Numerical null controllability of the 1D heat equation: primal and dual algorithms

ENRIQUE FERNÁNDEZ-CARA* and ARNAUD MÜNCH†

Abstract

This paper deals with the numerical computation of distributed null controls for the 1D heat equation, with Dirichlet boundary conditions. The goal is to compute a control that drives (a numerical approximation of) the solution from a prescribed initial state at $t = 0$ to zero at $t = T$. This problem is generically ill-posed, a feature which is closely related to the strong regularization effect of the heat equation. Using ideas from Fursikov and Imanuvilov [20], we consider the control that minimizes over the class of admissible null controls a functional that involves weighted integrals of the state and the control, with weights that blow up near $T$. We address this minimization problem first with primal and then with dual methods.

In view of the results in [20], the optimality system is equivalent to a differential problem that is fourth-order in space and second-order in time. We first address the numerical solution of the corresponding variational formulation by introducing a space-time finite element that is $C^1$ in space and $C^0$ in time. We prove a strong convergence result for the approximate controls and then we present some numerical experiments. In order to circumvent $C^1$ finite elements, we also introduce a mixed variational formulation and we prove well-posedness through a suitable inf-sup condition. We introduce a (non-conformal) $C^0$ finite element approximation and we provide new numerical results. In both cases, thanks to an appropriate change of variable, we observe a polynomial dependance of the condition number with respect to the discretization parameter. Furthermore, with this second method, the initial and final conditions are satisfied exactly.

Then, we address the controllability problem using duality arguments, extending the earlier contribution of Carthé, Glowinski and Lions [8] devoted to the computation of minimal $L^2$-norm controls. We formulate some constrained extremal problems (each of them corresponding again to the minimization of a functional that involves weighted integrals of the state and the control) and we apply appropriate duality techniques. We introduce appropriate numerical approximations of the associated dual problems and we solve them by applying conjugate gradient techniques. Finally, we present several experiments, we highlight the influence of the weights and we compare this approach and the first one in tems of robustness and efficiency.

Keywords: one-dimensional heat equation, null controllability, finite element methods, mixed finite elements, Carleman inequalities, dual methods.

Mathematics Subject Classification (2010)- 35K35, 65M12, 93B40.

*Dpto. EDAN, University of Sevilla, Aptdo. 1160, 41080 Sevilla, Spain. E-mail: car@us.es. Partially supported by grants MTM2006-07932 and MTM2010-15592 (DGES, Spain).
†Laboratoire de Mathématiques, Université Blaise Pascal (Clermont-Ferrand 2), UMR CNRS 6620, Campus des Cézeaux, 63177 Aubière, France. E-mail: arnaud.munch@math.univ-bpclermont.fr. Partially supported by grant MTM2010-15592 (DGES, Spain).
1 Introduction. The null controllability problem

We are mainly concerned in this work with the null controllability problem for the 1D heat PDE. The state equation is the following:

\[
\begin{aligned}
  &y_t - (a(x)y_x)_x + A(x, t) y = v 1_{\omega}, & (x, t) \in (0, 1) \times (0, T) \\
  &y(x, t) = 0, & (x, t) \in \{0, 1\} \times (0, T) \\
  &y(x, 0) = y_0(x), & x \in (0, 1).
\end{aligned}
\]  

(1)

Here, \( \omega \subset (0, 1) \) is a (small) non-empty open interval, \( 1_{\omega} \) is the associated characteristic function, \( T > 0, a \in L^\infty((0, 1) \times (0, T)) \) and \( y_0 \in L^2(0, 1) \).

In the sequel, for any \( T > 0 \), we will denote by \( Q_T, \Sigma_T \) and \( q_T \) the sets \( (0, 1) \times (0, T), \{0, 1\} \times (0, T) \) and \( \omega \times (0, T) \), respectively. We will also use the following notation:

\[
Ly := y_t - (a(x)y_x)_x + A(x, t) y, \quad L^* z := -z_t - (a(x)z_x)_x + A(x, t) z.
\]

For any \( y_0 \in L^2(0, 1) \) and \( v \in L^2(q_T) \), it is well-known that there exists exactly one solution \( y \) to (1), with

\[
y \in C^0([0, T]; L^2(0, 1)) \cap L^2(0, T; H^1_0(0, 1)).
\]

Accordingly, for any final time \( T > 0 \), the associated null controllability problem (at \( T \)) is the following: for each \( y_0 \in L^2(0, 1) \), find \( v \in L^2(q_T) \) such that the associated solution to (1) satisfies

\[
y(x, T) = 0, \quad x \in (0, 1).
\]  

(2)

The controllability of PDEs is an important area of research and has been the subject of many papers in recent years. Some relevant references are [11][12][33] and [1]. For heat equations, the existence of null controls was first obtained, by different methods, by Lebeau and Robbiano in [32] and Imanuvilov in [25], under appropriate assumptions on \( a \) and \( A \). In the particular case of the 1D heat equation (1) with \( A \equiv 0 \), this has been proved by Alessandrini and Escauriaza in [11] when the diffusion coefficient \( a \) “only” belongs to \( L^\infty(0, 1) \); for nonzero \( A \), the best known result is due to Le Rousseau [29] and holds for coefficients \( a \in BV(0, 1) \cap L^\infty(0, 1) \); see also [34] and [4] for other related results.

The control analysis of linear systems of the kind [11] may also be viewed as a first step for the controllability of other more complicate, maybe nonlinear, systems; for semilinear systems, the first contributions have been given in [41][28][11][25] and [20].

This paper is devoted to design and analyze efficient numerical methods for the previous null controllability problem.

The numerical approximation of null controls for (1) is a difficult issue. As shown below, this is mainly due to the strong regularization property of the heat kernel, that renders the numerical problem severely ill-posed.

So far, the approximation of the control of minimal \( L^2 \) norm has focused most of the attention. The first contribution was due to Carthel, Glowinski and Lions in [8], who made use of duality arguments. In this way, the original constrained minimization problem can be replaced by an unconstrained extremal (dual) problem, and \textit{a priori} easier to solve. However, the resulting problem involves some dual spaces which are very difficult (if not impossible) to approximate numerically.
More precisely, the null control of minimal norm in $L^2(q_T)$ is given by $v = \phi 1_\omega$, where $\phi$ solves the backward heat equation

\[
\begin{cases}
- \phi_t - (a(x)\phi_x)_x + A(x,t)\phi = 0, & (x,t) \in (0,1) \times (0,T) \\
\phi(x,t) = 0, & (x,t) \in \{0,1\} \times (0,T) \\
\phi(x,T) = \phi_T(x), & x \in (0,1)
\end{cases}
\]

and $\phi_T$ minimizes the strictly convex and coercive functional

\[
\mathcal{I}(\phi_T) = \frac{1}{2}\|\phi\|_{L^2(q_T)}^2 - (\phi(\cdot,0),y_0)_{L^2(0,1)}
\]

over the Hilbert space $\mathcal{H}$ defined by the completion of $L^2(0,1)$ with respect to the norm $\|\phi\|_{L^2(q_T)}$.

The coercivity of $\mathcal{I}$ in $\mathcal{H}$ is a consequence of the so-called observability inequality

\[
\|\phi(\cdot,0)\|_{L^2(0,1)}^2 \leq C \int_{Q_T} |\phi|^2 \, dx \, dt \quad \forall \phi_T \in L^2(0,1),
\]

that holds for some constant $C = C(\omega, T)$ and, in turn, this is a consequence of some appropriate global Carleman inequalities; see [20] and [16].

As discussed in length in [38] (see also [20] [35]), the minimization of $\mathcal{I}$ is numerically ill-posed, essentially because of the hugeness of $\mathcal{H}$. Notice that, in particular, $H^{-s}(0,1) \subset \mathcal{H}$ for any $s > 0$; see also [2], where the degree of ill-posedness is investigated in the boundary situation.

All this explains why in [8] the approximate controllability problem is considered and $\mathcal{I}$ is replaced by $\mathcal{I}_\epsilon$, where

\[
\mathcal{I}_\epsilon(\phi_T) := \mathcal{I}(\phi_T) + \epsilon\|\phi_T\|_{L^2(0,1)}
\]

for any $\epsilon > 0$. Now, the minimizer $\phi_{T,\epsilon}$ belongs to $L^2(0,1)$ and the corresponding control $v_\epsilon$ produces a state $y_\epsilon$ with $\|y_\epsilon(\cdot,T)\|_{L^2(0,1)} \leq \epsilon$. But, as $\epsilon \to 0^+$, high oscillations are observed for the controls $v_\epsilon$ near the controllability time $T$, see [38].

In this paper, we will consider the following extremal problem, introduced by Fursikov and Imnulov in [20]:

\[
\begin{cases}
\text{Minimize } J(y,v) = \frac{1}{2} \int_{Q_T} \rho^2 |y|^2 \, dx \, dt + \frac{1}{2} \int_{Q_T} \rho_0^2 |v|^2 \, dx \, dt \\
\text{Subject to } (y,v) \in \mathcal{C}(y_0,T).
\end{cases}
\]

Here, we denote by $\mathcal{C}(y_0,T)$ the linear manifold

\[
\mathcal{C}(y_0,T) = \{ (y,v) : v \in L^2(q_T), y \text{ solves } (1) \text{ and satisfies } (2) \}
\]

and we assume (at least) that

\[
\begin{cases}
\rho = \rho(x,t), \rho_0 = \rho_0(x,t) \text{ are continuous and } \rho_0 > 0 \text{ in } Q_T \\
\rho, \rho_0 \in L^\infty(Q_T - \delta) \forall \delta > 0
\end{cases}
\]

(hence, they can blow up as $t \to T^-$).

In order to find a solution to (6), we can apply methods of two kinds:

- Primal methods, that provide an optimal couple $(y,v)$ satisfying the constraint $(y,v) \in \mathcal{C}(y_0,T)$ and usually rely on the characterization of optimality. This will be the first family of methods considered in this paper; see Sections 3 and 4.
• Dual methods, in the spirit of the pioneering contribution of Carthe, Glowinski and Lions in [8] (see also [24]), that rely on appropriate reformulations of (6) as unconstrained problems and use new (dual) variables. Strategies of this kind will be also considered here, in Sections 5 and 6.

This paper is organized as follows.

In Section 2, we recall some results from [20] and we present the details of the variational approach to the null controllability problem. The optimal pair \( y \) and \( v \) are written in terms of a new function \( p \), the unique solution to a boundary value problem (17), that is second-order in time and fourth-order in space. Notice that \( p \) solves the variational formulation (15), that is well-posed when the weights \( \rho \) and \( \rho_0 \) are given by (10) and in particular blow up exponentially as \( t \to T^- \).

In Section 3, we analyze the numerical approximation of the variational formulation (22), that is obtained from (15) after a change of variables, see (18). The main advantage of (22) is that it involves no weight growing exponentially explicitly. The approximation makes use of a finite element that is \( C^1 \) in space and \( C^0 \) in time. We prove a convergence result as the discretization parameters go to zero and then we present some numerical experiments. Thanks to the change of variable, we observe that the condition number of the system depends polynomially on the discretization parameters. This is in contrast with the exponential behavior observed without this change of variables and in the literature, see [2, 38].

In order to avoid \( C^1 \) in space finite elements, we introduce in Section 4 the mixed variational formulation (58), which is equivalent to (22) and we prove well-posedness, see theorem 4.1. Some numerical experiments, based on a non conformal \( C^0 \) finite element, are discussed in Section 4.2 and highlight once again a polynomial dependance of the condition number.

In Section 5, we apply results from the Fenchel-Rockafellar duality theory to (6). To this end, we first introduce some approximations to (6) that lead to well-posed dual problems (see proposition 5.2). We also prove that the solutions to the latter converge, in an appropriate sense, to the solution to the original problem (6) (see propositions 5.1 and 5.3).

In Section 6, we apply gradient methods in this dual framework. More precisely, Section 6.1 is concerned with a conjugate gradient type algorithm, while Section 6.2 deals with the finite dimensional approximation of the control problems. In Section 6.3, we present several numerical experiments that show that the behavior of the considered algorithms is satisfactory.

Finally, some further comments, additional results and concluding remarks are given in Section 7.

2 A variational approach to the null controllability problem

In the sequel, unless otherwise specified, it will be assumed that

\[ a \in C^1([0, 1]), \quad a(x) \geq a_0 > 0 \quad \forall x \in [0, 1]. \]

Under this assumption, for any \( A \in L^\infty(Q_T) \), the linear system (1) is null-controllable. In other words, for any \( y_0 \in L^2(0, 1) \), there exist controls \( v \in L^2(q_T) \) such that the associated states \( y \) satisfy (2); see [20].

Let \( \rho \) and \( \rho_0 \) be functions satisfying (7) and let us consider the extremal problem (6). Then we have the following:

**Theorem 2.1** For any \( y_0 \in L^2(0, 1) \) and \( T > 0 \), there exists exactly one solution to (6).
The proof is simple. Indeed, from the null controllability of \([1], C(y_0, T)\) is non-empty. Furthermore, it is a closed convex set of \(L^2(Q_T) \times L^2(q_T)\); in fact, it is a closed linear manifold, whose supporting space is the set of all \((z, w)\) such that \(w \in L^2(q_T),
\[
\begin{aligned}
  z_t - (a(x)z_x)_x + A(x,t)z &= w 1_\omega, & (x,t) &\in (0, 1) \times (0,T) \\
  z(x, t) &= 0, & (x,t) &\in \{0,1\} \times (0, T) \\
  z(x,0) &= 0, & x &\in (0,1)
\end{aligned}
\]
and
\[z(x, T) = 0, \ x \in (0,1).\]
On the other hand, \((y,v) \mapsto J(y,v)\) is strictly convex, proper and lower semi-continuous on the space \(L^2(Q_T) \times L^2(q_T)\) and
\[J(y,v) \to +\infty \text{ as } \|y,v\|_{L^2(Q_T) \times L^2(q_T)} \to +\infty.\]
Hence, the extremal problem [10] certainly possesses a unique solution.

Since we are looking for controls such that the associated states satisfy [2], it is a good idea to choose weights \(\rho\) and \(\rho_0\) blowing up to \(+\infty\) as \(t \to T^-\). Indeed, this can be viewed as a reinforcement of the constraint [2]. In fact, by prescribing \(\rho\) and \(\rho_0\), we are even telling how fast \(y(\cdot, t)\) and \(v(\cdot, t)\) must decay to zero as \(t \to T^-\).

When [9] holds, there exist “good” weight functions \(\rho\) and \(\rho_0\) that blow up at \(t = T\) and provide a very suitable solution to the original null controllability problem. They were determined and systematically used by Fursikov and Imanuvilov and are the following:
\[
\begin{aligned}
  \rho(x, t) &= \exp \left( \frac{\beta(x)}{T-t} \right), \quad \rho_0(x, t) = (T-t)^{3/2} \rho(x, t), \quad \beta(x) = K_1 \left( e^{K_2} - e^{\beta_0(x)} \right)
\end{aligned}
\]
where the \(K_i\) are sufficiently large positive constants (depending on \(T, a_0\) and \(\|a\|_{C^1}\)) (10)
and \(\beta_0 \in C^\infty([0,1]), \beta_0 > 0\) in \((0,1), \beta_0(0) = \beta_0(1) = 0, |\beta_0| > 0\) outside \(\omega\).

The roles of \(\rho\) and \(\rho_0\) are clarified by the following arguments and results, which are mainly due to Fursikov and Imanuvilov [20]. First, let us set
\[P_0 = \{ q \in C^\infty(\overline{Q}_T) : q = 0 \text{ on } \Sigma_T \}.\]
In this linear space, the bilinear form
\[(p,q)_P := \iint_{Q_T} \rho^{-2} L^* p L^* q \, dx \, dt + \iint_{q_T} \rho_0^{-2} p q \, dx \, dt\]
is a scalar product. Indeed, if we have \(q \in P_0, L^* q = 0\) in \(Q_T\) and \(q = 0\) in \(q_T\), then, by the well known unique continuation property, we necessarily have \(q \equiv 0\).

Let \(P\) be the completion of \(P_0\) for this scalar product. Then \(P\) is a Hilbert space and the following results hold:

**Lemma 2.1** Assume that \(a\) satisfies [3] and \(\rho\) and \(\rho_0\) are given by [10]. Let us also set
\[
\rho_1(x,t) = (T-t)^{1/2} \rho(x,t), \quad \rho_2(x,t) = (T-t)^{-1/2} \rho(x,t).
\]
Then there exists \(C > 0\) only depending on \(\omega, T, a_0, \|a\|_{C^1}, \|A\|_{L^\infty}\), such that one has the following for all \(q \in P\):
\[
\begin{aligned}
  \iint_{Q_T} \rho_2^{-2} \left( |q_t|^2 + |q_x|^2 \right) + \rho_1^{-2} |q_x|^2 + \rho_0^{-2} |q|^2 \, dx \, dt \\
  &\leq C \left( \iint_{Q_T} \rho^{-2} |L^* q|^2 \, dx \, dt + \iint_{q_T} \rho_0^{-2} |q|^2 \, dx \, dt \right).
\end{aligned}
\]
The proof is given in [20]; see also [16].

**Lemma 2.2** Let the assumptions of lemma 2.1 hold. Then, for any $\delta > 0$, one has
\[ P \hookrightarrow C^0([0, T - \delta]; H^1_0(0, 1)), \]
where the embedding is continuous. In particular, there exists $C > 0$, only depending on $\omega$, $T$, $a_0$, $\|a\|_{C^1}$ and $\|A\|_{L^\infty}$, such that
\[ \|q(\cdot, 0)\|_{H^1_0(0, 1)}^2 \leq C \left( \iint_{Q_T} \rho^{-2} |L^* q|^2 \, dx \, dt + \iint_{q_T} \rho_0^{-2} |q|^2 \, dx \, dt \right) \] (13)
for all $q \in P$.

**Proof:** Let $\delta > 0$ be given and let us consider the Banach space $C^0([0, T - \delta]; L^2(0, 1))$. Let $q$ be given in $P$. Then, in view of lemma 2.1 and the fact that all the weights $\rho_i$ are bounded from above in $Q_{T-\delta}$, we see that $q, q_t, q_x, q_{xx} \in L^2(Q_{T-\delta})$, with norms in this space bounded by a constant times $\|q\|_P$.

In particular, $t \mapsto q(\cdot, t)$ and $t \mapsto q_t(\cdot, t)$, respectively regarded as a $H^2((0, 1))$-valued and a $L^2((0, T))$-valued function, are square-integrable. This implies that $t \mapsto q(\cdot, t)$, regarded as a $H^1_0((0, 1))$-valued function, is continuous on $[0, T)$.

**Proposition 2.1** Assume that $a$ satisfies (8) and let $\rho$ and $\rho_0$ be given by (10). Let $(y, v)$ be the corresponding optimal pair. Then there exists $p \in P$ such that
\[ y = \rho^{-2} L^* p, \quad v = -\rho_0^{-2} p|_{q_T}. \] (14)

The function $p$ is the unique solution to
\[ \begin{cases} \iint_{Q_T} \rho^{-2} L^* p L^* q \, dx \, dt + \iint_{q_T} \rho_0^{-2} p q \, dx \, dt = \int_0^1 y(x) q(x, 0) \, dx \\ \forall q \in P; \quad p \in P. \end{cases} \] (15)

**Proof:** In view of lemma 2.2 and the well known Lax-Milgram lemma, there exists exactly one solution $p$ to (15). Let us introduce $y$ and $v$ according to (14). We will check that $(y, v)$ solves (6); this will prove the result.

First, notice that $y \in L^2(Q_T)$ and $v \in L^2(q_T)$. Also, in view of (15), we have
\[ \begin{cases} \iint_{Q_T} y L^* q \, dx \, dt = \iint_{q_T} v q \, dx \, dt + \int_0^1 y(x) q(x, 0) \, dx \\ \forall q \in P; \quad y \in L^2(Q_T). \end{cases} \] (16)

But this means that $y$ is the solution to (1) in the transposition sense. Since $y_0 \in L^2(0, 1)$ and $v \in L^2(q_T)$, $y$ must coincide with the unique weak solution to (1). In particular, $y \in C^0([0, T]; L^2(0, 1))$ and, taking into account (14), we find that (2) holds. In other words, $(y, v) \in C(y_0, T)$.
Finally, let \((z, w) \in C(y_0, T)\) be such that \(J(z, w) < +\infty\). Then, it is immediate that

\[
J(z, w) \geq J(y, v) + \int\int_{Q_T} \rho^2 y (z - y) \, dx \, dt + \int\int_{Q_T} \rho_0^2 v (w - v) \, dx \, dt
\]

\[
= J(y, v) - \int\int_{Q_T} L^* p (z - y) \, dx \, dt - \int\int_{Q_T} p (w - v) \, dx \, dt
\]

\[
= J(y, v).
\]

Hence, \((y, v)\) solves (6). This ends the proof. 

\[\Box\]

**Remark 1** In this proposition, the regularity assumption on the diffusion coefficient \(a\) can be relaxed. Indeed, when \(a\) is piecewise \(C^1\) and satisfies the ellipticity hypothesis \(a \geq a_0 > 0\), it is also possible to construct weights \(\rho\) and \(\rho_0\) such that the previous results hold; see [3].

**Remark 2** In view of (14) and (15), it is clear that the function \(p\) furnished by proposition 2.1 solves, at least in the distributional sense, the following differential problem, that is second-order in time and fourth-order in space:

\[
\begin{align*}
L(\rho^{-2} L^* p) + \rho_0^{-2} p 1_{\omega} &= 0, & (x, t) \in (0, 1) \times (0, T) \\
p(x, t) &= 0, & (x, t) \in \{0, 1\} \times (0, T) \\
(\rho^{-2} L^* p)(x, 0) &= y_0(x), & (\rho^{-2} L^* p)(x, T) = 0, & x \in (0, 1).
\end{align*}
\]

Notice that the “boundary” conditions at \(t = 0\) and \(t = T\) are of the Neumann kind. Also, notice that no information is obtained on \(p(\cdot, T)\).

\[\Box\]

3 A first method: solving a variational equality

In the sequel, we will take \(\rho\) and \(\rho_0\) as in (10).

In account of proposition 2.1, a strategy to find the solution \((y, v)\) to (6) is to first solve (15) and then use (14); recall that (15) is the weak formulation of the differential problem (17).

We know from lemma 2.2 that the solution \(p\) to (15) belongs to \(C^0([0, T]; H^1_0(0, 1))\). However, for any \(s \geq 0\), there is no reason to have \(p(\cdot, t)\) bounded in \(H^{-s}(0, 1)\) as \(t \to T^-\). This means that it can be difficult to approximate with robustness the variational equality (15). It will appear very efficient, in terms of numerical robustness and stability, to perform a change of variable, so as to, somehow, we “normalize” the space \(P\).

3.1 An equivalent variational reformulation

The idea is to rewrite the variational equality (15) in terms of a new variable \(z\), given by

\[
z(x, t) = (T - t)^{-\alpha} \rho_0^{-1} (x, t) p(x, t), \quad (x, t) \in Q_T
\]

for some appropriate \(\alpha \geq 0\). We define \(Z\) as the completion of \(P_0\) for the scalar product

\[
(z, \overline{z})_Z := \int\int_{Q_T} \rho^{-2} L^* ((T - t)^{\alpha} \rho_0 z) L^* ((T - t)^{\alpha} \rho_0 \overline{z}) \, dx \, dt + \int\int_{Q_T} (T - t)^{2\alpha} z \overline{z} \, dx \, dt
\]

or, equivalently, we set

\[
Z = \{ (T - t)^{-\alpha} \rho_0 p : p \in P \}.
\]
We see that

$$\rho^{-1}L^*((T - t)^\alpha \rho_0 z) = A_1 z + A_2 z_t + A_3 z_x + A_4 z_{xx},$$

where the $A_i = A_i(x, t)$ satisfy:

$$\begin{cases}
A_1 = \left( \frac{3}{2} + \alpha \right)(T - t)^{\alpha + 1/2} - \beta(T - t)^{\alpha - 1/2} \\
- c \left( (T - t)^{\alpha + 1/2} \beta_{xx} + \beta_x^2(T - t)^{\alpha - 1/2} \right) + A(T - t)^{\alpha + 3/2},
\end{cases}$$

Consequently, the variational equality (22) can be rewritten as follows:

$$\begin{align*}
&\iint_{Q_T} (A_1 z + A_2 z_t + A_3 z_x + A_4 z_{xx})(A_1 \overline{\omega} + A_2 \overline{\omega}_t + A_3 \overline{\omega}_x + A_4 \overline{\omega}_{xx}) \, dx \, dt \\
&\quad + \iint_{Q_T} (T - t)^{2\alpha} \overline{\omega} \, dx \, dt = \int_0^1 y_0(x) \rho_0(x, 0) \overline{\omega}(x, 0) \, dx \\
&\quad \forall \overline{\omega} \in Z; \ z \in Z.
\end{align*}$$

The well-posedness of this formulation is an obvious consequence of the well-posedness of (15).

**Proposition 3.1** The variational equality (22) possesses exactly one solution $z \in Z$. Moreover, the unique solution $(y, v)$ to (6) is given by

$$y = \rho^{-1}(A_1 z + A_2 z_t + A_3 z_x + A_4 z_{xx}), \quad v = -(T - t)^{\alpha} \rho_0^{-1} z L,$$

where $z \in Z$ solves (22).

In order to have all the coefficients $A_i$ in $L^\infty(Q_T)$, it suffices to take $\alpha \geq 1/2$. Notice that, thanks to the previous change of variable, the functions $\rho$ and $\rho_0$ in $\rho^{-1}L^*p = \rho^{-1}L^*((T - t)^\alpha \rho_0 z)$ compensate each other, so that no exponential function appears anymore in (22).

### 3.2 Numerical analysis of the variational equality

Let us introduce the bilinear form $m(\cdot, \cdot)$, with

$$m(z, \overline{\omega}) := \iint_{Q_T} (A_1 z + A_2 z_t + A_3 z_x + A_4 z_{xx})(A_1 \overline{\omega} + A_2 \overline{\omega}_t + A_3 \overline{\omega}_x + A_4 \overline{\omega}_{xx}) \, dx \, dt$$

$$+ \iint_{Q_T} s^{2\alpha} \overline{\omega} \, dx \, dt$$

and the linear form $\ell$, with

$$\langle \ell, \overline{\omega} \rangle := \int_0^1 y_0(x) \rho_0(x, 0) \overline{\omega}(x, 0) \, dx \, dt.$$

Then (22) reads as follows:

$$m(z, \overline{\omega}) = \langle \ell, \overline{\omega} \rangle \quad \forall \overline{\omega} \in Z; \ z \in Z.$$
3.2.1 Finite dimensional approximation

For any finite dimensional space $Z_h \subset Z$, we can introduce the following problem:

$$m(z_h, \tau_h) = \langle \ell, \tau_h \rangle \quad \forall \tau_h \in Z_h; \quad z_h \in Z_h.$$  \hfill (27)

Obviously, (27) is well-posed. Furthermore, we have the following classical result:

**Lemma 3.1** Let $z \in Z$ be the unique solution to (26) and let $z_h \in Z_h$ be the unique solution to (27). We have

$$\|z - z_h\|_Z \leq \inf_{\tau_h \in Z_h} \|z - \tau_h\|_Z.$$  \hfill (28)

**Proof:** We write that

$$\|z_h - z\|_Z^2 = m(z_h - z, z_h - z) = m(z_h - z, z_h - \tau_h) + m(z_h - z, \tau_h - z).$$

The first term vanishes for all $\tau_h \in Z_h$. The second one is bounded by $\|z_h - z\|_Z\|\tau_h - z\|_Z$. So, we get

$$\|z - z_h\|_Z \leq \|z - \tau_h\|_Z \quad \forall \tau_h \in Z_h$$

and the result follows.

As usual, this result can be used to prove the convergence of $z_h$ towards $z$ as $h \to 0$ when the spaces $Z_h$ are chosen appropriately.

More precisely, assume that $\mathcal{H} \subset \mathbb{R}^d$ is a net (i.e. a generalized sequence) that converges to zero and $Z_h$ is as above for each $h \in \mathcal{H}$. Let us introduce the interpolation operators $\Pi_h : P_0 \to Z_h$ and let us assume that the finite dimensional spaces $Z_h$ are chosen such that

$$\|\Pi_h z - z\|_Z \to 0 \quad \text{as} \quad h \in \mathcal{H}, \ h \to 0, \quad \forall z \in P_0.$$  \hfill (29)

We then have:

**Proposition 3.2** Let $z \in Z$ be the solution to (26) and let $z_h \in Z_h$ be the solution to (27) for each $h \in \mathcal{H}$. Then

$$\|z - z_h\|_Z \to 0 \quad \text{as} \quad h \in \mathcal{H}, \ h \to 0.$$  \hfill (30)

**Proof:** Let us choose $\epsilon > 0$. From the density of $P_0$ in $Z$, there exists $z_\epsilon \in P_0$ such that $\|z - z_\epsilon\|_Z \leq \epsilon$. Therefore, from lemma 3.1 we find that

$$\|z - z_h\|_Z \leq \|z - \Pi_h z_\epsilon\|_Z$$

$$\quad \leq \|z - z_\epsilon\|_Z + \|z_\epsilon - \Pi_h z_\epsilon\|_Z$$

$$\quad \leq \epsilon + \|z_\epsilon - \Pi_h z_\epsilon\|_Z.$$  \hfill (31)

From (29), $\|z_\epsilon - \Pi_h z_\epsilon\|_Z$ goes to zero as $h \in \mathcal{H}, \ h \to 0$ and the result follows.

3.2.2 The finite dimensional spaces $Z_h$

We will now indicate which are the good spaces $Z_h$.

The spaces $Z_h$ have to be chosen so that $\rho^{-1}L^*((T - t)^\alpha \rho_0 z_h)$ belongs to $L^2(Q_T)$ for any $z_h \in Z_h$. This means that $z_h$ must possess first-order time derivatives and up to second-order
spatial derivatives in $L^2_{loc}(Q_T)$. Therefore, an approximation based on a standard triangulation of $Q_T$ requires spaces of functions that must be $C^0$ in $t$ and $C^1$ in $x$.

For large integers $N_x$ and $N_t$, we set $\Delta x = 1/N_x$, $\Delta t = T/N_t$ and $h = (\Delta x, \Delta t)$. We introduce the associated uniform triangulations $Q_h$, with $Q_T = \bigcup_{k \in \mathbb{Q}_h} K$ and we assume that $\{\mathbb{Q}_h\}$ is a regular family. Then, we introduce the space $Z_h$ as follows:

$$ Z_h = \{ z_h \in C^{1,0}_{x,t}(\bar{Q}_T) : z_h|_K \in \mathbb{P}(K) \, \forall K \in \mathbb{Q}_h, \, z_h = 0 \text{ on } \Sigma_T \}. $$

(32)

Here, $C^{1,0}_{x,t}(\bar{Q}_T)$ is the space of the functions in $C^0(\bar{Q}_T)$ that are continuously differentiable with respect to $x$ in $\bar{Q}_T$ and $\mathbb{P}(K)$ denotes the following space of polynomial functions in $x$ and $t$:

$$ \mathbb{P}(K) = (\mathbb{P}_{3,x} \otimes \mathbb{P}_{1,t})(K), $$

(33)

where $\mathbb{P}_{\ell,\xi}$ is the space of polynomial functions of order $\ell$ in the variable $\xi$.

Obviously, $Z_h$ is finite dimensional subspace of $Z$.

According to the specific geometry of $Q_T$, we shall analyze the situation for a uniform triangulation $Q_h$. Each element $K \in \mathbb{Q}_h$ is of the form

$$ K_{kl} = (x_k, x_{k+1}) \times (t_l, t_{l+1}), $$

with

$$ x_{k+1} = x_k + \Delta x, \quad t_{l+1} = t_l + \Delta t, \quad \text{for } k = 1, \ldots, N_x, \quad l = 1, \ldots, N_t. $$

It is then easy to see that a function $z_h \in \mathbb{P}(K_{kl})$ is uniquely determined by the real numbers $\{z_h(x_{k+m}, t_{l+n})\}$ and $\{(z_h)_x(x_{k+m}, t_{l+n})\}$, with $m, n = 0, 1$.

More precisely, let us introduce the functions

$$ L_{0k}(x) = \frac{(\Delta x + 2x - 2x_k)(\Delta x - x + x_k)^2}{(\Delta x)^3}, \quad L_{1k}(x) = \frac{(x - x_k)^2(-2x + 2x_k + 3\Delta x)}{(\Delta x)^3}, $$

$$ L_{2k}(x) = \frac{(x - x_k)(\Delta x - x + x_k)^2}{(\Delta x)^2}, \quad L_{3k}(x) = \frac{-(x - x_k)^2(h - x + x_k)}{(\Delta x)^2}, $$

(34)

and

$$ L_{0l}(t) = \frac{t_l - t + \Delta t}{\Delta t}, \quad L_{1l}(t) = \frac{t - t_l}{\Delta t}. $$

(35)

Then, the following result is not difficult to prove:

**Lemma 3.2** Let $u \in P_0$ and let us define the function $\Pi_h u$ as follows: on each $K_{kl} = (x_k, x_{k+1}) \times (t_l, t_{l+1})$, we set

$$ \Pi_h u(x, t) := \sum_{i,j=0}^1 L_{ik}(x) L_{jl}(t) u(x_{i+k}, t_{j+l}) + \sum_{i,j=0}^1 L_{i+k,2h}(x) L_{jl}(t) u_{x}(x_{i+k}, t_{j+l}). $$

(36)

Then $\Pi_h u$ is the unique function in $Z_h$ that satisfies

$$ \Pi_h u(x_k, t_l) = u(x_k, t_l), \quad (\Pi_h u(x_k, t_l))_x = u_x(x_k, t_l), \quad \forall k, l. $$

(37)

The linear mapping $\Pi_h : P_0 \to Z_h$ is by definition the interpolation operator associated to $Z_h$.

In the next Section, we will use the following result:
Lemma 3.3 For any $u \in P_0$ and any $(x, t) \in K_{kl}$, one has:

\[ u - \Pi_h u = \sum_{i,j=0}^1 m_{ij} u_x(x_{i+k}, t_{j+l}) + \sum_{i,j=0}^1 L_{ik} L_{jl} \mathcal{R}[u; x_{i+k}, t_{j+l}], \]  

(38)

where

\[ m_{ij}(x, t) := (L_{ik}(x)(x - x_i) - L_{i+2,k}(x)) L_j(t) \]

and

\[
\mathcal{R}[u; x_{i+k}, t_{j+l}](x, t) := \int_{t_{j+l}}^t u_t(x_{i+k}, s) ds + (x - x_{i+k}) \int_{t_{j+l}}^t (t - s) u_{xt}(x_{i+k}, s) ds \\
+ \int_{x_{i+k}}^x (x - s) u_{xx}(s, t) ds.
\]

Proof: The equality (38) is a consequence of the following Taylor expansion with integral remainder:

\[ u(x, t) = u(x_i, t_j) + (x - x_i) u_x(x_i, t_j) + \int_{t_j}^t u_t(x_i, s) ds \\
+ (x - x_i) \int_{t_j}^t (t - s) u_{xt}(x_i, s) ds + \int_{x_i}^x (x - s) u_{xx}(s, t) ds \]

(39)

and the fact that $\sum_{i,j=0}^1 L_{ik}(x) L_{jl}(t) = 1$. ☐

3.2.3 An estimate of $\|z - \Pi_h z\|_Z$ and some consequences

We will now prove that (29) holds.

Thus, let us fix $z \in P_0$ and let us first see that

\[
\iint_{Q_T} (T - t)^{2\alpha} |z - \Pi_h z|^2 \, dx \, dt \to 0 \quad \text{as} \quad \Delta x, \Delta t \to 0^+.
\]

(40)

For each $K_{kl} \in Q_h$ (denoted by $K$ in the sequel), we write:

\[
\iint_K ((T - t)^{2\alpha} |z - \Pi_h z|^2 \, dx \, dt \leq T^{2\alpha} \iint_K |z - \Pi_h z|^2 \, dx \, dt.
\]

(41)

Using lemma 3.3, we have:

\[
\iint_K |z - \Pi_h z|^2 \, dx \, dt = \iint_K \left( \sum_{i,j} m_{ij} z_x(x_i, t_j) + \sum_{i,j} L_{i} L_{j} \mathcal{R}[z; x_{i+k}, t_{j+l}] \right)^2 \, dx \, dt \\
\leq 8\|z_x\|_{L^\infty(K)}^2 \sum_{i,j} \iint_K |m_{ij}|^2 \, dx \, dt + 8 \sum_{i,j} \iint_K |L_{i} L_{j} \mathcal{R}[z; x_{i+k}, t_{j+l}]|^2 \, dx \, dt
\]

(42)

where we have omitted the indices $k$ and $l$.

Moreover,

\[
|\mathcal{R}[z; x_{i+k}, t_{j+l}]|^2 \leq 3|z_t(x_i, \cdot)|_{L^2(t_1, t_2)}^2 |t - t_j| + (x - x_i)^2 |t - t_j|^3 \|z_{xt}(x_i, \cdot)\|_{L^2(t_1, t_2)}^2 \\
+ |x - x_i|^3 \|z_{xx}(\cdot, t)\|_{L^2(x_i, x_2)}^2.
\]

(43)
Consequently, we get:
\[
\sum_{i,j} \int_K |L_i \mathcal{L}_j \mathcal{R}[z; x_{i+k}, t_{j+i}]|^2 \, dx \, dt \\
\leq 3 \sup_{x \in (x_1, x_2)} \|z_t(x, \cdot)\|_{L^2(x_1, x_2)}^2 \int_K |L_i(x) \mathcal{L}_j(t)|^2 |t - t_j| \, dx \, dt \\
+ \sup_{x \in (x_1, x_2)} \|z_{tx}(x, \cdot)\|_{L^2(x_1, x_2)}^2 \int_K |L_i(x) \mathcal{L}_j(t)|^2 |t - t_j|^3 (x - x_i)^2 \, dx \, dt \\
+ \sup_{t \in (t_1, t_2)} \|z_{xx}(\cdot, t)\|_{L^2(x_1, x_2)}^2 \int_K |L_i(x) \mathcal{L}_j(t)|^2 |x - x_i|^3 \, dx \, dt. 
\]

After some tedious computations, one finds that
\[
\sum_{i,j} \int_K |m_{ij}|^2 \, dx \, dt = \frac{8}{945} (\Delta x)^3 \Delta t \sum_{i,j} \int_K |L_i(x) \mathcal{L}_j(t)|^2 |t - t_j| \, dx \, dt = \frac{13}{105} \Delta x (\Delta t)^2; 
\]
\[
\sum_{i,j} \int_K |L_i(x) \mathcal{L}_j(t)|^2 |t - t_j|^3 (x - x_i)^2 \, dx \, dt = \frac{19}{9450} (\Delta x)^3 (\Delta t)^4, 
\]
\[
\sum_{i,j} \int_K |L_i(x) \mathcal{L}_j(t)|^2 |x - x_i|^3 \, dx \, dt = \frac{11}{630} (\Delta x)^4 \Delta t. 
\]

This leads to the estimate
\[
\int_K |z - \Pi_h z|^2 \, dx \, dt \leq \frac{64}{945} (\Delta x)^3 \Delta t \|z_x\|^2_{L^\infty(K)} \\
+ \frac{312}{105} \Delta x (\Delta t)^2 \sup_{x \in (x_1, x_2)} \|(r^{-1} p)_t(x, \cdot)\|^2_{L^2(x_1, x_2)} \\
+ \frac{152}{9450} (\Delta x)^3 (\Delta t)^4 \sup_{x \in (x_1, x_2)} \|z_{tx}(x, \cdot)\|^2_{L^2(x_1, x_2)} \\
+ \frac{88}{630} (\Delta x)^4 \Delta t \sup_{t \in (t_1, t_2)} \|z_{xx}(\cdot, t)\|^2_{L^2(x_1, x_2)}.
\]

We deduce that
\[
\int_{q_T} |z - \Pi_h z|^2 \, dx \, dt \leq K_1 |q_T| \|z_x\|^2_{L^\infty(q_T)} (\Delta x)^2 + K_2 |\omega| \|z_t\|^2_{L^2(0, T; L^\infty)} (\Delta t)^2 \\
+ K_3 |\omega| \|z_{tx}\|^2_{L^2(0, T; L^\infty)} (\Delta x)^2 (\Delta t)^4 \\
+ K_4 T \|z_{xx}\|^2_{L^2(0, T; L^2)} (\Delta x)^4
\]

for some finite positive constants $K_i$. Hence, for any $z \in P_0$ one has
\[
\int_{q_T} (T - t)^{2\alpha} |z - \Pi_h (z)|^2 \, dx \, dt \to 0 \quad \text{as} \quad \Delta x, \Delta t \to 0.
\]

On the other hand, taking $\alpha = 1/2$ (in order to get bounded coefficients $A_i$), after similar
computations, we get:

\[
\iint_{K} |A_1(z - \Pi_h z) + A_2(z - \Pi_h z)_t + A_3(z - \Pi_h z)_x + A_4(z - \Pi_h z)_{xx}|^2 \, dx \, dt \\
\leq 4\|A_1\|_{L^\infty(Q_T)} \iint_{K} |z - \Pi_h z|^2 \, dx \, dt \\
+ 4\|A_2\|_{L^\infty(Q_T)} \iint_{K} |(z - \Pi_h z)_t|^2 \, dx \, dt \\
+ 4\|A_3\|_{L^\infty(Q_T)} \iint_{K} |(z - \Pi_h z)_x|^2 \, dx \, dt \\
+ 4\|A_4\|_{L^\infty(Q_T)} \iint_{K} |(z - \Pi_h z)_{xx}|^2 \, dx \, dt
\]

and, proceeding as above, we see that all these quantities go to 0 as \( h = (\Delta x, \Delta t) \to (0,0) \) (it suffices to differentiate (38) with respect to \( t \) and \( x \); for more details, we refer the reader to \([18]\)).

This proves the convergence of \( z_h \) towards \( z \) in the space \( Z \), that is, (29).

Consequently, we have the following result:

**Proposition 3.3** Let \( z_h \in Z_h \) be the unique solution of (27) and let \( y_h, v_h \) be the functions defined by

\[
y_h = \rho^{-1}(A_1 z_h + A_2 z_{h,t} + A_3 z_{h,x} + A_4 z_{h,xx}), \quad v_h = -(T-t)^{\alpha} \rho_0^{-1} z_h 1_{\omega}.
\]

Then,

\[
\|v - v_h\|_{L^2(Q_T)} \to 0 \quad \text{and} \quad \|y - y_h\|_{L^2(Q_T)} \quad \text{as} \quad h \to 0.
\]

![Figure 1](image1.png)

**Figure 1**: **Left**: The function \( \beta_{0,s} \) on \((0,1)\) for \( s = 0.1, s = 0.25 \) and \( s = 0.5 \). **Right**: The non-constant \( C^1 \) diffusion coefficient \( a \) used in Section 4.2 (first case).

In order to take into account the numerical approximation of the weights and the data that we necessarily have to perform in practice, we will consider in the next section a third problem:

\[
m_h(\hat{z}_h, z_h) = \langle \ell_h, z_h \rangle \quad \forall z_h \in Z_h; \quad \hat{z}_h \in Z_h,
\]

(52)
where

\[ m_h(z_h, \pi_h) := \iint_{Q_T} ((\pi_h A_1)z_h + (\pi_h A_2)z_{h,t} + (\pi_h A_3)z_{h,x} + (\pi_h A_4)z_{h,xx}) \]

\[ + \iint_{Q_T} \pi_{\Delta t}(T - t)^{2\alpha}z_h \pi_h \, dx \, dt \]

(53)

and

\[ \langle \ell_h, \pi_h \rangle := T^\alpha \int_0^1 \pi_{\Delta x}(y_0 \rho_0(\cdot, 0))\pi_h(x, 0) \, dx. \]

(54)

Here, for any function \( f \in L^\infty(Q_T) \), \( \pi_h f \) denotes the piecewise linear function which coincides with \( f \) at all vertices of \( Q_h \). Similar (self-explanatory) meanings can be assigned to \( \pi_{\Delta x} f \) and \( \pi_{\Delta t} f \).

We refer the reader to \([18]\), where the strong convergence of \( \hat{\beta} \) is proved, together with some \textit{a priori} estimates, explicit in \( h \) and \( ||z||_Z \).

### 3.3 Numerical experiments (I)

We present now some numerical experiments concerning the solution of (52), which can in fact be viewed as a linear system involving a sparse, definite positive and symmetric matrix of order \( 2N_x N_t \).

We denote by \( M_h \) this matrix, so that \((z_h, \pi_h)_{Z_h} = (M_h \{ z_h \}, \{ \pi_h \})\). Once the variable \( z_h \) is known, the control \( v_h \) is given by

\[ v_h = -\pi_h((T - t)^\alpha \rho_0^{-1})z_h \omega. \]

(55)

The corresponding controlled state may be first obtained from (23). Then, this approximation \( y_h \) satisfies the controllability requirement (2) (that is, \( y_h(\cdot, T) = 0 \)), but not exactly the initial condition. Instead, in order to check the action of the control function \( v_h \), the approximation \( y_h \) may be obtained by solving (1) using a finite element method in space and time in the standard way.

Let us now define the function \( \beta_0 \) appearing in the weights \( \rho \) and \( \rho_0 \) and the coefficients \( A_i \) for \( 0 \leq i \leq 4 \), see (10) and (21). For any \( s \in (0, 1) \), we consider the function \( \beta_{0,s} \):

\[ \beta_{0,s}(x) = \frac{x(1 - x)e^{(x-c_s)^2}}{s(1-s)e^{(s-c_s)^2}}, \quad c_s = s \frac{1 - 2s}{2s(1-s)}. \]

(56)

If \( s \) belongs to \( \omega \), we easily check that \( \beta_{0,s} \) satisfies the conditions in (10). Indeed, notice that \( \beta_{0,s}(0) = \beta_{0,s}(1) = 0 \), \( \beta_{0,s} > 0 \) in \((0, 1)\) and \( |\beta'_{0,s}| > 0 \) except at \( x = s \). In the numerical experiments, we will take \( \rho \) and \( \rho_0 \) as in (10) with \( \beta_0 = \beta_{0,s} \), \( s \) being the middle point of \( \omega \), \( K_1 = 1/10 \) and \( K_2 = 2\|\beta_0\|_{L^\infty(0,1)} = 2 \).

The function \( \beta_{0,s} \) is plotted in Figure 1 (Left) for \( s = 1/10, 1/4 \) and \( s = 1/2 \). The weights \( \rho^{-2} \) and \( \rho_0^{-2} \) corresponding to \( s = 1/2 \) are displayed in Figure 2.

We use an exact integration method in order to compute the components of \( M_h \) and Gauss method to solve the corresponding linear system.

Let us consider a constant diffusion function \( a \equiv a_0 = 10^{-1} \) in \((0, 1)\). The initial state \( y_0 \) is the first eigenfunction of the Dirichlet-Laplacian, that is \( y_0(x) \equiv \sin(\pi x) \) and \( T = 1/2 \). We also take \( A \equiv 1 \) and \( \alpha = 1/2 \), so that all the coefficients appearing in the formulation belong to \( L^\infty(Q_T) \).

Tables 1 and 2 collect relevant numerical values for \( \omega = (0.2, 0.8) \) and \( \omega = (0.3, 0.6) \) respectively. For \( \omega = (0.2, 0.8) \), we take \( \beta_0 = \beta_{0,1/2} \). For \( \omega = (0.3, 0.6) \), we take \( \beta_0 = \beta_{0,0.45} \). Moreover, for simplicity, we always set \( \Delta x = \Delta t \).
observe that bounded with $\|v\|$ while for $K = 0.1$ and $K_2 = 2\|\beta_0\|_L^\infty\) of the size of $\omega$ the norm of the control is also emphasized. The absolute errors, displayed in the last two rows, are computed assuming that $h = (1/320, 1/320)$ provides a reference solution.

For $\omega = (0.2, 0.8)$, we observe that $\|y - y_h\|_{L^2(Q_T)} \approx O(h^{1.19})$ and $\|v - v_h\|_{L^2(Q_T)} \approx O(h^{1.25})$, while for $\omega = (0.3, 0.6)$ we observe a slightly slower convergence: $\|y - y_h\|_{L^2(Q_T)} \approx O(h^{0.95})$ and $\|v - v_h\|_{L^2(Q_T)} \approx O(h^{0.85})$.

We also check that the null controllability requirement (2) is very well satisfied: indeed, we observe that $\|y_h(\cdot, T) - y(\cdot, T)\|_{L^2(0,1)} \approx O(h^{1.97})$ and $\|y_h(\cdot, T) - y(\cdot, T)\|_{L^2(0,1)} \approx O(h^{1.65})$ for $\omega = (0.2, 0.8)$ and $\omega = (0.3, 0.6)$, respectively.

Notice that, as a consequence of the change of variable (18), the $L^2$ norm of $z_h(\cdot, T)$ remains bounded with $h$.

<table>
<thead>
<tr>
<th>$\Delta x, \Delta t$</th>
<th>$1/20$</th>
<th>$1/40$</th>
<th>$1/80$</th>
<th>$1/160$</th>
<th>$1/320$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa(M_h)$</td>
<td>$1.28 \times 10^7$</td>
<td>$1.00 \times 10^{11}$</td>
<td>$7.04 \times 10^{12}$</td>
<td>$4.71 \times 10^{14}$</td>
<td>$3.07 \times 10^{16}$</td>
</tr>
<tr>
<td>$|z_h|_{L^2(Q_T)}$</td>
<td>1.804</td>
<td>2.083</td>
<td>2.309</td>
<td>2.462</td>
<td>2.559</td>
</tr>
<tr>
<td>$|z_h(\cdot, T)|_{L^2(0,1)}$</td>
<td>$1.18 \times 10^{-1}$</td>
<td>$4.08 \times 10^{-2}$</td>
<td>$4.64 \times 10^{-3}$</td>
<td>$4.98 \times 10^{-3}$</td>
<td>$1.41 \times 10^{-3}$</td>
</tr>
<tr>
<td>$|v_h|_{L^2(Q_T)}$</td>
<td>0.97</td>
<td>1.002</td>
<td>1.023</td>
<td>1.035</td>
<td>1.041</td>
</tr>
<tr>
<td>$|y_h(\cdot, T)|_{L^2(0,1)}$</td>
<td>$2.01 \times 10^{-1}$</td>
<td>$1.998 \times 10^{-1}$</td>
<td>$1.990 \times 10^{-1}$</td>
<td>$1.986 \times 10^{-1}$</td>
<td>$1.984 \times 10^{-1}$</td>
</tr>
<tr>
<td>$|y - y_h|_{L^2(Q_T)}$</td>
<td>$1.13 \times 10^{-3}$</td>
<td>$3.00 \times 10^{-4}$</td>
<td>$7.59 \times 10^{-5}$</td>
<td>$1.89 \times 10^{-5}$</td>
<td>$4.74 \times 10^{-6}$</td>
</tr>
<tr>
<td>$|v - v_h|_{L^2(Q_T)}$</td>
<td>$6.47 \times 10^{-3}$</td>
<td>$3.52 \times 10^{-3}$</td>
<td>$1.59 \times 10^{-3}$</td>
<td>$3.53 \times 10^{-4}$</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: $\omega = (0.2, 0.8)$, $y_0(x) \equiv \sin(\pi x)$, $a(x) = a_0 \equiv 10^{-1}$ - $\alpha = 1/2$.

In Tables 3 and 4 we consider the case $\alpha = 0$, once again with $\omega = (0.2, 0.8)$ and $\omega = (0.3, 0.6)$. In this case, the coefficient $A_1$ is weakly singular, like $(T - t)^{-1/2}$; see (21). This is simply handled by numerically replacing $(T - t)$ by $(T - t + 10^{-10})$. The approximation $z_h$ is different here, but we check that the control function $v_h$ and the corresponding solution $y_h$ are independent of $\alpha$, so that the rates of convergence are very similar.

A relevant feature of these Tables is that they show that the condition number $\kappa(M_h)$ of the
A FIRST METHOD: SOLVING A VARIATIONAL EQUALITY

Table 2: $\omega = (0.3,0.6)$, $y_0(x) \equiv \sin(\pi x)$, $a(x) = a_0 \equiv 10^{-1} - \alpha = 1/2$.

<table>
<thead>
<tr>
<th>$\Delta x, \Delta t$</th>
<th>$1/20$</th>
<th>$1/40$</th>
<th>$1/80$</th>
<th>$1/160$</th>
<th>$1/320$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa(M_h)$</td>
<td>$1.36 \times 10^9$</td>
<td>$1.05 \times 10^{11}$</td>
<td>$6.87 \times 10^{12}$</td>
<td>$4.67 \times 10^{14}$</td>
<td>$3.07 \times 10^{16}$</td>
</tr>
<tr>
<td>$|z_h|_{L^2(Q_T)}$</td>
<td>$9.58$</td>
<td>$16.18$</td>
<td>$24.22$</td>
<td>$33.46$</td>
<td>$44.11$</td>
</tr>
<tr>
<td>$|z_h(\cdot,T)|_{L^2(0,1)}$</td>
<td>$6.84 \times 10^{-1}$</td>
<td>$1.90 \times 10^{-1}$</td>
<td>$1.92 \times 10^{-2}$</td>
<td>$8.05 \times 10^{-3}$</td>
<td>$1.63 \times 10^{-5}$</td>
</tr>
<tr>
<td>$|v_h|_{L^2(Q_T)}$</td>
<td>$1.596$</td>
<td>$2.005$</td>
<td>$2.334$</td>
<td>$2.571$</td>
<td>$2.729$</td>
</tr>
<tr>
<td>$|y_h|_{L^2(Q_T)}$</td>
<td>$1.881 \times 10^{-1}$</td>
<td>$1.837 \times 10^{-1}$</td>
<td>$1.827 \times 10^{-1}$</td>
<td>$1.827 \times 10^{-1}$</td>
<td>$1.829 \times 10^{-1}$</td>
</tr>
<tr>
<td>$|y_h(\cdot,T)|_{L^2(0,1)}$</td>
<td>$4.09 \times 10^{-3}$</td>
<td>$1.65 \times 10^{-3}$</td>
<td>$5.65 \times 10^{-4}$</td>
<td>$1.68 \times 10^{-4}$</td>
<td>$4.62 \times 10^{-5}$</td>
</tr>
<tr>
<td>$|y - y_h|_{L^2(Q_T)}$</td>
<td>$7.92 \times 10^{-2}$</td>
<td>$5.01 \times 10^{-2}$</td>
<td>$2.70 \times 10^{-2}$</td>
<td>$1.07 \times 10^{-2}$</td>
<td>-</td>
</tr>
<tr>
<td>$|v - v_h|_{L^2(Q_T)}$</td>
<td>$1.580$</td>
<td>$1.064$</td>
<td>$0.613$</td>
<td>$0.258$</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3: $\omega = (0.2,0.8)$, $y_0(x) \equiv \sin(\pi x)$, $a(x) = a_0 \equiv 10^{-1} - \alpha = 0$.

<table>
<thead>
<tr>
<th>$\Delta x, \Delta t$</th>
<th>$1/20$</th>
<th>$1/40$</th>
<th>$1/80$</th>
<th>$1/160$</th>
<th>$1/320$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa(M_h)$</td>
<td>$1.69 \times 10^{12}$</td>
<td>$5.05 \times 10^{13}$</td>
<td>$1.22 \times 10^{14}$</td>
<td>$2.88 \times 10^{15}$</td>
<td>$1.06 \times 10^{16}$</td>
</tr>
<tr>
<td>$|z_h|_{L^2(Q_T)}$</td>
<td>$1.005$</td>
<td>$1.113$</td>
<td>$1.191$</td>
<td>$1.239$</td>
<td>$1.266$</td>
</tr>
<tr>
<td>$|z_h(\cdot,T)|_{L^2(0,1)}$</td>
<td>$5.37 \times 10^{-10}$</td>
<td>$2.61 \times 10^{-10}$</td>
<td>$6.27 \times 10^{-11}$</td>
<td>$7.07 \times 10^{-12}$</td>
<td>$2.87 \times 10^{-13}$</td>
</tr>
<tr>
<td>$|v_h|_{L^2(Q_T)}$</td>
<td>$0.971$</td>
<td>$1.003$</td>
<td>$1.023$</td>
<td>$1.035$</td>
<td>$1.041$</td>
</tr>
<tr>
<td>$|y_h|_{L^2(Q_T)}$</td>
<td>$2.011 \times 10^{-1}$</td>
<td>$1.998 \times 10^{-1}$</td>
<td>$1.990 \times 10^{-1}$</td>
<td>$1.986 \times 10^{-1}$</td>
<td>$1.984 \times 10^{-1}$</td>
</tr>
<tr>
<td>$|y_h(\cdot,T)|_{L^2(0,1)}$</td>
<td>$6.17 \times 10^{-4}$</td>
<td>$1.56 \times 10^{-4}$</td>
<td>$3.83 \times 10^{-5}$</td>
<td>$9.44 \times 10^{-6}$</td>
<td>$2.35 \times 10^{-6}$</td>
</tr>
<tr>
<td>$|y - y_h|_{L^2(Q_T)}$</td>
<td>$6.27 \times 10^{-3}$</td>
<td>$3.43 \times 10^{-3}$</td>
<td>$1.56 \times 10^{-3}$</td>
<td>$5.28 \times 10^{-4}$</td>
<td>-</td>
</tr>
<tr>
<td>$|v - v_h|_{L^2(Q_T)}$</td>
<td>$1.36 \times 10^{-1}$</td>
<td>$7.26 \times 10^{-2}$</td>
<td>$3.25 \times 10^{-2}$</td>
<td>$1.09 \times 10^{-2}$</td>
<td>-</td>
</tr>
</tbody>
</table>

The computed state and control $y_h$ and $v_h$ for $h = (1/80,1/80)$ and $\omega = (0.3,0.6)$ are displayed in Figures 3 and 4.

From these results, we see that the finite dimensional formulation (52) provides an efficient and robust method to approximate null controls for the heat equation (1). Let us mention two drawbacks:

- First, for any fixed $h$, the controlled state computed by solving (1) numerically does not
### Table 4: \( \omega = (0.3, 0.6) \), \( y_0(x) \equiv \sin(\pi x) \), \( a(x) = a_0 \equiv 10^{-1} \) - \( \alpha = 0 \).

<table>
<thead>
<tr>
<th>( \Delta x, \Delta t )</th>
<th>1/20</th>
<th>1/40</th>
<th>1/80</th>
<th>1/160</th>
<th>1/320</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \kappa(\mathcal{M}_h) )</td>
<td>3.70 \times 10^{12}</td>
<td>1.12 \times 10^{13}</td>
<td>3.33 \times 10^{13}</td>
<td>1.01 \times 10^{14}</td>
<td>3.03 \times 10^{14}</td>
</tr>
<tr>
<td>( | z_h |_{L^2(Q_T)} )</td>
<td>4.664</td>
<td>7.725</td>
<td>10.98</td>
<td>14.30</td>
<td>17.63</td>
</tr>
<tr>
<td>( | z_h(\cdot, T) |_{L^2(0,1)} )</td>
<td>3.98 \times 10^{-9}</td>
<td>1.36 \times 10^{-9}</td>
<td>3.05 \times 10^{-10}</td>
<td>1.19 \times 10^{-11}</td>
<td>3.40 \times 10^{-13}</td>
</tr>
<tr>
<td>( | v_h |_{L^2(\Omega_T)} )</td>
<td>1.597</td>
<td>2.023</td>
<td>2.348</td>
<td>2.58</td>
<td>2.733</td>
</tr>
<tr>
<td>( | y_h |_{L^2(Q_T)} )</td>
<td>1.879 \times 10^{-1}</td>
<td>1.834 \times 10^{-1}</td>
<td>1.826 \times 10^{-1}</td>
<td>1.827 \times 10^{-1}</td>
<td>1.829 \times 10^{-1}</td>
</tr>
<tr>
<td>( | y - y_h |_{L^2(Q_T)} )</td>
<td>4.96 \times 10^{-3}</td>
<td>1.82 \times 10^{-3}</td>
<td>5.91 \times 10^{-4}</td>
<td>1.71 \times 10^{-4}</td>
<td>4.65 \times 10^{-5}</td>
</tr>
<tr>
<td>( | u - v_h |_{L^2(\Omega_T)} )</td>
<td>7.52 \times 10^{-2}</td>
<td>4.82 \times 10^{-2}</td>
<td>2.62 \times 10^{-2}</td>
<td>1.04 \times 10^{-2}</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 5: \( \omega = (0.3, 0.6) \) - First line: \( \kappa(\mathcal{M}_h) \) with \( \alpha = 0 \) - (\( z_h, \pi_h \)) \( z_h = (\mathcal{M}_h \{ z_h \}, \{ z_h \}) - \kappa(\mathcal{M}_h) \approx O(h^{-1.58}) \) - Second line: \( \kappa(\mathcal{M}_{1,h}) - (p_h, \bar{p}_h) p_h = (\mathcal{M}_{1,h} \{ p_h \}, \{ p_h \}) - \kappa(\mathcal{M}_{1,h}) \approx O(e^{h^{-0.87}}) \)

<table>
<thead>
<tr>
<th>( \Delta x, \Delta t )</th>
<th>1/20</th>
<th>1/40</th>
<th>1/80</th>
<th>1/160</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \kappa(\mathcal{M}_h) )</td>
<td>3.70 \times 10^{12}</td>
<td>1.12 \times 10^{13}</td>
<td>3.33 \times 10^{13}</td>
<td>1.01 \times 10^{14}</td>
</tr>
<tr>
<td>( \kappa(\mathcal{M}_{1,h}) )</td>
<td>3.52 \times 10^{15}</td>
<td>2.56 \times 10^{27}</td>
<td>2.13 \times 10^{50}</td>
<td>2.48 \times 10^{95}</td>
</tr>
</tbody>
</table>

### Table 6: \( \omega = (0.3, 0.6) \) and \( \alpha = 0 \) - Iterations numbers of the GMRES algorithm to solve the variational formulation (22) (Second line) and the variational formulation (15) (Third line).

<table>
<thead>
<tr>
<th>( \Delta x, \Delta t )</th>
<th>( 2N_x N_t )</th>
<th>#GMRES(( \mathcal{M}_h ))</th>
<th>#GMRES(( \mathcal{M}_{1,h} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \kappa(\mathcal{M}_h) )</td>
<td>462</td>
<td>385</td>
<td>411</td>
</tr>
<tr>
<td>( \kappa(\mathcal{M}_{1,h}) )</td>
<td>1722</td>
<td>1297</td>
<td>1501</td>
</tr>
<tr>
<td>#GMRES(( \mathcal{M}_h ))</td>
<td>6642</td>
<td>4141</td>
<td>5890</td>
</tr>
<tr>
<td>#GMRES(( \mathcal{M}_{1,h} ))</td>
<td>26082</td>
<td>7201</td>
<td>11567</td>
</tr>
</tbody>
</table>

Figure 3: \( \omega = (0.3, 0.6) \). The state \( y_h \).
satisfy exactly the null controllability condition at time \( t = T \); this is mainly explained by the fact that, in (17), this requirement appears as a Neumann condition; it is not the result of prescribing Dirichlet data.

• Secondly (and above all), the method requires a finite element approximation that must be \( C^1 \) in space. This is rather straightforward in the present 1D framework but, in higher dimension, involves the use of specific (and complex) finite elements; we refer to [9] for more details.

These two points are circumvented in the next section by introducing a mixed reformulation of (26).

4 A second method: solving a mixed variational formulation

Let us introduce the new variables \( m = \rho^{-1}L^*p \) and \( r = \rho_0^{-1}p \) and let us rewrite (15) in the form

\[
\begin{cases}
\iint_{Q_T} m \, \overline{m} \, dx \, dt + \iint_{Q_T} r \, \overline{r} \, dx \, dt = \int_0^1 \rho_0(x,0)y_0(x)\overline{r}(x,0) \, dx \\
\forall (\overline{m}, \overline{r}) \text{ with } \rho^{-1}L^*(\rho_0 \overline{r}) - m = 0 \text{ and } \overline{r} \in \rho_0^{-1}P; \quad \rho^{-1}L^*(\rho_0 r) - m = 0 \text{ and } r \in \rho_0^{-1}P.
\end{cases}
\] (57)

Let us introduce the spaces \( M = L^2(Q_T), \ R := \rho_0^{-1}P \) and \( \tilde{M} := (T-t)^{1/2} M \), the bilinear forms

\[
\begin{align*}
& a((m, r), (\overline{m}, \overline{r})) = \iint_{Q_T} m \, \overline{m} \, dx \, dt \ + \ \iint_{Q_T} r \, \overline{r} \, dx \, dt \quad \forall (m, r), (\overline{m}, \overline{r}) \in M \times R \\
& b((\overline{m}, \overline{r}), \mu) = \iint_{Q_T} (\rho^{-1}L^*(\rho_0 \overline{r}) - m) \, \overline{r} \, dx \, dt \quad \forall (\overline{m}, \overline{r}) \in M \times \tilde{M}.
\end{align*}
\]
and the linear form

\[ \langle \ell, (\overline{m}, \overline{r}) \rangle = \int_0^1 \rho_0(x, 0) y_0(x) \overline{\tau}(x, 0) \, dx \quad \forall (\overline{m}, \overline{r}) \in M \times R. \]

Then, it is not difficult to check that \( a(\cdot, \cdot), b(\cdot, \cdot) \) and \( \ell \) are well-defined and continuous and the announced mixed formulation is the following:

\[
\begin{aligned}
&\left\{ \begin{array}{ll}
\alpha((m, r), (\overline{m}, \overline{r})) + b((\overline{m}, \overline{r}), \lambda) = \langle \ell, (\overline{m}, \overline{r}) \rangle & \forall (\overline{m}, \overline{r}) \in M \times R \\
b((m, r), \mu) = 0 & \forall \mu \in \tilde{M}
\end{array} \right.
\end{aligned}
\] (58)

\((m, r) \in M \times R, \lambda \in \tilde{M}.

We can now state and prove an existence and uniqueness result:

**Theorem 4.1** There exists a unique solution \((m, r, \lambda)\) to (58). Moreover, \( y := \rho^{-1}m \) is, together with \( v := -\rho_0^{-1}r|_{QT} \), the unique solution to (6).

**Proof:** It is clear that, if \((m, r, \lambda)\) solves (58), then \( m = \rho^{-1}L^*(\rho_0 r), p = \rho_0 r \) is the unique solution to (15) and, consequently, the unique solution to (6) is given by \( y = \rho^{-1}m \) and \( v = -\rho_0^{-1}r|_{QT} \).

Let us introduce the space

\[
V = \{ (m, r) \in M \times R : b((m, r), \mu) = 0 \quad \forall \mu \in \tilde{M} \} = \{ (m, r) \in M \times R : m = \rho^{-1}L^*(\rho_0 r) \}.
\]

In order to prove that (58) possesses exactly one solution, we will apply a general result concerning mixed variational problems. More precisely, we will check that

- \( a(\cdot, \cdot) \) is coercive on \( V \), that is:
\[
a((\overline{m}, \overline{r}), (\overline{m}, \overline{r})) \geq \kappa_1 \| (\overline{m}, \overline{r}) \|_{M \times R}^2 \quad \forall (\overline{m}, \overline{r}) \in V, \quad \kappa_1 > 0.
\] (59)

- \( b(\cdot, \cdot) \) satisfies the usual “inf-sup” condition with respect to \( M \times R \) and \( \tilde{M} \), i.e.
\[
\kappa_2 := \inf_{\mu \in \tilde{M}} \sup_{(\overline{m}, \overline{r}) \in M \times R} \frac{b((\overline{m}, \overline{r}), \mu)}{\| (\overline{m}, \overline{r}) \|_{M \times R} \| \mu \|_{\tilde{M}}} > 0.
\] (60)

This will suffice to ensure existence and uniqueness; see for instance (7).

The proofs of (59) and (60) are straightforward. Indeed, we first notice that, for any \((\overline{m}, \overline{r}) \in V, \overline{m} = \rho^{-1}L^*(\rho_0 \overline{r})\) and thus

\[
a((\overline{m}, \overline{r}), (\overline{m}, \overline{r})) = \frac{1}{2} \iint_{QT} |\overline{m}|^2 \, dx \, dt + \frac{1}{2} \iint_{QT} \rho^{-2}|L^*(\rho_0 \overline{r})|^2 \, dx \, dt + \iint_{QT} |\overline{r}|^2 \, dx \, dt \geq \frac{1}{2} \| (\overline{m}, \overline{r}) \|_{M \times R}^2.
\]

This proves (59).

On the other hand, for any \( \mu \in \tilde{M} \) there exists \((\tilde{m}, \tilde{r}) \in M \times R\) such that

\[
b((\tilde{m}, \tilde{r}), \mu) = \iint_{QT} (T-t)^{-1}|\mu|^2 \, dx \, dt \quad \text{and} \quad \| (\tilde{m}, \tilde{r}) \|_{M \times R} \geq C \| \mu \|_{\tilde{M}}.
\]
For any $4.1$ A non-conformal mixed finite element approximation

In other words, we can replace (21), that we may easily cancel by replacing the multiplier $\rho L$ term $\rho L (\rho^{-1} r)$ in the bilinear form $b(\cdot, \cdot)$ possesses a singularity of the kind $(T - r)^{-1/2}$, see (20–21), that we may easily cancel by replacing the multiplier $\mu$ by $(T - r)^{\gamma} \mu$, for any $\gamma \geq 1/2$. In other words, we can replace $b(\cdot, \cdot)$ by $b_\gamma(\cdot, \cdot)$, with

\[
b_\gamma((\bar{m}, \bar{r}), \mu) := \int_{Q_T} (T - t)^\gamma (\rho^{-1} L^*(\rho_0 \bar{r}) - \bar{m}) \mu dx dt.
\]

The previous analysis remains essentially the same with the following new definition: $\bar{M} := (T - r)^{1/2 - \gamma} M$. Moreover, if we assume that $\mu_\gamma \in L^2(Q_T)$ and $\mu_{|\Sigma_T} = 0$, i.e. $\mu \in L^2(0, T; H^1_0(0, 1))$, we find that $b_\gamma((\bar{m}, \bar{r}), \mu)$ can be written in the form

\[
b_\gamma((\bar{m}, \bar{r}), \mu) = \int_{Q_T} (B_1 \mu + B_2 \bar{r} \mu_x + B_3 \bar{r}_x \mu + B_4 \bar{r}_t \mu - (T - t)^\gamma \bar{m}) \mu dx dt
\]

with the coefficients $B_i = B_i(x, t)$ given by

\[
B_1 = 3(T - t)^{-1/2} A; \quad B_2 = (T - t)^{-1/2} a \beta_x; \quad B_3 = (T - t)^{3/2} a; \quad B_4 = -(T - t)^{3/2}.
\]

\[\square\]

4.4 A non-conformal mixed finite element approximation

For any $h = (\Delta x, \Delta t)$ as before, let us consider again the associated uniform quadrangulation $Q_h$. We now introduce the following finite dimensional spaces:

\[
M_h = \{ z_h \in C^0(\overline{Q}_T) : z_h|_K \in (P_{1, x} \otimes P_{1, t})(K) \ \forall K \in Q_h \}, \tag{62}
\]

\[\square\]

In fact, one of these methods will be considered below, in Section 7 in the context of a problem similar to (68) in higher spatial dimension.
A SECOND METHOD: SOLVING A MIXED VARIATIONAL FORMULATION

Let the mixed finite element approximation of (58) be the following:

\[ C \text{ method. Once the triplet } (m, r, \lambda) \text{ and symmetric. Once again, its components, as well as those in the right hand side, are computed with (52), where only the variable } z \text{ is still missing. This will be one of the tasks in the next future. See some related comments in Section 7.} \]

We present in this Section some numerical experiments obtained by solving (64).

4.2 Numerical experiments (II)

In practice, the numerical experiments we have performed show that (64) possesses exactly one solution \((m_h, r_h, \lambda_h) \in M_h \times Q_h \times \tilde{M}_h\) for each \(h\) that is stable and converges appropriately as \(h \to 0^+\). However, a rigorous analysis of the existence, uniqueness and stability of \((m_h, r_h, \lambda_h)\) is still missing. This will be one of the tasks in the next future. See some related comments in Section 7.

The good property of this approach is that, thanks to the introduction of the new variables \(m_h\) and \(r_h\), in the integrals in (64) there is no weight growing exponentially as \(t \to T^-\). The worst behavior is found in the computation of \(\rho^{-1}(\rho_0 \tilde{r}_h)x\mu_h\), which behaves at most like \((T - t)^{-1/2}\), but this singularity is weak, numerically acceptable and can be easily removed (see Remark 3).

In practice, the numerical experiments we have performed show that (64) possesses exactly one solution \((m_h, r_h, \lambda_h) \in M_h \times Q_h \times \tilde{M}_h\) for each \(h\) that is stable and converges appropriately as \(h \to 0^+\). However, a rigorous analysis of the existence, uniqueness and stability of \((m_h, r_h, \lambda_h)\) is still missing. This will be one of the tasks in the next future. See some related comments in Section 7.

4.2 Numerical experiments (II)

We present in this Section some numerical experiments obtained by solving (64).

Three unknown functions, \(m_h, r_h\) and \(\mu_h\), are involved in the formulation (to be compared with (52), where only the variable \(z_h\) appears). Since the three unknowns are approximated by \(C^0\) finite elements, the order of the corresponding matrix is of order \(3N_x N_t\). The matrix is sparse and symmetric. Once again, its components, as well as those in the right hand side, are computed with exact integration formulæ. Moreover, as before, the linear system is solved using the Gauss method. Once the triplet \((m_h, r_h, \mu_h)\) is computed, the numerical solution \((y_h, v_h)\) is given directly by

\[ y_h = \pi_h(\rho^{-1})m_h, \quad v_h = -\pi_h(\rho_0^{-1})r_h, \quad (x,t) \in Q_T. \]

First, we consider again the data of Section 3.3, that is, \(y_0(x) \equiv \sin(\pi x), a \equiv a_0 = 1/10\) and \(T = 1/2\). We also take \(\gamma = 1/2\) (see Remark 3).

Tables 7 and 8 give the norms of \(v_h\) and \(y_h\) for various \(h\) and \(\omega = (0.2, 0.8)\) and \(\omega = (0.3, 0.6)\), respectively. The numerical values agree with those obtained with the previous method, see Tables 1 and 3. If we compare closer, the case \(\omega = (0.3, 0.6)\) suggests a faster convergence of the norms \(\|v_h\|_{L^2(Q_T)}\) and \(\|y_h\|_{L^2(Q_T)}\). Assuming again that \(h = (1/320, 1/320)\) provides a reference
solution, we see that \( \|v - v_h\|_{L^2(Q_T)} = \mathcal{O}(h^{1.23}) \) and \( \|y - y_h\|_{L^2(Q_T)} = \mathcal{O}(h^{1.23}) \) for \( \omega = (0.2, 0.8) \) and \( \|v - v_h\|_{L^2(Q_T)} = \mathcal{O}(h^{1.15}) \) and \( \|y - y_h\|_{L^2(Q_T)} = \mathcal{O}(h^{1.02}) \) for \( \omega = (0.3, 0.6) \). Very similar values are observed for \( \gamma = 0 \), for which the coefficient \( B_1 \) is weakly singular at \( t = T \).

The main difference observed with respect to the method described in Section 3 is the size of the condition numbers \( \kappa(M_h) \), which is significantly reduced. Once again, the \( \kappa(M_h) \) behave polynomially with respect to \( h \).

We also report in Table 10 the number of iterates leading to the convergence of GMRES versus \( h \). These values can be compared to those in Table 6.

Let us emphasize that, here, as a consequence of (65), the null controllability condition is exactly satisfied, that is, \( y_h(\cdot, T) = 0 \) on \( (0, 1) \).

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\Delta x, \Delta t & \frac{1}{20} & \frac{1}{40} & \frac{1}{80} & \frac{1}{160} & \frac{1}{320} \\
\kappa(M_h) & 1.47 \times 10^5 & 8.30 \times 10^5 & 6.48 \times 10^6 & 5.02 \times 10^7 & - \\
\|v_h\|_{L^2(Q_T)} & 0.974 & 1.006 & 1.025 & 1.036 & 1.041 \\
\|y_h\|_{L^2(Q_T)} & 2.001 \times 10^{-1} & 1.996 \times 10^{-1} & 1.989 \times 10^{-1} & 1.986 \times 10^{-1} & 1.984 \times 10^{-1} \\
\|y - y_h\|_{L^2(Q_T)} & 6.32 \times 10^{-3} & 3.21 \times 10^{-3} & 1.41 \times 10^{-3} & 4.75 \times 10^{-4} & - \\
\|v - v_h\|_{L^2(Q_T)} & 1.27 \times 10^{-2} & 6.56 \times 10^{-2} & 2.90 \times 10^{-2} & 9.72 \times 10^{-3} & - \\
\hline
\end{array}
\]

Table 7: \( \omega = (0.2, 0.8) \), \( y_0(x) = \sin(\pi x) \), \( a(x) = a_0 \equiv 10^{-1} \) - \( \gamma = 1/2 \).

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\Delta x, \Delta t & \frac{1}{20} & \frac{1}{40} & \frac{1}{80} & \frac{1}{160} & \frac{1}{320} \\
\kappa(M_h) & 1.39 \times 10^5 & 8.78 \times 10^5 & 6.62 \times 10^6 & 4.76 \times 10^7 & - \\
\|v_h\|_{L^2(Q_T)} & 1.865 & 2.339 & 2.651 & 2.830 & 2.936 \\
\|y_h\|_{L^2(Q_T)} & 1.836 \times 10^{-1} & 1.814 \times 10^{-1} & 1.817 \times 10^{-1} & 1.822 \times 10^{-1} & 1.826 \times 10^{-1} \\
\|y - y_h\|_{L^2(Q_T)} & 7.13 \times 10^{-2} & 3.82 \times 10^{-2} & 1.78 \times 10^{-2} & 6.33 \times 10^{-3} & - \\
\|v - v_h\|_{L^2(Q_T)} & 1.56 & 0.957 & 0.489 & 0.182 & - \\
\hline
\end{array}
\]

Table 8: \( \omega = (0.3, 0.6) \), \( y_0(x) = \sin(\pi x) \), \( a(x) = a_0 \equiv 10^{-1} \) - \( \gamma = 1/2 \).

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\Delta x, \Delta t & \frac{1}{20} & \frac{1}{40} & \frac{1}{80} & \frac{1}{160} & \frac{1}{320} \\
\kappa(M_h) & 4.18 \times 10^5 & 5.02 \times 10^5 & 6.04 \times 10^6 & 1.17 \times 10^6 & - \\
\|v_h\|_{L^2(Q_T)} & 0.976 & 1.007 & 1.026 & 1.036 & 1.041 \\
\|y_h\|_{L^2(Q_T)} & 2.008 \times 10^{-1} & 1.996 \times 10^{-1} & 1.989 \times 10^{-1} & 1.986 \times 10^{-1} & 1.984 \times 10^{-1} \\
\|y - y_h\|_{L^2(Q_T)} & 6.24 \times 10^{-3} & 3.17 \times 10^{-3} & 1.41 \times 10^{-3} & 4.73 \times 10^{-4} & - \\
\|v - v_h\|_{L^2(Q_T)} & 1.25 \times 10^{-1} & 6.48 \times 10^{-2} & 2.87 \times 10^{-2} & 9.67 \times 10^{-3} & - \\
\hline
\end{array}
\]

Table 9: \( \omega = (0.2, 0.8) \), \( y_0(x) = \sin(\pi x) \), \( a(x) = a_0 \equiv 10^{-1} \) - \( \gamma = 0 \).

As we have seen, the measure of the support \( |\omega| \) may affect the convergence of the approximation. Contrarily, due to the regularizing effect of the heat operator, the regularity of the initial condition has no impact in practice. More determinant are the norm (and the sign) of the potential \( A \), the size of the controllability time \( T \) and, of course, the size of the diffusion coefficient.

Let us consider a much more stiff situation that leads to larger variations of the control and the corresponding controlled solution.

The diffusion coefficient will be a non-constant \( C^1 \) function: we take \( D_1 = (0, 0.45) \), \( D_2 = (0.55, 1) \), \( a_1 = 1 \) and \( a_2 = 1/15 \) and we assume that \( a \) is the \( C^1 \) function that coincides with a
polynomial of the third order in \((0,1) \setminus (D_1 \cup D_2)\) and satisfies

\[
a(x) \equiv a_1 \quad \text{in} \quad D_1.
\]

In particular, \(\min(a_1, a_2) \leq a(x) \leq \max(a_1, a_2)\) in \((0,1)\); the function \(a\) is displayed in Figure 4. Right.

We take \(\omega = (0.2, 0.4)\) (where the diffusion is higher), \(\beta_0 = \beta_{0.0.3}\), \(T = 1/2\) and we localize \(y_0\) in \(D_2\), where the diffusion is low: \(y_0(x) = e^{-100(x-3/4)^2}1_{(0,1)}\). Finally, we take a negative potential \(A \equiv -1\). Of course, the effect of \(A\) is opposite to the effect of diffusion and therefore enhances the action of the control.

### Table 11: \(\omega = (0.2, 0.4)\) - Numerical norms for a stiff case. \(\gamma = 0\).

<table>
<thead>
<tr>
<th>(\Delta x, \Delta t)</th>
<th>1/40</th>
<th>1/80</th>
<th>1/160</th>
<th>1/320</th>
<th>1/640</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\kappa(M_h))</td>
<td>1.05 \times 10^8</td>
<td>8.28 \times 10^9</td>
<td>2.32 \times 10^{11}</td>
<td>7.06 \times 10^{13}</td>
<td>-</td>
</tr>
<tr>
<td>(3N_x N_t)</td>
<td>861</td>
<td>3321</td>
<td>13041</td>
<td>51681</td>
<td>205761</td>
</tr>
<tr>
<td>#GMRES((M_h))</td>
<td>631</td>
<td>2912</td>
<td>10211</td>
<td>33091</td>
<td>-</td>
</tr>
<tr>
<td>(|v_h|_{L^2(Q_T)})</td>
<td>14.44</td>
<td>19.70</td>
<td>23.48</td>
<td>25.41</td>
<td>26.01</td>
</tr>
<tr>
<td>(|y_h|_{L^2(Q_T)})</td>
<td>3.58 \times 10^{-1}</td>
<td>4.67 \times 10^{-1}</td>
<td>5.59 \times 10^{-1}</td>
<td>6.18 \times 10^{-1}</td>
<td>6.43 \times 10^{-1}</td>
</tr>
<tr>
<td>(|y - y_h|_{L^2(Q_T)})</td>
<td>6.30 \times 10^{-1}</td>
<td>3.21 \times 10^{-1}</td>
<td>9.15 \times 10^{-2}</td>
<td>4.76 \times 10^{-2}</td>
<td>-</td>
</tr>
<tr>
<td>(|v - v_h|<em>{L^2(Q_T)}/|v|</em>{L^2(Q_T)})</td>
<td>1.21</td>
<td>0.45</td>
<td>0.23</td>
<td>0.09</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 11 collects some numerical values. The numerical convergence as \(h \to 0\) is observed. Using now the solutions associated to \(h = (1/640, 1/640)\) as a reference solution, we see that \(\|v - v_h\| = O(h^{0.94})\) and \(\|y - y_h\| = O(h^{1.1})\).

In Figures 5 and 6, the corresponding controlled state \(y_h\) and null control \(v_h\) are displayed for \(h = (1/80, 1/80)\). The action of the control required to drive to rest the function \(y_h\) is here much stronger than in previous examples. In particular, \(\|v_h\|_{L^\infty(Q_T)} \approx 2.58 \times 10^2\). As a consequence, the function \(y_h\) takes relatively large values for \(t < T\): \(\|y_h\|_{L^\infty(Q_T)} \approx 3.24\), while \(\|y_0\|_{L^\infty(0,1)}\) is only equal to one.

This situation can be amplified for semilinear heat equations for which uncontrolled solutions may blow up in finite time; see [17] for more details.

Finally, let us mention the work [36] that also rely on a variational reformulation of the controllability and allow to obtain both boundary and inner controls.
Figure 5: $\omega = (0.2, 0.4) - y_0(x) = e^{-300(x-3/4)^2}$ The state $y_h$ on $Q_T$

Figure 6: $\omega = (0.2, 0.4) - y_0(x) = e^{-300(x-3/4)^2}$. The control $v_h$ on $(0.1, 0.5) \times (0, T)$
5 Duality and approximation

We now consider other methods, that rely on a different viewpoint. Roughly speaking, in accordance with the previous contribution of Carthel, Glowinski and Lions [8], we will try to use the Fenchel-Rockafellar duality approach to convex optimization (see [15, 40]; see also [12]) in order to formulate the null controllability problem as an unconstrained extremal problem with good properties.

The main reason for using duality in the context of (6) is that it is difficult to construct minimizing sequences; in fact, it is already difficult to construct couples \((y, v)\) in \(C(y_0, T)\).

However, it is not clear how to apply the Fenchel-Rockafellar techniques to (6) directly, mainly because \(\rho\) blows up as \(t \to T^-\); recall that the problem considered in [8] corresponds to \(\rho \equiv 0\) and \(\rho_0 \equiv 1\). Consequently, we will first work with well chosen approximations depending on appropriate parameters and, then, we will try to see what happens in the limit.

In some sense, the arguments we are going to present generalize those in [8] and [23].

For each \(R > 0\), we first consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad J_R(y, v) = \frac{1}{2} \int \int_{Q_T} \rho_R^2 |y|^2 \, dx \, dt + \frac{1}{2} \int \int_{Q_T} \rho_0^2 |v|^2 \, dx \, dt \\
\text{Subject to} & \quad (y, v) \in C(y_0, T).
\end{align*}
\]

Here, we have used the notation \(\rho_R = T_R(\rho) := \min(\rho, R)\). Notice that (66) is a new constrained extremal problem; again, it possesses exactly one solution \((y_R, v_R)\).

**Proposition 5.1** For any \(R > 0\), let \((y_R, v_R)\) be the unique minimizer of \(J_R\) in \(C(y_0, T)\) and let us denote by \((\hat{y}, \hat{v})\) the unique solution to (6). Then\(\]
\[v_R \to \hat{v} \quad \text{strongly in} \quad L^2(Q_T) \quad \text{and} \quad y_R \to \hat{y} \quad \text{strongly in} \quad L^2(Q_T) \quad \text{as} \quad R \to +\infty. \quad (67)\]

**Proof:** First, notice that

\[J_R(\hat{y}, \hat{v}) = \frac{1}{2} \int \int_{Q_T} \rho_R^2 |\hat{y}|^2 \, dx \, dt + \int \int_{Q_T} \rho_0^2 |\hat{v}|^2 \, dx \, dt \leq J(\hat{y}, \hat{v})\]

for all \(R > 0\). Consequently, the solutions to the problem (66) satisfy

\[J_R(y_R, v_R) = \frac{1}{2} \int \int_{Q_T} \rho_R^2 |y_R|^2 \, dx \, dt + \int \int_{Q_T} \rho_0^2 |v_R|^2 \, dx \, dt \leq J(\hat{y}, \hat{v}).\]

This shows that \(\rho_R y_R\) is uniformly bounded in \(L^2(Q_T)\) and \(\rho_0 v_R\) is uniformly bounded in \(L^2(Q_T)\). Therefore, at least for some subsequence one has

\[\rho_0 v_R \to w \quad \text{weakly in} \quad L^2(Q_T) \quad \text{and} \quad \rho_R y_R \to z \quad \text{weakly in} \quad L^2(Q_T). \quad (68)\]

Let us set \(\tilde{y} = \rho^{-1} z\) and \(\tilde{v} = \rho^{-1} w\). Then, it is clear from (68) that

\[v_R = \rho^{-1}_0 (\rho_0 v_R) \to \tilde{v} \quad \text{weakly in} \quad L^2(Q_T) \quad \text{and} \quad y_R = \rho^{-1}_R (\rho_R y_R) \to \tilde{y} \quad \text{weakly in} \quad L^2(Q_T).\]

In fact, \(\tilde{y}\) is the state associated to \(\tilde{v}\) and \(y_R\) converges strongly to \(\tilde{y}\).

For every \((y', v') \in C(y_0, T)\), one has

\[
\begin{align*}
J(\tilde{y}, \tilde{v}) & \leq \frac{1}{2} \lim \inf_{R \to +\infty} \left( \int \int_{Q_T} \rho_R^2 |y_R|^2 \, dx \, dt + \int \int_{Q_T} \rho_0^2 |v_R|^2 \, dx \, dt \right) \\
& \leq \frac{1}{2} \lim \inf_{R \to +\infty} \left( \int \int_{Q_T} \rho_R^2 |y'|^2 \, dx \, dt + \int \int_{Q_T} \rho_0^2 |v'|^2 \, dx \, dt \right) \\
& = J(y', v').
\end{align*}
\]

Therefore, \(\tilde{y}\) is an admissible minimizer of \(J\) on \(C(y_0, T)\).
Hence, \((\tilde{y}, \tilde{v}) = (\hat{y}, \hat{v})\). Finally, we also deduce from (68) that
\[
\limsup_{R \to +\infty} \left( \int_{Q_T} \rho_R^2 |y_R|^2 \, dx \, dt + \int_{Q_T} \rho_0^2 |v_R|^2 \, dx \, dt \right) \leq J(\tilde{y}, \tilde{v}),
\]
whence we see that (67) holds. \(\square\)

Once again, it is difficult to construct a minimizing sequence for (66). On the other hand, as shown below, the constraint \(y(\cdot, T) = 0\) is related to the existence of multipliers in a (very) large space, difficult to handle in practice.

For these reasons, it is also convenient to consider for any \(R > 0\) and \(\varepsilon > 0\) the following unconstrained and penalized problem:
\[
\begin{cases}
\text{Minimize} & J_{R, \varepsilon}(y, v) = \frac{1}{2} \int_{Q_T} \rho_R^2 |y|^2 \, dx \, dt + \frac{1}{2} \int_{Q_T} \rho_0^2 |v|^2 \, dx \, dt + \frac{1}{2\varepsilon} \|y(\cdot, T)\|_{L^2}^2 \\
\text{Subject to} & (y, v) \in \mathcal{A}(y_0, T)
\end{cases}
\]
(70)

where
\[
\mathcal{A}(y_0, T) = \{ (y, v) : v \in L^2(q_T), \ y \text{ solves (1)} \}.
\]

The analysis of problems (66) and (70) will provide useful indications on what to do in order to solve (6).

To this end, let us denote by \(\overline{y}\) the solution to (1) with \(v = 0\) and let us introduce the operators \(M \in \mathcal{L}(L^2(q_T); L^2(Q_T))\) and \(B \in \mathcal{L}(L^2(q_T); L^2(0, 1))\), with
\[
Mv = z_v, \quad Bv = z_v(\cdot, T)
\]
for all \(v \in L^2(q_T)\), where \(z_v\) is the solution to
\[
\begin{cases}
Lz = v1_\omega, & (x, t) \in Q_T \\
z(x, t) = 0, & (x, t) \in \Sigma_T, \quad z(x, 0) = 0, \quad x \in (0, 1)
\end{cases}
\]
(71)

Accordingly, the solution \(y\) of (1) can be decomposed in the form
\[
y = Mv + \overline{y}.
\]
(72)

Clearly, \(M\) and \(B\) are linear bounded operators on \(L^2(q_T)\). The adjoints \(M^*\) and \(B^*\) are given as follows:

- For each \(\mu \in L^2(Q_T)\), \(M^*\mu = \zeta|_{q_T}\), where \(\zeta\) is the solution to the backwards system
\[
\begin{cases}
L\zeta = \mu, & (x, t) \in Q_T \\
\zeta(x, t) = 0, & (x, t) \in \Sigma_T, \quad \zeta(x, T) = 0, \quad x \in (0, 1)
\end{cases}
\]
(73)

- For each \(\varphi_T \in L^2(0, 1)\), \(B^*\varphi_T = \varphi|_{q_T}\), where \(\varphi\) is the solution to
\[
\begin{cases}
L^*\varphi = 0, & (x, t) \in Q_T \\
\varphi(x, t) = 0, & (x, t) \in \Sigma_T, \quad \varphi(x, T) = \varphi_T(x), \quad x \in (0, 1)
\end{cases}
\]
(74)
Then we associate to (70) the following dual (well-posed) problem:

\[
\begin{aligned}
\text{Minimize} & \quad J_{R,\varepsilon}(\mu, \varphi_T) = \frac{1}{2} \left( \int_{Q_T} \rho_R^2 |\mu|^2 dx \, dt + \int_{Q_T} \rho_0^2 |\varphi|^2 dx \, dt \right) \\
& \quad + \int_0^1 \psi(x, 0) y_0(x) dx + \frac{\varepsilon}{2} \|\varphi_T\|^2_{L^2}
\end{aligned}
\]

Subject to \((\mu, \varphi_T) \in V\)

where, for any \((\mu, \varphi_T) \in L^2(Q_T) \times L^2(0, 1)\), we have set \(\psi = M^*\mu + B^*\varphi_T\), i.e. \(\psi\) is the solution to

\[
\begin{aligned}
L^* \psi &= \mu, \quad (x, t) \in Q_T \\
\psi &= 0, \quad (x, t) \in \Sigma_T, \quad \psi(x, T) = \varphi_T(x), \quad x \in (0, 1)
\end{aligned}
\]

In the following result, we explain how (75) is related to (70):

**Proposition 5.2** The unconstrained extremal problem (75) is dual to (70) in the sense of the Fenchel-Rockafellar theory. Furthermore, (75) is stable and possesses a unique solution. Finally, if we denote by \((y_{R,\varepsilon}, v_{R,\varepsilon})\) the unique solution to (70), we denote by \((\mu_{R,\varepsilon}, \varphi_{T,R,\varepsilon})\) the unique solution to (75) and we set \(\psi_{R,\varepsilon} = M^*\mu_{R,\varepsilon} + B^*\varphi_{T,R,\varepsilon}\), then the following relations hold:

\[
v_{R,\varepsilon} = \rho_0^2 \psi_{R,\varepsilon}|_{Q_T}, \quad y_{R,\varepsilon} = -\rho_R^2 \mu_{R,\varepsilon}, \quad \varphi_{R,\varepsilon}(\cdot, T) = -\varepsilon \varphi_{T,R,\varepsilon}.
\]

**Proof:** In view of the decomposition (72), we can write that \(J_{R,\varepsilon}(y, v) = F(Mv, Bv) + G(y)\) for any \((y, v) \in A(y_0, T)\). Here, we have introduced the functions \(F\) and \(G\), with

\[
F(z, z_T) = \frac{1}{2} \int_{Q_T} \rho_R^2 |z + y|^2 dx \, dt + \frac{1}{2\varepsilon} \int_0^1 |z_T(x) + \|\varphi(x, T)\|^2 dx
\]

and

\[
G(v) = \frac{1}{2} \int_{Q_T} \rho_0^2 |v|^2 dx dt.
\]

The functions \(F : L^2(Q_T) \times L^2(q_T) \mapsto \mathbb{R}\) and \(G : L^2(q_T) \mapsto \mathbb{R}\) are both convex and continuous and we can apply the duality Theorem of W. Fenchel and T.R. Rockafellar; see Theorem 4.2 p. 60 in [12]. We deduce that

\[
\inf_{A(y_0, T)} J_{R,\varepsilon}(y, v) = -\inf_{V} \left\{ G^*(M^*\mu + B^*\varphi_T) + F^*(-(\mu, \varphi_T)) \right\},
\]

where \(F^*\) and \(G^*\) are the convex conjugate of \(F\) and \(G\), respectively.

Notice that

\[
F^*(\mu, \varphi_T) = \sup_V \left\{ \int_{Q_T} \mu z dx \, dt + \int_0^1 \varphi_T(x) z_T(x) dx - F(z, z_T) \right\}
\]

\[
= \frac{1}{2} \int_{Q_T} \rho_R^2 |\mu|^2 dx \, dt - \int_{Q_T} \rho_0^2 |\varphi_T|^2 dx \, dt - \frac{\varepsilon}{2} \|\varphi_T\|^2_{L^2} - \int_0^1 \varphi_T(x) \|\varphi(x, T)\|^2 dx
\]

for all \((\mu, \varphi_T) \in V\). On the other hand,

\[
G^*(w) = \frac{1}{2} \int_{Q_T} \rho_0^2 |w|^2 dx \, dt
\]
for all \( w \in L^2(Q_T) \). Therefore,

\[
G^*(M^*\mu + B^*\varphi_T) + F^*(-(\mu, \varphi_T)) = \frac{1}{2} \int_{Q_T} \rho_R^{-2}|\mu|^2 \, dx \, dt + \frac{1}{2} \int_{Q_T} \rho_0^{-2}|\varphi|^2 \, dx \, dt + \frac{\varepsilon}{2} \|\varphi_T\|_{L^2}^2
\]

+ \int_{Q_T} \mu \, \bar{y} \, dx \, dt + \int_0^1 \varphi_T(x) \, \bar{y}(x, T) \, dx
\]

where we have used again the notation \( \psi = M^*\mu + B^*\varphi_T \).

Finally, multiplying the state equation of \( (70) \) by \( \bar{y} \) and integrating by parts, we obtain that

\[
\int_{Q_T} \mu \, \bar{y} \, dx \, dt + \int_0^1 \varphi_T(x) \, \bar{y}(x, T) \, dx = \int_0^1 \psi(x, 0) \, y_0(x) \, dx,
\]

whence

\[
G^*(M^*\mu + B^*\varphi_T) + F^*(-(\mu, \varphi_T)) = \frac{1}{2} \left( \int_{Q_T} \rho_R^{-2}|\mu|^2 \, dx \, dt + \int_{Q_T} \rho_0^{-2}|\varphi|^2 \, dx \, dt \right)
\]

+ \int_0^1 \varphi(x, 0) \, y_0(x) \, dx + \frac{\varepsilon}{2} \|\varphi_T\|_{L^2}^2.
\]

This proves that \( (75) \) is the dual of \( (70) \).

It is also easy to check that \( (70) \) and \( (75) \) are stable and possess unique solutions. Indeed, the hypotheses of Theorem 4.2 in [12] are satisfied for \( (70) \) (notice that this is not the case for \( (66) \), since the interior of the constraint set \( C(y_0, T) \) is empty).

Finally, let us deduce that the optimality conditions \( (77) \) hold.

Let us set \( (y, v) = (y_{R, \varepsilon}, v_{R, \varepsilon}) \) and \( (\mu, \varphi_T) = (\mu_{R, \varepsilon}, \varphi_{T, R, \varepsilon}) \). Then, since \( (75) \) and \( (70) \) are dual to each other, one has:

\[
0 = \frac{1}{2} \int_{Q_T} \rho_R^2|y|^2 \, dx \, dt + \frac{1}{2} \int_{Q_T} \rho_0^2|v|^2 \, dx \, dt + \frac{1}{2\varepsilon} \|y(\cdot, T)\|_{L^2}^2
\]

+ \frac{1}{2} \int_{Q_T} \rho_R^2|\mu|^2 \, dx \, dt + \frac{1}{2} \int_{Q_T} \rho_0^{-2}|\varphi|^2 \, dx \, dt + (\psi(\cdot, 0), y_0) + \frac{\varepsilon}{2} \|\varphi_T\|_{L^2}^2
\]

= \frac{1}{2} \int_{Q_T} \rho_R^2|y| + \rho_R^{-2}|\mu|^2 \, dx \, dt + \frac{1}{2} \int_{Q_T} \rho_0^2|v - \rho_0^{-2}\varphi|^2 \, dx \, dt + \frac{1}{2\varepsilon} \|y(\cdot, T) + \varepsilon\varphi_T\|_{L^2}^2
\]

- \int_{Q_T} \rho_R^2\mu \, dx \, dt + \int_{Q_T} \rho_0^2v \psi \, dx \, dt - (y(\cdot, T), \varphi_T)_{L^2} + (\psi(\cdot, 0), y_0)_{L^2}.
\]

But the terms in the last line cancel, since \( \psi = M^*\mu + B^*\varphi_T \). Consequently,

\[
\int_{Q_T} \rho_R^2|y| + \rho_R^{-2}|\mu|^2 \, dx \, dt + \int_{Q_T} \rho_0^2|v - \rho_0^{-2}\varphi|^2 \, dx \, dt + \frac{1}{\varepsilon} \|y(\cdot, T) + \varepsilon\varphi_T\|_{L^2}^2 = 0
\]

and we get \( (77) \).

We now justify the introduction of the parameter \( \varepsilon \) by analyzing the behavior of the solutions to the problems \( (70) \) as \( \varepsilon \to 0^+ \).

**Proposition 5.3** With the notation of proposition 5.2 for each fixed \( R > 0 \) one has

\[
v_{R, \varepsilon} \rightarrow v_R \text{ strongly in } L^2(q_T) \text{ and } y_{R, \varepsilon} \rightarrow y_R \text{ strongly in } L^2(Q_T) \text{ as } \varepsilon \to 0^+. \tag{78}
\]
5 DUALITY AND APPROXIMATION

PROOF: First, notice that, for each \( R > 0 \) and \( \varepsilon > 0 \), one has

\[
\iint_{Q_T} \rho_R^2 |y_{R,\varepsilon}|^2 \, dx \, dt + \iint_{q_T} \rho_0^{-2} |\psi_{R,\varepsilon}|^2 \, dx \, dt + \frac{1}{\varepsilon} \|y_{R,\varepsilon}(\cdot, T)\|_{L^2}^2 = (\psi_{R,\varepsilon}(\cdot, 0), y_0)_{L^2}. \tag{79}
\]

Indeed, taking into account the equations satisfied by \( y_{R,\varepsilon} \) and \( \psi_{R,\varepsilon} \) and the identities (77), we find that the sum of the two integrals in the left hand side of (79) is equal to

\[
\iint_{Q_T} (L^* \psi_{R,\varepsilon} y_{R,\varepsilon} - \psi_{R,\varepsilon} L y_{R,\varepsilon}) \, dx \, dt = - (\psi_{R,\varepsilon}(\cdot, t), y_{R,\varepsilon}(\cdot, t)_{L^2})_{t=0}^{t=T} = (\psi_{R,\varepsilon}(\cdot, 0), y_0)_{L^2} - \frac{1}{\varepsilon} \|y_{R,\varepsilon}(\cdot, T)\|_{L^2}.
\]

Now, from Lemma 2.2 applied to \( \psi_{R,\varepsilon} \), we deduce that the left hand side of (79) is uniformly bounded. Indeed, we have

\[
\iint_{Q_T} \rho_R^2 |y_{R,\varepsilon}|^2 \, dx \, dt + \iint_{q_T} \rho_0^{-2} |\psi_{R,\varepsilon}|^2 \, dx \, dt + \frac{1}{\varepsilon} \|y_{R,\varepsilon}(\cdot, T)\|_{L^2}^2 \\
\leq \|\psi_{R,\varepsilon}(\cdot, 0)\|_{L^2} \|y_0\|_{L^2} \\
\leq C \|y_0\|_{L^2} \left( \iint_{Q_T} \rho^{-2} \rho_R^1 |y_{R,\varepsilon}|^2 \, dx \, dt + \iint_{q_T} \rho_0^{-2} |\psi_{R,\varepsilon}|^2 \, dx \, dt \right)^{1/2} \\
\leq C \|y_0\|_{L^2} \left( \iint_{Q_T} \rho_R^2 |y_{R,\varepsilon}|^2 \, dx \, dt + \iint_{q_T} \rho_0^{-2} |\psi_{R,\varepsilon}|^2 \, dx \, dt \right)^{1/2}.
\]

Therefore, \( \rho_R y_{R,\varepsilon} \) is uniformly bounded in \( L^2(Q_T) \), \( \rho_0 v_{R,\varepsilon} = \rho_0^{-1} \psi_{R,\varepsilon} |_{q_T} \) is uniformly bounded in \( L^2(q_T) \), \( \|y_{R,\varepsilon}(\cdot, T)\|_{L^2} \leq C \varepsilon^{1/2} \) and, at least for some subsequence, one has

\[
\rho_R y_{R,\varepsilon} \to z_R = \rho_R \tilde{z}_R \text{ weakly in } L^2(Q_T) \quad \text{and} \quad \rho_0 v_{R,\varepsilon} \to w_R = \rho_0 \tilde{v}_R \text{ weakly in } L^2(q_T) \tag{80}
\]
as \( \varepsilon \to 0^+ \).

Obviously, \( \tilde{y}_R \) is the state associated to \( \tilde{v}_R \) and \( y_{R,\varepsilon} \) converges strongly to \( \tilde{y}_R \) in \( L^2(Q_T) \). Moreover, \( \tilde{y}(\cdot, T) = 0 \), that is, \( (\tilde{y}_R, \tilde{v}_R) \in C(y_0, T) \).

Now, arguing as in the proof of proposition 5.1 it is not difficult to check that \( (\tilde{y}_R, \tilde{v}_R) \) is the unique optimal pair of (66), i.e. \( (\tilde{y}_R, \tilde{v}_R) = (y_R, v_R) \) and \( v_{R,\varepsilon} \) also converges strongly. \( \square \)

**Remark 4** As a consequence of the way we have decided to penalize the constraint (2), \( J_{R,\varepsilon} \) is explicitly quadratic in \( \|\tilde{y}\|_{L^2(0,1)} \). In particular, this avoids the use of operator-splitting methods (see [23], Section 1.8.8). This does not affect the asymptotic limit in \( \varepsilon \), since from (79), one may show that the state \( y_{R,\varepsilon} \) associated to \( v_{R,\varepsilon} = \rho_0^{-2} \psi_{R,\varepsilon} \leq 1 \) satisfies

\[
\|y_{R,\varepsilon}(\cdot, T)\|_{L^2(0,1)} \leq C_R \varepsilon^{1/2} \|y_0\|_{L^2(0,1)}
\]

for some \( C_R > 0 \) that is uniformly bounded with respect to \( R \). \( \square \)

**Remark 5** There are other ways to apply duality techniques to (6). For instance, we can use the fact that, if the first integral in (6) is finite, then necessarily (2) is satisfied. This leads to the extremal problem

\[
\begin{align*}
\text{Minimize} \quad & \frac{1}{2} \iint_{q_T} \rho_0^{-2} |\zeta|^2 \, dx \, dt + \frac{1}{2} \iint_{q_T} \rho_0^{-2} |\mu|^2 \, dx \, dt + \int_0^1 \zeta(x, 0) y_0(x) \, dx \\
\text{Subject to} \quad & \mu \in L^2(Q_T)
\end{align*}
\tag{81}
\]

where, for each \( \mu \in L^2(Q_T) \), we have set \( \zeta = M^* \mu \); recall (73). However, this formulation is formal since the unique minimizer \( \mu \) may not belong to \( L^2(Q_T) \). \( \square \)
Remark 6 More interestingly, we can also get a dual problem to (70) where the unique (dual) variable is \( \varphi_T \). Indeed, using again that \( y = Mv + \overline{y} \), we can decompose \( J_{R,\varepsilon} \) as follows:

\[
J_{R,\varepsilon}(y,v) = F_1(v) + F_2(Bv),
\]

where

\[
F_1(v) = \frac{1}{2} \int_{Q_T} \rho_R^2 |Mv + \overline{y}|^2 \, dx \, dt + \frac{1}{2} \int_{Q_T} \rho_0^2 |v|^2 \, dx \, dt
\]

and

\[
F_2(Bv) = \frac{1}{2\varepsilon} \|Bv + \overline{y}(\cdot, T)\|^2,
\]

so that

\[
\inf_{L^2(\mathbb{R}_T)} \{ F_1(v) + F_2(Bv) \} = - \inf_{L^2(0,1)} \{ F_1^*(B^* \varphi_T) + F_2^*(-\varphi_T) \}.
\]

By introducing the mappings \( \mathcal{B} \) and \( \mathcal{A} \), with \( \mathcal{B} \varphi_T := B^* \varphi_T - M^*(\rho_R^2 \overline{y}) \) and \( \mathcal{A} := M^*(\rho_R^2 M) + \rho_0^2 1_\omega \), it is not difficult to check that

\[
F_1^*(B^* \varphi_T) = \frac{1}{2} \int_{Q_T} (\mathcal{A}^{-1} \mathcal{B}(\varphi_T)) \mathcal{B}(\varphi_T) \, dx \, dt - \frac{1}{2} \int_{Q_T} \rho_R^2 |\overline{y}|^2 \, dx \, dt
\]

and

\[
F_2^*(-\varphi_T) = \frac{1}{2 \varepsilon} \|\varphi_T\|^2 + \int_0^1 \varphi_T(x) \overline{y}(x,T) \, dx
\]

for all \( \varphi_T \in L^2(0,1) \). Consequently, an extremal problem that can be put in duality with (70) is the following:

\[
\begin{cases}
\text{Minimize} & \frac{1}{2} \int_{Q_T} (\mathcal{A}^{-1} \mathcal{B}(\varphi_T)) \mathcal{B}(\varphi_T) \, dx \, dt + \int_0^1 \varphi_T(x) \overline{y}(x,T) \, dx + \frac{1}{2 \varepsilon} \|\varphi_T\|^2 \\
\text{Subject to} & \varphi_T \in L^2(0,1).
\end{cases}
\]

When we start from a problem similar to (70) with the weights \( \rho \) and \( \rho_0 \) respectively replaced by 0 and 1, we find again a dual problem of this kind, with \( \mathcal{A} \) and \( \mathcal{B} \) respectively replaced by \( 1_\omega \) and \( B^* \). This is just the formulation considered in [8].

Problem (82) involves minimization only with respect to the variable \( \varphi_T \in L^2(0,1) \). However, it requires the inversion of a nonlocal operator \( \mathcal{A} \) and is therefore a priori harder to solve. \( \Box \)

In view of these convergence results, it seems that an appropriate way to solve (76) is to first find the solution to (75) and then apply the relations (77) for small \( \varepsilon \) and large \( R \). This is confirmed by the experiments in Section 6.3.

We also observe that problems (66) and (70) are close for small \( \varepsilon \). In fact, the experiments below will show that the parameter \( \varepsilon \) is in some sense useless, since the presence of the weighted integral of \( \mu \) in the functional of (75) suffices by itself to stabilize this extremal problem. The term in \( \varepsilon \) ensures that the second argument of the optimal pair \( (\mu_{R,\varepsilon}, \varphi_{T,R,\varepsilon}) \) belongs to \( L^2(0,1) \) and therefore allows to apply duality techniques in a rigorous way without using “abstract” or “nonstandard” spaces. Nevertheless, we can state that, in the limiting case, the analogous of the functional I (see (4)) is the functional

\[
J^*(\mu, \varphi_T) = \frac{1}{2} \int_{Q_T} \rho_R^{-2} |\psi|^2 \, dx \, dt + \frac{1}{2} \int_{Q_T} \rho_0^{-2} |\mu|^2 \, dx \, dt + \int_0^1 \psi(x,0) y_0(x) \, dx,
\]

(83)

to be minimized over the abstract space defined as the completion of \( \mathcal{D}(Q_T) \times \mathcal{D}(0,1) \) with respect to the norm \( (\int_{Q_T} \rho_R^{-2} |\psi|^2 \, dx \, dt + \int_{Q_T} \rho_0^{-2} |\mu|^2 \, dx \, dt)^{1/2} \).
6 Conjugate gradient and numerical approximation

In this section, we address the numerical solution to the minimization problem (75). Following [8], the method combines conjugate gradient algorithms with finite difference and finite element approximations.

The problem we want to solve reads as follows: for given \( \varepsilon, R > 0, y_0 \in L^2(0, 1) \) and \( T > 0 \), minimize over the Hilbert space \( V = L^2(Q_T) \times L^2(0, 1) \) the functional

\[
J_{R, \varepsilon}(\mu, \varphi_T) = \frac{1}{2} \int\int_{Q_T} \rho_0^{-2} |\psi|^2 \, dx \, dt + \frac{1}{2} \int\int_{Q_T} \rho_R^{-2} |\mu|^2 \, dx \, dt + \int_0^1 \psi(x, 0) y_0(x) \, dx + \frac{\varepsilon}{2} \|\varphi_T\|_{L^2}^2,
\]

where \( \psi = M^* \mu + B^* \varphi_T \), that is, \( \psi \) is the solution to (76). By definition, it will be said that \( \psi \) is the adjoint state associated to \( \mu \) and \( \varphi_T \).

Notice that, in view of the optimality condition (77), the optimal \( \mu_{R, \varepsilon} \) satisfies \( \mu_{R, \varepsilon} + \rho_R^2 y_{R, \varepsilon} = 0 \) and therefore must vanish on \( \Sigma_T \).

Our aim is to apply a conjugate gradient method. First of all, notice that the Fréchet derivative of \( J_{R, \varepsilon} \) at \( (\mu, \varphi_T) \) in the direction \( (\mu', \varphi_T') \) is given by

\[
DJ_{R, \varepsilon}(\mu, \varphi_T) \cdot (\mu', \varphi_T') = \int\int_{Q_T} (z + \rho_R^2 \mu) \mu' \, dx \, dt + \int_0^1 (z(x, T) + \varepsilon \varphi_T(x)) \varphi_T'(x) \, dx,
\]

where \( z \) is the unique solution to the following system:

\[
\begin{align*}
Lz &= \rho_0^{-2} \psi 1_\omega, \quad (x, t) \in Q_T \\
z(x, t) &= 0, \quad (x, t) \in \Sigma_T, \quad z(x, 0) = y_0(x), \quad x \in (0, 1)
\end{align*}
\]  \( 84 \)

and \( \psi = M^* \mu + B^* \varphi_T \) is the solution to (76).

Consequently, the gradient of \( J_{R, \varepsilon} \) at \( (\mu, \varphi_T) \) is \( (z + \rho_R^2 \mu, z(\cdot, T) + \varepsilon \varphi_T) \), where \( z \) is the solution to (84). This allows to apply a classical gradient (steepest descent) method:

\[
\begin{align*}
\langle \mu_0^0, \varphi_T^0 \rangle & \text{ is given in } V, \\
\langle \mu_{n+1}, \varphi_{T}^{n+1} \rangle &= \langle \mu^*, \varphi_T^n \rangle - \eta^n (z^n + \rho_R^2 \mu^n, z^n(\cdot, T) + \varepsilon \varphi_T^n), \quad n \geq 0.
\end{align*}
\]  \( 85 \)

Here, \( \eta^n \) is the minimizer of the function \( \eta \mapsto J_{R, \varepsilon}(\langle \mu^n, \varphi_T^n \rangle - \eta (z^n + \rho_R^2 \mu^n, z^n(\cdot, T) + \varepsilon \varphi_T^n)) \). However, due to the lack of uniform coercivity of \( J_{R, \varepsilon} \) with respect to \( R \) and \( \varepsilon \), this gradient method is not sufficiently robust and does not provide good results as soon as the discretization parameters are small enough.

6.1 The conjugate gradient algorithm

Let us introduce the following symmetric and continuous bilinear form on \( V \):

\[
a_{R, \varepsilon}((\mu, \varphi_T), (\mu', \varphi_T')) = \int\int_{Q_T} \rho_0^{-2} (M^* \mu + B^* \varphi_T)(M^* \mu' + B^* \varphi_T') \, dx \, dt + \int\int_{Q_T} \rho_R^{-2} \mu \mu' \, dx \, dt + \varepsilon \int_0^1 \varphi_T(x) \varphi_T'(x) \, dx, \quad \forall (\mu, \varphi_T), (\mu', \varphi_T') \in V.
\]

Then one has

\[
J_{R, \varepsilon}(\mu, \varphi_T) = \frac{1}{2} a_{R, \varepsilon}((\mu, \varphi_T), (\mu, \varphi_T)) + \int_0^1 (M^* \mu + B^* \varphi_T)(x, 0) y_0(x) \, dx, \quad \forall (\mu, \varphi_T) \in V.
\]
For any $\varepsilon$ and $R$, the bilinear form $a_{R,\varepsilon}(\cdot, \cdot)$ is coercive with respect to the norm of $V$. It is therefore appropriate to apply conjugate gradient methods to (75); see [24]. The Polak-Ribièrè version reads as follows:

**Step 0: Initialization**

Let $\sigma$ be a small and strictly positive real number.

We choose $(\mu^0, \varphi^n_T) \in V$ and we compute the gradient $g^0$ of $J_{R,\varepsilon}^*$ at $(\mu^0, \varphi^n_T)$. Notice that $g^0 = (g_1^0, g_2^0)$, where $g_1^0$ and $g_2^0$ are respectively given by

$$
\begin{align*}
g_1^0 &= z^0 + \rho_R^{-2} \mu^0, \\
g_2^0 &= z^0(\cdot, T) + \varepsilon \varphi^n_T
\end{align*}
$$

and $z^0$ solves, together with $\psi^0$, the cascade system

$$
\begin{cases}
L^* \psi^0 = \mu^0 & \text{in } Q_T, \\
L z^0 = \rho_0^{-2} \psi^0 1_\omega & \text{in } Q_T,
\end{cases}
\quad \psi^0 = 0 \text{ on } \Sigma_T, \quad \psi^0(\cdot, T) = \varphi^n_T
$$

If $\|g^0\|_V/\|(\mu^0, \varphi^n_T)\|_V \leq \sigma$, then we take $(\mu, \varphi_T) = (\mu^0, \varphi^n_T)$ and we stop; otherwise, we set $w^0 = (w^0_1, w^0_2) = g^0$.

Then, for $n \geq 0$, assuming that $(\mu^n, \varphi^n_T)$, $g^n$ and $w^n$ are given, with $g^n \neq 0$ and $w^n \neq 0$, we compute $(\mu^{n+1}, \varphi^{n+1}_T)$, $g^{n+1}$ and (if necessary) $w^{n+1}$ performing the following steps.

**Step 1: Steepest descent**

We set

$$
\eta^n = \frac{D J_{R,\varepsilon}^*(\mu^n, \varphi^n_T) \cdot w^n}{a_{R,\varepsilon}(w^n, w^n)}
$$

and we take

$$(\mu^{n+1}, \varphi^{n+1}_T) = (\mu^n, \varphi^n_T) - \eta^n w^n.$$

Then, we compute the gradient $g^{n+1}$ of $J_{R,\varepsilon}^*$ at $(\mu^{n+1}, \varphi^{n+1}_T)$. Now, $g^{n+1} = (g_1^{n+1}, g_2^{n+1})$, where $g_1^{n+1}$ and $g_2^{n+1}$ are respectively given by

$$
\begin{align*}
g_1^{n+1} &= z^{n+1} + \rho_R^{-2} \mu^{n+1}, \\
g_2^{n+1} &= z^{n+1}(\cdot, T) + \varepsilon \varphi^n_T
\end{align*}
$$

and $z^{n+1}$ is, together with $\psi^{n+1}$, the solution to the cascade system

$$
\begin{cases}
L^* \psi^{n+1} = \mu^{n+1} & \text{in } Q_T, \\
L z^{n+1} = \rho_0^{-2} \psi^{n+1} 1_\omega & \text{in } Q_T,
\end{cases}
\quad \psi^{n+1} = 0 \text{ on } \Sigma_T, \quad \psi^{n+1}(\cdot, T) = \varphi^{n+1}_T
$$

**Step 2: Convergence test and construction of the new direction**

If $\|g^{n+1}\|_V/\|g^n\|_V \leq \sigma$, then we take $(\mu, \varphi_T) = (\mu^{n+1}, \varphi^{n+1}_T)$ and we stop; otherwise, we compute

$$
\gamma_n = \frac{(g^{n+1} - g^n, g^{n+1})_V}{\|g^n\|_V^2},
$$

and we take

$$
w^{n+1} = g^{n+1} + \gamma_n w^n
$$

and we return to Step 1 with $n$ replaced by $n+1$. 


Remark 7 In the present quadratic-linear situation, by construction the gradients $g^n$ are conjugate to each other, that is, $(g^n,g^n)_V = 0$ for all $m,n \geq 0$, $m \neq n$. Consequently, the parameter $\gamma_n$ given by (86) can also be written in the form
\[
\gamma_n = \frac{\|g^{n+1}\|^2_V}{\|g^n\|^2_V}.
\] (87)

For non necessarily quadratic-linear extremal problems, the choices (86) and (87) are not equivalent; they respectively lead to the Polak-Ribiere and the Fletcher-Reeves conjugate gradient algorithms.

In our case, due to the numerical approximation, the orthogonality of the $g^n$ is lost and strongly accentuated for small values of $\varepsilon$ and large values of $R$. In that stiff case, the Polak-Ribiere version, mainly used in nonlinear situations, appears much more robust. \hfill \Box

Remark 8 With $\rho_R$ and $\rho_0$ respectively replaced by the constants 0 and 1, we obtain exactly the conjugate gradient algorithm considered in Section 1.8 in [24], designed for the computation of the control of minimal norm in $L^2(q_T)$. Notice that the present situation does not lead to a significative increase of the computational cost. \hfill \Box

6.2 Full discrete approximations

For “large” integers $N_x$ and $N_t$, we set, as in Section 3.2.2, $\Delta x = 1/N_x$, $\Delta t = T/N_t$ and $h = (\Delta x, \Delta t)$. Let us denote by $P_{\Delta x}$ the uniform partition of $[0,1]$ associated to $\Delta x$ and let us denote by $Q_h$ the uniform quadrangulation of $Q_T$ associated to $h$ so that in particular $Q_T = \bigcup_{K \in Q_h} K$.

The following (conformal) finite element approximation of $L^2(0,T;H^1(0,1))$ is introduced:
\[
X_h = \{ \varphi_h \in C^0([0,1] \times [0,T]) : \varphi_h|_K \in (P_{1,x} \otimes P_{1,t})(K) \ \forall K \in Q_h \}.
\]

Here again, $P_{m,\xi}$ denotes the space of polynomial functions of order $m$ in the variable $\xi$. Accordingly, the functions in $X_h$ reduce on each quadrangle $K \in Q_h$ to a linear polynomial in both $x$ and $t$. The space $X_h$ is a conformal approximation of $L^2(Q_T)$. We also consider the space
\[
X_{0h} = \{ \varphi_h \in X_h : \varphi_h(0,t) = \varphi_h(1,t) = 0 \ \forall t \in (0,T) \}.
\]

$X_{0h}$ is a finite-dimensional subspace of $L^2(0,T;H^1_0(0,1))$ and the functions $\varphi_h \in X_{0h}$ are uniquely determined by their values at the nodes $(x_j,t_j)$ of $Q_h$ such that $0 < x_j < 1$.

Let us now introduce other finite dimensional spaces. First, we set
\[
\Phi_{\Delta x} = \{ z \in C^0([0,1]) : z|_k \in P_{1,x}(k) \ \forall k \in P_{\Delta x} \}.
\]

Then, $\Phi_{\Delta x}$ is a finite dimensional subspace of $L^2(0,1)$ and the functions in $\Phi_{\Delta x}$ are uniquely determined by their values at the nodes of $P_{\Delta x}$.

Secondly, since the variable $\mu$ appears in the right hand side of the backward equation $L^*\psi = \mu$, it is natural to approximate $\mu \in L^2(Q_T)$ by a piecewise constant function. Thus, let $M_h$ be the space defined by
\[
M_h = \{ \mu_h \in L^2(Q_T) : \mu_h|_K \in (P_{0,x} \otimes P_{0,t})(K) \ \forall K \in Q_h \}.
\]

$M_h$ is a finite dimensional subspace of $L^2(Q_T)$ and the functions in $M_h$ are uniquely determined by their (constant) values on the quadrangles $K \in Q_h$.
For any $h$, we therefore consider the following approximation of $(75)$:

\[
\begin{aligned}
\text{Minimize } J_{R,\varepsilon,h}(\mu_h, \varphi_{\Delta x,T}) &= \frac{1}{2} \left( \int_{Q_T} \pi_h (\rho_R^{-2}) |\mu_h|^2 \, dx \, dt + \int_{Q_T} \pi_h (\rho_0^{-2}) |\psi_h|^2 \, dx \, dt \right) \\
&\quad + \int_0^1 \varphi_h(x,0) \pi_{\Delta x}(y_0(x)) \, dx + \frac{\varepsilon}{2} \| \varphi_{\Delta x,T} \|^2_{L^2(0,1)} \\
\text{Subject to } (\mu_h, \varphi_{\Delta x,T}) &\in M_h \times \Phi_{\Delta x}.
\end{aligned}
\]  

(88)

In (88), for every $\mu_h \in M_h$ and every $\varphi_{\Delta x,T} \in \Phi_{\Delta x}$, we have denoted by $\psi_h$ the associated discrete adjoint state. By definition, $\psi_h \in X_{0h}$ is given as follows:

(i) Let us introduce the times $t_j = j \Delta t$. We have $T = t_{N_t}$ and we first set $\psi_h|_{t=T} = \varphi_{\Delta x,T}$.

(ii) Secondly, $\psi_h|_{t=t_{N_t-1}}$ is the solution to the linear problem

\[
\begin{aligned}
\int_0^1 \frac{1}{\Delta t} (\Psi - \varphi_{\Delta x,T}) z \, dx + \frac{1}{2} \int_0^1 (\pi_{\Delta x} (a(x)) \Psi_x z_x + \pi_{\Delta x} A(x, t_{N_t-1}) \Psi z) \, dx \\
&\quad + \frac{1}{2} \int_0^1 (\pi_{\Delta x} (a(x)) \varphi_{\Delta x,T,T-x} z_x + \pi_{\Delta x} A(x, t_{N_t}) \varphi_{\Delta x,T,T-x} z) \, dx \\
&= \frac{1}{2} \int_0^1 (\mu_h(x, t_{N_t-1}) + \mu_h(x, t_{N_t})) z(x) \, dx \quad \forall z \in \Phi_{\Delta x}, \quad \Psi \in \Phi_{\Delta x}.
\end{aligned}
\]

(iii) Then, for given $n = N_t - 1, \ldots, 2$, $\Psi^* = \varphi_h|_{t=t_{n+1}}$ and $\overline{\Psi} = \varphi_h|_{t=t_n}$, $\varphi_h|_{t=t_{n-1}}$ is the solution to the linear problem

\[
\begin{aligned}
\int_0^1 \frac{1}{2\Delta t} (3\Psi^* - 4\overline{\Psi} + \Psi^*) z \, dx + \int_0^1 (\pi_{\Delta x} (a(x)) \Psi_x z_x + \pi_{\Delta x} (A(x, t_{n-1})) \Psi z) \, dx \\
= \int_0^1 \mu_h(x, t_{n-1}) z(x) \, dx \quad \forall z \in \Phi_{\Delta x}, \quad \Psi \in \Phi_{\Delta x}.
\end{aligned}
\]

We are thus using the two-step implicit Gear algorithm as a numerical tool to solve numerically the adjoint problem $(75)$. As advocated in [8], where the influence of the time discretization is highlighted, it has been observed that this second order scheme ensures a better behavior of the underlying conjugate gradient algorithm than, for instance, the implicit Euler scheme.

For the computation of the gradient of $J_{R,\varepsilon,h}$, we also need to solve numerically systems of the form (88). This is done in a similar way.

For any $R$ and $\varepsilon$, the functional $J_{R,\varepsilon,h}$ enjoys the same properties than $J_{*,\varepsilon}$ when $V$ is replaced by $V_h := X_h \times \Phi_{\Delta x}$. In particular $J_{R,\varepsilon,h}$ is coercive in $V_h$, uniformly with respect to $h$. Hence, (88) may be solved with the conjugate gradient algorithm stated in Section 6.1.

We do not present here any convergence result for the variables $\mu_h, \varphi_{\Delta x,T,R,\varepsilon}$ (the minimizer of $J_{R,\varepsilon,h}$) as $h \to 0$. Actually, only partial results have been obtained and concern the particular case of minimal $L^2$-minimal norm case, that is $\rho_0 = 1$ and $\rho = 0$. There, the main issue (in the limit case $\varepsilon = 0$ and $R = +\infty$) is to analyze the behavior of the constant arising at the discrete level in the observability inequality [5].

In this context, we mention the work of [34], where the null controllability for the heat equation with constant diffusion is proved for finite difference schemes in one spatial dimension on uniform meshes. In higher dimensions, discrete eigenfunctions may be an obstruction to the null controllability; see [45], where a counter-example for finite differences due to O.Kavian is described. A result of null controllability for a constant portion of the lower part of the discrete spectrum is
given in [5]. In [27], in the context of approximate controllability, a relaxed observability inequality is given for general semi-discrete (in space) schemes, with the parameter \( \varepsilon \) of the order of \( \Delta x \). The work [6] extends the results in [27] to the full discrete situation and proves the convergence of full discrete (approximated) controls toward a semi discrete one, as the time step \( \Delta t \) tends to zero. Let us also mention [13], where the authors prove that any controllable parabolic equation, be it discrete or continuous in space, is null controllable after time discretization upon the application of an appropriate filtering of the high frequencies.

To our knowledge, in the framework of duality, a convergence result similar to Proposition 3.3 for a sequence of discrete controls towards a null control of the infinite dimensional system 1 is still missing.

6.3 Numerical experiments (III)

We present in this Section some numerical experiments for problem (88). Using the same data as in Section 3.3, we first briefly discuss the behavior of the computed control with respect to \( \varepsilon, R \). Then, we analyze the influence of the weights \( \rho, \rho_0 \) on the behavior of the conjugate gradient method as \( h \to (0, 0) \). We also consider a change of variable similar to the one introduced in Section 3.1 and discuss its influence on the behavior of the algorithm.

6.3.1 Experiment 1: Behavior of \((v_{R, \varepsilon, h}, y_{R, \varepsilon, h})\) as \( R \to +\infty \) and \( \varepsilon \to 0 \)

For a fixed value of \( h \) sufficiently close to \((0, 0)\), we first illustrate the convergence of the numerical solution \( \rho_R^{-2} \mu_{R, \varepsilon} \) and \( \rho_0^{-2} \psi_{R, \varepsilon} \) for \( R \to +\infty \) and \( \varepsilon \to 0 \), in the sense stated in Propositions 5.1 and 5.3.

Once the unique minimizer \((\mu_{R, \varepsilon, h}, \varphi_{T, R, \varepsilon, h})\) of \( J_{R, \varepsilon, h}^\ast \) is obtained through the conjugate gradient algorithm described in Section 6.1, we compute the associated discrete adjoint solution \( \psi_{R, \varepsilon, h} \) using a \( P_1 \)-finite element method in space and a second order implicit scheme in time, as discussed in Section 6.2. The control is then given by \( v_{R, \varepsilon, h} = \rho_0^{-2} \psi_{R, \varepsilon, h} \). Finally, the controlled solution \( y_{R, \varepsilon, h} \) is given by \( y_{R, \varepsilon, h} = -\rho_R^{-2} \mu_{R, \varepsilon, h} \), in accordance with the optimality relations (77).

In order to compare the approaches, we fix the numerical data already used in Section 3.3, that is: \( a = a_0 = 1/10 \), \( \omega = (0.3, 0.6) \), \( A = 1, T = 1/2 \) and \( y_0(x) \equiv \sin(\pi x) \). Moreover, we take \( \Delta x = \Delta t \).

For \( h = (10^{-2}, 10^{-2}) \), we show in Table 12 the behavior of the norms of \( \mu_{R, \varepsilon, h} \) and \( \varphi_{T, R, \varepsilon, h} \) with respect to \( \varepsilon \) and \( R \). For each value of these parameters, we use the conjugate gradient algorithm with \( \sigma = 10^{-4} \). This is small enough to guarantee a good approximation of the control but, obviously, does not allow to fulfill exactly the optimality conditions (77). Notice however that the fact that \( \sigma \) (and \( h \)) is strictly positive allows to consider the limit cases \( R = +\infty \) (for which \( \rho_R^2 = \rho^2 \)) and \( \varepsilon = 0 \) as well.

The algorithm is initialized with \( \mu^0 \equiv 0 \) and \( \varphi_T^0 \equiv 0 \).

Table 12 reports \( \|\rho_0^{-2} \psi_{R, \varepsilon, h}\|_{L^2(Q_T)} \) and \( \|\rho_R^{-2} \mu_{R, \varepsilon, h}\|_{L^2(Q_T)} \) for \( \varepsilon \in \{10^{-4}, 10^{-6}, 10^{-8}, 0\} \) and \( R \in \{10^4, 10^6, 10^8, +\infty\} \). We check that these norms are uniformly bounded with respect to \( \varepsilon \) and \( R \) and both possess a limit as \( \varepsilon \to 0 \) and \( R \to \infty \), in agreement with propositions 5.1 and 5.3.

For small values of \( \varepsilon \) (near \( \varepsilon = 10^{-8} \)), we observe that the parameter \( R \) has only a weak influence on the norm of \( \rho_0^{-2} \psi_{R, \varepsilon, h} \); conversely, as soon as \( R \) is large enough (near \( R = 10^6 \)), the norm of \( \rho_R^{-2} \mu_{R, \varepsilon, h} \) is almost independent of \( \varepsilon \). This is due to the choice of the weights \( \rho \) and \( \rho_0 \) and that any small \( \varepsilon \) and any large \( R \) reinforce in a suitable sense the null controllability property.

Table 13 provides the number of iterates needed to achieve \( \|g_h^0\|_V/\|g_h^0\|_V \leq \sigma = 10^{-4} \), where \( g_h^0 \) is the gradient of \( J_{R, \varepsilon, h}^\ast \). In agreement with the results and conclusions in [8] and [38], this
number increases as $\varepsilon \to 0$ and/or $R \to +\infty$, which must be viewed as a numerical confirmation of the lack of uniform coercivity of $J_{R,\varepsilon}$ in $V$. On the other hand, as soon as $\sigma$ is small enough, depending on $a_0$, $T$ and the size of $\omega$, the conjugate algorithm fails to converge.

<table>
<thead>
<tr>
<th>$R$</th>
<th>$\varepsilon = 10^{-4}$</th>
<th>$\varepsilon = 10^{-6}$</th>
<th>$\varepsilon = 10^{-8}$</th>
<th>$\varepsilon = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^4$</td>
<td>4.57 $\times 10^1$</td>
<td>2.71 $\times 10^1$</td>
<td>2.23 $\times 10^1$</td>
<td>2.04 $\times 10^1$</td>
</tr>
<tr>
<td>$10^6$</td>
<td>3.36 $\times 10^2$</td>
<td>1.94 $\times 10^2$</td>
<td>2.95 $\times 10^2$</td>
<td>3.22 $\times 10^2$</td>
</tr>
<tr>
<td>$10^8$</td>
<td>4.40 $\times 10^2$</td>
<td>2.81 $\times 10^2$</td>
<td>3.64 $\times 10^2$</td>
<td>4.67 $\times 10^2$</td>
</tr>
<tr>
<td>$+\infty$</td>
<td>4.41 $\times 10^2$</td>
<td>2.82 $\times 10^2$</td>
<td>3.63 $\times 10^2$</td>
<td>4.60 $\times 10^2$</td>
</tr>
</tbody>
</table>

Table 14: $L^2(Q_T)$-norm of $\mu_{R,\varepsilon,h}$ (Top) and $L^2(0,1)$-norm of $\varphi_{T,R,\varepsilon,h}$ vs. $R$ and $\varepsilon$.

In agreement with the lack of uniform coercivity in $V$, the results in Tables indicate that $\mu_{R,\varepsilon,h}$ is not uniformly bounded in $L^2(Q_T)$ with respect to $R$ and $\varphi_{T,R,\varepsilon,h}$ is not uniformly bounded in $L^2(0,1)$ with respect to $\varepsilon$. Contrarily, we observe that the norm of $\mu_{R,\varepsilon,h}$ is bounded with respect to $\varepsilon$ and the norm of $\varphi_{T,R,\varepsilon}$ is bounded with respect to $R$ (Tables). This is due to the fact that, by definition of $J_{R,\varepsilon}$, the weight $\rho_R$ mainly acts on the variable $\mu_{R,\varepsilon,h}$ while $\varepsilon^{-1}$ mainly acts on $\varphi_{T,R,\varepsilon}$.

In the limit as $\varepsilon \to 0$, the $L^2$-norm of $\varphi_{T,R,\varepsilon}$, which can be viewed as a multiplier associated to the constraint $y(\cdot,T) = 0$, does not belong anymore to $L^2(0,1)$. This is what we observe when we
use the primal direct approach described in Section 3 and solve the formulation (15); as $h \to (0,0)$, we observe arbitrarily large values of the $L^2$-norm of $p_h(\cdot,T)$.

<table>
<thead>
<tr>
<th>$R$</th>
<th>$\varepsilon = 10^{-4}$</th>
<th>$\varepsilon = 10^{-6}$</th>
<th>$\varepsilon = 10^{-8}$</th>
<th>$\varepsilon = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R = 10^4$</td>
<td>$3.96 \times 10^{-3}$</td>
<td>$4.34 \times 10^{-4}$</td>
<td>$5.76 \times 10^{-5}$</td>
<td>$3.00 \times 10^{-5}$</td>
</tr>
<tr>
<td>$R = 10^6$</td>
<td>$6.19 \times 10^{-4}$</td>
<td>$3.07 \times 10^{-4}$</td>
<td>$4.62 \times 10^{-5}$</td>
<td>$3.09 \times 10^{-5}$</td>
</tr>
<tr>
<td>$R = 10^8$</td>
<td>$3.29 \times 10^{-4}$</td>
<td>$2.62 \times 10^{-4}$</td>
<td>$4.28 \times 10^{-5}$</td>
<td>$3.09 \times 10^{-5}$</td>
</tr>
<tr>
<td>$R = +\infty$</td>
<td>$3.27 \times 10^{-4}$</td>
<td>$2.62 \times 10^{-4}$</td>
<td>$4.32 \times 10^{-5}$</td>
<td>$3.10 \times 10^{-5}$</td>
</tr>
</tbody>
</table>

Table 15: $h = (10^{-2}, 10^{-2})$, $\omega = (0.3, 0.6)$, $y_0(x) \equiv \sin(\pi x)$. The $L^2(0,1)$-norm of $y_h(\cdot,T)$ vs. $R$ and $\varepsilon$.

Table 15 depicts the $L^2$-norm of the computed state at time $T$. Note that this solution satisfies $y_h(\cdot,0) = y_0$ and $a$ priori differs from the function $-\rho^2_{R,h} \mu R,\varepsilon,h$. As expected, the weight $\rho^2_R$ reinforces slightly the null controllability constraint (2) as $R$ increases.

These Tables suggest that it is not actually necessary to take $R = +\infty$ and $\varepsilon = 0$ to achieve a good approximation of the controls. Due to the weights, the norms of the computed controls and controlled solutions change only slightly with respect to these parameters. The singular case $R = +\infty$ and $\varepsilon = 0$ ensures a better approximation of the null controllability requirement, but leads to a significative increase of iterates, as the coercivity of $J^*_{R,\varepsilon}$ is lost.

6.3.2 Experiment 2: Influence of the weights on the algorithm

We now discuss with more depth the influence of the weights $\rho$ and $\rho_0$ on the behavior of the conjugate gradient algorithm. We take $R = +\infty$ and $\varepsilon = 0$. At the numerical level, this limit case still makes sense since, for any $h > 0$, the minimizer of $J^*_{+\infty,0,h}$ obtained via a conjugate gradient method depends on the stopping parameter $\sigma$ and does not actually satisfy the constraint $y_h(\cdot,T) = 0$ exactly. Note also that the numerical approximation we described in Section 6.2 remains consistent in that case, since the finite dimensional space $M_h \times \Phi_{\Delta x}$ is still a conformal approximation of the abstract space where $J^*_{+\infty,0}$ is coercive, namely the completion of $D(Q_T) \times D(0,1)$ for the norm $\| (\mu, \varphi_T) \| := (\int_{Q_T} \rho_0^2 \varphi^2 dx dt + \int_{Q_T} \rho^2 \mu^2 dx dt)^{1/2}$.

We use the same data as in the previous section, except that we begin with a larger domain control $\omega = (0.2, 0.8)$. This allows to reach gradients closer to zero, i.e. to prescribe smaller values of $\sigma$.

<table>
<thead>
<tr>
<th>$\Delta x, \Delta t$</th>
<th>1/40</th>
<th>1/80</th>
<th>1/160</th>
<th>1/320</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{#}{\text{CG iterates}}$</td>
<td>559</td>
<td>383</td>
<td>471</td>
<td>504</td>
</tr>
<tr>
<td>$| u_h |_{L^2(Q_T)}$</td>
<td>$9.89 \times 10^{-1}$</td>
<td>$1.006 \times 10^{-1}$</td>
<td>$1.015 \times 10^{-1}$</td>
<td>$1.021 \times 10^{-1}$</td>
</tr>
<tr>
<td>$| y_h |_{L^2(Q_T)}$</td>
<td>$2.01 \times 10^{-1}$</td>
<td>$2.004 \times 10^{-1}$</td>
<td>$1.999 \times 10^{-1}$</td>
<td>$1.996 \times 10^{-1}$</td>
</tr>
<tr>
<td>$| \mu_h |_{L^2(Q_T)}$</td>
<td>$9.207$</td>
<td>$9.293$</td>
<td>$13.29$</td>
<td>$18.99$</td>
</tr>
<tr>
<td>$| \varphi_{T,h} |_{L^2(0,1)}$</td>
<td>$3.81 \times 10^2$</td>
<td>$3.83 \times 10^2$</td>
<td>$3.94 \times 10^2$</td>
<td>$3.77 \times 10^1$</td>
</tr>
<tr>
<td>$| y_h(\cdot,T) |_{L^2(0,1)}$</td>
<td>$2.24 \times 10^{-5}$</td>
<td>$2.80 \times 10^{-5}$</td>
<td>$3.01 \times 10^{-5}$</td>
<td>$3.00 \times 10^{-5}$</td>
</tr>
</tbody>
</table>

Table 16: $\omega = (0.2, 0.8) - \sigma = 10^{-4}$.

In Tables 16 and 17 we collect some relevant results obtained respectively for $\sigma = 10^{-4}$ and $\sigma = 10^{-5}$. The conjugate gradient method is initialized with $\mu^0 = 0$ and $\varphi^0_T = 0$. The behavior of the method is shown for various $h = (\Delta x, \Delta t)$. In particular, the convergence of the
control \( v_h \) as well the as the state \( y_h \) as \( h \to (0, 0) \) becomes clear. These numerical results are very similar to those obtained with the primal direct method, see Tables 1 and 7.

These control functions approximate in a satisfactory way the null controllability requirement: we obtain \( \| y_h(\cdot, T) \|_{L^2(\Omega)} \) of the order of \( 10^{-5} \) and \( 10^{-6} \) for \( \sigma = 10^{-5} \) and \( \sigma = 10^{-6} \), respectively. The number of iterates increases when \( \sigma \) is reduced, but we observe that this number is weakly dependent of the discretization parameter \( h \). For \( \sigma = 10^{-5} \) and \( h = 1/160 \), the evolution in \( \log_{10} \) scale of the relative residue \( r^n_h = \| g^n_h \|_V / \| g^0_h \|_V \) is displayed in Figure 7. The evolution of the residues is nonlinear with respect to the iterates, as is usual for ill-posed parabolic problems, see for instance [38, 39]. Precisely, the slope reduces significantly after the first iterations and may even vanish for \( h \) too close to \( (0, 0) \).

\[
\begin{array}{|c|c|c|c|}
\hline
\Delta x, \Delta t & 1/80 & 1/160 & 1/320 \\
\hline
\text{CG iterates} & 3762 & 3620 & 3465 \\
\|v_h\|_{L^2(\Omega_T)} & 1.016 \times 10^{-1} & 1.027 \times 10^{-1} & 1.032 \times 10^{-1} \\
\|y_h\|_{L^2(\Omega_T)} & 1.997 \times 10^{-1} & 1.992 \times 10^{-1} & 1.990 \times 10^{-1} \\
\|\mu_h\|_{L^2(\Omega)} & 4.66 \times 10^{-1} & 5.99 \times 10^{-1} & 7.66 \times 10^{-1} \\
\|\psi_{T,h}\|_{L^2(\Omega)} & 1.05 \times 10^2 & 1.74 \times 10^2 & 1.53 \times 10^2 \\
\|y_h(\cdot, T)\|_{L^2(\Omega_T)} & 2.84 \times 10^{-6} & 3.14 \times 10^{-6} & 3.19 \times 10^{-6} \\
\hline
\end{array}
\]

Table 17: \( \omega = (0.2, 0.8) - \sigma = 10^{-5} \).

The first iterates are devoted to compute the lower frequencies of the unknowns \( \mu_h \) and \( \psi_{T,h} \): according to the regularizing effect of the operator \( L^* \), these low frequencies correspond for the backward solution \( \psi_h \) to the points \((x, t) \in Q_T\) far enough from \( t = T \). As we can see from Figure 7, this computation is achieved after a small number of iterates, almost independent of \( h \). The remaining iterates are devoted to compute the high frequencies of the unknowns \( \mu_h \) and \( \psi_{T,h} \), unavoidable and harder to capture. For \( \psi_h \), this corresponds to a neighborhood of \( t = T \), say \( (T - \delta, T] \) for some \( \delta > 0 \). This phenomenon, once again usual for ill-posed parabolic situations, is amplified by the behavior of the weights \( \rho^{-1} \) and \( \rho_{0}^{-1} \) near \( t = T \). More precisely, since \( \rho^{-1} \) and

![Figure 7: Evolution of \( \log_{10}(r^n_h) \) with respect to the iterates with \( \omega = (0.2, 0.8) \) and \( y_0(x) \equiv \sin(\pi x) \) for \( h = (1/160, 1/160) \).](image-url)
$\rho_0^{-1}$ are exponentially close to zero in $(0,1) \times (T - \delta, T)$, these high frequencies have a very weak effect on the values of $J^*$. The high frequency components of the unknowns $\phi_{T,h}$, which really do exist since the minimizer $\phi_T$ lives in abstract space much larger than $L^2(0,1)$, are damped out from $t = T$ to $t = T - \delta$ and, therefore, again does not affect the value of $J_h$ significatively. This very low dependence explains the difficulty to capture such frequencies with a gradient method.

But the crucial point from the numerical viewpoint is that, since these high frequencies are damped out where the weight vanishes, they are not necessary to achieve a good approximation of the control $v_h$, the state controlled $y_h$, and the associated cost. Consequently, a reasonable value of $\sigma$ suffices. For instance, from Table 16 (for which $\sigma = 10^{-4}$) and Table 17 (where $\sigma = 10^{-5}$), we see that, for $h$ close to $(0,0)$, $\|v_h\|_{L^2(Q_T)}$ and $\|y_h\|_{L^2(Q_T)}$ are unchanged in practice. Contrarily, the values of $\|\phi_{T,h}\|_{L^2(0,1)}$ do change when $\sigma$ is divided by 10, as it contains more high frequency modes. Table 18 displays relevant numerical values for $\sigma = 10^{-3}, 10^{-4}, 10^{-5}$ and $\sigma = 10^{-6}$ and emphasizes together with Table 17 that even if $\mu_h$ and $\psi_h$ do not converge in $L^2(Q_T)$, the weighted functions $\rho^{-1}_h\mu_h$ and $\rho^{-2}_{0,h}\psi_h$ do.

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>$10^{-3}$</th>
<th>$10^{-4}$</th>
<th>$10^{-5}$</th>
<th>$10^{-6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG iterates</td>
<td>16</td>
<td>471</td>
<td>3620</td>
<td>25631</td>
</tr>
<tr>
<td>$|y_h(\cdots,T)|_{L^2(0,1)}$</td>
<td>$3.18 \times 10^{-4}$</td>
<td>$3.01 \times 10^{-5}$</td>
<td>$3.14 \times 10^{-6}$</td>
<td>$2.81 \times 10^{-7}$</td>
</tr>
<tr>
<td>$|\rho_0^{-2}\psi_h|_{L^2(Q_T)}$</td>
<td>1.0022</td>
<td>1.0159</td>
<td>1.0274</td>
<td>1.0309</td>
</tr>
<tr>
<td>$|\rho^{-2}<em>h\mu_h|</em>{L^2(Q_T)}$</td>
<td>$2.0083 \times 10^{-1}$</td>
<td>$1.9995 \times 10^{-1}$</td>
<td>$1.9924 \times 10^{-1}$</td>
<td>$1.9904 \times 10^{-1}$</td>
</tr>
<tr>
<td>$|\psi_h|_{L^2(Q_T)}$</td>
<td>2.64</td>
<td>5.89</td>
<td>1.85 $\times 10^1$</td>
<td>2.48 $\times 10^1$</td>
</tr>
<tr>
<td>$|\phi_h(\cdots,T)|_{L^2(0,1)}$</td>
<td>$1.51 \times 10^1$</td>
<td>$3.94 \times 10^1$</td>
<td>$1.74 \times 10^2$</td>
<td>$2.46 \times 10^2$</td>
</tr>
<tr>
<td>$|\mu_h|_{L^2(Q_T)}$</td>
<td>7.55</td>
<td>1.32 $\times 10^1$</td>
<td>5.99 $\times 10^1$</td>
<td>1.62 $\times 10^2$</td>
</tr>
</tbody>
</table>

Table 18: $\omega = (0.2,0.8)$ and $h = (1/160,1/160)$.

When we try to compute the null control of minimal $L^2$-norm, that is, when we try to solve a problem like $[\text{(6)}]$ with $\rho \equiv 0$ and $\rho_0 = 1$, the conjugate gradient method for the associated dual problem, which is much more sensitive to the numerical approximation, behaves very differently. This was discussed in length in [38]. In this case, the control depends much more strongly on the final adjoint state $\phi_T$. Consequently, smaller values of the tolerance $\sigma$ are needed, so as to capture high frequencies and this leads to a larger number of iterates. Moreover, the control of minimal $L^2$-norm, defined simply as $v = \varphi 1_{Q_T}$, exhibits a highly oscillatory behavior in the time direction near $t = T$. It results that, for any fixed and small enough $\sigma$, the number of CG iterates is no more constant with respect to the discretization parameter, but blows up exponentially as $h \rightarrow (0,0)$.

In the present situation, the weight $\rho_0^{-2}$ has the effect to destroy such oscillatory behavior so that $v_h = \rho_0^{-2}1_{Q_T}$ is smooth near $T$ (see for instance Figure 1).

For $(\rho,\rho_0) = (0,1)$, $h = (1/160,1/160)$, Figure 5 depicts the evolution of the residue $r_h^T$ and the corresponding final adjoint state, which minimizes $I$, see [4]. The evolution is similar to the one observed for weighted integrals (Figure 7), but the stopping test for $\sigma = 10^{-5}$ is achieved only after 9671 iterates, instead of 3620. The other numerical values (to be compared with those in Table 17 second column) are the following: $\|v_h\|_{L^2(\Omega_T)} \approx 7.45 \times 10^{-1}$, $\|y_h\|_{L^2(Q_T)} \approx 1.52 \times 10^{-1}$, $\|y_h(\cdots,T)\|_{L^2(0,1)} \approx 2.61 \times 10^{-6}$ and $\|\varphi_T\|_{L^2(0,1)} \approx 5.05 \times 10^2$.

The first conclusion is that the weights have a clear influence on the behavior of the iterative conjugate gradient algorithm. Also, for $\varepsilon = 0$ and $R = \infty$, the minimization of $J^*$ is numerically ill-posed, since the unique minimizer $(\mu, \varphi_T)$ lives in a singular and very large space, hard to approximate by a finite dimensional approach. Thus, for more severe data like the one considered at the end of Section 4.2 (see Table 11 and Figures 5 and 9) for which $\omega = (0.2,0.4)$, we observe,
7 Further comments and concluding remarks

7.1 Common aspects of the primal and dual problems and methods

The solution to the variational formulation (15) is also the unique minimizer of the functional $I : P \mapsto \mathbb{R}$, with

$$I(p) := \frac{1}{2} \int_{Q_T} \rho^{-2} |L^* q|^2 \, dx \, dt + \frac{1}{2} \int_{Q_T} \rho_0^{-2} |q|^2 \, dx \, dt - (y_0, q(\cdot, 0))_{L^2(0,1)}.$$  (89)

This is similar to the conjugate functional $J^*$ found in Section 5, which is written here for $\varepsilon = 0$, $R = \infty$:

$$J^*(\mu, \varphi_T) := \frac{1}{2} \int_{Q_T} \rho^{-2} |\mu|^2 \, dx \, dt + \frac{1}{2} \int_{Q_T} \rho_0^{-2} |\psi|^2 \, dx \, dt + (y_0, \psi(\cdot, T))_{L^2(0,1)},$$  (90)

where $\psi$ solves the backwards problem: $L^* \psi = \mu$ in $Q_T$, $\psi = 0$ on $\Sigma_T$ and $\psi(\cdot, T) = \varphi_T$.

Obviously, $J^*(\mu, \varphi_T) = I(\psi)$ for all $(\mu, \varphi_T) \in V$. Note that, according to the precise definition of the space $P$, this analogy allows to affirm that the space $V$ (where the minimum of $J^*_R,\varepsilon$ is attained) degenerates as $\varepsilon \to 0$ and $R \to +\infty$ towards a space related to the completion of $L^2(Q_T) \times L^2(0,1)$ for the norm in $P$.

Moreover, if we view the previous backwards problem for $\psi$ as a set of constraints for $\mu$ and $\varphi_T$ and we introduce associated multipliers, we can find for (90) a Lagrangian analogous to (61). Therefore, all the extremal problems we have introduced having (6) as starting point are connected to each others. Those in Sections 3 and 4 belong to the framework of elliptic variational problems in two dimensions and are well tailored for a resolution with finite elements. Those in Sections 5 are of parabolic nature. Accordingly, the time variable is kept explicit and time integration is required.

Figure 8: $\omega = (0.2, 0.8) - (\rho, \rho_0) = (0, 1) - \sigma = 10^{-5} - h = (1/160, 1/160)$. Evolution of the residue $r^n_h$ with respect to $n$ (Left) and corresponding final adjoint state $\phi_{T,h}$ (Right).

for $h = (1/160, 160)$, that more 100 000 iterates of the conjugate gradient algorithm are needed to get a relative residue $r^n_h$ lower than $\sigma = 10^{-4}$.
7.2 The lack of regularity of $p(\cdot, T)$

An important feature of the problem satisfied by $p$ is that it gives no information on the regularity of $p(\cdot, T)$. There is an argument that justifies this lack of information even when $\alpha$ is constant and $p$ and $\rho_0$ are regular and bounded in $Q_T$. It is the following.

Let us assume that $\omega \neq (0,1)$. Let us set $j(w) = J(y_w, w)$, where $y_w$ is the (unique) solution to (1) with $v$ replaced by $w$ and let us set $Kw = z_w(\cdot, T)$, where $z_w$ is the solution to (9). Then, as a consequence of the irreversibility and the regularizing effect of the heat equation, $K$ can be viewed as a linear continuous and surjective mapping on $L^2(q_T)$ with values in a “very small” Hilbert space $R(K)$, a dense subspace of $L^2(0,1)$. From the Lagrange multipliers theorem, we know that $(y,v)$ is the unique solution to (6) if and only if

- There exists $\lambda \in R(K)'$ such that
  \[ (j'(v), w)_{L^2(q_T)} + \langle \lambda, z_w(\cdot, T) \rangle = 0 \quad \forall w \in L^2(q_T), \quad (91) \]
  where $\langle \cdot, \cdot \rangle$ stands for the duality pairing for $R(K)'$ and $R(K)$ and
  - $y$ solves (1).

Let $(y,v)$ be an optimal pair and let $p$ be the solution to (17); we know that $(y,v)$ and $p$ satisfy (14). Let us assume that $p(\cdot, T) \in L^2(0,1)$ and let $\lambda \in R(K)'$ satisfy (91). In principle, there is no reason to have $\lambda \in L^2(0,1)$ (of course, $L^2(0,1)$ can be viewed as a “very small” part of $R(K)'$).

However, it is clear from (14) and (17) that

\[
(j'(v), w)_{L^2(q_T)} + \langle \lambda, z_w(\cdot, T) \rangle = \int_{Q_T} (p + \rho_0^2 v) w \, dx \, dt + \langle \lambda - p(\cdot, T), z_w(\cdot, T) \rangle = \langle \lambda - p(\cdot, T), z_w(\cdot, T) \rangle
\]

for all $w \in L^2(q_T)$. Consequently, we should have

\[ \lambda = p(\cdot, T), \]

which is in contradiction with the fact that $\lambda$ does not necessarily belong to $L^2(0,1)$.

Thus, except in the particular case where the control acts on the whole space domain, the function $p(\cdot, T)$ (that can be viewed as a multiplier associated to the constraint $y(\cdot, T) = 0$) does not belong to $L^2(0,1)$.

7.3 The role of the weights

The explicit introduction of $y$ in the functional $J$ in (5) allows to give expressions of the optimal control and state in terms of the solution $p$ to (15). With $\rho = 0$, this would have not been possible, so that the search of the control of minimal $L^2$-norm would require, for instance, a dual method. The exponential behavior of these weights gives a meaning to the variational formulation (15), reinforces the controllability requirement (through the weight $\rho$) and regularizes the behavior of the control near $t = T$ (through $\rho_0$), in contrast with the evolution of the control of minimal $L^2$-norm, that is highly oscillatory near $T$.

Carleman estimates ensure the well-posedness of the variational formulation for these specific weights, that blow up exponentially as $t \to T$. Numerically, the use of other weights in (15) leads to non-convergent sequences. This specific behavior may be relaxed for the approximate controllability situation, which can be treated within the primal approach as well.
7.4 Numerical analysis and error estimates

At the numerical level, the weights play a major role. The variational formulation, within the framework of finite elements, leads to strong convergence results for the approximate sequence \(\{v_h\}\) as \(h\) goes to zero. To our knowledge, this is the first convergence result for the numerical approximation to the null controllability problem for the heat equation. After an appropriate regularity analysis of the solution to \((15)\), one may further obtain estimates of \(\|v - v_h\|_{L^2(Q_T)}\) in terms of \(|h|\) and a suitable norm of \(p\); we refer to \([18]\).

The mixed approach provides numerical results for which the null controllability requirement is very well satisfied, even when the diffusion function is only piecewise constant. This is in contrast with the existing literature, mainly devoted to the approximate controllability issue. The method is also robust enough to check the exponential behavior of the norm \(\|v_h\|_{L^2(Q_T)}\) with respect to the constant diffusion \(a\). A similar behavior is observed with respect to the final time \(T\).

As mentioned above, it will be interesting to analyze rigorously \([6]\) from the viewpoints of stability and convergence. In particular, a relevant question is whether inequalities like \((59)\) and \((60)\) hold at the finite dimensional level, with constants \(\kappa_1\) and \(\kappa_2\) independent of \(h\).

We also emphasize that the analysis in Section 3.2 and in particular proposition 3.2 remain true in higher dimensions in space. The approximation used here leads to a general convergence result, whatever be in particular the regularity of the initial datum \(y_0\). This is of course related to the behavior of the weight \(\rho_0^{-1}\) that, among other effects, serve to damp out the high frequencies of the backwards solution.

In that direction, taking into account that the primal and the dual approaches are closely linked, it is reasonable to suspect that convergence results of the same kind, namely the strong convergence of \(v_h = -\pi_h(\rho_0^{-2})\phi_h\), can be obtained in the framework of Section 6.2. This issue is open for the minimal \(L^2\)-norm situation (i.e. \(\rho = 0\) and \(\rho_0 = 1\)); actually, whether or not the strong convergence holds in this case is far from being clear.

7.5 The change of variables

It is clear that the variable \(p\), corresponding to the solution to \((15)\), is not appropriate in the numerical context. The same holds for the variable \(\psi\), associated to the backwards problem \((76)\).

In particular, we get that the \(L^\infty\)-norms of \(p_h\) and \(\psi_h\) are not bounded at time \(T\) uniformly in \(h\). According to that behavior, once again related to the high regularizing effect of the heat kernel, it appears very relevant to perform a change of variable assuming that \(p\) blows up exponentially at \(t\) goes to \(T^-\) and therefore search for example \(p\) under the form \(p = (T - t)^{\alpha}\rho_0 z\), where \(\alpha\) must be chosen taking into account the powers of the polynomials in \(T - t\) appearing in the global Carleman estimate \([12]\).

The new variable \(z\) enjoys better regularity properties. Within the primal approach, we observed a significant gain on the conditioning number of the matrix.

We also emphasize that this idea may be used in the context of the dual approach (this is not done here in order to keep the dual formulation as close as possible to the one initially introduced in \([8]\)). There, considering the limit case \(\varepsilon = 0\) and \(R = \infty\), we introduce the variables \(\tilde{\psi}\) and \(\tilde{m}\) as follows:

\[
\tilde{\psi} = (T - t)^{-\alpha}\rho_0^{-1}\psi, \quad \tilde{m} = \rho^{-1}\mu, \quad \alpha \in \mathbb{R}
\]

and we consider the minimization of \(J^*\), with

\[
J^*(\tilde{m}, \tilde{\psi}_T) := \frac{1}{2} \int_{Q_T} (T - t)^{2\alpha}|\tilde{\psi}|^2 \, dx \, dt + \frac{1}{2} \int_{Q_T} |\tilde{m}|^2 \, dx \, dt + T^\alpha < \rho_0(\cdot, 0)\tilde{\psi}(\cdot, 0), y_0 >_{L^2(0,1)},
\]

(92)
where \( \tilde{\psi} \) solves the backward problem:

\[
\begin{aligned}
\rho^{-1} L((T - t)^a \rho_0 \tilde{\psi}) &= \tilde{m}, & (x, t) &\in Q_T \\
\tilde{\psi}(x, t) &= 0, & (x, t) &\in \Sigma_T \\
\tilde{\psi}(x, T) &= \tilde{\psi}_T(x), & x &\in (0, 1)
\end{aligned}
\] (93)

over the completion of \( D(Q_T) \times D(0, 1) \) for the norm \( \left( \|(T - t)^a \tilde{\psi}\|_{L^2(Q_T)}^2 + \|\tilde{m}\|_{L^2(Q_T)}^2 \right)^{1/2} \). The control and the corresponding controlled state are then given by \( v = -\rho_0^{-1}(T - t)\tilde{\psi}1_\omega \) and \( y = \rho^{-1}\tilde{m} \), respectively.

As before, \( J^* \) may be minimized using a conjugate gradient method. Very likely, according to our observations on the primal problem, this rescaling must have an impact on the behavior of the algorithm. This remains to be analyzed. Notice that a similar strategy may be employed for the control of minimal \( L^2 \)-norm as well.

### 7.6 Primal versus dual methods

When the (exact) null controllability is under consideration, the primal method is more efficient. In particular, the mixed approach allows to reproduce exactly the null controllability requirement. Very interestingly, the robustness seems independent of the sizes of the control domain and the diffusion coefficient.

Furthermore, the primal approach benefits at the theoretical-numerical level from finite element theory. In particular, as we have seen, we can get convergence results easily.

On the other hand, we may want to adapt (and refine) the mesh of \( Q_T \) in order to improve convergence and such adaptation is less simple in the dual approach, where \( t \) is “conserved” as a time variable.

Also, using finite element tools, we can without any additional difficulty get results in the case where the sub-domain \( \omega \) varies in time, that is, non-cylindrical domains \( q_T \). It is not difficult to prove that null controllability holds as well for any time \( T > 0 \) when the control is exerted on any open set \( q_T \) defined by

\[
q_T = \{(x, t) \in Q_T : g(t) < x < h(t), t \in (0, T)\},
\]

where \( g \) and \( h \) are smooths functions on \([0, T]\), with \( 0 \leq g \leq h \leq 1 \) and \( g(t) \neq h(t) \). This opens the possibility to optimize numerically the domain \( q_T \), as was done in a cylindrical situation in [37].

The intrinsic ill-posedness of this problem is enhanced within the dual approach, at least when the variable \( \psi \) is considered. There, as soon as the control support is sufficiently small, the conjugate gradient fails to converge. Indeed, for the dual problem to work reasonably, one has to be very careful, in particular in the time integration process.

Notice that, contrarily, the primal mixed formulation is again easy to implement and leads to definite positive, symmetric and (very) sparse matrices.

### 7.7 Some extensions

The methods used in this paper can be extended to cover null controllability problems for linear heat equations in higher spatial dimensions. More precisely, let \( \Omega \subset \mathbb{R}^N \) be a regular, bounded, connected open set and let us consider the linear system

\[
\begin{aligned}
y_t - \nabla \cdot (a(x) \nabla y) + A(x, t) y &= v1_\Omega, & (x, t) &\in \Omega \times (0, T) \\
y(x, t) &= 0, & (x, t) &\in \partial \Omega \times (0, T) \\
y(x, 0) &= y_0(x), & x &\in \Omega
\end{aligned}
\] (94)

\( 0 < a(x) \leq a_0 \) and \( A(x, t) \) is a measurable function with \( |A(x, t)| \leq m \) a fixed constant, \( 0 < m < \infty \). The null controllability holds if and only if the associated adjoint state satisfies

\[
\begin{aligned}
\rho^{-1} L((T - t)^a \rho_0 \tilde{\psi}) &= \tilde{m}, & (x, t) &\in Q_T \\
\tilde{\psi}(x, t) &= 0, & (x, t) &\in \Sigma_T \\
\tilde{\psi}(x, T) &= \tilde{\psi}_T(x), & x &\in (0, 1)
\end{aligned}
\]

with \( \tilde{m} \) satisfying the above equation. The control and the corresponding controlled state are then given by \( v = -\rho_0^{-1}(T - t)\tilde{\psi}1_\omega \) and \( y = \rho^{-1}\tilde{m} \), respectively. As before, \( J^* \) may be minimized using a conjugate gradient method. Very likely, according to our observations on the primal problem, this rescaling must have an impact on the behavior of the algorithm. This remains to be analyzed. Notice that a similar strategy may be employed for the control of minimal \( L^2 \)-norm as well.

Over the completion of \( D(Q_T) \times D(0, 1) \) for the norm \( \left( \|(T - t)^a \tilde{\psi}\|_{L^2(Q_T)}^2 + \|\tilde{m}\|_{L^2(Q_T)}^2 \right)^{1/2} \). The control and the corresponding controlled state are then given by \( v = -\rho_0^{-1}(T - t)\tilde{\psi}1_\omega \) and \( y = \rho^{-1}\tilde{m} \), respectively. As before, \( J^* \) may be minimized using a conjugate gradient method. Very likely, according to our observations on the primal problem, this rescaling must have an impact on the behavior of the algorithm. This remains to be analyzed. Notice that a similar strategy may be employed for the control of minimal \( L^2 \)-norm as well.
where \( a \in C^1(\overline{\Omega}) \) with \( a(x) \geq a_0 > 0 \), \( A \in L^\infty(\Omega \times (0,T)) \), \( O \subset \Omega \) is a (small) non-empty open set, \( v \in L^2(\Omega \times (0,T)) \) is the control and \( y_0 \in L^2(\Omega) \) is the initial state. The null controllability problem for (94) is to find, for each \( y_0 \in L^2(\Omega) \), a control \( v \) such that the associated solution satisfies
\[
y(x,T) = 0, \quad x \in \Omega.
\]

The situation is more involved.

First, notice that a result similar to Theorem 2.1 holds (for piecewise constant coefficients one still has null controllability results under some hypotheses, see [30, 31, 4]; however, stronger regularity is needed in order to get Carleman estimates). Secondly, observe that the analog of the space \( P_h \) in (32) is considerably more complex than in the 1D case. As in Section 3.2 in order to construct finite dimensional subspaces of \( P_h \), we have in practice to work with a mesh of \( \Omega \times (0,T) \) and with functions that locally reduce to polynomials in \((x,t)\) (for instance) and are globally continuous and \( C^1 \) in the spatial variables. But this is not straightforward when \( N \geq 2 \).

Fortunately, this can be avoided using mixed formulations, as in Section 4. On the other hand, the dual approach allows to keep the time-dependent viewpoint and avoids the introduction of elliptic systems of higher order. Therefore, since the related computational work is reasonable, it is also expectable that dual methods can be adapted and extended to this setting.

Appropriate weight functions have to be used. In practice, what we have to be able to construct is a positive function in \( \Omega \) that vanishes on \( \partial \Omega \) and possesses nonvanishing gradient in \( \Omega \setminus \omega \). Such a function always exists (a result by Imanuvilov) and is relatively easy to construct for instance when \( \Omega \) is convex.

In order to illustrate the situation, let us present the results of an experiment. We solve numerically the null controllability problem for (94) with \( N = 2, \Omega = (0,1) \times (0,1), O = (0.2,0.6) \times (0.2,0.6), T = 1, a(x) \equiv 1, A(x,t) \equiv 1 \) and \( y_0(x) \equiv 1000 \).

The space-time domain and the mesh are displayed in Figure 9.

We have used a mixed formulation similar to (64), where \( M_h, Q_h \) and \( \tilde{M}_h \) are standard finite element \( P_2 \)-Lagrange spaces. The resulting system, in view of its size and structure, has been solved with the Arrow-Hurwicz method, that provides good results, better than a direct solver; see for instance [21, 22]. Indeed, recall that solving (64) is equivalent to the computation of the saddle-points of a Lagrangian, see the related argument in Section 4.

The iterates have been stopped for a relative error of two consecutive iterates less than \( 10^{-5} \). The computed control and state are shown in Figures 10–13. The computations have been performed with the FreFem++ package, see http://www.freefem.org/ff++. For more information, a detailed analysis and other similar numerical experiments, see [19].

The primal and dual approaches presented in Sections 3 to 6 can also be extended to cover many other controllable systems for which appropriate Carleman estimates are available.

Thus, we can analyze and solve that way non-scalar parabolic systems, nonlinear system, Stokes and Stokes-like systems, etc. We refer again to [19]. It is also possible to extend the previous arguments and methods to the boundary null controllability case and to the exact controllability to trajectories (with distributed or boundary controls).

Actually, as first noticed in [20] and using in part the results by [39] and [43], the approach may also work for linear equations of the hyperbolic kind, where the practical computation of exact controls remains a challenge (see [10]). This work also opens the possibility to address the numerical solution of nonlinear control problems, the optimization of the control support \( \omega \), etc.

Let us finally mention that many non-linear situations can be considered through a suitable linearization and an iterative process. We refer to the recent work of the authors [17] for the numerical approximations of controls for a semi-linear heat equation.
Figure 9: The space-time domain and the mesh: 2,800 vertices. Total number of unknowns (the values of $m_h$, $r_h$ and $\lambda_h$ at the nodal points, see (64)): $6846 \times 3 = 20,538$.

Figure 10: Control and state iso-lines at $x_1 = 0.28$. Minimal and maximal values for $u_h$: $-9.89 \times 10^2$, $9.45 \times 10^1$. Minimal and maximal values for $y_h$: $-2.84 \times 10^3$, $3.70 \times 10^3$. 
Figure 11: Control and state iso-lines at $x_1 = 0.44$. Minimal and maximal values for $u_h$: $-1.26 \times 10^3$, $1.17 \times 10^2$. Minimal and maximal values for $y_h$: $-2.84 \times 10^3$, $3.70 \times 10^3$.

References


Figure 12: The surface $u_h(x, t) = 0$ in the $(x, t)$ space. The region below this surface is the real support of the computed control.

Figure 13: The surface $y_h(x, t) = 0$ in the $(x, t)$ space. In practice, above this surface $y_h$ vanishes.
REFERENCES


REFERENCES


