}$ are the so-called mass and stiffness matrices, respectively.

As a result, the numerically implemented function, as a function of $z \in \mathbb{R}^{M_h}$, is $J_{\Delta t,h}(p_{fh}) = J_{\Delta t,h}(I_h z)$, where

$$J_{\Delta t,h}(I_h z) = \Delta t \sum_{n=1}^{N_T} \sigma_{B\mathcal{U}}(I_h(\mathrm{Id} + \Delta t \mathcal{M}^{-1}\mathcal{K})^{-n} z) - \mathrm{Re}\,\langle y_f, I_h z\rangle + \mathrm{Re}\,\langle y_0, I_h(\mathrm{Id} + \Delta t \mathcal{M}^{-1}\mathcal{K})^{-N_T} z\rangle\,.$$

In order to actually implement the above function, we will assume that $B, \sigma_{\mathcal{U}}$, y_0 and y_f are such that

- we may compute $\sigma_{B\mathcal{U}}(I_h z)$ explicitly as a function of $z \in \mathbb{R}^{M_h}$,
- we may compute $\mathrm{Re}\langle y_f, I_h z\rangle$ and $\mathrm{Re}\langle y_0, I_h z\rangle$ explicitly as a function of $z \in \mathbb{R}^{M_h}$, which by linearity is equivalent to having explicit access to $\mathrm{Re}\langle y_0, \psi_i\rangle$, $\mathrm{Re}\langle y_f, \psi_i\rangle$ for $i \in \{1, \dots, M_h\}$. In the case of a set $\mathcal{Y}_f$, we need to be able to compute $\sigma_{\mathcal{Y}_f}(I_h z)$ as a function of $z \in \mathbb{R}^{M_h}$.

2.4 Methodology

We here give the overall methodology underlying our method in full detail.

Control problem. The control problem of the form ($\mathcal{S}$) is defined by the state space X, the control space U, the operator $(A, \mathcal{D}(A))$, the bounded control operator $B \in \mathcal{L}(U, X)$, and the bounded convex and closed constraint set $\mathcal{U}$ containing 0.

We are given $y_0 \in X$, $y_f \in X$, and a final time $T > 0$, and we are interested in establishing that y_f is not $\mathcal{U}$-reachable from y_0 in time T.

We first need to check that $(-A^*, \mathcal{D}(A^*))$ (with $\mathcal{D}(A^*) \subset V$, V densely and continuously embedded in X) satisfies (8), where we must explicitly compute a_0, a_1.

Space discretisation. We then choose a family of approximation spaces $V_h \subset V$ such that the couple formed by $(-A^*, \mathcal{D}(A^*))$ and V_h should satisfy ($\mathcal{V}_1$), where we must explicitly compute C_0.

For a suitable basis $(\psi_1, \dots, \psi_{M_h})$ of V_h, we compute the mass and stiffness matrices associated to the discretisation method, given by (18).

Then, letting $I_h : z \mapsto \sum_{i=1}^{M_h} z_i \psi_i$, assuming that these can be computed explicitly, we compute $\sigma_{B\mathcal{U}}(I_h z)$, $\mathrm{Re}\langle y_f, I_h z\rangle$ and $\mathrm{Re}\langle y_0, I_h z\rangle$ explicitly as a function of $z \in \mathbb{R}^{M_h}$ – closed formulae are available for most classical control constraints and discretisation spaces.

Minimising the discrete functional. At this stage, we have access to the discretised functional $z \mapsto J_{\Delta t,h}(p_{fh}) = J_{\Delta t,h}(I_h z)$.

We perform a primal-dual algorithm (see Subsection 3.2 for more details) to try and minimise the functional $z \mapsto J_{\Delta t,h}(I_h z)$, in order to find $z \in \mathbb{R}^{M_h}$ such that $J_{\Delta t,h}(I_h z) < 0$.

Choosing a possible dual certificate. Assuming we have found $z \in \mathbb{R}^{M_h}$ such that $J_{\Delta t,h}(I_h z) < 0$, we then aim at applying Theorem 11. We hence need to come up with a possible choice of dual certificate $p_f \in \mathcal{D}(A^*)$ based on $p_{fh} := I_h z$.

In this case, we are faced with the following alternative:

- if $V_h \subset \mathcal{D}(A^*)$, then we may directly choose $p_f = p_{fh}$, in which case we directly apply the estimate of Theorem 11 since $\|p_f - p_{fh}\| = 0$,
- if the above inclusion does not hold, we interpolate into p_{fh} into some $p_f \in \mathcal{D}(A^*)$ by an appropriately chosen procedure. In this case, we need to bound $\|p_f - p_{fh}\|$ explicitly.

In both cases, we end up with a bound

$$|J(p_f) - J_{\Delta t,h}(p_{fh})| \leq C,$$

where C is known explicitly.

Checking that the proposed certificate is valid. Once p_f has been chosen, we compute the quantity $J_{\Delta t,h}(p_{fh}) + C$ by means of interval arithmetic. This computation leads to an interval, and if its upper bound is negative, so is $J(p_f)$, thereby concluding the proof that y_f is not $\mathcal{U}$-reachable from y_0 in time T.

3 Application to 1D parabolic problems

We now apply the general methodology outlined in Subsection 2.4 to several 1D parabolic equations and systems with the Dirichlet Laplacian as the main operator.

We shall demonstrate the versatility of the method with different examples, namely

- the 1D heat equation with Dirichlet boundary conditions and internal control, with two different types of constraints: symmetric L^2 constraints, then nonnegativity and L^∞ constraints,
- on a underactuated system of 1D coupled heat equations, where only one equation is internally controlled (with bilateral constraints).

3.1 Preliminaries on discretisation methods

To apply our method to the aforementioned examples, one need to choose appropriate discretisation spaces V_h, compatible with the hypotheses underlying Theorem 11. First, we need the couple formed by $(-A^*, \mathcal{D}(A^*))$ and V_h to satisfy $(\mathcal{V}_1)$.

Since we will be dealing with (variations around) the Dirichlet Laplacian, we will be using standard $\mathbb{P}_1$ finite elements. Estimates for these are readily available, albeit usually not with explicit and optimised constants.

For problems involving the Dirichlet Laplace operator, these will lead to functions $p_{fh} \notin \mathcal{D}(A^*)$. As mentioned in Remark 14, we will then employ interpolation to obtain $p_f \in \mathcal{D}(A^*)$ from p_{fh}. This procedure will be carried out using cubic splines; the rationale behind this choice is Lemma 16 below.

Discretisation using $\mathbb{P}_1$ finite elements

Let us gather the few main results needed regarding $\mathbb{P}_1$ finite elements – we refer for instance to [36, Section 3] for a more precise setup. First, in view of the fact that the main operator of interest will be the Dirichlet Laplacian, we shall recall the following standard estimates with explicit constants.

Proposition 15. *Let $N_h \geq 1$, $h = \frac{1}{N_h}$ and $x_i = ih$ for $i \in \{0, \dots, N_h\}$. Denoting by $(\psi_i)_{i\in\{1,\dots,N_h-1\}}$ the usual $\mathbb{P}_1$ finite element basis (with Dirichlet boundary conditions), we have for all $f \in H^2(0,1) \cap H_0^1(0,1)$,*

$$\left\| f - \sum_{i=1}^{N_h-1} f(x_i)\psi_i \right\|_{H_0^1} \leq \frac{h}{\sqrt{2}} \|f''\|_{L^2},$$

and

$$\left\| f - \sum_{i=1}^{N_h-1} f(x_i)\psi_i \right\|_{L^2} \leq \frac{h^2}{2\sqrt{2}} \|f''\|_{L^2}. \tag{19}$$

The proof is postponed to Appendix A.

Interpolation using cubic splines

A function p_{fh} obtained by the $\mathbb{P}_1$ finite element method satisfies $p_{fh} \in H_0^1(0,1)$, and hence is not in $\mathcal{D}(\Delta) = H^2(0,1) \cap H_0^1(0,1)$ in general (here, Δ stands for the Dirichlet Laplacian).

As already explained, it will be necessary to interpolate this function to build $p_f \in \mathcal{D}(\Delta)$ so as to make use of Theorem 11. Such an interpolation should be done while making the discretisation error negligible. For a given choice of discretisation parameters Δt, and h, this will be all the more likely that both terms $\|p_f - p_{fh}\|$ and $\|A^* p_f\|$ are small. If $A = \Delta$, then by the estimate (19), $\|p_f - p_{fh}\|_{L^2}$ can be controlled by $\|p_f''\|_{L^2}$.

All in all, a natural requirement is to interpolate p_{fh} into p_f in a way that makes $\|p_f''\|_{L^2}$ as small as possible. Using the following Lemma (see [22, Chapter V, Theorem (5-7)] for details), we are then led to using cubing splines.

Lemma 16. *Using the notations of Proposition 15, given a vector $(q_i)_{i\in\{0,\dots,N_h\}}$, the following optimisation problem*

$$\inf_{\substack{f\in H^2(0,1) \\ \forall i\in\{0,\dots,N_h\}, f(x_i)=q_i}} \|f''\|_{L^2}.$$

has a unique minimiser given by a cubic spline (that is, it is a cubic polynomial on each (x_i, x_{i+1}), $i \in \{0, \dots, N_h - 1\}$*). Furthermore, it is a* $C^2([0,1])$ *function satisfying* $f''(0) = f''(1) = 0$.

As a result, if $p_{fh} \in H_0^1(0,1)$ is a function associated to $\mathbb{P}_1$ finite elements, we choose the cubic spline associated to previous lemma, imposing $p_f(0) = p_f(1) = 0$ and $p_f(x_i) = p_{fh}(x_i)$ for all $i \in \{1, \dots, N_h - 1\}$. The resulting function is of classe $\mathcal{C}^2$ and vanishes at the boundary, hence it is in $\mathcal{D}(\Delta)$.

3.2 Finding p_{fh} satisfying $J_{\Delta t,h}(p_{fh}) < 0$

Given the discretised functional $J_{\Delta_t,h}$ defined by (15), a key step is to efficiently find $p_{fh} \in V_h$ at which it takes negative values, if it ever exists. This can be a computationally challenging step that must consequently be carried out with care. First, it is a crucial feature that this step can be done *completely independently* from certifying the negativity of $J(p_f)$: any method and acceleration available to minimise $J_{\Delta t,h}$ can (and should) therefore be used.

In particular, one should not minimise the functional within interval arithmetic, which considerably slows computations down. Instead, interval arithmetic should only be used once the pair (p_{fh}, p_f) has been determined to certify the value of $J_{\Delta t,h}(p_{fh})$ and of $J(p_f)$.

Another trick to help with the minimisation process is noticing that ultimately, $J_{\Delta t,h}$ is nothing more than a proxy to get to J. Intuitively, it follows that if for $h > 0$, V_h is well-enough crafted, an optimal $p_{fh} \in V_h$ should be 'close' to a good candidate $p_f \in \mathcal{D}(A^*)$. Along this line, it can greatly improve the efficiency of the minimisation process to first minimise $J_{\Delta t_1,h_1}$ to obtain a numerical p_{fh_1}, which can then be interpolated into $p_f \in \mathcal{D}(A^*)$ using the methods mentioned in Section 3.1, and discretised again into $p_{fh_2} \in V_{h_2}$ to evaluate the much more finely discretised $J_{\Delta t_2,h_2}$, with $0 < h_2 < h_1$ and $0 < \Delta t_2 < \Delta t_1$. For example, the dual certificate in Corollary 21 has been computed with $h_1 = 10^{-2}, \Delta t_1 = 5 \cdot 10^{-4}$, then reinterpolated so that the final result has been computed with discretisation parameters $h_2 = 1.6 \cdot 10^{-3}, \Delta t_2 = 9 \cdot 10^{-6}$, a mesh size that would have significantly slowed the minimisation stage.

Fenchel duality context. First, let us notice that one can address the question of finding $p_f \in X$ (in the non-discretised setting for now) by considering the optimisation problem

$$d := \inf_{p_f \in X} J(p_f) = \inf_{p_f \in X} \sigma_{E_\mathcal{U}}(L_T^* p_f) - \mathrm{Re}\langle y_f, p_f \rangle + \mathrm{Re}\langle y_0, S_T^* p_f \rangle.$$

It is critical to see that J is positively 1-homogeneous: as a result, if there exists p_f such that $J(p_f) < 0$, then $d = -\infty$. If not, then $d = 0$. In other words, we are faced with the following alternative: $d = 0$ if and only if y_f is $\mathcal{U}$-reachable from y_0 in time T, and $d = -\infty$ if and only if it is not.

Now, this optimisation problem is the Fenchel-dual to the following primal problem

$$\pi := \inf_{u \in L^2(0,T;U)} \delta_{E_\mathcal{U}}(u) + \delta_{\{y_f - S_T y_0\}}(L_T u),$$

where δ_C is the convex indicator of a set C (taking the value 0 in C, and $+\infty$ outside of C). By our previous arguments, we have $d = -\pi$; such a strong duality can also be proved by using the Fenchel-Rockafellar theorem. This duality structure calls for tailored minimisation algorithms: across this section, we shall use the Chambolle-Pock primal-dual algorithm [13].

Because the functional J is unbounded below in the case of non-reachability, we do not have convergence guarantees for this algorithm. Hence, some stopping criterion must be chosen in order to chose p_{fh}. In practice, two such criteria are used in the numerical experiments: a numerical check of whether $J_{\Delta t,h}$ seems negative enough, and another of whether the algorithm is still moving, in the sense that $\| \frac{p_{i+1}}{\|p_{i+1}\|} - \frac{p_i}{\|p_i\|} \|$ is small enough. The result of that 'minimisation' step will abusively be called 'minimiser of $J_{\Delta t,h}$'.

Regularisation. Recalling that the discretisation error ultimately depends on how smooth p_f is (more precisely, it is controlled by the quantity $\|A^*p_f\|$, see Theorem 11), one might consider a regularisation of the primal-dual problem. Notice that such a regularisation might provide convergence guarantees of minimisation algorithms as well. An interesting choice is to consider the following dual functional (and its discretisation):

$$\forall p_f \in \mathcal{D}(A^*), \quad J_\lambda(p_f) = J(p_f) + \frac{\lambda}{2}\|A^*p_f\|^2, \tag{20}$$

where $\lambda > 0$ can be thought of as a regularisation parameter. In that case, one can show using (8) and the continuous embedding of X into V' that we have the existence and uniqueness of a minimiser[2] of J_λ on $\mathcal{D}(A^*)$.

The major advantage of choosing such a regularisation term can be seen in Theorem 11. Indeed, the computed-assisted proof of non-reachability essentially boils down to

$$\exists (p_{fh}, p_f) \in V_h \times \mathcal{D}(A^*), \quad \frac{J_{\Delta t,h}(p_{fh})}{\|A^*p_f\|} < -C,$$

where $C > 0$ is a constant depending on several parameters, including Δt and h. This follows from an upper bound of $\|p_f - p_{fh}\|$ using ($\mathcal{H}_1$). In that context, it is natural to try and minimise $\|A^*p_f\|$ as well, hence the regularisation term above. However, since both $J_{\Delta t,h}$ and $p_f \mapsto \|A^*p_f\|$ are 1-homogeneous, the important component of p_f is not its norm, but its direction. One can thus wonder if it is possible to optimise the choice of λ: as it turns out, the following lemma proves that the choice of λ does not influence the direction of the minimiser.

Lemma 17. *Let H be a Hilbert space and $f, g : H \to \mathbb{R} \cup \{+\infty\}$ be two convex continuous functions such that*

$$\forall \alpha \geq 0,\ \forall x \in H, \quad \begin{cases} f(\alpha x) = \alpha f(x) \\ g(\alpha x) = \alpha^2 g(x). \end{cases}$$

Assume furthermore that g is positive outside of 0. Then, denoting $S_g = \{s \in H,\ g(s) = 1\}$:

$$\inf_{x \in H} f(x) + \frac{\lambda}{2} g(x) = -\frac{1}{2\lambda} \sup_{s \in S_g} \left(\min(f(s), 0)\right)^2,$$

and if the sup *is reached, then:*

$$\exists p \in H,\ \forall \lambda > 0,\ \exists r > 0, \quad f(rp) + \frac{\lambda}{2} g(rp) = \inf_{x \in H} f(x) + \frac{\lambda}{2} g(x).$$

The proof is postponed to Appendix A. Therefore the choice of $\lambda > 0$ has no influence over the direction of J_λ's 'minimiser'. Note however that choosing a 'reasonable value' of λ may have numerical advantages – increasing its value in turn increases the strong convexity of the functional and thus the rate of convergence of minimisation algorithms.

3.3 Examples of computer-assisted proofs

First, let us recall that Lemma 5 provides a basic estimate for the reachable set and hence a benchmark we can compare our approach to. The corresponding bound can easily be turned into a computer-assisted proof, but its interest for us will be to make sure that all the results we obtain are coherent with this estimate, and to prove 'new' results, namely results that were not already known by applying this basic rule.

In this section, all examples will be made using real initial states, targets, operators and controls, and thus so will be the dual certificates. We shall therefore drop the Re on the scalar products in this subsection.

[2] Indeed, one can show that J_λ is strongly convex and continuous on the Banach space $\mathcal{D}(A^*)$ endowed with the graph norm $\|\varphi\|_{\text{graph}} := \left(\|\varphi\|_X^2 + \|A^*\varphi\|_X^2\right)^{1/2}$. It thus follows that J_λ has a unique minimiser on $\mathcal{D}(A^*)$.

3.3.1 1D heat equation

We here consider the 1D heat equation with Dirichlet boundary conditions, namely

$$\begin{cases} \partial_t y(t,x) - \partial_{xx} y(t,x) = \chi_\omega u(t,x) & (t,x) \in (0,T] \times [0,1], \\ y(0,x) = y_0(x) & x \in [0,1], \\ y(t,0) = y(t,1) = 0 & t \in (0,T]. \end{cases} \tag{S}$$

The control problem is set with $X = L^2(0,1)$, $V = H_0^1(0,1)$, $A = \partial_{xx}$ is the Dirichlet Laplace operator with domain $\mathcal{D}(A) = H^2(0,1) \cap H_0^1(0,1)$, and the control operator is $B = \chi_\omega$, so that $U = L^2(0,1)$.

First, it is classical that $-A^*$ is self-adjoint, namely $-A^* = -A$ with domain $H^2(0,1) \cap H_0^1(0,1)$, and that it satisfies (8) with $a_0 = a_1 = 1$, and hence is mα accretive with $\alpha = 0$.

As mentioned in the previous section, we shall discretise $-A^*$ using $\mathbb{P}_1$ finite elements, in which case Proposition 15 yields $C_0 = \sqrt{2}$. We will now present the versatility of our method on different possible constraints. To avoid confusion, we will sometimes highlight the dependence of the functional J with respect to the target or the final time, by writing $J(p_f; y_f)$ or $J(p_f; T)$.

Toy example. The goal of this example is to consider a situation where we may compare our approach to known results obtained by basic calculations. To that end, we consider the system (S) with

$$y_0 = 0, \quad \omega = \Omega, \quad \mathcal{U} = \{u \in U,\ \|u\|_U \leq 1\}. \tag{21}$$

In this setting, note that $B = \mathrm{Id}$ and $M_{B\mathcal{U}} = 1$, and a simple calculation provides

$$\forall v \in X, \quad \sigma_{B\mathcal{U}}(v) = \|v\|_X.$$

Finally, we focus on the case where $y_f = \lambda \psi_k$, with $\lambda \in \mathbb{R}$ and $\psi_k := \sin(k\pi \cdot)$. In this simplified setting, we obtain an explicit characterization of non-reachability.

Lemma 18. *The following statements are equivalent:*

- $\lambda\psi_k$ *is* $\mathcal{U}$*-reachable from* 0 *in time* $T > 0$
- *There holds*

$$|\lambda| \leq \lambda_k^\star := \sqrt{2} M_{B\mathcal{U}} \frac{1 - e^{-k^2\pi^2 T}}{k^2\pi^2}.$$

Proof. We first prove the direct implication, assuming that $\lambda\psi_k$ is reachable, through some control u. Let us decompose $u(t,x)$ as follows: $u(t,x) = \alpha(t)\psi_k(x) + \beta(t)v(t,x)$, where the function $v(t,\cdot)$ satisfies the orthogonality condition: for a.e. $t \in (0,T)$, $\langle \psi_k, v(t,\cdot)\rangle = 0$. Then

$$\begin{aligned} \frac{\lambda}{2} &= \langle y(T), \psi_k \rangle = \langle L_T u, \psi_k \rangle = \langle u, L_T^* \psi_k \rangle = \int_0^T \langle u(t), S_{T-t}^* \psi_k \rangle \mathrm{d}t \\ &= \int_0^T \langle \alpha(t)\psi_k + \beta(t)v(t,\cdot), e^{-k^2\pi^2(T-t)}\psi_k \rangle \mathrm{d}t = \frac{1}{2}\int_0^T \alpha(t) e^{-k^2\pi^2(T-t)} \mathrm{d}t. \end{aligned}$$

Moreover, from the constraints (21), we have that for a.e. $t \in (0,T)$, $\|u(t)\| \leq 1$. Thus, for a.e. $t \in (0,T)$, $|\alpha(t)| \leq \sqrt{2} M_{B\mathcal{U}}$, which implies that

$$|\lambda| \leq \sqrt{2} M_{B\mathcal{U}} \int_0^T e^{-k^2\pi^2(T-t)} \mathrm{d}t = \sqrt{2} M_{B\mathcal{U}} \frac{1 - e^{-k^2\pi^2 T}}{k^2\pi^2} =: \lambda_k^\star.$$

We have thus shown that if $\lambda\psi_k$ is reachable, then $|\lambda| \leq \lambda_k^\star$. The converse implication follows easily by computing $L_T u$, for u defined by $u(t,x) := \frac{\lambda}{\lambda_k^\star} \sqrt{2} M_{B\mathcal{U}} \psi_k$. □

Building upon this analytic upper bound, we discuss the effectiveness of the method in Figure 1. In this table, we calculate certified upper bounds $\lambda_k^{\Delta t,h}$ of $\lambda_k^\star$ using the aforementioned method for different discretisation parameters, for $T = 1$ and $M = 1$. Those $\lambda_k^{\Delta t,h}$ are computed as a close upper-bound of

$$\inf\left\{\lambda \geq 0, J_{\Delta t,h}((\psi_k)_d; y_f = \lambda\psi_k) + e_r((\psi_k)_d) + e_r(\psi_k, \Delta t, h) < 0\right\}.$$

k	1	2	3	4	5	6	7	8
$\lambda_k^\star$	0.1433	0.0358	0.0159	0.009	0.0057	0.004	0.0029	0.0022
$(\Delta t, h) = (5\text{e–}4, 1\text{e–}2)$	0.1797	0.1810	0.3438	0.5953	0.9298	1.3486	1.8561	2.4584
$(\Delta t, h) = (1.25\text{e–}4, 5\text{e–}3)$	0.1525	0.0721	0.0975	0.1541	0.2328	0.3317	0.4502	0.5886
$(\Delta t, h) = (2\text{e–}5, 2\text{e–}3)$	0.1449	0.0418	0.0291	0.0322	0.0420	0.0562	0.0740	0.0951
$(\Delta t, h) = (5\text{e–}6, 1\text{e–}3)$	0.1437	0.0373	0.01920	0.0148	0.0148	0.0170	0.0207	0.0255

Figure 1: Comparison of theoretical and numerically certified upper-bounds of non-reachability of $\lambda\psi_k$

As we can see, the method is very effective for small frequencies, but its precision decreases quickly with higher frequencies. This can be easily understood since the error term in Theorem 11 depends on the second derivative of the interpolant, and thus increases with the square of the frequency. Here, no minimisation process was required to find an optimal dual certificate: indeed, one can show that in this context, if $y_f = \lambda\psi_k$, $p_f = \frac{y_f}{\|y_f\|}$ satisfies $J(p_f; y_f) < 0$ if and only if y_f is not reachable. Actually, one can easily show that $p_f = y_f = \psi_k$ (or any positive multiple thereof) is an optimal dual certificate of non-reachability. Let us emphasise that this is only the case because of such simple control constraints, and especially because $B = \mathrm{Id}$. This result is confirmed by numerical experiments: when minimising $J_{\Delta t,h}(\cdot; \psi_k)$, the algorithm stabilises around the direction given by ψ_k very quickly.

L^∞ constraints. Let us now consider a more involved example associated to the control problem (S), where we let

$$y_0 = 0, \quad \omega = (\tfrac{1}{5}, \tfrac{2}{5}) \cup (\tfrac{4}{5}, 1), \tag{22}$$

and

$$\mathcal{U} = \{u \in U,\ 0 \leq u(x) \leq 1 \text{ for a.e. } x \in [0,1]\}. \tag{23}$$

We also compute that $M_{B\mathcal{U}} = \sqrt{|\omega|} = \sqrt{\frac{2}{5}}$. We compute (letting $z_+ = \max(z, 0)$ for $z \in \mathbb{R}$)

$$\forall v \in X, \quad \sigma_{B\mathcal{U}}(v) = \sup_{u\in\mathcal{U}} \langle u, \chi_\omega v\rangle = \int_\omega \sup_{0\leq y\leq 1} yv(x)\mathrm{d}x = \int_\omega v_+(x)\mathrm{d}x.$$

In this context, aside from the dual method introduced in this article, two sufficient conditions can help determine whether a given target is non-reachable.

- First, the crude inclusion of the reachable set in a ball of centre $S_T y_0$ based on Lemma 5 shows that $\|y_f - S_T y_0\| \leq M_{B\mathcal{U}} T$ is a necessary condition for y_f to be $\mathcal{U}$-reachable from 0 in time T.
- Second, the parabolic comparison principle leads to $S_T y_0 \leq y(T) = S_T y_0 + L_T u \leq S_T y_0 + L_T \overline{u}$ for all controls taking values in $\mathcal{U}$, where $\overline{u}(t, x) = 1$ for a.e. $t \in (0, T)$, $x \in (0, 1)$. In the case where $y_0 = 0$, this yields a necessary condition for a target y_f to be $\mathcal{U}$-reachable from 0 in time $T > 0$, given by

$$0 \leq y_f \leq L_T \overline{u}. \tag{24}$$

For simple enough subsets ω of $[0,1]$, one can easily approximate $L_T\bar{u}$ with L^∞ estimates, thus allowing for proofs of non-reachability using the parabolic comparison principle. Therefore, to emphasise the usefulness of our method, all the computer-assisted results of non-reachability we will provide below satisfy (24), at least numerically.

That is why we will here focus on the case where $y_f = \lambda\psi_1$, with $\lambda > 0$, which is not – for λ small enough – trivially non-reachable using the parabolic comparison principle. We first present some results of non-reachability of half-lines: the following lemma proves that if $y_0 = 0$ (in which case 0 is reachable) then for any non-reachable targets y_f, so is the full half-line starting at y_f and moving away from 0.

Lemma 19. *If $0 \in \mathcal{U}$ and $y_0 = 0$, then for all y_f, p_f such that $J(p_f; y_f) < 0$, we have $J(p_f; \lambda y_f) < 0$ for all $\lambda \geq 1$.*

Proof. If $0 \in \mathcal{U}$, then $\forall v \in U$, $\sigma_{\mathcal{U}}(v) \geq 0$, and thus $J(p_f; y_f) < 0$. This implies $\langle y_f, p_f \rangle < 0$, which leads to $J(p_f; \lambda y_f) \leq J(p_f; y_f) < 0$. □

Its corollary below has been computed with a discretisation of $2,000,000$ points in time, and $2,000$ in space:

Corollary 20. *For the system (S) with operator B associated to ω given by (22) and constraints (23), the target $\lambda\psi_1$ is not $\mathcal{U}$-reachable for $T = 1$ if $\lambda \geq 0.035$. Indeed, there exists $p_f \in L^2(0,1)$ such that*

$$J(p_f; 0.035\psi_1) \in [-0.000718, -0.000111] < 0.$$

The dual certificate p_f associated to the previous result is shown in Figure 2.

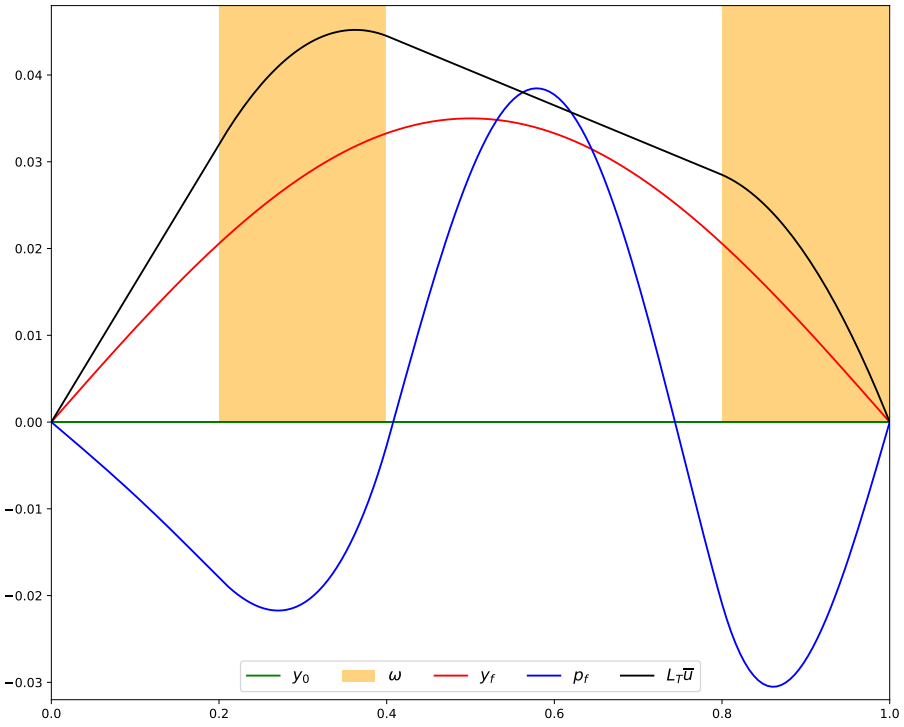

Figure 2: Target and optimal dual certificate associated to Corollary 20.

Notice that, unlike the results proved in (3.3.1), the dual certificate p_f is quite different from the target y_f (see Figure 2). This is mainly due to the presence of the operator $B = \chi_\omega$, which introduces

higher space heterogeneity. The optimisation step tends to create a dual certificate negative on ω, so as to reduce the value of $\sigma_{E_{\mathcal{U}}}(L_T^* p_f)$, while maximising $\langle y_f, p_f \rangle$, which translates in maximising the values of p_f on $[0,1]\backslash\omega$, since here $y_f = \psi_1 \geq 0$.

The result proved in Corollary 20 happens to be near the limit of what can be done with the computer-assisted proofs developed in this article, in the sense that the very fine discretisation allowed for very small discretisation errors (approximately $2.72 \cdot 10^{-4}$). However, a more precise discretisation comes at the cost of higher computational costs (here, several hours on a standard desk computer), and often with higher rounding errors: here, they amounted to $3.13 \cdot 10^{-5}$. In general, it is a general feature of our method that a compromise must be found to balance discretisation and rounding errors.

Another interesting application of this method is to compute certified lower-bounds of minimal times of reachability, which has been proved with a $500,000$-point discretisation in time, and 1000-point in space:

Corollary 21 (of Lemma 6)**.** *Let $y_f = \frac{1}{50}\psi_1$. With constraints (22) and (23), the minimal time $T^\star \in (0, +\infty]$ needed to steer (S) from $y_0 = 0$ to y_f satisfies $T^\star \geq T = 0.13$. Indeed, there exists $p_f \in L^2(0,1)$ for which*

$$J(p_f; T) \in [-0.000192, -0.0000268] < 0.$$

The dual certificate p_f associated to the previous result is identical to the one of corollary 20 shown in Figure 2.

The method can also be applied to less smooth targets, such as absolute value based functions (see Figure 3). However, minimising $J(\cdot; y_f)$ might provide less smooth dual certificates p_{fh}, which would lead to huge $\|A^* p_f\|$ after interpolation into p_f. Indeed, as stated before, the minimiser p_f will try to maximise the scalar product $\langle y_f, p_f \rangle$, and therefore mimic the regularity of y_f. To circumvent this, one might minimise the regularised functional introduced in (20).

For example, the minimisation of J and J_λ for y_f, leading to the non-reachibility result of Proposition 22 can be seen in Figure 3. One can calculate that $J(p_f) \simeq -0.0014$ and $J(p_f^{\mathrm{reg}}) \simeq -0.0013$, as expected since p_f^{reg} is obtained through the interpolation of a minimiser of J_λ with $\lambda = 10^{-12}$, but $\|A^* p_f\| \simeq 1350$ where $\|A^* p_f^{\mathrm{reg}}\| \simeq 125$. Overall, although p_f is a better minimiser of J, the massive difference of discretisation error simplifies the proof of non-reachability of y_f.

Considering a discretisation of $5,000,000$ points in time and $3,250$ points of space, we have obtained the following result. In particular, it is interesting because it has only H_0^1 regularity overall, and analytic regularity exactly outside of the control support: this is a pivot case where the reachability of the target is unknown for unconstrained controls (see [14]), let alone for the constrained case.

Proposition 22. *Consider the target*

$$y_f : x \mapsto \begin{cases} \frac{1}{40}\frac{5x}{4} & \text{if } 0 \leq x \leq \frac{4}{5} \\ \frac{1}{40}(5 - 5x) & \text{if } \frac{4}{5} \leq x \leq 1. \end{cases}$$

Then y_f is not reachable from $y_0 = 0$ in time $T = 1$. Indeed, with p_f^{reg} plotted in Figure 3, we have that

$$J(p_f^{\mathrm{reg}}) \in [-0.0017, -0.0009] < 0.$$

Once again, the limits of our computer-assisted method showed: such a precise discretisation induced high computations and rounding errors ($\simeq 2 \cdot 10^{-4}$), of the same level as discretisation errors ($\simeq 1.8 \cdot 10^{-4}$). It follows that a finer discretisation probably would likely not have allowed us to prove a stronger result.

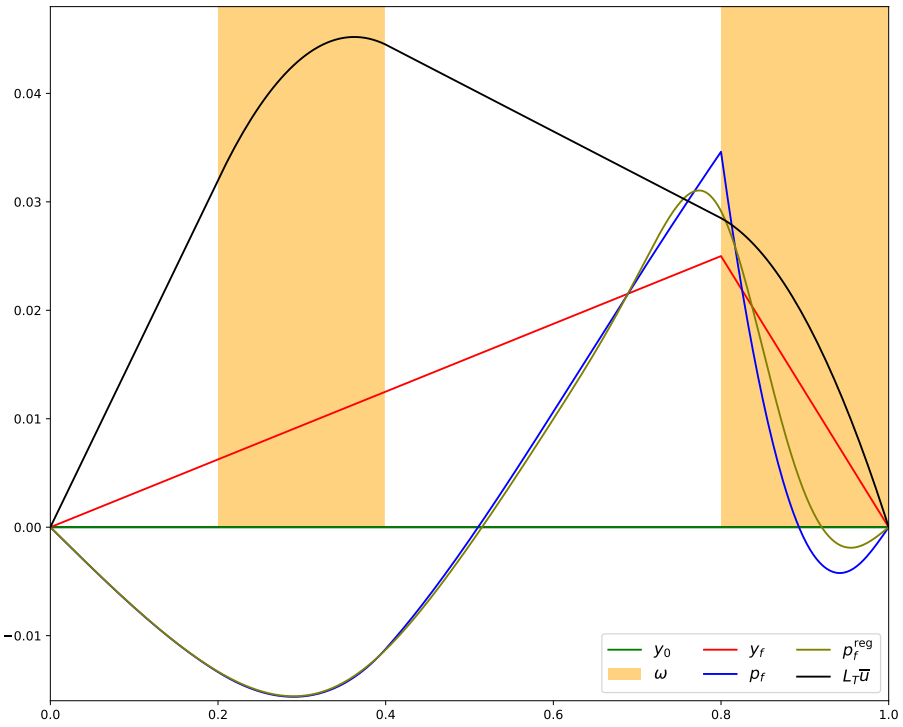

Figure 3: Target and optimal dual certificates for Proposition 22.

3.3.2 Coupled 1D heat equation

In this section we shall consider the following dynamical system

$$\begin{cases} \partial_t y_1 - \kappa_1 \partial_{xx} y_1 = a y_1 + b y_2 & (t,x) \in (0,T] \times [0,1] \\ \partial_t y_2 - \kappa_2 \partial_{xx} y_2 = c y_1 + d y_2 + \chi_\omega u & (t,x) \in (0,T] \times [0,1], \\ y_i(0,x) = y_{0,i}(x) & i \in \{1,2\},\ x \in [0,1], \\ y_i(t,0) = y_i(t,1) = 0 & i \in \{1,2\},\ t \in (0,T]. \end{cases}$$

Here, the control problem is set with $X = (L^2(0,1))^2$, $V = (H_0^1(0,1))^2$,

$$A := \begin{pmatrix} \kappa_1 \partial_{xx} + a & b \\ c & \kappa_2 \partial_{xx} + d \end{pmatrix}$$

with domain $\mathcal{D}(A) = (H^2(0,1) \cap H_0^1(0,1))^2$. Finally, $U = L^2(0,1)$ and the control operator is defined by

$$\forall u \in U, \quad Bu = \begin{pmatrix} 0 \\ \chi_\omega u \end{pmatrix} \qquad \text{and} \qquad \forall \begin{pmatrix} x \\ y \end{pmatrix} \in X, \quad B^* \begin{pmatrix} x \\ y \end{pmatrix} = \chi_\omega y.$$

We consider the case where

$$\omega = (0, \tfrac{1}{2}), \quad \mathcal{U} := \{u \in U,\ -1 \le u \le 2\}.$$

Hence, we have $M_{\mathcal{U}} = 2$, and $M_{B\mathcal{U}} = M_{\mathcal{U}} \sqrt{|\omega|} = \sqrt{2}$. With the notation $z_+ = \max(z,0)$, $z_- = \min(z,0)$,

$$\forall \begin{pmatrix} x \\ y \end{pmatrix} \in X, \quad \sigma_{B\mathcal{U}}\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \int_\omega (2y_+(x) - y_-(x)) \mathrm{d}x.$$

Here, we discretise by using $\mathbb{P}_1$ finite elements. Letting $W_h \subset H_0^1(0,1)$ be the discretisation spaces associated to $\mathbb{P}_1$ finite elements, this means we set $V_h := (W_h)^2 \subset V$.

Proposition 23. *Let $(a,b,c,d) \in \mathbb{R}^4$ and $\kappa_1, \kappa_2 > 0$. Let $\sigma_{\max}$ denote the largest singular value[3] of $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, and for $i \in \{1,2,3\}$, $\lambda_{\max}(S_i)$ denote the largest eigenvalue[4] of the matrix S_i defined by*

$$S_1 = \begin{pmatrix} a & \frac{1}{2}|b+c| \\ \frac{1}{2}|b+c| & d \end{pmatrix} \qquad S_2 = \begin{pmatrix} 2a\kappa_1 & |b\kappa_1 + c\kappa_2| \\ |b\kappa_1 + c\kappa_2| & 2d\kappa_2 \end{pmatrix} \qquad S_3 = \begin{pmatrix} 2a\kappa_1 & |c\kappa_1 + b\kappa_2| \\ |c\kappa_1 + b\kappa_2| & 2d\kappa_2 \end{pmatrix}.$$

Let us assume that the coefficients a, b, c, d, κ_1 and κ_2 are such that

$$\begin{aligned} \lambda_{\max}(S_1) \leq & \pi^2 \min(\kappa_1, \kappa_2) \\ \max\left(\lambda_{\max}(S_2), \lambda_{\max}(S_3)\right) \leq & 4\pi^2 \min(\kappa_1^2, \kappa_2^2). \end{aligned}$$

A possible choice of the continuity constant a_0 and the coercivity constant a_1 introduced in (8) are

$$\begin{aligned} a_0 &= \min(\kappa_1, \kappa_2) - \max\left(\frac{\lambda_{\max}(S_1)}{\pi^2}, 0\right) \\ a_1 &= \max(\kappa_1, \kappa_2) + \frac{\sigma_{\max}}{\pi^2}. \end{aligned}$$

If $a_0 > 0$, $(-A^*, \mathcal{D}(A^*))$ is hence $m\alpha$-accretive with $\alpha = \arccos(\frac{a_0}{a_1})$. Furthermore, $(\mathcal{V}_1)$ holds with

$$C_0 = \frac{1}{2\pi\sqrt{2}} \left(\sqrt{4\pi^2 \min(\kappa_1^2, \kappa_2^2) - \max\left(\lambda_{\max}(S_2), 0\right)} + \sqrt{4\pi^2 \min(\kappa_1^2, \kappa_2^2) - \max\left(\lambda_{\max}(S_3), 0\right)} \right).$$

Proof. Recall that here the state space is $X = (L^2(0,1))^2$ and $V = (H_0^1(0,1))^2$.

Computation of a_1. Using the Cauchy-Schwarz Poincaré inequalities[5], one can prove for all $(v,w) \in \mathcal{D}(A^*) \times V$

$$\begin{aligned} |\langle -A^* v, w \rangle| &= \left| \left\langle \partial_x \begin{pmatrix} \kappa_1 v_1 \\ \kappa_2 v_2 \end{pmatrix}, \partial_x \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \right\rangle - \left\langle \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \right\rangle \right| \\ &\leq \max(\kappa_1, \kappa_2) \|v\|_V \|w\|_V + \left| \left\langle \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \right\rangle \right| \\ &\leq \max(\kappa_1, \kappa_2) \|v\|_V \|w\|_V + \sigma_{\max} \|v\| \|w\| \\ &\leq \left(\max(\kappa_1, \kappa_2) + \frac{\sigma_{\max}}{\pi^2} \right) \|v\|_V \|w\|_V. \end{aligned}$$

[3] In other words,

$$\sigma_{\max} = \frac{1}{\sqrt{2}} \sqrt{a^2 + b^2 + c^2 + d^2 + \sqrt{(a^2 + b^2 - c^2 - d^2)^2 + 4(ac + bd)^2}}.$$

[4] In other words,

$$\begin{aligned} \lambda_{\max}(S_1) &= \frac{1}{2}(a + d + \sqrt{(a-d)^2 + (b+c)^2}), \\ \lambda_{\max}(S_2) &= a\kappa_1 + d\kappa_2 + \sqrt{(a\kappa_1 - d\kappa_2)^2 + (b\kappa_1 + c\kappa_2)^2}, \\ \lambda_{\max}(S_3) &= a\kappa_1 + d\kappa_2 + \sqrt{(a\kappa_1 - d\kappa_2)^2 + (c\kappa_1 + b\kappa_2)^2}. \end{aligned}$$

[5] Recall that for every $v \in H_0^1(0,1)$, there holds

$$\pi^2 \|v\|^2_{L^2(0,1)} \leq \|v'\|^2_{L^2(0,1)}$$

and the constant π on the left-hand side is sharp.

Computation of a_0. For $v \in \mathcal{D}(A^*)$, we compute

$$\begin{aligned}\mathrm{Re}(\langle -A^*v, v\rangle) &= \kappa_1\|\partial_x v_1\|_{L^2}^2 + \kappa_2\|\partial_x v_2\|_{L^2}^2 - \mathrm{Re}\left\langle \begin{pmatrix} a & c \\ b & d\end{pmatrix}\begin{pmatrix} v_1 \\ v_2\end{pmatrix}, \begin{pmatrix} v_1 \\ v_2\end{pmatrix}\right\rangle \\ &\geq \kappa_1\|\partial_x v_1\|_{L^2}^2 + \kappa_2\|\partial_x v_2\|_{L^2}^2 - \left\langle S_1\begin{pmatrix}\|v_1\| \\ \|v_2\|\end{pmatrix}, \begin{pmatrix}\|v_1\| \\ \|v_2\|\end{pmatrix}\right\rangle \\ &\geq \kappa_1\|\partial_x v_1\|^2 + \kappa_2\|\partial_x v_2\|^2 - \lambda_{\max}(S_1)\|v\|^2 \\ &\geq \left(\min(\kappa_1,\kappa_2) - \max\left(\frac{\lambda_{\max}(S_1)}{\pi^2}, 0\right)\right)\|v\|_V^2.\end{aligned}$$

Computation of C_0. First recall that Proposition 15 gives us,

$$\forall g \in H^2(0,1), \quad \inf_{v_h\in V_h}\|g - v_h\|_{H_0^1} \leq \frac{h}{\sqrt{2}}\|g''\|_{L^2},$$

which translates into, since $\forall f \in \mathcal{D}(A)$, $A^{-1}f \in V$ and $(A^*)^{-1}f \in \mathcal{D}(A^*)$

$$\forall f \in X, \quad \inf_{v_h\in V_h}\|A^{-1}f - v_h\|_V + \inf_{v_h\in V_h}\|(A^*)^{-1}f - v_h\|_V \leq \frac{h}{\sqrt{2}}\left(\|\partial_{xx}(A^{-1}f)\| + \|\partial_{xx}((A^*)^{-1}f)\|\right). \quad (25)$$

It follows that it is enough to show

$$\forall f \in X, \quad \left(\|\partial_{xx}(A^{-1}f)\| + \|\partial_{xx}((A^*)^{-1}f)\|\right) \leq C_0\sqrt{2}\|f\|.$$

To this aim, by setting $g = A^{-1}f$ (resp. $g = (A^{-1})^*f$) we will prove that

$$\forall g \in \mathcal{D}(A), \quad \min(\|Ag\|, \|A^*g\|) \geq \frac{C_0}{\sqrt{2}}\|\partial_{xx}g\|.$$

Let $g \in \mathcal{D}(A)$. We write

$$\begin{aligned}\|Ag\|^2 &= \kappa_1^2\|\partial_{xx}g_1\|_{L^2}^2 + \kappa_2^2\|\partial_{xx}g_2\|_{L^2}^2 + \|ag_1 + cg_2\|_{L^2}^2 + \|bg_1 + dg_2\|_{L^2}^2 \\ &\quad - 2a\kappa_1\|\partial_x g_1\|_{L^2}^2 - 2d\kappa_2\|\partial_x g_2\|_{L^2}^2 - 2(b\kappa_1 + c\kappa_2)\,\mathrm{Re}\langle\partial_x g_1, \partial_x g_2\rangle \\ &\geq \kappa_1^2\|\partial_{xx}g_1\|_{L^2}^2 + \kappa_2^2\|\partial_{xx}g_2\|_{L^2}^2 - 2a\kappa_1\|\partial_x g_1\|_{L^2}^2 - 2d\kappa_2\|\partial_x g_2\|_{L^2}^2 - 2|b\kappa_1 + c\kappa_2||\langle\partial_x g_1, \partial_x g_2\rangle| \\ &\geq \left(\kappa_1^2\|\partial_{xx}g_1\|_{L^2}^2 + \kappa_2^2\|\partial_{xx}g_2\|_{L^2}^2\right) - \left\langle\begin{pmatrix} 2a\kappa_1 & |b\kappa_1 + c\kappa_2| \\ |b\kappa_1 + c\kappa_2| & 2d\kappa_2\end{pmatrix}\begin{pmatrix}\|\partial_x g_1\|_{L^2} \\ \|\partial_x g_2\|_{L^2}\end{pmatrix}, \begin{pmatrix}\|\partial_x g_1\|_{L^2} \\ \|\partial_x g_2\|_{L^2}\end{pmatrix}\right\rangle \\ &\geq \left(\kappa_1^2\|\partial_{xx}g_1\|_{L^2}^2 + \kappa_2^2\|\partial_{xx}g_2\|_{L^2}^2\right) - \lambda_{\max}(S_2)\left(\|\partial_x g_1\|_{L^2}^2 + \|\partial_x g_2\|_{L^2}^2\right).\end{aligned}$$

Since $g_1, g_2 \in H_0^1(0,1)$ both are continuous and satisfy $\int_0^1 \partial_x g_i(t)\mathrm{d}t = g_i(1) - g_i(0) = 0$, using the Poincaré-Wirtinger inequality, we have $\|\partial_x g_i\|_{L^2} \leq \frac{1}{2\pi}\|\partial_{xx}g_i\|_{L^2}$, and thus

$$\|Ag\|^2 \geq \left(\kappa_1^2\|\partial_{xx}g_1\|_{L^2}^2 + \kappa_2^2\|\partial_{xx}g_2\|_{L^2}^2\right) - \frac{1}{4\pi^2}\max\left(\lambda_{\max}(S_2), 0\right)\left(\|\partial_{xx}g_1\|_{L^2}^2 + \|\partial_{xx}g_2\|_{L^2}^2\right).$$

It follows that

$$\|Ag\| \geq \sqrt{\min(\kappa_1^2,\kappa_2^2) - \frac{1}{4\pi^2}\max\left(\lambda_{\max}(S_2), 0\right)}\|\partial_{xx}g\|. \quad (26)$$

In a similar manner, one can show that

$$\|A^*g\| \geq \sqrt{\min(\kappa_1^2,\kappa_2^2) - \frac{1}{4\pi^2}\max\left(\lambda_{\max}(S_3), 0\right)}\|\partial_{xx}g\|. \quad (27)$$

Applying (26) with $g = A^{-1}f$ and (27) with $g = (A^*)^{-1}f$ to (25) yields ($\mathcal{V}_1$) with

$$C_0 = \frac{1}{2\pi\sqrt{2}}\left(\sqrt{4\pi^2\min(\kappa_1^2,\kappa_2^2) - \max\left(\lambda_{\max}(S_2), 0\right)} + \sqrt{4\pi^2\min(\kappa_1^2,\kappa_2^2) - \max\left(\lambda_{\max}(S_3), 0\right)}\right).$$

□

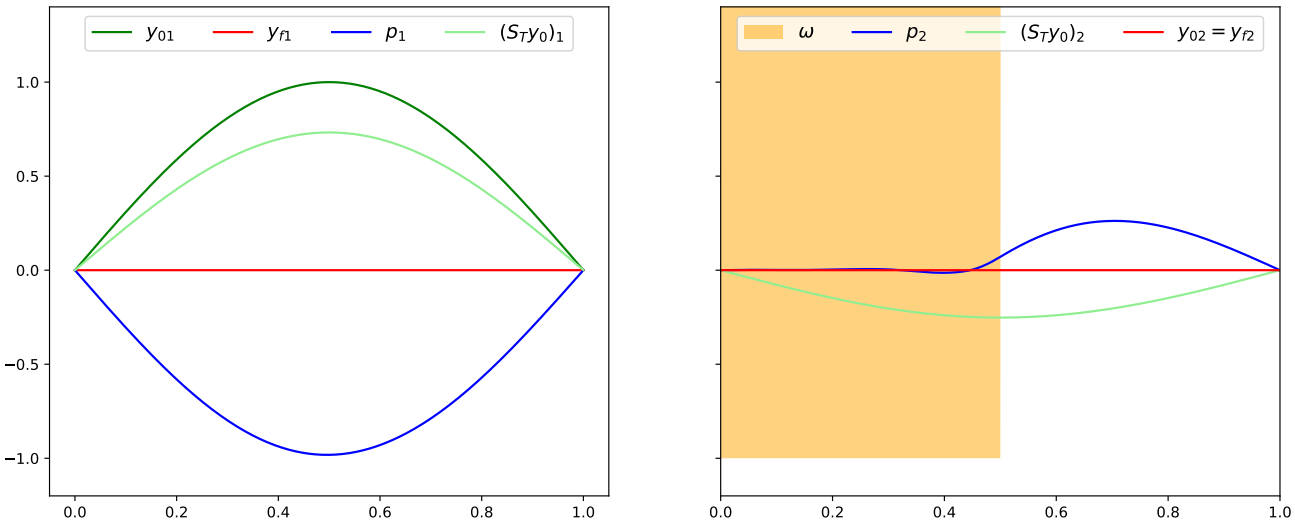

Figure 4: Initial and final state without control, target, control domain and dual certificate for Proposition 24.

For the following examples, we will use the constants

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} -1 & 2 \\ -2 & -1 \end{pmatrix}, \qquad \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1.5 \end{pmatrix}.$$

In particular, these coefficients satisfy the hypotheses of Proposition 23 and we have $C_0 = \sqrt{2}$, $a_0 = 1$, $a_1 = 1.5 + \frac{\sqrt{5}}{\pi^2}$. Notice that in this context, because of the competition term induced by $c < 0 < b$, no comparison principle is applicable. Therefore, aside from our methodology, only Lemma 5 provides reachability estimates in this setting.

Non-reachability of a ball: in the spirit of Remark 3, we shall here try to prove the non-reachability of sets of targets. For example, consider, for $y_f \in X$, $\varepsilon > 0$, the set

$$\mathcal{Y}_f = \{y \in X, \|y - y_f\| \leq \varepsilon\}.$$

One can then apply Remark 3 with $M_{\mathcal{Y}_f} = \|y_f\| + \varepsilon$ and use the same method to try and prove the non-reachability of all elements of $\mathcal{Y}_f$ under the constraints 3.3.2. In this context, one has that

$$\forall p_f \in X, \quad \sigma_{\mathcal{Y}_f}(p_f) = \langle y_f, p_f \rangle + \varepsilon \|p_f\|.$$

The following result was computed using a time-discretisation with $11,300$ points and a space discretisation with 750 points.

Proposition 24. *Let*

$$y_0 = \begin{pmatrix} \sin(\pi \cdot) \\ 0 \end{pmatrix}, \quad y_f = 0 \quad \textit{and} \quad \varepsilon = 10^{-2}.$$

The set $\mathcal{Y}_f := \{y \in X, \|y\| \leq \varepsilon\}$ is not reachable in time $T = 0.32$. Indeed, for $p_f = \binom{p_1}{p_2}$, whose graphs are pictured in Fig. 4, we have that

$$J(p_f; \mathcal{Y}_f, T) \in [-0.0262, -0.0014] < 0.$$

Notice that once more, the dual certificate $\binom{p_1}{p_2}$ tends to resemble $\mathcal{Y}_f + \{-S_T y_0\}$ so as to minimise $\sigma_{\mathcal{Y}_f}(-p_f) + \langle S_T y_0, p_f \rangle$, while staying close to 0 on ω, in order to minimise $\int_0^T \sigma_{B\mathcal{U}}(S_t^* p_f) \mathrm{d}t$. Also note

that discretisation parameters to prove this result are rather large: yet, discretisation errors (here, $8.1 \cdot 10^{-3}$) and round-off errors ($4.2 \cdot 10^{-3}$) are of the same order of magnitude. This is mainly due to the additional term $S_T^* p_f$ which considerably increases discretisation errors, and to the fact that space discretisation has doubled complexity with respect to the single equation case: two intervals $[0,1]$ are discretised with 750 points, thereby increasing rounding errors.

Remark 25. *In the spirit of Lemma 6, since* $\ker(B) \cap \mathcal{U} \neq \emptyset$ *and* $\|S_t y\| \leq \|y\|$ *for all* $y \in X$ *and* $t > 0$*, one can show that such a time* T *is a certified lower-bound of the minimal time* $T^\star(y_0, \mathcal{Y}_f)$ *needed to reach* $\mathcal{Y}_f$ *from* y_0*, where "minimal" means that for any time* $t > T^\star(y_0, \mathcal{Y}_f)$*, the target set* $\mathcal{Y}_f$ *is* $\mathcal{U}$*-reachable from* y_0 *in time* t*.*

Non-reachability of an unbounded set $\mathcal{Y}_f$. Based on Remark 3, one can also tackle the non-reachability of an unbounded set. As an example, consider the case where we want to make sure that, at time T, the first equation is not equal to some $y_1 \in L^2(0,1)$, without any restriction when it comes to the second equation. In other words, we set

$$\mathcal{Y}_f := \{y_1\} \times L^2(0,1).$$

In this context, we find

$$\forall p_f = \begin{pmatrix} p_f^1 \\ p_f^2 \end{pmatrix} \in X, \quad \sigma_{\mathcal{Y}_f}(p_f) = \begin{cases} \langle y_1, p_f^1 \rangle & \text{if } p_f^2 = 0 \\ +\infty & \text{if } p_f^2 \neq 0. \end{cases}$$

Since J now takes infinite values, we shall consider its restriction to $X_1 := L^2(0,1) \times \{0\}$ on which it takes finite values. Denoting $y_f = \left(\begin{smallmatrix} y_1 \\ 0 \end{smallmatrix}\right)$ we still have

$$\forall p_f \in X_1, \quad J(p_f; \mathcal{Y}_f) = \int_0^T \sigma_{\mathcal{U}}(L_T^* p_f(t)) \mathrm{d}t - \langle y_f, p_f \rangle,$$

for which the estimates of Theorem 11 can be applied. Using this result, one can prove the following proposition, computed with $80,000$ points in time, and 1000 points in space:

Proposition 26. *Let*

$$y_1 = -\frac{1}{50} \sin(2\pi \cdot) \quad \textit{and} \quad \mathcal{Y}_f := \{y_1\} \times L^2(0,1).$$

$\mathcal{Y}_f$ *is not reachable in time* $T = 1$*. Indeed, for* $p_f = \left(\begin{smallmatrix} y_1 \\ 0 \end{smallmatrix}\right)$*, we have that*

$$J(p_f; \mathcal{Y}_f) \in [-0.0262, -0.0008] < 0.$$

Here, we have computed the functional for $p_f = \left(\begin{smallmatrix} y_1 \\ 0 \end{smallmatrix}\right)$, which might not seem optimal. However, notice that this dual certificate does minimise $\sigma_{\mathcal{Y}_f}(-p_f)$ at fixed norm, and has low value of $\int_0^T \sigma_{B\mathcal{U}}(S_t^* p_f) \mathrm{d}t$. The 'true minimiser' of $J(\cdot, \mathcal{Y}_f)$ is very close to p_f, with p_1 very slightly modified so as to account for the time-integral term.

Acknowledgements. All four authors acknowledge the support of the ANR project TRECOS, grant number ANR-20-CE40-0009. We express our sincere gratitude to Michel Crouzeix for the time he devoted to us and for his valuable insights on the use of sectorial operators in numerical analysis.

Appendix

A Proof of minor results

Proof of Proposition 15. Let $p = \sum_{i=1}^{N-1} f(x_i)\psi_i$, and let $i \in \{0, \dots, N-1\}$. Remark that $p'(x) = \frac{f(x_{i+1})-f(x_i)}{x_{i+1}-x_i}$ for all $x \in (x_i, x_{i+1})$. Since f is in $H^2(0,1)$, it is in $C^1([0,1])$ and the mean value theorem then guarantees the existence of $\theta_i \in (x_i, x_{i+1})$ such that $f'(\theta_i) = \frac{f(x_{i+1})-f(x_i)}{x_{i+1}-x_i} = p'(x)$ for all $x \in (x_i, x_{i+1})$. The Cauchy-Schwarz inequality then entails

$$\begin{aligned}
\int_{x_i}^{x_{i+1}} |f'(x) - p'(x)|^2 \mathrm{d}x = \int_{x_i}^{x_{i+1}} |f'(x) - f'(\theta_i)|^2 \mathrm{d}x &= \int_{x_i}^{x_{i+1}} \left| \int_{\theta_i}^{x} f''(t) \mathrm{d}t \right|^2 \mathrm{d}x \\
&\leq \int_{x_i}^{x_{i+1}} |\theta_i - x| \int_{\theta_i}^{x} |f''(t)|^2 \mathrm{d}t \mathrm{d}x \\
&\leq \frac{1}{2} \left((x_{i+1} - \theta_i)^2 + (\theta_i - x_i)^2 \right) \int_{x_i}^{x_{i+1}} |f''(t)|^2 \mathrm{d}t \\
&\leq \frac{1}{2} (x_{i+1} - x_i)^2 \int_{x_i}^{x_{i+1}} |f''(t)|^2 \mathrm{d}t.
\end{aligned}$$

Summing over i achieves the first inequality. The second one is a direct consequence of the first, using the Cauchy-Schwarz inequality on two subintervals to reduce the constant:

$$\int_{x_i}^{x_{i+1}} |f(x) - p(x)|^2 \mathrm{d}x = \int_{x_i}^{\frac{1}{2}(x_i+x_{i+1})} |f(x) - p(x)|^2 \mathrm{d}x + \int_{\frac{1}{2}(x_i+x_{i+1})}^{x_{i+1}} |f(x) - p(x)|^2 \mathrm{d}x,$$

where

$$\begin{aligned}
\int_{x_i}^{\frac{1}{2}(x_i+x_{i+1})} |f(x) - p(x)|^2 \mathrm{d}x &= \int_{x_i}^{\frac{1}{2}(x_i+x_{i+1})} \left| \int_{x_i}^{x} (f'(t) - p'(t)) \mathrm{d}t \right|^2 \mathrm{d}x \\
&\leq \int_{x_i}^{\frac{1}{2}(x_i+x_{i+1})} |x - x_i| \int_{x_i}^{x} |f'(t) - p'(t)|^2 \mathrm{d}t \mathrm{d}x \\
&\leq \frac{1}{8} (x_{i+1} - x_i)^2 \int_{x_i}^{x_{i+1}} |f'(t) - p'(t)|^2 \mathrm{d}t \\
&\leq \frac{(x_{i+1} - x_i)^4}{16} \int_{x_i}^{x_{i+1}} |f''(t)|^2 \mathrm{d}t.
\end{aligned}$$

Similarly, one can prove the same upper-bound for the integral on the second subinterval, which leads to

$$\int_{x_i}^{x_{i+1}} |f(x) - p(x)|^2 \mathrm{d}x \leq \frac{(x_{i+1} - x_i)^4}{8} \int_{x_i}^{x_{i+1}} |f''(t)|^2 \mathrm{d}t.$$

Summing over i and taking the square root, we arrive at the second inequality. □

Proof of Lemma 17. Since g is positive outside of 0, for all $x \in H$ such that $g(x) < +\infty$, there exists

$r > 0, s \in S_g$ such that $x = rs$. Letting $\lambda > 0$, we may write

$$\begin{aligned}
\inf_{x\in H} f(x) + \frac{\lambda}{2} g(x) &= \inf_{\substack{x\in H \\ g(x)<+\infty}} f(x) + \frac{\lambda}{2} g(x) \\
&= \inf_{\substack{r>0 \\ s\in S_g}} rf(s) + \frac{\lambda}{2} r^2 \\
&= \inf_{s\in S_g} \inf_{r>0} rf(s) + \frac{\lambda}{2} r^2 \\
&= \inf_{s\in S_g} \frac{-1}{2\lambda} \big(\min(f(s), 0)\big)^2 \\
&= -\frac{1}{2\lambda} \sup_{s\in S_g} \left(\min(f(s), 0)\right)^2 .
\end{aligned}$$

Therefore, should the infimum be reached, the minimiser's direction is independent of λ, which concludes the proof. □

B Proof of the main results

The content of this section is drawn almost entirely from personal correspondence with Michel Crouzeix, to whom we are grateful for generously providing his personal notes. Our main contribution is to compute and optimise as much as possible all constants involved in the process.

Throughout this section, we assume that we are in the setting of Section 2, with an operator $(\mathcal{A}, \mathcal{D}(\mathcal{A}))$ and discretisation subspaces V_h satisfying (8) and ($\mathcal{V}_1$).

Let α and β satisfy $0 \leq \alpha < \alpha + \beta < \frac{\pi}{2}$. Recall the sectors defined by equations (9) and (10),and thus that $\mathcal{A}$ is a mα-accretive operator. A classical consequence of the Lumer-Phillips theorem is that $-\mathcal{A}$ generates of a semigroup of contraction, and thus for all $t \in \mathcal{S}_\beta$, the function $S_t := \exp(-t\mathcal{A})$ is well defined.

Before considering the discretisation of (5), we will first prove some theoretical results regarding its solution.

B.1 Theoretical results on solutions of the adjoint equation

Theorem 27. *For all $t \in \mathcal{S}_\beta$, with $t \neq 0$ and $0 \leq \alpha < \alpha + \beta < \frac{\pi}{2}$, and for every integer $k \geq 0$, the operator $S_t \in \mathcal{L}(X, \mathcal{D}(\mathcal{A}^k))$. Moreover, the map $t \mapsto S_t$, from $\mathcal{S}_{\frac{\pi}{2}-\alpha}$ into $\mathcal{L}(X)$, is holomorphic on the interior of $\mathcal{S}_{\frac{\pi}{2}-\alpha}$, and we have the estimate:*

$$\forall t > 0, \quad \|S_t^{(k)}\|_{\mathcal{L}(X)} = \|\mathcal{A}^k S_t\|_{\mathcal{L}(X)} \leq \frac{k!}{t^k \cos^k \alpha}.$$

Proof. (a) Let $f_k(z) = (1+z)^k e^{-tz}$. Since f_k is the uniform limit in $\mathcal{S}_\alpha$ of $(1+z)^k(1+tz/n)^{-n}$, we can define $f_k(\mathcal{A}) \in \mathcal{L}(X)$. Since $e^{-tz} = (1+z)^{-k} f_k(z)$ and $(\mathrm{Id} + \mathcal{A})^{-k} \in \mathcal{L}(X, \mathcal{D}(\mathcal{A}^k))$, we get:

$$S_t = (\mathrm{Id} + \mathcal{A})^{-k} f_k(\mathcal{A}) \in \mathcal{L}(X, \mathcal{D}(\mathcal{A}^k)).$$

(b) Let $K_2 = \sup\{|\zeta^2 e^{-\zeta}|, \zeta \in \mathcal{S}_{\alpha+\beta}\}$. For $t, s \in \mathcal{S}_\beta$, both nonzero, the Taylor expansion yields:

$$\forall z \in \mathcal{S}_\alpha, \quad |e^{-tz} - e^{-sz} + (t-s)ze^{-sz}| \leq \frac{K_2}{2} \max_{0\leq x\leq 1} \frac{1}{|xt + (1-x)s|^2} |t-s|^2.$$

Using Theorem 8:

$$\|S_t - S_s + (t-s)\mathcal{A}S_s\|_{\mathcal{L}(X)} \leq C_\alpha \frac{K_2}{2} \max_{0\leq x\leq 1} \frac{1}{|xt + (1-x)s|^2} |t-s|^2,$$

so $t \mapsto S_t$ is differentiable with $S_t' = -\mathcal{A}S_t$, and by induction:

$$\forall k \in \mathbb{N}, \quad S_t^{(k)} = (-\mathcal{A})^k S_t.$$

(c) Fix $t > 0$, let $s(\theta) = t + re^{i\theta}$, with $r = t\sin\beta$, $\theta \in [0, 2\pi)$. Then $s(\theta) \in \mathcal{S}_\beta$. For $u, v \in X$, define $\varphi(z) = \langle S_z u, v\rangle$. From the last paragraph, φ is holomorphic in $\mathcal{S}_\beta$, and:

$$\forall z \in \mathcal{S}_\beta, \quad |\varphi(z)| \leq \|S_z\|_{\mathcal{L}(X)}\|u\|\|v\| \leq \|u\|\|v\|.$$

Cauchy's residue theorem gives:

$$\frac{1}{k!}\varphi^{(k)}(t) = \frac{1}{2\pi i}\int_0^{2\pi} \frac{\varphi(s(\theta))}{(s(\theta))^{k+1}} \mathrm{d}s(\theta),$$

leading to:

$$|\varphi^{(k)}(t)| \leq \frac{k!}{2\pi}\int_0^{2\pi} \frac{\|u\|\|v\|}{r^k}\mathrm{d}\theta = \frac{k!\|u\|\|v\|}{(t\sin\beta)^k}.$$

Since $\varphi^{(k)}(t) = \langle S_t^{(k)}u, v\rangle$, we conclude:

$$\|S_t^{(k)}\|_{\mathcal{L}(X)} \leq \frac{k!}{(t\sin\beta)^k}.$$

Letting $\beta \to \frac{\pi}{2} - \alpha$ yields the result. □

Theorem 28. *Given $z_0 \in X$, the unique solution $z \in \mathcal{C}^1((0,\infty); \mathcal{D}(\mathcal{A})) \cap \mathcal{C}^0([0,\infty); X)$ to*

$$\begin{cases} z'(t) = -\mathcal{A}z(t) \\ z(0) = z_0 \end{cases} \tag{28}$$

satisfies $z \in \mathcal{C}^\infty((0,\infty); \mathcal{D}(\mathcal{A}^k))$ for all $k \in \mathbb{N}$. Moreover, if $z_0 \in \mathcal{D}(\mathcal{A}^\ell)$, then $z \in \mathcal{C}^0([0,\infty); \mathcal{D}(\mathcal{A}^\ell)) \cap \mathcal{C}^\ell([0,\infty); X)$ and

$$\forall k, \ell \geq 0, \forall t > 0, \quad \|z^{(k+l)}(t)\| = \|\mathcal{A}^k z^{(l)}(t)\| \leq \frac{k!}{(t\cos(\alpha))^k}\|\mathcal{A}^\ell z_0\|. \tag{29}$$

Proof. It is clear that $t \mapsto z(t)$ is a solution of (P) since $S_t' + \mathcal{A}S_t = 0$. The C^∞ regularity on $(0,\infty)$ follows from Theorem 27, while continuity at $t = 0$ in X results from the strong continuity of the semigroup. Since the problem is linear, to prove uniqueness of the solution to (28), it suffices to show that $z_0 = 0$ implies $z(t) = 0$. Indeed, if z is a solution of (28) and $z_0 = 0$, then:

$$\frac{d}{dt}\|z(t)\|^2 = \langle z'(t), z(t)\rangle + \langle z(t), z'(t)\rangle = 2\,\mathrm{Re}\langle z'(t), z(t)\rangle = -2\,\mathrm{Re}\langle \mathcal{A}z(t), z(t)\rangle \leq 0.$$

It follows that $\|z(t)\|^2 \leq \|z(0)\|^2 = 0$, hence uniqueness.

Now, if $z_0 \in \mathcal{D}(\mathcal{A})$, one easily verifies that:

$$\left(\mathrm{Id} + \frac{t}{n}\mathcal{A}\right)^{-1}\mathcal{A}z_0 = \mathcal{A}\left(\mathrm{Id} + \frac{t}{n}\mathcal{A}\right)^{-1} z_0,$$

and by induction:

$$\left(\mathrm{Id} + \frac{t}{n}\mathcal{A}\right)^{-n}\mathcal{A}z_0 = \mathcal{A}\left(\mathrm{Id} + \frac{t}{n}\mathcal{A}\right)^{-n} z_0.$$

Passing to the limit (recalling that $S_t = \lim_{n\to\infty}\left(\mathrm{Id} + \frac{t}{n}\mathcal{A}\right)^{-n}$), we deduce:

$$\mathcal{A}S_t z_0 = S_t \mathcal{A}z_0.$$

Similarly, if $z_0 \in \mathcal{D}(\mathcal{A}^\ell)$, then:

$$\mathcal{A}^\ell S_t z_0 = S_t \mathcal{A}^\ell z_0.$$

We then observe that $z^{(\ell)}$ is a solution of problem (28) with z_0 replaced by $(-\mathcal{A})^\ell z_0$, from which it follows that $z^{(\ell)} \in \mathcal{C}^0([0,\infty); X)$. Estimate (29) then follows from Theorem 27. □

B.2 Discretisation of the adjoint equation

In this section, we consider the discretisation described in Section 2, developing time- and space-discretisation estimates, ultimately providing the the proof of Proposition 9. First, the following lemma proves that a consequence of $(\mathcal{V}_1)$ is the approximation in a $\mathcal{O}(h^2)$ manner in the weaker X-norm for $u \in \mathcal{D}(\mathcal{A})$. In a similar manner, one could prove the same lemma for $u \in \mathcal{D}(\mathcal{A}^*)$.

Lemma 29. *Under $(\mathcal{V}_1)$ and the hypotheses on V_h and $\mathcal{A}_h$ described in Subsection 2.2, we have*

$$\forall f \in X, \quad \|(\mathcal{A}^{-1} - \mathcal{A}_h^{-1}P_h)f\| \leq C_1 h^2 \|f\|, \tag{$\mathcal{H}_1$}$$

where

$$C_1 = \frac{a_1^2 C_0^2}{a_0} \quad \Big(\textit{resp. } C_1 = \frac{a_1^{3/2} C_0^2}{4\sqrt{a_0}} \textit{ if } (\mathcal{A}, \mathcal{D}(\mathcal{A})) \textit{ is self-adjoint}\Big).$$

Proof. We shall only prove the lemma in the first case. If $\mathcal{A}$ is self-adjoint, the adjusted constant follows from the exact same method but using Céa's Lemma which yields $C_c = \sqrt{\frac{a_1}{a_0}}$, and noticing that $(\mathcal{V}_1)$ simplifies into

$$\forall f \in X, \quad \inf_{v_h \in V_h} \|\mathcal{A}^{-1}f - v_h\|_V \leq \frac{1}{2} C_0 h \|f\|.$$

Let $P_h : X \to V_h$ be the orthogonal projection onto V_h, so that by definition

$$\forall (v, w_h) \in X \times V_h, \quad P_h v \in V_h \quad \text{and} \quad \langle P_h v, w_h \rangle = \langle v, w_h \rangle.$$

Let $f \in X$. Applying Céa's Lemma to the maps

$$a : \begin{cases} V \times V \to \mathbb{C} \\ (v, w) \mapsto \langle \mathcal{A}v, w \rangle \end{cases} \qquad L : \begin{cases} V \to \mathbb{C} \\ v \mapsto \langle v, f \rangle \end{cases}$$

we get

$$\|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_h f\|_V \leq C_c \inf_{v_h \in V_h} \|\mathcal{A}^{-1}f - v_h\|, \tag{30}$$

with $C_c := \frac{a_1}{a_0}$. Notice that for all $v_h \in V_h$

$$\begin{aligned} \langle \mathcal{A}(\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_h f), v_h \rangle &= \langle f, v_h \rangle - \langle \mathcal{A}\mathcal{A}_h^{-1}P_h f, v_h \rangle \\ &= \langle f, v_h \rangle - \langle \mathcal{A}_h \mathcal{A}_h^{-1}P_h f, v_h \rangle \text{ since } \mathcal{A}_h^{-1}P_h f \in V_h \\ &= \langle f, v_h \rangle - \langle P_h f, v_h \rangle \\ &= 0. \end{aligned} \tag{31}$$

Using the Aubin-Nitsche trick, we let $g \in X$ such that $\|g\| = 1$ and write

$$\begin{aligned} \langle g, \mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_h f \rangle &= \langle \mathcal{A}^*(\mathcal{A}^*)^{-1}g, \mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_h f \rangle \\ &= \langle (\mathcal{A}^*)^{-1}g - v_h, \mathcal{A}(\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_h f) \rangle \quad \forall v_h \in V_h \text{ using (31)} \\ &\leq a_1 \|(\mathcal{A}^*)^{-1}g - v_h\|_V \|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_h f\|_V \quad \forall v_h \in V_h \\ &\leq a_1 \|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_h f\|_V \inf_{v_h \in V_h} \|(\mathcal{A}^*)^{-1}g - v_h\|_V. \end{aligned}$$

Hence

$$\begin{aligned} \|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_h f\| &= \sup_{\substack{g \in X \\ \|g\|=1}} \langle g, \mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_h f \rangle \\ &\leq a_1 \|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_h f\|_V \sup_{\|g\|=1} \inf_{v_h \in V_h} \|(\mathcal{A}^*)^{-1}g - v_h\|_V. \end{aligned}$$

Using the bounds ($\mathcal{V}_1$) and (30), we obtain

$$\begin{aligned}\|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_h f\| &\le a_1\|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_h f\|_V \sup_{\|g\|=1} C_0 h\|g\| \text{ using } (\mathcal{V}_1)\\ &= a_1 C_0 h\|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_h f\|_V\\ &\le a_1 C_0 C_c h\sqrt{a_0}h \inf_{v_h\in V_h}\|\mathcal{A}^{-1}f - v_h\|_V \text{ using (30)}\\ &\le a_1 C_0^2 C_c h^2\|f\| \text{ using } (\mathcal{V}_1).\end{aligned}$$

The result is proved, with $C_1 = a_1 C_0^2 C_c = \frac{a_1^2 C_0^2}{a_0}$. □

The time approximation described in the following Proposition 31 first requires a complex analytic lemma, which could be avoided at the cost of higher constants.

Lemma 30. *Let $z \in \mathbb{C}$ such that $\mathrm{Re}(z) \le 0$. Then*

$$\left|e^z - \frac{1}{1-z}\right| \le \frac{1}{2}|z|^2.$$

Proof. Define

$$F(z) := \frac{1}{z^2}\left(e^z - \frac{1}{1-z}\right) \quad \text{for } z \neq 0, \qquad \text{and} \qquad F(0) = \frac{1}{2}.$$

A standard series expansion yields that 0 is a removable singularity so that F is holomorphic on the half-plane $\{\mathrm{Re}\, z < 0\}$. Fix $R > 1$ and consider the rectangle

$$\mathcal{R}_R := \{x + iy : \ -R \le x \le 0, \ |y| \le R\}.$$

By the maximum modulus principle, the maximum of $|F|$ on the compact set $\overline{\mathcal{R}_R}$ is attained on its boundary $\partial\mathcal{R}_R$. Let us estimate $|F|$ on each part of $\partial\mathcal{R}_R$.

On the imaginary side $x = 0$, $y \in [-R, R]$. For $z = it$ with $t \in \mathbb{R}, t \neq 0$, a direct computation yields

$$F(it) = \frac{(1-it)e^{it} - 1}{t^2(1-it)} \qquad \text{and} \qquad |F(it)|^2 = \frac{t^2 + 2(1 - \cos t - t\sin t)}{t^4(1+t^2)}.$$

To prove that $|F(it)| \le \frac{1}{2}$, it suffices to prove that $4\big(t^2 + 2(1-\cos t - t\sin t)\big) \le t^4(1+t^2)$ for all $t \in \mathbb{R}$. Define

$$G(t) := 4\big(t^2 + 2(1 - \cos t - t\sin t)\big) - t^4(1+t^2).$$

A direct differentiation gives

$$G'(t) = 2t\big(-3t^4 - 2t^2 - 4\cos t + 4\big) = 2t\big(-3t^4 - 2t^2 + 4(1-\cos t)\big).$$

Since $1 - \cos t \le \frac{t^2}{2}$, we obtain

$$-3t^4 - 2t^2 + 4(1-\cos t) \le -3t^4 - 2t^2 + 2t^2 = -3t^4 \le 0,$$

so $G'(t) \le 0$ for $t \ge 0$. Since $G(0) = 0$ and $G' \le 0$ on $[0, \infty)$, we have $G(t) \le 0$ for all $t \ge 0$. We conclude by using the evenness of G.

On the left vertical side $x = -R$, $|y| \le R$. For $\mathrm{Re}\, z = -R$, we have $|e^z| \le e^{-R}$. Also

$$\left|\frac{1}{1-z}\right| = \frac{1}{|1-z|} \le \frac{1}{|z|-1} \le \frac{2}{|z|}$$

for R large, say $R > 2$, so

$$|e^z - \tfrac{1}{1-z}| \leq e^{-R} + \frac{2}{|z|}.$$

Thus on this side

$$|F(z)| = \frac{|e^z - \frac{1}{1-z}|}{|z|^2} \leq \frac{e^{-R}}{R^2} + \frac{2}{R^3} \xrightarrow[R\to\infty]{} 0.$$

On the top and bottom horizontal sides $-R \leq x \leq 0$, $y = \pm R$. Write $z = x \pm iR$. Then $|z| \geq R$ and $|e^z| \leq e^x \leq 1$. Also $|1/(1-z)| \leq 2/|z|$ for R large, so

$$|F(z)| \leq \frac{1 + 2/|z|}{|z|^2} \leq \frac{2}{R^2} \xrightarrow[R\to\infty]{} 0$$

uniformly in $x \in [-R, 0]$.

Combining the three last inequalities, the maximum of $|F|$ on $\overline{\mathcal{R}_R}$ is $1/2$ whenever R is chosen large enough. The expected conclusion follows. □

Proposition 31 (Time approximation)**.** *For $z_0 \in \mathcal{D}(\mathcal{A})$, we consider*

$$\begin{cases} \dot{z}(t) = -\mathcal{A}z(t) \\ z(0) = z_0. \end{cases}$$

Define $N_T \in \mathbb{N}^*$, $\Delta t = \frac{T}{N_T}$, and for $n \in \{0, \dots, N_T\}$, $t_n = n\Delta t$. Discretise the system using the implicit Euler scheme:

$$\begin{cases} z_0 = z_0 \\ (\mathrm{Id} + \Delta t\mathcal{A})z_{n+1} = z_n \quad \forall n \in \{0, \dots, N_T - 1\}. \end{cases}$$

Then

$$\forall n \in \{0, \dots, N_T\}, \quad \|z(t_n) - z_n\| \leq \Delta t \, \frac{C_\alpha}{\cos\alpha} \|\mathcal{A}z_0\|.$$

where $C_\alpha \leq 2 + \frac{2}{\sqrt{3}}$.

Proof. By definition, for all $n \in \{0, \dots, N_T\}$, we have

$$z(t_n) - z_n = \left[e^{-t_n\mathcal{A}} - (\mathrm{Id} + \Delta t\mathcal{A})^{-n}\right] z_0.$$

Hence me may write

$$z(t_n) - z_n = \Delta t \varphi_n(\Delta t\mathcal{A})\mathcal{A}z_0,$$

where for $n \in \mathbb{N}^*$, the function φ_n is defined for $z \in \mathcal{S}_\alpha$ by

$$\varphi_n(z) = \frac{e^{-nz} - (1+z)^{-n}}{z},$$

extended by continuity at $z = 0$ by $\varphi_n(0) = 0$.

Using Theorem 8, we find

$$\|z(t_n) - z_n\| \leq \Delta t \|\varphi_n(\Delta t\mathcal{A})\|_{\mathcal{L}(X)} \|\mathcal{A}z_0\| \leq \Delta t C_\alpha \sup_{z\in\mathcal{S}_\alpha} |\varphi_n(z)| \|\mathcal{A}z_0\|. \tag{32}$$

Let us finish with the proof that $\sup_{z\in\mathcal{S}_\alpha} |\varphi_n(z)| \leq \frac{1}{\cos(\alpha)}$. Let $z \in \mathcal{S}_\alpha$, $z \neq 0$ (the case $z = 0$ is obvious). Then $\mathrm{Re}(z) > 0$, hence by Lemma 30 applied to $-z$,

$$|\varphi_1(z)| = \frac{1}{|z|}\left[e^{-z} - \frac{1}{1+z}\right] \leq \frac{1}{2}|z|.$$

Since

$$\varphi_n(z) = \frac{1}{z}\left[e^{-z} - \frac{1}{1+z}\right]\sum_{k=0}^{n-1} e^{-kz}\left(\frac{1}{1+z}\right)^{n-k-1},$$

we get

$$|\varphi_n(z)| \leq \frac{1}{2}|z|\sum_{k=0}^{n-1}\left|e^{-z}\right|^k\left|\frac{1}{1+z}\right|^{n-k-1}.$$

We may write $z = \rho e^{i\theta}$, where $\rho > 0,\ \theta \in [-\alpha, \alpha]$. It follows that

$$|1+z| = |1+\rho e^{i\theta}| \geq |1+\rho e^{i\alpha}| \geq 1+\rho\cos(\alpha).$$

Thus

$$\left|\frac{1}{1+z}\right| \leq \frac{1}{1+\rho\cos(\alpha)}.$$

Since $|\theta| \leq \alpha \leq \frac{\pi}{2}$, $\cos(\alpha) \leq \cos(\theta)$, we get

$$|\varphi_n(z)| \leq \frac{1}{2}\rho\sum_{k=0}^{n-1}\left(e^{-\rho\cos(\alpha)}\right)^k\left(\frac{1}{1+\rho\cos(\alpha)}\right)^{n-k-1}.$$

Since $\rho\cos(\alpha) \geq 0$, we have

$$e^{\rho\cos(\alpha)} \geq 1+\rho\cos(\alpha), \qquad e^{-\rho\cos(\alpha)} \leq \frac{1}{1+\rho\cos(\alpha)}.$$

Finally

$$|\varphi_n(z)| \leq \frac{1}{2}\rho n\frac{1}{(1+\rho\cos(\alpha))^{n-1}},$$

and maximising the right-hand side for $\rho > 0$ yields the desired result if $n \geq 2$. As for $n = 1$, if $|z| \leq 2$ the upper bound from Lemma 30 proves that $|\varphi_1(z)| \leq 1 \leq \frac{1}{\cos(\alpha)}$, and similarly and if $|z| \geq 2$, $|\varphi_1(z)| \leq \frac{2}{|z|} \leq 1 \leq \frac{1}{\cos(\alpha)}$. □

Proposition 32 (Spatial approximation)**.** *Let $z_{h,0} \in V_h$, $z_h : t \mapsto S_{t,h}z_{h,0}$, where $S_{t,h} = \exp(-t\mathcal{A}_h)$. For $z_0 \in \mathcal{D}(\mathcal{A})$, $z(t) = \exp(-t\mathcal{A})z_0$, there holds*

$$\forall t > 0, \quad \|z(t) - z_h(t)\| \leq \|z_0 - z_{h,0}\| + \left(6 + \frac{4\ln(2)}{\cos(\alpha)}\right)C_1h^2\|\mathcal{A}z_0\|.$$

Proof. For $t > 0$, we write

$$\begin{aligned}\|z(t) - P_hz(t)\| &\leq \|z(t) - \mathcal{A}_h^{-1}P_h\mathcal{A}z(t)\| \text{ since } \mathcal{A}_h^{-1}P_h\mathcal{A}z(t) \in V_h\\ &= \|(\mathcal{A}^{-1} - \mathcal{A}_h^{-1}P_h)(\mathcal{A}z(t))\|\\ &\leq C_1h^2\|\mathcal{A}z(t)\| \text{ using } (\mathcal{H}_1).\end{aligned}$$

Similarly,

$$\begin{aligned}\|z'(t) - P_hz'(t)\| &\leq \|z'(t) - \mathcal{A}_h^{-1}P_h\mathcal{A}z'(t)\| \leq C_1h^2\|\mathcal{A}z'(t)\|\\ &\leq \frac{C_1h^2}{t\cos(\alpha)}\|\mathcal{A}z_0\| \text{ using Theorem 28.}\end{aligned}$$

Using the triangle inequality, we deduce the upper bounds

$$\|(P_h - \mathcal{A}_h^{-1}P_h\mathcal{A})z(t)\| \leq 2C_1h^2\|\mathcal{A}z(t)\| \quad \text{and} \quad \|(P_h - \mathcal{A}_h^{-1}P_h\mathcal{A})z'(t)\| \leq 2\frac{C_1h^2}{t\cos(\alpha)}\|\mathcal{A}z_0\|. \tag{33}$$

Denote $e_h : t \mapsto \mathcal{A}_h^{-1} P_h \mathcal{A} z(t) - z_h(t)$. Thus

$$\|z(t) - z_h(t)\| \leq \|e_h(t)\| + \|z(t) - \mathcal{A}_h^{-1} P_h \mathcal{A} z(t)\| \leq \|e_h(t)\| + C_1 h^2 \|\mathcal{A} z(t)\|.$$

Let us bound $e_h(t)$. Since $z_h'(t) = -\mathcal{A}_h z_h(t)$, we have

$$e_h'(t) + \mathcal{A}_h e_h(t) = \mathcal{A}_h^{-1} P_h \mathcal{A} z'(t) + P_h \mathcal{A} z(t) = (\mathcal{A}_h^{-1} P_h \mathcal{A} - P_h) z'(t).$$

Duhamel's formula yields

$$e_h(t) = S_{t,h} e_h(0) + \int_0^t S_{t-\sigma,h} (\mathcal{A}_h^{-1} P_h \mathcal{A} - P_h) z'(\sigma) \mathrm{d}\sigma =: E_1 + E_2 + E_3,$$

where $E_1 = S_{t,h} e_h(0)$, $E_2 = \int_{t/2}^t S_{t-\sigma,h} (\mathcal{A}_h^{-1} P_h \mathcal{A} - P_h) p'(\sigma) \mathrm{d}\sigma$ and $E_3 = \int_0^{t/2} S_{t-\sigma,h} (\mathcal{A}_h^{-1} P_h \mathcal{A} - P_h) z'(\sigma) \mathrm{d}\sigma$.

First, we have, using $\|S_{t,h}\|_{\mathcal{L}(X)} \leq 1$ (see Theorem 27 applied to $\mathcal{A}_h$) and ($\mathcal{H}_1$)

$$\begin{aligned} \|E_1\| &\leq \|e_h(0)\| = \|z_{h,0} - \mathcal{A}_h^{-1} P_h \mathcal{A} z_0\| \leq \|z_0 - z_{h,0}\| + \|(\mathcal{A}^{-1} - \mathcal{A}_h^{-1} P_h)(\mathcal{A} z_0)\| \\ &\leq \|z_0 - z_{h,0}\| + C_1 h^2 \|\mathcal{A} z_0\|. \end{aligned}$$

Secondly,

$$\begin{aligned} \|E_2\| &\leq \int_{t/2}^t \|S_{t-\sigma,h} (\mathcal{A}_h^{-1} P_h \mathcal{A} - P_h) z'(\sigma)\| \mathrm{d}\sigma \leq \int_{t/2}^t \|(\mathcal{A}_h^{-1} P_h \mathcal{A} - P_h) z'(\sigma)\| \mathrm{d}\sigma \\ &\leq \int_{t/2}^t 2 \frac{C_1 h^2}{\sigma \cos(\alpha)} \|\mathcal{A} z_0\| \mathrm{d}\sigma = \frac{2 \ln(2)}{\cos(\alpha)} C_1 h^2 \|\mathcal{A} z_0\|, \end{aligned}$$

where we used (33).

For the last term E_3, we first integrate by parts, then use (33) to uncover

$$\begin{aligned} \|E_3\| &= \left\| S_{t/2,h} (\mathcal{A}_h^{-1} P_h \mathcal{A} - P_h) z(t/2) - S_{t,h} (\mathcal{A}_h^{-1} P_h \mathcal{A} - P_h) z(0) + \int_0^{t/2} \mathcal{A}_h S_{t-\sigma,h} (\mathcal{A}_h^{-1} P_h \mathcal{A} - P_h) z(\sigma) \mathrm{d}\sigma \right\| \\ &\leq 2 C_1 h^2 \|\mathcal{A} z_0\| + 2 C_1 h^2 \|\mathcal{A} z_0\| + 2 C_1 h^2 \|\mathcal{A} z_0\| \int_0^{t/2} \|\mathcal{A}_h S_{t-\sigma,h}\|_{\mathcal{L}(X)} \mathrm{d}\sigma. \end{aligned}$$

It follows from Theorem 28 that

$$\|\mathcal{A}_h S_{t-\sigma,h}\|_{\mathcal{L}(X)} \leq \frac{1}{(t-\sigma)\cos(\alpha)},$$

hence

$$\|E_3\| \leq \left(4 + \frac{2 \ln(2)}{\cos(\alpha)} \right) C_1 h^2 \|\mathcal{A} z_0\|.$$

Gathering all the estimates, we obtain

$$\|z(t) - z_h(t)\| \leq \|E_1\| + \|E_2\| + \|E_3\| + C_1 h^2 \|\mathcal{A} z_0\| \leq \|z_0 - z_{h,0}\| + \left(6 + 4 \frac{\ln(2)}{\cos(\alpha)} \right) C_1 h^2 \|\mathcal{A} z_0\|.$$

□

Proof of Proposition 9. We keep the notations of the previous proof. Given $n \geq 1$, we have

$$\|z(t_n) - z_{h,n}\| \leq E_1 + E_2 + E_3$$
$$\text{with} \quad \begin{cases} E_1 = \|z(t_n) - S_{t_n,h}\mathcal{A}_h^{-1}P_h\mathcal{A}z_0\| \\ E_2 = \|(S_{t_n,h} - (\mathrm{Id} + \Delta t\mathcal{A}_h)^{-n})\mathcal{A}_h^{-1}P_h\mathcal{A}z_0\| \\ E_3 = \|(\mathrm{Id} + \Delta t\mathcal{A}_h)^{-n}(\mathcal{A}_h^{-1}P_h\mathcal{A}z_0 - z_{h,0})\| \end{cases}$$

It follows from Proposition 32 with $z_{h,0} = \mathcal{A}_h^{-1}P_h\mathcal{A}z_0$ and ($\mathcal{H}_1$) that

$$E_1 \leq \|z_0 - \mathcal{A}_h^{-1}P_h\mathcal{A}z_0\| + \left(6 + \frac{4\ln(2)}{\cos(\alpha)}\right) C_1 h^2 \|\mathcal{A}z_0\| \leq \left(7 + \frac{4\ln(2)}{\cos(\alpha)}\right) C_1 h^2 \|\mathcal{A}z_0\|.$$

Using Proposition 31 with $\mathcal{A}$ replaced by $\mathcal{A}_h$,

$$E_2 \leq \frac{\Delta t}{\cos(\alpha)} C_\alpha \|\mathcal{A}_h\mathcal{A}_h^{-1}P_h\mathcal{A}z_0\| \leq \frac{\Delta t}{\cos(\alpha)} C_\alpha \|\mathcal{A}z_0\|.$$

Using Theorem 8, ($\mathcal{H}_1$), and since $\sup_{z\in\mathcal{S}_\alpha} |\frac{1}{1+z}| = 1$,

$$\begin{aligned} E_3 &\leq C_\alpha \|\mathcal{A}_h^{-1}P_h\mathcal{A}z_0 - z_{h,0}\| \\ &\leq C_\alpha (\|\mathcal{A}_h^{-1}P_h\mathcal{A}z_0 - z_0\| + \|z_0 - z_{h,0}\|) \\ &\leq C_\alpha C_1 h^2 \|\mathcal{A}z_0\| + C_\alpha \|z_0 - z_{h,0}\|. \end{aligned}$$

Finally,

$$\|z(t_n) - z_{h,n}\| \leq C_\alpha \|z_0 - z_{h,0}\| + \left(C_1 \left(7 + \frac{4\ln(2)}{\cos(\alpha)} + C_\alpha\right) h^2 + \frac{C_\alpha}{\cos(\alpha)} \Delta t\right) \|\mathcal{A}z_0\|.$$

Noticing that $\alpha = \arccos(\frac{a_0}{a_1})$ in the case we consider, the final result is obtained, with the constants

$$C_2 = C_1 \left(7 + 4\ln(2)\frac{a_1}{a_0} + C_\alpha\right), \qquad C_3 = \frac{a_1}{a_0} C_\alpha.$$

□

References

[1] M. Althoff, G. Frehse, and A. Girard. Set Propagation Techniques for Reachability Analysis. Annual Review of Control, Robotics, and Autonomous Systems, 4:369–395, 2021. Publisher: Annual Reviews.

[2] H. Antil, U. Biccari, R. Ponce, M. Warma, and S. Zamorano. Controllability properties from the exterior under positivity constraints for a 1-d fractional heat equation. Evolution Equations and Control Theory, 13(3):893–924, 2024.

[3] A. Balakrishnan. Introduction to Optimization Theory in a Hilbert Space. Lecture Notes in Operations Research and Mathematical Systems, Economics, Computer Science, Information and Control. Springer, 1971.

[4] I. Balázs, J. B. van den Berg, J. Courtois, J. Dudás, J.-P. Lessard, A. Vörös-Kiss, J. Williams, and X. Y. Yin. Computer-assisted proofs for radially symmetric solutions of pdes. Journal of Computational Dynamics, 5(1&2):61–80, 2018.

[5] J. B. v. d. Berg, M. Breden, J.-P. Lessard, and L. v. Veen. Spontaneous periodic orbits in the navier–stokes flow. Journal of Nonlinear Science, 31(2):41, Mar. 2021.

[6] J. B. v. d. Berg, M. Breden, and R. Sheombarsing. Validated integration of semilinear parabolic PDEs. Numerische Mathematik, 156(4):1219–1287, Aug. 2024.

[7] L. Berrahmoune. A variational approach to constrained controllability for distributed systems. Journal of Mathematical Analysis and Applications, 416(2):805–823, 2014. Publisher: Elsevier.

[8] L. Berrahmoune. Constrained null controllability for distributed systems and applications to hyperbolic-like equations. ESAIM: Control, Optimisation and Calculus of Variations, 25:32, 2019. Publisher: EDP Sciences.

[9] L. Berrahmoune. A variational approach to constrained null controllability for the heat equation. European Journal of Control, 52:42 – 48, 2020.

[10] F. Boyer. Controllability of linear parabolic equations and systems. Feb. 2022.

[11] R. F. Brammer. Controllability in linear autonomous systems with positive controllers. SIAM Journal on Control, 10(2):339–353, 1972. Publisher: SIAM.

[12] E. Casas and K. Kunisch. Boundary control of semilinear parabolic equations with non-smooth point-wise-integral control constraints in time-space. In 2022 American Control Conference (ACC), pages 284–289, 2022.

[13] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of mathematical imaging and vision, 40(1):120–145, 2011. Publisher: Springer.

[14] M. Chen and L. Rosier. Reachable states for the distributed control of the heat equation. Comptes Rendus. Mathématique, 360:627–639, 2022. Publisher: Académie des sciences, Paris.

[15] M. Chen and C. J. Tomlin. Hamilton–Jacobi Reachability: Some Recent Theoretical Advances and Applications in Unmanned Airspace Management. Annual Review of Control, Robotics, and Autonomous Systems, 1(1):333–358, 2018. _eprint: https://doi.org/10.1146/annurev-control-060117-104941.

[16] M. Crouzeix and B. Delyon. Some estimates for analytic functions of strip or sectorial operators. Archiv der Mathematik, 81:559–566, 2003. Publisher: Springer.

[17] M. Crouzeix and A. Greenbaum. Spectral sets: numerical range and beyond. SIAM Journal on Matrix Analysis and Applications, 40(3):1087–1101, 2019. Publisher: SIAM.

[18] M. Crouzeix and C. Palencia. The numerical range is a (1+2)-spectral set. SIAM Journal on Matrix Analysis and Applications, 38(2):649–655, 2017. Publisher: SIAM.

[19] J. Dardé and S. Ervedoza. On the reachable set for the one-dimensional heat equation. SIAM Journal on Control and Optimization, 56(3):1692–1715, 2018. _eprint: https://doi.org/10.1137/16M1093215.

[20] P. Dario and E. Zuazua. Controllability under positivity constraints of semilinear heat equations. Mathematical Control & Related Fields, 8(3&4):935–964, 2018.

[21] S. Day, Y. Hiraoka, K. Mischaikow, and T. Ogawa. Rigorous numerics for global dynamics: A study of the swift–hohenberg equation. SIAM Journal on Applied Dynamical Systems, 4, Mar. 2004.

[22] C. De Boor. A practical guide to splines; rev. ed. Applied mathematical sciences. Springer, Berlin, 2001.

[23] S. Ervedoza, K. Le Balc'h, and M. Tucsnak. Reachability results for perturbed heat equations. Journal of Functional Analysis, 283(10):109666, 2022. Publisher: Elsevier.

[24] H. Fattorini. The time-optimal control problem in banach spaces. Applied Mathematics and Optimization, 1(2):163–188, 1974.

[25] A. Hartmann, K. Kellay, and M. Tucsnak. From the reachable space of the heat equation to Hilbert spaces of holomorphic functions. Journal of the European Mathematical Society, 22(10):3417–3440, 2020.

[26] I. Hasenohr, C. Pouchol, Y. Privat, and C. Zhang. Computer-assisted proofs of nonreachability for finite-dimensional linear control systems. SIAM Journal on Control and Optimization, 63(5):3272–3296, 2025.

[27] K. Kellay, T. Normand, and M. Tucsnak. Sharp reachability results for the heat equation in one space dimension. Analysis & PDE, 15(4):891–920, 2022. Publisher: Mathematical Sciences Publishers.

[28] H. Kong, E. Bartocci, and T. A. Henzinger. Reachable set over-approximation for nonlinear systems using piecewise barrier tubes. In H. Chockler and G. Weissenbacher, editors, Computer Aided Verification, pages 449–467, Cham, 2018. Springer International Publishing.

[29] G. Lebeau and L. Robbiano. Contrôle exact de l'équation de la chaleur. Séminaire Équations aux dérivées partielles (Polytechnique), pages 1–11, 1994. Publisher: Ecole Polytechnique, Centre de Mathématiques.

[30] J. Lohéac, E. Trélat, and E. Zuazua. Minimal controllability time for finite-dimensional control systems under state constraints. Automatica, 96:380–392, 2018. Publisher: Elsevier.

[31] J. Lohéac, E. Trélat, and E. Zuazua. Nonnegative control of finite-dimensional linear systems. Annales de l'Institut Henri Poincaré C, Analyse non linéaire, 38(2):301–346, 2021. Publisher: Elsevier.

[32] S. Micu and E. Zuazua. An introduction to the controllability of partial differential equations. Quelques questions de théorie du contrôle. Sari, T., ed., Collection Travaux en Cours Hermann, to appear, 2004.

[33] I. Mitchell, A. Bayen, and C. Tomlin. A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games. IEEE Transactions on Automatic Control, 50(7):947–957, 2005.

[34] C. Pouchol, E. Trélat, and C. Zhang. Approximate control of parabolic equations with on-off shape controls by Fenchel duality. Annales de l'Institut Henri Poincaré C, pages 1–43, 2024.

[35] S. Prajna and A. Jadbabaie. Safety Verification of Hybrid Systems Using Barrier Certificates. In R. Alur and G. J. Pappas, editors, Hybrid Systems: Computation and Control, pages 477–492, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[36] A. Quarteroni, R. Sacco, and F. Saleri. Numerical mathematics, volume 37. Springer Science & Business Media, 2006.

[37] W. Rudin. Functional Analysis. International series in pure and applied mathematics. McGraw-Hill, 1991.

[38] S. M. Rump. INTLAB—interval laboratory. In Developments in reliable computing, pages 77–104. Springer, 1999.

[39] M. Tucsnak and G. Weiss. Observation and control for operator semigroups. Springer Science & Business Media, 2009.

[40] G. Wang. l^∞-null controllability for the heat equation and its consequences for the time optimal control problem. Siam Journal on Control and Optimization - SIAM, 47:1701–1720, Jan. 2008.