Ill-Conditioning versus Ill-Posedness for the Boundary Controllability of the Heat Equation
Faker Ben Belgacem, Sidi-Mahmoud Kaber

To cite this version:

HAL Id: hal-01026447
https://hal.sorbonne-universite.fr/hal-01026447
Submitted on 22 Jul 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Ill-Conditioning versus Ill-Posedness for the Boundary Controllability of the Heat Equation.

Faker Ben Belgacem * Sidi Mahmoud Kaber †

December 20, 2013

Abstract

Ill-posedness and/or Ill-conditioning are features users have to deal with appropriately in the controllability of diffusion problems for secure and reliable outputs. We investigate those issues in the case of a boundary Dirichlet control, in an attempt to underline the origin of the troubles arising in the numerical computations and to shed some light on the difficulties to obtain good quality simulations. The exact controllability is severely ill-posed while, in spite of its well-posedness, the null-controllability turns out to be very badly ill-conditioned. Theoretical and numerical results are stated on the heat equation in one dimension to illustrate the specific instabilities of each problem. The main tools used here are first a characterization of the subspace where the HUM control lies and the study of the spectrum of some structured matrices, of Pick and Löwner type, obtained from the Fourier calculations on the state and adjoint equations.

KEYWORDS: boundary controllability, Dirichlet control, Structured matrices, ill-conditioning, ill-posedness.

*Sorbonne Universités, UTC, EA 2222, Laboratoire de Mathématique Appliquée de Compiègne, F-60205 Compiègne, France.
†Sorbonne Universités, UPMC Univ Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France,
1 Introduction

The subject is the null and exact controllability problems of the heat equation where the control variable is a boundary Dirichlet data. We focus on the instabilities and their effects users should cope with in the numerical computations. The null-controllability is well-posed and the exact-controllability is ill-posed. The purpose is to investigate the one-dimensional model to underline the specific difficulties generated by the instabilities of both problems. Following [24], we use a Fourier analysis of the heat equation and we show how the Gramians of the exact and null-controllability problems may be expressed by structured matrices of infinite dimensions. The study of the spectrum of these matrices provides the desired information on the instabilities of both problems. This procedure has been first exposed by the authors in [2].

To be more specific, let $I$ be the segment $(0, \pi)$ of the real axis and $T > 0$ be a fixed real-number. We set $Q = I \times [0, T]$. The generic point in $I$ is denoted by $x$ and the generic time variable is $t$. Let $y_0 = y_0(x), x \in I$ be an initial state and $v = v(t), t \in (0, T)$ be a Dirichlet boundary data. Then we consider $y_v$ the solution of the heat equation

\begin{align*}
\partial_t y_v - \partial_{xx} y_v = 0 & \quad \text{in } Q, \\
y_v(0, \cdot) = 0, \quad y_v(\pi, \cdot) = v & \quad \text{on } (0, T), \\
y_v(\cdot, 0) = y_0 & \quad \text{on } I.
\end{align*}

The null-controllability is when the final state is zero. The problem is then to: find Dirichlet condition $u(\cdot)$ such that

\begin{equation}
y_u(x, T) = 0, \quad \forall x \in I
\end{equation}

The exact controllability problem may be expressed as follows: find $u(\cdot)$ such that

\begin{equation}
y_u(x, T) = y_T(x), \quad \forall x \in I,
\end{equation}

where $y_T$ is a state fixed in advance. Whenever we deal with problem (5), we set the initial condition to zero ($y_0 = 0$). This is not a restriction and is only chosen for simplification.

HUM method applied to the null-controllability problem allow to prove the existence of the $v$ solution of (4). This control depends continuously on
the initial data $y_0$ (see [6, 8, 7, 31]). We refer to [19, 18, 20, 4, 23, 21] for valuable discussion about related numerical issues. Studies for the problem (5) are exposed in [15, 12, 13] and conclude to the failure of the failure of the exact-controllability. Some states $y_T$ are not reachable. Moreover, recent developments realized in [2] gives a better insight about its severe ill-posedness. The basic tool for the proof is an precise spectral analysis of structured matrices of Pick, L"owner and Cauchy types (see [29, 26]).

The aim of this contribution is to take benefit of the ideas initiated in [2] to investigate further the instabilities related to exact and null controllability problems and to identify their precise nature. This way to proceed is based on a use of the elegant structured matrices and is an alternative to the analysis of bi-orthogonal sequences used in a wide literature. We put therefore the controllability problems (4) and (5) under algebraic form. The infinite-dimensional-vectors, solutions of each of the resulting equations allow the reconstruction of the HUM control, the one with minimal norm in $L^2(0,T)$, the space of square integrable functions defined on $(0,T)$. The process we follow is constructive and consists in expressing the Gramians of both problems by means of Pick and L"owner matrices. To make so, we use a Fourier basis in $L^2(I)$ and a system of exponential functions that forms a total system in a closed subspace of $L^2(0,T)$, where the active controls are sought for. Each of those matrices enjoys a particular structure, the one corresponding to (4) is of a L"owner type and the one associated to (5) is a Pick matrix which a sum of two Cauchy matrices. Their ‘instability’ can be clarified after realizing a sharp study of their spectrum. Writing them as the solution of a Sylvester or a Lyapunov equation, and using estimates of Penzl’s type on their eigenvalues (see [25]) allow us to illustrate the (extremely) high ill-conditioning of the null-controllability and the severe ill-posedness of the exact-controllability. These instabilities, though fundamentally of different nature, are causes of tremendous complications in the numerical simulations to approximate the control.

An outline of the paper is as follows. In Section 2, we describe the exact-and null-controllability of the heat equation and we provide an accurate representation of the HUM control that has an important role in our analysis. Section 3 summarizes and completes the results recently elaborated in [2] and [19] about the exact controllability problem (5). Following the lines of [6, 24], we put the controllability problem under an algebraic form with an infinite structured matrix. The ill-posedness degree is therefore derived after
studying the eigenvalues of that matrix. Applying Penzl's type estimate to
this matrix yields the severe ill-posedness of the exact-controllability prob-
lem. Section 4 is dedicated to the null-controllability problem (4) which is
well posed. That result is known [6], [2]. The objective we have in mind here
is to investigate the spectrum of the null-controllability problem so to eventually evaluate the condition number of the problem. Tools we
use have been developed for the matrices that are solutions of Lyapunov or
Sylvester equations (see [22, 25] and references therein). The analysis con-
ducted here concludes to a very high ill-conditioning of the problem. This
confirms why solving the null-controllability causes high numerical troubles.
To close this work, some Matlab computations in Section 5 allow us to check
out numerically the theoretical results developed here.

Notation — Let $X$ be a Banach space endowed with its norm $\| \cdot \|_X$. We
denote by $C([0, T]; X)$ the space of continuous functions $v$ from $[0, T]$ in $X$
such that
\[
\|v\|_{C([0, T]; X)} = \sup_{t \in [0, T]} \|v(t)\|_X < +\infty.
\]
We need also in some places the Hilbert space $\ell^2(\mathbb{R})$ of square summable real
sequences. Finally, we consider the Sobolev space $H^1(I)$ of all the functions
that belong to $L^2(I)$ together with their first derivatives. The space $H^1_0(I)$
is then the closure in $H^1(I)$ of the space $\mathcal{D}(I)$ of infinitely differentiable
functions with a compact support in $I$. We denote by $H^{-1}(I)$ its dual space.

2 Control’s Representation

Let the Dirichlet data $v$ be given in $L^2(0, T)$. The parabolic problem (1)-(3)
has a unique solution $y_v$ in $L^2(Q)$ which belongs to $C([0, T]; H^{-1}(I))$ (see [14,
Theorem 9.1]). It may occur that for some $v \in L^2(0, T)$, the state $y_v(\cdot, T)$
does not belong to $L^2(I)$, we are therefore led to work with the unbounded
operator $B$, defined by
\[
Bv = y_v(\cdot, T).
\]
Its domain
\[
\mathcal{D}(B) = \left\{ v \in L^2(0, T); \quad y_v(\cdot, T) \in L^2(I) \right\}
\]
is dense in $L^2(0, T)$ and following [1, Lemma 2.4], $B$ is closed. The adjoint
operator $B^*$ is well defined and is specified as follows (see [12, 1]): let $\psi$ be
given in $L^2(I)$ and $q_\psi$ the solution of the backward problem (see [16, Chap. 4])

$$-\partial_t q_\psi - \partial_{xx} q_\psi = 0 \quad \text{in } Q, \quad (6)$$

$$q_\psi(0, \cdot), \quad q_\psi(\pi, \cdot) = 0 \quad \text{on } (0, T), \quad (7)$$

$$q_\psi(\cdot, T) = \psi \quad \text{on } I. \quad (8)$$

The domain of $B^*$

$$\mathbb{D}(B^*) = \left\{ \psi \in L^2(I); \quad \partial_\pi q_\psi(\pi, \cdot) \in L^2(0, T) \right\},$$

is dense and we have

$$B^* \psi = -\partial_\pi q_\psi(\pi, \cdot).$$

The properties of both operators $B$ and $B^*$ have been listed in [2]. The operator $B^*$ turns out to be injective:

$$\text{Ker } B^* = \{0\}.$$ 

The direct consequence is that the range $\mathcal{R}(B)$ is dense in $L^2(I)$. The set of reachable states is thus dense in $L^2(I)$; this is the approximate controllability (see [12]). The set of non-reachable states $L^2(I) \setminus \mathcal{R}(B)$ is also dense in $L^2(I)$, see [2, Lemma 3.4]. $B$ has therefore a non-closed range and cannot be continuously invertible. This means that any deviation of a reachable state $y_T \in \mathcal{R}(B)$, may produce a new state which is not reachable ($\notin \mathcal{R}(B)$). Then, equation (5) may fail. The determination of the closure of the range of $B^*$ and by then of the kernel of $B$ is a little bit more complicated. It will be obtained after carrying out some Fourier computations and using a sharp variant of Müntz theorem.

The methodology followed here has been used earlier in [6, 19]. For a given $\psi \in L^2(I)$, we consider the Fourier expansion,

$$\psi(x) = \sum_{k \geq 1} \psi_k \sin(kx) \quad \forall x \in I.$$ 

Plugging this series into problem (6)-(8) yields that $B^*$ may be expressed as

$$(B^* \psi)(t) = \sum_{k \geq 1} (-1)^{k+1} ke^{-k^2(T-t)} \psi_k, \quad \forall t \in (0, T).$$
As a consequence, we deduce another characterization of the domain of $B^*$

$$\mathcal{D}(B^*) = \left\{ \psi \in L^2(I); \quad \sum_{k \geq 1} \sum_{m \geq 1} (-1)^{k+m} \frac{km}{k^2 + m^2} (1 - e^{-(k^2 + m^2)T}) \psi_k \psi_m < \infty \right\}. $$

Notice that the family $(e^{-k^2(T-t)})_{k \geq 1}$ is total in $\mathcal{R}(B^*)$. Actually the following important result holds. We refer to [2, Lemma 3.4, Remark 4.2], for the proof.

**Lemma 2.1** The co-dimension of $\mathcal{R}(B^*)$ in $L^2(0,T)$ is infinite. Moreover, all the functions in $\overline{\mathcal{R}(B^*)}$ are analytic in $[0,T]$ and we have the following representation

$$\overline{\mathcal{R}(B^*)} = \left\{ v = \sum_{k \geq 1} v_k e^{-k^2(T-t)}, \quad \sum_{k \geq 1} \sum_{m \geq 1} \frac{1 - e^{-(k^2 + m^2)T}}{k^2 + m^2} v_k v_m < \infty \right\}. $$

The next step consists in deriving a closed form of the operator $B$ itself. It may be obtained by duality (see [2]). Indeed, let $v \in \mathcal{D}(B)$. We set for convenience $v_T(\cdot) = v(T - \cdot)$ and denote by $v_T$ the trivial extension of $v_T$ to the semi-axis $\mathbb{R}_+$. The operator $B$ may be written under the following form

$$(Bv)(x) = \frac{2}{\pi} \sum_{k \geq 1} (-1)^{k+1} \left[ k \int_0^T v(t)e^{-k^2(T-t)} dt \right] \sin(kx),$$

$$= \frac{2}{\pi} \sum_{k \geq 1} (-1)^{k+1} \left[ k \mathcal{L}(v_T)(k^2) \right] \sin(kx). \quad (9)$$

The symbol $\mathcal{L}(v_T)$ denotes the Laplace transform of the extended function $v_T \in L^2(\mathbb{R}_+)$. An explicit form of $\mathcal{D}(B)$ is therefore

$$\mathcal{D}(B) = \left\{ v \in L^2(0,T); \quad \sum_{k \geq 1} [k \mathcal{L}(v_T)(k^2)]^2 < \infty \right\}. $$

Above all, the kernel of $B$ is given as follows

$$\text{Ker } B = \left\{ v \in \mathcal{D}(B); \quad \mathcal{L}(v_T)(k^2) = 0, \quad \forall k \geq 1 \right\}. $$

Thus, Ker $B$ has a countably infinite dimension. Moreover, if $y_T$ is a reachable state, then the HUM control $u^* \in \overline{\mathcal{R}(B^*)}$, solution of (5), is analytic in $[0,T[$. The same result obviously holds for the HUM control $u^\dagger$, solution of the null-controllability (4).
3 Exact-Controllability: Ill-posedness

The issue of ill-posedness degree of problem (5) has been substantially clarified recently in [2] which concludes to the exponential ill-posedness. The proofs here are based on the properties of some structured matrices. We summarize and complete in this section the ideas exposed in [2]. We introduce a direct transformation of problem (5) which leads to the analysis of a linear system of infinite dimension. To proceed with the details we fix the initial condition to zero, \( y_0 = 0 \) for simplification. Using expression (9) for an expansion of \( y_T \in L^2(I) \)

\[
y_T(x) = \sum_{k \geq 1} (y_T)_k \sin(kx) \quad \text{in } I,
\]

we derive an algebraic system on the control \( u \)

\[
\mathcal{L}(y_T)(k^2) = (-1)^{k+1} \frac{\pi}{2} \frac{(y_T)_k}{k}, \quad \forall k \geq 1. \tag{10}
\]

This is a moment problem (see [6]). Users are most often interested in the HUM control \( u^\dagger \), the unique (hypothetic) one that lies in \( \mathcal{R}(B^* \mathcal{T}) \) and is the cheaper control with respect to the energy cost. Owing to Lemma 2.1, we may write

\[
u^\dagger(t) = \sum_{m \geq 1} (u^\dagger)_m e^{-m^2(T-t)}, \quad \forall t \in (0,T).
\]

Replacing in equation (10) implies

\[
\sum_{m \geq 1} \frac{1 - e^{-(k^2+m^2)T}}{k^2 + m^2} (u^\dagger)_m = (-1)^{k+1} \frac{\pi}{2} \frac{(y_T)_k}{k}, \quad \forall k \geq 1. \tag{11}
\]

Let us thus define the infinite vectors \( u^\dagger = ((u^\dagger)_m) \) and \( y_T = ((y_T)_m) \) in \( \mathbb{R}^{N^*} \), and the infinite symmetric matrices \( C_T \) and \( D \) in \( \mathbb{R}^{N^* \times N^*} \) with entries

\[
c_{k,m} = \frac{(1 - e^{-(k^2+m^2)T})}{k^2 + m^2}, \quad \forall k, m \geq 1,
\]

\[
d_{k,m} = (-1)^{k+1} \frac{\pi \delta_{k,m}}{2} \frac{k}{k}, \quad \forall k, m \geq 1.
\]

\( \delta_{k,m} \) is the Krönecker symbol. The matrix \( C_T \) is a Pick type matrix (see [11]). With these new notations, using (11) we obtain formally

\[
u^\dagger = (C_T)^{-1} D y_T. \tag{12}
\]
To assign a mathematical sense to it, we have to specify vectors \( u^\dagger \) that are representative of functions \( u^\dagger \in L^2(0, T) \). The necessary and sufficient condition for this to hold is

\[
\|u^\dagger\|^2_{L^2(0, T)} = \sum_{k \geq 1} \sum_{m \geq 1} \frac{1 - e^{-(k^2 + m^2)T}}{k^2 + m^2} (u^\dagger)_m (u^\dagger)_k = (C u^\dagger, u^\dagger) < \infty.
\]

This may be reworded in terms of the data vector \( y_T \) as follows

\[
(D y_T, (C_T)^{-1} D y_T)_{L^2(\mathbb{R})} = (y_T, D(C_T)^{-1} D y_T)_{L^2(\mathbb{R})} < \infty.
\]

The infinite matrix \( D(C_T)^{-1} D \) is unbounded, positive definite and is not of a closed range. The inverse operator \( A_T = D^{-1} C_T D^{-1} \) is actually the Gramian of the boundary controllability of the heat equation and is also unbounded. It has been checked out in [2] that it may be viewed as representing the operator \( BB^* \). The failure of the exact-controllability can be explained in terms of the properties of this Gramian matrix \( A_T \). Indeed, there holds that

\[
\sup_{y \in L^2(\mathbb{R})} (y, (A_T)^{-1} y)_{L^2(\mathbb{R})} \|y\|_{L^2(\mathbb{R})} = \infty.
\]

Or equivalently

\[
\inf_{y \in L^2(\mathbb{R})} \frac{\sqrt{(y, A_T y)_{L^2(\mathbb{R})}}}{\|y\|_{L^2(\mathbb{R})}} = 0. \tag{13}
\]

This may be established after investigating the spectrum of the principle sub-matrices \((a_{k,m})_{1 \leq k, m \leq N}\) of dimension \( N \). The entries are provided by

\[
a_{k,m} = (-1)^{k+m} \frac{4}{\pi^2} \frac{km}{k^2 + m^2} (1 - e^{-(k^2 + m^2)T}), \quad \forall k, m \geq 1.
\]

From now on, we drop the alternating terms \((-1)^{k+m}\) and the resulting matrix is still called \( A_N \). We denote by \(((\mu_N)_k)_{1 \leq k \leq N}\) the eigenvalues of \( A_N \) ordered decreasingly. Their asymptotic expansions are given in the following Proposition which improves the one in [2, Proposition 4.1].

**Proposition 3.1** We have the following bounds for the largest eigenvalue

\[
\frac{2(1 - e^{-2T})}{\pi^2} \ln N \leq (\mu_N)_1 \leq \frac{2(1 - e^{-2N^2T})}{\pi^2} N.
\]
Moreover, the eigenvalues \((\mu_N)_k\) satisfy
\[
(\mu_N)_{2k+1} \leq \gamma_T(\mu_N)_1 \exp\left(-\frac{\pi^2 k}{2 \ln(2N)}\right), \quad 1 \leq k \leq [(N - 1)/2].
\] (14)

The symbol \([\cdot]\) stands for the integer part. The constant \(\gamma_T\) depends on \(T\).

Proof: First of all, we have
\[
(\mu_N)_1 \leq \|A_N\|_{\infty} = \max_{1 \leq k \leq N} \sum_{m=1}^{N} a_{k,m}.
\]

The upper bound follows from
\[
\frac{km}{k^2 + m^2}(1 - e^{-(k^2+m^2)T}) \leq \frac{1}{2}(1 - e^{-2N^2T}).
\]

For the lower bound, we use the Collatz-Wielandt formula to get
\[
(\mu_N)_1 \geq \frac{4}{\pi^2}(1 - e^{-2T}) \min_{1 \leq k \leq N} \sum_{m=1}^{N} \frac{km}{k^2 + m^2}.
\]

Comparing integrals and sums, we get the lower bound. The proof of (14) is given in [2]. The proof is complete.

**Remark 3.1** Lemma 3.1 together with Cauchy’s Interlace Theorem for symmetric matrices (see [9]) that is
\[
(\mu_{N+1})_1 > (\mu_N)_1 > (\mu_{N+1})_2 > (\mu_N)_2 > \cdots > (\mu_{N+1})_N > (\mu_N)_N > (\mu_{N+1})_{N+1}.
\]
yields that the spectrum of \(A_T\) is actually a positive sequence stretched from zero to infinity and clustered at these extreme points. The exponential ill-posedness degree mainly comes from the fact that for a fixed \(k\), the eigenvalue \((\mu_N)_{N-k}\) decays exponentially fast towards zero when \(N\) grows to infinity. In the other hand, the unboundedness is caused by the fact that \((\mu_N)_k\) grows up to infinity, at worst like a constant times \(N\), when \(N\) tends to \(+\infty\).
4 Null-Controllability: Ill-conditioning

The instability for the null-controllability problem (4) is of a completely different nature. In spite of its well-posedness established in [6], the problem turns out to be intrinsically highly ill-conditioned. This equally implies heavy complications to handle it numerically (see [19]). The aim now is to find a control that drives the solution state variable $y_u$ from an initial condition $y_0$ to zero at a given final time $T$. Resuming the approach exposed in the previous section, we obtain a problem that looks like (10), where the final state $y_T$ corresponds to a trajectory initiated at $y(\cdot, 0) = y_0 \in L^2(I)$, with a homogeneous Dirichlet condition at $x = \pi$. All computations achieved, we derive the equation

$$\mathcal{L}^T(u_T)(k^2) = (-1)\frac{k \pi e^{-k^2T}}{2} \delta_{k,m}(y_0)_k, \quad \forall k \geq 1.$$

The construction of the HUM control is here again based on the representation in Lemma 2.1, and the vector $u^\dagger = (u^\dagger_m)_{m \in \mathbb{N}}$ solves the following equation

$$\sum_{m \geq 1} \frac{1 - e^{-(k^2+m^2)T}}{k^2 + m^2}(u^\dagger)_m = (-1)\frac{k \pi e^{-k^2T}}{2} \delta_{k,m}(y_0)_k, \quad \forall k \geq 1.$$

Put under a condensed form, it may be rewritten as

$$C_T u^\dagger = Q_T y_0 \quad \text{(15)}$$

with $C_T \in \mathbb{R}^{N^*} \times \mathbb{R}^{N^*}$ defined in the previous section and $Q_T$ is the diagonal matrix with entries

$$q_{k,m} = (-1)\frac{k \pi e^{-k^2T}}{2} \delta_{k,m} \quad \forall k, m \geq 1.$$

As indicated earlier, $C_T$ is compact in $\ell^2(\mathbb{R})$ and so is the operator $Q_T$.

Remark 4.1 Problem (15) may be also reworded in the same form as (12)

$$u^\dagger = (\mathcal{L}_T)^{-1}Dy_0.$$

The matrix $\mathcal{L}_T = -\text{diag } (\lambda_n T)C_T$ is a Löwner type matrix.
The null controllability problem (4) is well posed (see [6, 24]). A particular consequence is the following inclusion on the range subspaces $\mathcal{R}(Q_T) \subset \mathcal{R}(C_T)$. Equation (15) has thus a unique solution
\[ u^\dagger = (C_T)^{-1} Q_T y_0. \]
Recall that $u^\dagger$ is the HUM control and according to [6] and [32] we have
\[ \|u^\dagger\|_{L^2(0,T)} \leq \alpha_T \|y_0\|_{L^2(I)}. \]
This stability is equivalent to
\[ \sqrt{\langle y_0, Q_T(C_T)^{-1} Q_T y_0 \rangle_{L^2(\mathbb{R})}} \leq \alpha_T \|y_0\|_{L^2(\mathbb{R})}. \]
The smallest constant $\alpha_T$ is the control cost. Be aware that $(C_T)^{-1}$ is unbounded as the inverse of a compact operator. Nevertheless, the estimate above is possible because of the compactness of $Q_T$ which compensates and even dominates the unboundedness of $C_T$. Setting $(B_T)^{-1} = Q_T(C_T)^{-1} Q_T$ which is a symmetric positive definite operator, the above stability reads as
\[ \tau := \inf_{y \in \ell^2(\mathbb{R})} \frac{\langle y, B_T y \rangle_{\ell^2(\mathbb{R})}}{\|y\|_{\ell^2(\mathbb{R})}^2} > 0. \]
The constant $\tau$ coincides with $\frac{1}{(\alpha_T)^2}$ and is dependent on the final time $T$. It may be too small for fast controls. Indeed, it is proved in [10] (see also [28]), that $\alpha_T$ increases exponentially fast with $\frac{1}{T}$ for small $T$.

We explain now the causes of the instabilities of the null-controllability, despite its well-posed. The point then is to prove the high ill-conditioning of the operator $B_T$. Actually, the result we pursue is concerned with the principal sub-matrices $B_N$ of dimension $N$. We aim at showing that their condition numbers blow-up exponentially fast with respect to $N$. We start by writing down the coefficients of $B_T$,
\[ b_{k,m} = (-1)^{k+m} \frac{4}{\pi^2} \frac{km}{k^2 + m^2} (e^{(k^2+m^2)T} - 1), \quad \forall k, m \geq 1. \]
The important issue is to investigate the spectrum of $B_N$. Remark that the matrix $B_N$ is equivalent to the one with positive entries, obtained as the
absolute values of \((b_{k,m})_{k,m \geq 1}\). We subsequently consider this new matrix that we still denote abusively \(B_N\). Its coefficients are provided by
\[
b_{k,m} = \frac{4}{\pi^2} \frac{km}{k^2 + m^2} (e^{(k^2 + m^2)T} - 1), \quad \forall k, m \geq 1.
\]

Let \(((\tau_N)_k)_{1 \leq k \leq N}\) be the eigenvalues of \(B_N\) ordered decreasingly. Using the min-max principle, it is possible to state that the smallest eigenvalue \((\tau_N)_1 \geq \tau\) decreases towards \(\tau\), with large \(N\). Next, to figure out the asymptotics of the largest eigenvalues of \(B_N\), we need to define some auxiliary matrices. Let us define the diagonal matrix \(D_N = \text{diag} \{k^2\}_{1 \leq k \leq N}\) and the vectors \(\ell_N = (ke^{k^2T})_{1 \leq k \leq N}\) and \(\ell_N' = (k)_{1 \leq k \leq N}\). It is easily checked out that the matrix \(B_N\) satisfies the Sylvester or Lyapunov equation
\[
D_N B_N + B_N D_N = L_N,
\] with \(L_N = \ell_N \ell_N^* - (\ell_N')^*(\ell_N')\). The asymptotics of the eigenvalues of \(B_N\) depend on the properties of the diagonal matrix \(D_N\) and the displacement rank of equation (16) which is the rank of \(L_N\). Estimates established by T. Penzl in [22, 2000], and then improved by several authors are well-fit to provide the desired estimates. We refer to J. Sabino [25, 2006] and to the wide bibliography therein. The following result holds

\textbf{Lemma 4.1} The eigenvalues \(((\tau_N)_k)_{1 \leq k \leq N}\) of \(B_N\) satisfy the bound
\[
\frac{(\tau_N)_{2k+1}}{(\tau_N)_1} \leq \gamma_T \exp \left(-\frac{\pi^2 k}{2 \ln(2N)}\right), \quad 1 \leq k \leq [(N-1)/2].
\]

Proof: After observing that the displacement rank of equation (16) equals two, we apply the estimate in [25, Theorem 2.1.1] (see also [25, Formula (2.14) in p. 46]). We get the following bounds (with \(\kappa(D_N)\), the condition number of \(D_N\):
\[
\frac{(\tau_N)_{2k+1}}{(\tau_N)_1} \leq \gamma_T \exp \left(-\frac{\pi^2 k}{\ln(4\kappa(D_N))}\right)
\]
with a constant \(\gamma_T\) independent of \(N\). The displacement rank of the Lyapunov equation influences the counting of the eigenvalues. The condition number \(\kappa(D_N)\) being equals to \(N^2\), we derive the bound. \(\blacksquare\)
Lemma 4.2 The eigenvalues $((\tau_N)_k)_{1 \leq k \leq N}$ of $B_N$ satisfy the bound

$$\frac{(\tau_N)_N}{(\tau_N)_{N-2k}} \leq \gamma_T \exp \left(-\frac{\pi^2 k}{2 \ln(2N)} \right), \quad 1 \leq k \leq [(N - 1)/2].$$

The constant $\gamma_T$ is independent of $N$.

Proof: The proof is a fac-simile of the previous one. The difference is that we need to transform the Lyapunov equation (16) as follows

$$D_N(B_N)^{-1} + (B_N)^{-1}D_N = (B_N)^{-1}L_N(B_N)^{-1}.$$ 

The displacement rank is not modified and remains equal to two. Applying once again [25, Theorem 2.1.1] yields that

$$\frac{[(\tau_N)_{N-2k}]^{-1}}{[(\tau_N)_N]^{-1}} = \frac{(\tau_N)_N}{(\tau_N)_{N-2k}} \leq \gamma_T \exp \left(-\frac{\pi^2 k}{2 \ln(2N)} \right).$$

The proof is complete.

Remark 4.2 An immediate consequence is that the condition number $\kappa_{B_N}$ of matrix $B_N$ blows up at least exponentially fast with $N$. This is why solving numerically the null-controllability is a hard computational task. To bound $\kappa_{B_N}$ from below we observe that, in view of the symmetry and the positivity of the matrix $B_N$, we have that

$$\kappa_{B_N} = \frac{(\tau_N)_1}{(\tau_N)_N}.$$ 

Selecting $k = [(N - 1)/2]$ in Lemma 4.1, we obtain the following preliminary bound

$$\kappa_{B_N} \geq \gamma_T \exp \left(\frac{\pi^2 N}{4 \ln(2N)} \right).$$

The general theory initiated by T. Penzl [22] about the spectrum of the solutions of Sylvester equations with a low displacement rank predicts that, for a given $N$, the eigenvalues $(\tau_N)_{N-k}$ grows faster than $\gamma_T \exp \left(\frac{\pi^2 k}{4 \ln(2N)} \right)$. This result is not optimal. Actually, more direct algebraic tools enable to derive an accurate result, at least about $(\tau_N)_1$. 

13
Proposition 4.1 We have
\[
\frac{1}{2N}(e^{(N^2+1)T} - 1) \leq (\tau_N)_1 \leq \frac{N}{2}(e^{2N^2T} - 1).
\]

Proof: Bounding \((\tau_N)_1\) may be obtained owing to the double formula (see [17, Chapter 8]),
\[
\min_{1 \leq k \leq N} \sum_{1 \leq m \leq N} \frac{km}{k^2 + m^2}(e^{(k^2+m^2)T} - 1) \leq (\tau_N)_1 \leq \max_{1 \leq k \leq N} \sum_{1 \leq m \leq N} \frac{km}{k^2 + m^2}(e^{(k^2+m^2)T} - 1).
\]
We deduce that
\[
\frac{N}{1+N^2}(e^{(N^2+1)T} - 1) \leq (\tau_N)_1 \leq \frac{N}{2}(e^{2N^2T} - 1).
\]
The proof is complete.

Remark 4.3 A direct result of the proposition is that the condition number \(\kappa_{BN}\) satisfies the more accurate minoration
\[
\kappa_{BN} \geq \frac{\gamma_T}{N} e^{N^2T}.
\]

5 Numerical Discussion

We conduct a numerical investigation of the Gramians of the controllability problems from the spectra point of view. The aim is to assess the theoretical predictions and to illustrate the instabilities that arise in the computations for the (boundary) controllability of diffusion process. It is also the opportunity to underline the important role the structured matrices have in elucidating those instabilities. We run first some computations on the eigenvalues \((\mu_k)_{k \in \mathbb{N}}\) of \(A_T\), the Gramian for the exact-controllability and \((\tau_k)_{k \in \mathbb{N}}\) of \(B_T\), the Gramian for the null-controllability. Next, we consider the exact-controllability problem where the final target \(y_T\) is non-smooth and the null-controllability with a given non-smooth initial condition. The control \(u^*\) is obtained by the inversion of the systems (15) and (10). All these computations are made within MATLAB.
Spectra of the Gramians. We rather check the truncated matrices $A_N$ and $B_N$ for some values of $N$. Computations are obtained for cut-off parameter $N$ ranging from 3 to 15. Figure 1 depicts the spectra of $A_N$ and $B_N$ for different control times $T = 0.25$, 1 and 1.5. The panels to the left are related to the exact-controllability and those to the right correspond to the null-controllability. The main feature of the spectra for the exact-controllability Gramian has already been pointed at in [2]. Most of the eigenvalues are clustered around zero when $N$ grows higher. Actually, as shown in the estimate (14), they decrease rapidly toward zero. This is a typical sign of severe ill-posedness which causes strong instabilities for the exact-controllability problem. Notice that the largest eigenvalues grow slowly towards infinity but has a tiny incidence on the instabilities of the problem. The final conclusion here is that the ill-posedness degree of the exact-controllability is exponential. A glance to the results for $T = 0.25$, shows that ill-posedness is aggravated when a fast controllability is aimed at.

For the null-controllability, the observations concerning the spectra are different, even though their structure is also cause of heavy lack of stability. For a moderate control time that is $T = 1$ or 1.5, the smallest eigenvalue are clearly bounded away from zero. This is a consequence (and also an indication) of the well-posedness of the null-controllability problem. More significantly, the information one may draw from the distribution of those eigenvalues are twofold. In one hand side, despite the well-posedness, the Gramian and then the problem (15) are highly ill-conditioned. The condition number of the principal sub-matrices $(B_N)^N$ of $B_T$ (with $T = 1$) increases tremendously fast to reach extremely high values; for instance, $\kappa_{B_{14}}$ comes close to $10^{200}$. This predicts that the corresponding algebraic problem is too hard to invert. In the other side, when a fast control is aimed at, as for $T = 0.25$, a fraction of the eigenvalues become dangerously small, and things happen as if the null-controllability were ill-posed. This sounds in agreement with the result found in [10] (see also [28]), where the control cost is proved to blow-up exponentially fast in $\frac{1}{T}$. Actually the singular behavior seems highly stiff with respect to $\frac{1}{T}$. To make an accurate estimation of that behavior, Figure 2 depicts in a semi-logarithmic scale the curve of the variations of the smallest eigenvalue $(\tau_N)_N$ ($N = 15$) with respect to $\frac{1}{T}$ for $T$ ranging in $[0.3, 1.5]$. The linear regression yields that $(\tau_N)_{T} \approx 70 \exp(-\frac{7.5}{T})$. This means that the control cost is such that $\gamma_T = \frac{1}{\sqrt{(\tau_N)_N}} \approx 0.12 \exp(3.75 T)$.

Before closing the chapter on the spectra of the Gramians there remains...
to find out the behavior of the eigenvalues $\mu_N^k$ and $\tau_N^k$ with respect to $k$, when $N$ is fixed. In Figure 3 are plotted, again in semi-logarithmic scales, the curves of those eigenvalues with respect to $k$ when $N$ is fixed to 15, and for two final control times $T = 0.25$ and $T = 1.0$. The panel to the left is for the exact-controllability spectra $(\mu_{15})_{1\leq k \leq 12}$ and the one to the right is for the null-controllability eigenvalues $(\tau_{15})_{1\leq k \leq 15}$. The shape of the curves suggests that $(\tau_{15})^k$ increases as $\exp(\eta_T k^2)$ while $(\mu_{15})^k$ decreases like $\exp(-\eta_T' k)$ with $\eta_T, \eta_T' > 0$. The slight bending of both curves of the exact-controllability makes us feel that $(\mu_{15})^k$ rather decreases like $k^\alpha \exp(-\eta_T' k)$ for some $\alpha, \eta_T' > 0$.

**Exact controllability.** We start by the investigating the exact-controllability problem (5). As already noticed in [2, Section 6], due to the exponential ill-posedness, incautiously processing the truncated system (11) without any regularization produces erratic results. The Lavrentiev procedure is therefore used for stabilizing the computations. We are hence lead to compute the solution $u_\tilde{p}$ of the regularized problem (compare with (11))

$$
\frac{\zeta(u_\tilde{p})_k}{\kappa^2 + m^2} + \sum_{m \geq 1} \frac{1 - e^{-(k^2 + m^2)T}}{\kappa^2 + m^2} (u_\tilde{p})_m = (-1)^{k+1} \pi \frac{(y_T)_k}{k}, \quad \forall k \geq 1. \quad (17)
$$

This system is symmetric positive definite as long as $\zeta > 0$. A judicious selection of $\zeta$, so to ensure a balancing between the accuracy and the stability of the computations, is an important point and has been widely addressed for ill-posed algebraic systems. We refer for instance to [5] for a wide exposition about that issue.

Now, we seek for a Dirichlet control function $u(\cdot)$ that allows to realize a prescribed target $y_T$ after a time duration $T = 1$. That state, we hope to reach (numerically), is represented in the left diagram of Figure 4 and is given by $y_T(x) = \chi(x)$, with

$$
\chi(x) = \begin{cases} 
  x, & \forall x \in [0, \frac{\pi}{2}], \\
  \frac{\pi}{2}, & \forall x \in [\frac{\pi}{2}, \pi],
\end{cases}
$$

The singularity at $\pi/2$, the mid point of the interval, contained in the state $y_T$ makes it non-exactly reachable. However, owing to the approximate controllability result, we may steer the trajectory arbitrarily close to that $y_T$. The curves to the left in Figure 4 plot the numerical states computed by
the Lavrentiev method as in equation (17). The cut-off frequency is fixed to $N = 150$ which guarantees a rich representation of $u(\cdot)$ in the exponential basis given in Proposition 2.1. Various values of the parameter $\varrho$ are experienced. Here we retain $\varrho = 10^{-3}, 10^{-4}$ and $10^{-5}$. Actually, $\varrho = 10^{-4}$ seems to provide the best numerical simulation. Nevertheless, one may consider that, in all these cases, the state $y_T$ is approximated with a satisfactory accuracy. In the contrary, a glance to the shape of the controls $u^\dagger_\varrho$, depicted in the right panel shows that it is sensitive to the variations of the parameter $\varrho$. The $L^2$-norm of $u^\dagger_\varrho$, recorded in Table 1, corroborates such an observation. This has to do with the severe ill-posedness. Before switching to the next example, let us point out the smoothness of the computed controls away from the final time $T$, which is in agreement with the discussion in Section 2.

<table>
<thead>
<tr>
<th>$\varrho$</th>
<th>$10^{-3}$</th>
<th>$10^{-4}$</th>
<th>$10^{-5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$|u^\dagger_\varrho|_{L^2(0,T)}$</td>
<td>6.17</td>
<td>7.25</td>
<td>7.75</td>
</tr>
</tbody>
</table>

Table 1: Dependence of the norm of the computed controls with respect to $\varrho$, for the exact-controllability.

**Null controllability.** Consider now the null-controllability problem (4). The initial state coincides with $\chi$, the one used in the previous simulation as the final state. Actually, we intend the construction of a Dirichlet control $u(\cdot)$ that drives the trajectory to zero after a time $T = 0.5$. The results are obtained by solving equation (15) after a Lavrentiev regularization as in (17). In the left panel of Figure 5, are plotted the initial state, the one related to the absence of control (i.e. with null Dirichlet condition at $x = \pi$) and the controlled states for three values of the regularizing parameter $\varrho = 10^{-3}, 10^{-4}$ and $10^{-5}$. The controls represented in the right panel and their norms provided in Table 2 illustrate the important sensitivity of the control $u^\dagger_\varrho$ to $\varrho$. The variations of the norm with respect to $\varrho$ is more significant here than in the first example, due to a shorter controllability time: here, $T = 0.5$ versus $T = 1$ previously. Before closing the numerical chapter, let us draw the attention that MATLAB fails in inverting the (Löwner) matrix $(Q_N)^{-1}C_N$ for $N \geq 38$, despite the fact that the exact-counterpart $(Q_T)^{-1}C_T$ is arithmetically continuously invertible. The reason is that the condition number of this matrix grows up violently with $N$ as predicted in Proposition 4.1. For an illustration, MATLAB evaluates the condition number of that matrix to
$0.24 \times 10^{28}$ for $N = 10$ and to $0.5 \times 10^{305}$ when $N = 37$.

<table>
<thead>
<tr>
<th>$\varrho$</th>
<th>$10^{-3}$</th>
<th>$10^{-4}$</th>
<th>$10^{-3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$|u^\varrho_k|_{L^2(0,T)}$</td>
<td>7.41</td>
<td>13.18</td>
<td>22.41</td>
</tr>
</tbody>
</table>

Table 2: The norm of the computed controls versus $\varrho$, for the null-controllability.

6 Conclusion

We aim to know a little more about the instabilities of the (boundary) controllability of the heat equation. This may help adapting, improving and strengthening (already) cautious numerical approximations. Understanding the exponential ill-posedness of the exact-controllability, initiated in [2], is deepened and sharpened. The analysis of the null-controllability problem unveils its intrinsic instability. Despite its well-posedness, this problem turns out to be badly-conditioned in a very high extent. The spectrum of the Gramian grows violently. Moreover, when fast controls are sought for, the smaller fraction of the spectrum is substantially shifted toward zero. The problem suffers then from two drawbacks. In addition to the stiffening of the condition number, it reacts somehow like a severe ill-posed problem as predicted in [10, 28]. The feasible approaches used nowadays by engineers for the null-controllability consist in tackling it as an exact-controllability problem with a reachable state. The alternative that is to design performing algorithms that would specifically take profit of the well-posedness is likely doomed to fail. The most credible approach, at least so far, seems to be the one exposed in [19], for instance.

References


Figure 1: Eigenvalues of $A_N$ (panels in the left) and $B_N$ (panels in the right) for various cut-offs $N$. The top diagrams are for $T = 0.25$, the middles for $T = 1$ and the bottoms for $T = 1.5$. Notice the scales in the horizontal axis, $10^{16}$ for $A_N$ and $10^{260}$ for $B_N$. 
Figure 2: The variation of the smallest eigenvalue \((\tau_N)_N\) with respect to \(\frac{1}{T}\) (we fix \(N = 15\)).
Figure 3: The left panel depicts the dependence of $(\mu_{15})_k$ with respect to $k$. The right panel is for the variation of $(\tau_{15})_k$ in $k$. 
Figure 4: The exact target $y_T$ and the numerical final states obtained by Lavrentiev’s method with various regularization parameters $\rho$ (left panel). The related controls (right panel).
Figure 5: The null-controllability example. The initial state \( y_0 \) (plain line), the one obtained after a time \( T = 0.5 \), without any control and the controlled states obtained by various regularization parameters \( q \) (left panel). The related controls (right panel).