Control in moving interfaces and deep learning

Borjan Geshkovski

Departamento de Matemáticas
Universidad Autónoma de Madrid

PhD Dissertation
Advisor: Enrique Zuazua Iriondo
2021

Acknowledgments

Avant tout je voudrais exprimer toute ma gratitude à mon directeur de thèse, Enrique Zuazua. J'ai pu bénéficier au long de cette thèse de sa vision claire des mathématiques, de sa capacité de réduire le plus difficile au plus simple, et de sa grande culture scientifique. Je lui suis aussi reconnaissant de sa disponibilité perpetuelle. Arduraren eta dedikazioren adibide izan zara, eskerrik asko.

Cette rencontre avec lui et son équipe, je la doit à Marius Tucsnak, qui m'a incité à le contacter lorsque j'étais en master; un grand merci à lui pour ce conseil fructueux, mais également pour son soutien depuis mes études de master.

I thank all the members of the department of mathematics of UAM, for their generosity, patience and understanding during my time as a PhD student. I highlight Antonio Cuevas – for his help and guidance as coordinator of the doctoral program and for his deep insights regarding my work, Mar Gonzalez – for the interest she took in my work and for her patience during my lengthy presentations, Rafa Orive – for inviting me to give my first ever conference talk (at ICMAT), Antonio Cuevas (again) and Jávier Carcamo for organizing my talk in the Statistics chair, as well as Alejandro Mas and Fran Mengual for being friendly office mates in despacho 613 módulo 17. I also thank C. Burgos and E. Touris, as well as Marta Sánchez Calvo for her help with sorting out the highly nontrivial bureaucratic challenges up to halfway my PhD.

I also thank all of my colleagues at CCM Deusto, from whom I've learned so much during these years. In particular, thank you Azahar M., Alejandro C., Darlis B., Dongnam K., Jan H., Jon Asier B., Jorge M., Sébastian Z., Sergi A., Sergio P., Swann M., Thibault L, and Umberto B. I especially thank Martin Lazar for his reactivity, kindness and the interest he showed in my work. I am particularly indebted to Miren – for being a friendly guide throughout, to Jésus – por su creatividad, to Idriss – pour sa culture mathématique et littéraire inépuisable, to Dario – for his tenacity, to Debayan – for being a great teacher, to Carlos – for always finding the correct argument, and to Domènec – for his chaotic brilliance and Catalan mettle.

I salute all the members of the ConFlex project of which I've been part these past two and a half years. I greatly enjoyed the two in-person workshops, and I in particular thank G. Weiss and H. Zwart, as well as A. Mattioni, A. Muñoz, J. Zhang, M. Artola, N. Skrepek and P. Lorenzetti, for illuminating discussions and advice.

I thank Günter Leugering, Martin Gugat, Lukas Pflug, Martin Burger, Daniel Tenbrinck and Leon Bungert, for all the insightful discussions during my stays at FAU in Erlangen. I also thank Daniel and Leon for inviting me to give a talk in their colloquium. I once again thank Günter Leugering for his kindness, his guidance in various situations, and for stimulating discussions.

I thank all the mathematicians and scientists who have directly or indirectly helped and guided my research during my PhD – I recall R. Nochetto, E. Trélat, C. Seis, P. Lissy, F. Sueur, M. Léautaud, E. Dupont, J.P. Borthagaray, and V. Perollaz.

I thank my friendly flatmates Brooke, Léo, Yasmine, Kate and Matí from *l'Auberge Espagnole* on Mazarredo, for always raising my spirits during the (little) time I was at home.

Je tiens a remercier la famille Rodriguez pour tous les bons moments passés lors de mes séjours en France durant l'été et à Noël. Eric, Valérie, Laure, Lucas, Christophe, Angelina, Thomas, Louise, Lucien et Maïté, merci à vous. Merci aussi à Laurence. Merci à Alex et Solenne, de m'avoir toujours tenu au courant des nouveaux développements culturels en France. Merci à Jérôme et Espérance pour leur soutien et encouragement tout au long de ces années.

Не би можел да заминам, пред речиси девет години, на студии во странство без поддршката од моето семејство. Ова дело им го посветувам ним. На мојата баба Блага и покоен дедо Мирко, кои секој пат беа насмеани при моите посети во Скопје. На мојот брат Вангели и сестра Арна, за нашиот заеднички хумор. На мојот покоен татко Васил, од кого научив англиски и математика. На мојата мајка Јасмина, за нејзината доблест, мудрост и љубов.

Enfin, j'adresse le plus grand merci à Charlotte.

Funding. This thesis has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No.765579-ConFlex.

Abstract

This thesis brings forth several contributions to the controllability theory of free boundary problems, to the turnpike property for nonlinear optimal control problems, and to the modern theory of deep supervised learning.

In Part I, we set-up a systematic methodology for the exact-controllability of free boundary problems, governed by diffusive partial differential equations, to specific, possibly nontrivial targets, by combining a careful study of the linearized problem and fixed point arguments. We distinguish problems wherein the linearization is either controllable by using spectral techniques for deriving the needed observability inequality (e.g. when controlling the one-dimensional porous medium equation to its self-similar Barenblatt trajectory) or by a combination of Carleman inequalities with compactness arguments (in the context of a free boundary problem for the one-dimensional viscous Burgers equation, where steering the free boundary is seen as a finite-dimensional constraint on the control). We emphasize the importance of controlling not only the solution of the PDE, but also the free boundary, to a prescribed configuration.

This analysis culminates with the controllability study of the one-phase Stefan problem with surface tension set in a strip-like geometry in two space dimensions. Using a control actuating along the flat bottom, under smallness conditions on the initial data, we prove the null-controllability of both the temperature and the position of the free boundary in any positive time. The null-controllability of the linearized problem is covered by means of a Fourier decomposition in the periodic horizontal variable, and null-controllability results uniform with respect to the Fourier parameter of the one-dimensional problems, obtained using spectral techniques for the non-zero Fourier modes, with the zero mode system being seen as a control problem with a finite-dimensional control constraint. We conclude by commenting on the feasibility (or rather, lack thereof) of performing a vanishing surface tension limit in view of addressing the control properties of the classical Stefan problem.

In Part II, we present a new proof of the turnpike property for nonlinear optimal control problems, when the running target is a stationary solution of the free dynamics. By using of sub-optimal quasi-turnpike trajectories (via a controllability assumption) and a bootstrap argument, and bypassing an analysis of the optimality system or linearization techniques, we are able to address finite-dimensional, control-affine systems with globally Lipschitz nonlinearities. We show that our methodology is applicable to controlled PDEs as well, such as the semilinear wave and heat equation with a globally Lipschitz nonlinearity.

In Part III, we study the behavior of supervised learning problems for neural ODEs when the final time horizon T is increased, a fact that may be interpreted as increasing the depth in the associated residual neural network (ResNet) setting.

For the classical L^2–regularized empirical risk minimization problem, under homogeneity assumptions on the neural ODE dynamics, we prove that the training error decays to zero with a (almost) polynomial rate when T goes to infinity. In the context of regression tasks, the optimal parameters are furthermore shown converge to minimal L^2–norm parameters in the interpolation regime. Moreover, a natural scaling between the time horizon T and the parameter regularization λ appears, and we therefore obtain the same

convergence results when λ goes to zero and the horizon is fixed. These results thus allow us to stipulate generalization properties in the overparametrized regime, and are aligned with results on regularization path convergence (λ to zero) and implicit regularization of gradient descent for linear models or two-layer perceptrons.

Following the insight of Part II, we also propose an augmented learning problem by adding an artificial regularization term of the state trajectory over the entire time horizon. Applying the turnpike results, we obtain an exponential rate of decay for the training error and for the optimal parameters in any time – an improved estimate for the depth required to reach almost perfect training accuracy.

In the context of the augmented learning problem with L^1–parameter regularization, and under homogeneity assumptions on the dynamics (typical for ReLU activations), we prove that any global minimizer is sparse, in the sense there exists a positive stopping time beyond which the optimal parameters vanish. In practical terms, when extrapolated to the ResNet context, a shorter time-horizon in the optimal control problem can be interpreted as considering a shallower ResNet, which lowers the computational cost of training. We may also provide quantitative estimates on the stopping time, and on the training error of the neural ODE trajectories at the stopping time. The latter stipulates a quantitative approximation property of neural ODE flows with sparse parameters.

Keywords. Controllability, free boundary problems, viscous Burgers equation, thin-film equation, porous medium equation, degenerate parabolic equation, Stefan problem, phase transitions, Gibbs-Thomson correction, surface tension, optimal control, turnpike theory, nonlinear systems, stabilization, deep learning, ResNets, neural ODEs, regularization path, generalization, sparsity.

Resumen

En esta tesis se aportan varias contribuciones a la teoría de control para problemas de frontera libre, a la teoría de Turnpike para problemas de control óptimo no lineales, y a la teoría de aprendizaje supervisado profundo (supervised deep learning).

En la Parte I, se establece una metodología sistemática para la controlabilidad exacta a estados específicos, posiblemente no triviales, de problemas de frontera libre gobernados por ecuaciones en derivadas parciales difusivas. Para ello se combina un estudio cuidadoso del problema linealizado con argumentos de punto fijo. Distinguimos dos tipos de problemas: en uno el problema linealizado es controlable mediante el uso de técnicas espectrales que permiten derivar la desigualdad de observabilidad necesaria (por ejemplo, cuando se controla la ecuación de medios porosos unidimensional a su trayectoria de Barenblatt auto-similar); en el otro se combinan desigualdades de Carleman con argumentos de compacidad (en el contexto de un problema de frontera libre para la ecuación de Burgers viscosa unidimensional, donde la dirección de la frontera libre se ve como una restricción de dimensión finita para el control). En este contexto, destacamos la importancia de controlar, no solo la solución de la EDP, sino también la frontera libre a una configuración prescrita.

El análisis se culmina con un estudio de la controlabilidad para un problema de Stefan de una fase con tensión superficial, planteado en un domino similar a una banda de dimensión dos. Utilizando un control que actúa a lo largo del fondo plano, en condiciones de pequeñez del dato inicial, probamos controlabilidad a cero en cualquier horizonte temporal, tanto de la temperatura como de la posición de la frontera libre. La controlabilidad a cero del problema linealizado se aborda mediante una descomposición de Fourier en la variable horizontal periódica y resultados de controlabilidad a cero uniformes respecto al parámetro de Fourier del problema unidimensional. Éstos se obtienen mediante técnicas espectrales para los modos de Fourier no nulos, con el sistema de modo cero visto como un problema de control con una restricción de control de dimensión finita. Concluimos con una discusión sobre la viabilidad (o más bien, la falta de ella) de implementar técnicas de tensión superficial evanescente para abordar la controlabilidad del problema clásico de Stefan.

En la Parte II, presentamos una nueva demostración de la propiedad de turnpike para problemas de control óptimo no lineales, para casos en los que el estado objetivo es una solución estacionaria de la dinámica libre. Combinando la construcción de trayectorias cuasi-Turnpike subóptimas (bajo hipótesis de controlabilidad) con un argumento de tipo bootstrap, y sin tener que depender del análisis del sistema de optimalidad o técnicas de linealización, somos capaces de establecer resultados de turnpike para sistemas no lineales en dimensión finita, con control afín, y con dinámica globalmente Lipschitz. Además, demostramos que nuestra metodología también es aplicable a EDPs controladas, como la ecuación de ondas semilineal y la ecuación del calor semilineal con no linealidad es globalmente Lipschitz.

En la Parte III, estudiamos el comportamiento de problemas de aprendizaje supervisado para EDOs neuronales cuando se incrementa el horizonte temporal T, hecho que se puede interpretar como un aumento de la profundidad de la red neuronal residual (ResNet) asociada.

Para el problema clásico de minimización del riesgo empírico con una regularización L^2 de los parámetros, bajo la hipótesis de homogeneidad de la dinámica, probamos que el error sobre el conjunto de datos de entrenamiento decae a cero con una tasa (casi) polinomial cuando T tiende a infinito. En el contexto de problemas de regresión, mostramos además que los parámetros óptimos convergen a los parámetros de norma L^2 mínima que interpolan los datos de entrenamiento. Además, como consecuencia de un cambio de escala entre el horizonte temporal T y el hyper-parámetro de regularización λ, los mismos resultados de convergencia se pueden obtener cuando λ tiende a cero y el horizonte temporal es fijo. Estos resultados nos permiten estipular propiedades de generalización en el régimen sobreparametrizado (overfitting), y se encuentran en la misma línea que otros resultados existentes de convergencia para la trayectoria regularizada límite (λ tiende a cero) y la regularización implícita del gradiente descendente para modelos lineales o perceptrones de dos capas.

Siguiendo las ideas de la Parte II, también proponemos un problema de aprendizaje aumentado agregando un término artificial de regularización para la trayectoria del estado en todo el intervalo de tiempo. Aplicando los resultados de turnpike, obtenemos una tasa de decaimiento exponencial para el error de entrenamiento y para los parámetros óptimos en toda la trayectoria. Esto da lugar a una estimación mejorada de la profundidad requerida para asegurar una precisión de entrenamiento prefijada.

En el contexto de los problemas de aprendizaje aumentado con regularización L^1 de los parámetros, y bajo supuestos de homogeneidad de la dinámica (típico de funciones de activación de tipo ReLU), demostramos que cualquier minimizador global es sparse or ralo, en el sentido de que existe un tiempo de parada, a partir del cual, los parámetros óptimos son nulos. En términos prácticos, cuando se extrapola al contexto ResNet, un horizonte temporal más corto en el problema de control óptimo puede interpretarse como una ResNet menos profunda, lo que reduce el coste computacional del entrenamiento. También proporcionamos estimaciones cuantitativas sobre el tiempo de parada y sobre el error de entrenamiento de las trayectorias óptimas de la EDO neuronal. Este resultado estipula una propiedad de aproximación cuantitativa para EDOs neuronales con parámetros sparse.

Palabras clave. Controlabilidad, problemas de frontera libre, ecuación de Burgers viscosa, ecuación de película delgada, ecuación de medios porosos, ecuaciónes parabólicas degeneradas, problema de Stefan, transiciones de fase, corrección de Gibbs-Thomson, tensión superficial, control óptimo, teoría de Turnpike, sistemas no lineales, estabilización, aprendizaje profundo, ResNets, EDO neuronales, trayectoria de regularización, generalización, sparsidad.

Contents

Chapter 1

Introduction

The field of *control* provides the principles and methods used to design inputs which ensure that systems – arising in common physical, biological or sociological contexts–, reach a desired configuration in some time, or maintain a desirable performance over time. By leveraging technological improvements in sensing and computing with breakthroughs in the underlying mathematics, control strategies have become ubiquitous in most applied fields, including manufacturing, electronics, communications, transportation, networks, and many military systems.

The core object of use in modern control theory are dynamical systems, namely systems which evolve over time, governed by *ordinary* or *partial differential equations*. In principle, they all take the form

$$\begin{cases} \dot{y}(t) = f(y(t), u(t)), \\ +\text{initial conditions} \end{cases}$$

where $y(t)$ represents the state of the system (which could be, for instance, the velocity of a fluid, the temperature of a body, the vibration of a string, and so on), and $u(t)$ represents the control input. Such formalisms go as back as the XVII$^{\text{th}}$ century and the back to inventors of differential calculus – Newton and Leibniz. For instance, Newton's second law, which relates the acceleration of an object with the forces which are applied to it, takes the form

$$\ddot{y}(t) = -g$$

where g is the gravitational constant. This is an ordinary differential equation, describing the free fall of a ball in air; it is called *ordinary* since the number of unknowns is finite (here, only the height of the ball). Partial differential equations, on the other hand, involve an infinite amount of unknowns. For instance, when a string of length equal to 1 vibrates, denoting $y(t, x)$ the displacement of the string at position x (with x between 0 and 1 denotes the horizontal position), one sees that there is an infinite amount of positions x such that $x \in [0, 1]$. The equation describing the evolution of the string thus involves not only the time derivatives, but also the spatial derivatives (and thus, derivatives are partial, as they are taken with respect to one out of multiple variables); it takes the form

$$\begin{cases} \partial_t^2 y - c^2 \partial_x^2 y = 0 & \text{for } \ (t, x) \in [0, T] \times [0, 1] \\ y(t, 0) = u(t), \quad y(t, 1) = 0 & \text{for } \ t \in [0, T] \\ +\text{initial conditions} \end{cases}$$

where c denotes the velocity of the waves in the string. The string equation is the first partial differential equation to be formulated in such form, as derived by d'Alembert in 1747. The control input $u(t)$, representing the height of the string at the left extremity, can be of *open-loop* (where it is independent of the displacement y) or *feedback* form

(where it is a function of the displacement y) – we shall solely focus on the former in this thesis.

In the open-loop context, the notion of *exact-controllability* is the most natural objective one could address: given an initial displacement y_0 and a target y_1, can one find a control input $u(t)$ which steers the state from y_0 in time 0 to y_1 in time T? One could naturally expect such an objective to be costly in practice, and it can be alleviated by considering different variants if needed, such as almost reaching the target (*approximate controllability*), or reaching the target in infinite time (*stabilization*). A general paradigm, pioneered successfully in many practical contexts, consists in formulating control problems via optimization – for instance, by minimizing some cost

$$\text{minimize} \quad \text{Cost}(y, u).$$

This is a rather natural problem, which traces its roots to the calculus of variations and continuum mechanics, and is nowadays at the core of *optimal control theory*, pioneered in the 1950s by figures such as Bellman (and his dynamic programming principle) and Pontryagin (and his maximum principle).

Note that the spatial domain in the string equation formulated above is one-dimensional. This is certainly not always the case in nature, where different quantities and phenomena may evolve in two or (generally) three space dimensions, in which case, one replaces the interval by a more general higher dimensional domain. Furthermore, the domain itself could be time-dependent, and in fact, an unknown of the problem itself. This is the case of *free boundary problems* (also called *moving interface problems*), wherein part of the boundary of the domain is an unknown, with its evolution being governed by another differential equation. These problems are omnipresent in nature and engineering. For instance, in models of oceanic water waves ([170, 171]), the unknown fluid velocity is governed by the Euler equations in the fluid domain, but the latter is bounded from above solely by the free surface of the fluid (which represents the free boundary), which evolves with the velocity. Free boundary problems constitute the prototypical models in contact problems in elasticity [109, 270], fluid-structure interaction [238, 98, 114], water waves and free surface flows [52, 53, 54], phase transitions [134, 68], stock pricing in finance [172, 58], transonic shock waves [59], and so on. Due to the immense impact of these systems in modern industrial and societal contexts such as cardiovascular modeling [106, 74], coastal engineering for wind and wave energy harvesting [29, 79], and ice sheet forecasts [240, 241], the control properties thereof are of particular interest, and represent one of the directions addressed in this thesis.

The advent of big data in the past decade and the exponential increase in computing power have led *data-driven* and *machine learning* methods to become a new frontier in both theoretical and applied research, due to their universal applicability. In certain contexts, they can supersede classical control methodologies for physically derived models based on partial differential equations. Part of this is due to the incredible success of *deep learning* [176], which makes use of models called *neural networks*, which represent iterated compositions of simple nonlinearities and parametric affine maps. Yet in most contexts, neural networks can be reinterpreted as controlled dynamical systems, with optimizable parameters playing the role of controls. This leads one to view many deep learning paradigms as compound optimal control problems. Mastering the stability and regulating and tuning the various free hyper-parameters to derive simplified architectures is a challenge in which the analytical and computational methods of control theory find a natural and important application, as illustrated in this thesis.

1.1 Contributions of the thesis

This thesis brings forth several contributions to the controllability theory of free boundary problems, to the long-time behavior of nonlinear optimal control problems, and to the modern theory of deep supervised learning. We review these contributions hereafter.

Part I. *Controllability of free boundary problems.*

- In Chapter 2, we address the local controllability of a one-dimensional free boundary problem for a fluid governed by the viscous Burgers equation. The free boundary manifests itself as one moving end of the interval, and its evolution is given by the value of the fluid velocity at this endpoint. We prove that, by means of a control actuating along the fixed boundary, we may steer the fluid to constant velocity in addition to prescribing the free boundary's position, provided the initial velocities and interface positions are close enough.

  Chapter 2 is taken from [116]:

  > *Controllability of one-dimensional viscous free boundary flows.*
  > B. Geshkovski and E. Zuazua, 2019.
  > https://hal.archives-ouvertes.fr/hal-02277740/
  > Accepted for publication in SIAM J. Control Optim.

- In Chapter 3, we investigate the null-controllability of a nonlinear degenerate parabolic equation, which is the equation satisfied by a perturbation around the self-similar solution of the porous medium equation in Lagrangian-like coordinates. We prove a local null-controllability result for a regularized version of the nonlinear problem, in which singular terms have been removed from the nonlinearity. We use spectral techniques and the source-term method to deal with the linearized problem and the conclusion follows by virtue of a Banach fixed-point argument. The spectral techniques are also used to prove a null-controllability result for the linearized thin-film equation, a degenerate fourth order analog of the problem under consideration.

  Chapter 3 is taken from [115]:

  > *Null-controllability of perturbed porous medium gas flow.*
  > B. Geshkovski.
  > ESAIM: COCV, **26**, 85-105, 2020.
  > https://doi.org/10.1051/cocv/2020009

- In Chapter 4, we study the controllability properties of the one-phase Stefan problem with surface tension set in a strip-like geometry in two space dimensions, a system which may be seen as a singular perturbation of the classical Stefan problem via a regularizing term on the free boundary. Using a control actuating along the fixed flat bottom, under smallness conditions on the initial data, we prove the null-controllability of both the temperature and the position of the free boundary in any positive time. Our techniques rely on a careful analysis of the linear problem, which is obtained after fixing the domain via a harmonic extension diffeomorphism. The null-controllability of the linearized problem is covered by means of a Fourier decomposition in the periodic horizontal variable, and null-controllability results uniform with respect to the Fourier parameter of the one-dimensional problems. The latter are obtained using spectral techniques for the non-zero Fourier modes, whereas the zero mode system is seen as a controllability problem with a finite-dimensional constraint. The nonlinear problem may be tackled by combining an adaptation of the so-called source-term method, and a Banach fixed-point argument. We comment on the feasibility (rather, lack thereof) of performing a vanishing surface tension limit in view of deriving the controllability of the classical Stefan problem.

  Chapter 4 is a work in collaboration with D. Maity.

Part II. *Long-time optimal control.*

- In Chapter 5, we present a new proof of the turnpike property for nonlinear optimal control problems, when the running target is a stationary solution of the free dynamics. Our strategy combines the construction of sub-optimal quasi-turnpike trajectories (via a controllability assumption) and a bootstrap argument, and does not rely on analyzing the optimality system or linearization techniques. This in turn allows us to address finite-dimensional, control-affine systems with globally Lipschitz (possibly nonsmooth) nonlinearities. We show that our methodology is generic and applicable to controlled PDEs as well, such as the semilinear wave and heat equation with a globally Lipschitz nonlinearity.

  Chapter 5 is taken from [96]:

  > *Turnpike in Lipschitz-nonlinear optimal control.*
  > C. Esteve, B. Geshkovski, D. Pighin and E. Zuazua, 2020.
  > https://arxiv.org/abs/2011.11091

Part III. *Interplay of deep learning and control.*

- In Chapter 6, we study the behavior of supervised learning problems for neural ODEs when the final time horizon T is increased, a fact that may be interpreted as increasing the depth in the associated residual neural network (ResNet) setting.

  For the classical L^2–regularized empirical risk minimization problem, under homogeneity assumptions on the neural ODE dynamics, we prove that when T goes to infinity, the training error decays to zero with a (almost) polynomial rate. In the context of regression tasks, the optimal parameters are also shown converge to minimal L^2–norm parameters which interpolate the dataset. Moreover, motivated by the fact that the L^2–regularization context, a natural scaling between the time horizon T and the regularization hyper-parameter λ appears, using similar arguments, we obtain the same convergence results when λ goes to zero and the horizon is fixed. These results thus allow us to stipulate generalization properties in the overparametrized regime – now seen from the large depth and neural ODE perspective–, and are aligned with results on regularization path convergence (i.e., λ to zero) and implicit regularization of gradient descent for linear models or two-layer perceptrons.

  To enhance the polynomial decay rates of the training error, we propose an augmented learning problem by adding an artificial regularization term of the state trajectory over the entire time horizon. We apply the turnpike and stabilization results of Chapter 5 to obtain an exponential rate of decay for the training error and for the optimal parameters in any time – an improved estimate for the depth required to reach almost perfect training accuracy.

  The aforementioned asymptotic regimes are also discussed in the context of continuous space-time neural networks taking the form of nonlinear integro-differential equations, which provide a framework for addressing ResNets with variable widths.

  Chapter 6 is taken from [95]:

  > *Large-time asymptotics in deep supervised learning.*
  > C. Esteve, B. Geshkovski, D. Pighin and E. Zuazua, 2020.
  > https://arxiv.org/abs/2008.02491

- Finally, in Chapter 7, following the supervised learning framework of Chapter 6, we focus on a cost consisting of an integral of the empirical risk over the time

horizon and L^1–parameter regularization, and under homogeneity assumptions on the dynamics (typical for ReLU activations), we prove that any global minimizer is sparse, in the sense that there exists a positive stopping time beyond which the optimal parameters vanish. Moreover, under appropriate interpolation assumptions of the model, we may provide quantitative estimates on the stopping time, and on the training error of the neural ODE trajectories at the stopping time. The latter stipulates a quantitative approximation property of neural ODE flows with sparse parameters. In practical terms, when extrapolated to the ResNet context, a shorter time-horizon in the optimal control problem can be interpreted as considering a shallower ResNet, which lowers the computational cost of training.

Chapter 7 is taken from [272]:

> *Sparse approximation in learning via neural ODEs.*
> C. Esteve Yagüe and B. Geshkovski, 2021.
> https://arxiv.org/abs/2102.13566

Guide. We proceed by providing a brief introduction of the main paradigms in play in each part (I to III), followed by a detailed summary of each individual chapter and select results. Notation is local to each chapter.

1.2 Part I: Controllability of free boundary problems

We shall focus on time-evolving and one-phase free boundary problems, namely, free boundary problems in which both the PDE and the free boundary evolve over time, and wherein the free boundary is (part of) the bounding hyper-surface of the phase governed by the PDE (with the complementary phase usually representing the vacuum). Time-evolving free boundary problems are sometimes referred to as *moving interface problems* in engineering contexts.

The origin of modern free boundary problems may perhaps be traced back to the famous Stefan problem, a model of phase transitions in liquid-solid systems, first considered in 1831 by Lamé and Clapeyron [66] in relation to the problems of ice formation in the polar seas. Its general physical setup consists in considering a domain Ω which is occupied by water, a part of whose boundary is some interface Γ describing contact with a deformable solid such as ice. Due to melting or freezing of the water, the regions occupied by the water and ice will change over time and, consequently, the interface Γ will also change its position and shape, which indeed leads to the appearance of a free boundary. The problem is named after J. Stefan [250], who formulated the problem circa 1890, and validated the model by virtue of experimental data. The Stefan problem has since found a variety of alternative applications, including in population dynamics ([90]) as a generalization of the Fisher-KPP equation, in probability ([121]) as a hydrodynamic limit of particle densities and random walks, and in computer graphics ([156]) for simulating ice dynamics.

The Stefan problem is the prototypical time-evolving free boundary problem. In the simplest, one-dimensional case, it may be written as

$$\begin{cases} v_t - v_{zz} = 0 & \text{in } (0,T)\times(0,\ell(t)) \\ v(t,0) = u(t), \quad v(t,\ell(t)) = 0 & \text{in } (0,T) \\ \ell'(t) = -v_z(t,\ell(t)) & \text{in } (0,T) \\ v(0,z) = v_0(z), \quad \ell(0) = \ell_0 & \text{in } (0,\ell_0). \end{cases} \tag{1.2.1}$$

Here $u = u(t)$ is the control, actuating at the fixed end $z = 0$, and $\ell(t)$ represents the moving free boundary. As we shall focus on free boundary problems such as (1.2.1), where namely the PDE is parabolic (heat-like), and since our main interest will be to

set-up a methodology for proving the controllability of both components of the system to some (e.g. zero, but possibly non-trivial) trajectory in time $T > 0$, we briefly illustrate the main ideas used in the proof of controllability for the pure heat equation.

Let $\Omega \subset \mathbb{R}^d$ be a bounded and regular domain, let $\omega \subset \Omega$ be any open and non-empty subset, let $T > 0$, and consider the controlled heat equation

$$\begin{cases} y_t - \Delta y = u\mathbf{1}_\omega & \text{in } (0,T) \times \Omega \\ y = 0 & \text{on } (0,T) \times \partial\Omega \\ y_{|t=0} = y_0 & \text{in } \Omega, \end{cases} \tag{1.2.2}$$

where $u = u(t,x)$ is the control, actuating within the subset ω, and $y_0 \in L^2(\Omega)$ is a given initial datum. The problem of null-controllability (which is equivalent to controllability to any trajectory, due to the linearity of the system), consists in – given any initial datum $y_0 \in L^2(\Omega)$ – finding a control $u \in L^2((0,T) \times \omega)$ such that the unique solution $y \in C^0([0,T]; L^2(\Omega)) \cap L^2(0,T; H^1(\Omega))$ to (1.2.2) satisfies $y(T,\cdot) = 0$ a.e in Ω.

The main trick in proving the null-controllability of[1] (1.2.2) lies in noting that the identity

$$\int_\Omega y(T)\varphi_T \,\mathrm{d}x = \int_0^T \int_\omega u\varphi \,\mathrm{d}x \,\mathrm{d}t + \int_\Omega y_0\varphi(0) \,\mathrm{d}x \tag{1.2.3}$$

holds for all $\varphi_T \in L^2(\Omega)$, where φ is the solution of the adjoint heat equation

$$\begin{cases} -\varphi_t - \Delta\varphi = 0 & \text{in } (0,T) \times \Omega \\ \varphi = 0 & \text{on } (0,T) \times \partial\Omega \\ \varphi_{|t=T} = \varphi_T & \text{in } \Omega. \end{cases} \tag{1.2.4}$$

One then sees that to have $y(T,\cdot) = 0$ a.e. in Ω, the right hand side in (1.2.3) needs to be zero for all φ_T. But the latter can be seen as the Euler-Lagrange equation for the minimizer $\widehat{\varphi}_T$ of the functional

$$J_{\text{obs}}(\varphi_T) := \frac{1}{2}\int_0^T \int_\omega \varphi^2 \,\mathrm{d}x \,\mathrm{d}t - \int_\Omega y_0\varphi(0) \,\mathrm{d}x,$$

with $u = -\widehat{\varphi}|_\omega$. Constructing the control u following the minimization of a convex functional over the solutions of the adjoint problem, as described in what precedes, is the goal of the *Hilbert Uniqueness Method* (HUM), introduced by J-L. Lions in [185, 184]. For the above heuristic for building the null-controls to be rigorous, one needs to ensure that a minimizer of J_{obs} does indeed exist. This can be addressed by means of the direct method in the calculus of variations, with the main issue lying in ensuring the continuity[2] and coercivity of the functional J_{obs} – this can in turn be guaranteed provided the *observability inequality*

$$\int_0^T \int_\omega |\varphi(t,x)|^2 \,\mathrm{d}x \,\mathrm{d}t \geqslant \mathfrak{C}(T,\omega) \int_\Omega |\varphi(0,x)|^2 \,\mathrm{d}x \tag{1.2.5}$$

holds for some $\mathfrak{C}(T,\omega) > 0$ and for all $\varphi_T \in L^2(\Omega)$, where φ solves (1.2.4). Reviewing the above discussion, one thus readily sees that, by virtue of the HUM, the problem of null-controllability for the heat equation is equivalent to proving the observability inequality for the solution of the adjoint heat equation.

There have been a number of works regarding methods for proving (1.2.5), including the use of elliptic Carleman inequalities for the Laplacian to obtain observability on spaces

[1] This idea is not specific to the heat equation and is easily seen to hold for general linear systems of the form $y' = Ay + Bu$, where A generates a strongly continuous semigroup in some Hilbert space and B is a bounded operator in an appropriate functional setup.

[2] In fact, by virtue of (1.2.5), the functional J_{obs} can be extended by continuity in a unique way on the completion of $L^2((0,T) \times (-1,1))$ by the norm $\|\varphi_T\|^2_{\text{obs}} := \int_0^T \int_\omega |\varphi|^2 \,\mathrm{d}x \,\mathrm{d}t$.

spanned by the first eigenfunctions, combined with the dissipation of the heat solution ([174, 175]), full Carleman inequalities for the heat operator ([113, 104]) and transmutation techniques from wave to heat ([208, 93]), for instance. In the one-dimensional case (or, for that matter, any geometrical setting where the spectrum of the linear operator may be computed explicitly, e.g., a rectangle), further arguments for proving (1.2.5) can be used, since the spectrum of the Laplacian is explicit. By virtue of the summability of the sequence of the inverses, and the uniform gap between consecutive eigenvalues, one may ensure an estimate of the form

$$\mathfrak{C}(T)\int_0^T \left|\sum_{k=1}^{+\infty} a_k\ e^{-\lambda_k t}\right|^2 \mathrm{d}t \geqslant \sum_{k=1}^{+\infty} |a_k|^2 e^{-2\lambda_k T} \tag{1.2.6}$$

for all $\{a_k\}_{k=1}^{\infty} \in \ell^2(\mathbb{N})$, by constructing a biorthogonal sequence to the real exponentials and making use of the Paley-Wiener theorem and estimates of entire functions ([97, 255]). Estimate (1.2.6) is sometimes referred to as *parabolic Ingham* in control theory folklore, due to its resemblance to the well-known Ingham inequality for complex exponentials ([144], itself used for the wave equation). One may apply (1.2.6) to the expression obtained after decomposing the solution in the orthonormal basis of eigenfunctions the Laplacian, and, provided a uniform lower bound of the $L^2(\omega)$–norm of the eigenfunctions, may conclude the proof of (1.2.5). We shall make use of this spectral technique for proving the observability inequality in a couple of instances throughout this thesis.

Further modern methods for proving the null-controllability of the heat equation include flatness [200], explicit characterization of the reachable space via complex analysis techniques [138], and backstepping [71].

Our objective in Part I will be to control *both components* of the state for several free boundary problems in the mould of (2.1.4), namely, also control the free boundary to some reference point in time T.

1.2.1 One-dimensional viscous free boundary flows (Chapter 2)

We begin by considering the following free boundary problem for the viscous Burgers equation:

$$\begin{cases} v_t - v_{zz} + vv_z = 0 & \text{in } (0,T)\times(0,\ell(t)) \\ v(t,0) = u(t), \quad v_z(t,\ell(t)) = 0 & \text{in } (0,T) \\ \ell'(t) = v(t,\ell(t)) & \text{in } (0,T) \\ v(0,z) = v_0(z), \quad \ell(0) = \ell_0 & \text{in } (0,\ell_0). \end{cases} \tag{1.2.7}$$

where the unknown is the pair (v,ℓ), with ℓ represents the free boundary; $\ell_0 > 0$, and $u = u(t)$ is a control actuating along the fixed boundary $z = 0$. Model (1.2.7) is presented and studied in [36, 37], where local-in-time existence and uniqueness of strong solutions are shown, supplemented by numerical studies. It may be seen as a one-dimensional simplification of the incompressible Navier-Stokes equations with a free surface set in $\mathbb{R}^d$ with $d = 2, 3$, as encountered in the works [19, 20], and [199], where particular emphasis is given on the application to *mould filling*. The state of (1.2.7) involves the velocity $v(t,z)$ of the one-dimensional fluid and the free boundary $\ell(t)$, whose counterpart in dimension $d \geqslant 2$ would represent the position of the free surface of the fluid. The fluid velocity is governed by the viscous Burgers equation, while the dynamics of the free boundary follow the fluid velocity, as per the equation $\ell'(t) = v(t,\ell(t))$.

We observe that for any $\ell_* > 0$, the pair $(\overline{v},\overline{\ell})$ with

$$\overline{v} \in \mathbb{R}, \quad \overline{\ell}(t) = \ell_* + \overline{v}t > 0 \quad \text{in } [0,T], \tag{1.2.8}$$

is an explicit, non-trivial solution to System (1.2.7) with $u \equiv \overline{v}$. The main goal of Chapter 2 is to prove the local exact-controllability for (1.2.7) to this particular trajectory. This is reflected in our main result.

Theorem 1.1 ([116]). *Let $T > 0$, $\ell_* > 0$ and $\overline{v} \in \mathbb{R}$ be such that $\overline{\ell}(t) = \ell_* + \overline{v}t > 0$ for all $t \in [0,T]$. There exists $r > 0$ such that for all $\ell_0 > 0$ and $v_0 \in H^1(0,\ell_0)$ satisfying*

$$\|v_0 - \overline{v}\|_{H^1(0,\ell_0)} + |\ell_0 - \ell_*| \leqslant r,$$

there exists a control $u \in H^{3/4}(0,T)$ such that the unique solution

$$\ell \in C^1([0,T]) \quad v \in L^2\big(0,T;H^2(0,\ell(\cdot))\big) \cap C^0\big([0,T];H^1(0,\ell(\cdot))\big)$$

of (1.2.7) satisfies

$$\inf_{t\in[0,T]} \ell(t) > 0 \quad \text{and} \quad \ell(T) = \overline{\ell}(T) \quad \text{and} \quad v(T,\cdot) = \overline{v} \quad \text{in } (0,\ell(T)).$$

Discussion. Let us provide some context regarding Theorem 1.1.

- **Related work on the Stefan problem.** The controllability of the one-dimensional Stefan problem (1.2.1) has been partially addressed, [99, 103], where a null controllability result where *only the first component* v is controlled, i.e., $v(T,\cdot) = 0$ in $(0,\ell(T))$, for small initial data v_0, is shown. Such results are partial, as they cannot ensure that the entire system would remain in the prescribed configuration beyond the final time horizon. The novelty of Theorem 1.1 with respect to these works is that it yields the controllability of both components of the system, to a non-trivial trajectory.

- **Links with fluid-structure interaction.** Free boundary problems which arise in fluid-structure interaction, such as the simplified piston problem

$$\begin{cases} v_t - v_{zz} + vv_z = 0 & \text{in } (0,T)\times(-1,\ell(t))\cup(\ell(t),1) \\ v(t,-1) = u_1(t), \quad v(t,1) = u_2(t) & \text{in } (0,T) \\ v(t,\ell(t)) = \ell'(t) & \text{in } (0,T) \\ m\ell''(t) = [v_z](t,\ell(t)) & \text{in } (0,T) \\ v(0,z) = v_0(z), \quad \ell(0) = \ell_0, \quad \ell'(0) = \ell_1 & \text{in } (-1,\ell_0)\cup(\ell_0,1), \end{cases} \tag{1.2.9}$$

  introduced by Vázquez & Zuazua [264, 265], have also been addressed in the literature. The null-controllability of (1.2.9) refers to controlling the fluid velocity $v(T,\cdot) = 0$, the particle velocity $\ell'(T) = 0$, and the particle position $\ell(T) = 0$. In [100], controls u_1, u_2 are used on both boundaries in view of applying a Carleman based strategy. Such an approach is not feasible when there is a control at only one end (i.e., $u_2 = 0$) because of the lack of connectivity of the fluid domain. This issue is solved in [191], where the authors introduce a methodology for tackling the null-controllability of parabolic systems in spite of source terms, without requiring Carleman inequalities (they thus use spectral techniques).

  There is a notable difference between problems of the type (1.2.9) and (1.2.7). Indeed, the former system has a stronger coupling than the latter systems due to the presence of two equations for the free boundary ℓ. This can be seen when linearizing both systems around their trivial trajectory (after fixing the domain). In the linearization of (1.2.7),

$$\begin{cases} y_t - y_{xx} = 0 & \text{in } (0,T)\times(0,1) \\ y(t,0) = u(t), \quad y_x(t,1) = 0 & \text{in } (0,T) \\ \ell'(t) = y(t,1) & \text{in } (0,T) \\ y(0,x) = y_0(x), \quad \ell(0) = \ell_0 & \text{in } (0,1), \end{cases}$$

the PDE and ODE components are decoupled, as the linear PDE may be solved without any knowledge of the ODE component. On the other hand, the linearization of (1.2.9) around the trivial solution (see [191])

$$\begin{cases} y_t - y_{xx} = 0 & \text{in } (0,T)\times(-1,0)\cup(0,1) \\ y(t,-1) = u(t), \quad y(t,1) = 0 & \text{in } (0,T) \\ y(t,0) = \ell'(t) & \text{in } (0,T) \\ m\ell''(t) = [y_x](t,0) & \text{in } (0,T) \\ y(0,x) = y_0(x), \quad \ell(0) = \ell_0, \quad \ell'(0) = \ell_1 & \text{in } (-1,\ell_0)\cup(\ell_0,1), \end{cases}$$

preserves the coupling of the PDE component and the ODE component because of the presence of two equations for the latter. In other words, one may write the linearized fluid structure system in a canonical systems form $\dot{z} = Az + Bu$, where $z = (y,h)$, and then add the integrator $\ell' = h$ as a bounded and compact perturbation. The same cannot be done for the linearization of (1.2.7).

- **Arbitrary trajectories.** The controllability to arbitrary trajectories is not a straightforward extension, as observed on the level of the linearized system which contains several non-local trace terms (see (2.5.1)). Consequently, in terms of the adjoint problem one obtains non-standard boundary conditions (see (2.5.3)) for which, to the best of our knowledge, observability inequalities are lacking. This is discussed in more detail in Section 2.5.1.

- **Global results.** Theorem 2.1 is a local result, as while the PDE component may possess a dissipative mechanism, the asymptotic position of the free boundary is not known for this system. This is in part due to the lack of conservation properties satisfied by the position of the free boundary ℓ, making its asymptotic position more difficult to determine when compared to similar problems with a stronger coupling and set on the whole line [264, 166]. By means of maximum principle arguments, it could be possible to show that the free boundary increases as time grows, which could in turn stipulate an asymptotic behavior of the velocity v to a self-similar profile of the form $\frac{1}{\sqrt{t}} f\left(\frac{x}{\sqrt{t}}\right)$, well known in the context of the viscous Burgers equation set on $\mathbb{R}$ (see e.g. [284]). We leave this issue open.

Sketch of the proof. Our proof combines several elements of the control of parabolic equations in a systematic and ordered way.

Step 1). Fixing the domain. To take advantage of a simplified functional setting, it is more advantageous to reformulate (1.2.7) in a domain which is time-independent. To this end, let us define the pull-back velocity function $w : (0,1) \longrightarrow \mathbb{R}$ by

$$w(t,x) = v(t,z), \quad x = \frac{z}{\ell(t)} \qquad \text{for } x \in (0,1).$$

A simple application of the chain rule gives the following system of equations for w:

$$\begin{cases} w_t - \dfrac{1}{\ell^2} w_{xx} - \dfrac{\ell'}{\ell} x w_x + \dfrac{1}{\ell} w w_x = 0 & \text{in } (0,T)\times(0,1) \\ w(t,0) = u(t), \quad w_x(t,1) = 0 & \text{in } (0,T) \\ \ell'(t) = w(t,1) & \text{in } (0,T) \\ w(0,x) = w_0(x), \quad \ell(0) = \ell_0 & \text{in } (0,1), \end{cases} \tag{1.2.10}$$

where $w_0(x) = v_0(\ell_0 x)$. As (1.2.7) and (1.2.10) are equivalent provided $\ell(t) > 0$ in $[0,T]$, we will henceforth concentrate on the latter system. Taking the previous transformations into account, writing $w = \overline{v} + y$ and $\ell = \overline{\ell} + h$, extending the physical domain $(0,1)$ to the

fictitious domain $(-1,1)$, and using a control actuating within an open and non-empty interval $\omega \subsetneq (-1,0)$, we first consider the distributed control problem

$$\begin{cases} y_t - a y_{xx} + b y_x = \mathcal{N}(y,h) + u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ y(t,-1) = y_x(t,1) = 0 & \text{in } (0,T) \\ h'(t) = y(t,1) & \text{in } (0,T) \\ y(0,x) = y_0(x), \quad h(0) = h_0 & \text{in } (-1,1). \end{cases} \tag{1.2.11}$$

Then, Theorem 1.1 would be a consequence of the null-controllability of both components of (1.2.11). Here all of the intervening coefficients are smooth, with $b(t,1) = 0$. The initial datum $y_0 \in H^1(0,1)$ is also extended to a datum $\widetilde{y}_0$ with $\|\widetilde{y}_0\|_{H^1(-1,1)} \leqslant \|y_0\|_{H^1(0,1)}$. By abuse of notation, we continue denoting the extended initial datum by y_0. Once the null-controllability problem for (1.2.11) is solved, $u(t) := y(t,0) + \overline{v}$ would provide the desired control for Problem (1.2.10), which in view of the previous discussion, also provides a solution to (1.2.7).

Step 2). Control of a linearized system. To prove the null-controllability for (1.2.11), we first consider the linear system

$$\begin{cases} y_t - a y_{xx} + b y_x = f + u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ y(t,-1) = y_x(t,1) = 0 & \text{in } (0,T) \\ h'(t) = y(t,1) & \text{in } (0,T) \\ y(0,x) = y_0(x), \quad h(0) = h_0 & \text{in } (-1,1), \end{cases} \tag{1.2.12}$$

where f is a given source term. We seek a trajectory (y,h) of the linearized problem (1.2.12) satisfying

$$y(T,\cdot) = 0 \quad \text{in } (-1,1) \quad \text{and} \quad h(T) = 0.$$

In (1.2.12) we are dealing with a cascade-like system, as knowing y immediately yields h. In other words, the null-controllability of (1.2.12), would follow from solving the linear control problem (recall that $a(t) > 0$ and $b(t,1) = 0$)

$$\begin{cases} y_t - a y_{xx} + b y_x = f + u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ y(t,-1) = y_x(t,1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1) \\ y(T,x) = 0 & \text{in } (-1,1) \end{cases} \tag{1.2.13}$$

subject to the linear finite-dimensional constraint

$$h_0 + \int_0^T y(\tau,1)\,\mathrm{d}\tau = 0. \tag{1.2.14}$$

We will see this as a constrained controllability problem, namely with a linear finite-dimensional constraint on the control u.

It is well-known that a Carleman inequality along with the HUM method yield the null-controllability of the linear heat equation (1.2.13) with a source term f in an exponentially weighted L^2–space (as in (2.3.4)). To control the second component h to zero at time T, we will reformulate the constraint (1.2.14) by introducing a heat equation with a non-homogeneous boundary condition at $x = 1$. The requirement $h(T) = 0$ may then be achieved by adding a corrector term to the HUM control for the heat equation.

Let us consider

$$\begin{cases} -\psi_t - a\psi_{xx} - (b\psi)_x = 0 & \text{in } (0,T)\times(-1,1) \\ \psi(t,-1) = 0, \quad \psi_x(t,1) = 1 & \text{in } (0,T) \\ \psi(T,x) = 0 & \text{in } (-1,1). \end{cases} \tag{1.2.15}$$

Multiplying the heat equation appearing in (1.2.12) by the unique weak solution $\psi \in L^2(0,T;H^1(-1,1)) \cap C^0([0,T];L^2(-1,1))$ of (1.2.15) and integrating, we see that due to (1.2.14), a control u is such that the corresponding solution of (1.2.12) satisfies $h(T)=0$ if and only if

$$\int_0^T \int_\omega u\psi \,\mathrm{d}x\,\mathrm{d}t = \underbrace{-\int_{-1}^1 y_0(x)\psi(0,x)\,\mathrm{d}x + h_0 - \int_0^T \int_{-1}^1 f\psi \,\mathrm{d}x\,\mathrm{d}t}_{:=M_0\in\mathbb{R}}. \tag{1.2.16}$$

Let us define the projector

$$\mathbb{P}(\zeta) := \left(\int_{(0,T)\times\omega} |\psi|^2 \,\mathrm{d}x\,\mathrm{d}t\right)^{-1} \int_{(0,T)\times\omega} \psi\zeta \,\mathrm{d}x\,\mathrm{d}t \qquad \text{for all } \zeta \in L^2(0,T;L^2(-1,1)).$$

The key property of the operator $\mathbb{P}(\cdot)$ is its finite-dimensional range (in fact, one-dimensional range). We adapt the HUM method to account for the constraint (1.2.16), by considering the convex functional

$$\begin{aligned} J_{\mathrm{obs}}(\zeta_T, g) :=& \frac{1}{2}\int_0^T \int_\omega |\zeta - \mathbb{P}(\zeta)\psi|^2 \,\mathrm{d}x\,\mathrm{d}t + \frac{1}{2}\int_0^T \int_{-1}^1 e^{-2s\alpha}|g|^2 \,\mathrm{d}x\,\mathrm{d}t \\ &- \int_0^T \int_{-1}^1 f\zeta \,\mathrm{d}x\,\mathrm{d}t - \int_{-1}^1 y_0(x)\zeta(0,x)\,\mathrm{d}x - \mathbb{P}(\zeta)M_0, \end{aligned}$$

for some $s>0$ large enough, initially defined for $(\zeta_T, g) \in L^2(-1,1)\times L^2(0,T;L^2(-1,1))$ with corresponding solution $\zeta \in L^2(0,T;H^1(-1,1)) \cap C^0([0,T];L^2(-1,1))$ to the adjoint heat equation

$$\begin{cases} -\zeta_t - a\zeta_{xx} - (b\zeta)_x = g & \text{in } (0,T)\times(-1,1) \\ \zeta(t,-1) = \zeta_x(t,1) = 0 & \text{in } (0,T) \\ \zeta(T,x) = \zeta_T(x) & \text{in } (-1,1) \end{cases} \tag{1.2.17}$$

and ψ being the solution to (1.2.15). Should a minimizer to J_{obs} exist, it can then be used to build the desired control – state pair for (1.2.12). In fact, by writing the Euler-Lagrange equation at the minimizer $(\widehat{\zeta_T}, \widehat{g})$, one can find that the control takes the form

$$u := \left[\left(-\widehat{\zeta} - \mathbb{P}(\widehat{\zeta})\psi\right) + M_0\left(\int_0^T \int_\omega \psi^2 \,\mathrm{d}x\,\mathrm{d}t\right)^{-1} \psi\right]\Big|_\omega.$$

To guarantee the existence of a minimizer, we use the direct method in the calculus of variations, which requires the coercivity of J_{obs}. This in turn is guaranteed by an improved observability inequality of the form

$$\begin{aligned} &\int_0^T \int_{-1}^1 \theta^3 e^{-2s\alpha}|\zeta|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_{-1}^1 |\zeta(0,x)|^2 \,\mathrm{d}x + |\mathbb{P}(\zeta)|^2 \\ &\qquad \lesssim_{T,\omega} \left(\int_0^T \int_{-1}^1 e^{-2s\alpha}|g|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_0^T \int_\omega |\zeta - \mathbb{P}(\zeta)\psi|^2 \,\mathrm{d}x\,\mathrm{d}t\right), \end{aligned}$$

the proof of which follows by combining the classical Carleman inequality for the heat operator with a compactness-uniqueness argument.

Step 3). Fixed-point argument. After some (albeit nontrivial) technical estimates of the nonlinear terms in the Carleman weighted spaces, we may conclude by applying a Banach fixed-point argument to the source term f appearing in the linear system, provided taking the initial data in some small-enough ball.

1.2.2 Perturbed porous-medium gas flow (Chapter 3)

The *porous medium equation*

$$\partial_t h - \partial_z^2(h^m) = 0 \tag{1.2.18}$$

where $m > 1$, is a prototypical model for the density distribution of a gas flowing in a porous medium ($h(t,z)$ representing the density), or the evolution of a thin liquid film deposited onto a solid substrate ($h(t,z)$ representing the height of the film). By developing the diffusion term, it is readily seen that equation (1.2.18) degenerates when h approaches zero. Thus, any solution with compactly supported initial datum retains the compact support in any finite time. In physical terms, the diffusing gas does not reach any point in space instantaneously, but rather propagates with finite speed. This property results in the fact that the porous medium equation is a free boundary problem, the free boundary being given by $\partial\{h > 0\}$.

In view of the known asymptotic behavior for large times (see [268, Chapter 18]) and the desired positivity of the state, a natural question which arises is whether one may control the state $h(t,z)$, as well as its interface, to the self-similar Barenblatt trajectory

$$h_B(t,z) = (t+1)^{-\frac{1}{m+1}}\left(1 - \frac{m-1}{2m(m+1)}\frac{z^2}{(t+1)^{\frac{2}{m+1}}}\right)^{\frac{1}{m-1}} \qquad \text{in } \{h_B > 0\}$$

in a given finite time $T > 0$ by means of an additional forcing control term. To the best of our knowledge, this kind of exact-controllability to trajectories question has not been addressed in the existing literature on the porous medium equation.

An important difficulty when tackling this question is the moving time-dependent support of the solution h and the target Barenblatt trajectory h_B. As the two are defined in different domains, perturbations of the form $h_B + y$ around Barenblatt are difficult to define in view of linearizing, a key step in proving controllability. It is more convenient to look at the equation satisfied by the *pressure* $v = \frac{m}{m-1}h^{m-1}$ in self-similar coordinates, namely

$$\begin{cases} \partial_t v - v\partial_z^2 v - (\sigma+1)((\partial_z v)^2 + x\partial_z v) - v = 0 & \text{in } \{v > 0\} \\ v(0,z) = v_0(z) & \text{in } \{v_0 > 0\}, \end{cases} \tag{1.2.19}$$

(see [244, Section 1.2]) where $\sigma = -\frac{m-2}{m-1} > -1$. Indeed, in these coordinates, the Barenblatt solution $\rho := v_B$ is stationary and supported in the unit interval:

$$\rho(z) = \frac{1}{2}\left(1 - z^2\right) \qquad \text{for } z \in (-1,1). \tag{1.2.20}$$

The motivation behind our work is to know whether one can steer the state $v(t,z)$ and its interface to the stationary Barenblatt solution $\rho(z)$ in a given time $T > 0$, by means of an additional forcing control term in the equation.

To overcome the difficulty of the moving domain, a Lagrangian-like change of variables (thus depending on the solution, and called *von Mises transformation*, see Section 3.7.2) introduced by Koch [160] may be applied, mapping the moving support of the solution onto the support of the Barenblatt profile, now the interval $(-1,1)$. The change of coordinates is a diffeomorphism provided the solution is small enough in Lipschitz norm, and in these new variables the Barenblatt reduces to the constant 1.

After the von Mises transform and after considering perturbations around the transformed Barenblatt, we are brought to consider the control problem for the transformed perturbation equation (see [244, Section 3]):

$$\begin{cases} \partial_t y - \rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x y) = \mathcal{N}(y) + u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ (\rho^{\sigma+1}\partial_x y)(t,\pm 1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1), \end{cases} \tag{1.2.21}$$

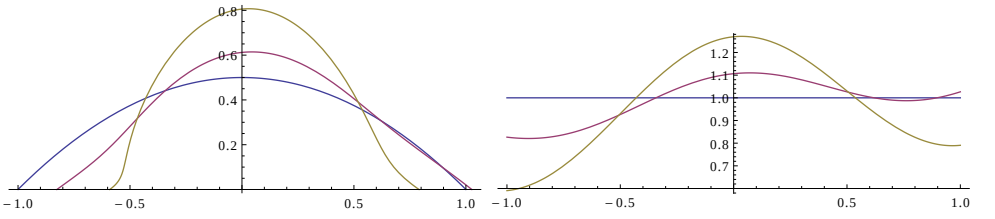


Figure 1.1: (*Left*) The graph of the Barenblatt pressure (blue curve) and two perturbations in the (z, v)–coordinates. (*Right*) The corresponding graphs after the von Mises transformation (Section 3.7.2) in the $(x, 1+y)$–coordinates.

where $T > 0$ and $\sigma > -1$, and the nonlinearity $\mathcal{N}(y) = \mathcal{N}(y, \partial_x y)$ is of the form

$$\mathcal{N}(y) = \rho F(y, \partial_x y) - \rho^{-\sigma}\partial_x(\rho^{\sigma+1} x F(y, \partial_x y)), \quad F(p,q) = \frac{q^2}{1+p+xq}, \quad p, q \in \mathbb{R}. \tag{1.2.22}$$

The distributed control u appearing in (1.2.21) actuates within $\omega = (a, b) \subsetneq (-1, 1)$. The solution $y(t, x)$ is a perturbation around the Barenblatt in the new variables (see Remark 3.7.2). Consequently, the null-controllability of (1.2.21) would correspond to the exact-controllability of the pressure $v(t, z)$ and its free boundary of a controlled version of (1.2.19) to the original Barenblatt $\rho(z)$, after reverting the von Mises transformation. While the nonlinearity in (1.2.21) is essentially quadratic in a neighborhood of the origin, the denominator may be singular and applying a fixed-point argument using only a (weighted) Sobolev space theory is not straightforward. We concentrate on a truncated version – we multiply the nonlinear terms by a smooth cut-off which vanishes at points where y and/or $\partial_x y$ are large; the truncated equation would thus be linear at such points. Let $\chi : [0, \infty) \longrightarrow [0, 1]$ be a smooth cut-off function, supported on $[0, 4)$ with $\chi(x) \equiv 1$ on $[0, 1]$. Let $0 < \varepsilon, \delta < 1$ satisfying $4(\varepsilon + \delta) < 1$ be fixed. For $p, q \in \mathbb{R}$, and recalling the definition of F in (1.2.22), we define

$$F_{\varepsilon,\delta}(p, q) = \chi\left(\frac{p^2}{\delta^2}\right)\chi\left(\frac{q^2}{\varepsilon^2}\right) F(p, q). \tag{1.2.23}$$

We will henceforth only be interested in (1.2.21) wherein $\mathcal{N}$ is replaced by $\rho F_{\varepsilon,\delta}$, namely

$$\begin{cases} \partial_t y - \rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x y) = \rho F_{\varepsilon,\delta}(y, \partial_x y) + u\mathbf{1}_\omega & \text{in } (0, T) \times (-1, 1) \\ (\rho^{\sigma+1}\partial_x y)(t, \pm 1) = 0 & \text{in } (0, T) \\ y(0, x) = y_0(x) & \text{in } (-1, 1). \end{cases} \tag{1.2.24}$$

Our main result is the following.

Theorem 1.2 ([115]). *Let $T > 0$, let $\omega \subsetneq (-1, 1)$ be an open, non-empty interval, and let $\sigma \in (-1, 0)$. Then there exists $r > 0$ such that for every $y_0 \in \mathcal{H}^1$ satisfying $\|y_0\|_{\mathcal{H}^1} \leqslant r$, there exists a control $u \in L^2(0, T; L^2(\omega))$ for which the unique solution $y \in L^2(0, T; \mathcal{H}^2) \cap C^0([0, T]; \mathcal{H}^1)$ of (1.2.24) satisfies $y(0, \cdot) = y_0$ and $y(T, \cdot) = 0$.*

Discussion. Let us put Theorem 1.2 into context with existing literature on the porous medium equation and degenerate parabolic equations.

- **Weighted Sobolev spaces.** The natural functional space for addressing (1.2.24) is that of the weighted Spaces

$$\mathcal{H}^k := \left\{ f \in L^1_{\text{loc}}(-1, 1) \colon \|f\|^2_{\mathcal{H}^k} := \sum_{j=0}^{k} \int_{-1}^{1} \rho^{\sigma+j} |\partial_x f|^2 \, \mathrm{d}x < \infty \right\},$$

for $k \geqslant 1$. Note that when σ is positive, $L^2(-1,1) \subset \mathcal{H}^0$ – thus, several results, making use of *new adapted Hardy inequalities* (see Section 3.2) are needed to make sense of the boundary traces in (3.1.7), among other things.

- **The free boundary problem.** Theorem 1.2 is a priori not sufficient to deduce a local controllability result (for both the state and the interface) to the stationary Barenblatt parabola ρ for the free boundary system (1.2.19) with a distributed control. However, if (1.2.24) is null-controllable with the nonlinearity $\mathcal{N}(y)$ as in (1.2.21), then one could deduce such a result.

  To achieve this, one only needs to remove the cut-off factor $\chi(p^2/\delta^2)\chi(q^2/\varepsilon^2)$ and add the high order nonlinear term. The cut-off is identically 1 whenever the solution y satisfies

  $$\|y\|_{C^{0,1}([0,T]\times[0,1])} \ll 1$$

  and this regularity is also what is needed to guarantee that the von Mises transformation (Section 3.7.2) is a diffeomorphism. However, Theorem 1.2 does not provide this regularity, and one should look to use more regular controls than just $L^2_{t,x}$ (this could perhaps be done by using a penalized HUM method) combined with maximal regularity results for the linear problem. See Remark 3.7.2 for more details.

- **Related work.** In [70], Coron et al. prove the null-controllability of the porous medium equation set on $(0,1)$ using Dirichlet boundary controls on both ends as well as a scalar forcing control. This differs from the original motivation behind our work, which was to control the pressure and its free boundary to the non-trivial Barenblatt profile (instead of the null-state).

  Our work may also be seen as a novel contribution to the controllability theory of linear degenerate parabolic equations. Indeed, while the differential operator in (1.2.26) may be rewritten as $-\rho\partial_x^2 y + (\sigma+1)x\partial_x y$, the weighted Neumann boundary conditions have not been considered in works on problems in non-divergence form. In particular, we do not consider the same weight and functional framework as in [47, 107], since $\frac{b}{a} = \frac{2(\sigma+1)x}{1-x^2} \notin L^1(-1,1)$ in our case. While we use spectral techniques, a Carleman inequality for our functional setting is lacking.

Elements of proof. Looking[3] at (1.2.24), it is natural to first study the null-controllability of the corresponding linear problem, where the nonlinear term is replaced by a source term:

$$\begin{cases} \partial_t y - \rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x y) = f + u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ (\rho^{\sigma+1}\partial_x y)(t,\pm 1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1). \end{cases} \tag{1.2.25}$$

The nonlinear term would be seen as a small perturbation, and be dealt with by means of a fixed-point argument. The latter argument will rely on the particular structure of the nonlinearity, which is now non-singular and essentially quadratic due to the cut-off factor.

To prove the null-controllability of Problem (1.2.25), we will make use of the so-called *source-term method*, first introduced by Liu, Takahashi & Tucsnak [191]. Roughly speaking, the strategy involves first showing the null-controllability of the homogeneous problem

$$\begin{cases} \partial_t y - \rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x y) = u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ (\rho^{\sigma+1}\partial_x y)(t,\pm 1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1), \end{cases} \tag{1.2.26}$$

[3] The requirement $\sigma \in (-1,0)$ only appears when estimating the nonlinear term in the weighted spaces (see Section 3.4). The null-controllability and well-posedness of the linearized problem (3.1.8) holds true for any $\sigma > -1$. We recall that σ is related to the nonlinearity exponent of the porous medium equation by $m = \frac{\sigma+2}{\sigma+1}$.

and the null-controllability of Problem (1.2.25) follows provided the source term f vanishes with appropriate decay as $t \nearrow T$. More specifically, the decay of the source term should be quick enough near the final time compared to the control cost in small time. The null-controllability of problem (1.2.26) is done by combining HUM induced duality with spectral techniques, making use of the explicit spectrum of the linear operator $\mathcal{A} = -\rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x)$ computed in the works of Seis [243, 244].

The linearized thin film equation. The *thin-film equation*

$$\partial_t h + \partial_z(h^n \partial_z^3 h) = 0 \quad \text{in } \{h > 0\}$$

where $n \in (0,3)$ represents a more accurate model for the evolution of a liquid film over a solid substrate in a regime known as lubrication approximation. Much like its second order counterpart, the PME (1.2.18), it is a free boundary problem whenever the initial datum is compactly supported (a physical phenomenon known as *droplets*). Our strategy for proving the null-controllability of the linearized problem (1.2.26) can also be applied to obtain a null-controllability result (see Section 3.5) for the thin-film equation linearized around its self-similar solution, which is a fourth-order degenerate parabolic equation.

For $n = 1$ (known as linear mobility regime), McCann & Seis [205, 245] replicate the ideas used for the PME in [78, 243, 244] to compute the spectrum of the full linearization of the thin-film equation around its own self-similar (Smyth-Hill) solution. Namely, after an analog rescaling and von Mises transformation, the control problem for the equation linearized around the self-similar solution is of the form

$$\begin{cases} \partial_t y + \mathcal{A}^2 y + \mathcal{A} y = u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ (\rho y)(t,\pm 1) = (\rho^2 \partial_x y)(t,\pm 1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1). \end{cases} \tag{1.2.27}$$

where $T > 0$ and $\mathcal{A} = -\rho^{-1}\partial_x(\rho^2\partial_x)$ is the operator governing the linearized porous medium equation (1.2.26) with $\sigma = 1$. As the eigenfunctions of $\mathcal{L} = \mathcal{A}(\mathcal{A} + \mathrm{Id})$ and $\mathcal{A}$ coincide, and the control operator B is the same as in Section 3.3, we may deduce the following null-controllability result for (1.2.27).

Theorem 1.3 ([115]). *Let $T > 0$, $\omega \subsetneq (-1,1)$ be an open, non-empty interval, and $\sigma = 1$. Then, for any $y_0 \in \mathcal{H}^0$, there exists a control $u \in L^2(0,T;L^2(\omega))$ such that the unique solution $y \in L^2(0,T;\mathcal{H}^2) \cap C^0([0,T];\mathcal{H}^0)$ of (1.2.27) satisfies $y(0,\cdot) = y_0$ and $y(T,\cdot) = 0$.*

1.2.3 The Stefan problem with surface tension (Chapter 4)

Let $\mathbb{T} := \mathbb{R}/(2\pi\mathbb{Z})$ denote the one-dimensional torus, which we identify with $[0, 2\pi]$, and set

$$\Omega := \mathbb{T} \times (0,1).$$

We also set $\Gamma_{\mathrm{bot}} := \mathbb{T}\times\{0\}$ and $\Gamma_{\mathrm{top}} := \mathbb{T}\times\{1\}$. We recall that in the Stefan problem, a heat-conducting liquid fills a time-varying domain $\Omega(t) \subset \mathbb{R}^2$ for $t \geqslant 0$ – we will assume that the boundary $\partial\Omega(t)$ of the liquid consists of two components, namely a time-varying component (the *free boundary* $\Gamma(t)$) and a fixed component. More specifically,

$$\Omega(t) := \big\{z = (z_1, z_2) \in \mathbb{T}\times\mathbb{R} : \quad 0 < z_2 < 1 + h(t, z_1)\big\},$$

where $h = h(t, z_1)$ is the unknown *height function*, and represents the displacement of the free boundary away from the reference boundary Γ_{top} (see fig. 1.10). The free boundary is consequently given by

$$\Gamma(t) := \big\{z = (z_1, z_2) \in \mathbb{T}\times\mathbb{R} : \quad z_2 = 1 + h(t, z_1)\big\}.$$

Given a time horizon $T > 0$, the one-phase Stefan problem with surface tension (or with Gibbs-Thomson correction) takes the form

$$\begin{cases} \partial_t \vartheta - \Delta \vartheta = 0 & \text{in } (0,T) \times \Omega(t) \\ \partial_t h = -\sqrt{1+|\partial_{z_1} h|^2}\, \nabla \vartheta_{|_{\Gamma(t)}} \cdot \mathbf{n} & \text{on } (0,T) \times \mathbb{T} \\ \vartheta = -\sigma \kappa(h) & \text{on } (0,T) \times \Gamma(t) \\ \vartheta = u & \text{on } (0,T) \times \Gamma_{\text{bot}} \\ (\vartheta, h)_{|_{t=0}} = (\vartheta^0, h^0) & \text{in } \Omega(0) \times \mathbb{T}, \end{cases} \tag{1.2.28}$$

where $\vartheta(t,z)$ is the unknown temperature, $h(t,z_1)$ is the unknown height function defining the free boundary, $u(t,z_1)$ denotes the control, while $\mathbf{n} = \mathbf{n}(t,z_1)$ given by

$$\mathbf{n}(t,z_1) = \frac{1}{\sqrt{1+|\partial_{z_1} h(t,z_1)|^2}} \begin{bmatrix} -\partial_{z_1} h(t,z_1) \\ 1 \end{bmatrix}, \tag{1.2.29}$$

denotes the unit normal to $\Gamma(t)$ outward $\Omega(t)$.

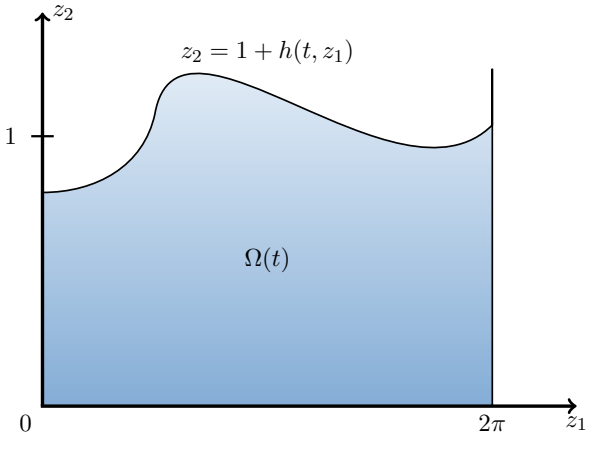


Figure 1.2: The moving domain $\Omega(t)$ representing the liquid, and the free boundary $\Gamma(t)$, delimiting the liquid-solid region, parametrized by the height function $h(t,z_1)$.

Note that the control u actuates along the whole fixed bottom boundary Γ_{bot}. On the other hand, the constant $\sigma > 0$ represents the surface tension coefficient, whereas $\kappa(h)$ denotes the mean curvature of the free boundary $\Gamma(t)$, and is defined by

$$\kappa(h) = \frac{\partial_{z_1}^2 h}{\left(1+|\partial_{z_1} h|^2\right)^{3/2}}.$$

The initial domain $\Omega(0)$ is given by

$$\Omega(0) = \left\{ z = (z_1, z_2) \in \mathbb{T} \times \mathbb{R} : \quad 0 < z_2 < 1 + h_0(z_1) \right\}.$$

We note that when $\sigma = 0$, (1.2.28) reduces to the classical Stefan problem, namely one has $\vartheta = 0$ instead of the Gibbs-Thomson condition along the interface $\Gamma(t)$. The main physical reason for introducing the Gibbs-Thomson correction stems from the need to account for supercooling effects, in which a fluid permits temperatures below its freezing point, or dendrite formation, in which simple shapes evolve into complex fingering patterns.

Since $(\vartheta^*, h^*) = (0,0)$ is an equilibrium configuration for (1.2.28), a natural question one may ask is whether – given a time horizon $T > 0$, a surface tension coefficient $\sigma > 0$, and initial data (ϑ^0, h^0), which, due to the nonlinear effects, ought to be small enough in an appropriate topology –, there exists a control $u = u(t,z_1)$ actuating along the fixed boundary Γ_{bot} and a corresponding solution (y,h) to (1.2.28) which satisfy

$$\vartheta(T,\cdot)=0 \quad \text{in } \Omega, \qquad \min_{[0,T]\times\mathbb{T}}\big(1+h(t,z_1)\big)>0, \qquad h(T,\cdot)=0 \quad \text{on } \mathbb{T}. \tag{1.2.30}$$

We anticipate such a result to hold true mainly due to the fact that the null-controllability (for both components) of the linearized system

$$\begin{cases} \partial_t y-\Delta y=0 & \text{in } (0,T)\times\Omega \\ \partial_t h(t,x_1)=\partial_{x_2} y(t,x_1,1) & \text{on } (0,T)\times\mathbb{T} \\ y(t,x_1,0)=u(t,x_1) & \text{on } (0,T)\times\mathbb{T} \\ y(t,x_1,1)=\sigma\partial_{x_1}^2 h(t,x_1) & \text{on } (0,T)\times\mathbb{T} \\ (y,h)_{|t=0}=\big(y^0,h^0\big) & \text{in } \Omega\times\mathbb{T} \end{cases} \tag{1.2.31}$$

does indeed hold (see Theorem 1.4) in any time $T>0$. We provide a full sketch of the proof methodology in what follows.

Discussion. Let us however begin by providing some context to the above property and its novelty with regard to the existing literature.

- **Novel contribution.** A controllability result as would be the first of its kind for free boundary problems governed by diffusive equations where the free boundary is a space-dependent function. In this sense, the problem we tackle differs from existing works on the controllability of fluid-rigid body models (e.g. [143, 31, 229, 30, 180]), as therein, the free boundary is a time-only dependent function, thus, after linearization, the controllability condition for the free boundary may be seen as a finite dimensional constraint on the control, similarly as Chapter 2.

  In addition, the spatial regularity of the height function h plays a crucial role in the analysis (or even existence) results. One needs to possibly consider very regular initial data (ϑ^0,h^0) in order to guarantee the smoothness of the domain.

  In fact, up to the best of our knowledge, even the controllability result regarding the linearized system (see Theorem 1.4 just below) is new in the literature.

- **Water waves.** Albeit for a system of different nature to ours, we refer to [5] for the local exact controllability result of the velocity and the free surface elevation of the water-waves equations in two dimensions, by means of a single control actuating along an open subset of the free surface. The two-dimensional geometrical strip-like setting is the same as ours. The authors use the Dirichlet to Neumann map to define the problem on a fixed domain. This procedure is closely tied to the equations under consideration, and is not applicable in our setting. After (quasi)linearization, a dispersive equation is obtained, which is shown to be controllable by means of Ingham-like techniques. Due to the lack of regularizing effect, the nonlinear problem is tackled by using a Nash-Moser iteration.

Methodology. To stimulate the plausibility of a result of the mould of (1.2.3), let us provide a brief sketch of the proof methodology, based on linearization techniques, with a statement and proof of the controllability of the linearized problem.

Step 1). Fixing the domain. We begin by fixing the domain, as it will allow us to work in a time-independent spatial setup. We emphasize that in the two-dimensional geometrical setup we consider here, the free boundary depends on the spatial variable z_1, hence the regularity of the domain $\Omega(t)$ depends on the spatial regularity of the height function $h(t,z_1)$. To avoid requiring high order Sobolev regularity on h, we shall fix the domain via a transformation which gains a $\frac{1}{2}$–order of regularity with respect to h. This is done by a harmonic extension of the boundary, namely by defining

$$\Psi(t,x):=(x_1,x_2+\psi(t,x)) \quad y(t,x)=\vartheta(t,\Psi(t,x)) \qquad \text{for } \ (t,x)\in[0,T]\times\Omega,$$

for all $t \geqslant 0$ given $h \in C^0([0,T];H^s(\mathbb{T}))$ for some $s \geqslant 0$, with $\psi(t,\cdot) \in H^{s+1/2}(\Omega)$ being the solution to

$$\begin{cases} \Delta\psi(t,\cdot) = 0 & \text{in } \Omega \\ \psi(t,x_1,0) = 0 & \text{on } \mathbb{T} \\ \psi(t,x_1,1) = h(t,x_1) & \text{on } \mathbb{T}. \end{cases}$$

This leads us to the system in the reference domain Ω:

$$\begin{cases} \partial_t y - \Delta y = \mathcal{N}_1(y,h) & \text{in } (0,T)\times\Omega, \\ \partial_t h = (\nabla y \cdot \mathbf{e}_2)_{\Gamma_{\text{top}}} + (\mathcal{N}_3(y,h)\cdot \mathbf{e}_2)_{\Gamma_{\text{top}}} & \text{on } (0,T)\times\Gamma_{\text{top}}, \\ y = \sigma\partial_{x_1}^2 h + \mathcal{N}_2(y,h) & \text{on } (0,T)\times\Gamma_{\text{top}}, \\ y = u & \text{on } (0,T)\times\Gamma_{\text{bot}}, \\ (y,h)\,|_{t=0} = \big(y^0,h^0\big) & \text{in } \Omega\times\mathbb{T}, \end{cases} \tag{1.2.32}$$

where the nonlinear terms $\{\mathcal{N}_i\}_{i=1}^3$ are all quadratic.

Step 2). The linearized system. As commonly done in literature on the controllability of nonlinear parabolic problems, we will first concentrate on proving the controllability of the system linearized around the target $(0,0)$, and then view the nonlinear terms in (1.2.32) as a small perturbation which may be dealt with by means of a fixed-point argument. Moreover, to avoid working with boundary control systems, we extend the physical reference domain Ω to the fictitious domain $\mathcal{O} := \mathbb{T}\times(-1,1)$ and consider a distributed control, actuating inside an open and nonempty subset $\omega := \mathbb{T}\times(a,b)$ with $(a,b)\subset(-1,0)$. In other words, the distributed control problem for the linearized Stefan problem with Gibbs-Thomson correction takes the form

$$\begin{cases} \partial_t y - \Delta y = u\mathbf{1}_\omega & \text{in } (0,T)\times\mathcal{O} \\ \partial_t h(t,x_1) = \partial_{x_2} y(t,x_1,1) & \text{on } (0,T)\times\mathbb{T} \\ y(t,x_1,0) = 0 & \text{on } (0,T)\times\mathbb{T} \\ y(t,x_1,1) = \sigma\partial_{x_1}^2 h(t,x_1) & \text{on } (0,T)\times\mathbb{T} \\ (y,h)_{|t=0} = \big(y^0,h^0\big) & \text{in } \mathcal{O}\times\mathbb{T}. \end{cases} \tag{1.2.33}$$

We prove the following result, which to the best of our knowledge, is also new in the literature on the control of parabolic systems.

Theorem 1.4 (Linear control)**.** *Let $T>0$ and $\sigma>0$. For any $(y^0,h^0)\in L^2(\mathcal{O})\times H^1(\mathbb{T})$, there exists a control $u\in L^2((0,T)\times\omega)$ such that the corresponding unique solution $y\in C^0([0,T];L^2(\mathcal{O}))$ and $h\in C^0([0,T];H^1(\mathbb{T}))$ of (1.2.33) satisfies*

$$y(T,\cdot)=0 \quad \text{in } \mathcal{O} \qquad \text{and} \qquad h(T,\cdot)=0 \quad \text{in } \mathbb{T}.$$

Moreover, there exists a positive constant $\mathfrak{C}(T,\sigma)=\mathfrak{C}(T,\sigma,\omega,\mathcal{O})>0$ such that

$$\|u\|_{L^2((0,T)\times\omega)} \leqslant \mathfrak{C}(T,\sigma)\,\big\|(y^0,h^0)\big\|_{L^2(\mathcal{O})\times H^1(\mathbb{T})}.$$

The proof[4] of Theorem 1.4 is a cornerstone of our work. Due to the difficulty in obtaining

[4] In our proof, we shall deduce that the derived constant $\mathfrak{C}(T,\sigma) \longrightarrow +\infty$ as $\sigma \searrow 0$ (this is not necessarily true the control cost, which is an even smaller constant). This is a curious observation – it may be difficult to derive the null-controllability of the classical Stefan problem ($\sigma=0$) – which is known to be the (macroscopic) limit case in the zero surface tension limit without control [130] –, from that of the Gibbs-Thomson system. In fact, the controllability of the classical Stefan problem does not appear obvious. The one-dimensional techniques of Chapter 2 do not directly apply, as the height function manifests itself as an infinite-dimensional propagator and thus cannot be covered by compactness. On another hand, by proceeding via Fourier techniques as for the Gibbs-Thomson system, we fall upon a non-self adjoint operator in the linearized one-dimensional case, which we are only able to tackle using the techniques of Chapter 2 for fixed $n\in\mathbb{Z}$, but fail to obtain uniform estimates with respect to n due to the usage of a compactness-uniqueness argument. These observations are not sufficient to conclude on the possible null-controllability (or lack thereof) of the classical Stefan problem in a strip-like geometry, which for the time being, remains open.

a clear formulation of the adjoint problem, a direct proof via HUM and an observability inequality does not appear straightforward. Instead, we exploit the periodicity of the control and the unknowns with respect to the $x_1 \in \mathbb{T}$ variable, write the unknowns in Fourier series, prove that each Fourier coefficient is null-controllable with a control cost uniform in the Fourier parameter, and then paste all the coefficients together to deduce Theorem 1.4. Such ideas have been used in the control literature, see Beauchard et al. [24, 21] for instance. To be more precise, for any $n \in \mathbb{Z}$, the system satisfied by each Fourier coefficient is

$$\begin{cases} \partial_t \widehat{y}_n - \partial_{x_2}^2 \widehat{y}_n + n^2 \widehat{y}_n = \widehat{u}_n \mathbf{1}_{(a,b)} & \text{in } (0,T)\times(-1,1) \\ \widehat{h}_n'(t) = \partial_{x_2}\widehat{y}_n(t,1) & \text{in } (0,T) \\ \widehat{y}_n(t,-1) = 0 & \text{in } (0,T) \\ \widehat{y}_n(t,1) = -\sigma n^2 \widehat{h}_n(t) & \text{in } (0,T) \\ \left(\widehat{y}_n, \widehat{h}_n\right)_{|t=0} = \left(\widehat{y}_n^0, \widehat{h}_n^0\right) & \text{in } (-1,1). \end{cases} \tag{1.2.34}$$

The null-controllability of (4.1.8) (Proposition 4.3.4) is then done in two parts, distinguishing in (4.1.8) the case $n \neq 0$, where the governing linear operator is self-adjoint in an appropriate product space and the observability inequality follows from an explicit computation of the spectrum, and the case $n = 0$, in which $\widehat{y}_n$ is independent of $\widehat{h}_n$, and the controllability of $\widehat{h}_n$ is seen as a finite-dimensional constraint on the linear heat control, and may be covered using improved observability inequalities done by compactness-uniqueness arguments as in [116].

Step 3). The nonlinear system. To tackle the nonlinear system, we look to apply a Banach fixed-point argument over the source-terms decoying the nonlinear terms in (1.2.32). To obtain the required null-controllability result for the problem with given source terms, we make use of an adaptation of the *source-term method* [191, 173, 115] in fractional Sobolev spaces[5] (see Theorem 4.3 for our proof). The Banach fixed-point argument is then performed inside small enough balls of these exponentially weighted energy spaces – it remains only to be shown that the quadratic nonlinear terms are indeed elements of these weighted energy spaces provided by the source-term method. We are led to propose the following conjecture.

Conjecture 1.2.1 (Nonlinear control)**.** *Let $T > 0$ and $\sigma > 0$. There exists $\varepsilon > 0$ such that for every (ϑ^0, h^0) satisfying*

$$h^0 \in H^{5/2}(\mathbb{T}), \qquad \min_{z_1 \in \mathbb{T}} \left(1 + h^0(z_1)\right) > 0, \qquad \vartheta^0 \in H^1(\Omega(0)),$$

the compatibility condition

$$\vartheta^0\left(z_1, 1 + h^0(z_1)\right) = -\sigma\kappa\left(h^0(z_1)\right) \qquad \text{for } z_1 \in \mathbb{T},$$

and

$$\left\|\vartheta^0\right\|_{H^1(\Omega(0))} + \left\|h^0\right\|_{H^{5/2}(\mathbb{T})} < \varepsilon,$$

there exists $u \in L^2(0,T;H^{3/2}(\mathbb{T})) \cap H^{3/4}(0,T;L^2(\mathbb{T}))$ such that the corresponding unique solution pair $\vartheta \in L^2(0,T;H^2(\Omega(\cdot))) \cap C^0([0,T];H^1(\Omega(\cdot)))$ and $h \in L^2(0,T;H^{7/2}(\mathbb{T})) \cap H^{3/4}(0,T;H^2(\mathbb{T})) \cap H^1(0,T;H^1(\mathbb{T})) \cap H^{5/4}(0,T;L^2(\mathbb{T}))$ to (1.2.28) satisfies (1.2.3).

1.3 Part II: Long-time optimal control

The *turnpike property* reflects the fact that, for suitable optimal control problems set in a sufficiently large time horizon, any optimal solution thereof remains, during most of the

[5] We define fractional-in-time Sobolev spaces via the standard Gagliardo seminorm, whereas the fractional Sobolev spaces on the torus $\mathbb{T}$ are defined via Fourier analysis.

time, close to the optimal solution of a corresponding "static" optimal control problem. This optimal static solution is referred to as *the turnpike* – the name stems from the idea that a turnpike is the fastest route between two points which are far apart, even if it is not the most direct route. In many cases, the turnpike property is described by an exponential estimate – for instance, the optimal trajectory $y_T(t)$ is $\mathcal{O}\left(e^{-\mu t} + e^{-\mu(T-t)}\right)$–close to the optimal static solution $\overline{y}$, for $t \in [0, T]$ and for some $\mu > 0$.

The notion that optimal strategies, when considered over long time periods, are constant for most of the time, traces back to work of von Neumann [267]. The terminology *turnpike* was introduced in the context of economics by Samuelson et al. [82] to interpret the full evolutionary phenomenon.

There has been an ever-increasing need, brought by applications in *deep learning* via *residual neural networks* (ResNets) (see [89, 95, 140]), of turnpike results for nonlinear optimal control problems without smallness conditions on the data or the running target, and for systems with globally Lipschitz-continuous but possibly nonsmooth nonlinearities. Such results are, to our knowledge, not known in the literature. As an intermezzo to Part III, let us motivate the above observation further ahead of presenting the mathematical setup and theory. In deep learning (see the subsequent sections for more details), one wishes to find a map which interpolates a dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ where $\vec{x}_i \in \mathbb{R}^{d_x}$ and $\vec{y}_i \in \mathbb{R}^{d_y}$ and gives accurate predictions on unknown points $\vec{x} \in \mathbb{R}^{d_x}$. Such a task may be accomplished, for instance, by minimizing

$$\int_0^T \sum_{i=1}^N \|P\mathbf{x}_i(t) - \vec{y}_i\|^2 \, \mathrm{d}t + \int_0^T \|u(t)\|^2 \, \mathrm{d}t, \tag{1.3.1}$$

where $u := [w, b]$ and $P : \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ is a non-zero surjective map, subject to

$$\begin{cases} \dot{\mathbf{x}}_i(t) = w(t)\sigma(\mathbf{x}_i(t)) + b(t) & \text{in } (0, T) \\ \mathbf{x}_i(0) = \vec{x}_i, \end{cases} \tag{1.3.2}$$

with $w \in L^2(0, T; \mathbb{R}^{d_x \times d_x})$ and $b \in L^2(0, T; \mathbb{R}^{d_x})$ designating the controls, whereas $\sigma \in \mathrm{Lip}(\mathbb{R})$ with $\sigma(0) = 0$ is a scalar nonlinear function, defined component-wise in (1.3.2). A typically used nonlinearity in practical applications is $ReLU$: $\sigma(x) = \max\{x, 0\}$ (see Figure 1.4). Variants of (1.3.2) may also be used. Optimizing u over $N \gg 1$ different initial data establishes robustness, so that (1.3.2) may correctly perform future predictions on unknown points.

In Figure 1.3, we see stabilization for the trajectories to some points $\overline{\mathbf{x}}_i \in P^{-1}(\{\vec{y}_i\})$, which are uncontrolled steady states of (1.3.2). This motivates the choice of running target as a steady control-state pair we consider ((5.1.4)), which would then entail bounds for (1.3.1) (see [95]). The practical interest of the turnpike and stabilization analysis when $T \gg 1$ presented herein is presented in Part III. Briefly, it consists in the link to the large-layer regime (common setting for many deep learning applications [176]) and approximation capacity of ResNets, which are the forward Euler discretization of (1.3.2) (see [89]).

1.3.1 Turnpike in Lipschitz-nonlinear optimal control (Chapter 5)

Given $T > 0$, we focus on *control-affine* systems, namely canonical nonlinear systems

$$\dot{y} = f(y, u) \qquad \text{in } (0, T) \tag{1.3.3}$$

with a nonlinearity f of the form

$$f(y, u) = f_0(y) + \sum_{j=1}^m u_j f_j(y) \qquad \text{for } (y, u) \in \mathbb{R}^d \times \mathbb{R}^m, \tag{1.3.4}$$

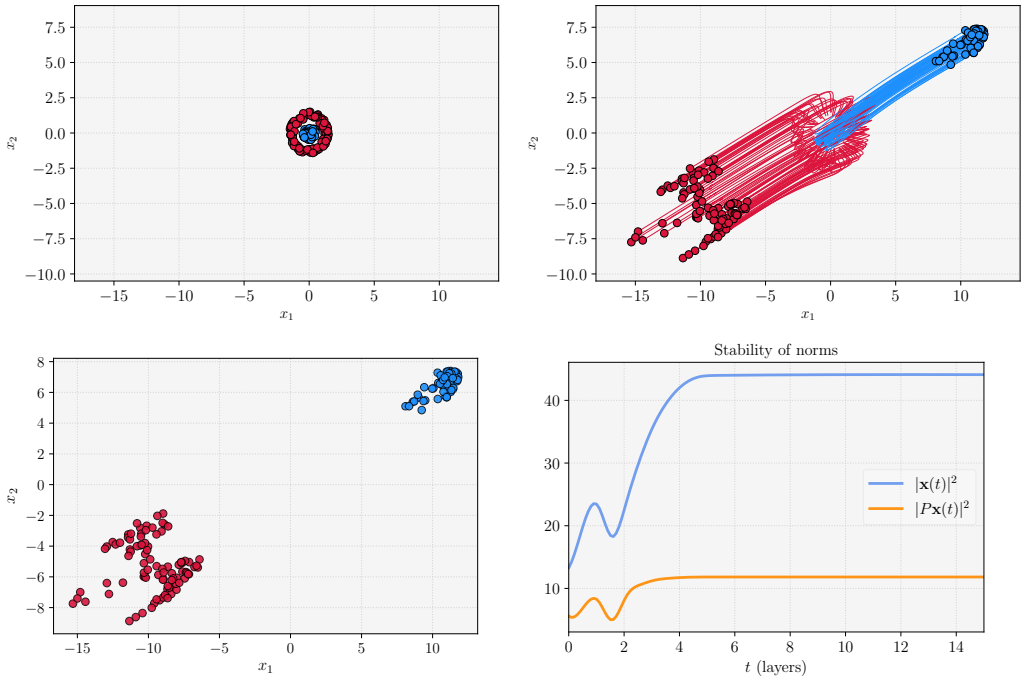


Figure 1.3: *A binary classification task in deep learning.* One aims to separate the data points $\{\vec{x}_i\}_i$ in $\mathbb{R}^2$ (*top left*) with respect to their color by using the controlled flow of (1.3.2) at time $T = 15$, here done by minimizing (1.3.1) ($\vec{y}_i = \pm 1$ for red/blue). We visualize the evolution of the trajectories of (1.3.2) (*top right*) and their output (*bottom left*). We see a stabilization property for the projections, but also the trajectories to some points $\overline{\mathbf{x}}_i \in P^{-1}(\{\vec{y}_i\})$ (*bottom right*).

where the vector fields $f_0, \dots, f_m \in \mathrm{Lip}(\mathbb{R}^d; \mathbb{R}^d)$ are only assumed to be globally Lipschitz continuous. This formulation includes (1.3.2). Given $y^0 \in \mathbb{R}^d$, we will investigate the behavior when $T \gg 1$ of global minimizers $u_T \in L^2(0, T; \mathbb{R}^m)$ to nonnegative functionals of the form

$$J_T(u) := \phi(y(T)) + \int_0^T \|y(t) - \overline{y}\|^2 \, \mathrm{d}t + \int_0^T \|u(t)\|^2 \, \mathrm{d}t, \tag{1.3.5}$$

and of the corresponding solution y_T to (1.3.3) with $y_T(0) = y^0$. Here, $\phi \in C^0(\mathbb{R}^d; \mathbb{R}_+)$ is a given final cost, while $\overline{y} \in \mathbb{R}^d$ is a given running target which we select as an *uncontrolled steady state* of the nonlinear dynamics, namely

$$f_0(\overline{y}) = 0. \tag{1.3.6}$$

Now note that, due to the assumption (1.3.6) on the running target $\overline{y}$, and the form of the nonlinearity f in (1.3.4), it can be seen that $(u_s, y_s) \equiv (0, \overline{y})$ designates the unique optimal stationary solution, namely the unique solution to

$$\inf_{u \in \mathbb{R}^m} \|y - \overline{y}\|^2 + \|u\|^2 \qquad \text{subject to} \quad f(y, u) = 0. \tag{1.3.7}$$

We will henceforth assume that (5.2.1) is *controllable with linear cost* in some time $T_0 > 0$ by means of some control $u \in L^2(0, T_0; \mathbb{R}^m)$; by the latter, we mean that there exists an $r > 0$ and a constant $C(T_0) > 0$ such that

$$\inf_{\substack{u \\ \text{such that} \\ y(0)=y^0,\, y(T_0)=\overline{y}}} \|u\|_{L^2(0,T_0;\mathbb{R}^m)} \leqslant C(T_0) \left\| y^0 - \overline{y} \right\|, \tag{1.3.8}$$

and

$$\inf_{\substack{u \\ \text{such that} \\ y(0)=\overline{y},\, y(T_0)=y^1}} \|u\|_{L^2(0,T_0;\mathbb{R}^m)} \leqslant C(T_0) \left\|y^1 - \overline{y}\right\|, \tag{1.3.9}$$

hold for any $y^0, y^1 \in \left\{x \in \mathbb{R}^d : \|x - \overline{y}\| \leqslant r\right\}$, where $\overline{y} \in \mathbb{R}^d$ is fixed as in (1.3.6).

We may now state our main result.

Theorem 1.5 (Turnpike, [96]). *Assume that $f_0, \dots, f_m \in \mathrm{Lip}(\mathbb{R}^d; \mathbb{R}^d)$ in (1.3.4), and assume that (1.3.3) is controllable in some time $T_0 > 0$ with linear cost estimates (1.3.8) – (1.3.9). Let $y^0 \in \mathbb{R}^d$ be given, and let $\overline{y} \in \mathbb{R}^d$ be as in (1.3.6). Then there exists a time $T^* > 0$ and constants $C_1, C_2, \mu > 0$ such that for any $T \geqslant T^*$, any global minimizer $u_T \in L^2(0,T;\mathbb{R}^m)$ to J_T defined in (1.3.5) and corresponding optimal state y_T solution to (1.3.3) with $y_T(0) = y^0$ satisfy*

$$\|y_T(t) - \overline{y}\| \leqslant C_1 \left(e^{-\mu t} + e^{-\mu(T-t)}\right) \tag{1.3.10}$$

for all $t \in [0,T]$, and

$$\|u_T\|_{L^2(0,T;\mathbb{R}^m)} \leqslant C_2. \tag{1.3.11}$$

Discussion. Let us put the above global nonlinear turnpike result into context.

- The prevalent argument in the literature for proving exponential turnpike results relies on a thorough analysis of the optimality system provided by the Pontryagin Maximum Principle. In the case of nonlinear dynamics, such arguments require nonlinearities which are continuously differentiable: a linearization argument is used, the linear study and a fixed point argument provide nonlinear results under smallness assumptions on the initial data and the target ([222, 261]). The smallness conditions on the initial data can be removed in some specific cases (see e.g. [218]), but to the best of our knowledge, the assumptions on the running target have not been as of yet (albeit, they may be removed under restrictive assumptions, such as strict dissipativity, uniqueness of minimizers and C^2–regular nonlinearities, see [259]).

  Our contribution is a new methodology for proving the turnpike property for nonlinear optimal control problems which does not use the Pontryagin Maximum Principle or require differentiable dynamics, thus being applicable to systems of practical importance such as ReLU activated neural networks.

- In Chapter 5, we illustrate the flexibility of the finite-dimensional arguments presented just below and employ them to the semilinear wave and heat equation. As a matter of fact, the only difference between the finite and infinite dimensional setting is in the proof of uniform control and state bounds by means of quasi-turnpike strategies. The specific proof of turnpike is identical in both cases.

- The rate $\mu > 0$ appearing in (1.3.10) depends on the datum y^0 due to the multiplicative form of the control, but is uniform with respect to y^0 when the control is *additive*, namely, when $f_1, \dots, f_m$ are nonzero constants. This is due to the form of the constant provided by Grönwall arguments (e.g., in Lemma 5.4.1 and Lemma 5.5.2).

When one considers an optimal control problem for J_T without a final cost for the endpoint $y(T)$, Theorem 1.5 can in fact be improved to an *exponential stabilization* estimate to the running target $\overline{y}$.

Corollary 1.3.1 (Stabilization, [96]). *Suppose that $\phi \equiv 0$ in J_T defined in (5.2.3). Under the assumptions of Theorem 1.5, there exists a time $T^* > 0$, and constants $C_1, C_2, \mu > 0$ such that for any $T \geqslant T^*$, any global minimizer $u_T \in L^2(0,T;\mathbb{R}^m)$ to J_T defined in (1.3.5) and corresponding optimal state y_T solution to (1.3.3) with $y_T(0) = y^0$ satisfy (1.3.11) as well as*

$$\|y_T(t) - \overline{y}\| \leqslant C_1 e^{-\mu t} \tag{1.3.12}$$

for all $t \in [0,T]$.

On another hand, when the underlying dynamics (1.3.3) are of *driftless control affine* form (namely, $f_0 \equiv 0$ in (1.3.4)), we can obtain an exponential decay for the optimal controls as well. Note that in this case, any $\overline{y} \in \mathbb{R}^d$ is an admissible running target for J_T, since $f(\overline{y}, 0) = 0$ for any $\overline{y} \in \mathbb{R}^d$.

Corollary 1.3.2 (Control decay, [96]). *Suppose that $f_0 \equiv 0$ in (1.3.4). Under the assumptions of Theorem 1.5, there exists a time $T^* > 0$, and constants $C, \mu > 0$ such that for any $T \geqslant T^*$, any global minimizer $u_T \in L^2(0,T;\mathbb{R}^m)$ to J_T defined in (1.3.5) and corresponding optimal state y_T solution to (1.3.3) with $y_T(0) = y^0$ satisfy (1.3.10) as well as*

$$\|u_T(t)\| \leqslant C\Big(e^{-\mu t} + e^{-\mu(T-t)}\Big) \tag{1.3.13}$$

for a.e. $t \in [0,T]$.
If moreover, $\phi \equiv 0$ in J_T defined in (1.3.5), in addition to (1.3.12), there exist constants $C_1, \mu_1 > 0$ independent of T such that

$$\|u_T(t)\| \leqslant C_1 e^{-\mu_1 t} \tag{1.3.14}$$

holds for a.e. $t \in [0,T]$.

The proof of Corollary 1.3.2 (see Section 5.5.4) will follow by firstly using a specific suboptimal control (constructed using the time-scaling specific to driftless systems) to estimate $J_T(u_T)$ and obtain

$$\int_t^{t+h} \|u_T(s)\|^2 \,\mathrm{d}s \leqslant \int_t^{t+h} \|y_T(s) - \overline{y}\|^2 \,\mathrm{d}s$$

for h small enough, an estimate which, coupled with the turnpike estimates of Theorem 1.5 – Corollary 1.3.1 and the Lebesgue differentiation theorem, will suffice to conclude.

Sketch of the proof of Theorem 1.5. The proof of Theorem 1.5 follows the following scheme. For simplicity, suppose that $T \geqslant 2T^*$.

1). By controllability, we first construct a suboptimal quasi-turnpike control u^1 which is such that the associated state y^1 satisfies $y^1(T_0) = \overline{y}$, and $u^1(t) = 0$ for $t \in [T_0, T]$. Thus, $y^1(t) = \overline{y}$ for $t \in [T_0, T]$. Due to the form of J_T in (1.3.5), this would imply that $J_T(u^1)$ is independent of T, and by using $J_T(u_T) \leqslant J_T(u^1)$, would also entail a uniform bound of $J_T(u_T)$ with respect to T. A Grönwall argument ensures that, moreover,

$$\|y_T - \overline{y}\|_{L^2(0,T;\mathbb{R}^d)} + \|y_T(t) - \overline{y}\| \leqslant C_0 \qquad \text{for all } t \in [0,T] \tag{1.3.15}$$

for some $C_0 > 0$ independent of T. (1.3.15) alone is enough to obtain the desired exponential estimates for $t \in [0,T^*] \cup [T - T^*, T]$, an interval whose length is independent of T. More details can be found in Lemma 5.5.1.

2). Since $T^* \leqslant \frac{T}{2}$, by a simple contradiction argument (see Lemma 5.5.3), there exist $\tau_1 \in [0, T^*)$ and $\tau_2 \in (T - T^*, T]$ such that

$$\|y_T(\tau_i) - \overline{y}\| \leqslant \frac{\|y_T - \overline{y}\|_{L^2(0,T;\mathbb{R}^d)}}{\sqrt{T^*}} \overset{(1.3.15)}{\leqslant} \frac{C_0}{\sqrt{T^*}}. \tag{1.3.16}$$

3). On $[\tau_1, \tau_2]$, the optimal control u_T will minimize a functional without the final cost $\phi(y_T(T))$ but with a terminal constraint on the state y_T. By controllability, using a second suboptimal quasi-turnpike control u^2 satisfying linear control cost estimates (as those in Definition 6.4.1), and using $J_T(u_T) \leqslant J_T(u^2)$ along with a Grönwall argument, one shows an estimate of the form

$$\|y_T(t) - \overline{y}\| \leqslant C_1 \Big(\big\|y_T(\tau_1) - \overline{y}\big\| + \big\|y_T(\tau_2) - \overline{y}\big\| \Big) \tag{1.3.17}$$

$$\overset{(1.3.16)}{\leqslant} \frac{2C_1^2}{\sqrt{T^*}} \tag{1.3.18}$$

for all $t \in [\tau_1, \tau_2]$, thus also for $t \in [T^*, T - T^*] \subset [\tau_1, \tau_2]$ where $C_1 > 0$ is independent of T. The linear control cost estimates of Definition 6.4.1 are used precisely in this step, and are essential in obtaining an estimate of the mould of (1.3.17). For more details, see Lemma 5.5.2.

4). A *bootstrap argument* (Section 5.5.2): estimate (1.3.18) can be iterated by shrinking the time interval to obtain an estimate of the form

$$\|y_T(t) - \overline{y}\| \leqslant \left(\frac{2C_1^2}{\sqrt{T^*}} \right)^n \qquad \text{for} \quad [nT^*, T - nT^*] \tag{1.3.19}$$

for "suitable" $n \geqslant 1$. Then taking $T^* > 4C_1^4$ and a suitable choice of n in (1.3.19) will yield the exponential estimate for $t \in [T^*, T - T^*]$.

1.4 Part III: Interplay of deep learning and control

The goal of supervised machine learning is to conceive models and algorithms that can learn models from a set of labeled examples in an automatized manner, in order to make predictions on new (unlabeled) examples – formally speaking, supervised learning addresses the problem of predicting an unknown function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ from N known and possibly noisy samples $\{\vec{x}_i, \vec{y}_i = f(\vec{x}_i)\}_{i=1}^N$. Depending on the nature of the space of labels $\mathcal{Y}$, one distinguishes two types of supervised learning tasks, namely that of *classification* (labels take values in a finite set of m classes, e.g. $\mathcal{Y} = \{1, \dots, m\}$) and *regression* (labels take continuous values in $\mathcal{Y} \subset \mathbb{R}^m$). Heuristically, supervised learning consists in constructing a map

$$f_{\text{approx}} : \mathcal{X} \longrightarrow \mathcal{P}(\mathcal{Y}),$$

which, desirably, is such that for any $x \in \mathcal{X}$ and for any Borel measurable $A \subset \mathcal{Y}$, $f_{\text{approx}}(x)(A) \simeq 1$ whenever $f(x) \in A$, and $f_{\text{approx}}(x)(A) \simeq 0$ whenever $f(x) \notin A$; here, $\mathcal{P}(\mathcal{Y})$ denotes the space of probability measures on $\mathcal{Y}$. In other words, one looks for a map f_{approx} which approximates the map $x \longmapsto \delta_{f(x)}$ where δ_z stands for the Dirac measure centered at z. In modern machine learning, the map f_{approx} is generally chosen from a class of parametric functions. As one only has N samples of f, the parameters are tuned in order to fit f_{approx} to these data by minimizing a specific loss functional.

Compared to more traditional methods in statistics, this focus on prediction in machine learning has led to many empirical successes on complex tasks where both the data and the models are very high-dimensional. Indeed, in many applications, the dimension d of each sample $\vec{x}_i$ is big compared to the number/dimension m of the labels – in image

classification for instance, a sample of the ImageNet dataset [167], which has $m = 1000$ classes, is an element of $\mathbb{R}^{65536}$. Further examples of such complex tasks include object recognition or scene segmentation in images, speech recognition in audio signals, and natural language understanding, all of which fall under the common umbrella of *artificial intelligence* (AI).

A plethora of methods for finding $f(\cdot)$ efficiently with theoretical and empirical guarantees have been developed and investigated in the machine learning literature in recent decades. Prominent examples, to name a few, include linear parametric methods (e.g., linear or logistic regression), kernel-based methods (e.g. support vector machines), tree-based methods (e.g., decision trees) and so on. We refer to [119] for a comprehensive presentation of these topics.

At the heart of many recent successes on complex tasks are *deep learning models*, which are typically parametric models which iterate and transform the input data over a large sequence of elementary modules. Deep neural networks are such parametrized computational architectures which propagate each individual sample of the input data $\{\vec{x}_i\}_{i=1}^N \in \mathbb{R}^{d\times N}$ across a sequence of linear parametric operators and simple nonlinearities. The so-called *residual neural networks* (ResNets, [140]) may, in the simplest case, be cast as schemes of the mould

$$\begin{cases} \mathbf{x}_i^{k+1} = \mathbf{x}_i^k + \sigma(w^k \mathbf{x}_i^k + b^k) & \text{for } k \in \{0, \dots, N_{\text{layers}} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i \in \mathbb{R}^d \end{cases} \tag{1.4.1}$$

for all $i \in [N]$, where we set $[N] := \{1, \dots, N\}$. The unknowns are the states $\mathbf{x}_i^k \in \mathbb{R}^d$ for any $i \in [N]$, σ is an explicit scalar, Lipschitz continuous nonlinear activation function defined component-wise in (1.4.1) (see Figure 1.4 for examples), $\left\{w^k, b^k\right\}_{k=0}^{N_{\text{layers}}-1}$ are optimizable parameters (controls) with $w^k \in \mathbb{R}^{d\times d}$ – called weights, and $b^k \in \mathbb{R}^d$ – called biases, and $N_{\text{layers}} \geqslant 1$ designates the number of layers referred to as the depth. The training process consists in finding optimal parameters steering all of the network outputs $\mathbf{x}_i^{N_{\text{layers}}}$ as close as possible to the corresponding labels $\vec{y}_i$ by solving

$$\min_{\{w^k, b^k\}_{k=0}^{N_{\text{layers}}-1}} \frac{1}{N} \sum_{i=1}^{N} \text{loss}\left(P\mathbf{x}_i^{N_{\text{layers}}}, \vec{y}_i\right),$$

whilst guaranteeing reliable performance on unseen data (ensuring *generalization*). Here $\text{loss}(\cdot,\cdot)$ is a given continuous and nonnegative function which may change depending on the task in hand – for instance $\text{loss}(x, y) := \|x - y\|_{\ell^p}^p$ for $p = 1, 2$ is commonly used for regression tasks, while $\text{loss}(x, y) = \log(1 + \exp(-yx))$ may be used for binary classification, namely when $\vec{y}_i \in \{-1, +1\}$ (we refer to (6.2.8) for more general settings). On the other hand, $P : \mathbb{R}^d \longrightarrow \mathbb{R}^m$ is an affine map which in practice is either part of the optimizable parameters or may be chosen at random. We shall assume that P is given and specified on a case-by-case basis.

Historically, neural networks as known and applied nowadays date back to the perceptron, introduced by Rosenblatt [231]. However, the major success and breakthrough which has spurred the flurry of works in deep learning over the past decade is the work of Krizhevsky et al. [167] on the ImageNet challenge. This work employs several engineering tricks for training deep neural networks, combined with immense amounts of training data and computing power. Developments of this kind are experimental in nature, and with the increasing availability of computing power and use of deep networks in the past decade, the gap between the theoretical understanding and experimental design has increased. Neural networks are perceived as powerful but complex black boxes developed through engineering craftsmanship, but there is a lack of an in depth theoretical understanding of their fundamental working mechanisms, and in particular, of the choice of various hyperparameters (e.g., the depth and width of the network, the regularization amplitude of the trainable parameters, the learning rate, etc.).

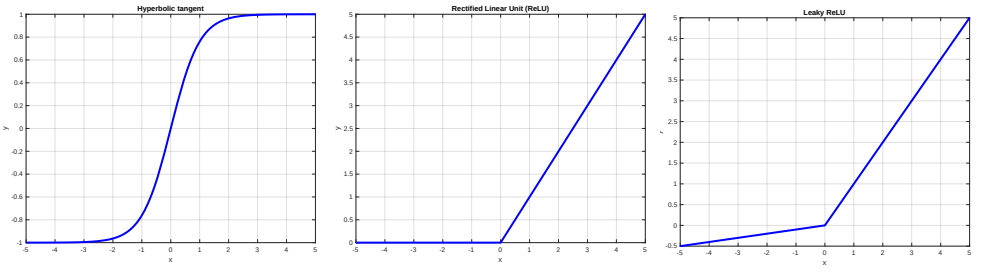


Figure 1.4: Commonly used activation functions include sigmoids such as $\sigma(x) = \tanh(x)$ (*left*), and rectifiers such as $\sigma(x) = \max\{x, \alpha x\}$ with $\alpha \in [0, 1)$; here $\alpha = 0$ (*middle*) and $\alpha = 0.1$ (*right*).

Due to the inherent dynamical systems nature of ResNets, several recent works have aimed at studying an associated continuous-time formulation in some detail, a trend started with the works [89, 129]. This perspective is motivated by the simple observation that for any $i \in [N]$ and for $T > 0$, (1.4.1) is roughly the forward Euler scheme for the neural ordinary differential equation (neural ODE)

$$\begin{cases} \dot{\mathbf{x}}_i(t) = \sigma(w(t)\mathbf{x}_i(t) + b(t)) & \text{for } t \in (0, T) \\ \mathbf{x}_i(0) = \vec{x}_i \in \mathbb{R}^d. \end{cases} \tag{1.4.2}$$

This observation establishes a clear link between supervised learning via neural ODEs and nonlinear optimal control.

Generalization in machine learning, namely learning meaningful and accurate representations from data, relies on underlying forms of simplicity. For instance, the folkloric *no free lunch theorem* [271] states that no single model can succeed on all possible tasks – it is therefore important for one to enforce a form of simplicity in the model or algorithm (following the heuristic implied by Occam's razor), which may (then called inductive bias of the model) or may not rely on a priori knowledge of the data at hand. In the context of neural networks and neural ODEs, such constraints can be induced by enforcing the smoothness and smallness of parameters and the simplicity of the neural network output or neural ODE flow, or by sparsity, which can be achieved by the choice of architecture such as using convolutional neural networks [177], or by seeking regularizing the parameters to seek models which only rely on a small subset of relevant variables/parameters among the large set of available variables/parameters [258].

Motivated by this discussion, we present several original contributions ensuring possible simplicity of learning problems for neural ODEs in various asymptotic regimes of the model hyperparameters.

1.4.1 Large-time asymptotics in deep learning (Chapter 6)

The role of the final time horizon $T > 0$, which plays a key role in the control of dynamical systems, has not been discussed in the context of supervised learning problems via models such as (1.4.2). As each time-step of a discretization to (1.4.2) represents a different layer of the derived neural network (e.g. (1.4.1)), the time horizon $T > 0$ in (1.4.2) may serve as an indicator of the number of layers N_{layers} in the discrete-time context (1.4.1). Thus, a good a priori knowledge of the dynamics of the learning problem over longer time horizons is needed. Such an understanding would lead to potential rules for choosing the number of layers, and enlighten the possible generalization properties when the number of layers is large. We aim to bridge this gap by proposing several insights and an analysis of the role of the time horizon T.

We shall henceforth denote

$$d_u := d \times (d+1), \qquad d_x := d \times N.$$

Moreover, given $w \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$, we shall write

$$\mathbf{w} := \begin{bmatrix} w & & \\ & \ddots & \\ & & w \end{bmatrix} \in \mathbb{R}^{d_x \times d_x}, \qquad \mathbf{b} := \begin{bmatrix} b \\ \vdots \\ b \end{bmatrix} \in \mathbb{R}^{d_x}. \tag{1.4.3}$$

We will consider stacked neural ODEs in $\mathbb{R}^{d_x}$ of the form

$$\begin{cases} \dot{\mathbf{x}}(t) = \sigma(\mathbf{w}(t)\mathbf{x}(t) + \mathbf{b}(t)) & \text{for } t \in (0,T) \\ \mathbf{x}(0) = \mathbf{x}^0 \in \mathbb{R}^{d_x}, \end{cases} \tag{1.4.4}$$

and

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{w}(t)\sigma(\mathbf{x}(t)) + \mathbf{b}(t) & \text{for } t \in (0,T) \\ \mathbf{x}(0) = \mathbf{x}^0 \in \mathbb{R}^{d_x}. \end{cases} \tag{1.4.5}$$

Empirical risk minimization

We first consider the classical supervised learning problem, namely that of regularized empirical risk minimization:

$$\begin{aligned} &\inf_{\substack{[w,b]\in H^k(0,T;\mathbb{R}^{d_u}) \\ \mathbf{x}_i \text{ solves } (7.1.2)}} J_{T,\lambda}(w,b) \\ &= \inf_{\substack{[w,b]\in H^k(0,T;\mathbb{R}^{d_u}) \\ \mathbf{x}_i \text{ solves } (7.1.2)}} \underbrace{\frac{1}{N}\sum_{i=1}^{N} \operatorname{loss}\big(P\mathbf{x}_i(T),\, \vec{y}_i\big)}_{\text{training error}:=\mathcal{E}(\mathbf{x}(T))} + \underbrace{\lambda \Big\| [w,b] \Big\|^2_{H^k(0,T;\mathbb{R}^{d_u})}}_{\text{regularization}} \end{aligned} \tag{1.4.6}$$

with $k = 0, 1$, and we begin by considering the case wherein P and $\operatorname{loss}(\cdot,\cdot)$ are chosen in (1.4.6) so that $\operatorname{loss}(x,x) = 0$. This is for instance the case when loss is a distance inferred by a norm[6] (e.g., $\operatorname{loss}(x,y) = \|x-y\|^p_{\ell^p}$, $p = 1, 2$), and P is an affine map. Such modeling assumptions are typically made in the context of regression tasks[7], wherein when minimizing the training error, one looks to interpolate the training data by means of the projected neural ODE flow.

In this context, one can hopefully expect that the training error of the trajectory associated to the optimal parameters converges to zero, i.e., the model asymptotically interpolates/fits the dataset, and the parameters themselves converge to some limit which satisfies desirable properties. We say that (1.4.5) (resp. (1.4.4)) *interpolates/fits the dataset* $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ *in some time* $T > 0$ if there exists a time $T > 0$ and some parameters $[w,b] \in L^2(0,T;\mathbb{R}^{d_u})$ (resp. in $H^1(0,T;\mathbb{R}^{d_u})$) such that the unique solution $\mathbf{x}$ to (1.4.5) (resp. (1.4.4)), noting (6.3.1).intro, satisfies

$$P\mathbf{x}_i(T) = \vec{y}_i \qquad \text{for all} \quad i \in [N].$$

Clearly, in view of the definition of $\mathcal{E}$, with loss and P as above, if interpolation holds, then the minimum of $\mathcal{E}$ (equal to 0) is attained.

We may state our main result in this context.

[6] We view matrices and tensors as vectors and consider the entry-wise norm throughout this thesis.

[7] Albeit, one may address classification tasks by considering an appropriately chosen (Lipschitz and monotonic) nonlinear $P : \mathbb{R}^m \longrightarrow [-1,1]$, for instance, by truncating $\tanh(x)$.

Theorem 1.6. *Let $\lambda > 0$ be fixed. Suppose that $P : \mathbb{R}^d \longrightarrow \mathbb{R}^m$ is any non-zero affine map, and suppose that $\mathrm{loss} \in C^0(\mathbb{R}^m \times \mathbb{R}^m; \mathbb{R}_+)$ is such that $\mathrm{loss}(x, x) = 0$. Assume that (1.4.5) (resp. (1.4.4) with σ 1–homogeneous) interpolates the dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ in time 1. For any $T \geqslant 1$ let $[w_T, b_T] \in L^2(0, T; \mathbb{R}^{d_u})$ (resp. in $H^1(0, T; \mathbb{R}^{d_u})$) be any pair of global minimizers to $J_{\lambda,T}$, and let $\mathbf{x}_T$ be the unique associated solution to (1.4.5) (resp. (1.4.4)), noting (1.4.3). The following properties then hold.*

1. *There exists a constant $C = C(\mathbf{x}^0, \vec{y}, \lambda) > 0$ independent of T such that*

$$\mathcal{E}(\mathbf{x}_T(T)) \leqslant \frac{C}{T}.$$

2. *There exists a sequence $\{T_n\}_{n=1}^{+\infty}$, with $T_n > 0$ and $T_n \xrightarrow[n \longrightarrow +\infty]{} +\infty$, and some $\mathbf{x}_\circ \in \mathbb{R}^{d_x}$ with $\mathcal{E}(\mathbf{x}_\circ) = 0$ such that, along a subsequence,*

$$\mathbf{x}_{T_n}(T_n) \xrightarrow[n \longrightarrow +\infty]{} \mathbf{x}_\circ. \tag{1.4.7}$$

3. *For any $n \geqslant 1$, set*

$$\begin{aligned} w_n(t) &:= T_n \, w_{T_n}(t \, T_n) \qquad \text{for } t \in [0, 1], \\ b_n(t) &:= T_n \, b_{T_n}(t \, T_n) \qquad \text{for } t \in [0, 1]. \end{aligned}$$

Then along a subsequence,

$$\Big\| [w_n, b_n] - [w^*, b^*] \Big\|_{H^k(0,1;\mathbb{R}^{d_u})} \xrightarrow[n \longrightarrow +\infty]{} 0,$$

where $[w^*, b^*] \in H^k(0, 1; \mathbb{R}^{d_u})$ is some solution to the minimization problem

$$\inf_{\substack{[w,b] \in H^k(0,1;\mathbb{R}^{d_u}) \\ \mathbf{x} \text{ solves (1.4.4) (resp. (1.4.5)) in } [0,1] \\ \text{and} \\ P\mathbf{x}_i(1) = \vec{y}_i \quad \forall i}} \Big\| [w, b] \Big\|^2_{H^k(0,1;\mathbb{R}^{d_u})}.$$

The main underlying idea is to use the homogeneity of the dynamics and the fact that the squared L^2–norm scales like $\frac{1}{T}$ when one performs the inherent change of variable. In fact, exploiting this idea in the case of (1.4.5) and thus $k = 0$, we see that

$$\begin{aligned} &\inf_{\substack{u_T = [w_T, b_T] \in L^2(0,T;\mathbb{R}^{d_u}) \\ \mathbf{x}_T \text{ solves (1.4.5)}}} \mathcal{E}(\mathbf{x}_T(T)) + \lambda \int_0^T \|u_T(t)\|^2 \, \mathrm{d}t \\ &= \inf_{\substack{u_T = [w_T, b_T] \in L^2(0,T;\mathbb{R}^{d_u}) \\ \mathbf{x}_T \text{solves (1.4.5)}}} \mathcal{E}(\mathbf{x}_T(T)) + \frac{\lambda}{T} \int_0^1 \|T u_T(sT)\|^2 \, \mathrm{d}s \\ &= \inf_{\substack{u^1 = [w^1, b^1] \in L^2(0,1;\mathbb{R}^{d_u}) \\ \mathbf{x}^1 \text{solves (1.4.5) on } [0,1]}} \mathcal{E}(\mathbf{x}^1(1)) + \frac{\lambda}{T} \int_0^1 \big\|u^1(s)\big\|^2 \, \mathrm{d}s. \end{aligned}$$

This computation indicates that one may consider the behavior when $T \longrightarrow +\infty$ for fixed $\lambda > 0$ and that when $\lambda \searrow 0$ for fixed $T > 0$ in the same fashion. Although this scaling is specific to the L^2–regularization setting, it motivates completing Theorem 1.6 with the following result.

Theorem 1.7. *Under the assumptions of Theorem 1.6, fix $T > 0$, and for any $\lambda > 0$, let $[w_\lambda, b_\lambda] \in L^2(0,T;\mathbb{R}^{d_u})$ (resp. $H^1(0,T;\mathbb{R}^{d_u})$) be any pair of global minimizers to $J_{\lambda,T}$, and let $\mathbf{x}_\lambda$ be the unique associated solution to (1.4.5) (resp. (1.4.4)), noting (1.4.3). The following properties then hold.*

1. *There exists a constant $C = C(\mathbf{x}^0, \vec{y}, T) > 0$ independent of $\lambda > 0$ such that*

$$\mathcal{E}(\mathbf{x}_\lambda(T)) \leqslant C\lambda.$$

2. *There exists a sequence $\{\lambda_n\}_{n=1}^{+\infty}$, with $\lambda_n > 0$ and $\lambda_n \xrightarrow[n\to+\infty]{} 0$, and some $\mathbf{x}_\circ \in \mathbb{R}^{d_x}$ with $\mathcal{E}(\mathbf{x}_\circ) = 0$ such that, along a subsequence*

$$\mathbf{x}_{\lambda_n}(T) \xrightarrow[n\longrightarrow+\infty]{} \mathbf{x}_\circ.$$

3. *Moreover, along a subsequence,*

$$\Big\| [w_{\lambda_n}, b_{\lambda_n}] - [w^*, b^*] \Big\|_{H^k(0,T;\mathbb{R}^{d_u})} \xrightarrow[n\longrightarrow+\infty]{} 0,$$

where $[w^*, b^*]^\top \in H^k(0,T;\mathbb{R}^{d_u})$ is some solution to the minimization problem

$$\inf_{\substack{[w,b]\in H^k(0,T;\mathbb{R}^{d_u}) \\ \mathbf{x} \text{ solves (1.4.4) (resp. (1.4.5))} \\ \text{and} \\ P\mathbf{x}_i(T) = \vec{y}_i \quad \forall i}} \Big\| [w,b] \Big\|^2_{H^k(0,T;\mathbb{R}^{d_u})}.$$

We now consider the standard setting of *classification tasks*, wherein the labels $\vec{y}_i$ take values in a set of $m \geqslant 2$ classes – unless otherwise stated, we henceforth consider $\vec{y}_i \in [m]$ for all $i \in [N]$. We will focus on the cross-entropy loss in (6.3.4), which we recall, reads

$$\text{loss}(P\mathbf{x}_i(T), \vec{y}_i) := -\log\left(\frac{e^{P\mathbf{x}_i(T)_{\vec{y}_i}}}{\sum_{j=1}^m e^{P\mathbf{x}_i(T)_j}}\right), \tag{1.4.8}$$

where $P : \mathbb{R}^d \longrightarrow \mathbb{R}^m$ is made precise later on. An important feature of the cross-entropy loss is the fact that it is not coercive with respect to the first variable – namely, as $P\mathbf{x}_i(T)_{\vec{y}_i}$ goes to infinity, the loss goes to zero. This is quite in line with intuition regarding classification tasks, as one looks to separate the features with respect to their individual class in the label space $\mathbb{R}^m$.

The problem consisting of classifying a given dataset is closely tied to the following notion of *separability*: we say that (1.4.5) (resp. (1.4.4)) separates the dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ with respect to P if there exists a time $T > 0$ and some parameters $[w,b] \in L^2(0,T;\mathbb{R}^{d_u})$ (resp. in $H^1(0,T;\mathbb{R}^{d_u})$) such that the unique solution $\mathbf{x}$ to (1.4.5) (resp. (1.4.4)) satisfies

$$P\mathbf{x}_i(T)_{\vec{y}_i} - \max_{\substack{j\in[N] \\ j\neq\vec{y}_i}} P\mathbf{x}_i(T)_j > 0 \qquad \text{for all } i \in [N].$$

In other words, a parametrized neural ODE flow separates the given dataset if the corresponding *margin* $\gamma_{[w,b]}$, defined as

$$\gamma_{[w,b]} := \min_{i\in[N]} \left(P\mathbf{x}_i(T)_{\vec{y}_i} - \max_{\substack{j\in[N] \\ j\neq\vec{y}_i}} P\mathbf{x}_i(T)_j \right) \tag{1.4.9}$$

is positive. We may now state our main result[8] in the classification context.

[8] Just as in the regression context, one may show the analog result when $T > 0$ is fixed and $\lambda \searrow 0$ (see Theorem 6.4).

Theorem 1.8. *Let $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ be a given dataset with $\vec{x}_i \in \mathbb{R}^d$ and $\vec{y}_i \in [m]$. Let $\lambda > 0$ be fixed, and let $\mathfrak{Q} : \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^d$ be a non-zero affine map such that $\mathfrak{Q}\vec{x}_i \geqslant 0$ for $i \in [N]$. Set*

$$\mathbf{x}_i^0 := \mathfrak{Q}\vec{x}_i \qquad \text{for } i \in [N],$$

and let $P \in \mathbb{R}^{m\times d}$ be any non-zero matrix such that (1.4.4) with $\sigma(x) = \max\{x, 0\}$ separates the dataset $\{\mathbf{x}_i^0, \vec{y}_i\}_{i=1}^N$ with respect to P in some time $T_0 > 0$, and let γ denote the associated margin as defined in (1.4.9). For any $T \geqslant T_0$, let $[w_T, b_T] \in H^1(0, T; \mathbb{R}^{d_u})$ be any pair of global minimizers to $J_{\lambda,T}$ with cross-entropy loss, and let $\mathbf{x}_T$ be the associated unique solution to (1.4.4) with $\sigma(x) = \max\{x, 0\}$. Then, there exists a constant $C = C(\mathbf{x}^0, \vec{y}, \lambda) > 0$ independent of $T > 0$ such that

$$\mathcal{E}(\mathbf{x}_T(T)) \leqslant \log\left(1 + (m-1)e^{-\gamma\, e^{\frac{T^\alpha}{2}}}\right) + CT^{2\alpha-1} \tag{1.4.10}$$

holds for any $\alpha \in \left(0, \frac{1}{2}\right)$.

Discussion. Let us put all of the above asymptotics results into context.

- **Generalization.** The regularization path limit $\lambda \searrow 0$ has been addressed in some machine learning literature. This was initiated in [233, 232], where the authors study linear logistic regression, and show convergence to the max-margin (the classification analog of minimal norm parameters in the regression context; see the end of the Introduction for more details) as $\lambda \searrow 0$, under the assumption of linearly separable data. The max-margin, support vector machine solution, ([249]) is a special example among all solutions that fit the training data. Another example is the minimal ℓ^2–norm solution for linear regression, and both of these solutions can be shown to ensure generalization by virtue of explicit generalization error estimates [18, 151]. This insight stipulates a generalization capacity of our asymptotic limits as $T \longrightarrow +\infty$ or $\lambda \searrow 0$.

  Our results are an extension of the above-cited works to significantly more compound models such as neural ODEs and ResNets, as, using similar arguments as when $T \longrightarrow +\infty$, we obtain the same conclusions when $\lambda \searrow 0$ and T is fixed.

  Moreover, Theorem 1.6 also stipulates generalization properties – namely, optimizing with $T \gg 1$, which may be interpreted as a larger depth for ResNets, has the practically desirable effect of making the training error close to zero, but by means of almost optimal parameters in the interpolating regime.

- **Universal approximation.** The asymptotic results presented above may (heuristically) be interpreted as approximation results in the sense of the *universal approximation theory*. These are density results for neural networks, and in the simplest cases can be interpreted in terms of the elementary building blocks of measure theory such as the density of simple functions in Lebesgue spaces. The first result in this direction is the seminal work [76], which indicates that shallow neural networks with increasing width, i.e., a superposition of sufficiently many dilated and translated sigmoids, may approximate any continuous function on compact sets. We also refer to [142, 220] for an extension to multi-layer neural networks. Our results are somewhat dual to [76] – therein, to increase the approximation accuracy, the width is allowed to grow, whilst we fix the width and allow the depth to increase. We refer to the thesis [211] for results and a comprehensive review of universal approximation results for ResNets, and to the recent works [182] and [237], for universal approximation results for neural ODE and for observations on the latter's working mechanisms.

  A key caveat of universal approximation results is that there is no scalable method to compute the theoretically guaranteed parameters. On the other hand, our results

provide approximation properties for the trained parameters, albeit for a fixed dataset and not for the unknown underlying function.

- **Completing Theorem 1.8.** As it stands, Theorem 1.8 is very specific to neural ODEs of the form (6.3.2) with ReLU activations, and the specific form of the cross-entropy loss, from which the first term in the estimate (6.3.12) is derived. This is due to the proof strategy, which relies on using the positivity of the right hand side to, in some sense, obtain a linear equation for the projected output features for some auxiliary parameters constructed within the proof, and thus have an explicit solution for these parameters of the form $\sim e^t$. This stimulates the appearance of the second exponential within the log in (1.4.10).

  Moreover, unlike what was done in the regression setting, Theorem 1.8 does not provide a limit for the trained parameters. We refer to Section 1.5.2 for more details regarding this direction.

Augmented empirical risk minimization

We are now interested seeing whether one can obtain better quantitative estimates for the decay of the training error $\mathcal{E}$ to 0 with respect to the time horizon ($\sim$ number of layers) $T > 0$ – namely, improve the rate of convergence of the training error to 0 manifested in Theorem 1.6 and Theorem 1.8. We will henceforth solely concentrate on the ℓ^2–loss:

$$\mathcal{E}(\mathbf{x}) := \frac{1}{N}\sum_{i=1}^{N} \|P\mathbf{x}_i - \vec{y}_i\|^2 \tag{1.4.11}$$

for $\mathbf{x} \in \mathbb{R}^{d_x}$, where $P \in \mathrm{Lip}(\mathbb{R}^d; \mathbb{R}^m)$ is any given surjective and non-zero map, which, in the context of regression, is simply a non-zero affine map, while in the context of binary classification, may be an affine map composed with a sigmoid nonlinearity.

To obtain stronger quantitative estimates, we will introduce a slightly different learning problem, inspired from results in optimal control theory of Part II. For fixed $\lambda > 0$, we will study the behavior when $T \gg 1$ of global minimizers to the functional

$$J_T(w, b) := \mathcal{E}(\mathbf{x}(T)) + \int_0^T \|\mathbf{x}(t) - \overline{\mathbf{x}}\|^2 \, \mathrm{d}t + \lambda \Big\| [w, b] \Big\|^2_{H^k(0,T;\mathbb{R}^{d_u})}, \tag{1.4.12}$$

with $\mathcal{E}$ as in (6.4.1), and where $\overline{\mathbf{x}}_i \in P^{-1}(\{\vec{y}_i\})$ for all $i \in [N]$ are given. We note that, contrary to the case where we minimizing the training error at the final time T, here, the same scaling does not appear which allows us to deduce an equivalence with $\lambda \longrightarrow 0$. Hence, we will solely be interested in the behavior when $T \gg 1$.

We state our main result in the context of the augmented supervised learning problem consisting of minimizing (1.4.12).

Theorem 1.9 (Exponential decay/Turnpike). *Fix $\lambda > 0$, let $P \in \mathrm{Lip}(\mathbb{R}^d; \mathbb{R}^m)$ be any given non-zero and surjective map and let $\overline{\mathbf{x}} \in \mathbb{R}^{d_x}$ with $\overline{\mathbf{x}}_i \in P^{-1}(\{\vec{y}_i\})$ be arbitrary. Suppose that system (1.4.5) (resp. (1.4.4) with σ 1–homogeneous) is controllable with linear cost in some time $T_0 > 0$ in the sense of Definition 6.4.1.*
Then, there exists $T^* > 0$ and positive constants $C_1, C_2, \mu > 0$ depending on $\lambda, \vec{x}_i, \vec{y}_i, N$ such that for any $T \geqslant T^*$, any parameters $[w_T, b_T] \in H^k(0, T; \mathbb{R}^{d_u})$ minimizing (1.4.12), where $k = 0$ in the case of (1.4.5), and $k = 1$ in the case of (1.4.4) and the corresponding unique solution $\mathbf{x}_T$ to (1.4.4) (resp. (1.4.5)) satisfy

$$\big\|w_T(t)\big\| + \big\|b_T(t)\big\| \leqslant C_1 e^{-\mu t}$$

for a.e. $t \in [0, T]$ and

$$\mathcal{E}(\mathbf{x}_T(t)) + \|\mathbf{x}_T(t) - \overline{\mathbf{x}}\| \leqslant C_2 e^{-\mu t}$$

for all $t \in [0, T]$.

Discussion. Curiously enough, up to the best of our knowledge, this is the first theoretical insight in the machine learning literature for supervised learning problems via neural ODEs where one regularizes the empirical risk over the entire horizon $[0, T]$ (i.e., penalizes the features over the entire depth of the associated ResNet). Let us provide some additional comments.

- **Comparison with Theorem 1.6 and Theorem 1.8.** This result is in line with Theorem 1.6 and Theorem 1.8, but with a significantly improved rate of convergence, and thus a better estimate of the time horizon needed to be ε–close to the interpolation or separation regime for any given $\varepsilon > 0$. This ought to be compared with universal approximation results, in which, a key caveat is that there is no scalable method to compute the theoretically guaranteed parameters. In fact, the exponential decay estimate ensures that T need not be chosen too large to render the training error small. Due to the exponentially small global minimizers, numerical experiments show that the learned flow is simple, stipulating possible generalization properties.

- **Extensions.** The second estimate can also be shown to hold for more compound neural ODEs consisting of combinations of (1.4.5) and (1.4.4) (e.g. (6.3.13)). However, due to the lack of homogeneity of the dynamics with respect to the parameters in such cases, we do not know how to show the exponential decay of the optimal parameters.

- **An alternative.** Due to the nature of the proof of Theorem 1.9, which strongly relies on the fact that we may estimate the entire state $\mathbf{x}(t)$ via Grönwall arguments, we have restricted our study to an integral tracking term consisting of the squared $L^2(0, T; \mathbb{R}^{d_x})$–norm, albeit the final cost $\mathcal{E}(\mathbf{x}_T(T))$ allows us to study both classification and regression tasks. However, having to look for targets $\overline{\mathbf{x}}$ in the preimage of the labels $\vec{y}_i$ by P for any general task may not be ideal computationally.

  To alleviate this, at least numerically[9], we observe that the stabilization phenomenon for the output features (and also for the trajectories, although perhaps not with the same rate) persists when the term $\|\mathbf{x}(t) - \overline{\mathbf{x}}\|^2$ is replaced by the training error $\mathcal{E}(\mathbf{x}(t))$ with a general and possibly non-coercive loss, for instance, the cross-entropy loss on a multi-label classification tasks as seen in Figure 1.5 & Figure 1.8 (see the respective examples for modeling details). We stipulate this stabilization phenomenon (be it exponential or not) to possibly hold for global minimizers of functionals of the form

$$J_T(w, b) := \int_0^T \mathcal{E}(\mathbf{x}(t)) \, \mathrm{d}t + \lambda \Big\| [w, b] \Big\|^2_{H^k(0,T;\mathbb{R}^{d_u})}, \tag{1.4.13}$$

  with $\mathcal{E}$ as above, and loss being continuous and nonnegative, but otherwise arbitrary.

1.4.2 Sparse approximation in learning via neural ODEs (Chapter 7)

Sparsity is a highly desirable property in many machine learning and optimization tasks due to the inherent reduction of computational complexity. When induced by ℓ^1–regularization for instance, it has been used extensively for simplifying a machine learning task by selecting a strict subset of the available features to be used in an automatized manner. An illustrative example is the well-known Lasso (least absolute shrinkage

[9] Unless stated otherwise, all software experiments were done using `PyTorch` [214] (and may be found at `https://github.com/borjanG/dynamical.systems`), using the Adam optimizer [159] with learning rate equal to 10^{-3} and `TorchDiffEq` library [61]. Experiments were conducted on a personal MacBook Pro laptop (2.4 GHz Quad-Core Intel Core i5, 16GB RAM, Intel Iris Plus Graphics 1536 MB)

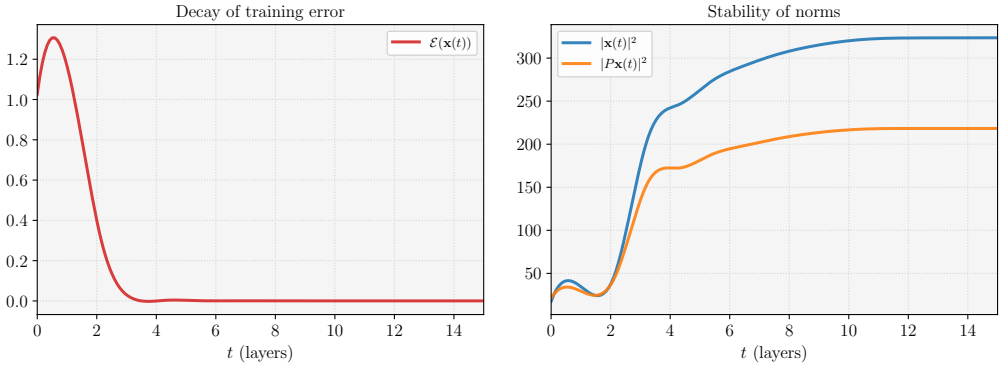


Figure 1.5: **Example 6.4.4:** The decay of the training error (*left*) and stabilization of optimal state trajectory (*right*) for (1.4.13) with cross-entropy loss. See Example 6.4.4 for modeling considerations.

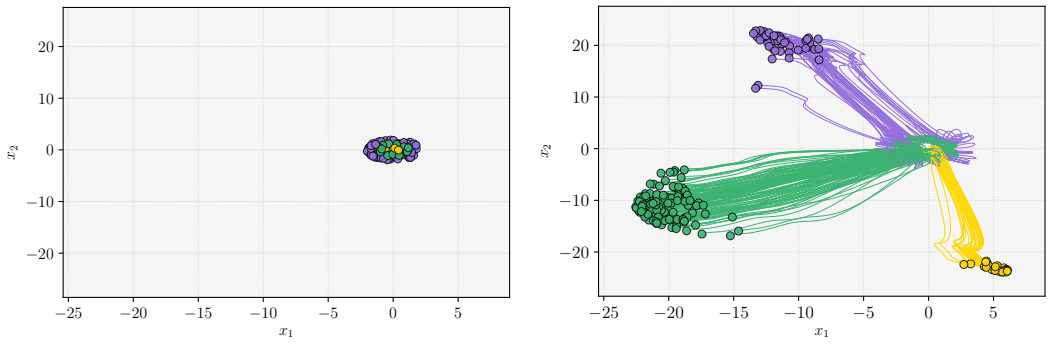


Figure 1.6: **Example 6.4.4:** The training dataset (*left*) and the evolution of the trained neural ODE trajectories $\mathbf{x}_{T,i}(t)$ (*right*) in the phase plane. See Example 6.4.4 for modeling considerations.

and selection operator, [239, 258]), which consists in minimizing a least squares cost function and an ℓ^1–penalty for an affine parametric model, and enforces a subset of the trainable parameters to become zero. As a consequence, the associated features may safely be removed.

Following this line of reasoning, we study supervised learning problems viewed from a continuous-time, neural ODE perspective, and we demonstrate the appearance of sparsity patterns for L^1–regularized minimization problems. More precisely, the supervised learning problem we address in this work consists in minimizing, for $\lambda > 0$ and $T > 0$, a functional of the form

$$J_T(u) := \int_0^T \mathcal{E}(\mathbf{x}(t))\,\mathrm{d}t + \lambda \int_0^T \|u(t)\|_1\,\mathrm{d}t, \tag{1.4.14}$$

over $u = [w, b] \in \mathfrak{U}_{\mathrm{ad},T}$, where $\mathcal{E}$ denotes the empirical risk defined by

$$\mathcal{E}(\mathbf{x}(t)) := \frac{1}{N}\sum_{i=1}^{N} \mathrm{loss}(P\mathbf{x}_i(t), \vec{y}_i). \tag{1.4.15}$$

Here $\mathbf{x} \in C^0([0,T];\mathbb{R}^{d_x})$ solves (1.4.5) (or (1.4.4) with σ 1–homogeneous), $P : \mathbb{R}^d \longrightarrow \mathbb{R}^m$ is a given affine map, and

$$\mathfrak{U}_{\mathrm{ad},T} := \Big\{ u \in L^1(0,T;\mathbb{R}^{d_u}) \colon \ \|u(t)\|_1 \leqslant M \ \text{ a.e. in } (0,T) \Big\}$$

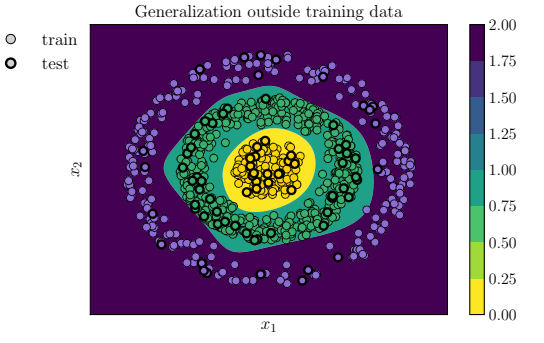


Figure 1.7: **Example 6.4.4:** Plot of the trained classifier on $[-2.5, 2.5]^2$ and its evaluation on the test dataset; the learned flow ensures satisfactory generalization as the shape of the dataset is captures adequately. See Example 6.4.4 for modeling considerations.

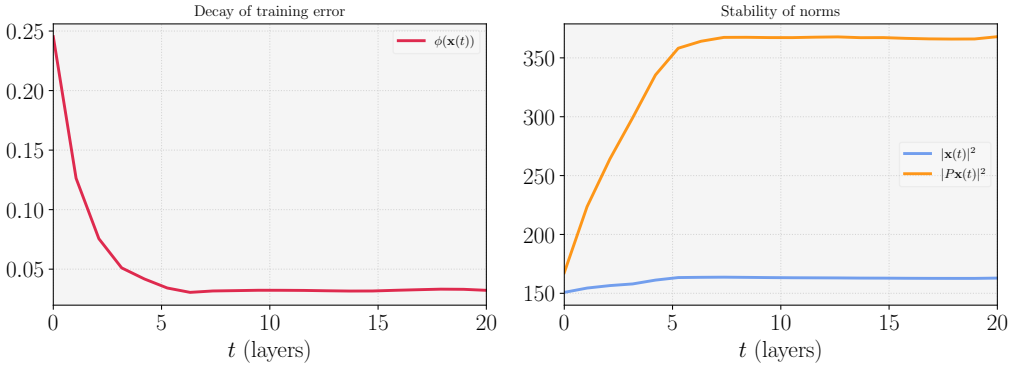


Figure 1.8: **Example 6.4.5:** The decay of the training error (*left*) and stabilization of optimal state trajectory (*right*) for (1.4.13) with cross-entropy loss. See Example 6.4.5 for modeling considerations.

for a fixed thresholding constant $M > 0$[10]. Finally, $\mathrm{loss}(\cdot,\cdot) : \mathbb{R}^m \times \mathcal{Y} \longrightarrow \mathbb{R}_+$ is assumed to satisfy

$$\mathrm{loss}(\cdot, y) \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}^m; \mathbb{R}_+) \qquad \text{and} \qquad \inf_{z \in \mathbb{R}^m} \mathrm{loss}(z, y) = 0, \qquad \text{for any } y \in \mathcal{Y}. \tag{1.4.16}$$

We shall make use of the ℓ^1–norm $\|\cdot\|_1$ on $\mathbb{R}^{d_u}$, defined element-wise as $\|u\|_1 := \sum_{k=1}^{d_u} |u_k|$ for $u = (u_1, \dots, u_{d_u}) \in \mathbb{R}^{d_u}$. We point out that our results would clearly hold for different norms on $\mathbb{R}^{d_u}$ (e.g., the euclidean norm or max norm) by the equivalence of norms.

We will assume that the neural ODE under consideration can interpolate the dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$, either in finite or in infinite time, namely, we shall suppose that there exist parameters such that its corresponding trajectory makes the training error $\mathcal{E}$ vanish, either in finite or in infinite time. More precisely,

1. We say that (1.4.5) (resp. (1.4.4) with σ 1–homogeneous) *interpolates* the dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ in time $T_0 > 0$ if there exist parameters $u \in L^\infty(0, T_0; \mathbb{R}^{d_u})$ such that the corresponding unique solution $\mathbf{x} \in C^0([0, T_0]; \mathbb{R}^{d_x})$ to (1.4.5) (resp. (1.4.4) with σ 1–homogeneous) satisfies

$$\mathcal{E}(\mathbf{x}(T_0)) := \frac{1}{N} \sum_{i=1}^{N} \mathrm{loss}(P\mathbf{x}_i(T_0), \vec{y}_i) = 0.$$

2. We say that (1.4.5) (resp. (1.4.4) with σ 1–homogeneous) *asymptotically interpolates* the dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ if there exist $T_0 > 0$, a function $h \in C^\infty([T_0, \infty); \mathbb{R}_+)$

[10] The L^1–regularization in (7.1.8) enforces the use of sparse parameters concentrated near $t = 0$. We include an L^∞–constraint in the definition of $\mathfrak{U}_{\mathrm{ad},T}$ in order to prevent degeneracy.

Figure 1.9: **Example 6.4.5:** We illustrate the evolution of three individual MNIST samples $\mathbf{x}_i(t) \in \mathbb{R}^{784}$ at times $t \in \{0, 2, 8, 15, 19\}$ – each trajectory stabilizes to some steady configuration after time $t \geqslant 8$. Curiously enough, the neural ODE tends to compress the input digit samples ahead of classifying via the softmax applied to the output features $P\mathbf{x}_i(t) \in \mathbb{R}^{10}$. See Example 6.4.5 for modeling considerations.

satisfying

$$h'(t) < 0 \quad \text{for } t \geqslant T_0 \quad \text{and} \quad \lim_{t \longrightarrow \infty} h(t) = 0,$$

and parameters $u \in L^\infty(\mathbb{R}_+; \mathbb{R}^{d_u})$ such that the corresponding unique solution $\mathbf{x} \in C^0([0, +\infty); \mathbb{R}^{d_x})$ to (1.4.5) (resp. (1.4.4) with σ 1–homogeneous) on $\mathbb{R}_+$ satisfies

$$\mathcal{E}(\mathbf{x}(t)) \leqslant h(t) \qquad \text{for } t \geqslant T_0.$$

We consider asymptotic interpolation due to the occurrence of non-coercive losses which do not attain their minimum, exemplified in the context of classification tasks with losses such as the cross entropy. In fact, in Proposition 7.4.2, we prove that, under suitable assumptions, the asymptotic interpolation property for the cross-entropy loss holds with

$$h(t) = \log\left(1 + (m-1)e^{-\gamma e^t}\right),$$

where $\gamma > 0$ is the *margin* defined by (we set $[m] := \{1, \dots, m\}$)

$$\gamma := \min_{i \in [N]} \left\{ P\mathbf{x}_i(T_0)_{\bar{y}_i} - \max_{\substack{j \in [m] \\ j \neq \bar{y}_i}} P\mathbf{x}_i(T_0)_j \right\}.$$

We may state our main result in this context.

Theorem 1.10. *Let $T > 0$, $\lambda > 0$ and $M > 0$ be fixed, and let $u_T \in U_{ad,T}$ be any (should it exist) global minimizer to J_T defined in (1.4.14), with $\mathcal{E}$ as in (1.4.15), loss satisfying (1.4.16). Let $\mathbf{x}_T \in C^0([0,T];\mathbb{R}^{d_x})$ be the corresponding solution to (1.4.5) (resp. (1.4.4) with σ 1–homogeneous). Then, there exists $T^* \in (0,T]$ such that*

$$\begin{aligned} \|u_T(t)\|_1 &= M \qquad \text{for a.e. } t \in (0,T^*), \\ \|u_T(t)\|_1 &= 0 \qquad \text{for a.e. } t \in (T^*,T) \end{aligned} \tag{1.4.17}$$

and

$$\mathcal{E}(\mathbf{x}_T(T^*)) \leqslant \mathcal{E}(\mathbf{x}_T(t)) \qquad \text{for } t \in [0,T]. \tag{1.4.18}$$

If moreover,

1. *(1.4.5) (resp. (1.4.4) with σ 1–homogeneous) interpolates the dataset in some time $T_0 > 0$, there exists $T_M > 0$ and $C_M > 0$ independent of T such that*

$$T^* \leqslant T_M \qquad \text{and} \qquad \mathcal{E}(\mathbf{x}_T(T^*)) \leqslant \frac{C_M}{T}.$$

2. *(1.4.5) (resp. (1.4.4) with σ 1–homogeneous) asymptotically interpolates the dataset, there exists $C_M > 0$ independent of T such that*

$$T^* \leqslant \frac{C_M}{M} h^{-1}\left(\frac{1}{T}\right) + \frac{1}{M} \qquad \text{and} \qquad \mathcal{E}(\mathbf{x}_T(T^*)) \leqslant \frac{C_M}{T} h^{-1}\left(\frac{1}{T}\right) + \frac{1}{T}.$$

Discussion. Let us comment on the insight provided by the above result.

- **Novel contribution.** Our main result ensures that any[11] minimizer u_T of J_T is sparse in the sense that $u_T \equiv 0$ on (T^*,T) for some $T^* \in (0,T]$. To the best of our knowledge, this is the first such result for nonlinear optimal control problems set in a finite-time horizon, and the first sparsity result in the machine learning literature regarding neural ODEs.

- **Coordinate-wise sparsity.** A related concept to sparsity is that of *coordinate-wise sparsity*, which is described by

$$u_j(t)u_k(t) = 0 \qquad \text{for } j,k \in \{1,\dots,d_u\},\ j \neq k$$

  for $t \in (0,T)$. Equivalently, this entails that at most one coordinate of $u(t)$ is non-zero at time t. This is itself in the spirit of *switching* – we refer the reader to [282] for a comprehensive overview of switching in the context of linear systems.

  In [153], the authors study the occurrence of switching for infinite-time horizon optimal control problems for ODE systems akin to ours, and stipulate that coordinate-wise switching occurs when one considers a regularization of the parameters such as

$$\int_0^T \left(\|u(t)\|_1 + 2 \sum_{\substack{j,k \in [d_u] \\ j \neq k}} |u_j(t)u_k(t)|^{1/2} \right) \mathrm{d}t = \int_0^T \left(\sum_{j=1}^{d_u} |u_j(t)|^{1/2} \right)^2 \mathrm{d}t.$$

  A relevant point of such sparse/switching parameters would be the possibility of allowing the discretized dynamics to alternate dimensions over different time instances, hence in the discrete, ResNet context, allow for a variable width interpretation of the neural ODE models.

[11] One can readily show that a minimizer exists when $\mathbf{f}$ is parametrized as in (1.4.5) by means of the direct method in the calculus of variations. However, for $\mathbf{f}$ as in (1.4.4), ensuring compactness does not appear straightforward.

Since our methodology for the proof of Theorem 1.10 is based on the homogeneity of the neural ODE with respect to the parameters, and the invariance of the $L^1(0,T;\mathbb{R}^{d_u})$–norm with respect to the induced scaling (and does not rely on analyzing the optimality system, thus allowing for Lipschitz-only activation functions), it is entirely plausible to stipulate the occurrence of coordinate-wise sparsity in our finite-time horizon context by applying our arguments presented below to this parameter regularization, since it also is invariant by the induced scaling. We however leave the proof for a forthcoming work.

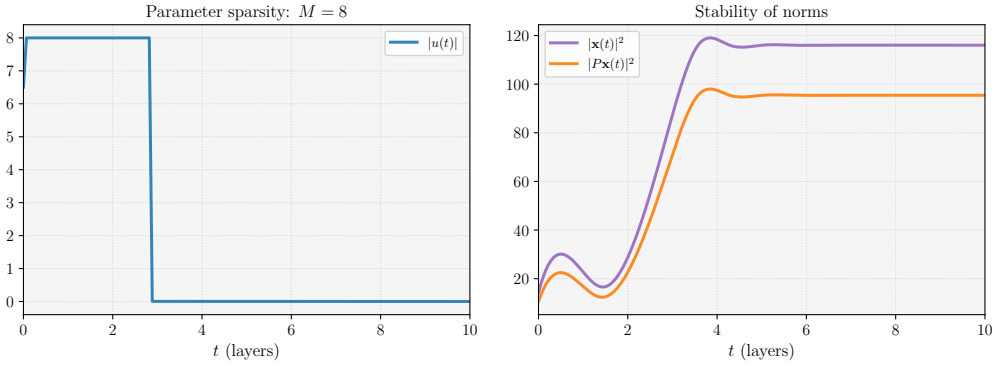


Figure 1.10: We visualize a manifestation of the first part of Theorem 1.10 for a binary classification task. *Left:* the sparsity of the optimal parameters $u_T = [w_T, b_T]$ over time/layer with $M = 8$; *Right:* The norms of the associated state trajectory and projected output (see Figure 1.12). One notes a phase transition at the stopping time $T^* \sim 3$.

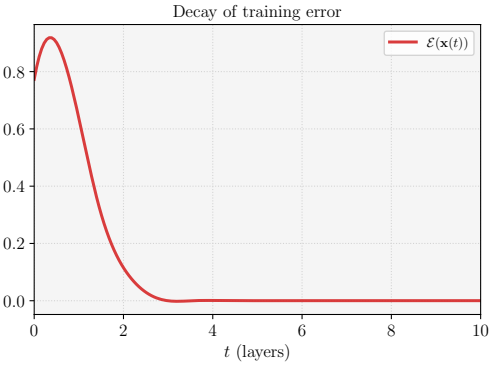


Figure 1.11: We also visualize a manifestation the second part of Theorem 1.10, which stipulates a bound of the training error at the stopping time $T^* \sim 3$ – we in fact see that the training error stabilizes beyond the stopping time.

1.5 A couple of open problems

Open problems specific to each chapter may be found at the end of each individual chapter. In what follows, we present a couple of open problems, related to our contributions, which we believe merit an in-depth investigation.

1.5.1 Controllability of the parabolic obstacle problem

The *parabolic obstacle problem* is the heat-like evolutionary analog of the *classical obstacle problem*, itself being the prototypical stationary free boundary problem [42]. It has seemingly found several applications in practice, mainly in finance [215], where it is used in the modeling of stock options, with the obstacle representing the stock payoff. The

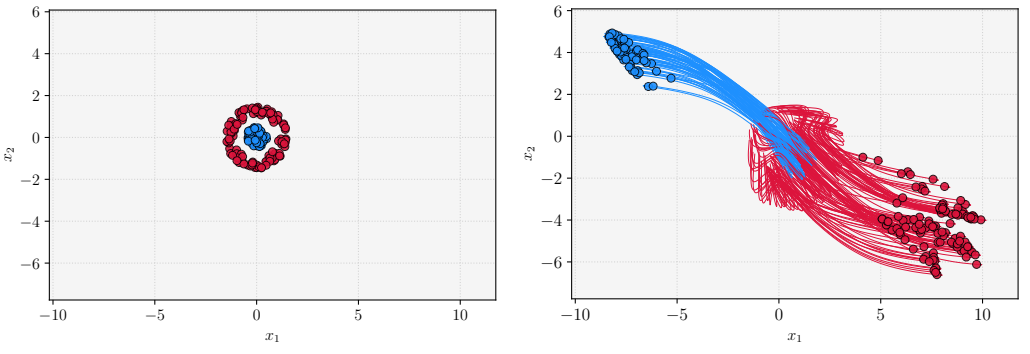


Figure 1.12: We visualize the evolution of the state trajectories of the neural ODE, in the setting of Figure 1.10. *Left:* Initial configuration of training data; *Right:* Evolution and the final configuration $\mathbf{x}_i(T)$ of the trajectories, for $i \in [N]$.

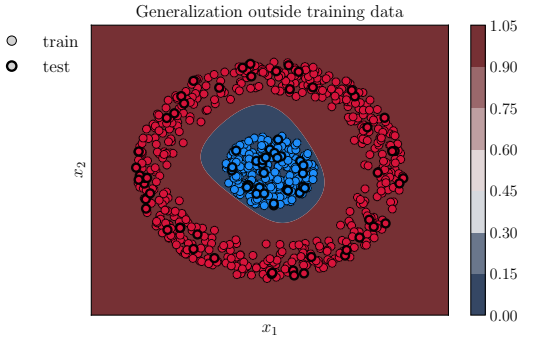


Figure 1.13: We see a satisfactory generalization capacity of the sparsely trained neural ODE flow on the test dataset and other generic points in the domain of the classifier.

parabolic obstacle problem is commonly formulated as a variational inequality: given $T > 0$, an open, bounded and smooth domain $\Omega \subset \mathbb{R}^d$, an *obstacle*

$$\psi \in C^2([0,T] \times \overline{\Omega}), \quad \text{with} \quad \psi < 0 \text{ on } (0,T) \times \partial\Omega \quad \text{and} \quad \psi_{|t=0} = 0 \text{ in } \Omega,$$

and $y_0 \in H_0^1(\Omega)$ satisfying $y_0 \geqslant 0$, find $y \in C^0([0,T]; H_0^1(\Omega)) \cap L^2(0,T; H^2(\Omega))$ satisfying

$$\begin{cases} \displaystyle\int_\Omega \partial_t y(t)(v(t) - y(t))\,\mathrm{d}x + \int_\Omega \nabla y(t) \cdot \nabla(v(t) - y(t))\,\mathrm{d}x \geqslant 0 & \text{in } (0,T) \\ y \geqslant \psi & \text{in } (0,T) \times \Omega \\ y_{t=0} = y_0 & \text{in } \Omega \end{cases} \tag{1.5.1}$$

for all $v \in C^0([0,T]; H_0^1(\Omega)) \cap L^2(0,T; H^2(\Omega))$ with $v \geqslant \psi$. Such a solution y may be found by using a penalization method to obtain a semilinear heat equation, a priori estimates, with the conclusion following by compactness to pass to the limit (see e.g. [32, 41, 15]). Whilst commonly in the literature, the parabolic obstacle problem is formulated with a time-independent obstacle ψ (oftentimes 0), we shall consider this specific case to illustrate an interesting open problem in the context of control.

By choosing specific test functions, one may see that the problem (1.5.1) with homogeneous Dirichlet boundary conditions may be rewritten as

$$\begin{cases} \min\{\partial_t y - \Delta y, y - \psi\} = 0 & \text{in } (0,T) \times \Omega \\ y \geqslant \psi & \text{in } (0,T) \times \Omega \\ y_{|t=0} = y_0 & \text{in } \Omega \end{cases} \tag{1.5.2}$$

Much like the classical obstacle problem, the parabolic obstacle problem is also a free boundary problem, the free boundary being the boundary of the non-contact set, i.e.

$\Gamma(t) := \partial\{y(t) > \psi(t)\}$, on which, it can be shown (see [105]) that, in addition to the matching $y(t) = \psi(t)$, the condition $|\nabla(y(t) - \psi(t))| = 0$ is satisfied. This implies that the free boundary must be an unknown of the problem, as otherwise, the solution would be zero by unique continuation.

An open problem is that of the exact-controllability of problem (1.5.1)/(1.5.2). The natural target would be the stationary solution, namely the solution $\overline{y} \in H^2(\Omega) \cap H_0^1(\Omega)$ to the elliptic obstacle problem:

$$\begin{cases} \min\{-\Delta\overline{y}, \overline{y} - \overline{\psi}\} = 0 & \text{in } \Omega \\ \overline{y} \geqslant \overline{\psi} & \text{in } \Omega, \end{cases} \tag{1.5.3}$$

given $\overline{\psi} \in C^2(\overline{\Omega})$ with $\overline{\psi} < 0$ on $\partial\Omega$.

To this end, one could for instance, consider, instead of homogeneous Dirichlet boundary condition, a boundary control acting on a subset or on the entire fixed boundary $\partial\Omega$. The issue in doing this is the lack of differentiability of the control-to-state map (even for $d = 1$), as illustrated in [217, Section 1.3, pp. 47] (see also [207]). On another hand, to our knowledge, there is no such impediment regarding the obstacle-to-state map $\psi \longmapsto y$. In fact, in the recent work [246], Serfaty and Serra show, for the elliptic obstacle problem defined on the whole $\mathbb{R}^d$, $d \geqslant 2$, that the contact set evolves in a differentiable manner with respect to perturbations of the obstacle, in the context of a Hölder functional framework. This could thus stimulate considering deformations of the obstacle, i.e., viewing the obstacle ψ as a control, with the end goal being the exact controllability of y to $\overline{y}$ in time T:

$$y(T, \cdot) = \overline{y}(\cdot) \qquad \text{in } \Omega.$$

Since $\overline{y}$ is an attractor of the solutions to the evolutionary problem ([67]), one may simply switch the time dependent obstacle ψ to $\overline{\psi}$ beyond time T to remain at $\overline{y}$.

The variational inequality formulation of the parabolic obstacle problem is not too convenient for making use of established controllability strategies for nonlinear problems, generally relying on combining linearization and the HUM. On another hand, to the best of our knowledge, the free boundary formulation does not appear to come along with an evolution equation for the free boundary (rather only the law $|\nabla(y(t) - \psi(t))| = 0$), should one look to parametrize it by the graph of a function. Therefore, the methods presented in Part I of the thesis do not seem (at least not immediately) applicable. On a related note, the exact controllability of the one-dimensional wave equation with a Signorini boundary condition (namely, an obstacle constraint but only at one end) is done in [13] by using a penalization method and uniform estimates to pass to the limit. In the context of the parabolic obstacle problem, one can indeed show the controllability under the obstacle constraint for the penalized problem (see [219]), but obtaining a uniform estimate of the control cost with respect to the penalization parameter does not appear straightforward.

1.5.2 Convergence to a max-margin separator

As discussed in what precedes, in works such as [233, 232, 139, 269], the authors prove the convergence of the normalized margin for ℓ^2–regularized classification problems with cross-entropy loss and homogeneous models (e.g., ReLU activated multi-layer perceptrons) converges to a max-margin classifier as the regularization hyper-parameter λ goes to zero. This is a very desirable property in the context of classification tasks, as max-margin classifiers can be shown to satisfy explicit generalization bounds, and have been seen to generalize well in practice.

The max-margin classifier is, as insinuated, the set of parameters which maximizes the margin between two separated classes, and in fact, the max-margin hyperplane, which is

the hyperplane lying in the middle of the margin, is in fact the solution provided by the (hard-margin) *support vector machine* algorithm (SVM), introduced in [73], wherein one explicitly seeks the hyperplane (see Figure 1.14) with the biggest margin between two (i.e. $\vec{y}_i \in \{-1, 1\}$) linearly separable[12]: we solve

$$\min_{\substack{[\mathbf{w},b]\in\mathbb{R}^d\times\mathbb{R} \\ \text{subject to} \\ \vec{y}_i\left(\mathbf{w}^\top \vec{x}_i - b\right)\geqslant 1 \text{ for } \ i\in[N]}} \|\mathbf{w}\|.$$

We have, however, only shown that the training error decays to zero when $T \longrightarrow +\infty$ (or equivalently $\lambda \searrow 0$) with an explicit rate in the context of ReLU activated neural ODEs and cross-entropy loss. The main difficulty in showing a convergence of the margin or the optimal parameters for neural ODEs – compared to existing works on multi-layer perceptrons–, is the lack of homogeneity of the ODE flow with respect to the parameters. Indeed, while the dynamics is homogeneous with respect to the parameters, there is no guarantee that this should hold for the full flow – in general, it is not even obvious to characterize the solution of the neural ODE with a multiplicative scaling with respect to the solution of the neural ODE with the original parameters.

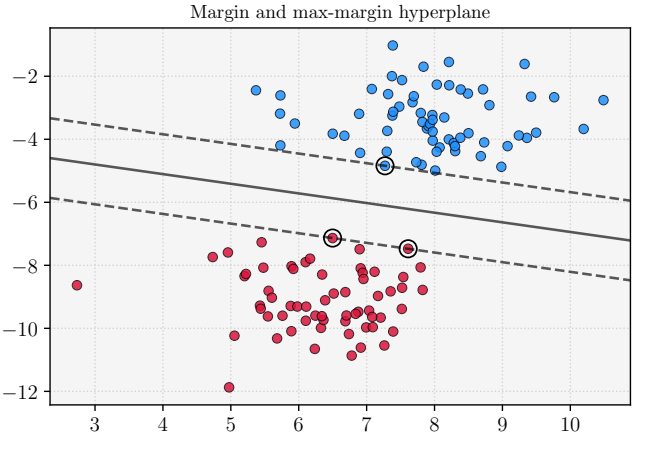


Figure 1.14: The *margin* denotes the distance between the two dashed lines, and the max-margin hyperplane is the line half-way, thus maximizing the distance between the two classes.

Let us provide more detail, and for this, let us consider the following setup. For simplicity, let us focus on the neural ODE (1.4.5). Now, given $u_T = [w_T, b_T]$, denote

$$\overline{u}_T := \frac{[w_T, b_T]}{\left\|[w_T, b_T]\right\|_{L^2(0,T;\mathbb{R}^{d_u})}},$$

and let $\overline{\mathbf{x}}_T$ denote the solution to (1.4.5) associated to $\overline{u}_T$. We recall that the *margin* of $\overline{u}_T$ is defined by

$$\gamma_{\overline{u}_T} := \min_{i\in[N]} \left\{ P\,\overline{\mathbf{x}}_{T,i}(T)_{\vec{y}_i} - \max_{\substack{j\in[m] \\ j\neq\vec{y}_i}} P\,\overline{\mathbf{x}}_{T,i}(T)_j \right\}. \tag{1.5.4}$$

We also define the *max-margin* as

$$\gamma^* := \sup_{\substack{\|u\|_{L^2(0,1;\mathbb{R}^{d_u})}\leqslant 1 \\ \mathbf{x} \text{ solves } (1.4.5)}} \gamma_u. \tag{1.5.5}$$

[12] This can be generalized to data which is not linearly separable by means of *soft-margin minimization*, or by the so-called *kernel trick*. We refer the interested reader to [119].

Note that $\gamma^* > 0$ if and only if the training dataset is separated in the sense of Definition 6.3.4. In view of the results of [233, 232, 139, 269], a natural question one may ask is whether

$$\lim_{T \longrightarrow +\infty} \gamma_{\overline{u}_T} = \gamma^*,$$

where $\overline{u}_T$ are the optimal normalized parameters.

A computation

Let us motivate the eventuality of the convergence result by providing a brief sketch of the proof when $\lambda \searrow 0$ (which is, by means of the scaling discussed in what precedes, equivalent to $T \longrightarrow +\infty$ for neural ODEs) in the context of the following ReLU-activated perceptron without bias[13]:

$$\Phi(x, u) = \mathbf{w}^2 \max\{\mathbf{w}^1 x, 0\},$$

where $u = [\mathbf{w}^1, \mathbf{w}^2]$ with $\mathbf{w}^2 \in \mathbb{R}^{m \times d_{\text{hid}}}$ and $\mathbf{w}^1 \in \mathbb{R}^{d_{\text{hid}} \times d}$. Note that

$$\Phi(x, \alpha u) = \alpha^2 \Phi(x, u), \tag{1.5.6}$$

which is the cornerstone of the subsequent computations. We shall consider the functional

$$J_\lambda(u) := \frac{1}{N} \sum_{i=1}^{N} - \log \left(\frac{e^{\Phi(\vec{x}_i, u)_{\vec{y}_i}}}{\sum_{j=1}^{m} e^{\Phi(\vec{x}_i, u)_{\vec{y}_j}}} \right) + \lambda \|u\|^2, \tag{1.5.7}$$

namely the ℓ^2–regularized empirical risk with cross-entropy loss. We denote by u_λ a global minimizer of this functional. Let $u^* = [\mathbf{w}^{1,*}, \mathbf{w}^{2,*}]$ be a max-margin separation solution, namely a pair of weights such that $\|u^*\| \leqslant 1$ and

$$\gamma_{u^*} = \max_{\|u\| \leqslant 1} \gamma_u := \gamma^*,$$

where γ_u denotes the normalized margin, as defined in (1.5.4):

$$\gamma_u = \min_{i \in [N]} \left\{ \Phi(\vec{x}_i, u)_{\vec{y}_i} - \max_{\substack{j \in [m] \\ j \neq \vec{y}_i}} \Phi(\vec{x}_i, u)_j \right\}. \tag{1.5.8}$$

Now note that for any $\alpha > 0$ and parameters u, due to (1.5.6),

$$\begin{aligned}
J_\lambda(\alpha u) &= \frac{1}{N} \sum_{i=1}^{N} - \log \left(\frac{e^{\alpha^2 \Phi(\vec{x}_i, u)_{\vec{y}_i}}}{\sum_{j=1}^{m} e^{\alpha^2 \Phi(\vec{x}_i, u)_j}} \right) + \lambda \alpha^2 \|u\|^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + \sum_{\substack{j \in [m] \\ j \neq \vec{y}_i}} e^{\alpha^2 \left(\Phi(\vec{x}_i, u)_j - \Phi(\vec{x}_i, u)_{\vec{y}_i} \right)} \right) + \lambda \alpha^2 \|u\|^2 \qquad (1.5.9) \\
&\leqslant \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + (m-1) e^{\alpha^2 \left(\max_{\substack{j \in [m] \\ j \neq \vec{y}_i}} \Phi(\vec{x}_i, u)_j - \Phi(\vec{x}_i, u)_{\vec{y}_i} \right)} \right) + \lambda \alpha^2 \|u\|^2 \\
&\leqslant \log \left(1 + (m-1) e^{-\alpha^2 \left(\Phi(\vec{x}_i, u)_{\vec{y}_i} - \max_{\substack{j \in [m] \\ j \neq \vec{y}_i}} \Phi(\vec{x}_i, u)_j \right)} \right) + \lambda \alpha^2 \|u\|^2. \qquad (1.5.10)
\end{aligned}$$

[13] One can readily adapt the subsequent computations to multi-layer perceptrons without bias, the only change being the exponent of the homogeneity.

On the other hand, in (1.5.9) we also observe that

$$\sum_{\substack{j\in[m]\\ j\neq\vec{y}_i}} e^{\alpha^2\left(\Phi(\vec{x}_i,u)_j-\Phi(\vec{x}_i,u)_{\vec{y}_i}\right)} \geqslant \exp\left(\alpha^2\left(\max_{\substack{j\in[m]\\ j\neq\vec{y}_i}} \Phi(\vec{x}_i,u)_j-\Phi(\vec{x}_i,u)_{\vec{y}_i}\right)\right),$$

and thus, we may lower bound in the identity (1.5.9) by

$$J_\lambda(\alpha u) \geqslant \frac{1}{N}\log\left(1+e^{-\alpha^2\left(\Phi(\vec{x}_i,u)_{\vec{y}_i}-\max_{\substack{j\in[m]\\ j\neq\vec{y}_i}} \Phi(\vec{x}_i,u)_j\right)}\right)+\lambda\alpha^2\|u\|^2. \tag{1.5.11}$$

Now picking $\alpha=\|u_\lambda\|$ and $u=u^*$ in (1.5.10), and noting that $\|u^*\|\leqslant 1$, we deduce

$$J_\lambda\left(u^*\|u_\lambda\|\right)\leqslant \log\left(1+(m-1)e^{-\|u_\lambda\|^2\ \gamma^*}\right)+\lambda\|u_\lambda\|^2. \tag{1.5.12}$$

We then use (1.5.11) with $\alpha=\|u_\lambda\|$ and $u=\overline{u}_\lambda:=\frac{u_\lambda}{\|u_\lambda\|}$ with the effect of

$$J_\lambda(u_\lambda)\geqslant \frac{1}{N}\log\left(1+e^{-\|u_\lambda\|^2\gamma_{\overline{u}_\lambda}}\right)+\lambda\|u_\lambda\|^2. \tag{1.5.13}$$

Combining (1.5.12), (1.5.13) with the optimality of u_λ, we deduce that

$$N\log\left(1+(m-1)e^{-\|u_\lambda\|^2\gamma^*}\right)\geqslant \log\left(1+e^{-\|u_\lambda\|^2\gamma_{\overline{u}_T}}\right). \tag{1.5.14}$$

This holds for any $\lambda>0$. Then, since $\|u_\lambda\|\longrightarrow+\infty$ as $\lambda\searrow 0$ (this can be shown by contradiction), we may Taylor expand (1.5.14), and after some elementary arguments, deduce that

$$\liminf_{\lambda\searrow 0}\ \gamma_{\overline{u}_\lambda}\ \geqslant\gamma^*.$$

Since by definition, $\gamma^*\geqslant\gamma_{\overline{u}_\lambda}$, this would lead to the convergence of the normalized margin.

Discussion

The main caveat with the above computations' applicability to ResNets and neural ODEs lies in the homogeneity of the output. In the neural ODE context, this would entail having a property of the form

$$P\mathbf{x}_T^\alpha(T)=\alpha^r P\mathbf{x}_T(T)\qquad\qquad\text{for all }\alpha>0, \tag{1.5.15}$$

for some $r>0$, where $\mathbf{x}_T$ is the solution to (6.3.3) associated to the optimal parameters $[w_T,b_T]$ (namely global minimizers $J_{\lambda,T}$, where $\lambda>0$ is fixed), whereas $\mathbf{x}_T^\alpha$ is the solution to (6.3.3) associated to $[\alpha w_T,\alpha b_T]$. This appears to be an unrealistic situation because of the presence of the multiplicative weight matrix $w(t)$, but perhaps, the polynomial homogeneity could be replaced by some function $\varphi(\alpha)$, with $\varphi:\ \mathbb{R}_+\longrightarrow\mathbb{R}_+$ increasing and bijective (e.g., of exponential form).

We strongly expect some convergence result for the margin associated to the optimal parameters to hold, at least in the context of ReLU activated neural ODEs and a setting akin to Theorem 6.3. We leave this topic open for further investigation.

Part I

Controllability of free boundary problems

Chapter 2

One-dimensional viscous free boundary flows

Abstract. In this chapter, we address the local controllability of a one-dimensional free boundary problem for a fluid governed by the viscous Burgers equation. The free boundary manifests itself as one moving end of the interval, and its evolution is given by the value of the fluid velocity at this endpoint. We prove that, by means of a control actuating along the fixed boundary, we may steer the fluid to constant velocity in addition to prescribing the free boundary's position, provided the initial velocities and interface positions are close enough.

Keywords. Controllability, free boundary problem, viscous Burgers equation.
AMS Subject Classification. 93B05, 35R35, 35Q35, 93C20.

This Chapter is taken from [116]:

> *Controllability of one-dimensional viscous free boundary flows.*
> B. Geshkovski and E. Zuazua, 2019.
> https://hal.archives-ouvertes.fr/hal-02277740/
> Accepted for publication in SIAM J. Control Optim.

Chapter Contents

2.1 Introduction and main result

Let $T > 0$ be a given positive time. We consider the following problem for the viscous Burgers equation:

$$\begin{cases} v_t - v_{zz} + vv_z = 0 & \text{in } (0,T) \times (0,\ell(t)) \\ v(t,0) = u(t), \quad v_z(t,\ell(t)) = 0 & \text{in } (0,T) \\ \ell'(t) = v(t,\ell(t)) & \text{in } (0,T) \\ v(0,z) = v_0(z), \quad \ell(0) = \ell_0 & \text{in } (0,\ell_0). \end{cases} \tag{2.1.1}$$

System (2.1.1) is a free boundary problem, where the unknown is the pair (v,ℓ), with ℓ representing the free boundary. Here $\ell_0 > 0$, and $u = u(t)$ is a control actuating along the fixed boundary $z = 0$. Henceforth and in the above, we use the notation $(0,T) \times (0,\ell(t))$ for the set $\{(t,z) \in (0,T) \times \mathbb{R} \colon 0 < z < \ell(t)\}$, with analogue notation for the closure of the latter.

Model (2.1.1) is presented and studied by Caboussat & Rappaz in [36, 37], where local-in-time existence and uniqueness of strong solutions are shown, supplemented by numerical studies. It may be seen as a one-dimensional simplification of the incompressible Navier-Stokes equations with a free surface set in $\mathbb{R}^d$ with $d = 2, 3$, as encountered in the works of Beale [19, 20], and Maronnier, Picasso & Rappaz [199], where particular emphasis is given on the application to *mould filling*. The state of System (2.1.1) involves the velocity $v(t,z)$ of the one-dimensional fluid and the free boundary $\ell(t)$, whose counterpart in dimension $d \geqslant 2$ would represent the position of the free surface of the fluid. The fluid velocity is governed by the viscous Burgers equation, while the dynamics of the free boundary follow the fluid velocity, as per the equation $\ell'(t) = v(t,\ell(t))$.

As the state of the system (2.1.1) consists of two components (v,ℓ), the natural exact-controllability problem, which is the main goal of this work, is to steer *both components* to a priori defined targets in a given time $T > 0$. This would ensure the entire system remains in such a configuration after the time T has elapsed. Formulated as such, this control problem has not been accurately addressed in the literature for systems where the coupling between the PDE and ODE components is only done through the boundary of the domain, as in (2.1.1). Through this work, we aim to present a systematic and ordered methodology for addressing such compound control problems.

The most general and feasible targets to which one may control both components of (2.1.1) are *time-dependent trajectories* of (2.1.1), namely free solutions to (2.1.1). The question of controllability to non-trivial trajectories is however not straightforward at all. This is observed on the level of the system linearized around the non-trivial target trajectory, which contains several non-local trace terms (see (2.5.1)). Consequently, in terms of the adjoint problem one obtains non-standard boundary conditions (see (2.5.3)) for which, up to the best of our knowledge, observability inequalities are lacking. This is discussed in more detail in Section 2.5.1, and the general problem of controllability to arbitrary trajectories remains open.

At this point, we observe that for any $\ell_* > 0$, the pair $(\overline{v},\overline{\ell})$ with

$$\overline{v} \in \mathbb{R}, \quad \overline{\ell}(t) = \ell_* + \overline{v}t > 0 \quad \text{in } [0,T], \tag{2.1.2}$$

is an explicit, non-trivial solution to System (2.1.1) with $u \equiv \overline{v}$. As discussed in Section 2.2, the system linearized around this trajectory does not manifest the issues appearing in the general trajectory case. The main goal of this work is to prove the local exact-controllability for (2.1.1) to this particular trajectory. To be more precise, given an arbitrary constant velocity $\overline{v}$ and an initial position ℓ_*, we want to show that whenever (v_0,ℓ_0) are sufficiently close to $(\overline{v},\ell_*)$ (see Figure 2.1), one can find a control $u(t)$ such that the corresponding trajectory (v,ℓ) to (2.1.1) connects (v_0,ℓ_0) to the target $(\overline{v},\ell_* + \overline{v}T)$ at time T. This is reflected in our main result.

Theorem 2.1. *Let $T > 0$, $\ell_* > 0$ and $\overline{v} \in \mathbb{R}$ be such that $\overline{\ell}(t) = \ell_* + \overline{v}t > 0$ for all $t \in [0,T]$. There exists $r > 0$ such that for all $\ell_0 > 0$ and $v_0 \in H^1(0,\ell_0)$ satisfying*

$$\|v_0 - \overline{v}\|_{H^1(0,\ell_0)} + |\ell_0 - \ell_*| \leqslant r,$$

there exists a control $u \in H^{3/4}(0,T)$ such that the unique solution

$$\ell \in C^1([0,T]) \quad v \in L^2\Big(0,T;H^2(0,\ell(\cdot))\Big) \cap C^0\Big([0,T];H^1(0,\ell(\cdot))\Big)$$

of (2.1.1) satisfies

$$\inf_{t\in[0,T]} \ell(t) > 0 \quad \text{and} \quad \ell(T) = \overline{\ell}(T) \quad \text{and} \quad v(T,\cdot) = \overline{v} \quad \text{in } (0,\ell(T)).$$

Moreover, one has

$$\|u\|_{H^{3/4}(0,T)} \lesssim_T \|v_0 - \overline{v}\|_{H^1(0,\ell_0)} + |\ell_0 - \ell_*|.$$

Our proof combines several elements of control of parabolic equations in a systematic and ordered way, in view of establishing a well-defined and clear methodology for tackling controllability problems for free boundary systems such as (2.1.1).

A couple of remarks are in order.

Remark 2.1.1. *It is readily seen that Theorem 2.1 also covers the case of null-controllability of the state and prescribing the position of the interface, by considering $(\overline{v},\overline{\ell}) = (0,\ell_*)$ with $\ell_* > 0$. Aside from the trivial solution $(0,\ell_*)$, we may also look to potentially control to the* stationary solutions *of (2.1.1), namely, time-independent solutions. In other words, given $\ell_* > 0$ and $\bar{v} \in \mathbb{R}$ we seek to compute the solutions to*

$$\begin{cases} -v_{zz} + vv_z = 0 & \text{in } (0,\ell_*) \\ v(0) = \bar{v}, \quad v(\ell_*) = 0, \quad v_z(\ell_*) = 0. \end{cases} \tag{2.1.3}$$

It may be checked that the only solution to the second-order equation in (2.1.3) is $v \equiv 0$, which enhances our interest in time-dependent trajectories as targets.

Remark 2.1.2. *The result we prove here is local (a global result is not known also for similar problems such as (2.1.4), (2.1.5)). One may think of combining this local result with a stabilization argument, which, should stabilization hold, would allow to steer System (2.1.1) to a neighborhood of the target wherein the local controllability result applies. However, while the PDE component may possess an inherent dissipative mechanism, the asymptotic position of the free boundary is generally not known for problems of this nature. See Section 2.5.2 for more details.*

2.1.1 State of the art

The controllability aspects of one-dimensional, parabolic free-boundary problems similar to (2.1.1) have been addressed in several recent works (see e.g. [99, 103, 102, 115]). In [99, 103], Fernández-Cara et al. consider the one-phase Stefan problem

$$\begin{cases} v_t - v_{zz} = 0 & \text{in } (0,T)\times(0,\ell(t)) \\ v(t,0) = u(t), \quad v(t,\ell(t)) = 0 & \text{in } (0,T) \\ \ell'(t) = -v_z(t,\ell(t)) & \text{in } (0,T) \\ v(0,z) = v_0(z), \quad \ell(0) = \ell_0 & \text{in } (0,\ell_0). \end{cases} \tag{2.1.4}$$

We stress that in [99, 103], a null-controllability result where *only the first component* v is controlled is shown, i.e. $v(T,\cdot) = 0$ in $(0,\ell(T))$, for small initial data v_0. Such results are partial as they cannot ensure that the entire system remain in such the prescribed

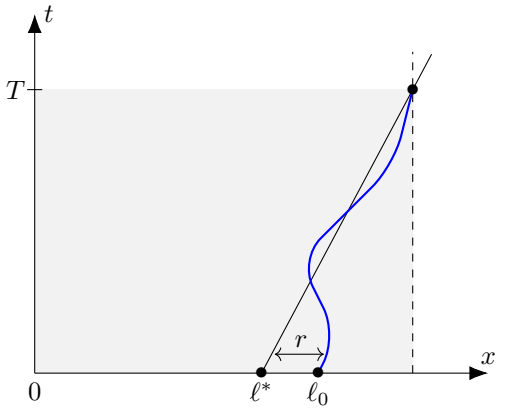


Figure 2.1: Controllability of the position of the free surface ℓ (blue curve) to the reference interface $\overline{\ell}$ (black) at time T, provided the initial positions are close enough.

configuration past the time horizon T. The authors' proof relies on fixing the free boundary $\ell \in C^1([0,T])$ (and removing the equation for the velocity ℓ'), and proving an observability inequality for the linear heat equation in the non-cylindrical domain $(0,T)\times(0,\ell(t))$, with a constant uniform in ℓ. The conclusion for (2.1.4) follows by means of a Schauder fixed-point argument applied to the map $\ell \longmapsto \ell_0 - \int_0^{\cdot} v_x^{\ell}(\tau,\ell(\tau))\,\mathrm{d}\tau$ in an appropriate subspace of $C^1([0,T])$. In [102], the authors obtain the same local controllability result by means of a different technique, which relies on a transformation to a fixed domain, a linear controllability test and an inverse function argument. Our strategy of proof for the controllability of both states of (2.1.1) has some resemblance to that in [102], but with several important technical differences. Moreover, with small adjustments, the control strategy we present herein also yields a local null-controllability result for both the solution and the free boundary of the Stefan problem (2.1.4), namely $\ell(T)=\ell_*$ and $v(T,\cdot)=0$ in $(0,\ell_*)$ whenever v_0 and $\ell_0-\ell_*$ are small enough.

Comparison with fluid-structure interaction problems

Free boundary problems which arise in fluid-structure interaction have also been addressed. Doubova & Fernández-Cara [100] as well as Liu, Takahashi & Tucsnak [191] consider the system

$$\begin{cases} v_t - v_{zz} + vv_z = 0 & \text{in } (0,T)\times(-1,\ell(t))\cup(\ell(t),1) \\ v(t,-1)=u_1(t), \quad v(t,1)=u_2(t) & \text{in } (0,T) \\ v(t,\ell(t))=\ell'(t) & \text{in } (0,T) \\ m\ell''(t)=[v_z](t,\ell(t)) & \text{in } (0,T) \\ v(0,z)=v_0(z), \quad \ell(0)=\ell_0, \quad \ell'(0)=\ell_1 & \text{in } (-1,\ell_0)\cup(\ell_0,1), \end{cases} \tag{2.1.5}$$

which is first introduced by Vázquez & Zuazua [264, 265], where global in-time well-posedness, self-similar asymptotics and particle collision are addressed (see also [195] for a related study). The free boundary $\ell(t)$ represents the displacement/position of a solid particle of mass $m>0$, which splits the domain in two parts. The null-controllability of (2.1.5) refers to controlling three components: the fluid velocity $v(T,\cdot)=0$, the particle velocity $\ell'(T)=0$, and the particle's position $\ell(T)=0$.

In [100], controls u_1, u_2 are used on both boundaries in view of applying a Carleman based strategy. Such an approach is not feasible when there is a control at only one end (i.e. $u_2=0$) because of the lack of connectivity of the fluid domain. This issue was mended in [191], where the authors introduce a systematic methodology for tackling the null-controllability of parabolic systems in spite of source terms, without requiring Carleman

inequalities (they thus use spectral techniques). We also refer to the work of Cindea, Micu, Roventa and Tucsnak [65], where the authors consider a control actuating only on the moving particle: $m\ell''(t) = [v_z](t, \ell(t)) + u(t)$. They prove global null-controllability (in large time) for the fluid and particle velocities, and approximate controllability for the particle's position. We refer to the recent work [228] for a technical improvement of this result. The lack of connectivity of the fluid domain does not appear in two and three dimensions, and the Carleman-based approach has been successfully applied for proving local null-controllability results for fluid-rigid-body systems (see [30, 143] and the references therein) where the control is generally actuating along a part of the fixed boundary.

Remark 2.1.3. *At this point we remark that there is a notable difference between problems of the type* (2.1.5) *and* (2.1.1). *Indeed, the former system has a stronger coupling than the latter systems due to the presence of two equations for the free boundary ℓ. This can be seen when linearizing both systems around their trivial trajectory (after fixing the domain). In the linearization of* (2.1.1) *(see* (2.5.1) *with $a \equiv 1$, $b, c, d, e \equiv 0$ and Section 2 for details),*

$$\begin{cases} y_t - y_{xx} = 0 & \text{in } (0,T)\times(0,1) \\ y(t,0) = u(t), \quad y_x(t,1) = 0 & \text{in } (0,T) \\ \ell'(t) = y(t,1) & \text{in } (0,T) \\ y(0,x) = y_0(x), \quad \ell(0) = \ell_0 & \text{in } (0,1), \end{cases}$$

the PDE and ODE components are decoupled, as the linear PDE may be solved without any knowledge of the ODE component. On the other hand, the linearization of (2.1.5) *around the trivial solution (see [191])*

$$\begin{cases} y_t - y_{xx} = 0 & \text{in } (0,T)\times(-1,0)\cup(0,1) \\ y(t,-1) = u(t), \quad y(t,1) = 0 & \text{in } (0,T) \\ y(t,0) = \ell'(t) & \text{in } (0,T) \\ m\ell''(t) = [y_x](t,0) & \text{in } (0,T) \\ y(0,x) = y_0(x), \quad \ell(0) = \ell_0, \quad \ell'(0) = \ell_1 & \text{in } (-1,\ell_0)\cup(\ell_0,1), \end{cases}$$

preserves the coupling of the PDE component and the ODE component because of the presence of two equations for the latter.

In the above-cited works on fluid-structure problems, the controllability problem addressed is that of controlling the PDE component to zero and the ODE component(s) to some given reference points. For the case of non-trivial stationary solutions and trajectories as targets, much less is known. In [17], Badra & Takahashi prove feedback stabilization to non-trivial stationary solutions for (2.1.5). Therein, it can also be seen that the question of controllability to non-trivial stationary solutions is not straightforward. This is observed on the level of the system linearized around the target, which contains several trace terms (as in (2.5.1)). As a result, in terms of the adjoint problem, one obtains non-local boundary conditions (similar to (2.5.3)), for which observability inequalities are lacking.

We also refer to Dunbar et al. [85, 84] for motion planning and flatness control, and Krstic et al. [161, 162, 163, 164] and the references therein for feedback stabilization via backstepping design of the Stefan problem (2.1.4), see also Phan & Rodrigues [216] for stabilization to trajectories for general parabolic problems.

As discussed in what precedes, up to the best of our knowledge, the question of controllability to non-trivial trajectories (or even non-trivial stationary states) for parabolic free boundary problems such as (2.1.1), (2.1.4), (2.1.5) has not been addressed in the literature. We aim to present some of the difficulties which appear in solving this kind of control problem through this work.

2.1.2 Scope

In Section 2, we reformulate the control problem (2.1.1) on the time-independent domain $(0,1)$. We give the linearization of (2.1.1) around the target trajectory (see Section 5 for the general linearization and a brief discussion on the possible strategies for the general controllability to trajectories problem). In Section 3, we prove the null-controllability of the system linearized around $(\overline{v},\overline{\ell})$. The PDE component is a linear heat equation with a source term, and the ODE component is simply an integrator of the heat solution's Dirichlet trace. The controllability requirement for the second component may thus be seen as a finite-dimensional constraint on the control. An improved observability inequality along with an adaptation of the HUM method provide the desired controllability result for both components of the linearized system. In Section 4, we come back to the nonlinear problem by means of a Banach fixed point argument.

2.2 Reformulation of the problem

Transformation

To take advantage of a simplified functional setting, it is more advantageous to reformulate (2.1.1) in a domain which is time-independent. In view of linearizing, perturbations around the target trajectory would be defined in the same domain.

To this end, let us define the pull-back velocity function $w:(0,1)\to\mathbb{R}$ by

$$w(t,x)=v(t,z),\quad x=\frac{z}{\ell(t)}\qquad \text{for } x\in(0,1). \tag{2.2.1}$$

A simple application of the chain rule gives the following system of equations for w:

$$\begin{cases} w_t-\dfrac{1}{\ell^2}w_{xx}-\dfrac{\ell'}{\ell}xw_x+\dfrac{1}{\ell}ww_x=0 & \text{in } (0,T)\times(0,1)\\ w(t,0)=u(t),\quad w_x(t,1)=0 & \text{in } (0,T)\\ \ell'(t)=w(t,1) & \text{in } (0,T)\\ w(0,x)=w_0(x),\quad \ell(0)=\ell_0 & \text{in } (0,1), \end{cases} \tag{2.2.2}$$

where $w_0(x)=v_0(\ell_0 x)$. As (2.1.1) and (2.2.2) are equivalent provided $\ell(t)>0$ in $[0,T]$, we will henceforth concentrate our controllability analysis on the latter system.

Linearization

We shall now linearize the transformed system (2.2.2) around the target trajectory $(\overline{v},\overline{\ell})$ given in (2.1.2). In order to illustrate some key difficulties related to the controllability to general trajectories for free boundary problems such as (2.1.1), we postpone to Section 2.5.1 the linearization of (2.2.2) around an arbitrary smooth time-dependent trajectory $(\overline{w},\overline{\ell})$, associated to initial and boundary data $(\overline{w}_0,\overline{\ell}_0,\overline{u})$.

To proceed with the linearization around $(\overline{v},\overline{\ell})$, we write $w=\overline{v}+y$ and $\ell=\overline{\ell}+h$, and keep all the terms which are linear with respect to (y,h). The nonlinear problem satisfied by the perturbation variables reads

$$\begin{cases} y_t-ay_{xx}+by_x=\mathcal{N}(y,h) & \text{in } (0,T)\times(0,1)\\ y(t,0)=u(t)-\overline{v},\quad y_x(t,1)=0 & \text{in } (0,T)\\ h'(t)=y(t,1) & \text{in } (0,T)\\ y(0,x)=y_0(x),\quad h(0)=h_0 & \text{in } (0,1). \end{cases} \tag{2.2.3}$$

where $y_0(\cdot)=w_0(\cdot)-\overline{v}$, $h_0=\ell_0-\overline{\ell}(0)$, and the smooth, bounded coefficients are given by

$$a(t)=\frac{1}{\overline{\ell}(t)^2},\qquad b(t,x)=\frac{\overline{v}-\overline{\ell}'(t)x}{\overline{\ell}(t)}\qquad \text{in } [0,T]\times[0,1]. \tag{2.2.4}$$

and the nonlinear term is of the form

$$\mathcal{N}(y,h) = a\big(- h^2 y_t - 2h\bar{\ell} y_t + h' h x y_x + h'\bar{\ell} x y_x + \bar{\ell} h x y_x - h y y_x - h\bar{v} y_x - \bar{\ell} y y_x\big).$$

It is important to note that since $\bar{\ell}'(t) = \bar{v}$, from (2.2.4) it follows that $b(t,1) = 0$. Moreover, the nonlinearity $\mathcal{N}(\cdot,\cdot)$ only consists of (at least) quadratic terms, which will facilitate the application of a Banach fixed point argument. The linearized problem corresponds to (2.2.3) with $\mathcal{N} \equiv 0$.

Remark 2.2.1. *At this point we notice that the linearized problem, namely (2.2.3) with $\mathcal{N} \equiv 0$, the PDE component y and the ODE component h are decoupled – namely, y can be solved independently of h, and thus the coupling between the PDE and ODE is done solely through the nonlinear term. As seen in Section 2.5.1, the problem linearized around an arbitrary trajectory, namely (2.5.1), contains the terms dh' and eh, which are non-local as they may be expressed in terms of the Dirichlet trace of y at $x = 1$. As these terms act on a single point in space, at the level of the adjoint problem one could expect to obtain a non-local integral boundary condition over all points in space (see (2.5.3)). See Section 2.5.1 for more details.*

Distributed control problem

Taking the previous transformations into account, Theorem 2.1 would in essence be a consequence of the null-controllability of System (2.2.3). To prove the latter, using common methodology for parabolic equations, we will first consider the distributed control problem

$$\begin{cases} y_t - a y_{xx} + b y_x = \mathcal{N}(y,h) + u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ y(t,-1) = y_x(t,1) = 0 & \text{in } (0,T) \\ h'(t) = y(t,1) & \text{in } (0,T) \\ y(0,x) = y_0(x), \quad h(0) = h_0 & \text{in } (-1,1) \end{cases} \tag{2.2.5}$$

where $\omega \subsetneq (-1,0)$ is an open and non-empty interval. The initial datum $y_0 \in H^1(0,1)$ is also extended to a datum $\widetilde{y}_0$ with $\|\widetilde{y}_0\|_{H^1(-1,1)} \leqslant \|y_0\|_{H^1(0,1)}$. By abuse of notation, we continue denoting the extended initial datum by y_0. Once the null-controllability problem for (2.2.5) is solved, $u(t) := y(t,0) + \overline{v}$ would provide the desired control for Problem (2.2.2), which in view of the previous discussion, also provides a solution to (2.1.1).

To prove the null-controllability for system (2.2.5), we will first consider the associated linear system

$$\begin{cases} y_t - a y_{xx} + b y_x = f + u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ y(t,-1) = y_x(t,1) = 0 & \text{in } (0,T) \\ h'(t) = y(t,1) & \text{in } (0,T) \\ y(0,x) = y_0(x), \quad h(0) = h_0 & \text{in } (-1,1), \end{cases} \tag{2.2.6}$$

where f is a given source term. The null-controllability at time T of the linearized system is the goal of the next section. The nonlinear term appearing in (2.2.5) will be seen as a small perturbation and will be dealt with by means of a Banach fixed-point argument.

2.3 Null-controllability of the linearized system

In this Section, given $T > 0$, arbitrarily large initial data (y_0, ℓ_0), and a source term f with appropriate decay as $t \nearrow T$, we seek a trajectory (y,h) of the linearized problem (2.2.6) satisfying

$$y(T,\cdot) = 0 \quad \text{in } (-1,1) \quad \text{and} \quad h(T) = 0.$$

In (2.2.6) we are dealing with a cascade-like system, as knowing y immediately yields h, with the latter being reduced to the integrator

$$h(t) = h_0 + \int_0^t y(\tau, 1)\,\mathrm{d}\tau.$$

In other words, the null-controllability of (2.2.6), would follow from solving the linear control problem (recall that $a(t) > 0$ and $b(t, 1) = 0$)

$$\begin{cases} y_t - ay_{xx} + by_x = f + u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ y(t,-1) = y_x(t,1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1) \\ y(T,x) = 0 & \text{in } (-1,1) \end{cases} \tag{2.3.1}$$

subject to the linear finite-dimensional constraint

$$h_0 + \int_0^T y(\tau, 1)\,\mathrm{d}\tau = 0. \tag{2.3.2}$$

We will see this as a constrained controllability problem, namely with a linear finite-dimensional constraint on the control u.

Carleman weights

Let us recall that $\omega = (\gamma_1, \gamma_2) \subsetneq (-1, 0)$. We take (a_0, b_0) with $\gamma_1 < a_0 < b_0 < \gamma_2$ and introduce a function $\alpha_0 \in C^2([-1, 1])$ such that

$$\alpha_0(x) > 0 \quad \text{in } (-1,1), \quad \alpha_0(\pm 1) = 0, \quad |\alpha_{0,x}| > 0 \quad \text{in } (-1,1)\setminus(a_0,b_0),$$

and for $\lambda \geqslant 1$ consider the function α defined by

$$\alpha(t,x) = \theta(t)\left(e^{2\lambda\|\alpha_0\|_{L^\infty}} - e^{\lambda\alpha_0(x)}\right), \quad \text{in } (0,T)\times(-1,1), \tag{2.3.3}$$

where $\theta \in C^2([0, T))$ is given by

$$\theta(t) = \begin{cases} \dfrac{4}{T^2} & \text{on } \left[0, \dfrac{T}{2}\right] \\ \dfrac{1}{t(T-t)} & \text{on } \left[\dfrac{T}{2}, T\right). \end{cases}$$

Notice that the weight $\theta(t)$ does not blow up as $t \searrow 0$. This is because in view of the fixed-point argument, we will need to work with source-terms which do not vanish at $t = 0$.

The main goal of this section is to prove the following result.

Theorem 2.2. *Let $T > 0$ be given. There exists $s \geqslant 1$ such that for any data $y_0 \in L^2(-1,1)$, $h_0 \in \mathbb{R}$ and $f \in L^2(0,T;L^2(-1,1))$ with*

$$\int_0^T \int_{-1}^1 \theta^{-3} e^{2s\alpha} |f|^2 \,\mathrm{d}x\,\mathrm{d}t < \infty, \tag{2.3.4}$$

there exists a control $u \in L^2(0,T;L^2(\omega))$ such that the associated solution

$$y \in L^2(0,T;H^1(-1,1)) \cap C^0([0,T];L^2(-1,1)) \quad \text{and} \quad h \in H^1(0,T)$$

of Problem (2.2.6) satisfies $y(T,\cdot) = 0$ and $h(T) = 0$. Moreover,

$$\begin{aligned} &\|u\|_{L^2(0,T;L^2(\omega))} + \left\|e^{s\alpha}y\right\|_{L^2(0,T;L^2(-1,1))} \\ &\qquad \leqslant C\left(\|y_0\|_{L^2(-1,1)} + |h_0| + \left\|\theta^{-3/2}e^{s\alpha}f\right\|_{L^2(0,T;L^2(-1,1))}\right) \end{aligned}$$

holds for some $C = C(T,\omega,s) > 0$.

It is well-known that a Carleman inequality (see Lemma 2.3.1) along with the HUM method yield the null-controllability of the linear heat equation (2.3.1) with a source term f as in (2.3.4).

To control the second component h to zero at time T, we will reformulate the constraint (2.3.2) by introducing an augmented adjoint problem for the heat equation with a non-homogeneous boundary condition at $x = 1$. The requirement $h(T) = 0$ may then be achieved by adding a corrector term to the HUM control for the heat equation. To guarantee the existence of this control by means of the HUM method, we will need to prove an improved observability inequality. This idea appears in the work of Nakoulima [213] (see also [91] for a recent generalization), and has been applied in works on fluid-structure interaction problems (see [30, 100] for instance) where the structure's displacement at time T is deduced after having controlled the fluid and structure velocities.

2.3.1 An improved observability inequality

We will make use of the following Carleman inequality for solutions to (recall that $b(t,1) = 0$) the adjoint heat equation

$$\begin{cases} -\zeta_t - a\zeta_{xx} - (b\zeta)_x = g & \text{in } (0,T)\times(-1,1) \\ \zeta(t,-1) = \zeta_x(t,1) = 0 & \text{in } (0,T) \\ \zeta(T,x) = \zeta_T(x) & \text{in } (-1,1), \end{cases} \tag{2.3.5}$$

and the weights defined in (2.3.3). The proof follows by combining the well-known Carleman inequality shown in Fursikov & Imanuvilov [113, Lemma 1] (see also [279]) with the parameters $s \geqslant s_0 \geqslant 1$ and $\lambda \geqslant \lambda_0 \geqslant 1$ appearing therein being henceforth fixed, and energy estimates (recall that $b(t,1) = 0$) as done in [101, Section 3].

Lemma 2.3.1. *Let $T > 0$. There exists $C = C(T,\omega,s,\lambda) > 0$ such that for every datum $\zeta_T \in L^2(-1,1)$ and source $g \in L^2(0,T;L^2(-1,1))$, the unique weak solution $\zeta \in L^2(0,T;H^1(-1,1)) \cap C^0([0,T];L^2(-1,1))$ to (2.3.5) satisfies*

$$\begin{aligned} &\int_0^T \int_{-1}^1 \theta^3 e^{-2s\alpha} |\zeta|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_{-1}^1 |\zeta(0,x)|^2 \,\mathrm{d}x \\ &\qquad \leqslant C\left(\int_0^T \int_{-1}^1 e^{-2s\alpha} |g|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_0^T \int_\omega \theta^3 e^{-2s\alpha} |\zeta|^2 \,\mathrm{d}x\,\mathrm{d}t \right). \end{aligned} \tag{2.3.6}$$

The Carleman inequality (2.3.6) guarantees the coercivity and continuity of the strictly convex HUM functional, the unique minimizer of which yields a solution to the adjoint heat equation (2.3.5) and subsequently a solution to the control problem (2.3.1) after investigating the corresponding Euler-Lagrange equation.

To take care of the constraint $h(T) = 0$, let us consider the augmented adjoint problem

$$\begin{cases} -\psi_t - a\psi_{xx} - (b\psi)_x = 0 & \text{in } (0,T)\times(-1,1) \\ \psi(t,-1) = 0, \quad \psi_x(t,1) = 1 & \text{in } (0,T) \\ \psi(T,x) = 0 & \text{in } (-1,1). \end{cases} \tag{2.3.7}$$

Multiplying the heat equation appearing in System (2.2.6) by the unique weak solution $\psi \in L^2(0,T;H^1(-1,1)) \cap C^0([0,T];L^2(-1,1))$ of (2.3.7) and integrating, we see that due to (2.3.2), a control u is such that the corresponding solution of (2.2.6) satisfies $h(T) = 0$ if and only if

$$\int_0^T \int_\omega u\psi \,\mathrm{d}x\,\mathrm{d}t = -\int_{-1}^1 y_0(x)\psi(0,x)\,\mathrm{d}x + h_0 - \int_0^T \int_{-1}^1 f\psi \,\mathrm{d}x\,\mathrm{d}t. \tag{2.3.8}$$

Let us define the projector

$$\mathbb{P}(\zeta) := \frac{\displaystyle\int_{(0,T)\times\omega} \psi\zeta \,\mathrm{d}x\,\mathrm{d}t}{\displaystyle\int_{(0,T)\times\omega} |\psi|^2 \,\mathrm{d}x\,\mathrm{d}t} \qquad \text{for all } \zeta \in L^2(0,T;L^2(-1,1)).$$

The key property of the operator $\mathbb{P}(\cdot)$ is its finite-dimensional range (in fact, one-dimensional range). Our next result is the desired improved observability inequality. The proof follows an indirect, compactness-uniqueness argument (following ideas in [30, 100]). We assume the setting of Lemma 2.3.1.

Proposition 2.3.2. *There exists a constant $C_{\mathrm{obs}} = C_{\mathrm{obs}}(T,\omega,s,\lambda) > 0$ such that for every datum $\zeta_T \in L^2(-1,1)$ and source $g \in L^2(0,T;L^2(-1,1))$, the unique weak solution $\zeta \in L^2(0,T;H^1(-1,1)) \cap C^0([0,T];L^2(-1,1))$ to (2.3.5) satisfies*

$$\int_0^T \int_{-1}^1 \theta^3 e^{-2s\alpha} |\zeta|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_{-1}^1 |\zeta(0,x)|^2 \,\mathrm{d}x + |\mathbb{P}(\zeta)|^2 \tag{2.3.9}$$
$$\leqslant C_{\mathrm{obs}} \left(\int_0^T \int_{-1}^1 e^{-2s\alpha} |g|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_0^T \int_\omega |\zeta - \mathbb{P}(\zeta)\psi|^2 \,\mathrm{d}x\,\mathrm{d}t \right).$$

Proof. We will begin by showing by means of an indirect argument that

$$\int_0^T \int_{-1}^1 \theta^3 e^{-2s\alpha} |\zeta|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_{-1}^1 |\zeta(0,x)|^2 \,\mathrm{d}x$$
$$\leqslant C_2 \left(\int_0^T \int_{-1}^1 e^{-2s\alpha} |g|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_0^T \int_\omega |\zeta - \mathbb{P}(\zeta)\psi|^2 \,\mathrm{d}x\,\mathrm{d}t \right) \tag{2.3.10}$$

for some $C_2 = C_2(T,\omega,s,\lambda) > 0$ and any (ζ_T, g) as in the statement, which would cover the two leftmost terms of the desired inequality (2.3.9). To do so, let us assume by contradiction that (2.3.10) is false, thus there exist two sequences $\{\zeta_T^k\}_{k=1}^\infty$ and $\{g^k\}_{k=1}^\infty$ such that

$$1 = \int_0^T \int_{-1}^1 \theta^3 e^{-2s\alpha} \left|\zeta^k\right|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_{-1}^1 |\zeta^k(0,\cdot)|^2 \,\mathrm{d}x$$
$$\geqslant k \left(\int_0^T \int_{-1}^1 e^{-2s\alpha} \left|g^k\right|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_0^T \int_\omega \left|\zeta^k - \mathbb{P}(\zeta^k)\psi\right|^2 \,\mathrm{d}x\,\mathrm{d}t \right), \tag{2.3.11}$$

for any $k \in \mathbb{N}$, with ζ^k being the corresponding solution to the adjoint problem (2.3.5). Elementary inequalities give

$$\frac{1}{2} \int_0^T \int_\omega \theta^3 e^{-2s\alpha} \left|\mathbb{P}(\zeta^k)\psi\right|^2 \,\mathrm{d}x\,\mathrm{d}t$$
$$\leqslant \int_0^T \int_\omega \theta^3 e^{-2s\alpha} \left|\zeta^k\right|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_0^T \int_\omega \theta^3 e^{-2s\alpha} \left|\zeta^k - \mathbb{P}(\zeta^k)\psi\right|^2 \,\mathrm{d}x\,\mathrm{d}t,$$

thus the left-most integral is uniformly bounded for any $k \in \mathbb{N}$ in view of (2.3.11) (recall also the definition of the weights in (2.3.3)). Hence, $\mathbb{P}(\zeta^k)$ is uniformly bounded in $\mathbb{R}$ with respect to $k \in \mathbb{N}$, whence it follows that

$$\mathbb{P}(\zeta^k) \longrightarrow \mathbb{P}_* \qquad \text{as } k \longrightarrow +\infty \tag{2.3.12}$$

for some $\mathbb{P}_* \in \mathbb{R}$, along some subsequence. From (2.3.11), we see that the functions ζ^k and $\zeta^k(0,\cdot)$ are uniformly bounded in $L^2(0,T-\varepsilon;L^2(-1,1))$ and $L^2(-1,1)$ respectively, for all $\varepsilon > 0$, as well as

$$\int_0^{T-\varepsilon} \int_{-1}^1 \left|g^k\right|^2 \,\mathrm{d}x\,\mathrm{d}t \lesssim \frac{1}{k}.$$

Whence, using the well-known energy estimates for the heat equation (recall that $b(t,1) = 0$), one also has that

$$\begin{aligned} \zeta^k &\rightharpoonup \zeta && \text{weakly in } L^2(0,T-\varepsilon;H^1(-1,1)) \\ \zeta^k_t &\rightharpoonup \zeta_t && \text{weakly in } L^2(0,T-\varepsilon;H^{-1}(-1,1)) \end{aligned}$$

along subsequences as $k \longrightarrow +\infty$. It can thus be seen that ζ satisfies

$$\begin{cases} -\zeta_t - a\zeta_{xx} - (b\zeta)_x = 0 & \text{in } (0,T)\times(-1,1) \\ \zeta(t,-1) = 0, \quad \zeta_x(t,1) = 0 & \text{in } (0,T). \end{cases}$$

In $(0,T)\times\omega$, we have $\zeta^k = (\zeta^k - \mathbb{P}(\zeta^k)\psi) + \mathbb{P}(\zeta^k)\psi$, so in view of (2.3.11) and (2.3.12) we have

$$\zeta^k \longrightarrow \mathbb{P}_*\psi \qquad \text{strongly in } L^2(0,T;L^2(\omega)) \tag{2.3.13}$$

as $k \longrightarrow +\infty$. The above convergence implies that $\zeta = \mathbb{P}_*\psi$ in $(0,T)\times\omega$. As ψ is also in the kernel of the heat operator (thus, so is $\mathbb{P}_*\psi$), by unique continuation we deduce that $\zeta = \mathbb{P}_*\psi$ in $(0,T)\times(-1,1)$. But this can only hold if $\zeta \equiv 0$ and $\mathbb{P}_* = 0$, since $\psi_x(t,1) = 1$.

From (2.3.13), we may thus deduce that

$$\zeta^k \longrightarrow 0 \qquad \text{strongly in } L^2(0,T;L^2(\omega))$$

as $k \longrightarrow +\infty$, and using (2.3.6) (noting that (2.3.11) is used for g^k) we deduce that

$$\int_0^T \int_{-1}^1 \theta^3 e^{-2s\alpha} \left|\zeta^k\right|^2 \mathrm{d}x\,\mathrm{d}t + \int_{-1}^1 \left|\zeta^k(0,x)\right|^2 \mathrm{d}x \longrightarrow 0$$

as $k \longrightarrow +\infty$, which contradicts (2.3.11). Consequently, (2.3.10) holds. Arguing as for (2.3.10), we can show

$$\left|\int_0^T \int_\omega \theta^3 e^{-2s\alpha}\zeta\psi\,\mathrm{d}x\,\mathrm{d}t\right|^2 \leqslant C_5\left(\int_0^T\int_{-1}^1 e^{-2s\alpha}|g|^2\,\mathrm{d}x\,\mathrm{d}t + \int_0^T\int_\omega |\zeta - \mathbb{P}(\zeta)\psi|^2\,\mathrm{d}x\,\mathrm{d}t\right) \tag{2.3.14}$$

for some $C_5 = C_5(T,\omega,s) > 0$. Indeed, setting up an assumption for (2.3.14) as in (2.3.11) and applying Cauchy-Schwarz, after following the lines of the previous step, it may be seen that this would provide the necessary contradiction. □

Remark 2.3.3. *While Proposition 2.3.2 yields the desired improved observability inequality for what follows, due to the indirect argument used for the proof an explicit dependence of the newly obtained constant on the parameters (T,ω) is not guaranteed.*

2.3.2 Proof of Theorem 2.2

We are now in a position to complete the proof of Theorem 2.2, which follows by adapting the well-known HUM arguments.

Proof of Theorem 2.2. For a solution ψ of (2.3.7), let us henceforth denote

$$M_0 := -\int_{-1}^1 y_0(x)\psi(0,\cdot)\,\mathrm{d}x + h_0 - \int_0^T\int_{-1}^1 f\psi\,\mathrm{d}x\,\mathrm{d}t. \tag{2.3.15}$$

We split the proof in three steps.

Step 1: *Minimization problem.* Consider the functional

$$\begin{aligned} J_{\mathrm{obs}}(\zeta_T,g) :=& \frac{1}{2}\int_0^T\int_\omega |\zeta - \mathbb{P}(\zeta)\psi|^2\,\mathrm{d}x\,\mathrm{d}t + \frac{1}{2}\int_0^T\int_{-1}^1 e^{-2s\alpha}|g|^2\,\mathrm{d}x\,\mathrm{d}t \\ &- \int_0^T\int_{-1}^1 f\zeta\,\mathrm{d}x\,\mathrm{d}t - \int_{-1}^1 y_0(x)\zeta(0,x)\,\mathrm{d}x - \mathbb{P}(\zeta)M_0, \end{aligned}$$

initially defined for $(\zeta_T, g) \in L^2(-1,1) \times L^2(0,T;L^2(-1,1))$ with corresponding solution $\zeta \in L^2(0,T;H^1(-1,1)) \cap C^0([0,T];L^2(-1,1))$ to the adjoint heat equation (2.3.5), and ψ being the solution to the augmented adjoint problem (2.3.7). We will show the existence of a minimizer to J_{obs}, which will consequently be used to build the desired control – state pair for Problem (2.2.6).

We remark that the quantity

$$\|(\zeta_T, g)\|_{\text{obs}}^2 = \int_0^T \int_\omega |\zeta - \mathbb{P}(\zeta)\psi|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_0^T \int_{-1}^1 e^{-2s\alpha}|g|^2 \,\mathrm{d}x\,\mathrm{d}t$$

defines a norm on $L^2(-1,1) \times L^2(0,T;L^2(-1,1))$. In order to have completeness, we thus introduce the space

$$X_{\text{obs}} := \overline{L^2(-1,1) \times L^2(0,T;L^2(-1,1))}^{\|\cdot\|_{\text{obs}}}.$$

The set X_{obs} is then endowed with the Hilbert structure given by the above norm.

On X_{obs}, the functional J_{obs} may be extended by continuity in a unique way. Indeed, the improved weighted observability inequality (2.3.9) implies (recall that f is assumed to satisfy (2.3.4))

$$\begin{aligned}\left|\int_0^T \int_{-1}^1 f\zeta \,\mathrm{d}x\,\mathrm{d}t\right| &\leqslant \left(\int_0^T \int_{-1}^1 \theta^{-3} e^{2s\alpha}|f|^2 \,\mathrm{d}x\,\mathrm{d}t\right)^{1/2} \left(\int_0^T \int_{-1}^1 \theta^3 e^{-2s\alpha}|\zeta|^2 \,\mathrm{d}x\,\mathrm{d}t\right)^{1/2} \\ &\leqslant C\left\|\theta^{-3/2}e^{s\alpha}f\right\|_{L^2(0,T;L^2(-1,1))} \|(\zeta_T, g)\|_{\text{obs}}, \end{aligned} \tag{2.3.16}$$

as well as

$$\begin{aligned}\left|\int_{-1}^1 y_0(x)\zeta(0,x) \,\mathrm{d}x\right| &\leqslant \left(\int_{-1}^1 |y_0|^2 \,\mathrm{d}x\right)^{1/2} \left(\int_{-1}^1 |\zeta(0,x)|^2 \,\mathrm{d}x\right)^{1/2} \\ &\leqslant C\|y_0\|_{L^2(-1,1)} \|(\zeta_T, g)\|_{\text{obs}} \end{aligned} \tag{2.3.17}$$

and

$$|\mathbb{P}(\zeta)| \leqslant C\|(\zeta_T, g)\|_{\text{obs}}. \tag{2.3.18}$$

Due to (2.3.16) – (2.3.17) – (2.3.18), it can be seen that the functional J_{obs} is also coercive. As J_{obs} is also strictly convex on X_{obs} (since $\|\cdot\|_{\text{obs}}$ is a Hilbert norm), it admits a unique minimizer $(\widehat{\zeta}, \widehat{g}) \in X_{\text{obs}}$ by the direct method.

Step 2: *Null-controllability requirements.* Now the unique minimizer $\left(\widehat{\zeta_T}, \widehat{g}\right) \in X_{\text{obs}}$ of J_{obs} satisfies the Euler-Lagrange equation

$$\begin{aligned} 0 = &\int_0^T \int_\omega \left(\widehat{\zeta} - \mathbb{P}(\widehat{\zeta})\,\psi\right)\varphi \,\mathrm{d}x\,\mathrm{d}t + \int_0^T \int_{-1}^1 e^{-2s\alpha}\widehat{g}F \,\mathrm{d}x\,\mathrm{d}t \\ &- \int_0^T \int_{-1}^1 f\varphi \,\mathrm{d}x\,\mathrm{d}t - \int_{-1}^1 y_0(x)\varphi(0,x) \,\mathrm{d}x - \mathbb{P}(\varphi)M_0 \end{aligned} \tag{2.3.19}$$

for all $(\varphi_T, F) \in X_{\text{obs}}$, where $\widehat{\zeta}$ and φ denote the solutions to (2.3.5) corresponding to $\left(\widehat{\zeta_T}, \widehat{g}\right)$ and (φ_T, F) respectively. Comparing (2.3.19) with (4.1.5), we are led to consider the control function

$$u := -\left(\widehat{\zeta} - \mathbb{P}(\widehat{\zeta})\psi\right) + M_0 \left(\int_0^T \int_\omega \psi^2 \,\mathrm{d}x\,\mathrm{d}t\right)^{-1} \psi$$

restricted to ω, where ψ is the unique solution to the augmented adjoint problem (2.3.7). Let $y \in L^2(0,T;H^1(-1,1)) \cap C^0([0,T];L^2(-1,1))$ be the solution to the heat equation in (2.2.6) with control u. Let us justify this choice. Noting that

$$\int_0^T \int_\omega u\varphi \,\mathrm{d}x\,\mathrm{d}t = -\int_0^T \int_\omega \left(\widehat{\zeta} - \mathbb{P}(\widehat{\zeta})\,\psi\right)\varphi \,\mathrm{d}x\,\mathrm{d}t + \mathbb{P}(\varphi)M_0,$$

we come back to (2.3.19) and deduce that

$$\begin{aligned} 0 = &-\int_0^T \int_{-1}^1 e^{2s\alpha}\widehat{g}F \,\mathrm{d}x\,\mathrm{d}t + \int_0^T \int_\omega u\varphi \,\mathrm{d}x\,\mathrm{d}t \\ &+ \int_0^T \int_{-1}^1 f\varphi \,\mathrm{d}x\,\mathrm{d}t + \int_{-1}^1 y_0\varphi(0,\cdot)\,\mathrm{d}x. \end{aligned} \tag{2.3.20}$$

On the other hand, multiplying the heat component in (2.2.6) by any φ weak solution of (2.3.5) with initial data φ_T and source term F, we see (modulo a density argument) that

$$\int_{-1}^1 y(T,\cdot)\varphi_T \,\mathrm{d}x = -\int_0^T \int_{-1}^1 (yF + f\varphi)\,\mathrm{d}x\,\mathrm{d}t + \int_{-1}^1 y_0\varphi(0,\cdot)\,\mathrm{d}x + \int_0^T \int_\omega u\varphi \,\mathrm{d}x\,\mathrm{d}t. \tag{2.3.21}$$

Comparing with (2.3.20), for all $(\varphi_T, F) \in L^2(-1,1) \times L^2(0,T;L^2(-1,1))$ one has

$$\int_{-1}^1 y(T,\cdot)\varphi_T \,\mathrm{d}x = \int_0^T \int_{-1}^1 \left(e^{2s\alpha}\widehat{g} - y\right) F \,\mathrm{d}x\,\mathrm{d}t.$$

As F is arbitrary, choosing $F \equiv 0$, we get the desired control requirement $y(T,\cdot) = 0$. On the other hand, as φ_T is arbitrary, choosing $\varphi_T \equiv 0$, we see also that

$$y = \widehat{g}e^{-2s\alpha}.$$

We now define $h \in H^1(0,T)$ by

$$h(t) := h_0 + \int_0^t y(\tau,1)\,\mathrm{d}\tau.$$

It remains to be seen that the above-defined control u is such that $h(T) = 0$. Recalling the definition of M_0 in (2.3.15), a straightforward computation shows that

$$\int_0^T \int_\omega u\psi \,\mathrm{d}x\,\mathrm{d}t = M_0,$$

which in view of (2.3.8) yields the conclusion $h(T) = 0$, as desired.

Step 3: *Estimates.* As $J_{\mathrm{obs}}\left(\widehat{\zeta}_T, \widehat{g}\right) \leqslant J_{\mathrm{obs}}(0,0) = 0$, straightforward estimates along with (2.3.16) – (2.3.18) give

$$\begin{aligned} &\left\|\widehat{\zeta} - \mathbb{P}(\widehat{\zeta})\,\psi\right\|_{L^2(0,T;L^2(\omega))} + \left\|e^{-s\alpha}\widehat{g}\right\|_{L^2(0,T;L^2(-1,1))} \\ &\qquad \leqslant C_1\left(\|y_0\|_{L^2(-1,1)} + |h_0| + \left\|\theta^{-3/2}e^{s\alpha}f\right\|_{L^2(0,T;L^2(-1,1))}\right) \end{aligned} \tag{2.3.22}$$

for some $C_1 > 0$. On another hand, it may easily be checked that

$$\int_0^T \int_\omega u^2 \,\mathrm{d}x\,\mathrm{d}t = \int_0^T \int_\omega \left(\widehat{\zeta} - \mathbb{P}(\widehat{\zeta})\psi\right)^2 \,\mathrm{d}x\,\mathrm{d}t + M_0^2\left(\int_0^T \int_\omega \psi^2 \,\mathrm{d}x\,\mathrm{d}t\right)^{-1} \tag{2.3.23}$$

Thus, in view of the definitions of the control u and the state y and (2.3.22) and (2.3.23) lead us to conclude that

$$\begin{aligned}\|u\|_{L^2(0,T;L^2(\omega))} + \|e^{s\alpha}y\|_{L^2(0,T;L^2(-1,1))} \\ \leqslant C_2\left(\|y_0\|_{L^2(-1,1)} + |h_0| + \left\|\theta^{-3/2}e^{s\alpha}f\right\|_{L^2(0,T;L^2(-1,1))}\right)\end{aligned}$$

for some $C_2 > 0$. This concludes the proof. □

The following Lemma gives additional estimates of the controlled trajectory in the weighted spaces provided more regular initial data.

Lemma 2.3.4. *Let (v, y, h) denote the control-state pair given by Theorem 2.2. Assume moreover that $y_0 \in H^1(-1,1)$. Then*

$$\begin{aligned}\left\|\theta^{-1}e^{s\alpha}y_x\right\|_{L^2(0,T;L^2(-1,1))} + \left\|\theta^{-2}e^{s\alpha}y_t\right\|_{L^2(0,T;L^2(-1,1))} \\ + \left\|\theta^{-2}e^{s\alpha}y_{xx}\right\|_{L^2(0,T;L^2(-1,1))} + \left\|\theta^{-2}e^{s\alpha}y\right\|_{L^\infty(0,T;H^1(-1,1))} \\ \leqslant C\left(\|y_0\|_{H^1(-1,1)} + |h_0| + \left\|\theta^{-3/2}e^{s\alpha}f\right\|_{L^2(0,T;L^2(-1,1))}\right)\end{aligned}$$

holds for some $C = C(T, \omega, s) > 0$.

Proof. The proof for estimating the first three norms follows standard energy estimate arguments, and we refer to [102, Lemma 3.4] for details. To obtain the weighted $L^\infty(H^1)$-estimate, we note that by interpolation

$$\left\|\theta^{-2}e^{s\alpha}y\right\|_{L^\infty(0,T;H^1(-1,1))} \lesssim \left\|\theta^{-2}e^{s\alpha}y\right\|^{1/2}_{L^2(0,T;H^2(-1,1))}\left\|\theta^{-2}e^{s\alpha}y\right\|^{1/2}_{H^1(0,T;L^2(-1,1))},$$

and the right-hand side is bounded by the properties of the Carleman weights and the three previous estimates. □

2.4 The nonlinear problem

We now look to conclude the proof of Theorem 2.1 by virtue of a fixed-point argument for nonlinear system

$$\begin{cases} y_t - ay_{xx} + by_x = \mathcal{N}(y,h) + u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ y(t,-1) = y_x(t,1) = 0 & \text{in } (0,T) \\ h'(t) = y(t,1) & \text{in } (0,T) \\ y(0,x) = y_0(x), \quad h(0) = h_0 & \text{in } (-1,1), \end{cases} \tag{2.4.1}$$

a restriction argument and reverting the transformations performed in Section 2.2. We recall that the nonlinear term in (2.4.1) is of the form

$$\mathcal{N}(y,h) = a\big(-hy_t(h+2\overline{\ell}) + h'y_x(hx+\overline{\ell}x) + hy_x(\overline{\ell}x+\overline{v}) - yy_x(h+\overline{\ell})\big), \tag{2.4.2}$$

only consisting of (at least) quadratic terms.

Let us consider the norm

$$\begin{aligned}\|y\|_{\mathcal{Y}} := \|e^{s\alpha}y\|_{L^2(0,T;L^2(-1,1))} + \left\|\theta^{-1}e^{s\alpha}y_x\right\|_{L^2(0,T;L^2(-1,1))} \\ + \left\|\theta^{-2}e^{s\alpha}y_t\right\|_{L^2(0,T;L^2(-1,1))} + \left\|\theta^{-2}e^{s\alpha}y_{xx}\right\|_{L^2(0,T;L^2(-1,1))} \\ + \left\|\theta^{-2}e^{s\alpha}y\right\|_{L^\infty(0,T;H^1(-1,1))}.\end{aligned}$$

We begin by the following lemma, which provides the appropriate estimates of each nonlinear term with respect to the $\|\cdot\|_{\mathcal{Y}}$ – norm.

Lemma 2.4.1 (Nonlinear estimates)**.** *For $y_0 \in H^1(-1,1)$, let (y,h) denote the controlled trajectory of the linearized problem (2.2.6) given by Theorem 2.2. Then*

$$\left\|\theta^{-3/2}e^{s\alpha}\mathcal{N}(y,h)\right\|_{L^2(0,T;L^2(-1,1))} \leqslant C\|y\|_{\mathcal{Y}}^2$$

holds for some $C = C(T,\omega,s) > 0$.

Proof. We begin by noting that $a \in L^\infty(0,T)$. Using interpolation estimates,

$$\|y\|_{L^\infty(L^\infty)} \leqslant \|y\|_{L^\infty(H^1)} \leqslant C\|y\|_{H^1(L^2)}^{1/2}\|y\|_{L^2(H^2)}^{1/2} \leqslant C\|y\|_{\mathcal{Y}}. \tag{2.4.3}$$

Let us begin by estimating the right-most term of (2.4.2). Since $h+\overline{\ell} \in L^\infty(0,T)$ as well as $\theta^{-1} \in L^\infty(0,T)$, using (2.4.3) one deduces

$$\begin{aligned}\left\|\theta^{-3/2}e^{s\alpha}(h+\overline{\ell})yy_x\right\|_{L^2(0,T;L^2(-1,1))} &\leqslant C\|y\|_{L^\infty(L^\infty)}\left\|\theta^{-3/2}e^{s\alpha}y_x\right\|_{L^2(0,T;L^2(-1,1))}\\ &\leqslant C\|y\|_{\mathcal{Y}}^2.\end{aligned} \tag{2.4.4}$$

To estimate the two middle terms in (2.4.2), we first observe that since $h(T) = 0$, for any $t \in [0,T]$ we may write

$$h(t) = h(t) - h(T) \leqslant C(T)\sup_{t\in[0,T]}|h'(t)|. \tag{2.4.5}$$

Moreover, as $h'(t) = y(t,1)$ for $t \in (0,T)$, $(h+\overline{\ell})\cdot \in L^\infty((0,T)\times(-1,1))$ and $\overline{\ell}\cdot + \overline{v} \in L^\infty((0,T)\times(-1,1))$ and $\theta^{-1} \in L^\infty(0,T)$, we may estimate the middle terms using (2.4.5) and (2.4.3) as follows:

$$\begin{aligned}\left\|\theta^{-3/2}e^{s\alpha}h'y_x(h+\overline{\ell})\right\|_{L^2(0,T;L^2(-1,1))} &+ \left\|\theta^{-3/2}e^{s\alpha}hy_x(\overline{\ell}+\overline{v})\right\|_{L^2(0,T;L^2(-1,1))}\\ &\leqslant C\|y\|_{L^\infty(L^\infty)}\left\|e^{-3/2}e^{s\alpha}y_x\right\|_{L^2(0,T;L^2(-1,1))}\\ &\leqslant C\|y\|_{\mathcal{Y}}^2.\end{aligned} \tag{2.4.6}$$

To estimate the leftmost term, we need further arguments. Indeed, arguing as above we deduce

$$\left\|\theta^{-3/2}e^{s\alpha}hy_t(h+2\overline{\ell})\right\|_{L^2(0,T;L^2(-1,1))} \leqslant C\left\|\theta^{1/2}h\right\|_{L^\infty(0,T)}\left\|\theta^{-2}e^{s\alpha}y_t\right\|_{L^2(0,T;L^2(-1,1))}.$$

The desired estimate would thus follow provided

$$\left\|\theta^{1/2}h\right\|_{L^\infty(0,T)} \lesssim \|y\|_{\mathcal{Y}} \tag{2.4.7}$$

holds. To prove (2.4.7), let $0 < \overline{\alpha} < \min_{x\in(-1,1)}(e^{2\lambda\|\alpha_0\|_{L^\infty}} - e^{\lambda\alpha_0})$ and we first notice that since $h(T) = 0$ and $e^{-\frac{s\overline{\alpha}\theta(T)}{2}} = 0$, by the Cauchy mean-value theorem

$$\left|\frac{h(t)}{e^{-\frac{s\overline{\alpha}\theta}{2}}}\right| = \left|\frac{h(t)-h(T)}{e^{-\frac{s\overline{\alpha}\theta(t)}{2}} - e^{-\frac{s\overline{\alpha}\theta(T)}{2}}}\right| \lesssim \left\|\frac{h'}{\left(e^{-\frac{s\overline{\alpha}\theta}{2}}\right)'}\right\|_{L^\infty(0,T)} \lesssim_T \left\|\frac{h'}{e^{-\frac{s\overline{\alpha}\theta}{2}}}\right\|_{L^\infty(0,T)} \tag{2.4.8}$$

for $t \in [0,T]$. We proceed in estimating the right-most term in (2.4.8). For $t \in [0,T]$, using trace estimates and the decay properties of the Carleman weights,

$$\begin{aligned}e^{s\overline{\alpha}\theta(t)}|h'(t)|^2 &= e^{s\overline{\alpha}\theta(t)}|y(t,1)|^2\\ &\lesssim \sup_{t\in[0,T]}\int_{-1}^1 e^{s\overline{\alpha}\theta(t)}|y(t,x)|^2\,\mathrm{d}x + \sup_{t\in[0,T]}\int_{-1}^1 e^{s\overline{\alpha}\theta(t)}|y_x(t,x)|^2\,\mathrm{d}x\\ &\lesssim_T \sup_{t\in[0,T]}\int_{-1}^1 \theta^{-4}e^{2s\alpha}|y|^2\,\mathrm{d}x + \sup_{t\in[0,T]}\int_{-1}^1 \theta^{-4}e^{2s\alpha}|y_x|^2\,\mathrm{d}x,\end{aligned} \tag{2.4.9}$$

and the right-most terms are bounded by Lemma 2.3.4. By (2.4.9), (2.4.8) holds, and the latter rewrites as

$$|h(t)| \lesssim_T e^{-\frac{s\overline{\alpha}\theta(t)}{2}} \left\| \frac{h'}{e^{-\frac{s\overline{\alpha}\theta}{2}}} \right\|_{L^\infty(0,T)}. \tag{2.4.10}$$

Consequently, (2.4.10) along with the decay properties of the Carleman weights yield (2.4.7), which concludes the proof. □

We are now in a position to state and prove the null-controllability result for Problem (2.4.1).

Theorem 2.3. *Let $T > 0$ and $\omega = (\gamma_1, \gamma_2) \subsetneq (-1, 0)$ be non-empty. There exists $r > 0$ such that for all $(y_0, h_0) \in H^1(-1,1) \times \mathbb{R}$ satisfying $\|y_0\|_{H^1(-1,1)} + |h_0| \leqslant r$, there exists a control $u \in L^2(0,T; L^2(\omega))$ such that the corresponding strong solution*

$$y \in L^2(0,T; H^2(-1,1)) \cap C^0([0,T]; H^1(-1,1)) \quad h \in H^1(0,T)$$

of (2.4.1) satisfies $y(T,\cdot) = 0$ in $(-1,1)$ and $h(T) = 0$.

The proof follows a Banach fixed point argument. For $r > 0$, we consider the associated ball of $H^1(-1,1)$:

$$\mathfrak{B}_r := \Big\{ y_0 \in H^1(-1,1) \colon \|y_0\|_{H^1(-1,1)} \leqslant r \Big\},$$

and we also set

$$\mathfrak{F}_r = \left\{ f \in L^2(0,T; L^2(-1,1)) \colon \left\| \theta^{-3/2} e^{s\alpha} f \right\|_{L^2(0,T;L^2(-1,1))} \leqslant r \right\}.$$

We construct a map $\mathcal{N} : \mathfrak{B}_r \times (-r,r) \times \mathfrak{F}_r \longrightarrow \mathfrak{F}_r$ by setting, for $y_0 \in \mathfrak{B}_r$, $h_0 \in (-r,r)$ and $f \in \mathfrak{F}_r$,

$$\mathcal{N}(y_0, h_0, f) = \mathcal{N}(y, h),$$

where (y,h) is the controlled trajectory provided by Theorem 2.2.

Proof of Theorem 2.3. We split the proof in 3 steps.

Step 1. For each $y_0 \in \mathfrak{B}_r$ and $h_0 \in (-r,r)$, the application $\mathcal{N}(y_0, h_0, \cdot)$ maps $\mathfrak{F}_r$ to itself whenever $r > 0$ is small enough. Indeed, by Lemma 2.4.1 and Lemma 2.3.4

$$\begin{aligned} \left\| \theta^{-3/2} e^{s\alpha} \mathcal{N}(y_0, h_0, f) \right\|_{L^2(0,T;L^2(-1,1))} &\leqslant C_1 \|y\|_{\mathcal{Y}}^2 \\ &\leqslant C_1 C_2^2 \left(\|y_0\|_{H^1(-1,1)} + |h_0| + \left\| \theta^{-3/2} e^{s\alpha} f \right\|_{L^2(0,T;L^2(-1,1))} \right)^2 \leqslant \frac{r}{2} \end{aligned}$$

whenever $r \leqslant \frac{1}{18 C_1 C_2^2}$ (where $C_1 > 0$ is the constant from Lemma 2.4.1 and $C_2 > 0$ the constant from Lemma 2.3.4).

Step 2. For each $y_0 \in \mathfrak{B}_r$ and $h_0 \in (-r,r)$ with $r > 0$ small enough, the application $\mathcal{N}(y_0, h_0, \cdot)$ is a contraction on $\mathfrak{F}_r$ with a uniform constant < 1. This follows by estimating similarly as in Lemma 2.4.1 and Step 1, and closely follows the estimates in [191].

Step 3. Thanks to the Banach fixed point theorem, given $r > 0$ small enough, for any $y_0 \in \mathfrak{B}_r$ and $h_0 \in (-r,r)$, the application $\mathcal{N}(y_0, h_0, \cdot)$ admits a unique fixed point $f \in \mathfrak{F}_r$, and consequently a unique solution to the control problem for (2.4.1). □

We may thus conclude the proof of Theorem 2.1.

Proof of Theorem 2.1. The result follows by virtue of the transformations performed in Section 2.2 and Theorem 2.3. Indeed, given initial data $(v_0, \ell_0) \in H^1(0,\ell_0) \times \mathbb{R}_+^*$, we consider $y_0(\cdot) := v_0(\ell_0 \cdot) - \overline{v}$ and $h_0 = \ell_0 - \ell_*$. As $y_0 \in H^1(0,1)$, we may extend it to a function $\widetilde{y}_0 \in H^1(-1,1)$, which coincides with y_0 on $(0,1)$. Let $\omega = (\gamma_1, \gamma_2) \subset (-1,0)$ be a non-empty set. By Theorem 2.3, there exists $r > 0$ such that whenever $\|\widetilde{y}_0\|_{H^1(-1,1)} + |h_0| \leqslant r$, there exists a control $\widetilde{u} \in L^2(0,T;L^2(\omega))$ such that the solution (y,h) to (2.4.1) satisfies $y(T,\cdot) = 0$ in $(-1,1)$ and $h(T) = 0$. This in turn implies that the control $u(t) := y(t,0) + \overline{v}$ guarantees the null-controllability of the boundary control system (2.2.3) on $(0,1)$, with initial data (y_0, h_0). We now set $w(t,x) := y(t,x) + \overline{v}$ in $[0,T] \times [0,1]$ and $\ell(t) = h(t) + \overline{\ell}(t)$ in $[0,T]$. It is readily seen that (w,ℓ) satisfy (2.2.2) for initial data $(v(\ell_0 \cdot), \ell_0)$, as well as $w(T,\cdot) = \overline{v}$ in $(0,1)$ and $\ell(T) = \overline{\ell}(T)$. As the result is local, one also has $\ell(t) > 0$ in $[0,T]$ by continuity, and thus reversing the transformation (2.2.1) gives the desired result. □

2.5 Concluding remarks

In this chapter, we addressed the local controllability of both components of the state of a one-dimensional free boundary problem governed by the viscous Burgers equation. By means of a control actuating along the fixed boundary, we showed that we may steer the fluid to constant velocity and also control the position of its free surface, whenever the difference between the initial velocities and the interface positions respectively is small enough. While the existence of this non-trivial trajectory is a particularity of the system under consideration, our result also implies its null-controllability.

We present hereinafter several topics closely related to our work.

2.5.1 Controllability to arbitrary trajectories

A challenging problem to which we *have not* given a solution herein is the controllability to arbitrary smooth trajectories for parabolic free boundary problems. Up to the best of our knowledge, this problem has not been addressed in the literature, even in the one-dimensional case. Let us give a brief overview of the issues that may arise in doing so for system (2.1.1).

We recall that as per Section 2.2, after fixing the domain for (2.1.1), we consider perturbations around a given smooth solution $(\overline{w}, \overline{\ell})$ of (2.2.2) – we write $w = \overline{w} + y$ and $\ell = \overline{\ell} + h$, and keep all the terms which are linear with respect to (y,h). The linearized system reads

$$\begin{cases} y_t - a y_{xx} + b y_x + cy + dh' + eh = 0 & \text{in } (0,T) \times (0,1) \\ y(t,0) = u(t) - \overline{u}(t), \quad y_x(t,1) = 0 & \text{in } (0,T) \\ h'(t) = y(t,1) & \text{in } (0,T) \\ y(0,x) = y_0(x), \quad h(0) = h_0 & \text{in } (0,1), \end{cases} \tag{2.5.1}$$

where $\overline{u}(t) = \overline{w}(t,0)$, $y_0(\cdot) = w_0(\cdot) - \overline{w}(0,\cdot)$, $h_0 = \ell_0 - \overline{\ell}(0)$, with a as in (2.2.4), and the remaining coefficients given by

$$b(t,x) = \frac{\overline{w}(t,x) - \overline{\ell}'(t)x}{\overline{\ell}(t)}, \quad c(t,x) = \frac{\overline{w}_x(t,x)}{\overline{\ell}(t)}$$

$$d(t,x) = -\frac{x\overline{w}_x(t,x)}{\overline{\ell}(t)}, \qquad e(t,x) = \frac{2\overline{w}_t(t,x)}{\overline{\ell}(t)} + \frac{\overline{w}(t,x)\overline{w}_x(t,x) - x\overline{\ell}'(t)\overline{w}_x(t,x)}{\overline{\ell}(t)^2},$$

in $[0,T] \times [0,1]$. We remark that by applying a Banach fixed-point argument to the source term $dh' + eh$, it can be shown that the linearized problem (2.5.1) is well-posed in the energy space $X_T = L^2(0,T;H^1(-1,1)) \cap C^0([0,T];L^2(-1,1))$.

Contrary to the specific case we treated in this paper, there is no reason as to why the factors d, e would vanish for an arbitrary trajectory $(\overline{v}, \overline{\ell})$, so the finite-dimensional constraint techniques presented herein are not applicable. Thus, as done in Section 2.2, let us first consider a distributed control system in the extended domain $(-1, 1)$:

$$\begin{cases} y_t - ay_{xx} + by_x + cy + dh' + eh = u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ y(t,-1) = y_x(t,1) = 0 & \text{in } (0,T) \\ h'(t) = y(t,1) & \text{in } (0,T) \\ y(0,x) = y_0(x), \quad h(0) = h_0 & \text{in } (-1,1), \end{cases} \tag{2.5.2}$$

where the coefficients and initial data are extended accordingly. The localized control $u = u(t,x)$ actuates inside some open, non-empty set $\omega \subsetneq (-1,0)$. Since we consider the case $d, e \not\equiv 0$, the PDE and ODE components remain coupled. Moreover the adjoint problem one obtains is more difficult to handle – multiplying (2.5.2) by a pair of smooth functions (ζ, s) and integrating leads us to

$$\begin{cases} -\zeta_t - a\zeta_{xx} - (b\zeta)_x + c\zeta = 0 & \text{in } (0,T)\times(-1,1) \\ \zeta(t,-1) = 0, \quad \zeta_x(t,1) = -\displaystyle\int_{-1}^{1} d\,\zeta\,\mathrm{d}x + s(t) & \text{in } (0,T) \\ s'(t) = \displaystyle\int_{-1}^{1} d\,\zeta\,\mathrm{d}x & \text{in } (0,T) \\ \zeta(T,x) = \zeta_T(x), \quad s(T) = s_T & \text{in } (-1,1). \end{cases} \tag{2.5.3}$$

The adjoint problem (2.5.3) is much like the forward problem appearing in certain works on population dynamics, see [196] for instance. The authors prove an observability inequality for (2.5.2), which in our case is the forward problem. Up to the best of our knowledge, an observability inequality for (2.5.3) has not been shown in the literature.

Another possible strategy for tackling the null-controllability of (2.5.2) is to "absorb" the nonlocal terms dh' and eh in the source term f. The fact that these terms are linear would raise an issue in proving the invariance of the fixed-point map (Step 1 in Proof of Theorem 2.3). An idea which is used in several papers on the controllability to trajectories for the non-homogeneous Navier-Stokes equations (see [92] and the references therein) is to keep the Carleman constants $s, \lambda \geqslant 1$ arbitrary throughout the proofs. Thus, when proving the fixed-point, one may appeal to these constants as an additional degree of freedom which could render the linear terms small. The main issue in applying this strategy is the compactness-uniqueness method used to prove the improved observability inequality in Proposition 2.3.2. Indeed, the indirect nature of this proof means that the explicit dependence of the new observability constant on the parameters s, λ is a priori unknown. Hence, taking s, λ arbitrarily large a posteriori may not be feasible.

2.5.2 Global results

As discussed in Remark 2.1.1, Theorem 2.1 is a local result, as while the PDE component may possess a dissipative mechanism, the asymptotic position of the free boundary is generally not known for problems of this nature. This is in part due to the lack of conservation properties satisfied by the position of the free boundary ℓ, making its asymptotic position significantly more difficult to determine when compared to similar problems with a stronger coupling and set on the whole line [264, 166]. In fact, by means of some maximum principle argument, it could be possible that the free boundary increases as time grows, which could in turn stipulate an asymptotic behavior of the velocity v to a self-similar profile of the form $\frac{1}{\sqrt{t}} f\left(\frac{x}{\sqrt{t}}\right)$, well known in the context of the viscous Burgers equation set on $\mathbb{R}$ (see e.g. [284]). Thus, even the set of attraction points of trajectories of (2.1.1) is not evident.

It would most certainly be interesting to know whether one may prove a global controllability result in large time. This question is in fact also open in the simpler case of the one-phase Stefan problem (2.1.4), and also in the fluid-structure problem (2.1.5).

2.5.3 Multi-dimensional problem

One may also consider an appropriate controllability problem for the incompressible Navier-Stokes equations with a free surface, as encountered in the works of Beale [19, 20]. This would represent a natural extension of our work to the multi-dimensional setting.

The main difference with the one-dimensional case presented herein and existing works on multi-dimensional fluid-rigid body control (see e.g. [143, 30]) is the fact that the free boundary would be given by the graph of a space-dependent function, whence the second component of the system would be governed by an infinite-dimensional ODE and controlling this component would not represent a finite-dimensional constraint. This is an obvious impediment to the direct application of the techniques presented herein. The null-controllability of the PDE component in the two-dimensional Stefan problem in a radial geometry has been addressed in [77], following the strategy of the one-dimensional counterpart presented in [103]. However, up to the best of our knowledge, the controllability of both components in such a geometrical setting has not been addressed in the literature.

Chapter 3

Perturbed porous-medium gas flow

Abstract. In this work, we investigate the null-controllability of a nonlinear degenerate parabolic equation, which is the equation satisfied by a perturbation around the self-similar solution of the porous medium equation in Lagrangian-like coordinates. We prove a local null-controllability result for a regularized version of the nonlinear problem, in which singular terms have been removed from the nonlinearity. We use spectral techniques and the source-term method to deal with the linearized problem and the conclusion follows by virtue of a Banach fixed-point argument. The spectral techniques are also used to prove a null-controllability result for the linearized thin-film equation, a degenerate fourth order analog of the problem under consideration.

Keywords. Null-controllability, degenerate parabolic equation, porous medium equation, thin-film equation.

AMS Subject Classification. 93B05, 35K65, 93C20, 35R35.

This Chapter is taken from [115]:

Null-controllability of perturbed porous medium gas flow.
B. Geshkovski.
ESAIM: COCV, **26**, 85-105, 2020.
https://doi.org/10.1051/cocv/2020009

Chapter Contents

3.1 Introduction

Due to their relevance in physics and engineering, much attention has been devoted in the scientific literature to fluid systems involving the evolution of a free moving boundary. We refer for example to [212] for models regarding the density of a gas penetrating a solid rock, and to [28, 254] for models on the evolution of thin liquid films in wetting and spreading phenomena. These examples appear in physical and industrial processes such as oil recovery, membranes in biophysics, and spin coating of microchips. Despite occurring in such diverse scientific fields, the mathematical modeling of these mechanisms is quite similar and understanding the control-theoretical aspects thereof is of high importance for applications.

An example of a simplified, applicable model is the *porous medium equation*

$$\partial_t h - \partial_z^2(h^m) = 0 \tag{3.1.1}$$

where $m > 1$. The state $h(t, z)$ may represent the density distribution of a gas flowing in a porous medium, or the height of a thin liquid film deposited onto a solid substrate. By developing the diffusion term, it is readily seen that equation (3.1.1) degenerates when the state h approaches zero. Thus, any solution with compactly supported initial datum retains the compact support in any finite time. In physical terms, the diffusing gas does not reach any point in space instantaneously, but rather propagates with finite speed. This property results in the fact that the porous medium equation is indeed a free boundary problem, the free boundary being given by $\partial\{h > 0\}$. In terms of thin films (see Section 3.5 for the related thin-film equation), it represents the interface separating the liquid, surrounding air and the adjacent solid, as in Figure 3.1.

While the analytical properties of (3.1.1) are well understood (particularly in the one dimensional case, see [268]), the literature on its control-theoretical aspects is rather scarce. In view of the known asymptotic behavior of the free boundary problem for large times (see [268, Chapter 18]) and the desired positivity of the state, a natural question which arises is whether one may control the state $h(t, z)$, as well as its interface, to the self-similar Barenblatt trajectory

$$h_B(t, z) = (t+1)^{-\frac{1}{m+1}} \left(1 - \frac{m-1}{2m(m+1)} \frac{z^2}{(t+1)^{\frac{2}{m+1}}}\right)^{\frac{1}{m-1}} \qquad \text{in } \{h_B > 0\}$$

in a given finite time $T > 0$ by means of an additional forcing control term. To the best of our knowledge, this kind of exact-controllability to trajectories question has not been addressed in the existing literature on the porous medium equation.

An important difficulty when tackling this question is the moving time-dependent support of the solution and the target Barenblatt trajectory. As the two are defined in different domains, perturbations of the form $h_B + y$ around Barenblatt are difficult to define in view of linearizing, a key step in proving controllability. Due to the slightly complex form of the Barenblatt, it is more convenient to look at the equation satisfied by the *pressure*

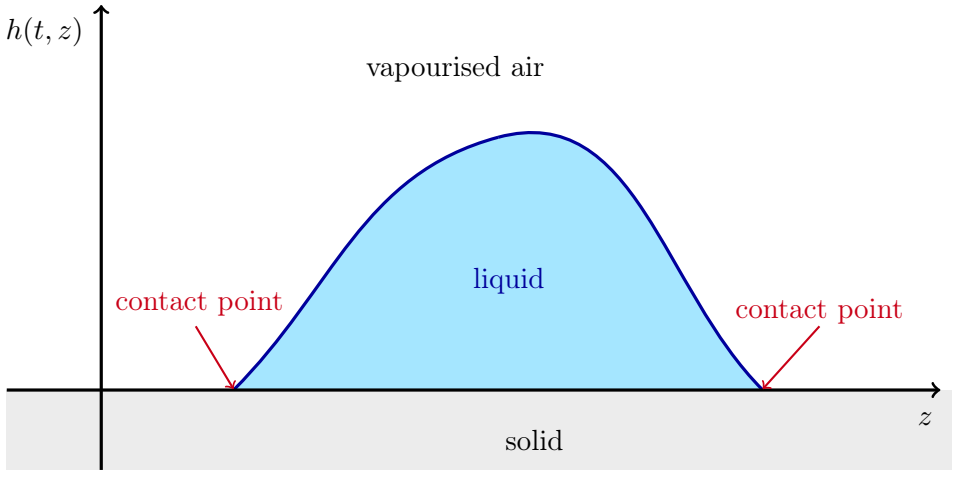


Figure 3.1: The free boundary represents the contact points where the three phases of gas, solid and liquid connect.

$v = \frac{m}{m-1}h^{m-1}$ in self-similar coordinates, namely

$$\begin{cases} \partial_t v - v\partial_z^2 v - (\sigma+1)((\partial_z v)^2 + x\partial_z v) - v = 0 & \text{in } \{v > 0\} \\ v(0,z) = v_0(z) & \text{in } \{v_0 > 0\}, \end{cases} \tag{3.1.2}$$

(see [244, Section 1.2]) where $\sigma = -\frac{m-2}{m-1} > -1$. In this case, the Barenblatt solution is stationary and supported in the unit interval:

$$\rho(z) = \frac{1}{2}\Big(1 - z^2\Big) \quad \text{for } z \in (-1,1). \tag{3.1.3}$$

The motivation behind our work is thus to know if one can steer the state $v(t,z)$ and its interface to the stationary Barenblatt solution $\rho(z)$ in a given time $T > 0$, by means of an additional forcing control term in the equation.

To overcome the difficulty of the moving domain, a Lagrangian-like change of variables (thus depending on the solution, and called *von Mises transformation*) may be applied, mapping the moving support of the solution onto the support of the Barenblatt profile, now the interval $(-1,1)$. The change of coordinates depends on the solution (and thus its smallness and regularity), and in these new variables the Barenblatt reduces to the constant 1. Since the transformed solution and Barenblatt are defined in the same fixed domain, it will be possible to consider perturbations around the latter. This transformation was introduced by Koch [160], who uses it to show the smoothness of the free boundary and of the pressure up to the interface in any space dimension (see also the work of Kienzler [155]). It is subsequently adapted and used by Seis [244] for quantifying the self-similar asymptotics of the equation close to Barenblatt by using the spectrum of the linearized operator and invariant manifolds. In all of the above-cited works, the authors consider compactly supported, Hölder continuous initial pressures v_0, with non-vanishing gradient. This last condition ensures avoidance of the *waiting-time phenomenon*, namely the existence of a positive time $T^* > 0$ up to which the free boundary is stationary, see [268, Chapter 14].

3.1.1 Problem formulation

After the von Mises transform and after cosidering perturbations around the transformed Barenblatt, we are brought to consider the control problem for the transformed pertur-

bation equation (see [244, Section 3]):

$$\begin{cases} \partial_t y - \rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x y) = \mathcal{N}(y) + u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ (\rho^{\sigma+1}\partial_x y)(t,\pm 1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1), \end{cases} \tag{3.1.4}$$

where $T > 0$ and $\sigma > -1$, and the nonlinearity $\mathcal{N}(y) = \mathcal{N}(y, \partial_x y)$ is of the form

$$\mathcal{N}(y) = \rho F(y,\partial_x y) - \rho^{-\sigma}\partial_x(\rho^{\sigma+1} x F(y,\partial_x y)), \quad F(p,q) = \frac{q^2}{1+p+xq}, \quad p,q \in \mathbb{R}. \tag{3.1.5}$$

The distributed control $u = u(t,x)$ appearing in (3.1.4) actuates inside an open, non-empty subset $\omega = (a,b) \subsetneq (-1,1)$. The solution $y(t,x)$ is a perturbation around the Barenblatt in the new variables (see Remark 3.7.2). Consequently, the null-controllability of (3.1.4) would heuristically correspond to the exact-controllability of the pressure $v(t,z)$ and its free boundary of a controlled version of (3.1.2) to the original Barenblatt $\rho(z)$, after reverting the von Mises transformation. As said above, $m = \frac{\sigma+2}{\sigma+1} > -1$.

Hereinafter, we will investigate the null-controllability of (3.1.4), namely the possibility of steering the solution y to 0 at time T by means of the control u. Considering the full nonlinear problem (3.1.4) requires high regularity of the trajectory, and thus of the control. Due to the peculiar functional setting detailed below, ensuring this regularity is not straightforward. Hence, in this work, we will prove a local null-controllability result for a regularized version of the nonlinear problem (3.1.4), in which the singular terms appearing in the denominator of (3.1.5) have been removed.

3.1.2 Functional setting

Recalling the definition of the degenerate coefficient ρ in (3.1.3), for $k \geqslant 0$ we consider spaces

$$\mathcal{H}^k := \{f \in L^1_{\mathrm{loc}}(-1,1) \colon \|f\|_{\mathcal{H}^k} < \infty\},$$

where $\|f\|^2_{\mathcal{H}^k} := \langle f, f\rangle_{\mathcal{H}^k}$ is the norm induced from the inner product

$$\langle f_1, f_2\rangle_{\mathcal{H}^k} := \sum_{j=0}^{k} \int_{-1}^{1} \rho^{\sigma+j}(\partial_x^j f_1)(\partial_x^j f_2)\,\mathrm{d}x.$$

As $\rho^\sigma \in L^1(-1,1)$ whenever $\sigma > -1$, the measure $\rho^\sigma\,\mathrm{d}x$ is a Radon measure, it is absolutely continuous with respect to the Lebesgue measure $\mathrm{d}x$ and possesses the same null-sets. For any $k \geqslant 0$, $\mathcal{H}^k$ are separable Hilbert spaces of which $C^\infty([-1,1])$ are dense subsets according to [243, Lemma 2], [244, Section 4.2]. Additionally, on any $\omega \subsetneq (-1,1)$ they coincide with the unweighted Sobolev spaces $H^k(\omega)$, $k \geqslant 0$.

3.1.3 The main results

While the nonlinearity in (3.1.4) is essentially quadratic in a neighborhood of the origin, the denominator may be singular and applying a fixed-point argument using only the weighted Sobolev space theory is not straightforward. To mend this issue, in this paper we concentrate on a truncated version of the nonlinearity. Namely, we multiply the nonlinear terms by a smooth cut-off function which vanishes at points where y and/or $\partial_x y$ are large; the truncated equation would thus be linear at such points.

Let $\chi : [0,\infty) \longrightarrow [0,1]$ be a smooth cut-off function, supported on $[0,4)$ with $\chi(x) \equiv 1$ on $[0,1]$. Let $0 < \varepsilon, \delta < 1$ satisfying $4(\varepsilon+\delta) < 1$ be fixed. For $p,q \in \mathbb{R}$, and recalling the definition of F in (3.1.5), we define

$$F_{\varepsilon,\delta}(p,q) = \chi\left(\frac{p^2}{\delta^2}\right)\chi\left(\frac{q^2}{\varepsilon^2}\right)F(p,q). \tag{3.1.6}$$

We will henceforth only be interested in Problem (3.1.4) wherein $\mathcal{N}$ is replaced by $\rho F_{\varepsilon,\delta}$, namely

$$\begin{cases} \partial_t y - \rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x y) = \rho F_{\varepsilon,\delta}(y,\partial_x y) + u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ (\rho^{\sigma+1}\partial_x y)(t,\pm 1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1). \end{cases} \tag{3.1.7}$$

We recall, as per (3.1.3), that $\rho(x) = \frac{1}{2}(1-x^2)$. The main result we claim in this work is the following.

Theorem 3.1. *Let $T > 0$, let $\omega \subsetneq (-1,1)$ be an open, non-empty interval, and let $\sigma \in (-1,0)$. Then there exists $r > 0$ such that for every $y_0 \in \mathcal{H}^1$ satisfying $\|y_0\|_{\mathcal{H}^1} \leqslant r$, there exists a control $u \in L^2(0,T;L^2(\omega))$ for which the unique solution $y \in L^2(0,T;\mathcal{H}^2)\cap C^0([0,T];\mathcal{H}^1)$ of (3.1.7) satisfies $y(0,\cdot) = y_0$ and $y(T,\cdot) = 0$.*

Remark 3.1.1. *While being a first step in this direction, Theorem 3.1 is not sufficient to deduce a local controllability result to the Barenblatt trajectory for an associated distributed control problem of the free boundary problem (3.1.2). If (3.1.7) is null-controllable with the nonlinearity $\mathcal{N}(y) = \rho F(y) - \rho^{-\sigma}\partial_x(\rho^{\sigma+1}xF(y))$ as in (3.1.4), then one could deduce such a result. To achieve this, one would need to remove the cut-off factor $\chi(p^2/\delta^2)\chi(q^2/\varepsilon^2)$, and add the high order nonlinear term. The cut-off is identically 1 whenever the solution is of sufficiently small $C^{0,1}([0,T]\times[0,1])$-norm, and this regularity is also sufficient to revert the von Mises transformation. However, Theorem 3.1 does not provide this regularity. Nonetheless, it is the best result that can be obtained by means of an only $L^2(L^2)$-regular control. See Remark 3.7.2 for more details.*

Looking at (3.1.7), it is natural to first study the null-controllability of the corresponding linear problem, where the nonlinear term is replaced by a source term:

$$\begin{cases} \partial_t y - \rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x y) = f + u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ (\rho^{\sigma+1}\partial_x y)(t,\pm 1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1). \end{cases} \tag{3.1.8}$$

The nonlinear term would be seen as a small perturbation, and be dealt with by means of a fixed-point argument. The latter argument will rely on the particular structure of the nonlinearity, which is now non-singular and essentially quadratic due to the cut-off factor.

Remark 3.1.2. *The requirement $\sigma \in (-1,0)$ only appears when estimating the nonlinear term in the weighted spaces (see Section 3.4). The null-controllability and well-posedness of the linearized problem (3.1.8) holds true for any $\sigma > -1$, as seen below. We recall that σ is related to the nonlinearity exponent of the porous medium equation by $m = \frac{\sigma+2}{\sigma+1}$.*

To prove the null-controllability of Problem (3.1.8), we will make use of the so-called *source-term method*, first introduced by Liu, Takahashi & Tucsnak [191]. Roughly speaking, the strategy involves first showing the null-controllability of the homogeneous problem

$$\begin{cases} \partial_t y - \rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x y) = u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ (\rho^{\sigma+1}\partial_x y)(t,\pm 1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1), \end{cases} \tag{3.1.9}$$

and the null-controllability of Problem (3.1.8) follows provided the source term f vanishes with appropriate decay as $t \nearrow T$. More specifically, the decay of the source term should be quick enough near the final time compared to the control cost in small time. The null-controllability of problem (3.1.9) is done by combining duality and spectral techniques, making use of the results obtained in the works of Seis [243, 244]. Namely, we prove the following result.

Theorem 3.2. *Let $T > 0$, $\omega \subsetneq (-1,1)$ be an open, non-empty interval, and $\sigma > -1$. Then, for any $y_0 \in \mathcal{H}^0$, there exists a control $u \in L^2(0,T;L^2(\omega))$ such that the unique solution $y \in L^2(0,T;\mathcal{H}^1)\cap C^0([0,T];\mathcal{H}^0)$ of (3.1.9) satisfies $y(0,\cdot) = y_0$ and $y(T,\cdot) = 0$.*

3.1.4 State of the art

In [70], Coron, Diáz, Drici & Mignazzini prove the null-controllability of the porous medium equation set on $(0,1)$ using Dirichlet boundary controls on both ends as well as a scalar forcing control. A control on one end can also be used as long as the other boundary condition is a Neumann one. The authors' strategy follows the *return method* to avoid the appearance of a free boundary, namely, the construction of an adequate non-trivial time-only dependent trajectory, starting and ending at 0, around which the problem is linearized. By a scaling argument, global null-controllability is achieved in arbitrarily small time, and the method guarantees non-negativity of the controls, and thus of the state for positive initial data. This differs from the original motivation behind our work, which was to control the pressure and its free boundary to the non-trivial Barenblatt profile (instead of the null-state). We also refer to the works of Liu & Gao [189, 190] for nonnegativity preserving approximate controllability results for the multi-dimensional porous medium equation set on a bounded domain by means of a distributed control.

Null-controllability results for one-dimensional parabolic equations which degenerate at the boundary such as

$$\partial_t y - \partial_x(x^\alpha \partial_x y) = u\mathbf{1}_\omega \quad \text{in } (0,T)\times(0,1),$$

where $\alpha \in [0,2)$ are shown in the works of Alabau-Boussouira, Cannarsa, Martinez & Vancostenoble [1, 49, 50] by using Carleman inequalities with degeneracy-adapted weights. In general, one distinguishes the *weak* ($\alpha \in [0,1)$) and *strong* ($\alpha \in [1,2)$) degeneracies, as the functional setting and boundary conditions are different for both cases. The case $\alpha \geqslant 2$ is excluded as null-controllability does not hold (only *regional* results are true, see [48]). We also refer to the monograph [51] for results on two dimensional problems of the above kind. The question of boundary null-controllability has also been addressed. For instance, Gueye [124] combines the transmutation method and spectral techniques for a weakly degenerate problem, and Moyano [210] makes use of the flatness method for a strongly degenerate problem.

These studies have been extended in the works of Cannarsa, Fragnelli & Rocchetti [46, 47, 107] to degenerate parabolic problems in non-divergence form (more alike (3.1.4)), such as

$$\partial_t y - a(x)\partial_x^2 y + b(x)\partial_x y + c(t,x)y = u\mathbf{1}_\omega \quad \text{in } (0,T)\times(0,1)$$

where $a \in C^0([0,1])$ may degenerate at $x=0$ and $x=1$. Therein, pure homogeneous Dirichlet and Neumann boundary conditions are considered, and null-controllability results are obtained by Carleman inequalities.

Our work may be seen as a further contribution to the controllability theory of linear degenerate parabolic equations. Indeed, while the differential operator in (3.1.9) may be rewritten as $-\rho\partial_x^2 y+(\sigma+1)x\partial_x y$, the weighted Neumann boundary conditions, which are the natural ones from the calculus of variations point of view, have not been considered in the above-cited papers on problems in non-divergence form. In particular, we do not consider the same weight and functional framework as in [47, 107], since $\frac{b}{a} = \frac{2(\sigma+1)x}{1-x^2} \notin L^1(-1,1)$ in our case. While we use spectral techniques, up to the best of our knowledge, a Carleman inequality for our functional setting is lacking.

Finally, we mention that our strategy for proving the null-controllability of the linearized problem (3.1.9) can also be applied to obtain a null-controllability result (see Section 3.5) for the thin-film equation linearized around its self-similar solution, which is a fourth-order degenerate parabolic equation. Up to the best of our knowledge, this has not been tackled in the literature.

3.1.5 Scope

We present the functional properties of the governing differential operator in Section 3.2. In Section 3.3, we use its explicit spectrum for proving Theorem 3.2. An adaptation

of the source-term method allows us to deduce the null-controllability of the linearized problem (3.1.8) (Theorem 3.4). In Section 3.4, we conclude the proof of Theorem 3.1 by means of a Banach fixed-point argument. Finally, in Section 3.5 we apply the linear controllability theory from Section 3.3 to deduce the null-controllability of the linearized thin-film equation, a fourth-order analog of (3.1.9).

3.1.6 Notation

Whenever the dependence on parameters of a constant is not specified, we will write $f \lesssim_S g$ whenever a constant $C \geqslant 1$, depending only on the set of parameters S, exists such that $f \leqslant Cg$.

3.2 The linear degenerate operator

This section is dedicated to a study of the functional and spectral properties of the linear operator $\mathcal{A} = -\rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x)$, which will be shown to be self-adjoint and with compact resolvents when viewed as an unbounded operator on the weighted Lebesgue space $\mathcal{H}^0$. The arguments will follow standard theory, starting by noting that symmetry holds as

$$\int_{-1}^{1} \rho^{\sigma}(\mathcal{A}f_1)f_2 \,\mathrm{d}x = \int_{-1}^{1} \rho^{\sigma+1}(\partial_x f_1)(\partial_x f_2)\,\mathrm{d}x \tag{3.2.1}$$

for all $f_1, f_2 \in C^\infty([-1,1])$ via integration by parts. To accurately characterize the domain of $\mathcal{A}$ we present some embedding results for the weighted Sobolev spaces $\mathcal{H}^k$.

3.2.1 Embeddings for weighted Sobolev spaces

The following two useful lemmas are adapted from the work of Gnann [118]. For the sake of completeness, we provide short proofs in Appendix 3.7.

Lemma 3.2.1. *Let $\alpha \in \mathbb{R}$. Then*

$$\left\|(1-x^2)^{\alpha} f\right\|^2_{C^0([-1,1])} \lesssim_\alpha \int_{-1}^{1} (1-x^2)^{2\alpha-1} f^2 \,\mathrm{d}x + \int_{-1}^{1} (1-x^2)^{2\alpha+1}(\partial_x f)^2 \,\mathrm{d}x \tag{3.2.2}$$

holds for all $f \in C^\infty([-1,1])$.

The following Lemma is as a Hardy-like inequality for the spaces $\mathcal{H}^k$.

Lemma 3.2.2. *Suppose $\alpha > -1$ and $\beta \in \mathbb{R}$. Then there exists $C = C(\alpha,\beta) > 0$ such that*

$$\int_{-1}^{1} (1-x^2)^{\alpha} f^2 \,\mathrm{d}x \leqslant C \int_{-1}^{1} \left((1-x^2)^{\beta} f^2 + (1-x^2)^{\alpha+2}(\partial_x f)^2\right) \mathrm{d}x \tag{3.2.3}$$

holds for all $f \in C^\infty([-1,1])$. The constant $C(\alpha,\beta)$ diverges as $\alpha \searrow -1$.

Remark 3.2.3. *We highlight that an inequality such as*

$$\int_{-1}^{1} (1-x^2)^{\alpha} f^2 \,\mathrm{d}x \lesssim_\alpha \int_{-1}^{1} (1-x^2)^{\alpha+2}(\partial_x f)^2 \,\mathrm{d}x$$

is not true, as any nonzero constant is a counterexample.

We combine the two previous lemmas to deduce the following result, which may also be seen as a weighted trace estimate.

Lemma 3.2.4. *Let $k \geqslant 1$, $\ell \geqslant 0$ and $\alpha \geqslant \frac{\sigma+1+\ell-k}{2}$ with $\alpha > 0$. Then there exists $C = C(k,\alpha) > 0$ such that*

$$\left\|(1-x^2)^\alpha \partial_x^\ell f\right\|_{C^0([-1,1])} \leqslant C\|f\|_{\mathcal{H}^{k+\ell}}$$

holds for all $f \in C^\infty([-1,1])$.

Proof. We may replace f by its derivatives, in view of the definition of the $\mathcal{H}^k$-norms. It is thus sufficient to prove the statement for $\ell = 0$. The latter fact follows by successive applications of (3.2.3) to (3.2.2), with the effect of

$$\|(1-x^2)^\alpha f\|^2_{C^0([-1,1])} \lesssim_{\alpha,k} \sum_{j=0}^{k} \int_{-1}^{1} (1-x^2)^{2\alpha+2k-1} (\partial_x^j f)^2 \,\mathrm{d}x. \tag{3.2.4}$$

As $(1-x^2) \leqslant 1$ in $[-1,1]$ and as, by definition,

$$\|f\|^2_{\mathcal{H}^k} := \sum_{j=0}^{k} \int_{-1}^{1} (1-x^2)^{\sigma+j} (\partial_x^j f)^2 \,\mathrm{d}x,$$

we see that if one picks $\alpha \geqslant \frac{\sigma+1-k}{2}$ with $\alpha > 0$ in (3.2.4), then $2\alpha + 2k - 1 \geqslant \sigma + j$ for all $0 \leqslant j \leqslant k$. Thus

$$\sum_{j=0}^{k} \int_{-1}^{1} (1-x^2)^{2\alpha+2k-1} (\partial_x^j f)^2 \,\mathrm{d}x \leqslant \|f\|^2_{\mathcal{H}^k},$$

which in view of (3.2.4) allows us to conclude. □

The inequality above fails when $k = \sigma + 1 + \ell$ due to the failure of the underlying Hardy inequality. Let us now illustrate (in the particular case of $\mathcal{H}^2$, and recall the definition of ρ in (3.1.3)) why the previous Lemma may be seen as a weighted trace estimate.

Lemma 3.2.5 (Boundary conditions)**.** *Let $\sigma > -1$. Then $(\rho^{\sigma+1}\partial_x f)(\pm 1) = 0$ for any $f \in \mathcal{H}^2$.*

Proof. By Lemma 3.2.4, $\rho^{\sigma+1}\partial_x f$ is continuous on $[-1,1]$. Thus there exists $\lambda \in \mathbb{R}$ such that $(\rho^{\sigma+\ell}\partial_x^\ell f)(x) \longrightarrow \lambda$ as $x \longrightarrow \pm 1$. Should $\lambda \neq 0$, then by continuity

$$\rho(x)^\sigma (\partial_x f(x))^2 \geqslant \frac{\lambda^2}{4\rho(x)^{\sigma+2}}$$

for x near ± 1. As $1/\rho^{\sigma+2} \notin L^1(-1,1)$ whenever $\sigma > -1$, the above inequality along with Lemma 3.2.2 contradict $f \in \mathcal{H}^2$. □

3.2.2 Spectrum of the linear operator

We henceforth fix $\sigma > -1$. We summarize the main functional and spectral properties of $\mathcal{A} = -\rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x)$. We begin by the following a priori estimate.

Lemma 3.2.6. *Let $f \in \mathcal{H}^0$ be given and let $u \in \mathcal{H}^1$ be a weak solution to $\mathcal{A}u = f$. There exist $C_1 > 0$ and $C_2 > 0$ such that*

$$(\sigma+1)^2 \int_{-1}^{1} \rho^\sigma x^2 (\partial_x u)^2 \,\mathrm{d}x \leqslant C_1 \int_{-1}^{1} \rho^\sigma f^2 \,\mathrm{d}x,$$

and

$$\int_{-1}^{1} \rho^{\sigma+2} (\partial_x^2 u)^2 \,\mathrm{d}x \leqslant C_2 \int_{-1}^{1} \rho^\sigma f^2 \,\mathrm{d}x.$$

Proof. Let $\{f_k\}_{k=0}^{\infty} \subset C^\infty([-1,1])$ be a sequence converging to $f \in \mathcal{H}^0$, and let u_k be a weak solution to $\mathcal{A}u_k = f_k$ for $k \geqslant 0$. By [165], $u_k \in C^\infty([-1,1])$ and we may thus work with smooth functions to conclude. For notational simplicity, we remove the subscripts k in what follows. We multiply $\mathcal{A}u = f$ by $(\sigma+1)\rho^\sigma x\partial_x u$ and integrate:

$$-(\sigma+1)\int_{-1}^{1} \partial_x(\rho^{\sigma+1}\partial_x u)x\partial_x u \,\mathrm{d}x = (\sigma+1)\int_{-1}^{1} \rho^\sigma f x\partial_x u \,\mathrm{d}x. \tag{3.2.5}$$

Integration by parts allows us to rewrite the left-most term as

$$-\int_{-1}^{1} \partial_x(\rho^{\sigma+1}\partial_x u)x\partial_x u \,\mathrm{d}x = \int_{-1}^{1} \rho^{\sigma+1}(\partial_x u)^2 \,\mathrm{d}x + \int_{-1}^{1} \rho^{\sigma+1}x(\partial_x u)(\partial_x^2 u) \,\mathrm{d}x.$$

We now integrate the right-most term by parts to deduce

$$\begin{aligned}\int_{-1}^{1} \rho^{\sigma+1}x(\partial_x u)(\partial_x^2 u) \,\mathrm{d}x = (\sigma+1)\int_{-1}^{1} \rho^\sigma x(\partial_x u)^2 \,\mathrm{d}x - \int_{-1}^{1} \rho^{\sigma+1}x(\partial_x^2 u)(\partial_x u) \,\mathrm{d}x \\ - \int_{-1}^{1} \rho^{\sigma+1}(\partial_x u)^2 \,\mathrm{d}x.\end{aligned}$$

Hence,

$$-\int_{-1}^{1} \partial_x(\rho^{\sigma+1}\partial_x u)x\partial_x u \,\mathrm{d}x = \frac{1}{2}\int_{-1}^{1} \rho^{\sigma+1}(\partial_x u)^2 \,\mathrm{d}x + \frac{\sigma+1}{2}\int_{-1}^{1} \rho^\sigma x^2(\partial_x u)^2 \,\mathrm{d}x. \tag{3.2.6}$$

As $\sigma > -1$, plugging (3.2.6) in (3.2.5) and applying the Young inequality to deduce

$$(\sigma+1)^2\int_{-1}^{1} \rho^\sigma x^2(\partial_x u)^2 \,\mathrm{d}x \leqslant \epsilon(\sigma+1)^2\int_{-1}^{1} \rho^\sigma x^2(\partial_x u)^2 \,\mathrm{d}x + \frac{1}{4\epsilon}\int_{-1}^{1} \rho^\sigma f^2 \,\mathrm{d}x$$

for all $\epsilon > 0$. Choosing $\epsilon < 1$ yields the desired conclusion.

For the second estimate, using $(a-b)^2 \leqslant 2a^2 + 2b^2$ we see that

$$\begin{aligned}\int_{-1}^{1} \rho^{\sigma+2}(\partial_x^2 u)^2 \,\mathrm{d}x &= \int_{-1}^{1} \rho^\sigma \left(f - (\sigma+1)x\partial_x u\right)^2 \,\mathrm{d}x \\ &\leqslant 2\int_{-1}^{1} \rho^\sigma f^2 \,\mathrm{d}x + 2(\sigma+1)^2\int_{-1}^{1} \rho^\sigma x^2(\partial_x u)^2 \,\mathrm{d}x.\end{aligned}$$

We may thus conclude using the first estimate in the statement. □

Proposition 3.2.7. *The operator $\mathcal{A} : \mathcal{H}^2 \longrightarrow \mathcal{H}^0$ is self-adjoint, nonnegative, and has compact resolvents.*

Proof. Let us first recall that any symmetric, densely defined operator on a Hilbert space $\mathcal{H}$ is closable, meaning the closure of its graph in $\mathcal{H} \oplus \mathcal{H}$ is again the graph of a linear, symmetric operator. Identity (3.2.1) shows that $\mathcal{A}_{|_{C^\infty([-1,1])}}$ is a symmetric, densely defined operator on the Hilbert space $\mathcal{H}^0$. Let us denote the closure of this operator by $\mathcal{A}$, with domain $\mathcal{D}(\mathcal{A})$. Our goal is to show that $\mathcal{A}$ is the unique self-adjoint extension of $\mathcal{A}_{|_{C^\infty([-1,1])}}$, with domain $\mathcal{D}(\mathcal{A}) = \mathcal{H}^2$.

A standard approximation argument yields $\mathcal{H}^2 \subset \mathcal{D}(\mathcal{A})$. We will show that $\mathbf{A} := \mathcal{A}_{|_{\mathcal{H}^2}}$ is a self-adjoint operator by proving $\mathbf{A}^* \subset \mathbf{A}$. The chain $\mathbf{A} \subset \mathcal{A} \subset \mathcal{A}^* \subset \mathbf{A}^*$ would then imply that $\mathcal{A} = \mathbf{A}$ is self-adjoint, and that any other self-adjoint extension of $\mathcal{A}_{|_{C^\infty([-1,1])}}$ would be jammed in-between $\mathcal{A}$ and $\mathcal{A}^*$ in the above inclusions, and hence coincide with $\mathcal{A}$.

Let $\mathbf{L} := \mathbf{A} + \mathrm{Id}$. The desired inclusion $\mathbf{A}^* \subset \mathbf{A}$ would follow by showing $\mathbf{L}^* \subset \mathbf{L}$. The latter requires us to show that if $u \in \mathcal{H}^0$ is such that $u \in \mathcal{D}(\mathbf{L}^*)$, then $u \in \mathcal{H}^2$.

To this end, we begin by observing that for $f \in \mathcal{H}^0$, the Poisson problem

$$\begin{cases} -\rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x w) + w = f & \text{in } (-1,1) \\ (\rho^{\sigma+1}\partial_x w)(\pm 1) = 0, \end{cases}$$

has a unique weak solution $w \in \mathcal{H}^1$ by Lax-Milgram, and $w \in \mathcal{H}^2$ by Lemma 3.2.6. The operator $\mathbf{L}$ is hence boundedly invertible. Let $u \in \mathcal{D}(\mathbf{L}^*)$. Thus there exists $f \in \mathcal{H}^0$ such that

$$\langle u, \mathbf{L}v\rangle_{\mathcal{H}^0} = \langle f, v\rangle_{\mathcal{H}^0} \quad \text{for all } v \in \mathcal{H}^2. \tag{3.2.7}$$

For this f, let w denote the weak solution to the Poisson problem above; w also satisfies (3.2.7) after integration by parts. Thus, taking the difference and considering the test function $v = \mathbf{L}^{-1}(u - w) \in \mathcal{H}^2$, we conclude that $u \in \mathcal{H}^2$.

Finally, by the estimate $\|u\|_{\mathcal{H}^1} \leqslant \|f\|_{\mathcal{H}^0}$ and the compact embedding $\mathcal{H}^1 \hookrightarrow \mathcal{H}^0$ (see [243, Lemma 4]), it is seen that $(\mathcal{A} + \mathrm{Id})^{-1}$ is compact, and we obtain the desired conclusion. □

By well-known results, we deduce that $\mathcal{A} : \mathcal{H}^2 \longrightarrow \mathcal{H}^0$ has a purely discrete spectrum consisting of an increasing sequence of nonnegative eigenvalues $\{\lambda_k\}_{k=0}^{\infty}$ with $\lim_{k\longrightarrow\infty} \lambda_k = \infty$, and an associated sequence of eigenfunctions that form an orthonormal basis of $\mathcal{H}^0$. In order to use spectral techniques for studying the null-controllability of problem (3.1.8), we need knowledge of the explicit spectrum of $\mathcal{A}$. The definition of the eigenfunctions involves the rising factorials (also called Pochhammer symbols):

$$(s)_j = s(s+1)\dots(s+j-1) \quad \text{for } j \in \mathbb{N} \quad \text{and } (s)_0 = 1 \quad \text{for } s \in \mathbb{R}.$$

For fixed $a, b, c, x \in \mathbb{C}$, we define the hypergeometric series ${}_2F_1(a, b; c; x)$ by

$${}_2F_1(a,b,c,x) := \sum_{j=0}^{\infty} \frac{(a)_j (b)_j}{(c)_j j!} x^j,$$

provided c is not an integer $\leqslant 0$. The series is convergent if $|x| < 1$, and terminates if $a \in \mathbb{Z}$ and becomes a polynomial (see [252, Chapter IV]). We also recall the standard integral definition of the Gamma function $\Gamma(z)$ for $z \in (0, \infty)$:

$$\Gamma(z) = \int_0^{\infty} e^{-t} t^{z-1}\, \mathrm{d}t.$$

Γ is a monotone increasing function on $(0, \infty)$, and this integral form shows that $\Gamma(1) = 1$ and that $z\Gamma(z) = \Gamma(z+1)$ holds for all $z \in (0, \infty)$.

The following, albeit reformulated result is shown by Seis [243, Theorem 1] and [244, Proposition 6.1] (in any dimension), following ideas from Denzler & McCann [78]. In the one-dimensional case, it may also be found in a previous work of Angenent [14].

Theorem 3.3. *The spectrum of $\mathcal{A}$ consists of simple nonnegative eigenvalues $\{\lambda_k\}_{k=0}^{\infty}$, given by*

$$\lambda_k = \frac{k^2}{2} + \frac{k}{2}(1 + 2\sigma)$$

for $k \geqslant 0$. The corresponding eigenfunctions $\{\varphi_k\}_{k=0}^{\infty}$ are of the form

$$\varphi_k(x) = {}_2F_1\left(-\frac{k}{2}, \sigma + \frac{k}{2} + \frac{1}{2}, \frac{1}{2}, x^2\right) \quad \textit{if } k \textit{ is even}$$

and

$$\varphi_k(x) = {}_2F_1\left(-\frac{k-1}{2}, \sigma + \frac{k}{2} + 1, \frac{3}{2}, x^2\right) x \quad \textit{if } k \textit{ is odd}$$

for $x \in (-1, 1)$. In particular, $\lambda_0 = 0$ with associated eigenfunction $\varphi_1(x) = 1$ since constants are in the domain of $\mathcal{A}$.

Let us comment the results of [243, 244] in the specific one-dimensional case we are treating here. A key observation is that the operator $\mathcal{A}$ commutes with the parity operator $\mathbb{P}$ defined by $(\mathbb{P}f)(x) = f(-x)$, which has two eigenvalues: ± 1. One may identify the restriction of $\mathcal{A}$ to even functions with $\ell = 0$, and odd functions with $\ell = 1$, and then aims to simultaneously diagonalize both operators. In [243], Seis computes the spectrum by relating the derived eigenvalue problem to a second-order Fuchsian ODE with three regular singular points. A point spectrum $\{\lambda_{\ell\kappa}\}_{\kappa=0}^{\infty}$ is obtained, which for convenience we merge here by setting $2\kappa = k$ for $\ell = 0$ and $2\kappa + 1 = k$ for $\ell = 1$.

On another note, as mentioned above the series defining the eigenfunctions φ_k terminates since $-\frac{k}{2} \in \mathbb{Z}$ when k is even (similarly $-\frac{k-1}{2} \in \mathbb{Z}$ when k is odd). Thus, φ_k are polynomials of degree k. It is more advantageous to represent the eigenfunctions in terms of classical orthogonal polynomials, for which explicit norm relations and asymptotic behavior are known. We may in fact relate the eigenfunctions to Jacobi polynomials $P_\ell^{(\alpha,\beta)}(\cdot)$, as:

$$ {}_2F_1(-\ell, \alpha+\beta+\ell+1, \alpha+1, x) = \frac{\ell!}{(\alpha+1)_\ell} P_\ell^{(\alpha,\beta)}(1-2x) \quad x \in (-1,1), $$

for $\alpha, \beta > -1$ and $\ell \geqslant 0$, see [252, Chapter IV] for instance. The Jacobi polynomials are orthogonal in $L^2(-1,1)$ with respect to the weight $(1-x)^\alpha(1+x)^\beta$:

$$ \int_{-1}^{1} (1-x)^\alpha (1+x)^\beta P_\ell^{(\alpha,\beta)}(x)^2 \,\mathrm{d}x = \frac{2^{\alpha+\beta+1}}{2\ell+\alpha+\beta+1} \frac{\Gamma(\ell+\alpha+1)\Gamma(\ell+\beta+1)}{\Gamma(\ell+\alpha+\beta+1)\ell!}, \tag{3.2.8} $$

see [252, Chapter IV, Section 4.1], which holds for $\alpha, \beta > -1$. Using this, relatively straightforward computations yield the normalized eigenfunctions of the form

$$ \overline{\varphi}_k(\cdot) = c_k \varphi_k(\cdot) \tag{3.2.9} $$

as per the following result.

Lemma 3.2.8. *Let $k \geqslant 0$, and let φ_k be the k-th eigenfunction of $\mathcal{A}$. Then*

$$ \|\varphi_{2\ell}\|_{\mathcal{H}^0}^2 = 2^{-\sigma} \Big(\frac{1}{2}\Big)_\ell^{-2} \frac{\ell!\Gamma(\ell+\frac{1}{2})\Gamma(\ell+\sigma+1)}{(2\ell+\sigma+\frac{1}{2})\Gamma(\ell+\sigma+\frac{1}{2})} $$

if $k = 2\ell$ is even, and

$$ \|\varphi_{2\ell+1}\|_{\mathcal{H}^0}^2 = 2^{-\sigma} \Big(\frac{3}{2}\Big)_\ell^{-2} \frac{\ell!\Gamma(\ell+\frac{3}{2})\Gamma(\ell+\sigma+1)}{(2\ell+\sigma+\frac{3}{2})\Gamma(\ell+\sigma+\frac{3}{2})} $$

if $k = 2\ell+1$ is odd.

Proof. Let $k = 2\ell$ be even. We write

$$ 2^{-\sigma} \int_{-1}^{1} (1-x^2)^\sigma \varphi_{2\ell}^2 \,\mathrm{d}x = 2^{-\sigma} \Big(\frac{1}{2}\Big)_\ell^{-2} (\ell!)^2 \int_{-1}^{1} (1-x^2)^\sigma P_\ell^{(-\frac{1}{2},\sigma)}(1-2x^2) \,\mathrm{d}x. $$

A simple change of variables yields

$$ \int_{-1}^{1} (1-x^2)^\sigma P_\ell^{(-\frac{1}{2},\sigma)}(1-2x^2) \,\mathrm{d}x = 2^{-\sigma-\frac{1}{2}} \int_{-1}^{1} (1-z)^{-\frac{1}{2}} (1+z)^\sigma P_\ell^{(-\frac{1}{2},\sigma)}(z) \,\mathrm{d}z. $$

Using the orthonormality relation (3.2.8), we obtain

$$ \int_{-1}^{1} (1-z)^{-\frac{1}{2}} (1+z)^\sigma P_\ell^{(-\frac{1}{2},\sigma)}(z) \,\mathrm{d}z = \frac{2^{\sigma+\frac{1}{2}}}{2\ell+\sigma+\frac{1}{2}} \frac{\Gamma(\ell+\frac{1}{2})\Gamma(\ell+\sigma+1)}{\ell!\Gamma(\ell+\sigma+\frac{1}{2})}. $$

We deduce

$$ \|\varphi_{2\ell}\|_{\mathcal{H}^0}^2 = 2^{-\sigma} \Big(\frac{1}{2}\Big)_\ell^{-2} \frac{\ell!\Gamma(\ell+\frac{1}{2})\Gamma(\ell+\sigma+1)}{(2\ell+\sigma+\frac{1}{2})\Gamma(\ell+\sigma+\frac{1}{2})}. $$

The case when k is odd follows from an analogous computation. □

3.3 Null-controllability of the linearized problem

Before proceeding with the proofs of the controllability results for the linearized problems, let us argue the well-posedness of

$$\begin{cases} \partial_t y - \rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x y) = f & \text{in } (0,T)\times(-1,1) \\ \big(\rho^{\sigma+1}\partial_x y\big)(t,\pm 1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1), \end{cases} \tag{3.3.1}$$

where $T>0$ and f is an arbitrary source term. The following result holds.

Proposition 3.3.1. *For every $y_0 \in \mathcal{H}^0$ and $f \in L^2(0,T;\mathcal{H}^0)$, there exists a unique weak solution*

$$y \in L^2(0,T;\mathcal{H}^1)\cap C^0([0,T];\mathcal{H}^0)$$

to Problem (3.3.1) satisfying the estimate

$$\|y\|_{C^0([0,T];\mathcal{H}^0)} + \|y\|_{L^2(0,T;\mathcal{H}^1)} \leqslant C_T\big(\|f\|_{L^2(0,T;\mathcal{H}^0)} + \|y_0\|_{\mathcal{H}^0}\big) \tag{3.3.2}$$

for some $C_T>0$. If moreover $y_0 \in \mathcal{H}^1$, then y is a strong solution enjoying maximal regularity

$$y \in L^2(0,T;\mathcal{H}^2)\cap H^1(0,T;\mathcal{H}^0)\cap C^0([0,T];\mathcal{H}^1)$$

along with the estimate

$$\|y\|_{C^0([0,T];\mathcal{H}^1)} + \|y\|_{L^2(0,T;\mathcal{H}^2)} \leqslant C_T\big(\|f\|_{L^2(0,T;\mathcal{H}^0)} + \|y_0\|_{\mathcal{H}^1}\big) \tag{3.3.3}$$

for some $C_T>0$.

Proof. The statement follows from well-known semigroup theory results (see for instance [26, Part II, Chapter 1, Section 3]). Indeed, Proposition 3.2.7 along with [26, Theorem 2.12, Section 2] imply that the self-adjoint operator $(\mathcal{A},\mathcal{D}(\mathcal{A}))$ generates an analytic semigroup in $\mathcal{H}^0$.

We remark that the semigroup theory results make use of the fact that $\mathcal{H}^1$ is the $(\frac{1}{2},2)$-interpolation space of $\mathcal{D}(\mathcal{A})=\mathcal{H}^2$ and $\mathcal{H}^0$. A proof of this may be found in [118, Lemma 3.6] and also [117, Lemma 1.7].

The constant C_T in estimates (3.3.2), (3.3.3) depends on T due to the fact that first eigenfunction of $\mathcal{A}$ is associated with the eigenvalue 0. Thus, the contribution of this first mode to the $L^2(0,T;\mathcal{H}^1)$-norm of y is not bounded as $T\to\infty$. □

As discussed in the introduction, the null-controllability of Problem (3.1.8) requires first proving Theorem 3.2, regarding the null-controllability of the homogeneous problem

$$\begin{cases} \partial_t y - \rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x y) = u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ \big(\rho^{\sigma+1}\partial_x y\big)(t,\pm 1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1). \end{cases} \tag{3.3.4}$$

3.3.1 The homogeneous problem

The main objective in what follows is to provide a proof to Theorem 3.2. Let us begin with a short review of some well-known notions on the null-controllability of linear systems. Let $\mathcal{H}$ and U be two Hilbert spaces. Consider the linear control system

$$\begin{cases} y' = Ay + Bu & \text{in } (0,T) \\ y(0,\cdot) = y_0 \in \mathcal{H}, \end{cases} \tag{3.3.5}$$

where $A:\mathcal{D}(A)\longrightarrow\mathcal{H}$ is the generator of a strongly continuous semigroup $\{e^{tA}\}_{t\geqslant 0}$ on $\mathcal{H}$ and $B\in\mathcal{L}(U,\mathcal{H})$. If (3.3.5) is null-controllable in time $T>0$ then the set

$$\mathcal{U}_{T,y_0} = \{u\in L^2(0,T;U)\colon y(T,\cdot)=0\}$$

is non-empty. The quantity

$$\kappa(T) := \sup_{\|y_0\|_{\mathcal{H}}=1} \inf_{u\in\mathcal{U}_{T,y_0}} \|u\|_{L^2(0,T;U)}$$

is called the *control cost* in time T. It is known that if (3.3.5) is null-controllable in any time $T > 0$, then $\kappa : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ is continuous and non-increasing, and $\lim_{T\searrow 0^+} \kappa(T) = \infty$. This namely implies that for every function $\gamma : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ with $\kappa(t) < \gamma(t)$ for every $t > 0$, for every $T > 0$ there exists a control driving the solution of (3.3.5) to rest im time T such that

$$\|u\|_{L^2(0,T;U)} \leqslant \gamma(T)\|y_0\|_{\mathcal{H}}.$$

Let us consider the adjoint problem

$$\begin{cases} -\zeta' = A^*\zeta & \text{in } (0,T) \\ \zeta(T,\cdot) = \zeta_T \in \mathcal{H}. \end{cases} \tag{3.3.6}$$

The following result is relatively standard and may be found in [191, Propositon 2.2], and originates from the work of Fattorini & Russell [97]. Due to a minimal change in the assumptions with respect to [191], we give a short proof below.

Lemma 3.3.2. *Assume that A is a negative operator*[1]*, with an orthonormal basis of eigenfunctions $\{\varphi_k\}_{k=0}^{\infty}$ and corresponding decreasing sequence of eigenvalues $\{-\lambda_k\}_{k=0}^{\infty}$ which satisfy*

$$\inf_{k\geqslant 0}(\lambda_{k+1} - \lambda_k) = s > 0 \tag{3.3.7}$$

$$\lambda_k = rk^2 + \mathcal{O}(k) \qquad \text{as } k \longrightarrow \infty \tag{3.3.8}$$

for some $r > 0$. Assume U is a separable Hilbert space and that there exists $\mu > 0$ such that

$$\|B^*\varphi_k\|_U \geqslant \mu \tag{3.3.9}$$

for all $k \geqslant 0$. Then there exists a constant $C_{\text{obs}} = C_{\text{obs}}(T) > 0$ such that the observability inequality

$$\|\zeta(0,\cdot)\|_{\mathcal{H}}^2 \leqslant C_{\text{obs}}^2 \int_0^T \|B^*\zeta\|_U^2 \,\mathrm{d}t \tag{3.3.10}$$

holds for any $\zeta_T \in \mathcal{H}$, where ζ is the corresponding solution to (3.3.6).

Proof. We may write the Fourier decomposition of ζ as

$$\zeta(t,x) = \sum_{k=0}^{\infty} e^{-\lambda_k(T-t)}\langle\zeta_T,\varphi_k\rangle_{\mathcal{H}}\,\varphi_k(x). \tag{3.3.11}$$

Since U is separable, it has an orthonormal basis $\{\psi_j\}_{j=0}^{\infty}$, which combined with identity (3.3.11) and the time-shift $T - t \longmapsto t$ gives

$$\int_0^T \|B^*\zeta\|_U^2 \,\mathrm{d}t = \sum_{j=0}^{\infty} \int_0^T \left| \sum_{k=0}^{\infty} e^{-\lambda_k t}\langle\zeta_T,\varphi_k\rangle_{\mathcal{H}}\langle B^*\varphi_k,\psi_j\rangle_U \right|^2 \mathrm{d}t \tag{3.3.12}$$

for $T > 0$ and $\zeta_T \in \mathcal{H}$. Now, making use of assumptions (3.3.7), (3.3.8), we deduce from [242, Theorem 1] that there exists $C(T) > 0$ such that $\lim_{T\searrow 0^+} C(T) = \infty$ and

$$C(T)\int_0^T \left|\sum_{k=0}^{\infty} a_k e^{-\lambda_k t}\right|^2 \mathrm{d}t \geqslant \sum_{k=0}^{\infty} |a_k|^2 e^{-2\lambda_k T}$$

[1] Meaning $\langle Ay, y\rangle_{\mathcal{H}} \leqslant 0$ for $y \in \mathcal{D}(A)$.

for all $T > 0$ and $\{a_k\}_{k=0}^{\infty} \in \ell^2(\mathbb{N})$. Applying this estimate in (3.3.12) gives

$$C(T) \int_0^T \|B^*\zeta\|_U^2 \,\mathrm{d}t \geqslant \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} e^{-2\lambda_k T} \langle \zeta_T, \varphi_k \rangle_{\mathcal{H}}^2 \langle B^*\varphi_k, \psi_j \rangle_U^2$$

for $T > 0$ and $\zeta_T \in \mathcal{H}$. This last estimate along with assumption (3.3.9) yields

$$C(T) \int_0^T \|B^*\zeta\|_U^2 \,\mathrm{d}t \geqslant \mu^2 \|\zeta(0,\cdot)\|_{\mathcal{H}}^2$$

for $\zeta_T \in \mathcal{H}$. The observability inequality (3.3.10) thus holds with $C_{\mathrm{obs}}(T) = \sqrt{\frac{C(T)}{\mu^2}}$. □

Remark 3.3.3 (Decay of the control cost)**.** *When the operator A is strictly negative*[2] *(and thus $\lambda_0 > 0$), then there exist $M_1, M_2 > 0$ such that*

$$\kappa(T) < M_1 e^{\frac{M_2}{T}} \qquad \text{for all } T > 0.$$

This will not hold in our case since $\lambda_0 = 0$ is an eigenvalue. The cost may be shown to be of the same exponential form for small time (see [242, Section 5.2]), but in the long time limit $T \to \infty$, it is rather of the order of a constant. The control cost plays a role in choosing the explicit time-weights in the source-term method, as seen below. For a thorough study, we refer to Tenenbaum & Tucsnak [256].

In our framework, we take $\mathcal{H} = \mathcal{H}^0$ and $U = \mathcal{H}^0(\omega) = L^2(\omega, \rho^\sigma dx)$. The control operator $B \in \mathcal{L}(U, \mathcal{H})$ is given by $Bu = u\mathbf{1}_\omega$, where $\omega = (a, b) \subsetneq (-1, 1)$ is non-empty. Hence, $B^*u = u|_\omega$.

Using Lemma 3.3.2 and the spectral results from Subsection 3.2.2, we are now in a position to prove the following result.

Lemma 3.3.4. *Let $\sigma > -1$. The eigenvalues $\{\lambda_k\}_{k=0}^{\infty}$ and associated normalized eigenfunctions $\{\overline{\varphi}_k\}_{k=0}^{\infty}$ of the negative operator $\mathcal{A} : \mathcal{H}^2 \longrightarrow \mathcal{H}^0$ satisfy conditions (3.3.7)-(3.3.9).*

Proof. Due to their form, it is readily seen that the eigenvalues given in Theorem 3.3 satisfy (3.3.8). The separation condition (3.3.7) follows from a simple computation:

$$\lambda_{k+1} - \lambda_k \geqslant k + \sigma + 1 \geqslant \sigma + 1 \qquad \text{for any } k \geqslant 0.$$

The main issue will be to show that the normalized eigenfunctions satisfy condition (3.3.9). Recall that they write

$$\overline{\varphi}_{2\ell}(x) = c_{2\ell} P_\ell^{(-\frac{1}{2}, \sigma)}(1 - 2x^2)$$

when $k = 2\ell$ is even and

$$\overline{\varphi}_{2\ell+1}(x) = c_{2\ell+1} P_\ell^{(\frac{1}{2}, \sigma)}(1 - 2x^2)$$

when $k = 2\ell + 1$ is odd, for $\ell \geqslant 0$, where

$$c_{2\ell}^2 = \frac{2^\sigma \ell! (2\ell + \sigma + \frac{1}{2}) \Gamma(\ell + \sigma + \frac{1}{2})}{\Gamma(\ell + \frac{1}{2}) \Gamma(\ell + \sigma + 1)}, \quad c_{2\ell+1}^2 = \frac{2^\sigma k! (2\ell + \sigma + \frac{3}{2}) \Gamma(\ell + \sigma + \frac{3}{2})}{\Gamma(\ell + \frac{3}{2}) \Gamma(\ell + \sigma + 1)},$$

(see Lemma 3.2.8 and (3.2.9)). In view of the fact that $B^*u = u|_\omega$ and since $L^2(a, b)$ and $\mathcal{H}^0(a, b)$ are topologically equivalent, (3.3.9) may be rewritten as

$$\int_a^b \overline{\varphi}_k^2 \,\mathrm{d}x \geqslant \mu$$

[2] Meaning there exists $\alpha > 0$ such that $\langle Ay, y \rangle_{\mathcal{H}} \leqslant -\alpha \|y\|_{\mathcal{H}}^2$ for all $y \in \mathcal{D}(A)$.

for some $\mu = \mu(a,b,\sigma) > 0$ independent of k. Now, for any fixed $k \geqslant 0$, since these eigenfunctions are nonzero solutions of a second order differential equation (they are nonzero polynomials), we have

$$\int_a^b \overline{\varphi}_k^2 \,\mathrm{d}x \geqslant \mu$$

for some $\mu = \mu(k,a,b,\sigma) > 0$. We are thus going to study the behavior of this quantity as $k \longrightarrow \infty$.

For technical purposes, let us first assume that $0 \notin (a,b)$. We will add 0 in (a,b) a posteriori, after observing that the asymptotic lower bound does not blow up as $a \longrightarrow 0$ or $b \longrightarrow 0$.

Let us assume that $a > 0$ (the cases $b < 0$ and $a < 0$, $b > 0$, $0 \notin (a,b)$ follow similar arguments). Let $k = 2\ell$ with $\ell \geqslant 0$ be even. We have

$$\int_a^b \overline{\varphi}_{2\ell}^2 \,\mathrm{d}x = c_{2\ell}^2 \int_a^b P_\ell^{(-\frac{1}{2},\sigma)}(1-2x^2)^2 \,\mathrm{d}x. \tag{3.3.13}$$

We look to reformulate the integral on the right-hand in view of using the following asymptotic formula:

$$P_n^{(\alpha,\beta)}(\cos\theta) = \frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{\pi}}\sin^{-\alpha-\frac{1}{2}}\left(\frac{\theta}{2}\right)\cos^{-\beta-\frac{1}{2}}\left(\frac{\theta}{2}\right)\cos(\theta\psi(n)-\phi)\right) + \mathcal{O}\big(n^{-\frac{3}{2}}\big), \tag{3.3.14}$$

for $n \geqslant 0$, $\alpha, \beta \in \mathbb{R}$ and $\theta \in (0,\pi)$, where

$$\psi(n) = \Big(n + \frac{1}{2}(\alpha+\beta+1)\Big) \quad \text{and} \quad \phi = \frac{\pi}{2}\Big(\alpha+\frac{1}{2}\Big),$$

see [252, Chapter VIII, Theorem 8.21.8] for instance. Performing the change of variable $\cos\theta = 1-2x^2$, whence $\mathrm{d}x = 2^{-\frac{3}{2}}\sqrt{1+\cos\theta}\,\mathrm{d}\theta$, gives

$$\int_a^b P_\ell^{(-\frac{1}{2},\sigma)}(1-2x^2)^2 \,\mathrm{d}x = 2^{-\frac{3}{2}}\int_{\gamma_1}^{\gamma_2} P_\ell^{(-\frac{1}{2},\sigma)}(\cos\theta)^2\sqrt{1+\cos\theta}\,\mathrm{d}\theta,$$

where $\gamma_1 = \arccos(1-2a^2)$ and $\gamma_2 = \arccos(1-2b^2)$, thus now $(\gamma_1,\gamma_2) \subset (0,\pi)$. We may use (3.3.14), which combined with the above identity gives

$$\begin{aligned}\int_a^b P_\ell^{(-\frac{1}{2},\sigma)}(1-2x^2)^2 \,\mathrm{d}x = 2^{-\frac{3}{2}}\int_{\gamma_1}^{\gamma_2} \frac{1}{\ell\pi}\frac{\cos^2(\theta(\ell+\frac{\sigma}{2}+\frac{1}{4}))}{\cos(\frac{\theta}{2})^{2\sigma+1}}\sqrt{1+\cos\theta}\,\mathrm{d}\theta \\ + \frac{2^{-\frac{1}{2}}}{\sqrt{\pi}}\int_{\gamma_1}^{\gamma_2}\mathcal{O}\left(\frac{1}{\ell^2}\right)\frac{\cos\left(\theta(\ell+\frac{\sigma}{2}+\frac{1}{4})\right)}{\cos(\frac{\theta}{2})^{\sigma+\frac{1}{2}}}\sqrt{1+\cos\theta}\,\mathrm{d}\theta \\ + \int_{\gamma_1}^{\gamma_2}\mathcal{O}\Big(\frac{1}{\ell^3}\Big)\sqrt{1+\cos\theta}\,\mathrm{d}\theta\end{aligned}$$

as $\ell \longrightarrow \infty$. Let us take a closer look at the right-hand side integrals. Using elementary trigonometric relations,

$$\begin{aligned}\int_{\gamma_1}^{\gamma_2}\frac{\cos^2(\theta(\ell+\frac{\sigma}{2}+\frac{1}{4}))}{\cos(\frac{\theta}{2})^{2\sigma+1}}\sqrt{1+\cos\theta}\,\mathrm{d}\theta &= 2^{\sigma+\frac{1}{2}}\int_{\gamma_1}^{\gamma_2}\frac{\cos^2(\theta(\ell+\frac{\sigma}{2}+\frac{1}{4}))}{(1+\cos\theta)^\sigma}\,\mathrm{d}\theta \\ &= 2^{\sigma-\frac{1}{2}}\int_{\gamma_1}^{\gamma_2}\frac{\big(1+\cos(\theta(2\ell+\sigma+\frac{1}{2}))\big)}{(1+\cos\theta)^\sigma}\,\mathrm{d}\theta.\end{aligned}$$

Similarly,

$$\mathcal{O}\left(\frac{1}{\ell^2}\right)\int_{\gamma_1}^{\gamma_2}\frac{\cos\left(\theta(\ell+\frac{\sigma}{2}+\frac{1}{4})\right)}{\cos(\frac{\theta}{2})^{\sigma+\frac{1}{2}}}\sqrt{1+\cos\theta}\,\mathrm{d}\theta = \mathcal{O}\left(\frac{1}{\ell^2}\right)\int_{\gamma_1}^{\gamma_2}\eta(\theta)\frac{\cos\left(\theta(\ell+\frac{\sigma}{2}+\frac{1}{4})\right)}{(1+\cos\theta)^{\frac{2\sigma-1}{4}}}\,\mathrm{d}\theta$$

where $\eta(\theta) = \operatorname{sgn}\big(\pi+\theta+4\pi\lfloor\frac{\pi-\theta}{4\pi}\rfloor\big) \in \{-1,1\}$. Putting together the three previous identities, we obtain

$$\begin{aligned}\int_a^b P_\ell^{(-\frac{1}{2},\sigma)}(1-2x^2)^2\,\mathrm{d}x &= \frac{2^{\sigma-2}}{\ell\pi}\int_{\gamma_1}^{\gamma_2}\frac{\big(1+\cos(\theta(2\ell+\sigma+\frac{1}{2}))\big)}{(1+\cos\theta)^\sigma}\,\mathrm{d}\theta\\ &\quad+\mathcal{O}\left(\frac{1}{\ell^2}\right)\int_{\gamma_1}^{\gamma_2}\eta(\theta)\frac{\cos\big(\theta(\ell+\frac{\sigma}{2}+\frac{1}{4})\big)}{(1+\cos\theta)^{\frac{2\sigma-1}{4}}}\,\mathrm{d}\theta\\ &\quad+\mathcal{O}\Big(\frac{1}{\ell^3}\Big)\end{aligned}$$

as $\ell\to\infty$. Going back to (3.3.13), we now have

$$\begin{aligned}\int_a^b \overline{\varphi}_{2\ell}^2\,\mathrm{d}x &= \frac{(\ell-1)!(2\ell+\sigma+\frac{1}{2})\Gamma(\ell+\sigma+\frac{1}{2})}{\Gamma(\ell+\frac{1}{2})\Gamma(\ell+\sigma+1)}\frac{2^{2\sigma-2}}{\pi}\int_{\gamma_1}^{\gamma_2}\frac{\big(1+\cos(\theta(2\ell+\sigma+\frac{1}{2}))\big)}{(1+\cos\theta)^\sigma}\,\mathrm{d}\theta \qquad (3.3.15)\\ &\quad+\frac{\ell!(2\ell+\sigma+\frac{1}{2})\Gamma(\ell+\sigma+\frac{1}{2})}{\Gamma(\ell+\frac{1}{2})\Gamma(\ell+\sigma+1)}\left(\mathcal{O}\left(\frac{1}{\ell^2}\right)\int_{\gamma_1}^{\gamma_2}\eta(\theta)\frac{\cos\big(\theta(\ell+\frac{\sigma}{2}+\frac{1}{4})\big)}{(1+\cos\theta)^{\frac{2\sigma-1}{4}}}\,\mathrm{d}\theta+\mathcal{O}\Big(\frac{1}{\ell^3}\Big)\right)\end{aligned}$$

as $\ell\longrightarrow\infty$. Making use of the relations $(\ell-1)!=\Gamma(\ell)$, $z\Gamma(z)=\Gamma(z+1)$ for $z\in\mathbb{C}$ as well as $\frac{\Gamma(\ell+\alpha)}{\Gamma(\ell+\beta)}\sim\ell^{\alpha-\beta}$ (a consequence of Stirling's formula), we obtain

$$\begin{aligned}\frac{(\ell-1)!}{\Gamma(\ell+\frac{1}{2})}\frac{(2\ell+\sigma+\frac{1}{2})\Gamma(\ell+\sigma+\frac{1}{2})}{\Gamma(\ell+\sigma+1)} &= \frac{\Gamma(\ell)}{\Gamma(\ell+\frac{1}{2})}\left(\ell\frac{\Gamma(\ell+\sigma+\frac{1}{2})}{\Gamma(\ell+\sigma+1)}+\frac{(\ell+\sigma+\frac{1}{2})\Gamma(\ell+\sigma+\frac{1}{2})}{\Gamma(\ell+\sigma+1)}\right)\\ &= \ell^{-\frac{1}{2}}\left(\ell\frac{\Gamma(\ell+\sigma+\frac{1}{2})}{\Gamma(\ell+\sigma+1)}+\frac{\Gamma(\ell+\sigma+\frac{3}{2})}{\Gamma(\ell+\sigma+1)}\right) \qquad (3.3.16)\\ &\sim 2,\end{aligned}$$

and similarly

$$\frac{\ell!(2\ell+\sigma+\frac{1}{2})\Gamma(\ell+\sigma+\frac{1}{2})}{\Gamma(\ell+\frac{1}{2})\Gamma(\ell+\sigma+1)}\sim 2\ell \qquad (3.3.17)$$

as $\ell\longrightarrow\infty$. Moreover, recall that for any interval $I\subseteq\mathbb{R}$, the sequence $\{\cos(n\cdot)\}_{n\in\mathbb{N}}$ converges weakly-$*$ to 0 in $L^\infty(I)$ as $n\longrightarrow\infty$ (an application of the Riemann-Lebesgue Lemma), meaning

$$\int_I\cos(nx)\phi(x)\,\mathrm{d}x\xrightarrow[n\to\infty]{}0 \quad \text{for all } \phi\in L^1(I). \qquad (3.3.18)$$

Since $(1+\cos(\cdot))^{-\sigma}\in L^1(\gamma_1,\gamma_2)$ and also $\eta(\theta)(1+\cos\theta)^{-\frac{2\sigma-1}{4}}\in L^1(\gamma_1,\gamma_2)$, using (3.3.16), (3.3.17) and (3.3.18) in (3.3.15), we deduce that

$$\int_a^b\overline{\varphi}_{2\ell}^2\,\mathrm{d}x\longrightarrow\frac{2^{2\sigma-1}}{\pi}\int_{\gamma_1}^{\gamma_2}\frac{1}{(1+\cos\theta)^\sigma}\,\mathrm{d}\theta$$

as $\ell\longrightarrow\infty$. A straightforward computation yields

$$\int_{\gamma_1}^{\gamma_2}\frac{1}{(1+\cos\theta)^\sigma}\,\mathrm{d}\theta=2^{1-\sigma}\int_a^b\frac{1}{(1-x^2)^{\sigma+\frac{1}{2}}}\,\mathrm{d}x,$$

thus we may conclude that

$$\int_a^b\overline{\varphi}_{2\ell}^2\,\mathrm{d}x\longrightarrow\frac{2^\sigma}{\pi}\int_a^b\frac{1}{(1-x^2)^{\sigma+\frac{1}{2}}}\,\mathrm{d}x \qquad (3.3.19)$$

as $\ell\longrightarrow\infty$.

The arguments differ very little when $k = 2\ell+1$ is odd, so we provide less detail. We have

$$\int_a^b \overline{\varphi}_{2\ell+1}^2 \,\mathrm{d}x = \frac{2^\sigma \ell!(2\ell+\sigma+\frac{3}{2})\Gamma(\ell+\sigma+\frac{3}{2})}{\Gamma(\ell+\frac{3}{2})\Gamma(\ell+\sigma+1)} \int_a^b x^2 P_\ell^{(\frac{1}{2},\sigma)}(1-2x^2)^2 \,\mathrm{d}x.$$

Applying the same change of variable as in the above computation yields

$$\int_a^b x^2 P_\ell^{(\frac{1}{2},\sigma)}(1-2x^2)^2 \,\mathrm{d}x = \frac{1}{2}\int_{\gamma_1}^{\gamma_2} P_\ell^{(\frac{1}{2},\sigma)}(1-2x^2)^2(1-\cos\theta)\sqrt{1+\cos\theta}\,\mathrm{d}\theta.$$

By virtue of the asymptotic formula (3.3.14) and elementary trigonometric identities, we now have

$$\begin{aligned}
\int_a^b x^2 P_\ell^{(\frac{1}{2},\sigma)}(1-2x^2)^2 \,\mathrm{d}x &= \frac{1}{2\ell\pi}\int_{\gamma_1}^{\gamma_2} \frac{\cos^2(\theta(\ell+\frac{\sigma}{2}+\frac{3}{4})-\frac{\pi}{2})}{\sin^2(\frac{\theta}{2})\cos(\frac{\theta}{2})^{2\sigma+1}}(1-\cos\theta)\sqrt{1+\cos\theta}\,\mathrm{d}\theta \\
&\quad + \mathcal{O}\left(\frac{1}{\ell^2}\right)\int_{\gamma_1}^{\gamma_2} \frac{\cos(\theta(\ell+\frac{\sigma}{2}+\frac{3}{4})-\frac{\pi}{2})}{\sin(\frac{\theta}{2})\cos(\frac{\theta}{2})^{\sigma+\frac{1}{2}}}(1-\cos\theta)\sqrt{1+\cos\theta}\,\mathrm{d}\theta \\
&\quad + \mathcal{O}\Big(\frac{1}{\ell^3}\Big) \\
&= \frac{2^\sigma}{\ell\pi}\int_{\gamma_1}^{\gamma_2} \frac{1+\cos(\theta(2\ell+\sigma+\frac{3}{2})-\pi)}{(1+\cos\theta)^\sigma}\,\mathrm{d}\theta \\
&\quad + \mathcal{O}\left(\frac{1}{\ell^2}\right)\int_{\gamma_1}^{\gamma_2} \sin\left(\theta\left(\ell+\frac{\sigma}{2}+\frac{3}{4}\right)\right)\frac{\sqrt{1-\cos\theta}}{(1+\cos\theta)^{\frac{2\sigma-1}{4}}}\eta(\theta)\,\mathrm{d}\theta \\
&\quad + \mathcal{O}\Big(\frac{1}{\ell^3}\Big)
\end{aligned}$$

as $\ell \longrightarrow \infty$, where $\eta(\theta) \in \{-1, 1\}$. Using the parity and periodicity of the cosine, we see that the computations reduce to an almost identical scenario as when k is even, and we may use (3.3.16), (3.3.17) and (3.3.18)[3] to deduce

$$\int_{-1}^1 \overline{\varphi}_{2\ell+1}^2 \,\mathrm{d}x \longrightarrow \frac{2^{\sigma+1}}{\pi}\int_a^b \frac{1}{(1-x^2)^{\sigma+\frac{1}{2}}}\,\mathrm{d}x \tag{3.3.20}$$

as $\ell \longrightarrow \infty$. As the limit bound in (3.3.19), (3.3.20) does not blow up as $a \to 0$, we may conclude the proof. □

We are now in a position to conclude the proof of Theorem 3.2.

Proof of Theorem 3.2. The conclusion follows from a well-known adaptation of the HUM method (Hilbert Uniqueness Method, [184, Chapitre 2]). We give brief details for the sake of completeness.

For fixed $\varepsilon > 0$, let us introduce the functional

$$J_{\varepsilon,\mathrm{obs}}(\zeta_T) = \frac{1}{2}\int_0^T\int_\omega \rho^\sigma|\zeta|^2\,\mathrm{d}x\,\mathrm{d}t + \int_{-1}^1 \rho^\sigma y_0\zeta(0,\cdot)\,\mathrm{d}x + \varepsilon\|\zeta_T\|_{\mathcal{H}^0}$$

for every $\zeta_T \in \mathcal{H}^0$, where ζ is the unique solution to the adjoint problem

$$\begin{cases} \partial_t\zeta + \rho^{-\sigma}\partial_x(\rho^{\sigma+1}\partial_x\zeta) = 0 & \text{in } (0,T)\times(-1,1) \\ (\rho^{\sigma+1}\partial_x\zeta)(t,\pm1) = 0 & \text{in } (0,T) \\ \zeta(T,x) = \zeta_T(x) & \text{in } (-1,1). \end{cases} \tag{3.3.21}$$

$J_{\varepsilon,\mathrm{obs}}$ can be shown to be strictly convex, continuous and coercive on $\mathcal{H}^0$ by virtue of the observability inequality (3.3.10) (which holds for solutions of (3.3.21) by Lemma

[3] Which also holds when $\cos(n\cdot)$ is replaced by $\sin(n\cdot)$.

3.3.4). $J_{\varepsilon,\mathrm{obs}}$ thus has a unique minimizer $\zeta_T^\varepsilon \in \mathcal{H}^0$ by the direct method. Following common practice, we introduce the control $u_\varepsilon = \zeta_\varepsilon \mathbf{1}_\omega$ where ζ_ε is the solution to (3.3.21) corresponding to ζ_T^ε. We denote by y_ε the solution to (3.3.4) associated to the control u_ε. Differentiating $J_{\varepsilon,\mathrm{obs}}$ at ζ_T^ε yields the Euler-Lagrange equation

$$\int_0^T \int_\omega \rho^\sigma \zeta_\varepsilon \varphi \,\mathrm{d}x\,\mathrm{d}t + \langle y_0, \varphi(0,\cdot)\rangle_{\mathcal{H}^0} + \varepsilon \left\langle \frac{\zeta_T^\varepsilon}{\|\zeta_T^\varepsilon\|_{\mathcal{H}^0}}, \varphi_T \right\rangle_{\mathcal{H}^0} = 0 \tag{3.3.22}$$

for every $\varphi_T \in \mathcal{H}^0$. (3.3.22) along with the observability inequality (3.3.10) for $\varphi_T = \zeta_T^\varepsilon$ give

$$\|u_\varepsilon\|_{L^2(0,T;L^2(\omega))} \lesssim C_{\mathrm{obs}} \|y_0\|_{\mathcal{H}^0}$$

uniformly in $\varepsilon > 0$. On the other hand, by duality

$$\int_0^T \int_\omega \rho^\sigma \zeta_\varepsilon \varphi \,\mathrm{d}x\,\mathrm{d}t = \langle y_\varepsilon(T,\cdot), \varphi_T\rangle_{\mathcal{H}^0} - \langle y_0, \varphi(0,\cdot)\rangle_{\mathcal{H}^0}, \tag{3.3.23}$$

which combined with (3.3.22) yields

$$\|y_\varepsilon(T,\cdot)\|_{\mathcal{H}^0} \leqslant \varepsilon. \tag{3.3.24}$$

Since $\{u_\varepsilon\}_{\varepsilon>0}$ is bounded in $L^2(0,T;L^2(\omega))$, it converges weakly (along a subsequence) to some $u \in L^2(0,T;L^2(\omega))$. Using analytic semigroup estimates such as [26, Theorem 2.12, Section 2] we deduce that along subsequences

$$y_\varepsilon \rightharpoonup y \qquad \text{weakly in } L^2(0,T;\mathcal{H}^1) \cap H^1(0,T;(\mathcal{H}^1)^*)$$

as $\varepsilon \longrightarrow 0$, where y is the solution to (3.3.4) with control u. By Aubin-Lions, this gives strong convergence of $\{y_\varepsilon(t,\cdot)\}_{\varepsilon>0}$ to $y(t,\cdot)$ in $\mathcal{H}^0$ for all $t \in [0,T]$, whence $y(T,\cdot) = 0$ in view of (3.3.24). □

3.3.2 Controllability in spite of a source term

The null-controllability of Problem (3.1.8) will follow by combining Theorem 3.2 with the source-term method. We review the latter in what follows.

Let $\gamma : (0,\infty) \longrightarrow [0,\infty)$ be a continuous, non-increasing function satisfying $\lim_{t\to 0}\gamma(t) = \infty$ and

$$\kappa(t) < \gamma(t) \quad \text{for all } t > 0.$$

We moreover assume that for some $M_1, M_2 > 0$,

$$\gamma(T) = M_1 e^{\frac{M_2}{T}} \quad \text{for } T \ll 1. \tag{3.3.25}$$

One may for instance consider the observability constant $t \longmapsto C_{\mathrm{obs}}(t)$ in (3.3.10), which satisfies the above assumptions, as per [242, Theorem 1 & Section 5.2] (see also [256]). We recall that in our case, condition (3.3.25) does not hold when the time horizon is large, as $\lambda_0 = 0$ (see Remark 3.3.3).

Let $T > 0$, $q \in (1,\sqrt{2})$ and $p > 0$ such that $2p > (1+p)q^2$ be fixed. We now consider the continuous, non-increasing function $\theta_{\mathcal{F}} : [0,T] \longrightarrow [0,\infty)$ defined by

$$\theta_{\mathcal{F}}(t) = \gamma\Big(\frac{q-1}{q^2}(T-t)\Big)^{-(p+1)} \quad \text{for } t \in [0,T]. \tag{3.3.26}$$

As $p > 0$, it is easily seen that $\theta_{\mathcal{F}}(T) = 0$. We then consider the continuous and non-increasing function $\theta_0 : \Big[T(1-q^{-2}), T\Big] \longrightarrow [0,\infty)$ defined by

$$\theta_0(t) = \theta_{\mathcal{F}}\big(q^2(t-T)+T\big)\gamma\big((q-1)(T-t)\big) \quad \text{for } t \in \Big[T(1-q^{-2}), T\Big], \tag{3.3.27}$$

which also satisfies $\theta_0(T) = 0$. We extend θ_0 (and use the same notation) to a continuous, non-increasing function on $[0, T]$ by setting

$$\theta_0(t) = \theta_0\Big(T\big(1 - q^{-2}\big)\Big) \quad \text{for } t \in \Big[0, T\big(1 - q^{-2}\big)\Big].$$

When dealing with the nonlinear problem, it will be important that the above-defined weights satisfy the condition

$$\frac{\theta_0^2}{\theta_{\mathcal{F}}} \in C^0([0, T]). \tag{3.3.28}$$

This is accomplished by the choice of $q > 1$ and $p > 0$ above. Indeed, notice that the only obstruction for having (3.3.28) is the behavior of this quotient near $t = T$, as one may see that

$$\frac{\theta_0^2}{\theta_{\mathcal{F}}}(t) = \frac{\gamma\Big(\frac{q-1}{q^2}(T-t)\Big)^{(p+1)}}{\gamma((q-1)(T-t))^{2p}}.$$

Thus, in view of condition (3.3.25), the choice $q \in (1, \sqrt{2})$ and $2p > (1+p)q^2$ has the effect of guaranteeing (3.3.28).

Remark 3.3.5. *Should $\lambda_0 > 0$, one may for instance consider the explicit weights*

$$\theta_{\mathcal{F}}(t) = e^{-\frac{\alpha}{(T-t)^2}}, \quad \theta_0(t) = M_1 e^{\frac{M_2}{(q-1)(T-t)} - \frac{\alpha}{q^4(T-t)^2}}$$

as in [191], where α, q are appropriately chosen for the fixed-point argument.

To the time-weight functions $\theta_0, \theta_{\mathcal{F}}$, we associate the time-weighted Hilbert spaces

$$\begin{aligned} \mathcal{F} &= \left\{ f \in L^2(0, T; \mathcal{H}^0) \colon \frac{f}{\theta_{\mathcal{F}}} \in L^2(0, T; \mathcal{H}^0) \right\}, \\ \mathcal{U} &= \left\{ u \in L^2(0, T; L^2(\omega)) \colon \frac{u}{\theta_0} \in L^2(0, T; L^2(\omega)) \right\}. \end{aligned} \tag{3.3.29}$$

The following Theorem is originally shown in [191, Proposition 2.3] (see also [173] and [23] for subsequent adaptations). We assume higher regularity for the initial datum a priori, and thus for the controlled trajectory, having in mind the fixed-point argument. For the sake of completeness, we give a proof below, and the proof follows the same time-splitting scheme of [191].

Theorem 3.4. *Let $T > 0$. There exists a constant $C = C(T) > 0$ and a continuous linear map $\mathfrak{L} : \mathcal{H}^1 \times \mathcal{F} \longrightarrow \mathcal{U}$ such that for any $y_0 \in \mathcal{H}^1$ and any $f \in \mathcal{F}$, the solution y of (3.1.8) with control $u = \mathfrak{L}(y_0, f)$ satisfies*

$$\left\| \frac{y}{\theta_0} \right\|_{C^0([0,T];\mathcal{H}^1)} + \left\| \frac{y}{\theta_0} \right\|_{L^2(0,T;\mathcal{H}^2)} + \|u\|_{\mathcal{U}} \leqslant C(\|f\|_{\mathcal{F}} + \|y_0\|_{\mathcal{H}^1}). \tag{3.3.30}$$

In particular, since θ_0 is a continuous function satisfying $\theta_0(T) = 0$, the above relation yields $y(T, \cdot) = 0$.

Proof. For $k \in \mathbb{N}$, we define $T_k := T(1 - q^{-k})$: On one hand, we set $a_0 := y_0$ and, for $k \in \mathbb{N}$, we define $a_{k+1} := y_f(T_{k+1}^-, \cdot)$ where y_f is the solution to

$$\begin{cases} \partial_t y_f + \mathcal{A} y_f = f & \text{on } (T_k, T_{k+1}) \\ y_f(T_k^+, \cdot) = 0. \end{cases}$$

From the energy estimate (3.3.2) in Proposition 3.3.1, we have

$$\|a_{k+1}\|_{\mathcal{H}^1} \leqslant \|y_f\|_{C^0([T_k, T_{k+1}];\mathcal{H}^1)} \leqslant C_T \|f\|_{L^2(T_k, T_{k+1};\mathcal{H}^0)}. \tag{3.3.31}$$

On the other hand, for $k \in \mathbb{N}$ we consider the homogeneous control system

$$\begin{cases} \partial_t y_u + \mathcal{A} y_u = u_k \mathbf{1}_\omega & \text{on } (T_k, T_{k+1}) \\ y(T_k^+, \cdot) = a_k, \end{cases}$$

where $u_k \in L^2(T_k, T_{k+1}; L^2(\omega))$ is such that $y_u(T_{k+1}^+, \cdot) = 0$ and

$$\|u_k\|^2_{L^2(T_k,T_{k+1};L^2(\omega))} \leqslant \gamma^2 (T_{k+1} - T_k) \|a_k\|^2_{\mathcal{H}^0}. \tag{3.3.32}$$

Such a u_k exists for any $k \in \mathbb{N}$ by virtue of Theorem 3.2. Now remark that by definition of the weights, one has

$$\theta_0(T_{k+2}) = \theta_{\mathcal{F}}(T_k)\gamma(T_{k+2} - T_{k+1}) \tag{3.3.33}$$

for $k \in \mathbb{N}$. Now combining (3.3.32), (3.3.31), and the fact that $\theta_{\mathcal{F}}$ is a non-increasing function, we obtain

$$\begin{aligned} \|u_{k+1}\|^2_{L^2(T_{k+1},T_{k+2};L^2(\omega))} &\leqslant \gamma^2 (T_{k+2} - T_{k+1}) \|a_{k+1}\|^2_{\mathcal{H}^0} \\ &\leqslant C_T^2 \gamma^2 \left((q-1) \frac{T}{q^{k+2}} \right) \theta^2_{\mathcal{F}}(T_k) \left\| \frac{f}{\theta_{\mathcal{F}}} \right\|^2_{L^2(T_k,T_{k+1};\mathcal{H}^0)} \end{aligned}$$

for any $k \in \mathbb{N}$. In view of the definition of θ_0 and the relation (3.3.33), we deduce that

$$\|u_{k+1}\|^2_{L^2(T_{k+1},T_{k+2};L^2(\omega))} \leqslant C_T^2 \theta_0^2(T_{k+2}) \left\| \frac{f}{\theta_{\mathcal{F}}} \right\|^2_{L^2(T_k,T_{k+1},\mathcal{H}^0)}.$$

Finally, since θ_0 is a non-increasing function, there exists a constant $C = C(T) > 0$ such that

$$\left\| \frac{u_{k+1}}{\theta_0} \right\|^2_{L^2(T_{k+1},T_{k+2};L^2(\omega))} \leqslant C \left\| \frac{f}{\theta_{\mathcal{F}}} \right\|^2_{L^2(T_k,T_{k+1};\mathcal{H}^0)} \tag{3.3.34}$$

for all $k \in \mathbb{N}$. We can now patch the controls u_k for $k \in \mathbb{N}$ all together by defining

$$u := \sum_{k=0}^{\infty} u_k \mathbf{1}_{[T_k,T_{k+1})}.$$

In particular, combining estimates (3.3.34) and (3.3.32) (with $k = 0$) yields a constant $C = C(T) > 0$ such that

$$\left\| \frac{u}{\theta_0} \right\|_{L^2(0,T;L^2(\omega))} \leqslant C \left(\left\| \frac{f}{\theta_{\mathcal{F}}} \right\|_{L^2(0,T;\mathcal{H}^0)} + \|y_0\|_{\mathcal{H}^1} \right),$$

for any $y_0 \in \mathcal{H}^1$ and any $f \in \mathcal{F}$. The state y can also be reconstructed by concatenation, namely $y = y_f + y_u$, continuous at each junction by construction. Indeed,

$$y(T_k^-, \cdot) = y_f(T_k^-, \cdot) + y_u(T_k^-, \cdot) = a_k = y_f(T_k^+, \cdot) + y_u(T_k^+, \cdot) = y(T_k^+, \cdot),$$

and so y satisfies (3.1.8). We now look to estimate the state y. We use the energy estimate (3.3.2) from Proposition 3.3.1 on each time interval to obtain

$$\|y_f\|^2_{C^0([T_k,T_{k+1}];\mathcal{H}^1)} + \|y_f\|^2_{L^2(T_k,T_{k+1};\mathcal{H}^2)} \leqslant C_T^2 \|f\|^2_{L^2(T_k,T_{k+1};\mathcal{H}^0)} \tag{3.3.35}$$

and

$$\begin{aligned} &\|y_u\|^2_{C^0([T_k,T_{k+1}];\mathcal{H}^1)} + \|y_u\|^2_{L^2(T_k,T_{k+1};\mathcal{H}^2)} \\ &\qquad \leqslant C_T^2 \|a_k\|^2_{\mathcal{H}^1} + C_T^2 \|u_k\|_{L^2(T_k,T_{k+1};L^2(\omega))} \end{aligned} \tag{3.3.36}$$

for $k \in \mathbb{N}$. Proceeding similarly as for estimating the control, we may deduce

$$\|y\|^2_{C^0([T_k,T_{k+1}];\mathcal{H}^1)} + \|y\|^2_{L^2(T_k,T_{k+1};\mathcal{H}^2)} \leqslant C_T^2 \theta_0^2(T_{k+1}) \left\| \frac{f}{\theta_{\mathcal{F}}} \right\|^2_{L^2(T_{k-1},T_{k+1};\mathcal{H}^0)}$$

for $k \geqslant 1$, and since θ_0 is a non-increasing function, using (3.3.32), (3.3.35), (3.3.36) (all for $k = 0$) we deduce

$$\left\| \frac{y}{\theta_0} \right\|_{C^0([0,T];\mathcal{H}^1)} + \left\| \frac{y}{\theta_0} \right\|_{L^2(0,T;\mathcal{H}^2)} \leqslant C \left(\left\| \frac{f}{\theta_{\mathcal{F}}} \right\|_{L^2(0,T;\mathcal{H}^0)} + \|y_0\|_{\mathcal{H}^1} \right).$$

In view of the above estimate and (3.3.34), we may conclude. □

3.4 The fixed-point argument

We look to conclude the proof of Theorem 3.1 by virtue of a Banach fixed point argument. Let $X_T := L^2(0,T;\mathcal{H}^2) \cap C^0([0,T];\mathcal{H}^1)$, and consider the time-weighted space

$$\mathcal{Y} := \left\{ y \in X_T \colon \frac{y}{\theta_0} \in X_T \right\},$$

which is endowed with the Hilbert norm

$$\|y\|^2_{\mathcal{Y}} := \int_0^T \theta_0^{-2}(t) \|y(t,\cdot)\|^2_{X_T} \, \mathrm{d}t.$$

We recall that the weights $\theta_0, \theta_{\mathcal{F}}$ are defined in (3.3.27) and (3.3.26) respectively. Let us also denote

$$M := \sup_{t \in [0,T]} \frac{\theta_0^2}{\theta_{\mathcal{F}}}(t),$$

which is finite due to (3.3.28), and consider the radius

$$r := \min \left\{ \frac{1}{2C(T)}, \frac{1}{8C(T)MC_\sigma} \right\}, \tag{3.4.1}$$

where $C(T) > 0$ is the constant appearing in the control-continuity estimate (3.3.30) and $C_\sigma > 0$ appears in the embedding given by Lemma 3.2.4 (see (3.4.3) below). We also consider the ball

$$\mathcal{Y}_r := \left\{ y \in \mathcal{Y} \colon \|y\|_{\mathcal{Y}} \leqslant r \right\}$$

Given $y_0 \in \mathcal{H}^1$, we may construct the nonlinear map $\mathcal{N} : \mathcal{Y}_r \longrightarrow \mathcal{Y}_r$ by setting

$$\mathcal{N}(\overline{y}) := y,$$

where y is the solution to the controlled problem

$$\begin{cases} \partial_t y - \rho^{-\sigma} \partial_x (\rho^{\sigma+1} \partial_x y) = \rho F_{\varepsilon,\delta}(\overline{y}, \partial_x \overline{y}) + u \mathbf{1}_\omega & \text{in } (0,T) \times (-1,1) \\ (\rho^{\sigma+1} \partial_x y)(t, \pm 1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1), \end{cases}$$

where we recall (see (3.1.6)) that $F_{\varepsilon,\delta}(p,q) = \frac{q^2}{1+p+xq}$ when $p^2 < \delta^2, q^2 < \varepsilon^2$, and $F_{\varepsilon,\delta} = 0$ whenever $p^2 \geqslant 4\delta^2$ or $q^2 \geqslant 4\varepsilon^2$. Also, we assumed that $4(\varepsilon + \delta) < 1$. We are now in a position to prove our main result.

Proof of Theorem 3.1. For the sake of cohesion, we split the proof in three steps.

First step. Fix $y_0 \in \mathcal{H}^1$. We first look to show that the map $\mathcal{N}$ is well-defined and leaves $\mathcal{Y}_r$ invariant. Given $\overline{y} \in \mathcal{Y}_r$ we consider the source term

$$\overline{f} := \rho F_{\varepsilon,\delta}(\overline{y}, \partial_x \overline{y}).$$

Let us first demonstrate that $\overline{f} \in \mathcal{F}$, with $\mathcal{F}$ being defined in (3.3.29). Recall that

$$F_{\varepsilon,\delta}(\overline{y}, \partial_x \overline{y}) = \chi\left(\frac{\overline{y}^2}{\delta^2}\right) \chi\left(\frac{(\partial_x \overline{y})^2}{\varepsilon^2}\right) \frac{(\partial_x \overline{y})^2}{1 + \overline{y} + x\partial_x \overline{y}},$$

where $\chi : [0, \infty) \longrightarrow [0, 1]$ is a smooth cut-off supported on $[0, 4)$ with $\chi(x) \equiv 1$ on $[0, 1]$. Since $F_{\varepsilon,\delta} \not\equiv 0$ if and only if $|\overline{y}| \leqslant 2\delta$ and $|\partial_x \overline{y}| \leqslant 2\varepsilon$ where $4(\varepsilon + \delta) < 1$ (and thus $2(\varepsilon + \delta) < \frac{1}{2}$), using the triangle inequality we have

$$\begin{aligned} |F_{\varepsilon,\delta}(\overline{y}, \partial_x \overline{y})| \leqslant \frac{(\partial_x \overline{y})^2}{|1 + \overline{y} + x\partial_x \overline{y}|} \leqslant \frac{(\partial_x \overline{y})^2}{1 - |\overline{y}| - |\partial_x \overline{y}|} \leqslant \frac{(\partial_x \overline{y})^2}{1 - 2(\varepsilon + \delta)} \\ \leqslant 2(\partial_x \overline{y})^2. \end{aligned}$$

Whence,

$$\begin{aligned} \int_0^T \int_{-1}^1 \rho^\sigma \frac{\overline{f}^2}{\theta_{\mathcal{F}}^2} \,\mathrm{d}x \,\mathrm{d}t &\leqslant 4 \int_0^T \int_{-1}^1 \rho^{\sigma+2} \frac{\theta_0^4}{\theta_{\mathcal{F}}^2} \frac{(\partial_x \overline{y})^4}{\theta_0^4} \,\mathrm{d}x \,\mathrm{d}t \\ &\leqslant 4M^2 \int_0^T \int_{-1}^1 \rho^{\sigma+2} \frac{(\partial_x \overline{y})^4}{\theta_0^4} \,\mathrm{d}x \,\mathrm{d}t. \end{aligned} \tag{3.4.2}$$

We now recall that from Lemma 3.2.4, the embedding

$$\left\| \rho^{\frac{1}{2}} \partial_x \overline{y} \right\|_{C^0([-1,1])} \leqslant C_\sigma \|\overline{y}\|_{\mathcal{H}^2} \tag{3.4.3}$$

holds for some $C_\sigma > 0$ whenever $\sigma \in (-1, 0)$. As moreover $\overline{y}\,\theta_0^{-1} \in C^0([0, T]; \mathcal{H}^1)$, going back to (3.4.4) we may apply estimate (3.4.3) to the effect of

$$\begin{aligned} \int_0^T \int_{-1}^1 \rho^{\sigma+2} \frac{(\partial_x \overline{y})^4}{\theta_0^4} \,\mathrm{d}x \,\mathrm{d}t &\leqslant C_\sigma^2 \int_0^T \theta_0^{-2} \|\overline{y}\|_{\mathcal{H}^2}^2 \int_{-1}^1 \rho^{\sigma+1} \frac{(\partial_x \overline{y})^2}{\theta_0^2} \,\mathrm{d}x \,\mathrm{d}t \\ &\leqslant C_\sigma^2 \left(\int_0^T \theta_0^{-2} \|\overline{y}\|_{\mathcal{H}^2}^2 \,\mathrm{d}t \right) \left(\sup_{t \in [0,T]} \int_{-1}^1 \rho^{\sigma+1} \frac{(\partial_x \overline{y})^2}{\theta_0^2} \,\mathrm{d}x \right) \\ &\leqslant C_\sigma^2 \left(\int_0^T \theta_0^{-2} \|\overline{y}(t, \cdot)\|_{X_T}^2 \,\mathrm{d}t \right)^2. \end{aligned} \tag{3.4.4}$$

Combining estimates (3.4.2) and (3.4.4), we deduce

$$\left\| \frac{\overline{f}}{\theta_{\mathcal{F}}} \right\|_{L^2(0,T;\mathcal{H}^0)} \leqslant 2MC_\sigma \|\overline{y}\|_{\mathcal{Y}}^2, \tag{3.4.5}$$

and so $\overline{f} \in \mathcal{F}$. Now, let $u := \mathfrak{L}(y_0, \overline{f})$, which is well-defined by Theorem 3.4, and consider the corresponding controlled trajectory $y \in \mathcal{Y}$. We aim to show that $y \in \mathcal{Y}_r$. From the control-continuity estimate (3.3.30) we have

$$\|y\|_{\mathcal{Y}} \leqslant C(T) \left(\left\| \frac{\overline{f}}{\theta_{\mathcal{F}}} \right\|_{L^2(0,T;\mathcal{H}^0)} + \|y_0\|_{\mathcal{H}^1} \right).$$

Inequality (3.4.5) leads us to

$$\|y\|_{\mathcal{Y}} \leqslant C(T) \Big(2M_0 C_\sigma \|\overline{y}\|_{\mathcal{Y}}^2 + \|y_0\|_{\mathcal{H}^1} \Big).$$

In view of (3.4.1), choosing $\|y_0\|_{\mathcal{H}^1} \leqslant r$ leads us to conclude that $y \in \mathcal{Y}_r$.

Second step. Let us now demonstrate that the map $\mathcal{N} : \mathcal{Y}_r \longrightarrow \mathcal{Y}_r$ is strictly contractive. Observe that for $x \in (-1,1)$, $(p_i, q_i) \in \mathbb{R}$ satisfying $p_i^2 < \delta^2 < 1$ and $q_i^2 < \varepsilon^2 < 1$, $i = 1, 2$, one has

$$\begin{aligned} |F_{\varepsilon,\delta}(p_1,q_1) - F_{\varepsilon,\delta}(p_2,q_2)| &\leqslant 4\big(q_1^2(1+p_2+xq_2) - q_2^2(1+p_1+xq_1)\big) \\ &\leqslant 4\big((1+p_1+q_1)(q_1^2-q_2^2) + q_1^2(p_2-p_1) + q_1^2(q_2-q_1)\big) \\ &\leqslant 6(q_1^2-q_2^2) + 4q_1(p_2-p_1) + 4q_1(q_2-q_1). \end{aligned} \tag{3.4.6}$$

Hence, using estimates (3.3.30), (3.4.6) and arguing as in Step 1, we may see that

$$\begin{aligned} \Big\|\mathcal{N}(y_1) - \mathcal{N}(y_2)\Big\|_{\mathcal{Y}} &\leqslant C(T)\Big\|\rho\Big(F_{\varepsilon,\delta}(y_1,\partial_x y_1) - F_{\varepsilon,\delta}(y_2,\partial_x y_2)\Big)\Big\|_{\mathcal{F}} \\ &\leqslant 4C(T)C_\sigma M r \left\|y_1 - y_2\right\|_{\mathcal{Y}}. \end{aligned}$$

In view of (3.4.1), we deduce that $\mathcal{N}$ is a strict contraction.

Third step. Thanks to the Banach fixed point theorem, for any $y_0 \in \mathcal{H}^1$ satisfying $\|y_0\|_{\mathcal{H}^1} \leqslant r$, the nonlinear operator $\mathcal{N} : \mathcal{Y}_r \longrightarrow \mathcal{Y}_r$ admits a unique fixed point $y \in \mathcal{Y}_r$. We may thus conclude the proof of Theorem 3.1. □

3.5 Null-controllability of the linearized thin-film equation

We give brief arguments as to see that the controllability study in Section 3.2 may also be applied to the one-dimensional thin-film equation linearized around its self-similar profile, derived in [205, 245]. The *thin-film equation*

$$\partial_t h + \partial_z(h^n \partial_z^3 h) = 0 \quad \text{in } \{h > 0\}$$

where $n \in (0,3)$ represents a more realistic model for the evolution of a liquid film over a solid substrate in a regime known as lubrication approximation. Much like its second order counterpart, the PME (3.1.1), it is a free boundary problem whenever the initial datum is compactly supported (a physical phenomenon known as *droplets*). We refer to [117, 118, 245] and the references therein for an overview of the well-posedness results, self-similar asymptotics and the role of boundary conditions.

For $n = 1$ (known as linear mobility regime), McCann & Seis [205, 245] replicate the ideas used for the PME in [78, 243, 244] to compute the spectrum of the full linearization of the thin-film equation around its own self-similar (Smyth-Hill) solution. Namely, after an analog rescaling and von Mises transformation, the control problem for the equation linearized around the self-similar solution is of the form

$$\begin{cases} \partial_t y + \mathcal{A}^2 y + \mathcal{A} y = u\mathbf{1}_\omega & \text{in } (0,T)\times(-1,1) \\ (\rho y)(t,\pm 1) = (\rho^2 \partial_x y)(t,\pm 1) = 0 & \text{in } (0,T) \\ y(0,x) = y_0(x) & \text{in } (-1,1). \end{cases} \tag{3.5.1}$$

where $T > 0$ and $\mathcal{A} = -\rho^{-1}\partial_x(\rho^2\partial_x)$ is the operator governing the linearized porous medium equation (3.3.4) with $\sigma = 1$. Replicating the linear theory from Section 3.2, we may deduce that the operator $\mathcal{L} = \mathcal{A}(\mathcal{A} + \mathrm{Id})$ is self-adjoint, non-negative with domain $\mathcal{D}(\mathcal{L}) = \mathcal{H}^4$, and has compact resolvents. Both boundary conditions are automatically satisfied by arguing as in Lemma 3.2.5. The operator thus generates an analytic semigroup on $\mathcal{H}^0$, which implies the following result (see [245, Lemma 3]).

Proposition 3.5.1. *Let $T > 0$. For any $y_0 \in \mathcal{H}^0$ and $u \in L^2(0,T;\mathcal{H}^0)$, there exists a unique solution $y \in L^2(0,T;\mathcal{H}^2) \cap C^0([0,T];\mathcal{H}^0)$ to (3.5.1).*

As done in Section 3.3, we use the explicit spectrum of the linearized operator $\mathcal{A}(\mathcal{A}+\mathrm{Id})$, given in [205], to demonstrate the null-controllability of (3.5.1). This is in essence an immediate consequence of Theorem 3.3.

Corollary 3.5.2 ([205], Corollary 3)**.** *The spectrum of $\mathcal{L} : \mathcal{H}^4 \longrightarrow \mathcal{H}^0$ consists of simple nonnegative eigenvalues $\{\mu_k\}_{k=0}^{\infty}$, given by*

$$\mu_k = \lambda_k^2 + \lambda_k$$

for $k \geqslant 0$, where λ_k denote the eigenvalues of $\mathcal{A}$ from Theorem 3.3. Moreover, $\mathcal{L}$ and $\mathcal{A}$ have the same normalized eigenfunctions $\{\overline{\varphi}_k\}_{k=0}^{\infty}$, which generate an orthonormal basis of $\mathcal{H}^0$.

As the eigenfunctions of $\mathcal{L} = \mathcal{A}(\mathcal{A} + \mathrm{Id})$ and $\mathcal{A}$ coincide, and the control operator B is the same as in Section 3.3, we may deduce the following null-controllability result for Problem (3.5.1).

Theorem 3.5. *Let $T > 0$, $\omega \subsetneq (-1, 1)$ be an open, non-empty interval, and $\sigma = 1$. Then, for any $y_0 \in \mathcal{H}^0$, there exists a control $u \in L^2(0, T; L^2(\omega))$ such that the unique solution $y \in L^2(0, T; \mathcal{H}^2) \cap C^0([0, T]; \mathcal{H}^0)$ of (3.5.1) satisfies $y(0, \cdot) = y_0$ and $y(T, \cdot) = 0$.*

Proof. In view of Lemmas 3.3.2 and 3.3.4, and since the eigenfunctions of $\mathcal{A}$ and $\mathcal{L}$ coincide, we only need to investigate the eigenvalues $\{\mu_k\}_{k=0}^{\infty}$ of the latter operator. Due to their form, it is readily seen the eigenvalues satisfy both the growth condition (3.3.8) and separation condition (3.3.7). We may thus conclude by using the HUM method as for the proof of Theorem 3.2. □

3.6 Concluding remarks

In this work, we addressed the null controllability a one-dimensional degenerate parabolic equation, which represents the problem satisfied by a perturbation around the Barenblatt profile of the free boundary porous medium equation after fixing the moving domain. We proved a local null-controllability result for this perturbed problem with a regularized version of the original nonlinearity. This allowed us to make use of only $L^2(0, T; L^2(\omega))$-regular controls for the fixed-point argument. The linear controllability theory was also applied for proving a null-controllability result for the linearized thin-film equation.

Let us present some directions on how our results may be extended, as well as some related open problems.

3.6.1 The full nonlinearity and free boundary problem

In this work we only addressed the case where the nonlinearity

$$\mathcal{N}(y) = \rho F(y, \partial_x y) - \rho^{-\sigma}\partial_x(\rho^{\sigma+1} x F(y, \partial_x y)), \quad F(y, \partial_x y) = \frac{(\partial_x y)^2}{1 + y + x\partial_x y}$$

is truncated as in (3.1.6) (and without the higher order term $-\rho^{-\sigma}\partial_x(\rho^{\sigma+1} x F_{\varepsilon,\delta})$). To derive a local null-controllability result for the full nonlinear perturbation equation (3.1.4), one would need to remove the cut-off, i.e. to ensure that $F_{\varepsilon,\delta} \equiv F$. To ensure this condition, higher regularity of the controlled trajectory y and consequently of the control u is needed. Namely, y should be $C^{0,1}([0, T] \times [-1, 1])$ and have small-enough norm. This will be considered in a future work. The Lipschitz regularity of the controlled trajectory is also required for reversing the von Mises transformation in view of obtaining the local controllability to the Barenblatt trajectory for the free boundary problem (see Remark 3.7.2).

3.6.2 The multi-dimensional case

The null-controllability of perturbation equation in arbitrary dimension is also worth investigating. In the linearized regime, it would read

$$\begin{cases} \partial_t y - \rho^{-\sigma}\nabla\cdot(\rho^{\sigma+1}\nabla y) = u\mathbf{1}_\omega & \text{in } (0,T)\times B_1 \\ (\rho^{\sigma+1}\partial_n y)(t,\pm 1) = 0 & \text{in } (0,T), \end{cases}$$

where B_1 is the open unit ball, $\rho(x) = \frac{1}{2}(1-|x|^2)$ and $\omega \subsetneq B_1$ is open and non-empty. The well-posedness follows from similar arguments as in the one-dimensional case, and is also argued in [244]. The spectral/Fourier techniques we used in this work are however restricted to the one-dimensional case. Thus, proving the desired observability inequality would likely require a Carleman estimate in the weighted $\mathcal{H}^k$ spaces. To the best of our knowledge, this has not been addressed in the literature.

3.6.3 The thin-film equation

The preceding questions are also open in the case of the (perturbed) thin-film equation of Section 3.5, for which we have only addressed the null-controllability of the one-dimensional, linearized equation.

3.7 Appendix

3.7.1 Hardy-type inequalities

Herein, we provide short proofs of Lemmas 3.2.1 and 3.2.2.

Proof of Lemma 3.2.1. Set $x = \tanh s$, and consider $g(s) = (1-\tanh^2 s)^\alpha f(\tanh s)$. Then we have $\|(1-x^2)^\alpha f\|_{C^0([-1,1])} = \|g\|_{L^\infty(\mathbb{R})}$. Along with the standard Sobolev embedding $\|g\|_{L^\infty(\mathbb{R})} \leqslant \|g\|_{H^1(\mathbb{R})}$ and $\frac{\mathrm{d}x}{\mathrm{d}s} = 1-x^2$, this yields

$$\begin{aligned} \|(1-x^2)^\alpha f\|^2_{C^0([-1,1])} &\leqslant \int_{\mathbb{R}} (g^2 + (\partial_s g)^2)\,\mathrm{d}s \\ &= \int_{-1}^1 \big((1-x^2)^{2\alpha-1} f^2 + (1-x^2)(\partial_x(1-x^2)^\alpha f)^2\big)\,\mathrm{d}x. \end{aligned}$$

Using the elementary estimate $(a-b)^2 \leqslant 2a^2+2b^2$, we conclude

$$\|(1-x^2)^\alpha f\|^2_{C^0([-1,1])} \leqslant (1+8\alpha^2)\int_{-1}^1 (1-x^2)^{2\alpha-1} f^2 + 2\int_{-1}^1 (1-x^2)^{2\alpha+1}(\partial_x f)^2.$$

□

We recall the following Hardy inequality, and refer to [117, Lemma A.1] for a proof (see [137] for the original).

Lemma 3.7.1 (Hardy)**.** *Let $\alpha \neq \frac{1}{2}$, and let $\|x^{\alpha+1}\partial_x f\|_{L^2(\mathbb{R}_+)} < \infty$. Suppose that $f(x_k) \longrightarrow 0$ for some sequence $x_k \longrightarrow c$ as $k\to\infty$, where $c=0$ if $\alpha < -\frac{1}{2}$ and $c=\infty$ if $\alpha > -\frac{1}{2}$. Then*

$$\|x^\alpha f\|_{L^2(\mathbb{R}_+)} \leqslant \frac{2}{|2\alpha+1|}\|x^{\alpha+1}\partial_x f\|_{L^2(\mathbb{R}_+)}. \tag{3.7.1}$$

One may choose f such that $f(x) = \log(\log\frac{1}{x})$ near $x=0$ to show that the assumption $\alpha \neq -\frac{1}{2}$ is necessary.

Proof of Lemma 3.2.2. The proof follows similar arguments to those for Lemma 3.2.1. We begin by writing

$$\int_{-1}^{1}(1-x^2)^\alpha f^2\,\mathrm{d}x=\int_{-1}^{0}(1-x^2)^\alpha f^2\,\mathrm{d}x+\int_{0}^{1}(1-x^2)^\alpha f^2\,\mathrm{d}x. \tag{3.7.2}$$

As both terms on the right-hand side of the identity (3.7.2) are symmetric, we will only look at the first one. Let $\eta\in C^\infty(\mathbb{R})$ be a cut-off function with $\eta(x)\equiv 1$ for $x\leqslant 0$ and $\eta(x)\equiv 0$ for $x\geqslant\frac{1}{2}$. Also, set $g(s)=f(x)\eta(x)$ for $s=1+x$. As $1\leqslant(1-x)\leqslant 2$ in $(-1,0)$, we have

$$\int_{-1}^{0}(1-x^2)^\alpha f^2\,\mathrm{d}x\leqslant C_\alpha\int_{-1}^{0}(1+x)^\alpha f^2\,\mathrm{d}x=C_\alpha\int_{0}^{1}s^\alpha g^2\,\mathrm{d}s\leqslant C_\alpha\int_{0}^{\infty}s^\alpha g^2\,\mathrm{d}s, \tag{3.7.3}$$

where $C_\alpha=2^\alpha$ if $\alpha>0$ and 1 otherwise. We make use of the Hardy inequality (3.7.1) on the right-most term in (3.7.3), which yields

$$\int_{-1}^{0}(1-x^2)^\alpha f^2\,\mathrm{d}x\leqslant\frac{C_\alpha}{(\alpha+1)^2}\int_{0}^{\infty}s^{\alpha+2}(\partial_s g)^2\,\mathrm{d}s\qquad\text{for }\alpha>-1. \tag{3.7.4}$$

Now, a straightforward computation gives

$$(\partial_s g)^2=(f\partial_x\eta+\eta\partial_x f)^2\leqslant 2((f\partial_x\eta)^2+(\eta\partial_x f)^2)$$

for $s<\frac{3}{2}$ i.e. $x<\frac{1}{2}$, and so from (3.7.4) we can deduce

$$\begin{aligned}\int_{-1}^{0}(1-x^2)^\alpha f^2\,\mathrm{d}x&\leqslant\frac{C_\alpha}{(\alpha+1)^2}\int_{-1}^{1}(1-x^2)^{\alpha+2}((f\partial_x\eta)^2+(\eta\partial_x f)^2)\,\mathrm{d}x\\&\leqslant\frac{C_\alpha}{(\alpha+1)^2}\left(\int_{0}^{\frac{1}{2}}f^2\,\mathrm{d}x+\int_{-1}^{1}(1-x^2)^{\alpha+2}(\partial_x f)^2\,\mathrm{d}x\right)\\&\lesssim_\beta\frac{C_\alpha}{(\alpha+1)^2}\int_{-1}^{1}\left((1-x^2)^\beta f^2+(1-x^2)^{\alpha+2}(\partial_x f)^2\right)\mathrm{d}x,\end{aligned}$$

where on the last line we used $1-x^2\geqslant\frac{3}{4}$ in $(0,\frac{1}{2})$. As per (3.7.2), this implies the desired result. □

3.7.2 Von Mises transformation

The von Mises change of variables of [244, Section 2] (see also [160, Section 5.4], [155]) transforms the free boundary problem (3.1.2) (in any dimension) into the degenerate-parabolic equation (3.1.4) (with $u\equiv 0$, in any dimension). It has the effect of fixing the moving domain to the reference domain which is the open unit ball B_1. For $t\geqslant 0$, the transformation of the spatial coordinates reads

$$x:=\frac{z}{\sqrt{2v(t,z)+|z|^2}}, \tag{3.7.5}$$

where $z\in\{v(t,\cdot)>0\}$. Hence $x\in B_1$, and the transformation reduces to the identity map when $v(t,z)$ is the Barenblatt $\rho(z)$. The unknown in the new variables is defined as

$$w(t,x):=\sqrt{2v(t,z)+|z|^2},$$

so that in these new variables, the Barenblatt reduces to the constant 1. As the interest is to linearize around the Barenblatt, perturbations of the form $w(t,x)=1+y(t,x)$ are considered. In other words,

$$1+y(t,x):=\sqrt{2v(t,z)+|z|^2}. \tag{3.7.6}$$

The inverse change of variables reads

$$z = (y(t,x)+1)x, \qquad v(t,z) - \rho(z) = y(t,x) + \frac{1}{2}y(t,x)^2 \tag{3.7.7}$$

for $t \geqslant 0$ and $x \in B_1$. The transformation (6.7.24), (3.7.6) and (3.7.7) is rigorously justified in [244, Section 3], and the transformation preserves the smallness of the data.

Remark 3.7.2. *If one may apply the above transformation given a null-controlled trajectory y of (3.1.4) (thus provided $\|y\|_{C^{0,1}([0,T]\times[-1,1])} < 1$, see [244, Lemma 3.2]), then $y(T,\cdot) = 0$ would imply $v(T,\cdot) = \rho(\cdot)$, along with equality of the interfaces, as originally desired. The control in the free boundary problem would a priori be actuating inside a moving subregion (due to the fact that the new spatial variable z depends on the state y). However since the results are local, it may be possible to exhibit a time-independent subregion in the new variables (see for instance [64, Lemma 2.10]). Finally, as one may rewrite $v(t,z) = \rho(x)(1+y(t,x))^2$ in (3.7.7), the transformation would moreover guarantee the non-negativity of the controlled trajectory v.*

Chapter 4

The Stefan problem with surface tension

Abstract. We study the controllability properties of the one-phase Stefan problem with surface tension set in a strip-like geometry in two space dimensions, a problem which may be seen as a singular perturbation of the classical Stefan problem via a regularizing term on the free boundary. Using a control actuating along the fixed flat bottom, under smallness conditions on the initial data, we prove the null-controllability of both the temperature and the position of the free boundary in any positive time. Our techniques rely on a careful analysis of the linear problem, which is obtained after fixing the domain via a harmonic extension of the boundary datum, which increases regularity with respect to the free boundary. The null-controllability of the linearized problem is covered by means of a Fourier decomposition in the periodic horizontal variable, and null-controllability results uniform with respect to the Fourier parameter of the one-dimensional problems. The latter are obtained using spectral techniques for the non-zero Fourier modes, whereas the zero mode system is seen as a controllability problem with a finite-dimensional constraint. The nonlinear problem may be tackled by combining an adaptation of the so-called source-term method, and a Banach fixed-point argument. We comment on the feasibility of performing a vanishing surface tension limit in view of addressing the control properties of the classical Stefan problem.

Keywords. Stefan problem, phase transitions, controllability, free boundary problem, Gibbs-Thomson correction, surface tension.

AMS Subject Classification. 93B05, 35R35, 35Q35, 93C20.

This Chapter is a work in collaboration with D. Maity.

Chapter Contents

4.1 Introduction and main result

4.1.1 Origins

The Stefan problem is the quintessential model of phase transitions in liquid-solid systems. The general physical setup of this model consists in considering a domain Ω which is occupied by water, a part of whose boundary is some interface Γ describing contact with a deformable solid such as ice. Due to melting or freezing of the water, the regions occupied by the water and ice will change over time and, consequently, the interface Γ will also change its position and shape. This leads to a *free boundary problem*.

In this paper, we shall solely focus on the *one-phase* Stefan problem, namely wherein the temperature of the ice is not an unknown. The strong formulation of the one-phase Stefan problem corresponds to a free boundary problem involving the linear heat equation $(\partial_t - \Delta)\vartheta = 0$ for the unknown temperature ϑ in the water phase $\Omega(t)$, and by the so-called Stefan condition – which accounts for the exchange of latent heat due to meting of solidifying –, at the unknown moving interface $\Gamma(t)$ separating the water and the ice. In the *classical* Stefan problem, to close the system, one also assumes that the temperature ϑ coincides with the melting temperature 0 at the interface $\Gamma(t)$, meaning

$$\vartheta = 0 \qquad \text{on} \quad \Gamma(t). \tag{4.1.1}$$

Molecular considerations on the mesoscopic level suggest however that the condition (4.1.1) on the free boundary $\Gamma(t)$ should be replaced by the *Gibbs-Thomson correction*

$$\vartheta = -\sigma\kappa \qquad \text{on} \quad \Gamma(t), \tag{4.1.2}$$

where σ is a positive constant, called *surface tension*, and where $\kappa(t)$ denotes the mean curvature of $\Gamma(t)$. The main physical reason for introducing the Gibbs-Thomson correction (4.1.2) stems from the need to account for supercooling effects, in which a fluid permits temperatures below its freezing point, or dendrite formation, in which simple shapes evolve into complex fingering patterns. The effect of supercooling can be on the order of hundreds of degrees for certain materials, see [56, Chapter 1] and [266].

The Stefan problem has become a classical topic in the mathematical literature over the last few decades (see [206, 235], [266, pp. 117–120] for a comprehensive literature review). The classical Stefan problem is by now well-known to admit unique long-time weak solutions characterized by parabolic variational inequalities[1], see for instance [108, 109, 154] and [169, pp. 496–503]. Continuity and regularity of such weak solutions is established in a plethora of works, see e.g. [38, 39, 40, 110, 157, 158, 201]. It is to be noted that classical solutions for the strong formulation of the Stefan problem with condition (4.1.1) – which is closer to the setup we consider here–, were first established in [136, 206]; see [148] for recent results in this direction.

Fewer analytical results concerning existence, regularity and qualitative properties of solutions are known for the Stefan problem with Gibbs-Thomson correction (4.1.2). Under the assumption of existence of smooth solutions for the classical Stefan problem, the authors in [111] prove existence and uniqueness of a weak solution for the linearized problem for $\sigma \ll 1$, and then investigate the effect of small surface tension on the shape of $\Gamma(t)$. Existence of long time weak solutions is established in [10, 194, 230]. A proof for

[1] In some literature on the *parabolic obstacle problem* $\min\{\partial_t y - \Delta y + 1, y\} = 0$ in $(0,T)\times\Omega$ (see e.g. [105]), the case wherein $\partial_t y \geqslant 0$ is interpreted as providing a weak solution to the Stefan problem via the so-called Duvaut transform. Herein, we rather take the perspective of the strong formulation, and directly work with the free boundary parametrized as the graph of a time-dependent function, the latter being viewed an unknown of the problem.

existence – without uniqueness – of local time classical solutions is obtained in [226, 227]. The case of a strip-like geometry, where the free surface is given as the graph of a function – much like the setup we consider herein –, is considered in [94] where existence as well as uniqueness of local time classical solutions is established. Moreover, solutions are shown to instantaneously regularize to become analytic in space and time, using maximal regularity theory. In [147], linearized stability and instability of equilibria are studied. In [132], a strip-like geometry (over the torus) is considered, and asymptotic stability of flat solutions is established via a high-order energy method, in addition to global-in-time existence and uniqueness. These ideas and results are extended in [131], where the question of long-time nonlinear stability of steady-state solutions to the two-phase Stefan problem with surface tension is addressed (see also [134, 133] for subsequent studies, including the zero surface-tension limit in [130]). We refer to the works [225] and the book [224] for further arguments involving linearization and maximal L^p–regularity theory (see also [223] for an extension to Navier-Stokes with a free surface and surface tension condition).

In spite of the breadth of analytical results on the existence, uniqueness and qualitative behavior of solutions to the multidimensional Stefan problem (with or without Gibbs-Thomson correction), very little is known on the controllability properties of this problem. Through this work, we aim to cover this gap and provide new results in this direction.

4.1.2 Formulation

We shall concentrate on the strong formulation of the two-dimensional one-phase Stefan problem, for reasons which will become clear in subsequent discussions. Let $\mathbb{T} := \mathbb{R}/(2\pi\mathbb{Z})$ denote the one-dimensional torus, which we identify with $[0, 2\pi]$, and set

$$\Omega := \mathbb{T} \times (0, 1).$$

The domain Ω will serve as the reference configuration in what follows. We also set

$$\Gamma_{\text{bot}} := \mathbb{T} \times \{0\}, \qquad \Gamma_{\text{top}} = \mathbb{T} \times \{1\}.$$

As mentioned in what precedes, in the one-phase Stefan problem a heat-conducting liquid fills a time-varying domain $\Omega(t) \subset \mathbb{R}^2$ for $t \geqslant 0$. We will assume that the boundary $\partial\Omega(t)$ of the liquid consists of two components, namely a time-varying component (the *free boundary* $\Gamma(t)$) and a fixed component. More specifically, for any $t \geqslant 0$, $\Omega(t)$ is assumed to have a flat, rigid bottom, while the free boundary will be described by the equation $1 + z_2 = h(t, z_1)$. In other words,

$$\Omega(t) := \{z = (z_1, z_2) \in \mathbb{T} \times \mathbb{R} : \quad 0 < z_2 < 1 + h(t, z_1)\},$$

where $h = h(t, z_1)$ is the *height function*, and represents the displacement of the free boundary away from the reference boundary Γ_{top} (see fig. 7.1).
The free boundary is consequently given by

$$\Gamma(t) := \{z = (z_1, z_2) \in \mathbb{T} \times \mathbb{R} : \quad z_2 = 1 + h(t, z_1)\}.$$

Given a time horizon $T > 0$, the one-phase Stefan problem with surface tension (i.e. with Gibbs-Thomson correction) takes the form

$$\begin{cases} \partial_t \vartheta - \Delta\vartheta = 0 & \text{in } (0,T) \times \Omega(t) \\ \partial_t h = -\sqrt{1 + |\partial_{z_1} h|^2}\, \nabla\vartheta_{|_{\Gamma(t)}} \cdot \mathbf{n} & \text{on } (0,T) \times \mathbb{T} \\ \vartheta = -\sigma\kappa(h) & \text{on } (0,T) \times \Gamma(t) \\ \vartheta = u & \text{on } (0,T) \times \Gamma_{\text{bot}} \\ (\vartheta, h)_{|_{t=0}} = (\vartheta^0, h^0) & \text{in } \Omega(0) \times \mathbb{T}, \end{cases} \tag{4.1.3}$$

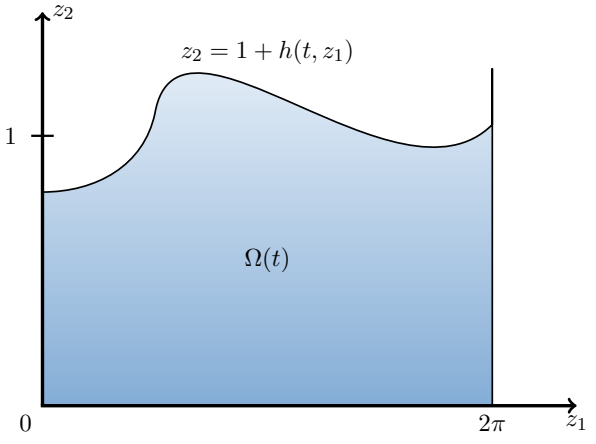


Figure 4.1: The moving domain $\Omega(t)$ representing the liquid, and the free boundary $\Gamma(t)$, delimiting the liquid-solid region, parametrized by the height function $h(t, z_1)$.

where $\vartheta(t, z)$ denotes the unknown temperature of the liquid, $h(t, z_1)$ describes the unknown height function defining the free boundary, $u(t, z_1)$ denotes the control force, while $\mathbf{n} = \mathbf{n}(t, z_1)$ given by

$$\mathbf{n} = \frac{1}{\sqrt{1 + |\partial_{z_1} h|^2}} \begin{bmatrix} -\partial_{z_1} h \\ 1 \end{bmatrix}, \tag{4.1.4}$$

denotes the unit normal to $\Gamma(t)$ outward $\Omega(t)$. The control u actuates along the whole fixed bottom boundary Γ_{bot}. The constant $\sigma > 0$ represents the surface tension coefficient, whereas $\kappa(h)$ denotes the mean curvature of the free boundary $\Gamma(t)$, and is defined by

$$\kappa(h) = \frac{\partial_{z_1}^2 h}{\left(1 + |\partial_{z_1} h|^2\right)^{3/2}}.$$

Finally, the initial domain $\Omega(0)$ is given by

$$\Omega(0) = \{z = (z_1, z_2) \in \mathbb{T} \times \mathbb{R} \colon \quad 0 < z_2 < 1 + h_0(z_1)\}.$$

We note that when $\sigma = 0$, (4.1.3) reduces to the classical Stefan problem, namely one has (4.1.1) instead of the Gibbs-Thomson condition along the interface $\Gamma(t)$.

4.1.3 Main results

One may observe that $(\vartheta^*, h^*) = (0, 0)$ is an equilibrium configuration for (4.1.3). Hence, the natural control problem we aim to address in this work is the following. Given a time horizon $T > 0$, a surface tension coefficient $\sigma > 0$, and initial data (ϑ^0, h^0), which are small enough in an appropriate topology, we look to find a control $u = u(t, z_1)$ actuating along the fixed boundary Γ_{bot} such that

$$\vartheta(T, \cdot) = 0 \qquad \text{in } \Omega = \mathbb{T} \times (0, 1) \qquad \text{and} \qquad h(T, \cdot) = 0 \qquad \text{on } \mathbb{T}. \tag{4.1.5}$$

This is reflected in what follows.

Conjecture 4.1.1. *Let $T > 0$ be a given time, and let $\sigma > 0$. There exists $\varepsilon > 0$ such that for every pair of initial data (ϑ^0, h^0) satisfying*

$$h^0 \in H^{5/2}(\mathbb{T}), \qquad \min_{z_1 \in \mathbb{T}} \left(1 + h^0(z_1)\right) > 0, \qquad \vartheta^0 \in H^1(\Omega(0)),$$

the compatibility condition

$$\vartheta^0 \left(z_1, 1 + h^0(z_1)\right) = -\sigma \kappa \left(h^0(z_1)\right) \qquad \text{for } z_1 \in \mathbb{T},$$

and

$$\left\|\vartheta^0\right\|_{H^1(\Omega(0))} + \left\|h^0\right\|_{H^{5/2}(\mathbb{T})} \leqslant \varepsilon,$$

there exists a control $u \in L^2(0,T;H^{3/2}(\mathbb{T})) \cap H^{3/4}(0,T;L^2(\mathbb{T}))$ *such that the corresponding solution* (ϑ, h) *to* (4.1.3) *satisfies* (4.1.5)*. Furthermore, the controlled trajectory satisfies the regularity*

$$\vartheta \in L^2\Big(0,T;H^2(\Omega(\cdot))\Big) \cap C^0\Big([0,T];H^1(\Omega(\cdot))\Big) \cap H^1\Big(0,T;L^2(\Omega(\cdot))\Big),$$

and

$$\begin{aligned} h \in L^2\Big(0,T;H^{7/2}(\mathbb{T})\Big) \cap H^{3/4}\Big(0,T;H^2(\mathbb{T})\Big) \cap H^1\Big(0,T;H^1(\mathbb{T})\Big) \\ \cap H^{5/4}\Big(0,T;L^2(\mathbb{T})\Big) \cap C^0\Big([0,T];H^{5/2}(\mathbb{T})\Big). \end{aligned}$$

We anticipate the above conjecture to hold – a result which would be the first of its kind for free boundary problems where the free boundary is a space-dependent function. To stimulate this conjecture, let us provide a brief sketch of the proof methodology.

Step 1). Fixing the domain. We begin by fixing the domain to render the analysis and control of the corresponding linear system more convenient, as it will allow us to work in a time-independent spatial setup. We emphasize that in the two-dimensional geometrical setup we consider here, the free boundary depends on the spatial variable z_1, hence the regularity of the domain $\Omega(t)$ depends on the spatial regularity of the height function $h(t,z_1)$. To avoid requiring high order spatial Sobolev regularity on h, we shall fix the domain via a transformation which gains a $\frac{1}{2}$–order of regularity with respect to h. This is done by defining

$$\Psi(t,x) := (x_1, x_2 + \psi(t,x)) \qquad y(t,x) = \vartheta(t,\Psi(t,x)) \qquad \text{for } \ (t,x) \in (0,T)\times\Omega,$$

given $h(t,\cdot) \in H^s(\mathbb{T})$ for some $s \geqslant 0$ and for all $t \geqslant 0$, where $\psi(t,\cdot) \in H^{s+1/2}(\Omega)$ solves

$$\begin{cases} \Delta\psi(t,\cdot) = 0 & \text{in } \ \Omega \\ \psi(t,x_1,0) = 0 & \text{on } \ \mathbb{T} \\ \psi(t,x_1,1) = h(t,x_1) & \text{on } \ \mathbb{T}. \end{cases}$$

This transformation is a diffeomorphism for sufficiently small $h(t,\cdot)$, and is also used in [130]. It leads us to the system in the reference domain Ω:

$$\begin{cases} \partial_t y - \Delta y = \mathcal{N}_1(y,h) & \text{in } (0,T)\times\Omega, \\ \partial_t h = (\nabla y \cdot \mathbf{e}_2)_{\Gamma_{\text{top}}} + (\mathcal{N}_3(y,h)\cdot\mathbf{e}_2)_{\Gamma_{\text{top}}} & \text{on } (0,T)\times\Gamma_{\text{top}}, \\ y = \sigma\partial_{x_1}^2 h + \mathcal{N}_2(y,h) & \text{on } (0,T)\times\Gamma_{\text{top}}, \\ y = u & \text{on } (0,T)\times\Gamma_{\text{bot}}, \\ (y,h)\,|_{t=0} = \big(y^0,h^0\big) & \text{in } \Omega\times\mathbb{T}, \end{cases} \tag{4.1.6}$$

where the nonlinear terms $\{\mathcal{N}_i\}_{i=1}^3$ are all quadratic.

Step 2). The linearized system. As commonly done in literature on the controllability of nonlinear parabolic problems, we will first concentrate on proving the controllability of the system linearized around the target $(0,0)$, and then view the nonlinear terms in (4.1.6) as a small perturbation which may be dealt with by means of a fixed-point argument. Moreover, to avoid working with boundary control systems, we extend the physical reference domain Ω to the fictitious domain $\mathcal{O} := \mathbb{T}\times(-1,1)$ and consider a distributed control, actuating inside an open and nonempty subset $\omega = \mathbb{T}\times(a,b)$ with $(a,b) \subset (-1,0)$. In other words, the distributed control problem for the linearized Stefan

problem with Gibbs-Thomson correction takes the form

$$\begin{cases} \partial_t y - \Delta y = u\mathbf{1}_\omega & \text{in } (0,T)\times\mathcal{O} \\ \partial_t h(t,x_1) = \partial_{x_2} y(t,x_1,1) & \text{on } (0,T)\times\mathbb{T} \\ y(t,x_1,0) = 0 & \text{on } (0,T)\times\mathbb{T} \\ y(t,x_1,1) = \sigma\partial_{x_1}^2 h(t,x_1) & \text{on } (0,T)\times\mathbb{T} \\ (y,h)_{|t=0} = \big(y^0,h^0\big) & \text{in } \mathcal{O}\times\mathbb{T}. \end{cases} \tag{4.1.7}$$

We prove the following result, which to the best of our knowledge, is also new in the literature on the control of parabolic systems.

Theorem 4.1. *Let $T>0$ and $\sigma>0$. For any $(y^0,h^0)\in L^2(\mathcal{O})\times H^1(\mathbb{T})$, there exists $u\in L^2((0,T)\times\omega)$ such that the corresponding unique solution $y\in C^0([0,T];L^2(\mathcal{O}))$ and $h\in C^0([0,T];H^1(\mathbb{T}))$ of (4.1.7) satisfies*

$$y(T,\cdot)=0 \quad \text{in } \mathcal{O} \qquad \text{and} \qquad h(T,\cdot)=0 \quad \text{in } \mathbb{T}.$$

Moreover, there exists a positive constant $\mathfrak{C}(T,\sigma)=\mathfrak{C}(T,\omega,\Omega,\sigma)>0$ such that

$$\|u\|_{L^2((0,T)\times\omega)} \leqslant \mathfrak{C}(T,\sigma)\,\big\|\big(y^0,h^0\big)\big\|_{L^2(\mathcal{O})\times H^1(\mathbb{T})}.$$

We provide a brief proof of the well-posedness of the linear system (4.1.7) in this simplified functional framework via analytic semigroup arguments in Proposition 4.3.1 & Corollary 4.3.2 – note that the ambient state space is chosen as $L^2(\mathcal{O})\times H^1(\mathbb{T})$.

The proof of Theorem 4.1 is a cornerstone of our work. Due to the nontrivial form of the adjoint problem (the governing linear operator is not self-adjoint), a direct proof via HUM and an observability inequality does not appear straightforward. Instead, we exploit the periodicity of the control and the unknowns with respect to the $x_1\in\mathbb{T}$ variable, write the unknowns in Fourier series, prove that each Fourier coefficient is null-controllable with a control cost uniform in the Fourier parameter, and then paste all the coefficients together to deduce Theorem 4.1. Such ideas have already been used in the control literature, see [24, 21] for instance. To be more precise, for any $n\in\mathbb{Z}$, the system satisfied by each Fourier coefficient is

$$\begin{cases} \partial_t \widehat{y}_n - \partial_{x_2}^2\widehat{y}_n + n^2\widehat{y}_n = \widehat{u}_n\mathbf{1}_{(a,b)} & \text{in } (0,T)\times(-1,1) \\ \widehat{h}_n'(t) = \partial_{x_2}\widehat{y}_n(t,1) & \text{in } (0,T) \\ \widehat{y}_n(t,-1)=0 & \text{in } (0,T) \\ \widehat{y}_n(t,1) = -\sigma n^2\widehat{h}_n(t) & \text{in } (0,T) \\ \Big(\widehat{y}_n,\widehat{h}_n\Big)_{|t=0} = \Big(\widehat{y}_n^0,\widehat{h}_n^0\Big) & \text{in } (-1,1). \end{cases} \tag{4.1.8}$$

The null-controllability of (4.1.8) (Proposition 4.3.4) is then done in two parts, distinguishing in (4.1.8) the case $n\neq 0$, where the governing linear operator is self-adjoint in an appropriate product space and the observability inequality follows from an explicit computation of the spectrum, and the case $n=0$, in which $\widehat{y}_n$ is independent of $\widehat{h}_n$, and the controllability of $\widehat{h}_n$ is seen as a finite-dimensional constraint on the linear heat control, and may be covered using improved observability inequalities done by compactness-uniqueness arguments as in [116].

Step 3). The nonlinear system. To tackle the nonlinear system, we will perform a Banach fixed-point argument over the source-terms decoying the nonlinear terms in (4.1.6). To obtain the required null-controllability result for the problem with given source terms, we make use of an adaptation of the *source-term method* (see e.g. [191, 173, 115]) in fractional Sobolev spaces (see Theorem 4.3). We recall that for the source term method, the decay of the source terms should be quick enough near the final time compared to the control cost in small time, and the Banach fixed-point argument is then performed inside small enough balls of these weighted energy spaces. To conclude the proof of Conjecture 4.1.1, it remains to be shown that the quadratic nonlinear terms are indeed elements of these weighted energy spaces provided by the source-term method.

4.1.4 Related work

The null-controllability result we prove in this work is among the first of its kind for multi-dimensional free-boundary problems where the free boundary depends on the spatial variable. In this sense, our problem differs from existing works on the controllability of fluid-structure interaction models (e.g. [143, 31, 229, 30, 180]), and the controllability of one-dimensional free boundary problems ([191, 102, 116, 115]), as therein, the free boundary is parametrized by the graph of a time-only dependent function. In particular, the spatial regularity of the height function h plays a crucial role in the analysis (or even existence) results. One needs to possibly consider very regular initial data (ϑ^0, h^0) in order to guarantee the smoothness of the domain.

A partial controllability result for the two-dimensional classical Stefan problem is shown in [77] (following the partial controllability result in the one-dimensional setting in [103]). Only the temperature ϑ is controlled to 0 without any consideration of the height function $h(t, z_1)$ defining the free boundary $\Gamma(t)$. In fact, the geometrical setting is also different, as the free boundary $\Gamma(t)$ manifests as the entire boundary of the fluid domain $\Omega(t)$. Moreover, the Stefan law governing the velocity of the height function is regularized by adding a Laplacian term, which simplifies the analysis.

Albeit for a system of different nature to ours, we also refer to the work [5] (see also [3, 4, 251] for related observability and stabilization results, and [2] for stabilization of water waves with surface tension) for an exact-controllability result of the velocity and the free surface elevation of the water-waves equations in two dimensions, by means of a single control actuating along an open subset of the free surface. In the aforementioned works, the two-dimensional geometrical strip-like setting of the free boundary problem is the same as ours. Since the fluid is assumed to be irrotational, the authors may work with the trace on the free surface and use the Dirichlet to Neumann map to redefine the problem on a fixed domain. This procedure is closely related to the equations under consideration, and is not applicable in our setting. After linearization, a dispersive equation is obtained, which is shown to be controllable in arbitrarily short time by means of Ingham-like techniques. Due to the lack of regularizing effect, the nonlinear problem is then tackled by using a Nash-Moser iteration. These results are extended to the three dimensional context in [278].

4.1.5 Outline

The remainder of this work is organized as follows.

- In Section 4.2, we define in more detail the transformation used to fix the domain, and the nonlinear terms which it entails.
- in Section 4.3, we take a look at the linearized version of the transformed problem, and we prove that it is well-posed in an appropriate Hilbertian setting, and that it is null-controllable by means of the methodology presented just above.
- In Section 4.4, we add the source-terms needed for the fixed-point argument by virtue of an adaptation of the source-term method for fractional Sobolev spaces.

Notation. We denote by $\mathbb{N}$ the set of non-negative integers, and $\mathbb{Z}^* = \mathbb{Z} \setminus \{0\}$. Given $T > 0$, we use the notations

$$(0,T) \times \Omega(t) := \bigcup_{0 \leqslant t \leqslant T} \{t\} \times \Omega(t), \qquad (0,T) \times \Gamma(t) := \bigcup_{0 \leqslant t \leqslant T} \{t\} \times \Gamma(t).$$

Whenever the dependence on parameters of a constant is not specified, we will write $f \lesssim_S g$ whenever a constant $C > 0$, depending only on the set of parameters S, exists such that $f \leqslant Cg$.

4.2 Fixing the domain

In this section, we present the transformation allowing us to pass from system (4.1.3) set in the moving domain $\Omega(t)$, to a nonlinear problem in the time-independent reference domain Ω. Before proceeding, we indicate the following elementary observation.

Remark 4.2.1. *Note that one simple change of variables which fixes the domain is the following. Assume $h(t,\cdot) \in H^s(\mathbb{T})$ for some $s \geqslant 0$ and for all $t \geqslant 0$. We define the map $\Psi(t,\cdot) : \overline{\Omega} \longrightarrow \overline{\Omega(t)}$ for $t \geqslant 0$ by*

$$\Psi(t,x) := \Big(x_1, \big(1 + h(t,x_1)\big)\, x_2\Big), \qquad x = (x_1,x_2) \in \overline{\Omega}.$$

Note however that if $h(t,\cdot) \in H^s(\mathbb{T})$, then $\Psi(t,\cdot) \in H^s(\Omega)$ – namely, the transformation $\Psi(t,\cdot)$ preserves the spatial regularity of h, which means that a higher regularity than necessary on h may be required.

In view of the above observation, we proceed by defining a slightly different transformation to fix the domain which entails higher spatial regularity than that of the height function $h(t,\cdot)$. Given $h \in C^0([0,T]; H^s(\mathbb{T}))$ for $s \geqslant 0$, for any $t \geqslant 0$, and recalling the definitions $\Gamma_{\text{top}} := \mathbb{T} \times \{1\}$ and $\Gamma_{\text{bot}} := \mathbb{T} \times \{0\}$, we consider the solution $\psi(t,\cdot)$ to

$$\begin{cases} \Delta\psi(t,\cdot) = 0 & \text{in } \Omega \\ \psi(t,x_1,0) = 0 & \text{on } \mathbb{T} \\ \psi(t,x_1,1) = h(t,x_1) & \text{on } \mathbb{T}, \end{cases}$$

and we define[2] the gauge $\Psi(t,\cdot) : \overline{\Omega} \longrightarrow \overline{\Omega(t)}$ by

$$\Psi(t,x) := \big(x_1, x_2 + \psi(t,x)\big).$$

In this case, if $h(t,\cdot) \in H^s(\mathbb{T})$ then $\Psi(t,\cdot) \in H^{s+1/2}(\Omega)$, which represents a gain in regularity of the transformation with respect to the input height function. Note that Ψ is similar to the transformation defined in [130, Eq. (1.6)]. From elliptic estimates, it can be seen that

$$\|\Psi(t,\cdot) - \mathrm{Id}\|_{H^{s+1/2}(\Omega)} \lesssim \|h(t,\cdot)\|_{H^s(\mathbb{T})}$$

for all $t \geqslant 0$, so whenever $h(t,\cdot)$ is sufficiently small, $\Psi(t,\cdot)$ is a diffeomorphism from $\overline{\Omega}$ onto $\overline{\Omega(t)}$ by the inverse function theorem. In this case, we denote by $X(t,\cdot) = [\Psi(t,\cdot)]^{-1}$ the inverse of $\Psi(t,\cdot)$ for all $t \geqslant 0$, and consider the following change of unknown

$$y(t,x) = \vartheta(t,\Psi(t,x)) \qquad \text{for } (t,x) \in (0,T) \times \Omega.$$

In other words,

$$\vartheta(t,z) = y(t,X(t,z)) \qquad \text{for } (t,z) \in (0,T) \times \Omega(t).$$

We also introduce the standard notation

$$\mathfrak{B}_\Psi := \mathrm{Cof}(\nabla\Psi), \qquad \text{and} \qquad \mathfrak{A}_\Psi := \frac{1}{\det(\nabla\Psi)} \mathfrak{B}_\Psi^\top \mathfrak{B}_\Psi,$$

where $\delta_\Psi := \det(\nabla\Psi)$ denotes the Jacobian determinant of $\nabla\Psi$, and $\mathrm{Cof}(M)$ denotes the cofactor matrix of M, satisfying $M(\mathrm{Cof}(M))^\top = (\mathrm{Cof}(M))^\top M = \det(M)\mathrm{Id}$. The system (4.1.3) can then be equivalently rewritten as

$$\begin{cases} \partial_t y - \Delta y = \mathcal{N}_1(y,h) & \text{in } (0,T) \times \Omega, \\ \partial_t h = (\nabla y \cdot \mathbf{e}_2)_{\Gamma_{\text{top}}} + (\mathcal{N}_3(y,h) \cdot \mathbf{e}_2)_{\Gamma_{\text{top}}} & \text{on } (0,T) \times \Gamma_{\text{top}}, \\ y = \sigma\partial_{x_1}^2 h + \mathcal{N}_2(y,h) & \text{on } (0,T) \times \Gamma_{\text{top}}, \\ y = u & \text{on } (0,T) \times \Gamma_{\text{bot}}, \\ (y,h)\,|_{t=0} = \big(y^0, h^0\big) & \text{in } \Omega \times \mathbb{T}, \end{cases} \tag{4.2.1}$$

[2] Defining a diffeomorphism via a harmonic extension of the boundary diffeomorphism is in the spirit of the Arbitrary Eulerian-Lagrangian (ALE) coordinates.

where $y^0(\cdot) := \vartheta^0(\Psi(0,\cdot))$, with the quadratic nonlinear terms having the form

$$\begin{aligned}\mathcal{N}_1(y,h) &= -(\det(\nabla\Psi)-1)\partial_t y + \operatorname{div}(\mathcal{N}_3),\\ \mathcal{N}_2(y,h) &= \sigma\left(\kappa(h) - \partial_{x_1}^2 h\right),\\ \mathcal{N}_3(y,h) &= (\mathfrak{A}_\Psi - \mathrm{Id})\,\nabla y.\end{aligned}$$

4.3 Control of the linear system

We will now investigate the null-controllability of (4.1.3) linearized around the equilibrium $(0,0)$, which taking (4.2.1) into account, reads:

$$\begin{cases}\partial_t y - \Delta y = 0 & \text{in } (0,T)\times\Omega\\ \partial_t h(t,x_1) = \partial_{x_2} y(t,x_1,1) & \text{on } (0,T)\times\mathbb{T}\\ y(t,x_1,0) = u(t,x_1) & \text{on } (0,T)\times\mathbb{T}\\ y(t,x_1,1) = \sigma\partial_{x_1}^2 h(t,x_1) & \text{on } (0,T)\times\mathbb{T}\\ (y,h)_{|t=0} = \left(y^0,h^0\right) & \text{in } \Omega(0)\times\mathbb{T}.\end{cases} \tag{4.3.1}$$

We recall that $\sigma>0$ is fixed and $\Omega = \mathbb{T}\times(0,1)$. The unknown state (y,h) is periodic with respect to the horizontal (i.e. x_1) variable. Our goal in what follows is to prove the *null-controllability* of (4.3.1) – namely, to find a control $u = u(t,x_1)$ such that the corresponding solution (y,h) of (4.3.1) satisfies

$$y(T,\cdot) = 0 \quad \text{in } \Omega \qquad \text{and} \qquad h(T,\cdot) = 0 \quad \text{on } \mathbb{T}. \tag{4.3.2}$$

Following common practice in the control of parabolic equations and systems, we will first extend the physical reference domain Ω to the fictitious domain $\mathcal{O} = \mathbb{T}\times(-1,1)$, and consider a distributed control for the linear heat equation set in $\mathcal{O}$, with the control u actuating inside the open subset[3] $\omega := \mathbb{T}\times\left(-\frac{3}{4},-\frac{1}{4}\right)$. We thus consider the distributed control problem

$$\begin{cases}\partial_t y - \Delta y = u\mathbf{1}_\omega & \text{in } (0,T)\times\mathcal{O}\\ \partial_t h(t,x_1) = \partial_{x_2} y(t,x_1,1) & \text{on } (0,T)\times\mathbb{T}\\ y(t,x_1,0) = 0 & \text{on } (0,T)\times\mathbb{T}\\ y(t,x_1,1) = \sigma\partial_{x_1}^2 h(t,x_1) & \text{on } (0,T)\times\mathbb{T}\\ (y,h)_{|t=0} = \left(y^0,h^0\right) & \text{in } \mathcal{O}\times\mathbb{T},\end{cases} \tag{4.3.3}$$

where (y^0,h^0) are appropriate extensions of the initial data in (4.3.1). Using a standard restriction argument, it can then be shown that the controllability of (4.3.3) implies the controllability of (4.3.1).

It should be noted however that the well-posedness of the linear systems (4.3.1) and (4.3.3), and in particular, the functional framework, is not immediately obvious. In view of this, we begin with a presentation a more in-depth analysis of this issue before proceeding with the control methodology.

4.3.1 The linear semigroup

We look to rewrite (4.3.3) in a canonical abstract control system evolving on state space

$$\mathcal{H} := L^2(\mathcal{O})\times H^1(\mathbb{T}).$$

We begin with some needed definitions. For a Hilbert space $\mathcal{X}$ and for $s\geqslant 0$, we define the fractional Sobolev space

$$H^s(\mathbb{T},\mathcal{X}) := \left\{ f:\mathbb{T}\to\mathcal{X} \mid f(x_1,x_2) = \sum_{n\in\mathbb{Z}} \widehat{f}_n(x_2)\varphi_n(x_1),\ \sum_{n\in\mathbb{Z}} |n|^{2s}\left\|\widehat{f}_n\right\|_{\mathcal{X}}^2 < \infty\right\},$$

[3] This specific choice of ω is done in view of simplifying subsequent spectral computations, but of course, the result would hold for any open, non-empty subset $\omega\subset\mathcal{O}$.

where $\{\varphi_n\}_{n\in\mathbb{Z}}$ generate the orthonormal basis of $\mathcal{X}$ and $\widehat{f}_n = \langle f, \varphi_n\rangle_{\mathcal{X}}$ denote the Fourier coefficients of f. The Sobolev space $H^s(\mathbb{T};\mathcal{X})$ is endowed with the norm

$$\|f\|_{H^s(\mathbb{T},\mathcal{X})} := \left(\sum_{n\in\mathbb{Z}} \left(1+|n|^{2s}\right) \left\|\widehat{f}_n\right\|_{\mathcal{X}}^2\right)^{1/2}.$$

This definition will be of use in the subsequent analysis.

To write (4.3.3) as an abstract control system evolving in $\mathcal{H}$, we begin by introducing the unbounded operator $\mathcal{A} : \mathcal{D}(\mathcal{A}) \to \mathcal{H}$, which governs the dynamics of (4.3.3), defined by its domain

$$\mathcal{D}(\mathcal{A}) = \Big\{(y,h) \in H^2(\mathcal{O}) \times H^{7/2}(\mathbb{T}) \mid y - \sigma\partial_{x_1}^2 h = 0 \text{ on } \Gamma_{\text{top}},\ y = 0 \text{ on } \Gamma_{\text{bot}},\ \partial_{x_2} y(\cdot,1) \in H^1(\mathbb{T})\Big\},$$

and

$$\mathcal{A}\begin{bmatrix} y \\ h \end{bmatrix} = \begin{bmatrix} \Delta y \\ \partial_{x_2} y_{|\Gamma_{\text{bot}}} \end{bmatrix}.$$

We also introduce the control operator $\mathcal{B} \in \mathcal{L}(L^2(\mathcal{O});\mathcal{H})$ defined by

$$\mathcal{B}u = \begin{bmatrix} \mathbf{1}_\omega u \\ 0 \end{bmatrix}.$$

By using the above definitions, we can clearly rewrite the system (4.3.3) as a first order system in the Hilbert state space $\mathcal{H}$:

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} y \\ h \end{bmatrix} = \mathcal{A}\begin{bmatrix} y \\ h \end{bmatrix} + \mathcal{B}u \quad \text{in } (0,T) \\ \begin{bmatrix} y \\ h \end{bmatrix}_{|t=0} = \begin{bmatrix} y^0 \\ h^0 \end{bmatrix}. \end{cases} \tag{4.3.4}$$

We now prove the following result, which using standard results from parabolic equations (see e.g. [26, Part II, Chap. 1, Sect. 3]), will entail the well-posedness of the linear system (4.3.4) (and thus (4.3.3), and also (4.3.1)).

Proposition 4.3.1. *The operator $\mathcal{A} : \mathcal{D}(\mathcal{A}) \to \mathcal{H}$ is the infinitesimal generator of an analytic semigroup $\left\{e^{t\mathcal{A}}\right\}_{t\geqslant 0}$ on $\mathcal{H}$.*

Proof of Proposition 4.3.1. It can readily be seen that the operator $\mathcal{A} : \mathcal{D}(\mathcal{A}) \to \mathcal{H}$ is closed and densely defined. On another hand, it is well-known that $\mathcal{A}$ would generate an analytic semigroup on $\mathcal{H}$ if $\mathcal{A}$ is *sectorial* (see e.g. [26, Part II, Chap. 1, Sect. 3]), in the sense that there exists $\beta_0 \in \left(\frac{\pi}{2},\pi\right)$ and $\beta_1 > 0$ such that the sector

$$\Sigma_{\beta_0,\beta_1} := \{\lambda \in \mathbb{C}\setminus\{0\} \mid |\arg(\lambda)| < \beta_0,\ |\lambda| \geqslant \beta_1\}$$

is a subset of the resolvent set $\rho(\mathcal{A})$ of $\mathcal{A}$, and if there exists a constant $C > 0$ such that the estimate

$$\left\|(\lambda - \mathcal{A})^{-1}\right\|_{\mathcal{L}(L^2(\mathcal{O})\times H^1(\mathbb{T}))} \leqslant \frac{C}{|\lambda|}$$

holds for all $\lambda \in \Sigma_{\beta_0,\beta_1}$.

We thus proceed in showing that $\mathcal{A}$ is sectorial. Let $(f,g) \in \mathcal{H}$. We consider the resolvent problem

$$\begin{cases} \lambda y - \Delta y = f & \text{in } \mathcal{O}, \\ y = \sigma\partial_{x_1}^2 h & \text{on } \Gamma_{\text{top}}, \\ y = 0 & \text{on } \Gamma_{\text{bot}}, \\ \lambda h = \partial_{x_2} y + g & \text{in } \Gamma_{\text{top}}. \end{cases}$$

We also decompose all functions appearing in the resolvent equation above in Fourier series with respect to the periodic, x_1–variable:

$$\begin{aligned} y(x_1,x_2) &= \frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}} \widehat{y}_n(x_2)e^{inx_1} && \text{for } (x_1,x_2)\in\mathcal{O},\\ h(x_1) &= \frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}} \widehat{h}_n e^{inx_1} && \text{for } \ x_1\in\mathbb{T},\\ f(x_1,x_2) &= \frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}} \widehat{f}_n(x_2)e^{inx_1} && \text{for } \ (x_1,x_2)\in\mathcal{O},\\ g(x_1) &= \frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}} \widehat{g}_n e^{inx_1} && \text{for } \ x_1\in\mathbb{T}, \end{aligned}$$

where we recall that the Fourier coefficients of any $\psi(x_1,x_2)$ are $\widehat{\psi}_n(x) := \langle \psi(x,\cdot), e^{in\cdot}\rangle_{L^2(\mathbb{T})}$, with the complex exponentials denoting the orthonormal basis of $L^2(\mathbb{T})$. Then, for each $n\in\mathbb{Z}$, the pair $(\widehat{y}_n,\widehat{h}_n)$ of Fourier coefficients solves

$$\begin{cases} \lambda\widehat{y}_n - \partial_{x_2}^2\widehat{y}_n + n^2\widehat{y}_n = \widehat{f}_n \quad \text{in } (-1,1),\\ \widehat{y}_n(1) = -\sigma n^2\widehat{h}_n\\ \widehat{y}_n(-1) = 0\\ \lambda\widehat{h}_n = \partial_{x_2}\widehat{y}_n(1) + \widehat{g}_n. \end{cases} \tag{4.3.5}$$

Now let $\beta_0\in\left(\frac{\pi}{2},\pi\right)$ and $\beta_1>0$ be fixed. Then for each $n\in\mathbb{Z}\setminus\{0\}$, we have $\Sigma_{\beta_0,\beta_1}\subset\rho(\mathcal{A}_n)$, (indeed, see Lemma 4.3.5) where $\mathcal{A}_n$ is the linear operator associated to the resolvent problem (4.3.5), with $\rho(\mathcal{A}_n)$ denoting the resolvent set of $\mathcal{A}_n$. To be more specific, $\mathcal{A}_n : \mathcal{D}(\mathcal{A}_n)\to L^2(-1,1)\times\mathbb{C}$ is defined by its domain

$$\mathcal{D}(\mathcal{A}_n) = \left\{(\zeta,r)\in H^2(-1,1)\times\mathbb{C} \mid \zeta(-1)=0,\ \zeta(1)=-\sigma n^2 r\right\},$$

and

$$\mathcal{A}_n\begin{bmatrix}\zeta\\ r\end{bmatrix} = \begin{bmatrix}\partial_{x_2}^2\zeta - n^2\zeta\\ \partial_{x_2}\zeta(1)\end{bmatrix}.$$

Now take $\lambda=\beta_1 e^{i\beta_0}\in\Sigma_{\beta_0,\beta_1}$. Multiplying the first equation in (4.3.5) by $e^{-i\beta_0/2}$ and taking inner product with $\widehat{y}_n$, we obtain

$$\beta_1 e^{i\beta_0/2}\int_{-1}^1 |\widehat{y}_n|^2\,\mathrm{d}x_2 + e^{-i\beta_0/2}\int_{-1}^1 |\partial_{x_2}\widehat{y}_n|^2\,\mathrm{d}x_2 - e^{-i\beta_0/2}\partial_{x_2}\widehat{y}_n(1)\widehat{y}_n(1) = e^{-i\beta_0/2}\int_{-1}^1 \widehat{f}_n\widehat{y}_n\,\mathrm{d}x_2.$$

Using the boundary conditions, the above identity can be rewritten as

$$\begin{aligned} \beta_1 e^{i\beta_0/2}\int_{-1}^1 |\widehat{y}_n|^2\,\mathrm{d}x_2 + e^{-i\beta_0/2}\int_{-1}^1 |\partial_{x_2}\widehat{y}_n|^2\,\mathrm{d}x_2 + \beta_1 e^{i\beta_0/2}\sigma n^2\left|\widehat{h}_n\right|^2 \\ = e^{-i\beta_0/2}\int_{-1}^1 \widehat{f}_n\widehat{y}_n\,\mathrm{d}x_2 + e^{-i\beta_0/2}\sigma n^2\widehat{h}_n\widehat{g}_n. \end{aligned}$$

By taking real part on both sides in the above identity, we find

$$\begin{aligned} \beta_1\cos\left(\frac{\beta_0}{2}\right)\left(\int_{-1}^1 |\widehat{y}_n|^2\,\mathrm{d}x_2 + n^2\left|\widehat{h}_n\right|^2\right) + \cos\left(\frac{\beta_0}{2}\right)\int_{-1}^1 |\partial_{x_2}\widehat{y}_n|^2\ \mathrm{d}x_2 \\ \leqslant \left\|\left(\widehat{y}_n, n\widehat{h}_n\right)\right\|_{L^2(-1,1)\times\mathbb{C}}\left\|\left(\widehat{f}_n, n\widehat{g}_n\right)\right\|_{L^2(-1,1)\times\mathbb{C}}. \end{aligned}$$

Taking into account the fact that $\cos\left(\frac{\beta_0}{2}\right)>0$, we deduce

$$|\lambda|\left\|\left(\widehat{y}_n, n\widehat{h}_n\right)\right\|_{L^2(-1,1)\times\mathbb{C}} \leqslant C\left\|\left(\widehat{f}_n, n\widehat{g}_n\right)\right\|_{L^2(-1,1)\times\mathbb{C}},$$

for all $n \in \mathbb{Z} \setminus \{0\}$ and for all $\lambda \in \Sigma_{\beta_0,\beta_1}$, and for some constant $C > 0$ which is independent of n and λ. On another hand, $\mathcal{A}_0$ also generates an analytic semigroup on $L^2(-1,1) \times \mathbb{C}$[4]. In particular, there exists $\beta_0 \in \left(\frac{\pi}{2}, \pi\right)$ and $\beta_1 > 0$ such that

$$|\lambda| \left\| \left(\widehat{y}_0, \widehat{h}_0\right) \right\|_{L^2(-1,1)\times\mathbb{C}} \leqslant C \left\| \left(\widehat{f}_0, \widehat{g}_0\right) \right\|_{L^2(-1,1)\times\mathbb{C}},$$

for all $\lambda \in \Sigma_{\beta_0,\beta_1}$. Combing both of the above estimates and summing up over all $n \in \mathbb{Z}$, we deduce that there exists some $\beta_0 \in \left(\frac{\pi}{2}, \pi\right)$ and $\beta_1 > 0$ such that,

$$|\lambda| \, \|(y,h)\|_{L^2(\mathcal{O})\times H^1(\mathbb{T})} \leqslant C \, \|(f,g)\|_{L^2(\mathcal{O})\times H^1(\mathbb{T})}, \qquad \text{for all } \lambda \in \Sigma_{\beta_0,\beta_1}.$$

This completes the proof. □

In view of the above result, and standard results from parabolic equations (see e.g. [26, Thm. 2.12, Sect. 2]), we deduce the well-posedness of the linear system (4.3.4) (and thus (4.3.3), and also (4.3.1)).

Corollary 4.3.2. *Let $T > 0$. For every $(y^0, h^0) \in L^2(\mathcal{O}) \times H^1(\mathbb{T})$ and $f \in L^2(0,T; L^2(\mathcal{O}))$, there exists a unique mild solution (y,h), with $y \in C^0([0,T]; L^2(\mathcal{O}))$ and $h \in C^0([0,T]; H^1(\mathbb{T}))$, to*

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} y \\ h \end{bmatrix} = \mathcal{A} \begin{bmatrix} y \\ h \end{bmatrix} + f \quad \text{in } (0,T) \\ \begin{bmatrix} y \\ h \end{bmatrix}_{|t=0} = \begin{bmatrix} y^0 \\ h^0 \end{bmatrix}. \end{cases}$$

Moreover, there exists a constant $C = C(T,\sigma) > 0$ such that

$$\|y\|_{C^0([0,T];L^2(\mathcal{O}))} + \|h\|_{C^0([0,T];H^1(\mathbb{T}))} \leqslant C \left(\left\| (y^0, h^0) \right\|_{L^2(\mathcal{O})\times H^1(\mathbb{T})} + \|f\|_{L^2(0,T;L^2(\mathcal{O}))} \right).$$

4.3.2 Null-controllability of the linearized system

Having shown that (4.3.3) is well-posed in an appropriate Hilbertian setting, we now aim to prove Theorem 4.1, namely the following controllability result for (4.3.3) which may be written in the form (4.3.4).

Theorem 4.2. *Let $T > 0$ and $\sigma > 0$. For any $(y^0, h^0) \in L^2(\mathcal{O}) \times H^1(\mathbb{T})$, there exists $u \in L^2((0,T) \times \omega)$ such that the corresponding unique solution $y \in C^0([0,T]; L^2(\mathcal{O}))$ and $h \in C^0([0,T]; H^1(\mathbb{T}))$ of (4.3.3) satisfies*

$$y(T,\cdot) = 0 \quad \text{in } \mathcal{O} \qquad \text{and} \qquad h(T,\cdot) = 0 \quad \text{in } \mathbb{T}.$$

Moreover, there exists a positive constant $\mathfrak{C}(T,\sigma) = \mathfrak{C}(T,\omega,\Omega,\sigma) > 0$ such that

$$\|u\|_{L^2((0,T)\times\omega)} \leqslant \mathfrak{C}(T,\sigma) \left\| (y^0, h^0) \right\|_{L^2(\mathcal{O})\times H^1(\mathbb{T})},$$

with $\mathfrak{C}(T,\sigma) = M_1 e^{\frac{M_2}{T}}$ for some constants $M_1 = M_1(\omega,\Omega,\sigma) > 0$ and $M_2 > 0$ whenever $T \ll 1$.

Remark 4.3.3. *Before proceeding with the proof, let us provide some relevant comments.*

- *Of course, the first idea for proving Theorem 4.2 one could have is to write the adjoint system and prove an observability inequality for all solutions of this system, which in turn would imply the coercivity and continuity of the HUM functional. Note that however, the explicit form of the adjoint of the linear system is not*

[4] We see $\mathcal{A}_0$ as a compact perturbation of the operator $\mathcal{A}^\sharp \begin{bmatrix} y \\ h \end{bmatrix} = \begin{bmatrix} \partial_{x_2}^2 y \\ 0 \end{bmatrix}$, which has the same domain and generates an analytic semigroup.

obvious, and one in particular should look for the adjoint in the state space $\mathcal{H} = L^2(\mathcal{O}) \times H^1(\mathbb{T})$*, which makes the computations less straightforward.*

In fact, the issue in the latter is very specific to the topology regarding the second component. Indeed, if we rather consider the state space as $\mathcal{H} = L^2(\mathcal{O}) \times H^1_{\text{mean}}(\mathbb{T})$*, where* $H^1_{\text{mean}}(\mathbb{T})$ *is the space of* H^1 *functions with zero mean, endowed with the inner product*

$$\left\langle \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \right\rangle_{\mathcal{H}} := \langle f_1, g_1 \rangle_{L^2(\mathcal{O})} + \sigma \langle \partial_{x_1} f_2, \partial_{x_1} g_2 \rangle_{L^2(\mathbb{T})},$$

then one can readily see that $\mathcal{A}$ *is symmetric and thus self-adjoint. But when one considers the canonical inner product on the state space* $\mathcal{H} = L^2(\mathcal{O}) \times H^1(\mathbb{T})$*, it does not appear obvious what one may do with the remainder term*

$$\int_{\mathbb{T}} \partial_{x_2} f_1(x_1, 1) g_2(x_1) \, \mathrm{d}x_1.$$

On the other hand, the caveat of working spaces of with mean zero functions appears in the study of the nonlinear problem, as our strategy will be oriented towards using a Banach fixed-point argument (rather than, say, Schauder, which is commonly used in works on the controllability of compressible Navier Stokes where mean zero spaces are ubiqutous). This motivates the usage of the Fourier decomposition arguments we use herein.

- *When* $\sigma = 0$*, the linear system* (4.3.1) *is of cascade type as in the one-dimensional case addressed in [102, 116], and the equation for* y *can be solved without knowing* h*. In particular,*

$$h(t, x_1) = h_0(x_1) + \int_0^t \partial_{x_2} y(\tau, x_1, 1) \, \mathrm{d}\tau \qquad \text{for } x_1 \in \mathbb{T}.$$

The above expression implies that the null-controllability requirement for the second component, i.e. $h(T, \cdot) = 0$ *in* $\mathbb{T}$*, can equivalently be rewritten as*

$$\int_0^T \partial_{x_2} y(\tau, x_1, 1) \, \mathrm{d}\tau = h_0(x_1) \qquad \text{for } x_1 \in \mathbb{T}. \tag{4.3.6}$$

It may thus be seen that solving the control problem (4.3.2) *for* (4.3.1) *is equivalent to the null-controllability of the linear heat equation with the trace constraint* (4.3.6)*. Such questions have been investigated for more general linear control problems when the constraint is finite-dimensional, see [91]. However, as in this geometrical setting the constraint is not finite-dimensional, it is not straightforward to say that the null-controllability of the second component, i.e.* (4.3.6)*, follows immediately by arguing as in the one-dimensional case.*

Moreover, the case $\sigma = 0$ *in particular removes the regularizing effect that* $\partial^2_{x_1} h$ *has on the problem, and thus one cannot expect to readily solve the full system in the state space* $L^2(\mathcal{O}) \times H^1(\mathbb{T})$ *– rather, the height function* h *should be sought in a fractional Sobolev space such as* $H^s(\mathbb{T})$ *for* $s \in \left[0, \frac{1}{2}\right)$*, where the linear operator generates an analytic semigroup.*

To prove theorem 4.2, we will make use of the periodicity with respect to the x_1 variable of the functions appearing in (4.3.3). Such ideas have been exploited in different control contexts, see [21] for instance. We write the Fourier series expansions

$$y(t, x_1, x_2) = \frac{1}{\sqrt{2\pi}} \sum_{n \in \mathbb{Z}} \widehat{y}_n(t, x_2) e^{i n x_1} \qquad \text{in } (0, T) \times \mathcal{O}$$

$$h(t, x_1) = \frac{1}{\sqrt{2\pi}} \sum_{n \in \mathbb{Z}} \widehat{h}_n(t) e^{i n x_1} \qquad \text{in } (0, T) \times \mathbb{T}$$

$$u(t, x_1, x_2) = \frac{1}{\sqrt{2\pi}} \sum_{n \in \mathbb{Z}} \widehat{u}_n(t, x_2) e^{i n x_1} \qquad \text{in } (0, T) \times \omega$$

where $\left\{(2\pi)^{-1/2}e^{in\cdot}\right\}_{n\in\mathbb{Z}}$ denotes the orthonormal basis of $L^2(\mathbb{T})$, and recall that the Fourier coefficients appearing above series are given by $\widehat{\psi}_n(x) = (2\pi)^{-1/2}\langle\psi(x,\cdot),\, e^{in\cdot}\rangle_{L^2(\mathbb{T})}$ for all $n\in\mathbb{Z}$. It is readily seen that for any $n\in\mathbb{Z}$, the Fourier coefficients defined just above satisfy the following system of equations

$$\begin{cases} \partial_t\widehat{y}_n - \partial_{x_2}^2\widehat{y}_n + n^2\widehat{y}_n = \widehat{u}_n\mathbf{1}_{\left(-\frac{3}{4},-\frac{1}{4}\right)} & \text{in } (0,T)\times(-1,1) \\ \widehat{h}_n'(t) = \partial_{x_2}\widehat{y}_n(t,1) & \text{in } (0,T) \\ \widehat{y}_n(t,-1) = 0 & \text{in } (0,T) \\ \widehat{y}_n(t,1) = -\sigma n^2\widehat{h}_n(t) & \text{in } (0,T) \\ \left(\widehat{y}_n,\widehat{h}_n\right)_{|_{t=0}} = \left(\widehat{y}_n^0,\widehat{h}_n^0\right) & \text{in } (-1,1), \end{cases} \tag{4.3.7}$$

where $\left(\widehat{y}_n^0,\widehat{h}_n^0\right)$ denote the Fourier coefficients of the initial datum $\left(y^0,h^0\right)$.

Our objective in what follows is to prove the following controllability result for the Fourier coefficients with a control cost which is uniform in $n\in\mathbb{Z}$. This will allow us to simply sum up all of the coefficients and deduce Theorem 4.2. This is reflected by the proposition just below.

Proposition 4.3.4. *Let $T>0$ and $\sigma>0$ be fixed, and suppose that Assumption 4.3.8 holds true. For any $n\in\mathbb{Z}$ and for any pair $\left(\widehat{y}_n^0,\widehat{h}_n^0\right)\in L^2(-1,1)\times\mathbb{R}$, there exists a control $\widehat{u}_n\in L^2\left((0,T)\times\left(-\frac{3}{4},-\frac{1}{4}\right)\right)$ such that the corresponding pair of solutions $\widehat{y}_n\in C^0([0,T];L^2(-1,1))$ and $\widehat{h}_n\in C^0([0,T])$ to (4.3.7) satisfy*

$$\widehat{y}_n(T,\cdot) = 0 \quad \text{in } (-1,1) \quad \text{and} \quad \widehat{h}_n(T) = 0.$$

Moreover, there exist a constant $\mathfrak{C}(T,\sigma)>0$ such that

$$\|\widehat{u}_n\|_{L^2\left((0,T)\times\left(-\frac{3}{4},-\frac{1}{4}\right)\right)} \leqslant \mathfrak{C}(T,\sigma)\left(\left\|\widehat{y}_n^0\right\|_{L^2(-1,1)} + \left|n\widehat{h}_n^0\right|\right)$$

holds for all $n\in\mathbb{Z}$, with $\mathfrak{C}(T,\sigma) = M_1 e^{\frac{M_2}{T}}$ for some constants $M_1 = M_1(\omega,\Omega,\sigma)>0$ and $M_2>0$ whenever $T\ll 1$.

To prove Proposition 4.3.4 on the other hand, when $n\neq 0$ we a customary duality argument induced by the Hilbert Uniqueness Method (HUM), which renders the controllability problem of Proposition 4.3.4 equivalent to a proof of an observability inequality for the adjoint system. The observability of the adjoint system will be shown by means of spectral arguments which come with a slight degree of difficulty. For the zeroth mode $n=0$, we shall note that the eigenfunctions of the governing linear operator are not orthogonal (it is not self-adjoint), but in particular, the system is of cascade type and falls into the setting of [116].

Let us begin by defining the problem setup. We consider the Hilbert state space

$$\mathcal{H}_{\sigma,n} = L^2(-1,1)\times\mathbb{R},$$

which we endow with the inner product

$$\langle f,g\rangle_{\mathcal{H}_{\sigma,n}} := \langle f_1,g_1\rangle_{L^2(-1,1)} + \sigma n^2 f_2 g_2,$$

for any $n\in\mathbb{Z}^*$ and the canonical inner product when $n=0$. We then define, for any $n\in\mathbb{Z}$, the operator $\mathcal{A}_n : \mathcal{D}(\mathcal{A}_n)\to\mathcal{H}_{\sigma,n}$ by

$$\mathcal{A}_n\begin{bmatrix} y \\ h\end{bmatrix} = \begin{bmatrix} \partial_{x_2}^2 y - n^2 y \\ \partial_{x_2} y(1)\end{bmatrix}, \qquad \begin{bmatrix} y \\ h\end{bmatrix}\in\mathcal{D}(\mathcal{A}_n)$$

with domain

$$\mathcal{D}(\mathcal{A}_n) = \left\{ \begin{bmatrix} y \\ h \end{bmatrix} \in H^2(-1,1) \times \mathbb{R} \;\middle|\; y(-1) = 0,\, y(1) = -\sigma n^2 h \right\}.$$

Let us also introduce the control operator $\mathcal{B} \in \mathcal{L}\left(L^2\left(-\frac{3}{4}, -\frac{1}{4}\right), X\right)$ defined by

$$\mathcal{B}u = \begin{bmatrix} u\mathbf{1}_{\left(-\frac{3}{4}, -\frac{1}{4}\right)} \\ 0 \end{bmatrix}$$

for $u \in L^2\left(-\frac{3}{4}, -\frac{1}{4}\right)$. We now note that, in fact, when $n \in \mathbb{Z}^*$, the operator $\mathcal{A}_n$ is self-adjoint due to the specific inner product we endowed to $\mathcal{H}_{\sigma,n}$ – this is illustrated in more detail in Lemma 4.3.5 just below. Since $\mathcal{A}_n : \mathcal{D}(\mathcal{A}_n) \to \mathcal{H}_{\sigma,n}$ clearly has compact resolvents, by the Hilbert-Schmidt theorem, it may be diagonalized in the sense that there exists an orthonormal basis of $\mathcal{H}_{\sigma,n}$ consisting of eigenfunctions of $\mathcal{A}_n$, associated to a decreasing sequence of eigenvalues. On another hand, when $n = 0$, we note that the adjoint $\mathcal{A}_0^* : \mathcal{D}(\mathcal{A}_0^*) \to X$ of $\mathcal{A}_0$ can be found to read

$$\mathcal{A}_0^* \begin{bmatrix} \zeta \\ r \end{bmatrix} = \begin{bmatrix} \partial_{x_2}^2 \zeta \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} \zeta \\ r \end{bmatrix} \in \mathcal{D}(\mathcal{A}_0^*)$$

with domain

$$\mathcal{D}(\mathcal{A}_0^*) = \left\{ \begin{bmatrix} \zeta \\ r \end{bmatrix} \in H^2(-1,1) \times \mathbb{R} \;\middle|\; \zeta(-1) = 0,\, \zeta(1) = -r \right\}.$$

Furthermore, the adjoint $\mathcal{B}^* \in \mathcal{L}\left(X, L^2\left(-\frac{3}{4}, -\frac{1}{4}\right)\right)$ of $\mathcal{B}$ is given by

$$\mathcal{B}^* \begin{bmatrix} \zeta \\ r \end{bmatrix} = \zeta_{|_{\left(-\frac{3}{4}, -\frac{1}{4}\right)}}$$

for $\begin{bmatrix} \zeta \\ r \end{bmatrix} \in X$.

To prove Proposition 4.3.4 it suffices to have explicit knowledge of the spectrum of $\mathcal{A}_n^*$ for $n \in \mathbb{Z}$, and in particular to track the its dependence on the parameter $n \in \mathbb{Z}$. We begin with the following elementary lemma.

Lemma 4.3.5. *Let $n \in \mathbb{Z}$ be fixed.*

1. *If $n \neq 0$, the operator $\mathcal{A}_n : \mathcal{D}(\mathcal{A}_n) \to \mathcal{H}_{\sigma,n}$ is self-adjoint, has compact resolvents, and its spectrum $\mathrm{sp}(\mathcal{A}_n)$ consists of only negative eigenvalues.*

2. *If $n = 0$, the spectrum $\mathrm{sp}(\mathcal{A}_0^*)$ of $\mathcal{A}_0^* : \mathcal{D}(\mathcal{A}_0^*) \to L^2(-1,1) \times \mathbb{R}$ consists of only nonpositive eigenvalues.*

Proof of Lemma 4.3.5. Let us first note that clearly $\mathcal{A}_n^* : \mathcal{D}(\mathcal{A}_n^*) \to \mathcal{H}_{\sigma,n}$ has compact resolvents for $n \in \mathbb{Z}$ – thus, it's spectrum $\mathrm{sp}(\mathcal{A}_n^*)$ is a discrete subset of $\mathbb{C}$. We separate the remainder of the proof in two parts distinguishing the value of n.

Part 1: $n \neq 0$**.** In this case, as mentioned just above, the operator $\mathcal{A}_n$ is self-adjoint. We nonetheless, for the sake of completeness, provide more detailed computations on the nature of the spectrum. Let us first show that, in fact, the spectrum $\mathrm{sp}(\mathcal{A}_n)$ is a subset of $\mathbb{R}$. Thus, let $\lambda \in \mathrm{sp}(\mathcal{A}_n)$ be arbitrary. So $\lambda \in \mathbb{C}$ is such that there exists a vector $(\zeta, r) \in \mathcal{D}(\mathcal{A}_n) \setminus \{0\}$ such that

$$\begin{cases} \lambda\zeta - \partial_{x_2}^2 \zeta + n^2 \zeta = 0 & \text{in } (-1,1) \\ \zeta(-1) = 0 \\ \zeta(1) = -\sigma n^2 r \\ \lambda r - \partial_{x_2} \zeta(1) = 0. \end{cases}$$

We now multiply the first equation by ζ and integrate, to obtain

$$\lambda \int_{-1}^{1} \zeta^2 \, \mathrm{d}x_2 + \int_{-1}^{1} |\partial_{x_2}\zeta|^2 \, \mathrm{d}x_2 - \partial_{x_2}\zeta(1)\zeta(1) + n^2 \int_{-1}^{1} \zeta^2 \, \mathrm{d}x_2 = 0.$$

Using the boundary conditions, this identity entails

$$\lambda \int_{-1}^{1} \zeta^2 \, \mathrm{d}x_2 + \int_{-1}^{1} |\partial_{x_2}\zeta|^2 \, \mathrm{d}x_2 + \lambda\sigma n^2 r^2 + n^2 \int_{-1}^{1} \zeta^2 \, \mathrm{d}x_2 = 0. \tag{4.3.8}$$

Taking the imaginary part in the above identity, we deduce

$$\Im(\lambda) \left(\int_{-1}^{1} \zeta^2 \, \mathrm{d}x_2 + \sigma n^2 r^2 \right) = 0.$$

Hence $\lambda \in \mathbb{R}$, and thus $\mathrm{sp}(\mathcal{A}_n) \subset \mathbb{R}$. Let us now conclude by showing that $\mathrm{sp}(\mathcal{A}_n) \subset (-\infty, 0)$. Suppose that $\lambda \geqslant 0$. From (4.3.8) we deduce

$$\lambda \left(\int_{-1}^{1} \zeta^2 \, \mathrm{d}x_2 + \sigma n^2 r^2 \right) < 0.$$

This is clearly a contradiction, and hence $\lambda \in (-\infty, 0)$.

Part 2: $n = 0$**.** Let $\lambda \in \mathrm{sp}(\mathcal{A}_0^*)$ be arbitrary – namely, $\lambda \in \mathbb{C}$ is such that there exists a vector $(\zeta, r) \in \mathcal{D}(\mathcal{A}_0^*) \setminus \{0\}$ such that

$$\begin{cases} \lambda\zeta - \partial_{x_2}^2 \zeta = 0 & \text{in } (-1,1) \\ \zeta(-1) = 0 \\ \zeta(1) = -r \\ \lambda r = 0. \end{cases}$$

We thus have two cases to distinguish: either $\lambda = 0$ and the conclusion follows; or $\lambda \neq 0$, in which case $r = 0$, and thus λ is an element of the spectrum of the Dirichlet Laplacian with $\zeta \not\equiv 0$, and hence, $\lambda \in (-\infty, 0)$. This concludes the proof. □

We however need to explicitly characterize the spectrum of $\mathcal{A}_n^*$. This is the goal of the following result.

Lemma 4.3.6. *Let $\sigma > 0$ and $n \in \mathbb{Z}^*$ be fixed. The sequence $\{\lambda_{n,k}\}_{k=0}^{+\infty}$, $\lambda_{n,k} < 0$, of eigenvalues of $\mathcal{A}_n : \mathcal{D}(\mathcal{A}_n) \longrightarrow \mathcal{H}_{\sigma,n}$ is regular uniformly in $n \in \mathbb{Z}^*$ in the sense that*

$$\inf_{k \geqslant 0} |\lambda_{n,k+1} - \lambda_{n,k}| > s, \tag{4.3.9}$$

for some $s > 0$ independent of n and σ. Moreover,

$$-\lambda_{n,k} = rk^2 + n^2 + \underset{k \to +\infty}{\mathcal{O}}(k) \tag{4.3.10}$$

for some $r > 0$ independent of n and σ. Furthermore, there exists a constant $c(\sigma) > 0$ such that for any $n \in \mathbb{Z}^*$ and $k \geqslant 0$, the normalized eigenfunctions $\Phi_{n,k}$ of $\mathcal{A}_n$ satisfy

$$\|\mathcal{B}^* \Phi_{n,k}\|_{L^2\left(-\frac{3}{4}, -\frac{1}{4}\right)} \geqslant c(\sigma). \tag{4.3.11}$$

Proof of Lemma 4.3.6. We recall that $\mathcal{A}_n : \mathcal{D}(\mathcal{A}_n) \to \mathcal{H}_{\sigma,n}$ is self-adjoint, has compact resolvents, and its spectrum consists of a decreasing sequence of negative eigenvalues, namely a sequence $\{\lambda_{n,k}\}_{k=0}^{+\infty}$ with $-\infty < \ldots \leqslant \lambda_{n,k} \leqslant \ldots \leqslant \lambda_{n,0} < 0$. We shall distinguish two different scenarios.

1). Should $\lambda < -n^2$. Suppose that $\lambda \in (-\infty, 0)$ is an eigenvalue of $\mathcal{A}_n$ satisfying $\lambda < -n^2$, so that there exists a vector $(\zeta, r) \in \mathcal{D}(\mathcal{A}_n) \setminus \{0\}$ such that

$$\begin{cases} -\partial_{x_2}^2 \zeta - (-\lambda - n^2)\zeta = 0 \quad \text{in} \quad (-1,1) \\ \zeta(-1) = 0 \\ \zeta(1) = -\sigma n^2 r \\ \partial_{x_2}\zeta(1) = \lambda r. \end{cases} \tag{4.3.12}$$

In other words, ζ would solve the mixed Dirichlet-Robin problem

$$\begin{cases} -\partial_{x_2}^2 \zeta - (-\lambda - n^2)\zeta = 0 \quad \text{in} \quad (-1,1) \\ \zeta(-1) = 0 \\ \zeta(1) - \dfrac{\sigma n^2}{-\lambda}\partial_{x_2}\zeta(1) = 0. \end{cases} \tag{4.3.13}$$

Since $-\lambda - n^2 > 0$, one may readily see that the solutions to (4.3.13) are of the form

$$\zeta(x_2) = c\sin(\nu(1 + x_2)),$$

with $c > 0$, where $\nu := \sqrt{-\lambda - n^2}$ is the positive root of the transcendental equation

$$\left(\frac{\nu^2}{n^2} + 1\right)\tan(2\nu) = \sigma\nu. \tag{4.3.14}$$

Studying the positive roots of this equation suggest studying the fixed points of $f(\nu) = \left(\frac{\nu^2}{n^2} + 1\right)\tan(2\nu)$, defined and non-decreasing on the union of consecutive intervals of the form

$$\bigcup_{k=1}^{+\infty} \left(\frac{\pi}{4} + \frac{(k-1)\pi}{2}, \frac{\pi}{4} + \frac{k\pi}{2}\right).$$

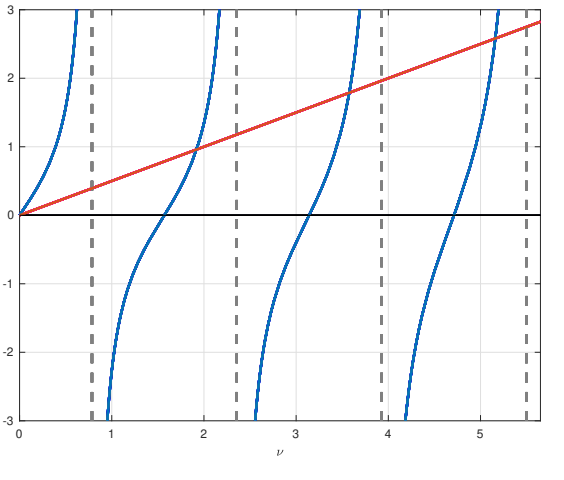


Figure 4.2: The function $f(\mu) = \left(\frac{\nu^2}{n^2} + 1\right)\tan(2\mu)$ in blue and $\nu \longmapsto \sigma x$ in red, with $n = 5$ and $\sigma = 0.5$. We see how the fixed points of f are localized over each subinterval.

Moreover, for $k \geqslant 1$,

$$\lim_{\nu \searrow \frac{\pi}{4} + \frac{(k-1)\pi}{2}} f(\nu) = -\infty, \quad f\left(\frac{k\pi}{2}\right) = 0, \quad \lim_{\nu \nearrow \frac{\pi}{4} + \frac{k\pi}{2}} f(\nu) = +\infty.$$

Thus, (4.3.14) has a sequence of positive roots $\{\nu_k\}_{k=1}^{+\infty}$ of the form

$$\nu_k = \frac{k\pi}{2} + \frac{\pi}{4} - \omega_k \tag{4.3.15}$$

for $k \geqslant 1$, where $\omega_k \in \left(0, \frac{\pi}{4}\right)$ may a priori depend on σ and n. Consequently, the eigenvalues $\lambda_{n,k}$ in this case are of the form

$$-\lambda_{n,k} = \left(\frac{k\pi}{2} + \frac{\pi}{4} - \omega_k\right)^2 + n^2 \tag{4.3.16}$$

for $k \geqslant 1$.

2). Should $\lambda \in [-n^2, 0)$. First of all, note that since $\sigma \in (0,1)$, $-n^2$ cannot be an eigenvalue of $\mathcal{A}_n$[5]. Hence we suppose that $\lambda \in (-n^2, 0)$ is an eigenvalue of $\mathcal{A}_n$, so that there exists a vector $(\zeta, r) \in \mathcal{D}(\mathcal{A}_n) \setminus \{0\}$ such that

$$\begin{cases} -\partial_{x_2}^2 \zeta - (-\lambda - n^2)\zeta = 0 \quad \text{in} \quad (-1,1) \\ \zeta(-1) = 0 \\ \zeta(1) = -\sigma n^2 r \\ \partial_{x_2}\zeta(1) = \lambda r. \end{cases}$$

Then ζ would again solve a mixed Dirichlet-Robin problem

$$\begin{cases} -\partial_{x_2}^2 \zeta - (-\lambda - n^2)\zeta = 0 \quad \text{in} \quad (-1,1) \\ \zeta(-1) = 0 \\ \zeta(1) - \dfrac{\sigma n^2}{-\lambda}\partial_{x_2}\zeta(1) = 0. \end{cases} \tag{4.3.17}$$

Since $-\lambda - n^2 < 0$, one may readily see that the solutions to (4.3.17) are of the form

$$\zeta(x_2) = ce^{\nu x_2} - ce^{-2\nu}e^{-\nu x_2},$$

with $c > 0$, where $\nu := \sqrt{n^2 + \lambda}$ is the positive root of the transcendental equation

$$e^{\nu} - e^{-3\nu} - \frac{\sigma n^2}{n^2 - \nu^2}\left(\nu e^{\nu} + \nu e^{-3\nu}\right) = 0$$

in $(0, |n|)$. We may rewrite the above equation as

$$\left(n^2 - \nu^2 - \sigma n^2 \nu\right) - \left(n^2 - \nu^2 + \sigma n^2 \nu\right) e^{-4\nu} = 0.$$

We claim that the function $f(\nu) = \left(n^2 - \nu^2 - \sigma n^2 \nu\right) - \left(n^2 - \nu^2 + \sigma n^2 \nu\right) e^{-4\nu}$ has a unique root[6] in $(0, |n|)$. Indeed, first note that by some elementary computations,

$$f(\nu) < 0 \qquad \text{for} \qquad \nu \in \left[\frac{1}{2}\left(|n|\sqrt{n^2\sigma^2 + 4} - n^2\sigma\right), |n|\right).$$

On another hand, we note that $f_1(\nu) := \left(n^2 - \nu^2 - \sigma n^2 \nu\right)$ satisfies

$$f_1'(\nu) = -(2\nu + \sigma n^2) < 0 \qquad \text{for} \quad \nu \in (0, |n|),$$

[5] Indeed, if this were the case, then ζ in (4.3.12) would be harmonic and thus an affine function, and its coefficients would equal zero unless $\sigma = |[-1,1]| = 2$.

[6] One may in fact try to compute this root by using special functions such as the Lambert W function; we omit this from our work as it is not necessary to our analysis and to avoid additional technical details.

whereas $f_2(\nu) := -\left(n^2 - \nu^2 + \sigma n^2 \nu\right) e^{-4\nu}$ satisfies

$$\begin{aligned} f_2'(\nu) &= (4n^2 - 4\nu^2 + 2\nu + 4\sigma n^2 \nu - \sigma n^2) e^{-4\nu} \\ &= -\left(\nu - \frac{1}{4}\left(2n^2\sigma - \sqrt{4n^4\sigma^2 + 16n^2 + 1} + 1\right)\right) \\ &\quad \times \left(\nu - \frac{1}{4}\left(2n^2\sigma + \sqrt{4n^4\sigma^2 + 16n^2 + 1} + 1\right)\right) e^{-4\nu}. \end{aligned}$$

One then sees that $f_2'(\nu) < 0$ for $\nu \in \left(0, \frac{1}{2}\left(|n|\sqrt{n^2\sigma^2 + 4} - n^2\sigma\right)\right)$, thus f can only have one root $\nu_0 \in (0, |n|)$, which in fact lies in the previous interval. Hence, the first eigenvalue $\lambda_{n,0}$ of $\mathcal{A}_n$ will have the form

$$\lambda_{n,0} = \nu_0^2 - n^2. \tag{4.3.18}$$

Note that since

$$\nu_0 \leqslant \frac{1}{2}\left(|n|\sqrt{n^2\sigma^2 + 4} - n^2\sigma\right) = \frac{4n^2}{2\left(|n|\sqrt{n^2\sigma^2 + 4} + n^2\sigma\right)} \leqslant \frac{1}{\sigma},$$

we see that $\nu_0 \in \left(0, \frac{1}{\sigma}\right)$ for all $n \in \mathbb{Z}^*$.

We thus collect the sequence of eigenvalues $\{\lambda_{n,k}\}_{k=0}^{+\infty}$, with $\lambda_{n,0} \in (-n^2, 0)$ defined in (4.3.18) and λ_k with $k \geqslant 1$ defined in (4.3.16). One thus readily sees that (4.3.10) holds. On another hand, since $\omega_k \in \left(0, \frac{\pi}{4}\right)$, we see that for $k \geqslant 1$,

$$\begin{aligned} -\lambda_{n,k+1} + \lambda_{n,k} &= \left(\frac{(k+1)\pi}{2} + \frac{\pi}{4} - \omega_{k+1}\right)^2 - \left(\frac{k\pi}{2} + \frac{\pi}{4} - \omega_k\right)^2 \\ &= (k\pi + \pi - \omega_{k+1} - \omega_k)\left(\frac{\pi}{2} - \omega_{k+1} + \omega_k\right) \\ &\geqslant \left(\frac{3\pi}{2}\right)\left(\frac{\pi}{4}\right) = \frac{3\pi^2}{8}. \end{aligned}$$

Furthermore,

$$-\lambda_{n,1} + \lambda_{n,0} = \left(\frac{\pi}{2} + \frac{\pi}{4} - \omega_k\right)^2 + \nu_0^2 \geqslant \frac{\pi^2}{4}.$$

Hence, (4.3.9) holds as well.

Let us finally prove (4.3.11) in the case $n \neq 0$. We recall that the normalized eigenfunctions $\Phi_{n,k}$ have the form

$$\Phi_{n,k} = \begin{bmatrix} \zeta_k \\ -\zeta_k(1) \end{bmatrix},$$

and are associated to an eigenvalue λ_k given by (4.3.16) when $k \geqslant 0$ and (4.3.18) when $k = 0$, while ζ_k is given by

$$\zeta_k(x_2) = c_k \sin\left(\sqrt{-\lambda_k - n^2}\,(1 + x_2)\right), \qquad x_2 \in (-1, 1)$$

for $k \geqslant 1$, and

$$\zeta_0(x_2) = c_0\left(e^{\sqrt{n^2+\lambda_0}\,x_2} - e^{-\sqrt{n^2+\lambda_0}(2+x_2)}\right) \qquad x_2 \in (-1, 1).$$

Let us first suppose $k \geqslant 1$. Reusing the notation $\nu_k := \sqrt{-\lambda_k - n^2} > 0$, we note that in order to ensure that the eigenfunctions $\Phi_{n,k}$ are of norm 1, $c_k > 0$ needs to satisfy

$$c_k^2\left(1 - \frac{\sin(4\nu_k)}{4\nu_k}\right) + c_k^2 \sin^2(2\nu_k) = 1$$

for all $k \geqslant 1$. On the other hand, again using $\mathcal{B}^*\Phi_{n,k} = \zeta_k|_{\left(-\frac{3}{4},-\frac{1}{4}\right)}$, we have

$$\|\mathcal{B}^*\Phi_{n,k}\|^2_{L^2\left(-\frac{3}{4},-\frac{1}{4}\right)} = \frac{c_k^2}{2}\left(1 + \frac{\sin\left(\frac{\nu_k}{2}\right)}{2\nu_k} - \frac{\sin\left(\frac{5\nu_k}{2}\right)}{2\nu_k}\right). \tag{4.3.19}$$

In view of (4.3.15), we deduce that there exists $\delta_1 > 0$ independent of $n \in \mathbb{Z}^*$ such that

$$1 + \frac{\sin\left(\frac{\nu_k}{2}\right)}{2\nu_k} - \frac{\sin\left(\frac{5\nu_k}{2}\right)}{2\nu_k} > \delta_1 \qquad \text{for all } k \geqslant 1. \tag{4.3.20}$$

Therefore, we see from (4.3.19) and (4.3.20) that, in order to obtain (4.3.11), it suffices to have an appropriate lower bound on c_k for all $k \geqslant 1$. To this end, we note that

$$c_k^2 = \left(\left(1 - \frac{\sin(4\nu_k)}{4\nu_k}\right) + \sin^2(2\nu_k)\right)^{-1}.$$

By virtue of (4.3.15), we see that

$$\frac{\sin(4\nu_k)}{4\nu_k} \underset{k\to\infty}{\longrightarrow} 0 \qquad \text{and} \qquad \sin^2(2\nu_k) \underset{k\to\infty}{\longrightarrow} 0,$$

hence

$$c_k^2 \geqslant c^\circ$$

for some $c^\circ > 0$ independent of n and σ. This concludes the proof of (4.3.11) when $k \geqslant 1$.

On another hand, when $k = 0$, we see that to ensure orthonormality, $c_0 > 0$ needs to satisfy

$$c_0^2 \frac{\sinh(2\nu_0) - 4\nu_0}{\nu_0} e^{-2\nu_0} + c_0^2\left(e^{\nu_0} - e^{-3\nu_0}\right)^2 = 1,$$

thus

$$c_0^2 = \left(\frac{\sinh(2\nu_0) - 4\nu_0}{\nu_0} e^{-2\nu_0} + \left(e^{\nu_0} - e^{-3\nu_0}\right)^2\right)^{-1}. \tag{4.3.21}$$

We also have

$$\|\mathcal{B}^*\Phi_{n,0}\|^2_{L^2\left(-\frac{3}{4},-\frac{1}{4}\right)} = c_0^2 \frac{\left(\sinh\left(\frac{3\nu_0}{2}\right) - \sinh\left(\frac{\nu_0}{2}\right) - \nu_0\right)}{\nu_0} e^{-2\nu_0}. \tag{4.3.22}$$

In view of (4.3.21), and since $\frac{\sinh(2x)}{x} \leqslant C_{1,\sigma}$ for $x \in \left(0, \frac{1}{\sigma}\right)$, we see that

$$c_0^2 \geqslant \left((C_{1,\sigma} - 4)e^{-\frac{4}{\sigma}} + 2e^{\frac{2}{\sigma}} + 1\right)^{-1}$$

for all $n \in \mathbb{Z}^*$. One may similarly, using the continuity and the positivity of the function $x \mapsto \frac{\sinh(\frac{3x}{2}) - \sinh(\frac{x}{2}) - x}{x}$ on $\left(0, \frac{1}{\sigma}\right)$, we conclude that there exists $C_{2,\sigma} > 0$ with $C_{2,\sigma} \longrightarrow 0$ as $\sigma \searrow 0$ such that

$$\|\mathcal{B}^*\Phi_{n,0}\|^2_{L^2\left(-\frac{3}{4},-\frac{1}{4}\right)} \geqslant C_{2,\sigma},$$

holds for all $n \in \mathbb{Z}^*$. This concludes the proof. □

Remark 4.3.7. *Before proceeding with the concluding proofs of the linear problems, let us comment on the above proof.*

- *Note that the lower bound $C_{2,\sigma} > 0$ of the quantity $\|\mathcal{B}^*\Phi_{n,0}\|^2_{L^2\left(-\frac{3}{4},-\frac{1}{4}\right)}$ appearing in the proof collapses as $\sigma \searrow 0$, i.e. $C_{2,\sigma} \longrightarrow 0$ as $\sigma \searrow 0$. Whilst this does not prove the lack of null-controllability of the linearized classical Stefan problem, namely the linear problem when $\sigma = 0$, it could stipulate a possible obstruction in obtaining this zero surface tension limit for the control problem.*

- *When $n = 0$, we see that $\mathcal{A}_0^*$ will render its components decoupled. In terms of the spectrum, let $\lambda \leqslant 0$ be an eigenvalue of $\mathcal{A}_0^*$, so that there exists a vector $(\zeta, r) \in \mathcal{D}(\mathcal{A}_0^*) \setminus \{0\}$ such that*

$$\begin{cases} \partial_{x_2}^2 \zeta = \lambda \zeta & \text{in } (-1,1) \\ \zeta(-1) = 0 \\ \zeta(1) = -r \\ \lambda r = 0. \end{cases} \tag{4.3.23}$$

It is readily seen that (4.3.23) yields the normalized eigenfunctions $\Phi_{0,k} = \begin{bmatrix} \zeta_k \\ r_k \end{bmatrix}$, where

$$\zeta_k(x_2) = \begin{cases} c_0 \zeta_0(x_2) & \text{for } k = 0 \\ c_k \sin\left(\dfrac{k\pi}{2}(1+x_2)\right) & \text{for } k \geq 1 \end{cases} \quad \text{and} \quad r_k = \begin{cases} c_0 r_0 & \text{for } k = 0 \\ 0 & \text{for } k \geq 0 \end{cases},$$

associated to the eigenvalues $\{\lambda_k\}_{k=0}^{+\infty} = \left\{-\frac{k^2\pi^2}{4}\right\}_{k=0}^{+\infty}$, for some $\zeta_0 \not\equiv 0$ and $r_0 \not\equiv 0$. One can readily see that the normalization constant $c_k > 0$ takes the form

$$c_k = \begin{cases} c_0 & \text{for } k = 0 \\ 1 & \text{for } k \geq 1. \end{cases}$$

Hence, the eigenfunctions of $\mathcal{A}_0$ are a priori not orthogonal, and we may not directly apply spectral techniques to deduce the controllability properties of the linear system which is governed by $\mathcal{A}_0$.

We may conclude this study with the proof of Proposition 4.3.4. We insist[7] that we will make use of the following assumption on the control cost for the zeroth mode solely to be able to add the source terms via the source-term method (which necessitates exponential cost in small times) ahead of the nonlinear study. Assumption 4.3.8 is a relatively pessimistic hypothesis, and we envisage to prove it using perturbation arguments as done in [135] for instance, by exploiting the uniform control cost of the remaining Fourier coefficients systems with respect to $n \in \mathbb{Z}^*$.

By virtue of [116], we know that the control $\widehat{u}_0$ steering $\widehat{y}_0$ and $\widehat{h}_0$ to 0 in time $T > 0$ is such that there exists a constant $\mathfrak{C}_0(T, \sigma) > 0$ such that

$$\|\widehat{u}_0\|^2_{L^2\left(0,T;L^2\left(-\frac{3}{4},-\frac{1}{4}\right)\right)} \leqslant \mathfrak{C}_0(T,\sigma) \left\| \left(\widehat{y}^0, \widehat{h}^0\right) \right\|^2_{L^2(-1,1)\times\mathbb{R}}. \tag{4.3.24}$$

Assumption 4.3.8 (Control cost of zeroth mode)**.** *We shall assume that there exist positive constants $M = M_1(\sigma) > 0$ and $M_2 = M_2(\sigma) > 0$ such that*

$$\mathfrak{C}_0(T,\sigma) = M_1 e^{\frac{M_2}{T}} \qquad \text{for } T \ll 1,$$

where $\mathfrak{C}_0(T,\sigma) > 0$ is the constant appearing in (4.3.24).

Proof of Proposition 4.3.4. We again split the proof in two separate cases.

Case 1: $n = 0$**.** The proof follows directly from the results shown in [116]. To obtain the exponential bound on the control cost for small times, we use Assumption 4.3.8.

[7] The exponential bound on the control cost entailed by Theorem 4.1 holds without any assumption if one furthermore supposes that y^0 and h^0 are of zero mean over $\mathbb{T}$ – this would entail that the zero mode $n = 0$ does not appear in the projected systems. However, as elaborated in a previous remark regarding the symmetry of the operator $\mathcal{A}$, looking for solutions which live in Sobolev spaces of zero mean is a clear impediment in the application of a Banach fixed-point argument for the nonlinear system, as there is no reason to guarantee that the quadratic nonlinearities will be of zero mean.

Case 2: $n \neq 0$**.** Let $n \neq 0$ be fixed. We rewrite (4.3.7) as

$$\begin{cases} \dot{z} = \mathcal{A}_n z + \mathcal{B}\,\widehat{u}_n & \text{in } (0,T) \\ z(0) = z^0, \end{cases} \tag{4.3.25}$$

where $z = \left(\widehat{y}_n, \widehat{h}_n\right)$ and $z^0 = \left(\widehat{y}_n^0, \widehat{h}_n^0\right)$. Since the operator $\mathcal{A}_n : \mathcal{D}(\mathcal{A}_n) \longrightarrow \mathcal{H}_{\sigma,n}$ is self-adjoint, by virtue of the Hilbert Uniqueness Method, (4.3.25) is null-controllable by means of a control $\widehat{u}_n$ satisfying

$$\|\widehat{u}_n\|^2_{L^2\left(0,T;L^2\left(-\frac{3}{4},-\frac{1}{4}\right)\right)} \leqslant M_1 e^{\frac{M_2}{T}} \underbrace{\left(\left\|\widehat{y}_n^0\right\|^2_{L^2(-1,1)} + \sigma\left|n\widehat{h}_n^0\right|^2\right)}_{=\|z^0\|^2_{\mathcal{H}_{\sigma,n}}},$$

for some $M_1 > 0$ and $M_2 > 0$ independent of n, if and only if the observability inequality

$$M_1 e^{\frac{M_2}{T}} \int_0^T \|\mathcal{B}^*\zeta(t,\cdot)\|_{L^2\left(-\frac{3}{4},-\frac{1}{4}\right)}\,\mathrm{d}t \geqslant \|\zeta(0,\cdot)\|^2_{\mathcal{H}_{\sigma,n}} \tag{4.3.26}$$

holds for some $M_1 > 0$ and $M_2 > 0$ independent of n, and for all $\zeta_T \in \mathcal{H}_{\sigma,n}$, where ζ is the solution to the adjoint system

$$\begin{cases} -\dot{\zeta} = \mathcal{A}_n\zeta & \text{in } (0,T) \\ \zeta(T) = \zeta_T. \end{cases} \tag{4.3.27}$$

Since the operator $\mathcal{A}_n : \mathcal{D}(\mathcal{A}_n) \longrightarrow \mathcal{H}_{\sigma,n}$ is self-adjoint and negative, with an orthonormal basis of eigenfunctions $\{\Phi_{n,k}\}_{k=0}^{+\infty}$ and corresponding decreasing sequence of negative eigenvalues $\{-\lambda_{n,k}\}_{k=0}^{+\infty}$, we may write the Fourier decomposition of ζ as

$$\zeta(t,x_2) = \sum_{k=0}^{+\infty} e^{-\lambda_{n,k}(T-t)}\left\langle \zeta_T, \Phi_{n,k}\right\rangle_{\mathcal{H}_{\sigma,n}} \Phi_{n,k}(x_2).$$

Denoting $\{\psi_j\}_{j=0}^{+\infty}$ the orthonormal basis of $L^2\left(-\frac{3}{4},-\frac{1}{4}\right)$, and via the shift $T - t \longmapsto t$, we obtain

$$\begin{aligned} &\int_0^T \|\mathcal{B}^*\zeta(t,\cdot)\|_{L^2\left(-\frac{3}{4},-\frac{1}{4}\right)}\,\mathrm{d}t \\ &\quad = \sum_{j=0}^{+\infty}\int_0^T \left|\sum_{k=0}^{+\infty} e^{-\lambda_{n,k}t}\left\langle \zeta_T, \Phi_{n,k}\right\rangle_{\mathcal{H}_{\sigma,n}} \left\langle \mathcal{B}^*\Phi_{n,k}, \psi_j\right\rangle_{L^2\left(-\frac{3}{4},-\frac{1}{4}\right)}\right|^2 \mathrm{d}t. \end{aligned} \tag{4.3.28}$$

Now, making use of (4.3.9) and (4.3.10), we deduce from [255, Cor. 3.6] that there exist $M_1 > 0$ and $M_2 > 0$ depending only on $r > 0$ and $s > 0$ such that

$$M_1 e^{\frac{M_2}{T}} \int_0^T \left|\sum_{k=0}^{+\infty} a_k e^{-(\lambda_{n,k}-n^2)t}\right|^2 \mathrm{d}t \geqslant \sum_{k=0}^{+\infty} |a_k|^2 e^{-2(\lambda_{n,k}-n^2)T}$$

for any $\{a_k\}_{k=0}^{+\infty} \in \ell^2(\mathbb{R})$, and hence

$$\begin{aligned} M_1 e^{\frac{M_2}{T}} \int_0^T \left|\sum_{k=0}^{+\infty} a_k e^{-\lambda_{n,k}t}\right|^2 \mathrm{d}t &\geqslant M_1 e^{\frac{M_2}{T}} e^{-n^2T} \int_0^T \left|\sum_{k=0}^{+\infty} a_k e^{-(\lambda_{n,k}-n^2)t}\right|^2 \mathrm{d}t \\ &\geqslant e^{-n^2T} \sum_{k=0}^{+\infty} |a_k|^2 e^{-2(\lambda_{n,k}-n^2)T} \\ &= \sum_{k=0}^{+\infty} |a_k|^2 e^{-2\lambda_{n,k}T}. \end{aligned}$$

The above estimate combined with (4.3.28) implies that

$$M_1 e^{\frac{M_2}{T}} \int_0^T \|\mathcal{B}^*\zeta(t,\cdot)\|^2_{L^2(-\frac{3}{4},-\frac{1}{4})}\,\mathrm{d}t \geqslant \sum_{j=0}^{+\infty}\sum_{k=0}^{+\infty} e^{-2\lambda_{n,k}T}\left|\langle \zeta_T, \Phi_{n,k}\rangle_{\mathcal{H}_{n,\sigma}}\right|^2 \left|\langle \mathcal{B}^*\Phi_{n,k}, \psi_j\rangle_{L^2(-\frac{3}{4},-\frac{1}{4})}\right|^2.$$

Applying (4.3.11) to the above estimate, we deduce that

$$M_1 e^{\frac{M_2}{T}} \int_0^T \|\mathcal{B}^*\zeta(t,\cdot)\|^2_{L^2(-\frac{3}{4},-\frac{1}{4})}\,\mathrm{d}t \geqslant c(\sigma)\|\zeta(0,\cdot)\|^2_{\mathcal{H}_{n,\sigma}},$$

which holds for all $\zeta_T \in \mathcal{H}_{n,\sigma}$. This concludes the proof of (4.3.26), and thus the proof of the proposition. □

We conclude this section with the proof of Theorem 4.1 / Theorem 4.2.

Proof of Theorem 4.1 / Theorem 4.2. Let us define the control $u \in L^2(0,T;L^2(\omega))$ by

$$u(t,x_1,x_2) = \frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}} \widehat{u}_n(t,x_2)e^{inx_1} \qquad \text{in } (0,T)\times\omega,$$

where $\widehat{u}_n \in L^2(0,T;L^2(-\frac{3}{4},-\frac{1}{4}))$ are the controls provided by Proposition 4.3.4, thus such that $\widehat{y}_n$ and $\widehat{h}_n$, which solve (4.3.7), vanish at time $T>0$. Defining y and h via Fourier series similarly as u above, we readily see that (y,h) is the unique solution to (4.3.3) (equiv. (4.3.4)). Moreover, since all the Fourier coefficients of y and h vanish at time T, then also (y,h) vanishes at time T. The estimate on the control follows by summing up the estimate of the Fourier coefficient controls over all n, and using the fact that all the constants intervening in this estimate are independent of n. This concludes the proof. □

4.4 Control in spite of source terms

In view of tackling the controllability of the nonlinear system, we look to add the source terms over which we aim to apply a fixed point argument. Let us hence consider the following linear system

$$\begin{cases} \partial_t y - \Delta y = f_1 & \text{in } (0,T)\times\mathcal{O},\\ \partial_t h = \partial_{x_2} y + f_3 & \text{in } (0,T)\times\Gamma_{\mathrm{top}},\\ y = 0 & \text{on } (0,T)\times\Gamma_{\mathrm{bot}},\\ y = \sigma\partial^2_{x_1} h + f_2 & \text{on } (0,T)\times\Gamma_{\mathrm{top}},\\ (y,h)_{|t=0} = (y^0,h^0) & \text{in } \mathcal{O}\times\mathbb{T}. \end{cases} \tag{4.4.1}$$

Before proceeding with the control analysis, let us provide a necessary regularity result. We consider the subset of initial data

$$\mathfrak{I} := \left\{ \begin{bmatrix} y^0 \\ h^0 \end{bmatrix} \in H^1(\mathcal{O})\times H^{5/2}(\mathbb{T}) \mid y^0 = 0 \text{ on } \Gamma_{\mathrm{bot}} \right\},$$

as well as the space of source terms

$$\mathbf{E}_f(0,T;\mathcal{O}) := \left\{ (f_1,f_2,f_3) \in L^2(0,T;L^2(\mathcal{O}))\times H^{3/2,3/4}((0,T)\times\mathbb{T})\times H^{1/2,1/4}((0,T)\times\mathbb{T}) \right\}.$$

We also introduce the energy spaces for the state y:

$$\mathbf{E}_y := L^2\Big(0,T;H^2(\mathcal{O})\Big)\cap H^1\Big(0,T;L^2(\mathcal{O})\Big)\cap C^0\Big([0,T];H^1(\mathcal{O})\Big),$$

and for h:

$$\mathbf{E}_h := L^2\Big(0,T;H^{7/2}(\mathbb{T})\Big)\cap H^{3/4}\Big(0,T;H^2(\mathbb{T})\Big)\cap H^1\Big(0,T;H^1(\mathbb{T})\Big)\\ \cap H^{5/4}\Big(0,T;L^2(\mathbb{T})\Big)\cap C^0\Big([0,T];H^{5/2}(\mathbb{T})\Big).$$

The following improved well-posedness result then holds.

Proposition 4.4.1. *Let $\sigma>0$ and $T>0$ be fixed. For any $\big(y^0,h^0\big)\in L^2(\mathcal{O})\times H^1(\mathbb{T})$ and $(f_1,f_2,f_3)\in\mathbf{E}_f$, (4.4.1) admits a unique mild solution $(y,h)\in C^0([0,T];L^2(\mathcal{O})\times H^1(\mathbb{T}))$, and there exists a constant $C_T=C(T,\sigma)>0$ such that*

$$\|(y,h)\|_{C^0([0,T];L^2(\mathcal{O})\times H^1(\mathbb{T}))}\leqslant C_T\left(\big\|\big(y^0,h^0\big)\big\|_{L^2(\mathcal{O})\times H^1(\mathbb{T})}+\|(f_1,f_2,f_3)\|_{\mathbf{E}_f}\right). \tag{4.4.2}$$

If moreover $\big(y^0,h^0\big)\in\mathfrak{I}$ satisfies the compatibility condition

$$y^0=\sigma\partial_{x_1}^2h^0+f_2(0)\qquad \text{on }\Gamma_{\text{top}}, \tag{4.4.3}$$

then (4.4.1) admits a unique strong solution $(y,h)\in\mathbf{E}_y\times\mathbf{E}_h$, together with the estimate

$$\|(y,h)\|_{\mathbf{E}_y\times\mathbf{E}_h}\leqslant C_T\left(\big\|\big(y^0,h^0\big)\big\|_{\mathfrak{I}}+\|(f_1,f_2,f_3)\|_{\mathbf{E}_f}\right), \tag{4.4.4}$$

for some $C_T=C(T,\sigma)>0$.[8]

Proof of Proposition 4.4.1. The uniqueness of solutions follows easily. Thus we just focus on the existence part. Using standard trace results (see for instance [186]), there exists $y^\sharp\in\mathbf{E}_y$ such that

$$y^\sharp=f_2\ \text{ on }\Gamma_{\text{top}},\quad y^\sharp=0\text{ on }\Gamma_{\text{bot}},\quad \partial_{x_2}y^\sharp=f_3\text{ on }\Gamma_{\text{top}}.$$

Moreover, there exits a positive constant $C_T>0$ such that

$$\big\|y^\sharp\big\|_{\mathbf{E}_y}\leqslant C_T\,\|(f_1,f_2,f_3)\|_{\mathbf{E}_f}. \tag{4.4.5}$$

We look for y in the form $y=y^\dagger+y^\sharp$. Thus $(y^\dagger,h)$ satisfies the following system

$$\begin{cases}\partial_t y^\dagger-\Delta y^\dagger=f_1^\dagger & \text{in }(0,T)\times\mathcal{O},\\ y^\dagger=\sigma\partial_{x_1}^2h & \text{on }(0,T)\times\Gamma_{\text{top}},\\ y^\dagger=0 & \text{on }(0,T)\times\Gamma_{\text{bot}},\\ \partial_t h=\partial_{x_2}y^\dagger & \text{in }(0,T)\times\Gamma_{\text{top}},\\ \big(y^\dagger,h\big)_{|t=0}=\big(y^{\dagger,0},h^0\big) & \text{in }\ \mathcal{O}\ \times\mathbb{T}\end{cases} \tag{4.4.6}$$

where

$$f_1^\dagger=f_1-\partial_t y^\sharp+\Delta y^\sharp,\quad y^{\dagger,0}=y^0-y^\sharp(0,\cdot).$$

Form (4.4.5), there exists a positive constant $C_T>0$ such that

$$\big\|f_1^\dagger\big\|_{L^2(0,T;L^2(\mathcal{O}))}+\big\|y^{\dagger,0}\big\|_{H^1(\mathcal{O})}\leqslant C_T\left(\big\|\big(y^0,h^0\big)\big\|_{\mathfrak{I}}+\|(f_1,f_2,f_3)\|_{\mathbf{E}_f}\right).$$

Moreover, the compatibility condition (4.4.3) implies that the corrected initial data lives in the interpolation space

$$\big(y^{\dagger,0},h^0\big)\in[\mathcal{D}(\mathcal{A}),\mathcal{H}]_{\frac12}.$$

Therefore, by standard maximal regularity results, we have

$$\big(y^\dagger,h\big)\in L^2(0,T;\mathcal{D}(\mathcal{A}))\cap H^1(0,T;\mathcal{H}).$$

Combining the above estimate with (4.4.5) and standard interpolation estimates, we deduce (4.4.2). □

[8] Note that the constant C_T is of the form e^T.; so, it does not blow up if T goes to zero.

4.4.1 Adding the source terms

We are now in a position to provide an adaptation of the source-term method first introduced in [191] (see also [173, 115]), in the specific setting of the problem we consider containing boundary source terms, which will allow us to then apply a fixed point method for tackling the nonlinear system.

Let $\gamma : (0,\infty) \longrightarrow [0,\infty)$ be a continuous and non-increasing function satisfying

$$\lim_{t\searrow 0} \gamma(t) = +\infty$$

and (note that $\mathfrak{C}(t,\sigma) > 0$ is the constant appearing in Theorem 4.2)

$$\mathfrak{C}(t,\sigma) < \gamma(t) \qquad \text{for all } \ t > 0. \tag{4.4.7}$$

Let $q \in (1,\sqrt{2})$ and $p > 0$ be fixed such that $2p > (1+p)q^2$ be fixed. Now consider the continuous and non-increasing function $\rho_{\mathfrak{F}} : [0,T] \longrightarrow [0,\infty)$ defined by

$$\rho_{\mathfrak{F}}(t) = \gamma\left(\frac{q-1}{q^2}(T-t)\right)^{-(1+p)} \qquad t \in [0,T].$$

As $p > 0$ it is easy to see that $\rho_{\mathfrak{F}}(T) = 0$. Next, we consider the continuous and non-increasing function $\rho_0 : [0,T] \to [0,\infty)$ defined by

$$\rho_0(t) = \begin{cases} \rho_{\mathfrak{F}}(0)\gamma\left(\dfrac{q-1}{q^2}T\right) & \text{for } t \in \left[0, T\left(1-q^{-2}\right)\right] \\ \rho_{\mathfrak{F}}\left(q^2(t-T)+T\right)\gamma((q-1)(T-t)) & \text{for } t \in \left[T\left(1-q^{-2}\right), T\right], \end{cases}$$

which also satisfies $\rho_0(T) = 0$.In what follows, due to the properties of the control cost of the linear system in small times, we can and shall assume that

$$\gamma(t) = M_1 e^{\frac{M_2}{t}}, \quad \rho_{\mathfrak{F}}(t) = e^{-\frac{\alpha}{(T-t)^2}}, \quad \rho_0(t) = M_1 e^{\frac{M_2}{(q-1)(T-t)} - \frac{\alpha}{q^4(T-t)^2}} \qquad \text{for } \ t \ll 1.$$

We then define the weighted space of source terms and controls

$$\mathfrak{F} := \left\{ f = (f_1,f_2,f_3) \in \mathbf{E}_f(0,T;\mathcal{O}) \ \middle| \ \frac{f}{\rho_{\mathfrak{F}}} \in \mathbf{E}_f(0,T;\mathcal{O}) \right\}$$
$$\mathfrak{U} := \left\{ u \in L^2(0,T;L^2(\omega)) \ \middle| \ \frac{u}{\rho_0\sqrt{\eta}} \ \in L^2(0,T;L^2(\omega)) \right\},$$

where $\eta : [0,T] \to [0,+\infty)$ is a non-decreasing function defined by

$$\eta(t) = 2 + \frac{4\alpha^2}{(T-t)^6} \qquad \text{for } \ t \ \in [0,T].$$

We also define the non-decreasing function $\psi : [0,T] \to [0,\infty)$ by

$$\psi(t) = 2 + \frac{4\alpha^2 q^6}{(T-t)^6} \qquad \text{for } \ t \ \in [0,T].$$

The following version of the source-term method then holds.

Theorem 4.3. *Let $T > 0$. There exists a constant $C(T) > 0$ and a continuous linear map $\mathfrak{L} : L^2(\mathcal{O}) \times H^1(\mathbb{T}) \times \mathfrak{F} \longrightarrow \mathfrak{U}$ such that for any $(y^0,h^0) \in L^2(\mathcal{O}) \times H^1(\mathbb{T})$ and $f = (f_1,f_2,f_3) \in \mathfrak{F}$, the unique solution (y,h) to (4.3.3) with control $u = \mathfrak{L}(y^0,h^0,f)$ satisfies*

$$\left\| \left(\frac{y}{\rho_0\sqrt{\psi}}, \frac{h}{\rho_0\sqrt{\psi}} \right) \right\|^2_{C^0([0,T];L^2(\mathcal{O})\times H^1(\mathbb{T}))} + \left\| \frac{u}{\rho_0\sqrt{\eta}} \right\|^2_{L^2(0,T;L^2(\omega))}$$
$$\leqslant C(T) \left(\left\| (y^0,h^0) \right\|^2_{L^2(\mathcal{O})\times H^1(\mathbb{T})} + \left\| \left(\frac{f_1}{\rho_{\mathfrak{F}}}, \frac{f_2}{\rho_{\mathfrak{F}}}, \frac{f_3}{\rho_{\mathfrak{F}}} \right) \right\|^2_{\mathbf{E}_f} \right).$$

In particular, $y(T,\cdot) = 0$ a.e. in $\mathcal{O}$ and $h(T,\cdot) = 0$ a.e. in $\mathbb{T}$.

Proof of Theorem 4.3. For $k \geqslant 0$, we define

$$T_k = T\left(1 - \frac{1}{q^k}\right).$$

We also set $a_0 = (y^0, h^0)$, and for $k \geqslant 0$ we define

$$a_{k+1} = \left(y_f(T_{k+1}^-, \cdot), h_f(T_{k+1}^-, \cdot)\right),$$

where (y_f, h_f) is the solution to the system

$$\begin{cases} \partial_t y_f - \Delta y_f = f_1 & \text{in } (T_k, T_{k+1}) \times \mathcal{O}, \\ y_f = \sigma \partial_{x_1}^2 h_f + f_2 & \text{on } (T_k, T_{k+1}) \times \Gamma_{\text{top}}, \\ y_f = 0 & \text{on } (T_k, T_{k+1}) \times \Gamma_{\text{bot}}, \\ \partial_t h_f = \partial_{x_2} y_f + f_3 & \text{in } (T_k, T_{k+1}) \times \Gamma_{\text{top}}, \\ (y_f, h_f)_{|t=T_k^+} = 0 & \text{in } \mathcal{O} \times \mathbb{T}. \end{cases}$$

From Proposition 4.4.1, we have

$$\|a_{k+1}\|^2_{L^2(\mathcal{O}) \times H^1(\mathbb{T})} \leqslant C_T \, \|(f_1, f_2, f_3)\|^2_{\mathbf{E}_f(T_k, T_{k+1})}. \tag{4.4.8}$$

On another hand, we consider the homogeneous control system

$$\begin{cases} \partial_t y_u - \Delta y_u = u_k \mathbf{1}_\omega & \text{in } (T_k, T_{k+1}) \times \mathcal{O}, \\ y_u = \sigma \partial_{x_2}^2 h_u & \text{on } (T_k, T_{k+1}) \times \Gamma_{\text{top}}, \\ y_u = 0 & \text{on } (T_k, T_{k+1}) \times \Gamma_{\text{bot}}, \\ \partial_t h_u = \partial_{x_2} y_u & \text{in } (T_k, T_{k+1}) \times \Gamma_{\text{top}}, \\ (y_u, h_u)_{|t=T_k^+} = a_k, & \text{in } \mathcal{O} \times \mathbb{T} \end{cases}$$

where $u_k \in L^2(T_k, T_{k+1}; L^2(\omega))$ is such that

$$(y_u, h_u)(T_{k+1}^-, \cdot) = 0, \text{ in } \mathcal{O} \times \mathbb{T}$$

and

$$\|u_k\|^2_{L^2(T_k, T_{k+1}; L^2(\omega))} \leqslant \gamma^2(T_{k+1} - T_k) \, \|a_k\|^2_{L^2(\mathcal{O}) \times H^1(\mathbb{T})}. \tag{4.4.9}$$

From the definition of ρ_0 and $\rho_{\mathfrak{F}}$, we see that

$$\rho_0(T_{k+2}) = \rho_{\mathfrak{F}}(T_k) \gamma(T_{k+2} - T_{k+1}).$$

Thus

$$\begin{aligned} \|u_{k+1}\|^2_{L^2(T_{k+1}, T_{k+2}; L^2(\omega))} &\leqslant \gamma^2(T_{k+2} - T_{k+1}) \, \|a_{k+1}\|^2_{L^2(\mathcal{O}) \times H^1(\mathbb{T})} \\ &\leqslant C_T \gamma^2(T_{k+2} - T_{k+1}) \, \|(f_1, f_2, f_3)\|^2_{\mathbf{E}_f(T_k, T_{k+1})}. \end{aligned} \tag{4.4.10}$$

Thus, we now need to provide estimates of the $\mathbf{E}_f(T_k, T_{k+1})$–norm appearing in (4.4.10). First of all, using product estimates, we can easily verify that the estimates

$$\|f_1\|^2_{L^2(T_k, T_{k+1}; L^2(\mathcal{O}))} \leqslant \|\rho_{\mathfrak{F}}\|^2_{L^\infty(T_k, T_{k+1})} \left\| \frac{f_1}{\rho_{\mathfrak{F}}} \right\|^2_{L^2(T_k, T_{k+1}; L^2(\mathcal{O}))},$$

and

$$\|f_2\|^2_{L^2(T_k, T_{k+1}; H^{3/2}(\mathbb{T}))} \leqslant \|\rho_{\mathfrak{F}}\|^2_{L^\infty(T_k, T_{k+1})} \left\| \frac{f_2}{\rho_{\mathfrak{F}}} \right\|^2_{L^2(T_k, T_{k+1}; H^{3/2}(\mathbb{T}))},$$

and

$$\|f_3\|^2_{L^2(T_k, T_{k+1}; H^{1/2}(\mathbb{T}))} \leqslant \|\rho_{\mathfrak{F}}\|^2_{L^\infty(T_k, T_{k+1})} \left\| \frac{f_3}{\rho_{\mathfrak{F}}} \right\|^2_{L^2(T_k, T_{k+1}; H^{1/2}(\mathbb{T}))},$$

all hold. Using the fact that $\rho_{\mathfrak{F}}$ is decreasing, we now compute the $H^{3/4}(0,T;L^2(\mathbb{T}))$ norm of f_2; we concentrate on estimating the Gagliardo semi-norm:

$$\begin{aligned}
\left[f_2\right]^2_{H^{3/4}(T_k,T_{k+1};L^2(\mathbb{T}))} &= \int_{\mathbb{T}}\int_{T_k}^{T_{k+1}}\int_{T_k}^{T_{k+1}} \frac{|f_2(t,x)-f_2(s,x)|^2}{|t-s|^{5/2}}\,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}x \\
&\lesssim \int_{\mathbb{T}}\int_{T_k}^{T_{k+1}}\int_{T_k}^{T_{k+1}} |\rho_{\mathfrak{F}}(t)|^2 \frac{\left|\frac{f_2(t,x)}{\rho_{\mathfrak{F}}(t)}-\frac{f_2(s,x)}{\rho_{\mathfrak{F}}(s)}\right|^2}{|t-s|^{5/2}}\,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}x \\
&\quad + \int_{\mathbb{T}}\int_{T_k}^{T_{k+1}}\int_{T_k}^{T_{k+1}} \left|\frac{f_2(s,x)}{\rho_{\mathfrak{F}}(s)}\right|^2 \frac{|\rho_{\mathfrak{F}}(t)-\rho_{\mathfrak{F}}(s)|^2}{|t-s|^{5/2}}\,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}x \\
&\lesssim \rho_{\mathfrak{F}}(T_k)^2 \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{H^{3/2}(T_k,T_{k+1};L^2(\mathbb{T}))} \\
&\quad + \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^\infty(T_k,T_{k+1};L^2(\mathbb{T}))} \int_{T_k}^{T_{k+1}}\int_{T_k}^{T_{k+1}} \frac{|\rho_{\mathfrak{F}}(t)-\rho_{\mathfrak{F}}(s)|^2}{|t-s|^{5/2}}\,\mathrm{d}t\,\mathrm{d}s \\
&\lesssim \rho_{\mathfrak{F}}(T_k)^2 \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{H^{3/2}(T_k,T_{k+1};L^2(\mathbb{T}))} \\
&\quad + \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^\infty(T_k,T_{k+1};L^2(\mathbb{T}))} \|\rho_{\mathfrak{F}}\|^2_{H^{3/4}(T_k,T_{k+1})}
\end{aligned}$$

On the other hand, we also have

$$\begin{aligned}
\|\rho_{\mathfrak{F}}\|^2_{H^{3/4}(T_k,T_{k+1})} &\lesssim \|\rho_{\mathfrak{F}}-\rho_{\mathfrak{F}}(T_k)\|^2_{H^{3/4}(T_k,T_{k+1})} + \rho_{\mathfrak{F}}(T_k)^2(T_{k+1}-T_k) \\
&\lesssim \|\rho_{\mathfrak{F}}-\rho_{\mathfrak{F}}(T_k)\|^2_{H^{1}(T_k,T_{k+1})} + \rho_{\mathfrak{F}}(T_k)^2(T_{k+1}-T_k) \\
&\lesssim \|\rho_{\mathfrak{F}}\|^2_{H^{1}(T_k,T_{k+1})} + \rho_{\mathfrak{F}}(T_k)^2 T
\end{aligned}$$

Combining the above estimates, we get

$$\begin{aligned}
\|f_2\|^2_{H^{3/2}(T_k,T_{k+1};L^2(\mathbb{T}))} \lesssim C_T\rho_{\mathfrak{F}}(T_k)^2 &\left(\left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{H^{3/2}(T_k,T_{k+1};L^2(\mathbb{T}))} + \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^\infty(T_k,T_{k+1};L^2(\mathbb{T}))}\right) \\
&+ \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^\infty(T_k,T_{k+1};L^2(\mathbb{T}))} \|\rho_{\mathfrak{F}}\|^2_{H^{1}(T_k,T_{k+1})}\,.
\end{aligned}$$

Making use of elementary Sobolev embeddings, we now estimate the $H^{1/4}(0,T;L^2(\mathbb{T}))$ norm of f_3 as follows:

$$\begin{aligned}
\|f_3\|^2_{H^{1/4}(T_k,T_{k+1};L^2(\mathbb{T}))} &= \left\|\frac{f_3}{\rho_{\mathfrak{F}}}(\rho_{\mathfrak{F}}-\rho_{\mathfrak{F}}(T_k)) + \frac{f_3}{\rho_{\mathfrak{F}}}\rho_{\mathfrak{F}}(T_k)\right\|^2_{H^{1/4}(T_k,T_{k+1};L^2(\mathbb{T}))} \\
&\lesssim \left\|\frac{f_3}{\rho_{\mathfrak{F}}}\right\|^2_{H^{1/4}(T_k,T_{k+1};L^2(\mathbb{T}))} \left(\|\rho_{\mathfrak{F}}-\rho_{\mathfrak{F}}(T_k)\|^2_{H^{1}(T_k,T_{k+1})} + \rho_{\mathfrak{F}}(T_k)^2\right) \\
&\lesssim C_T \left\|\frac{f_3}{\rho_{\mathfrak{F}}}\right\|^2_{H^{1/4}(T_k,T_{k+1};L^2(\mathbb{T}))} \left(\|\rho_{\mathfrak{F}}\|^2_{H^{1}(T_k,T_{k+1})} + \rho_{\mathfrak{F}}(T_k)^2\right).
\end{aligned}$$

From the definition of ρ_F and the fact that it is decreasing, we obtain

$$\begin{aligned}
\|\rho_{\mathfrak{F}}\|^2_{H^{1}(T_k,T_{k+1})} &= \int_{T_k}^{T_{k+1}} |\rho_{\mathfrak{F}}(t)|^2\ \mathrm{d}t + 4\alpha^2\int_{T_k}^{T_{k+1}} \frac{|\rho_{\mathfrak{F}}(t)|^2}{(T-t)^6}\,\mathrm{d}t \\
&\lesssim \rho_{\mathfrak{F}}(T_k)^2\left(1+\frac{4\alpha^2}{(T-T_{k+1})^6}\right)(T_{k+1}-T_k) \\
&\lesssim T\rho_{\mathfrak{F}}(T_k)^2\left(1+\frac{4\alpha^2}{(T-T_{k+1})^6}\right).
\end{aligned}$$

Let us define

$$\eta(t) := 2 + \frac{4\alpha^2}{(T-t)^6}.$$

Combining the above estimates, from (4.4.10), we infer that

$$\begin{aligned}\|u_{k+1}\|^2_{L^2(T_{k+1},T_{k+2};L^2(\omega))} \leqslant C_T\rho_0^2(T_{k+2})\eta(T_{k+1})\Bigg(&\left\|\frac{f_1}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_k,T_{k+1};L^2(\mathcal{O}))} \\ &+ \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_k,T_{k+1};H^{3/2}(\mathbb{T}))} + \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{H^{3/2}(T_k,T_{k+1};L^2(\mathbb{T}))} \\ &+ \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^\infty(T_k,T_{k+1};L^2(\mathbb{T}))} + \left\|\frac{f_3}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_k,T_{k+1};H^{1/2}(\mathbb{T}))} \\ &+ \left\|\frac{f_3}{\rho_{\mathfrak{F}}}\right\|^2_{H^{1/4}(T_k,T_{k+1};L^2(\mathbb{T}))}\Bigg).\end{aligned}$$

Thus, using the fact that ρ_0 is decreasing and η is increasing, by virtue of the above estimate, we deduce that

$$\begin{aligned}\left\|\frac{u_{k+1}}{\rho_0\sqrt{\eta}}\right\|^2_{L^2(T_{k+1},T_{k+2};L^2(\omega))} \leqslant C_T\Bigg(&\left\|\frac{f_1}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_k,T_{k+1};L^2(\mathcal{O}))} + \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_k,T_{k+1};H^{3/2}(\mathbb{T}))} \\ &+ \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{H^{3/2}(T_k,T_{k+1};L^2(\mathbb{T}))} + \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^\infty(T_k,T_{k+1};L^2(\mathbb{T}))} + \left\|\frac{f_3}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_k,T_{k+1};H^{1/2}(\mathbb{T}))} \\ &+ \left\|\frac{f_3}{\rho_{\mathfrak{F}}}\right\|^2_{H^{1/4}(T_k,T_{k+1};L^2(\mathbb{T}))}\Bigg), \qquad (4.4.11)\end{aligned}$$

holds for some constant $C_T > 0$ independent of k. We now define the control u by pasting all of the u_k:

$$u := \sum_{k=0}^{+\infty} u_k \mathbf{1}_{[T_k,T_{k+1}]}.$$

Note that, from (4.4.9), we have

$$\left\|\frac{u_0}{\rho_0\sqrt{\eta}}\right\|^2_{L^2(T_0,T_1;L^2(\omega))} \leqslant C_T \left\|a_0\right\|^2_{L^2(\mathcal{O})\times H^1(\mathbb{T})}.$$

Combining the above estimate with (4.4.11), we get

$$\left\|\frac{u}{\rho_0\sqrt{\eta}}\right\|^2_{L^2(0,T;L^2(\omega))} \leqslant C_T\left(\left\|(y^0,h^0)\right\|^2_{L^2(\mathcal{O})\times H^1(\mathbb{T})} + \left\|\left(\frac{f_1}{\rho_{\mathfrak{F}}},\frac{f_2}{\rho_{\mathfrak{F}}},\frac{f_3}{\rho_{\mathfrak{F}}}\right)\right\|^2_{\mathbf{E}_f}\right).$$

We now look to estimate the controlled state. Let us set $(y,h) = (y_f,h_f) + (y_u,h_u)$. Then clearly for every $k \geqslant 0$, (y,h) satisfies

$$\begin{cases} \partial_t y - \Delta y = u_k\mathbf{1}_\omega + f_1 & \text{in } (T_k,T_{k+1})\times\mathcal{O}, \\ y = \sigma\partial_{x_1}^2 h + f_2 & \text{on } (T_k,T_{k+1})\times\Gamma_{\text{top}}, \\ y = 0 & \text{on } (T_k,T_{k+1})\times\Gamma_{\text{bot}}, \\ \partial_t h = \partial_{x_2} y + f_3 & \text{in } (T_k,T_{k+1})\times\Gamma_{\text{top}}, \\ (y_u,h_u)_{|t=T_k^+} = a_k & \text{in } \mathcal{O}\times\mathbb{T}. \end{cases}$$

Moreover,

$$(y,h)(T_k^-) = (y_f,h_f)(T_k^-) + (y_u,h_u)(T_k^-) = a_k = (y_f,h_f)(T_k^+) + (y_u,h_u)(T_k^+) = (y,h)(T_k^+),$$

so that (y,h) is continuous at each T_k. Furthermore, by applying proposition 4.4.1, we have

$$\|(y,h)\|^2_{C^0([T_k,T_{k+1}];L^2(\mathcal{O})\times H^1(\mathbb{T}))} \lesssim \Big(\|a_k\|^2_{L^2(\mathcal{O})\times H^1(\mathbb{T})} + \|(f_1,f_2,f_3)\|^2_{\mathbf{E}_f(T_k,T_{k+1})} + \|u_k\|^2_{L^2(T_k,T_{k+1};L^2(\omega))} \Big). \tag{4.4.12}$$

Plugging estimate (4.4.9) in (4.4.12), we infer that

$$\|(y,h)\|^2_{C^0([T_k,T_{k+1}];L^2(\mathcal{O})\times H^1(\mathbb{T}))} \lesssim \Big(\|a_k\|^2_{L^2(\mathcal{O})\times H^1(\mathbb{T})} + \gamma^2(T_{k+1}-T_k)\,\|a_k\|^2_{L^2(\mathcal{O})\times H^1(\mathbb{T})} + \|(f_1,f_2,f_3)\|^2_{\mathbf{E}_f(T_k,T_{k+1})} \Big).$$

Using (4.4.8), the above estimate can be written as

$$\|(y,h)\|^2_{C^0([T_k,T_{k+1}];L^2(\mathcal{O})\times H^1(\mathbb{T}))} \lesssim \gamma^2(T_{k+1}-T_k)\,\|(f_1,f_2,f_3)\|^2_{\mathbf{E}_f(T_{k-1},T_{k+1})}.$$

By proceeding similarly as above, we find

$$\begin{aligned}\|(y,h)\|^2_{C^0([T_k,T_{k+1}];L^2(\mathcal{O})\times H^1(\mathbb{T}))} \lesssim \gamma^2(T_{k+1}-T_k)\rho^2_{\mathfrak{F}}(T_{k-1})\eta(T_{k+1})\Bigg(&\left\|\frac{f_1}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_{k-1},T_{k+1};L^2(\mathcal{O}))} \\ + \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_{k-1},T_{k+1};H^{3/2}(\mathbb{T}))} &+ \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{H^{3/2}(T_{k-1},T_{k+1};L^2(\mathbb{T}))} + \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^\infty(T_{k-1},T_{k+1};L^2(\mathbb{T}))} \\ + \left\|\frac{f_3}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_{k-1},T_{k+1};H^{1/2}(\mathbb{T}))} &+ \left\|\frac{f_3}{\rho_{\mathfrak{F}}}\right\|^2_{H^{1/4}(T_{k-1},T_{k+1};L^2(\mathbb{T}))}\Bigg).\end{aligned}$$

Therefore,

$$\begin{aligned}\|(y,h)\|^2_{C^0([T_k,T_{k+1}];L^2(\mathcal{O})\times H^1(\mathbb{T}))} \lesssim \rho^2_0(T_{k+1})\eta(T_{k+1})\Bigg(&\left\|\frac{f_1}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_{k-1},T_{k+1};L^2(\mathcal{O}))} \\ + \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_{k-1},T_{k+1};H^{3/2}(\mathbb{T}))} &+ \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{H^{3/2}(T_{k-1},T_{k+1};L^2(\mathbb{T}))} + \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^\infty(T_{k-1},T_{k+1};L^2(\mathbb{T}))} \\ + \left\|\frac{f_3}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_{k-1},T_{k+1};H^{1/2}(\mathbb{T}))} &+ \left\|\frac{f_3}{\rho_{\mathfrak{F}}}\right\|^2_{H^{1/4}(T_{k-1},T_{k+1};L^2(\mathbb{T}))}\Bigg).\end{aligned}$$

Let us define

$$\psi(t) := 2 + \frac{4\alpha^2 q^6}{(T-t)^6}.$$

Note that $\psi(T_k) = \eta(T_{k+1})$. Using this fact, we deduce from the last estimate and the fact that ρ_0 is decreasing and ψ is non-decreasing, that

$$\begin{aligned}\left\|\left(\frac{y}{\rho_0\sqrt{\psi}},\frac{h}{\rho_0\sqrt{\psi}}\right)\right\|^2_{C^0([T_k,T_{k+1}];L^2(\mathcal{O})\times H^1(\mathbb{T}))} \lesssim \Bigg(&\left\|\frac{f_1}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_{k-1},T_{k+1};L^2(\mathcal{O}))} \\ + \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_{k-1},T_{k+1};H^{3/2}(\mathbb{T}))} &+ \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{H^{3/2}(T_{k-1},T_{k+1};L^2(\mathbb{T}))} + \left\|\frac{f_2}{\rho_{\mathfrak{F}}}\right\|^2_{L^\infty(T_{k-1},T_{k+1};L^2(\mathbb{T}))} \\ + \left\|\frac{f_3}{\rho_{\mathfrak{F}}}\right\|^2_{L^2(T_{k-1},T_{k+1};H^{1/2}(\mathbb{T}))} &+ \left\|\frac{f_3}{\rho_{\mathfrak{F}}}\right\|^2_{H^{1/4}(T_{k-1},T_{k+1};L^2(\mathbb{T}))}\Bigg).\end{aligned}$$

Combining the above estimate together with (4.4.9) and (4.4.12) (for $k = 0$), we infer that

$$\left\| \left(\frac{y}{\rho_0 \sqrt{\psi}}, \frac{h}{\rho_0 \sqrt{\psi}} \right) \right\|^2_{C^0([0,T];L^2(\mathcal{O})\times H^1(\mathbb{T}))} \leqslant C_T \left(\left\| (y^0, h^0) \right\|^2_{L^2(\mathcal{O})\times H^1(\mathbb{T})} + \left\| \left(\frac{f_1}{\rho_{\mathfrak{F}}}, \frac{f_2}{\rho_{\mathfrak{F}}}, \frac{f_3}{\rho_{\mathfrak{F}}} \right) \right\|^2_{\mathbf{E}_f} \right).$$

□

4.5 Concluding remarks

We have proven that the linearized Stefan problem with surface tension (i.e. Gibbs-Thomson correction) is null-controllable (in the sense that both the temperature and the height function are controllable) in any time by means of controls actuating along the fixed bottom boundary, a result which stipulates that the nonlinear system is itself locally null-controllable. Moreover,

- **If $\sigma = 0$.** Interestingly enough, it is not obvious to say whether the classical Stefan problem, which is known to be the (macroscopic) limit case in the zero surface tension limit without control [130], is itself null-controllable. Clearly the one-dimensional techniques of [116] do not directly apply as remarked in what precedes, as in fact, the height function manifests as an infinite-dimensional propagator. On another hand, when looking at each individual Fourier mode of the linearized Gibbs-Thomson system, we have observed that the control strategy may collapse when $\sigma \searrow 0$. These observations are nonetheless not sufficient to conclude on the possible null-controllability (or lack thereof) of the classical Stefan problem, which for the time being, remains open.

- **Memory problems.** When $\sigma = 0$, the linear system (4.3.1) is akin to the one-dimensional case addressed in [102, 116], and the equation for y can be solved without knowing h. In particular, the null-controllability requirement for the second component, i.e. $h(T, \cdot) = 0$ in $\mathbb{T}$, can equivalently be rewritten as

$$\int_0^T \partial_{x_2} y(\tau, x_1, 1)\, \mathrm{d}\tau = h_0(x_1) \qquad \text{for } x_1 \in \mathbb{T}. \tag{4.5.1}$$

  As in this geometrical setting the constraint is not finite-dimensional, it is not straightforward to say that the null-controllability of the second component, i.e. (4.5.1), follows immediately by arguing as in the one-dimensional case. In fact, since the control cost of the linearized problem $\mathfrak{C}(T, \sigma) \longrightarrow +\infty$ as $\sigma \searrow 0$, one could look to see whether the linearized classical Stefan problem may be linked to *memory problems*, where it is well-known that, unless the control region moves with time in such a way that it covers the entire domain $\mathcal{O}$ over $[0, T]$, the system is not null-controllable (see e.g. [146, 123, 57]).

- **Localized controls.** As a further perspective, one may of course seek to consider the problem where the boundary controls are localized and actuate within some non-empty subset of the torus $\mathbb{T}$. This would however mean that our projection techniques are not immediately applicable, and a direct observability inequality needs to be shown. However, as observed in what precedes, even computing the adjoint of the linear operator $\mathcal{A}$ seems farfetched, and merits clarification.

- **Three-dimensional problem.** We have, for the time being, focused solely on the two-dimensional Stefan problem in a strip-like geometry. In fact, the dimensionality plays a key role in the regularity of the solutions, as an $L^2_{t,x}$–only regular control

for the linearized system suffices to establish a nonlinear control theory. This is not the case for the three-dimensional problem for instance, where a more regular control would be needed.

Part II

Long-time optimal control

Chapter 5

Turnpike in Lipschitz-nonlinear optimal control

Abstract. We present a new proof of the turnpike property for nonlinear optimal control problems, when the running target is a stationary solution of the free dynamics. Our strategy combines the construction of sub-optimal quasi-turnpike trajectories (via a controllability assumption) and a bootstrap argument, and does not rely on analyzing the optimality system or linearization techniques. This in turn allows us to address finite-dimensional, control-affine systems with globally Lipschitz (possibly nonsmooth) nonlinearities. We show that our methodology is generic and applicable to controlled PDEs as well, such as the semilinear wave and heat equation with a globally Lipschitz nonlinearity.

Keywords. Optimal control; Turnpike; Nonlinear systems; Stabilization; Deep learning.
AMS Subject Classification. 34H05; 34H15; 93C15; 93C20.

This Chapter is taken from [96]:

> *Turnpike in Lipschitz-nonlinear optimal control.*
> C. Esteve, B. Geshkovski, D.Pighin and E. Zuazua, 2020.
> https://arxiv.org/abs/2011.11091

Chapter Contents

5.1 Introduction

The *turnpike property* reflects the fact that, for suitable optimal control problems set in a sufficiently large time horizon, any optimal solution thereof remains, during most of the time, close to the optimal solution of a corresponding "static" optimal control problem. This optimal static solution is referred to as *the turnpike* – the name stems from the idea that a turnpike is the fastest route between two points which are far apart, even if it is not the most direct route. In many cases, the turnpike property is described by an exponential estimate – for instance, the optimal trajectory $y_T(t)$ is $\mathcal{O}\left(e^{-\mu t} + e^{-\mu(T-t)}\right)$–close to the optimal static solution $\overline{y}$, for $t \in [0, T]$ and for some $\mu > 0$.

The prevalent (but not exclusive) argument for proving exponential turnpike results relies on a thorough analysis of the optimality system provided by the Pontryagin Maximum Principle. In the context of linear quadratic optimal control problems, under appropriate controllability or stabilizability conditions, turnpike is established via properties of the optimality system characterizing the optimal controls and states through the coupling with the adjoint system, see Porretta & Zuazua [221].

In the case of nonlinear dynamics, this argument thus requires nonlinearities which are continuously differentiable. A linearization argument is used – the linear study and a fixed point argument provide nonlinear results under smallness assumptions on the initial data and the target, see Zuazua et al. [222, 261]. The smallness conditions on the initial data can be removed in some specific cases (see e.g. Pighin [218]), but to the best of our knowledge, the assumptions on the running target have not been as of yet (albeit, they may be removed under restrictive assumptions, such as strict dissipativity, uniqueness of minimizers and C^2–regular nonlinearities – see [259]). This is due to the lack of tools for showing that the linearized optimality system corresponds to a linear-quadratic control problem satisfying the turnpike property, when the running target of the original nonlinear control problem is large.

There has been an ever-increasing need however, brought by applications in *deep learning* via *residual neural networks* (ResNets) (see [89, 95, 140]), of turnpike results for nonlinear optimal control problems without smallness conditions on the data or the running target, and for systems with globally Lipschitz-continuous but possibly nonsmooth nonlinearities.

In deep learning, one wishes to find a map which interpolates a dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ where $\vec{x}_i \in \mathbb{R}^{d_x}$ and $\vec{y}_i \in \mathbb{R}^{d_y}$ and gives accurate predictions on unknown points $\vec{x} \in \mathbb{R}^{d_x}$. Such a task may be accomplished by minimizing

$$\int_0^T \sum_{i=1}^N \|P\mathbf{x}_i(t) - \vec{y}_i\|^2 \, \mathrm{d}t + \int_0^T \|u(t)\|^2 \, \mathrm{d}t, \tag{5.1.1}$$

where $u := [w, b]^\top$ and $P : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ is an affine surjective map, subject to

$$\begin{cases} \dot{\mathbf{x}}_i(t) = \sigma(w(t)\mathbf{x}_i(t) + b(t)) & \text{in } (0, T) \\ \mathbf{x}_i(0) = \vec{x}_i, \end{cases} \tag{5.1.2}$$

with $w \in L^2(0, T; \mathbb{R}^{d_x \times d_x})$ and $b \in L^2(0, T; \mathbb{R}^{d_x})$ designating the controls, whereas $\sigma \in \mathrm{Lip}(\mathbb{R})$ with $\sigma(0) = 0$ is a scalar nonlinear function, defined componentwise in (5.1.2). The most frequently used nonlinearities in practical applications are *rectifiers*: $\sigma(x) = \max\{\alpha x, x\}$ for $\alpha \in [0, 1)$, and *sigmoids*: $\sigma(x) = \tanh(x)$. The order of the nonlinearity σ and the affine map within may be permuted to obtain a driftless control-affine system

$$\begin{cases} \dot{\mathbf{x}}_i(t) = w(t)\sigma(\mathbf{x}_i(t)) + b(t) & \text{in } (0, T) \\ \mathbf{x}_i(0) = \vec{x}_i. \end{cases} \tag{5.1.3}$$

Combinations and variants of (5.1.2) and (5.1.3) may also be used, see e.g. [183]. Optimizing u over $N \gg 1$ different initial data establishes robustness, so that the neural

networks (5.1.2) and (5.1.3) may correctly perform future predictions on unknown points.

In Figure 5.1, we see stabilization for the trajectories to some points $\overline{\mathbf{x}}_i \in P^{-1}(\{\vec{y}_i\})$, which are uncontrolled steady states of (5.1.2) and (5.1.3). This motivates the choice of running target as a steady control-state pair we consider in this work ((5.1.4)), which would then entail bounds for (5.1.1) (see [95]). The practical interest of the turnpike and stabilization analysis when $T \gg 1$ presented herein is its link to the large-layer regime and approximation capacity (dual to [76]) of ResNets, which are the forward Euler discretizations of (5.1.2) and (5.1.3) (see [89]). This regime is the common setting for many deep learning applications [176]. We refer the reader to [95] for further details.

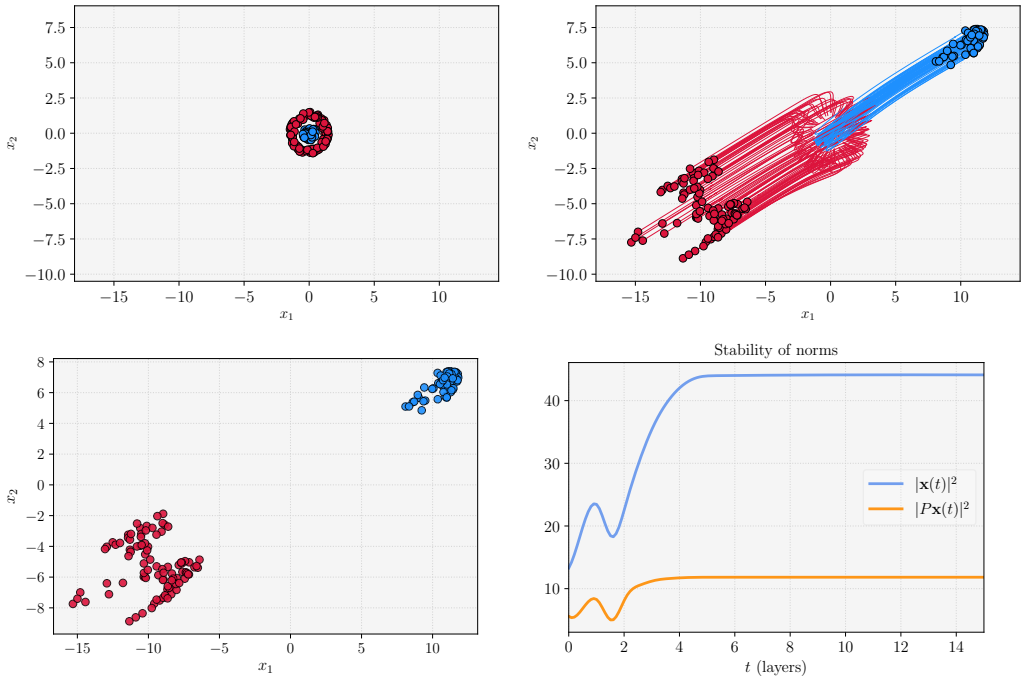
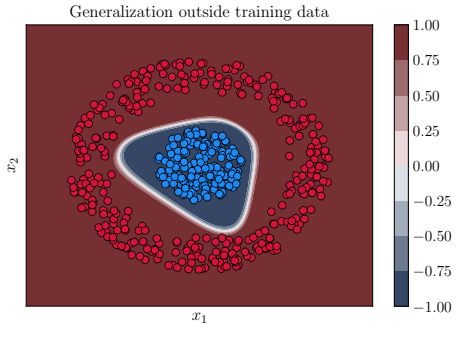


Figure 5.1: *A binary classification task in deep learning.* One aims to separate the data points $\{\vec{x}_i\}_i$ in $\mathbb{R}^2$ (top left) with respect to their color by using the controlled flow of (5.1.2) – (5.1.3) at time $T = 15$, here done by minimizing (5.1.1) ($\vec{y}_i = \pm 1$ for red/blue). We visualize the evolution of the trajectories of (5.1.3) (top right) and their output (bottom left). We see a stabilization property for the projections, but also the trajectories to some points $\overline{\mathbf{x}}_i \in P^{-1}(\{\vec{y}_i\})$ (bottom right). Displayed below is the inferred classifier on $[-2, 2]^2$, generalizing the shape of the dataset.



Our contributions. To answer this need, and motivated by problems as those above, in this work we provide a different perspective on the turnpike property in the context of nonlinear dynamics, and we bring forth the following contributions.

(1). In Section 5.2, we consider optimal control problems consisting of minimizing

$$J_T(u) := \phi(y(T)) + \int_0^T \|y(t) - \overline{y}\|^2 \, \mathrm{d}t + \int_0^T \|u(t) - \overline{u}\|^2 \, \mathrm{d}t \tag{5.1.4}$$

subject to $\dot{y} = f(y, u)$, where f is of control-affine form. Under the assumption that the running target $(\overline{u}, \overline{y})$ is a steady control-state pair, namely $f(\overline{y}, \overline{u}) = 0$, and that the system is controllable with an estimate on the cost (see Definition 6.4.1), in Theorem 5.1 we prove the exponential turnpike property described above. The main novelty lies in the fact that the nonlinearity f is only assumed to be globally Lipschitz continuous, and the result comes without any smallness conditions on the initial data or the specific running target. In this case, existing results such as those presented in Trélat & Zuazua [261, 259] do not apply, as they either require smallness assumptions or uniqueness of minimizers, and C^2–nonlinearities.

Moreover, whenever the functional to be minimized does not contain a final-time cost (such as $\phi(y(T))$ in J_T above), we can prove (see Corollary 5.2.4 below) that the exponential arc near the final time $t = T$ disappears, thus entailing an exponential stabilization property for the optimal state to the running target.

(2). In Section 5.3, the finite-dimensional results are extended to analogue optimal control problems for underlying PDE dynamics. This is illustrated in Theorem 5.2, Corollary 5.3.2 and Theorem 5.3 in the context of the semilinear wave and heat equation with globally Lipschitz–only nonlinearity, once again under the assumption that the running target is a steady control-state pair. We make no smallness assumptions neither on it, nor on the initial data, thus covering some cases where results from [122, 218, 222, 283] are not applicable.

Notation. We denote by $\|\cdot\|$ the standard euclidean norm, and $\mathbb{N} := \{1, 2, \dots\}$. We denote by $\mathrm{Lip}(\mathbb{R})$ (resp. $\mathrm{Lip}_{\mathrm{loc}}(\mathbb{R})$) the set of functions $f : \mathbb{R} \to \mathbb{R}$ which are globally (resp. locally) Lipschitz continuous.

5.2 Finite-dimensional systems

5.2.1 Setup

Let $d \geqslant 1$ and $m \geqslant 1$. We will consider differential control systems where the state $y(t)$ lives in $\mathbb{R}^d$ and the control input $u(t)$ in $\mathbb{R}^m$. Given $T > 0$, we focus on *control-affine* systems, namely canonical nonlinear systems

$$\dot{y} = f(y, u) \qquad \text{in} \quad (0, T) \tag{5.2.1}$$

with a nonlinearity f of the form

$$f(y, u) = f_0(y) + \sum_{j=1}^{m} u_j f_j(y) \qquad \text{for} \quad (y, u) \in \mathbb{R}^d \times \mathbb{R}^m, \tag{5.2.2}$$

where the vector fields $f_0, \dots, f_m \in \mathrm{Lip}(\mathbb{R}^d; \mathbb{R}^d)$ are only assumed to be globally Lipschitz continuous. This formulation includes (5.1.3) – see Remark 5.2.6 for possible extensions to (5.1.2).

For any given initial datum $y^0 \in \mathbb{R}^d$ and control input $u \in L^1(0, T; \mathbb{R}^m)$, system (5.2.1), with f as in (5.2.2), admits a unique solution $y \in C^0([0, T]; \mathbb{R}^d)$ with $y(0) = y^0$. This can be shown by means of a fixed point theorem and the Grönwall inequality applied to the integral formulation

$$y(t) = y^0 + \int_0^t f(y(s), u(s)) \, \mathrm{d}s.$$

Given $y^0 \in \mathbb{R}^d$, we will investigate the behavior when $T \gg 1$ of global minimizers $u_T \in L^2(0,T;\mathbb{R}^m)$ to nonnegative functionals of the form

$$J_T(u) := \phi(y(T)) + \int_0^T \|y(t) - \overline{y}\|^2 \,\mathrm{d}t + \int_0^T \|u(t)\|^2 \,\mathrm{d}t, \tag{5.2.3}$$

and of the corresponding solutions y_T to (5.2.1) with $y_T(0) = y^0$. Here, $\phi \in C^0(\mathbb{R}^d;\mathbb{R}_+)$ is a given final cost, while $\overline{y} \in \mathbb{R}^d$ is a given running target which we select as an *uncontrolled steady state* of the nonlinear dynamics, namely

$$f_0(\overline{y}) = 0. \tag{5.2.4}$$

We provide further comments on the specific choice of the running target just below, in Remark 5.2.1. Due to the coercivity of J_T and the explicit form of f in (5.2.2), the existence of a minimizer of J_T follows from the direct method in the calculus of variations.

Due to the presence of the state tracking term in the definition of J_T, which regulates the state over the entire time interval $[0,T]$, the well-known *turnpike property* is expected to hold: over long time horizons, the optimal control-state pair (u_T, y_T) should be "near" the optimal steady control-state pair (u_s, y_s), namely a solution to the problem

$$\inf_{u\in\mathbb{R}^m} \|y - \overline{y}\|^2 + \|u\|^2 \qquad \text{subject to} \quad f(y,u) = 0. \tag{5.2.5}$$

Now note that, due to the assumption (5.2.4) on the running target $\overline{y}$, and the form of the nonlinearity f in (5.2.2), it can be seen that $(u_s, y_s) \equiv (0, \overline{y})$ designates the unique optimal stationary solution, namely the unique solution to (7.1.3).

Remark 5.2.1 (Controlled steady states)**.** *The choice of the running target $\overline{y}$ in (5.2.4) is tailored to our proof strategy and the choice of the functional J_T in (5.2.3). The key feature our methodology requires is that the Lagrangian $\mathcal{L}(u,y) = \|y - \overline{y}\|^2 + \|u - \overline{u}\|^2$ equals zero when evaluated at the optimal steady state. In fact, we could more generally consider the functional*

$$J_T(u) := \phi(y(T)) + \int_0^T \|y(t) - \overline{y}\|^2 \,\mathrm{d}t + \int_0^T \|u(t) - \overline{u}\|^2 \,\mathrm{d}t$$

where $(\overline{u}, \overline{y}) \in \mathbb{R}^m \times \mathbb{R}^d$ is chosen so that $f(\overline{y}, \overline{u}) = 0$ (with f as in (5.2.2)), as discussed in the introduction. The results presented below could then readily be adapted to this case (by additionally changing (5.2.9) and Definition 6.4.1 to an L^2–bound of $u_T - \overline{u}$). We have taken $\overline{u} = 0$ for presentational simplicity.

In the context of nonlinear optimal control, such turnpike results have been shown by Trélat & Zuazua in [261] (see also [259]) for C^2–regular nonlinearities f. This order of regularity is required due to the proof strategy, which relies on linearizing the optimality system given by the Pontryagin Maximum Principle. As a consequence, the results in [261] are also local, in the sense that smallness conditions are assumed on the initial data and target in view of applying a fixed point argument. In this work, we take a further step and obtain global results for globally Lipschitz nonlinearities.

5.2.2 Main results

The notion of controllability plays a key role in the context of turnpike. Hence, before proceeding, we state the following assumption.

Assumption 5.2.2 (Controllability & cost estimate)**.** *We will assume that (5.2.1) is* controllable *in some time $T_0 > 0$, meaning that there exists some time $T_0 > 0$ such that for any $y^0, y^1 \in \mathbb{R}^d$, there exists a control $u \in L^2(0,T_0;\mathbb{R}^m)$ such that the corresponding solution $y \in C^0([0,T_0];\mathbb{R}^d)$ to (5.2.1) with $y(0) = y^0$ satisfies $y(T_0) = y^1$.*

We will moreover assume that there exists an $r > 0$ and a constant $C(T_0) > 0$ such that

$$\inf_{\substack{u \\ \text{such that} \\ y(0)=y^0,\, y(T_0)=\overline{y}}} \|u\|_{L^2(0,T_0;\mathbb{R}^m)} \leqslant C(T_0) \left\|y^0 - \overline{y}\right\|, \tag{5.2.6}$$

and

$$\inf_{\substack{u \\ \text{such that} \\ y(0)=\overline{y},\, y(T_0)=y^1}} \|u\|_{L^2(0,T_0;\mathbb{R}^m)} \leqslant C(T_0) \left\|y^1 - \overline{y}\right\|, \tag{5.2.7}$$

hold for any $y^0, y^1 \in \left\{x \in \mathbb{R}^d \,:\, \|x - \overline{y}\| \leqslant r\right\}$, where $\overline{y} \in \mathbb{R}^d$ is fixed as in (5.2.4).

We discuss the feasibility of this assumption later on, in Remark 5.2.7. Note that this is not a smallness assumption – it merely stipulates that, inside some ball centered at $\overline{y}$, the cost of controlling from and to $\overline{y}$ can be estimated by means of the distance to $\overline{y}$.

We may now state our first main result.

Theorem 5.1 (Turnpike)**.** *Assume that $f_0, \dots, f_m \in \mathrm{Lip}(\mathbb{R}^d; \mathbb{R}^d)$ in (5.2.2), and assume that (5.2.1) is controllable in some time $T_0 > 0$ in the sense of Definition 6.4.1. Let $y^0 \in \mathbb{R}^d$ be given, and let $\overline{y} \in \mathbb{R}^d$ be as in (5.2.4). Then there exists a time $T^* > 0$ and constants $C_1, C_2, \mu > 0$ such that for any $T \geqslant T^*$, any global minimizer $u_T \in L^2(0,T;\mathbb{R}^m)$ to J_T defined in (5.2.3) and corresponding optimal state y_T solution to (5.2.1) with $y_T(0) = y^0$ satisfy*

$$\|y_T(t) - \overline{y}\| \leqslant C_1 \left(e^{-\mu t} + e^{-\mu(T-t)}\right) \tag{5.2.8}$$

for all $t \in [0,T]$, and

$$\|u_T\|_{L^2(0,T;\mathbb{R}^m)} \leqslant C_2. \tag{5.2.9}$$

We sketch the idea of the proof (which may be found in Section 5.5.2) in Section 5.2.2 below. The rate $\mu > 0$ appearing in (5.2.8) depends on the datum y^0 due to the multiplicative form of the control, but is uniform with respect to y^0 when the control is *additive*, namely, when $f_1, \dots, f_m$ are nonzero constants. This is due to the form of the constant provided by Grönwall arguments (e.g. in Lemma 5.4.1 and Lemma 5.5.2).

Remark 5.2.3 (On (5.2.9))**.** *An exponential estimate for the optimal control u_T is a hallmark of turnpike results obtained by analyzing the optimality system. Therein, the optimal control can be characterized explicitly via the adjoint state, which, much like the optimal state, fulfills an exponential estimate. Since in this work we do not use the optimality system, we do not have as much information on $u_T(t)$ as we have on $y_T(t) - \overline{y}$. The latter quantity, in addition to being penalized by J_T, may be further estimated by using the system dynamics. In the context of* driftless systems*, we show that $u_T(t)$ too is in $\mathcal{O}\left(e^{-\mu t} + e^{-\mu(T-t)}\right)$ in Corollary 5.2.5, by using the homogeneity of the system with respect to the control.*

Before proceeding with further remarks, let us state a couple of important corollaries of Theorem 5.1.

Firstly, when one considers an optimal control problem for J_T without a final cost for the endpoint $y(T)$, namely taking $\phi \equiv 0$ in (5.2.3), Theorem 5.1 can in fact be improved to an *exponential stabilization* estimate to the running target $\overline{y}$.

Corollary 5.2.4 (Stabilization)**.** *Suppose that $\phi \equiv 0$ in J_T defined in (5.2.3). Under the assumptions of Theorem 5.1, there exists a time $T^* > 0$, and constants $C_1, C_2, \mu > 0$ such that for any $T \geqslant T^*$, any global minimizer $u_T \in L^2(0,T;\mathbb{R}^m)$ to J_T defined in (5.2.3) and corresponding optimal state y_T solution to (5.2.1) with $y_T(0) = y^0$ satisfy (5.2.9) as well as*

$$\|y_T(t) - \overline{y}\| \leqslant C_1 e^{-\mu t} \tag{5.2.10}$$

for all $t \in [0,T]$.

We refer to Section 5.5.3 for a proof. In fact, Corollary 5.2.4 may be proven independently of Theorem 5.1 by a simple adaptation of the proof strategy. This is illustrated in the proof of Theorem 5.3 in the context of the semilinear heat equation.

On another hand, when the underlying dynamics (5.2.1) are of *driftless control affine* form (namely, $f_0 \equiv 0$ in (5.2.2)), we can obtain an exponential decay for the optimal controls as well. Note that in this case, any $\overline{y} \in \mathbb{R}^d$ is an admissible running target for J_T, since $f(\overline{y}, 0) = 0$ for any $\overline{y} \in \mathbb{R}^d$.

Corollary 5.2.5 (Control decay). *Suppose that $f_0 \equiv 0$ in (5.2.2). Under the assumptions of Theorem 5.1, there exists a time $T^* > 0$, and constants $C, \mu > 0$ such that for any $T \geqslant T^*$, any global minimizer $u_T \in L^2(0,T;\mathbb{R}^m)$ to J_T defined in (5.2.3) and corresponding optimal state y_T solution to (5.2.1) with $y_T(0) = y^0$ satisfy (5.2.8) as well as*

$$\|u_T(t)\| \leqslant C\Big(e^{-\mu t} + e^{-\mu(T-t)}\Big) \tag{5.2.11}$$

for a.e. $t \in [0,T]$.

If moreover, $\phi \equiv 0$ in J_T defined in (5.2.3), in addition to (5.2.10), there exist constants $C_1, \mu_1 > 0$ independent of T such that

$$\|u_T(t)\| \leqslant C_1 e^{-\mu_1 t} \tag{5.2.12}$$

holds for a.e. $t \in [0,T]$.

Corollary 5.2.4 and Corollary 5.2.5 are in particular applicable for the continuous time analog (5.1.3) of ResNets (see Remark 5.2.6 for (5.1.2)).

The proof of Corollary 5.2.5 (see Section 5.5.4) will follow by firstly using a specific suboptimal control (constructed using the time-scaling specific to driftless systems) to estimate $J_T(u_T)$ and obtain

$$\int_t^{t+h} \|u_T(s)\|^2 \,\mathrm{d}s \leqslant \int_t^{t+h} \|y_T(s) - \overline{y}\|^2 \,\mathrm{d}s$$

for h small enough, an estimate which, coupled with the turnpike estimates of Theorem 5.1 – Corollary 5.2.4 and the Lebesgue differentiation theorem, will suffice to conclude.

Sketch of the proof of Theorem 5.1

The proof of Theorem 5.1 may be found in Section 5.5.2. It roughly follows the following scheme (see also Figure 5.2 just below). For simplicity, suppose that $T \geqslant 2T^*$.

1). By controllability, we first construct a suboptimal quasi-turnpike control u^1 which is such that the associated state y^1 satisfies $y^1(T_0) = \overline{y}$, and $u^1(t) = 0$ for $t \in [T_0, T]$. Thus $y^1(t) = \overline{y}$ for $t \in [T_0, T]$. Due to the form of J_T in (5.2.3), this would imply that $J_T(u^1)$ is independent of T, and by using $J_T(u_T) \leqslant J_T(u^1)$, would also entail a uniform bound of $J_T(u_T)$ with respect to T. A Grönwall argument ensures that, moreover,

$$\|y_T - \overline{y}\|_{L^2(0,T;\mathbb{R}^d)} + \|y_T(t) - \overline{y}\| \leqslant C_0 \qquad \text{for all } t \in [0,T] \tag{5.2.13}$$

for some $C_0 > 0$ independent of T. (5.2.13) alone is enough to obtain the desired exponential estimates for $t \in [0, T^*] \cup [T - T^*, T]$, an interval whose length is independent of T. More details can be found in Lemma 5.5.1.

2). Since $T^* \leqslant \frac{T}{2}$, by a simple contradiction argument (see Lemma 5.5.3), there exist $\tau_1 \in [0, T^*)$ and $\tau_2 \in (T - T^*, T]$ such that

$$\|y_T(\tau_i) - \overline{y}\| \leqslant \frac{\|y_T - \overline{y}\|_{L^2(0,T;\mathbb{R}^d)}}{\sqrt{T^*}} \overset{(5.2.13)}{\leqslant} \frac{C_0}{\sqrt{T^*}}. \tag{5.2.14}$$

3). On $[\tau_1, \tau_2]$, the optimal control u_T will minimize a functional without the final cost $\phi(y_T(T))$ but with a terminal constraint on the state y_T. By controllability, using a second suboptimal quasi-turnpike control u^2 satisfying estimates as those in Definition 6.4.1, and using $J_T(u_T) \leqslant J_T(u^2)$ along with a Grönwall argument, one shows an estimate of the form

$$\|y_T(t) - \overline{y}\| \leqslant C_1 \Big(\|y_T(\tau_1) - \overline{y}\| \ + \|y_T(\tau_2) - \overline{y}\| \Big) \tag{5.2.15}$$

$$\overset{(5.2.14)}{\leqslant} \frac{2C_1^2}{\sqrt{T^*}} \tag{5.2.16}$$

for all $t \in [\tau_1, \tau_2]$, thus also for $t \in [T^*, T - T^*] \subset [\tau_1, \tau_2]$ where $C_1 > 0$ is independent of T. Definition 6.4.1 is used precisely in this step, and is essential in obtaining an estimate of the mould of (5.2.15). For more details, see Lemma 5.5.2.

4). A *bootstrap argument* (Section 5.5.2): estimate (5.2.16) can be iterated by shrinking the time interval to obtain an estimate of the form

$$\|y_T(t) - \overline{y}\| \ \leqslant \left(\frac{2C_1^2}{\sqrt{T^*}} \right)^n \qquad \text{for} \quad [nT^*, T - nT^*] \tag{5.2.17}$$

for "suitable" $n \geqslant 1$. Then taking $T^* > 4C_1^4$ and a suitable choice of n in (5.2.17) will yield the exponential estimate for $t \ \in [T^*, T - T^*]$.

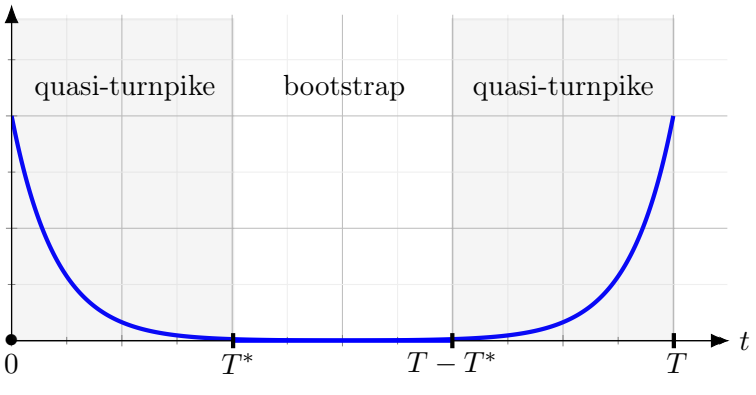


Figure 5.2: *A sketch of the scheme.* We use a quasi-turnpike control to bound $J_T(u_T)$ uniformly in T, which entails the exponential estimates on $[0, T^*] \cup [T - T^*, T]$. We then perform a bootstrap by iteratively shrinking symmetric intervals within $[T^*, T - T^*]$ in view of obtaining an estimate of the mould of (5.2.17).

5.2.3 Comments on the main results

Several pertinent remarks are in order.

Remark 5.2.6 (On the nonlinearity)**.** *With little modifications, Theorem 5.1 and Corollary 5.2.4 also apply to system* (5.2.1) *with nonlinearities f of the form*

$$f(y, u) = \sum_{j=1}^{m} f_j(u_j y) \qquad \text{for} \quad (y, u) \in \mathbb{R}^d \times \mathbb{R}^m \tag{5.2.18}$$

where the vector fields $f_1, \dots, f_m \in \mathrm{Lip}(\mathbb{R}^d; \mathbb{R}^d)$ are additionally assumed to be positively homogeneous of degree 1*, and an H^1–penalization instead of only L^2 of the control appears*

in the definition of J_T, in order to assert sufficient compactness for proving the existence of minimizers. Such nonlinearities are motivated by (5.1.2). Due to the homogeneity of the vector fields in (5.2.18), the corresponding optimal steady states coincide with those of the driftless case, namely $(u_s, y_s) = (0, \overline{y})$ for any $\overline{y} \in \mathbb{R}^d$.

Remark 5.2.7 (On Definition 6.4.1)**.** *Both parts of Definition 6.4.1 are needed in our strategy.*

- *In the driftless case ($f_0 = 0$ in (5.2.2)), the Chow-Rashevskii theorem (see [69, Chapter 3, Section 3.3]), characterized by iterated Lie brackets, is a necessary and sufficient condition for the global exact controllability of systems with smooth vector fields. But general necessary and sufficient conditions which ensure the exact controllability of control-affine systems are not known to our knowledge – see [69, Chapter 3]. This is mainly due to the drift term f_0, which affects the geometry of the problem and may pose obstructions to the controllability in arbitrary time – see [22] for a survey on these issues. We do insist however, that we merely require controllability in a possibly large time T_0, and not necessarily in any arbitrarily small time.*

- *The assumptions (5.2.6) – (5.2.7) are more commonly encountered in the linear systems setting, and thus also for nonlinear systems obtained by perturbation arguments. In such contexts, it is well-known (see e.g. [281, Remark 2.2]) that the minimal L^2–norm control u satisfies*

$$\|u\|_{L^2(0,T_0;\mathbb{R}^m)} \leqslant C(T_0)\left(\left\|y^0\right\| + \left\|y^1\right\|\right)$$

for some $C(T_0) > 0$. This makes Definition 6.4.1 entirely plausible in the settings mentioned above. Indeed, we consider $z := y - \overline{y}$, then either $z^0 = 0$ (if $y^0 = \overline{y}$) or $z^1 = 0$ (if $y^1 = \overline{y}$). The control u steering y from y^0 to y^1 in time T would then be the same as the one steering z from either 0 to $y^1 - \overline{y}$ or from $y^0 - \overline{y}$ to 0 in time T, and the above estimate would yield the desired assumption.

To complete this discussion, we refer to [95, Theorem 5.2], where in the context of driftless systems motivated by neural networks (see the Introduction), we prove a local controllability result along with estimates (5.2.6) – (5.2.7). The main caveat when comparing to the setting we consider here is that in neural networks, the control is typically a matrix of dimension $d \times d$ (eventhough we find a single control for $N \gg 1$ initial data), allowing us to access the entire state, whereas here it is a vector in $\mathbb{R}^m$, possibly with $m < d$. Nonetheless, driftless systems motivated by neural networks remain a case where our results apply.

5.3 Infinite-dimensional systems

We illustrate the flexibility of the finite-dimensional arguments and adapt them to the semilinear wave and heat equation. As a matter of fact, the only difference between the finite and infinite dimensional setting is in the proof of uniform control and state bounds by means of quasi-turnpike strategies. The specific proof of turnpike is identical in both cases. We distinguish the case of the wave and heat equation because of the validity of the PDE analog of Definition 6.4.1, as made more precise below.

5.3.1 Semilinear wave equation

Let $T > 0$ and let $\Omega \subset \mathbb{R}^d$ be a bounded and (at least C^2) regular domain. We will be interested in control systems of the form

$$\begin{cases} \partial_t^2 y - \Delta y + f(y) = u\mathbf{1}_\omega & \text{in } (0,T)\times\Omega \\ y = 0 & \text{on } (0,T)\times\partial\Omega \\ (y, \partial_t y)|_{t=0} = \mathbf{y}^0 & \text{in } \Omega. \end{cases} \tag{5.3.1}$$

Here $f \in \mathrm{Lip}(\mathbb{R})$, $\omega \subset \Omega$ is open (with geometric assumptions given in (5.3.4)), whereas $\mathbf{y}^0 = \left(y_1^0, y_2^0\right)$ is a given initial datum. It is well-known, by fixed-point arguments, that for any initial data $\mathbf{y}^0 = \left(y_1^0, y_2^0\right) \in H_0^1(\Omega) \times L^2(\Omega)$ and for any $u \in L^2((0,T) \times \omega)$, there exists a unique finite-energy solution $y \in C^0([0,T]; H_0^1(\Omega)) \cap C^1([0,T]; L^2(\Omega))$ to (5.3.1).

As in the finite-dimensional case, we will address the behavior when $T \gg 1$ of global minimizers $u_T \in L^2((0,T) \times \omega)$ to nonnegative functionals of the form

$$J_T(u) := \phi(y(T)) + \int_0^T \|y(t) - \overline{y}\|_{H_0^1(\Omega)}^2 \,\mathrm{d}t + \int_0^T \|\partial_t y(t)\|_{L^2(\Omega)}^2 \,\mathrm{d}t + \int_0^T \|u(t)\|_{L^2(\omega)}^2 \,\mathrm{d}t, \tag{5.3.2}$$

and of the corresponding solution y_T to (5.3.1). Here $\phi \in C^0(L^2(\Omega); \mathbb{R}_+)$ is a given final cost, while $\overline{y} \in H_0^1(\Omega)$ is a running target which we select as an uncontrolled steady state of (5.3.1), namely we assume that $\overline{y}$ is some solution[1] to

$$\begin{cases} -\Delta \overline{y} + f(\overline{y}) = 0 & \text{in } \Omega \\ \overline{y} \ = 0 & \text{on } \partial\Omega. \end{cases} \tag{5.3.3}$$

We henceforth moreover assume that f, Ω are such that a solution to (5.3.3) exists. This can be ensured in a variety of different cases, including, for instance (see [55, 187] for further results):

- If $f(0) = 0$, then clearly $\overline{y} \equiv 0$ is one solution. But if moreover there exist $p \in \left(1, \frac{d+2}{d-2}\right)$ ($p \in (1,\infty)$ for $d = 1, 2$), $\nu < \lambda_1(\Omega)$ and $\theta > 2$ such that

$$\begin{aligned} |f(s)| &\leqslant C(1 + |s|^p) && \text{for all } s \in \mathbb{R} \\ -\int_0^s f(\zeta)\,\mathrm{d}\zeta &\leqslant \frac{\nu}{2}\, s^2 && \text{for } |s| \text{ small} \\ 0 < -\theta \int_0^s f(\zeta)\,\mathrm{d}\zeta &\leqslant -s \int_0^s f(\zeta)\,\mathrm{d}\zeta && \text{for } |s| \text{ large,} \end{aligned}$$

  then a nontrivial solution $\overline{y} \in H_0^1(\Omega)$, $\overline{y} \not\equiv 0$ also exists. We refer to [55, Theorem 2.5.6]. This fact is a consequence of the mountain pass theorem. Here $\lambda_1(\Omega)$ denotes the first eigenvalue of the Dirichlet Laplacian $-\Delta$.

- When $d = 1$ and $\Omega = (-R, R)$, then both necessary and sufficient conditions on f can be provided ensuring the existence of nontrivial solutions – see [55, Theorem 1.2.3].

The case of a controlled steady state (namely adding $\overline{u}\mathbf{1}_\omega$ in (5.3.3)) may also be considered, under the condition that the functional J_T is modified appropriately as discussed in Remark 5.2.1. The existence of minimizers to J_T again follows by the direct method in the calculus of variations.

We note that, since $\overline{y}$ is fixed as above, the pair $(u_s, y_s) \equiv (0, \overline{y})$ is the unique solution to the steady-state optimal control problem

$$\inf_{u \in L^2(\omega)} \|y - \overline{y}\|_{H_0^1(\Omega)}^2 + \|u\|_{L^2(\omega)}^2 \quad \text{subject to} \quad \begin{cases} -\Delta y + f(y) = u\mathbf{1}_\omega & \text{in } \Omega \\ y \ = 0 & \text{on } \partial\Omega. \end{cases}$$

This is because the functional in the expression above attains its minimum, equal to 0, precisely at $(0, \overline{y})$, a pair which satisfies the constraint provided by the elliptic equation.

[1] There is no need for the solution of (5.3.3) to be unique.

Before proceeding, we need to define the appropriate geometric setup for ensuring the exact controllability of (5.3.1) when $d \geqslant 2$. For any fixed $x_\circ \in \mathbb{R}^d \setminus \overline{\Omega}$, we define

$$\Gamma_\circ := \{x \in \partial\Omega : (x - x_\circ) \cdot \nu(x) > 0\}$$

where $\nu(x)$ denotes the outward unit normal at $x \in \partial\Omega$. The set $\Gamma_\circ$ coincides with the subset of the boundary arising usually in the context of the multiplier method [184]. We will suppose that for some $\delta > 0$,

$$\omega = \mathcal{O}_\delta(\Gamma_\circ) \cap \Omega, \tag{5.3.4}$$

where $\mathcal{O}_\delta(\Gamma_\circ) := \{x \in \mathbb{R}^d : |x - x'| < \delta \text{ for some } x' \in \Gamma_\circ\}$. It is known that, under these geometric assumptions on ω, and since $f \in \mathrm{Lip}(\mathbb{R})$, there exists some time $T_{\min} = T_{\min}(\Omega, \omega) > 0$ such that the wave equation (5.3.1) is exactly controllable in any time $T_0 > T_{\min}$, see [112, 277] and also [88, Section 7.2] (see also the introduction of [150] for an ample survey of controllability results for semilinear wave equations). These results are extensions of the one-dimensional results in [280].

We may now state our main result in the context of the wave equation.

Theorem 5.2 (Turnpike). *Suppose that $f \in \mathrm{Lip}(\mathbb{R})$ and $\Omega \subset \mathbb{R}^d$ are such that (5.3.3) admits at least one solution, and let $\overline{y} \in H_0^1(\Omega)$ be any such solution. Let $\phi \in C^0(L^2(\Omega); \mathbb{R}_+)$, and suppose that ω is as in (5.3.4). For any $\mathbf{y}^0 \in H_0^1(\Omega) \times L^2(\Omega)$, there exists a time $T^* > T_{\min}(\omega, \Omega)$ and constants $C_1, C_2 > 0$ and $\mu > 0$, such that for any $T \geqslant T^*$, any global minimizer $u_T \in L^2((0,T) \times \omega)$ to J_T defined in (5.3.2) and corresponding optimal state y_T solution to (5.3.1) satisfy*

$$\|y_T(t) - \overline{y}\|_{H_0^1(\Omega)} + \|\partial_t y_T(t)\|_{L^2(\Omega)} \leqslant C_1 \left(e^{-\mu t} + e^{-\mu(T-t)}\right)$$

for all $t \in [0,T]$, and

$$\|u_T\|_{L^2((0,T)\times\omega)} \leqslant C_2.$$

Moreover, $\mu > 0$ is independent of $\mathbf{y}^0$.

The proof of turnpike (see Section 5.6) is identical to the finite-dimensional case. Some technical adaptations are however needed for obtaining the quasi-turnpike bounds, wherein one uses the Duhamel formula for mild solutions in view of applying an integral Grönwall argument, in the spirit of the ODE setting.

Remark 5.3.1 (On the choice of J_T). *We note that in existing turnpike results for the wave equation, e.g. [127, 260, 283], a slightly weaker functional is sometimes considered. For instance, in [283] for the linear wave equation, only the $L^2(0,T;H_0^1(\Omega))$–norm of $y - \overline{y}$ is penalized, and not the $L^2((0,T)\times\Omega)$–norm of $\partial_t y$, yet turnpike is shown to hold for the full state $(y, \partial_t y)$. This is justified by the equipartition of energy property, which states that, along a given time interval $[0,T]$, the energy concentrated on the y component in $H_0^1(\Omega)$ and on the $\partial_t y$ component in $L^2(\Omega)$ is comparably the same up to a compact remainder term. We choose to work with a functional penalizing the full state of the system due to the specificity of our proof strategy.*

Similarly to the finite-dimensional case, when $\phi \equiv 0$ in (5.3.2), Theorem 5.2 entails an exponential stabilization property for the optimal states, namely

Corollary 5.3.2 (Stabilization). *Suppose that $\phi \equiv 0$ in J_T defined in (5.3.2). Under the assumptions of Theorem 5.2, there exists a time $T^* > T_{\min}(\omega, \Omega)$ and constants $C_1, C_2, \mu > 0$ such that for any $T \geqslant T^*$, any global minimizer $u_T \in L^2((0,T)\times\omega)$ to J_T defined in (5.3.2) and corresponding optimal state y_T solution to (5.3.1) satisfy*

$$\|y_T(t) - \overline{y}\|_{H_0^1(\Omega)} + \|\partial_t y_T(t)\|_{L^2(\Omega)} \leqslant C_1 e^{-\mu t}$$

for all $t \in [0,T]$ and

$$\|u_T\|_{L^2((0,T)\times\omega)} \leqslant C_2.$$

Moreover, $\mu > 0$ is independent of $\mathbf{y}^0$.

5.3.2 Semilinear heat equation

To complete our presentation, we will also discuss control systems of the form

$$\begin{cases} \partial_t y - \Delta y + f(y) = u\mathbf{1}_\omega & \text{in } (0,T)\times\Omega \\ y = 0 & \text{on } (0,T)\times\partial\Omega \\ y|_{t=0} = y^0 & \text{in } \Omega, \end{cases} \tag{5.3.5}$$

were $f \in \mathrm{Lip}(\mathbb{R})$, $\omega \subset \Omega$ is any open, non-empty subset, whereas y^0 is a given initial datum. It is well-known that for any given $T > 0$, $y^0 \in L^2(\Omega)$ and $u \in L^2((0,T)\times\omega)$, there exists a unique globally-defined solution $y \in C^0([0,T];L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$ to (5.3.5).

We will again study global minimizers $u_T \in L^2((0,T)\times\omega)$ to nonnegative functionals of the form

$$J_T(u) := \int_0^T \|y(t)-\overline{y}\|^2_{L^2(\Omega)}\,\mathrm{d}t + \int_0^T \|u(t)\|^2_{L^2(\omega)}\,\mathrm{d}t, \tag{5.3.6}$$

and the corresponding solution y_T to (5.3.5) in the regime $T \gg 1$. Once again, $\overline{y} \in L^2(\Omega)$ is a running target which we select as an uncontrolled steady state, namely a solution to (5.3.3). The existence of minimizers to J_T defined in (5.3.6) follows by the direct method in the calculus of variations.

Theorem 5.3 (Stabilization). *Suppose that $f \in \mathrm{Lip}(\mathbb{R})$ and $\Omega \subset \mathbb{R}^d$ are such that (5.3.3) admits at least one solution, and let $\overline{y} \in H_0^1(\Omega)$ be any such solution. For any $y^0 \in L^2(\Omega)$, there exists $T^* > 0$ and constants $C_1, C_2, \mu > 0$ such that for any $T \geqslant T^*$, any global minimizer $u_T \in L^2((0,T)\times\omega)$ of J_T defined in (5.3.6) and corresponding optimal state y_T solution to (5.3.5) satisfy*

$$\|y_T(t)-\overline{y}\|_{L^2(\Omega)} \leqslant C_1 e^{-\mu t}$$

for all $t \in [0,T]$, and

$$\|u_T\|_{L^2((0,T)\times\omega)} \leqslant C_2.$$

Moreover, $\mu > 0$ is independent of y^0.

We refer to Section 5.7 for the proof.

We consider the heat equation in addition to the wave equation because of the validity of the PDE analog of Definition 6.4.1. The heat equation is exactly controllable to controlled trajectories, namely solutions $\widehat{y}$ to (5.3.5) for given controls $\widehat{u}$. Instead of an estimate such as (5.2.7), one has $\|u-\widehat{u}\|_{L^2((0,T_0)\times\omega)} \leqslant C(T_0)\left\|y^0-\widehat{y}(0)\right\|_{L^2(\Omega)}$ (see e.g. [219, Lemma 8.3] and the references therein) for minimal L^2–norm controls u steering y to $\widehat{y}$ in time T_0. Such an estimate does not suffice for applying our methodology, as we clearly need to estimate the minimal L^2–norm control by means of the distance of the initial data to the target. Nonetheless, we illustrate that the stabilization result can be shown independently of the turnpike result. Indeed, the proof closely follows that of Theorem 5.1, with the exception that we only need to perform the bootstrap forward in time, whence we do not require that the system is controllable to anything else but a steady state. We refer to Section 5.7 for more details.

The semilinear heat equation is a commonly used benchmark for nonlinear turnpike results, thus this example serves to compare with existing results, such as those in [218].

Remark 5.3.3 (On the nonlinearity). *The assumption that f is globally Lipschitz in (5.3.1) and (5.3.5) could perhaps be relaxed to a locally Lipschitz f (for which blow-up is avoided and controllability is ensured – for instance, $f(y) = y^3$), under the condition that one can show a uniform $L^\infty((0,T)\times\Omega)$–estimate of y_T with respect to $T > 0$. Arguments of this sort in the context of turnpike can be found in [218] under smallness assumptions on the target. We refer to the end of Section 5.8 for a discussion of a*

(possibly technical) impediment encountered in applying our methodology to the cubic heat equation. In addition to the controllability properties it entails for (5.3.1) – (5.3.5) *as blow-up is avoided,we use the Lipschitz character of f in the estimates in Lemma 5.6.1, Lemma 5.6.3 and Lemma 5.7.1.*

5.4 Preliminary results

We begin by presenting a couple of simple but important lemmas, containing bounds of the quantity $\|y(t) - \overline{y}\|$ for both the nonlinear ODE and PDE setting, solely by means of $\|y^0 - \overline{y}\|$ and the tracking terms appearing in the functional J_T. These bounds would thus imply that bounding the functional J_T uniformly in T would entail a bound for the desired quantity $\|y(t) - \overline{y}\|$.

Let us begin with the ODE estimate.

Lemma 5.4.1. *Let $T > 0$ be given, and let $\overline{y} \in \mathbb{R}^d$ be as in* (5.2.4). *For any data $u \in L^2(0,T;\mathbb{R}^m)$ and $y^0 \in \mathbb{R}^d$, let $y \in C^0([0,T];\mathbb{R}^d)$ be the solution to* (5.2.1) *with $y(0) = y^0$. Then there exist constants $C_1 = C_1(f,\overline{y}) > 0$ and $C_2 = C_2(f)$ independent of T such that*

$$\|y(t) - \overline{y}\| \leqslant C\Big(\big\|y^0 - \overline{y}\big\| + \|u\|_{L^2(0,T;\mathbb{R}^m)} + \|y - \overline{y}\|_{L^2(0,T;\mathbb{R}^d)} \Big)$$

holds for all $t \in [0,T]$, where

$$C := C_1 \exp\Big(C_2 \|u\|_{L^2(0,T;\mathbb{R}^m)} \Big).$$

As insinuated by the form of the constant in the estimate, the proof follows a Grönwall argument. However, as this constant depends on T only through the L^2–norm of the control u, we present the proof for the sake of clarity.

Proof of Lemma 5.4.1. Let us first suppose that $t \in [0,1]$. By integrating the equation satisfied by y, and using the fact that $f_0, \dots, f_m \in \mathrm{Lip}(\mathbb{R}^d;\mathbb{R}^d)$ and $t \leqslant 1$, as well as Cauchy-Schwarz, it may be seen that

$$\|y(t) - \overline{y}\| \leqslant C_0 \left(\big\|y^0 - \overline{y}\big\| + \|u\|_{L^2(0,T;\mathbb{R}^m)} \right)$$

for some $C_0 = C_0(f) > 0$.

Now suppose that $t \in (1,T]$. We begin by showing that for any such t, there exists a $t^* \in (t-1,t]$ such that

$$\|y(t^*) - \ \overline{y}\| \leqslant \|y - \overline{y}\|_{L^2(0,T;\mathbb{R}^d)}. \tag{5.4.1}$$

To this end, we argue by contradiction. Suppose that

$$\|y(t^*) - \overline{y}\| \ > \|y - \overline{y}\|_{L^2(0,T;\mathbb{R}^d)}$$

for all $t^* \in (t-1,t]$. Then

$$\|y - \overline{y}\|^2_{L^2(0,T;\mathbb{R}^d)} = \int_0^T \|y(t) - \overline{y}\|^2 \,\mathrm{d}t \geqslant \int_{t-1}^t \|y(\tau) - \overline{y}\|^2 \,\mathrm{d}\tau > \|y - \overline{y}\|^2_{L^2(0,T;\mathbb{R}^d)},$$

which contradicts the hypothesis. Thus (5.4.1) holds.

Consequently, we know that there exists $t^* \in (t-1, t]$ such that (5.4.1) holds. By integrating the equation satisfied by y in $[t^*, t]$, namely writing

$$\begin{aligned} y(t) - \overline{y} &= y(t^*) - \overline{y} + \int_{t^*}^{t} \left(f_0(y) + \sum_{j=1}^{m} u_j f_j(y) \right) \mathrm{d}\tau \\ &= y(t^*) - \overline{y} + \int_{t^*}^{t} \left(f_0(y) - f_0(\overline{y}) \right) \mathrm{d}\tau + \int_{t^*}^{t} \sum_{j=1}^{m} u_j \left(f_j(y) - f_j(\overline{y}) \right) \mathrm{d}\tau \\ &\quad + \int_{t^*}^{t} \sum_{j=1}^{m} u_j f_j(\overline{y}) \, \mathrm{d}\tau, \end{aligned}$$

we see that, by using the Lipschitz character of $f_0, \dots, f_m$ and Cauchy-Schwarz for the sums,

$$\|y(t) - \overline{y}\| \leqslant \|y(t^*) - \overline{y}\| + C_0(f) \int_{t^*}^{t} \Big(1 + \|u(\tau)\| \Big) \|y(\tau) - \overline{y}\| \, \mathrm{d}\tau + C_1(f, \overline{y}) \int_{t^*}^{t} \|u(\tau)\| \, \mathrm{d}\tau.$$

Now applying a combination of Cauchy-Schwarz, the fact that $t - t^* \leqslant 1$, (5.4.1), and the Grönwall inequality to the inequality just above, we obtain

$$\|y(t) - \overline{y}\| \leqslant C_2 \exp \left(C_3(f) \sqrt{1 + \int_{t^*}^{t} \|u(\tau)\|^2 \, \mathrm{d}\tau} \right) \left(\|y - \overline{y}\|_{L^2(0,T;\mathbb{R}^d)} + \|u\|_{L^2(0,T;\mathbb{R}^m)} \right),$$

for some $C_2(f, \overline{y}) > 0$ and $C_3(f) > 0$, from which, using $\sqrt{x^2 + y^2} \leqslant x + y$ for $x, y > 0$, the desired statement readily follows. □

Remark 5.4.2. *Let us make two brief observations.*

- *We note that in the case where the running target is $(\overline{u}, \overline{y})$ with $f(\overline{y}, \overline{u}) = 0$ and $\overline{u} \neq 0$, and thus we minimize J_T defined in (5.1.4), we argue as above to obtain a bound of the form*

  $$\|y(t) - \overline{y}\| \leqslant C \left(\left\| y^0 - \overline{y} \right\| + \|u - \overline{u}\|_{L^2(0,T;\mathbb{R}^m)} + \|y - \overline{y}\|_{L^2(0,T;\mathbb{R}^d)} \right)$$

  with $C \sim \exp\left(\|u - \overline{u}\|_{L^2(0,T;\mathbb{R}^m)} \right)$. Obtaining a dependence of the constant C with respect to $\|u - \overline{u}\|_{L^2(0,T;\mathbb{R}^m)}$ rather than just $\|u\|_{L^2(0,T;\mathbb{R}^m)}$ is important, as by using the functional and optimality arguments, we will be able to obtain a uniform bound with respect to T of the former, which does not necessarily entail a bound on the latter. The argument for deducing such a bound is identical to the proof of Lemma 5.4.1 – assume that $m = 1$ for notational simplicity, and observe that, since $f_0(\overline{y}) + \overline{u} f_1(\overline{y}) = 0$,

  $$\begin{aligned} y(t) - \overline{y} &= y(t^*) - \overline{y} + \int_{t^*}^{t} \left(f_0(y) - f_0(\overline{y}) \right) \mathrm{d}s + \int_{t^*}^{t} \left(u - \overline{u} \right) \left(f_1(y) - f_1(\overline{y}) \right) \mathrm{d}s \\ &\quad + \int_{t^*}^{t} \left(u - \overline{u} \right) f_1(\overline{y}) \, \mathrm{d}s + \int_{t^*}^{t} \overline{u} \left(f_1(y) - f_1(\overline{y}) \right) \mathrm{d}s. \end{aligned}$$

  One may then proceed as before.

- *It may readily be seen that if the control is of additive rather than multiplicative form, i.e. if $f_1, \dots, f_m$ are nonzero constants, then the constant appearing in the estimate provided by Lemma 5.4.1 will not depend on the time horizon T.*

We state and prove an analogous result for the semilinear heat equation (5.3.5). The proof is almost identical to the ODE case, but we sketch it for the sake of clarity.

Lemma 5.4.3. *Let $T > 0$ be given, and let $\overline{y}$ be as in (5.3.3). For any $u \in L^2((0,T)\times\omega)$ and $y^0 \in L^2(\Omega)$, let $y \in C^0([0,T];L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$ be the unique weak solution to (5.3.5). Then there exists a constant $C = C(f) > 0$ independent of T such that*

$$\|y(t) - \overline{y}\|_{L^2(\Omega)} \leqslant C\left(\left\|y^0 - \overline{y}\right\|_{L^2(\Omega)} + \|u\|_{L^2((0,T)\times\omega)} + \|y - \overline{y}\|_{L^2((0,T)\times\Omega)}\right)$$

holds for all $t \in [0,T]$.

Proof of Lemma 5.4.3. The proof closely follows that of Lemma 5.4.1. We first note that by uniqueness, $y - \overline{y}$ can be shown (see [12]) to coincide with the unique mild solution to

$$\begin{cases} \partial_t z - \Delta z + f(z+\overline{y}) - f(\overline{y}) = u\mathbf{1}_\omega & \text{in } (0,T)\times\Omega \\ z = 0 & \text{on } (0,T)\times\partial\Omega \\ z|_{t=0} = y^0 - \overline{y} & \text{in } \Omega \end{cases}$$

which is given by the Duhamel/variation by constants formula:

$$y(t) - \overline{y} = e^{t\Delta}(y^0 - \overline{y}) + \int_0^t e^{(t-s)\Delta}u(s)\mathbf{1}_\omega \,\mathrm{d}s - \int_0^t e^{(t-s)\Delta}(f(y) - f(\overline{y}))\,\mathrm{d}s, \tag{5.4.2}$$

where $\left\{e^{t\Delta}\right\}_{t>0}$ denotes the heat semigroup on $L^2(\Omega)$ generated by the Dirichlet Laplacian $-\Delta : H^2(\Omega)\cap H_0^1(\Omega) \to L^2(\Omega)$. Of course, (5.4.2) is interpreted as an identity in $L^2(\Omega)$. We may thus proceed and use (5.4.2) throughout.

First suppose that $0 < t \leqslant 1$. Using the well-known property $\left\|e^{t\Delta}\right\| \leqslant e^{-\lambda_1(\Omega)t} \leqslant 1$ of the heat semigroup (where $\lambda_1(\Omega) > 0$ denotes the first eigenvalue of the Dirichlet Laplacian), and the Lipschitz character of f, we find using (5.4.2) that

$$\begin{aligned} \|y(t) - \overline{y}\|_{L^2(\Omega)} &\leqslant \left\|e^{t\Delta}\left(y^0 - \overline{y}\right)\right\|_{L^2(\Omega)} + \int_0^t \left\|e^{(t-s)\Delta}u(s)\right\|_{L^2(\omega)}\,\mathrm{d}s \\ &\quad + \int_0^t \left\|e^{(t-s)\Delta}(f(y(s)) - f(\overline{y}))\right\|_{L^2(\Omega)}\,\mathrm{d}s \\ &\leqslant \left\|y^0 - \overline{y}\right\|_{L^2(\Omega)} + \int_0^t \|u(s)\|_{L^2(\omega)}\,\mathrm{d}s \\ &\quad + C_0\int_0^t \|y(t) - \overline{y}\|_{L^2(\Omega)}\,\mathrm{d}s, \end{aligned}$$

where $C_0 = C_0(f) > 0$ is the Lipschitz constant of f. As $t \leqslant 1$, we may use Cauchy-Schwarz and Grönwall to conclude.

Now suppose that $t \in (1,T]$. Arguing as in the proof of Lemma 5.4.1, we know that there exists a $t^* \in (t-1,t]$ such that

$$\|y(t^*) - \overline{y}\|_{L^2(\Omega)} \leqslant \|y - \overline{y}\|_{L^2((0,T)\times\Omega)} \tag{5.4.3}$$

holds. By writing the Duhamel formula for $y - \overline{y}$ in $[t^*,t]$, namely writing

$$y(t) - \overline{y} = e^{t\Delta}\left(y(t^*) - \overline{y}\right) + \int_{t^*}^t e^{(t-s)\Delta}u(s)\,\mathrm{d}s - \int_{t^*}^t e^{(t-s)\Delta}(f(y) - f(\overline{y}))\,\mathrm{d}s$$

we see just as before that

$$\|y(t) - \overline{y}\|_{L^2(\Omega)} \leqslant \|y(t^*) - \overline{y}\|_{L^2(\Omega)} + \int_0^t \|u(s)\|_{L^2(\omega)}\,\mathrm{d}s + C_0\int_{t^*}^t \|y(t) - \overline{y}\|_{L^2(\Omega)}\,\mathrm{d}s$$

where $C_0 = C_0(f) > 0$ is the Lipschitz constant of f. Using the fact that $t^* - t \leqslant 1$ and (5.4.3), we may, as before, apply Cauchy-Schwarz and Grönwall to conclude. □

We finally show the analog estimate for the semilinear wave equation, which is, after defining the proper functional setup, identical to the proof of Lemma 5.4.3.

Lemma 5.4.4. *Let $T > 0$ be given, and let $\overline{y}$ be as in (5.3.3). For any $u \in L^2((0,T)\times\omega)$ and $\mathbf{y}^0 = (y_1^0, y_2^0) \in H_0^1(\Omega) \times L^2(\Omega)$, let $y \in C^0([0,T]; H_0^1(\Omega)) \cap C^1([0,T]; L^2(\Omega))$ be the unique weak solution to (5.3.1). Then there exists a constant $C = C(f,\Omega) > 0$ independent of T such that*

$$\begin{aligned}&\|y(t) - \overline{y}\|_{H_0^1(\Omega)} + \|\partial_t y(t)\|_{L^2(\Omega)}\\ &\leqslant C\left(\left\|y_1^0 - \overline{y}\right\|_{H_0^1(\Omega)} + \left\|y_2^0\right\|_{L^2(\Omega)} + \|u\|_{L^2((0,T)\times\omega)} + \|y - \overline{y}\|_{H_0^1((0,T)\times\Omega)} + \|\partial_t y\|_{L^2((0,T)\times\Omega)}\right)\end{aligned}$$

holds for all $t \in [0,T]$.

Proof of Lemma 5.4.4. Once (5.3.1) is written as a first order evolution equation in an appropriate Hilbert space X, the proof is identical to that of Lemma 5.4.3. Define the energy space $X := H_0^1(\Omega) \times L^2(\Omega)$, and consider the closed, densely-defined operator

$$A := \begin{bmatrix} 0 & \mathrm{Id} \\ \Delta & 0 \end{bmatrix}, \qquad D(A) = D(\Delta) \times H_0^1(\Omega),$$

where $D(\Delta) = H^2(\Omega) \cap H_0^1(\Omega)$. The operator A is skew-adjoint and thus generates a strongly continuous semigroup $\left\{e^{tA}\right\}_{t>0}$ in X by virtue of the Stone-Lumer-Phillips theorem (see e.g. [262, Theorem 3.8.6]). We now denote

$$\mathbf{y} := \begin{bmatrix} y \\ \partial_t y \end{bmatrix}, \qquad \overline{\mathbf{y}} := \begin{bmatrix} \overline{y} \\ 0 \end{bmatrix}.$$

Analog arguments to those in Lemma 5.4.3 lead us to deduce that

$$\mathbf{y}(t) - \overline{\mathbf{y}} = e^{tA}\left(\mathbf{y}^0 - \overline{\mathbf{y}}\right) + \int_0^t e^{(t-s)A}\begin{bmatrix} 0 \\ u(s)\mathbf{1}_\omega - f(y(s)) + f(\overline{y}) \end{bmatrix} \mathrm{d}s \tag{5.4.4}$$

for $t > 0$ is the unique mild solution to the equation satisfied by the perturbation $\mathbf{y} - \overline{\mathbf{y}}$. Of course, (5.4.4) is interpreted as an identity in X. By virtue of the conservative character of the semigroup, namely $\left\|e^{tA}g\right\|_X = \|g\|_X$ for all $t > 0$ and $g \in X$, we see that one may apply precisely the same arguments as in the proof of Lemma 5.4.3, this time to the integral formulation (5.4.4) in X (with an intermediate application of the Poincaré inequality after using the Lipschitz character of f) to conclude. □

5.5 Proof of Theorem 5.1

In this section, we present the proof of Theorem 5.1, Corollary 5.2.4 and Corollary 5.2.5. The proof of Theorem 5.1 requires a couple of preliminary results. In particular, we will, by means of a quasi-turnpike control strategy, provide bounds – uniform with respect to the time horizon T– of the tracking terms appearing in the definition (5.2.3) of the functional J_T for the optimal control-state pairs (u_T, y_T).

5.5.1 Quasi-turnpike lemmas

Both of the following results are heavily based on the specific choice of target $\overline{y}$ as a steady state of the nonlinear system with 0 control, and on the Lipschitz character of the nonlinear terms.

We begin with the following lemma.

Lemma 5.5.1. *Let $y^0 \in \mathbb{R}^d$ be given, and assume that system (5.2.1) is controllable in some time $T_0 > 0$. Let $T > 0$ be fixed, and let $u_T \in L^2(0,T;\mathbb{R}^m)$ be a global minimizer to J_T defined in (5.2.3), with y_T denoting the associated solution to (5.2.1) with $y_T(0) = y^0$.*

Then, there exists a constant $C = C(f, \phi, T_0, \overline{y}, y^0) > 0$ *independent of* $T > 0$ *such that*

$$\|u_T\|_{L^2(0,T;\mathbb{R}^m)} + \|y_T - \overline{y}\|_{L^2(0,T;\mathbb{R}^d)} + \|y_T(t) - \overline{y}\| \leqslant C \tag{5.5.1}$$

holds for all $t \in [0, T]$.

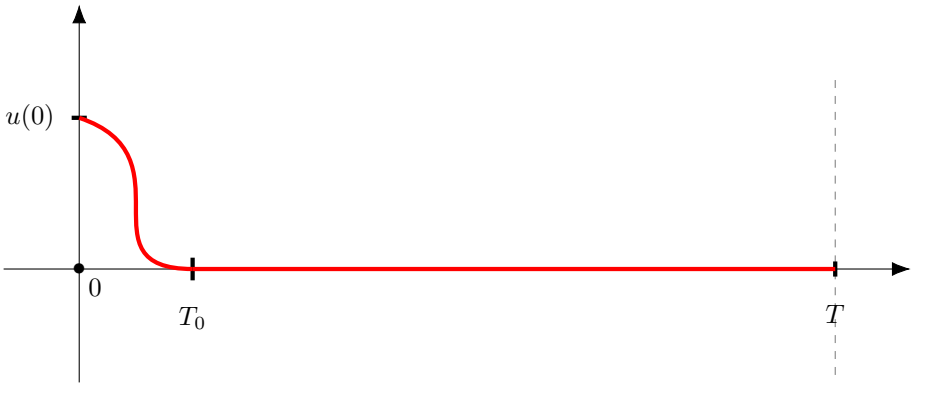
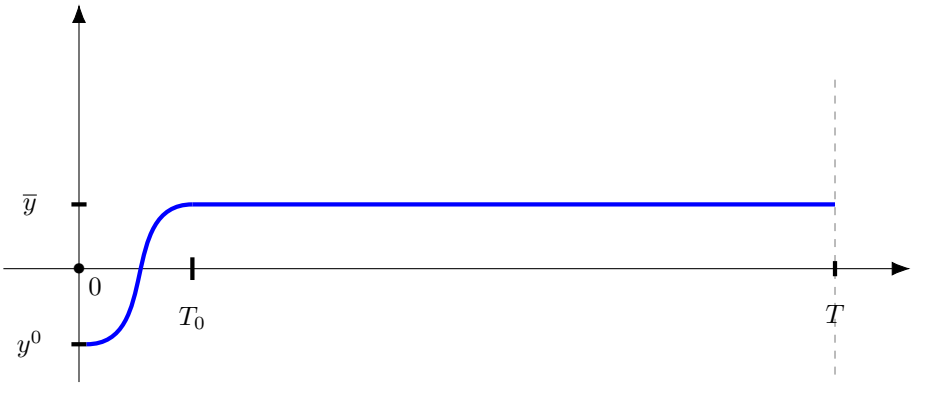


Figure 5.3: **Proof of Lemma 5.5.1.** The first two terms appearing in (5.5.1) also appear in the functional $J_T(u_T)$. We thus construct an admissible quasi-turnpike control u^{aux} (red), for which the corresponding state y^{aux} (blue) coincides with $\overline{y}$ over (T_0, T). In this way, as $J_T(u_T) \leqslant J_T(u^{\text{aux}})$, and $J_T(u^{\text{aux}})$ is independent of T, we can conclude. The estimate of the third term then follows from Lemma 5.4.1.



Proof of Lemma 5.5.1. We begin by considering the case $T \geqslant T_0$. Using the controllability assumption, we know that there exists a control $u^\dagger \in L^2(0, T_0; \mathbb{R}^m)$ such that the corresponding solution $y^\dagger$ to

$$\begin{cases} \dot{y}^\dagger = f\left(y^\dagger, u^\dagger\right) & \text{in } (0, T_0) \\ y^\dagger(0) = y^0 \end{cases}$$

satisfies $y^\dagger(T_0) = \overline{y}$. Now set

$$u^{\text{aux}}(t) := \begin{cases} u^\dagger(t) & \text{in } (0, T_0) \\ 0 & \text{in } (T_0, T) \end{cases}$$

and let y^{aux} be the corresponding solution to (5.2.1) with $y^{\text{aux}}(0) = y^0$. Clearly $y^{\text{aux}}(t) = \overline{y}$ for $t \in [T_0, T]$. Hence, using $\phi \geqslant 0$ and $J_T(u_T) \leqslant J_T(u^{\text{aux}})$, we see that

$$\|y_T - \overline{y}\|^2_{L^2(0,T;\mathbb{R}^d)} + \|u_T\|^2_{L^2(0,T;\mathbb{R}^m)} \leqslant \phi(\overline{y}) + \left\|y^\dagger - \overline{y}\right\|^2_{L^2(0,T_0;\mathbb{R}^d)} + \left\|u^\dagger\right\|^2_{L^2(0,T_0;\mathbb{R}^m)}.$$

As the right hand side in the above inequality is clearly independent of T, we conclude the proof by applying Lemma 5.4.1 after noting the uniform boundedness of $\|u_T\|_{L^2(0,T;\mathbb{R}^m)}$ with respect to $T > 0$.

Now suppose that $T \leqslant T_0$. In this case, we use $\phi \geqslant 0$ and the optimality inequality $J_T(u_T) \leqslant J_T(u_{T_0+1})$ with the effect of obtaining

$$\begin{aligned}\|y_T - \overline{y}\|^2_{L^2(0,T;\mathbb{R}^d)} + \|u_T\|^2_{L^2(0,T;\mathbb{R}^m)} \\ \leqslant \phi\left(y_{T_0+1}(T)\right) + \|y_{T_0+1} - \overline{y}\|^2_{L^2(0,T;\mathbb{R}^d)} + \|u_{T_0+1}\|^2_{L^2(0,T;\mathbb{R}^m)}\end{aligned}$$

Now the trajectory $y_{T_0+1} \in C^0([0, T_0+1];\mathbb{R}^d)$ is uniformly bounded with respect to T by virtue of the case presented just above. Whence, using the continuity of ϕ, as well as $T \leqslant T_0$, we may conclude that

$$\|y_T - \overline{y}\|^2_{L^2(0,T;\mathbb{R}^d)} + \|u_T\|^2_{L^2(0,T;\mathbb{R}^m)} \leqslant C$$

for some $C > 0$ independent of T. We may use Lemma 5.4.1 to conclude. □

We will now focus on an auxiliary control problem with *fixed* endpoints. Namely, given $y^{\tau_1}, y^{\tau_2} \in \mathbb{R}^d$, and $0 \leqslant \tau_1 < \tau_2 \leqslant T$, this problem consists in minimizing the nonnegative functional

$$J_{\tau_1,\tau_2}(u) := \int_{\tau_1}^{\tau_2} \|y(t) - \overline{y}\|^2 \,\mathrm{d}t + \int_{\tau_1}^{\tau_2} \|u(t)\|^2 \,\mathrm{d}t \tag{5.5.2}$$

over all $u \in \mathfrak{U}_{\mathrm{ad}}$, where $y \in C^0([\tau_1, \tau_2];\mathbb{R}^d)$ denotes the unique solution to

$$\begin{cases} \dot{y} = f(y,u) & \text{in } (\tau_1, \tau_2) \\ y(\tau_1) = y^{\tau_1} \end{cases} \tag{5.5.3}$$

where

$$\mathfrak{U}_{\mathrm{ad}} := \left\{u \in L^2(\tau_1, \tau_2;\mathbb{R}^m) \colon y(\tau_2) = y^{\tau_2}\right\}.$$

The following lemma is of key importance in what follows. It ensures that the optimal controls (for J_{τ_1,τ_2}) and trajectories are in fact bounded by means of the distance of the starting point y^{τ_1} and endpoint y^{τ_2} from the running target $\overline{y}$. This estimate will be the cornerstone of the bootstrap argument performed in the proof of Theorem 5.1.

Lemma 5.5.2. *Let $\overline{y} \in \mathbb{R}^d$ be as in (5.2.4), and assume that system (5.2.1) is controllable in some time $T_0 > 0$ in the sense of Definition 6.4.1. Let $r > 0$ be the radius provided by Definition 6.4.1, and let $y^{\tau_1}, y^{\tau_2} \in \mathbb{R}^d$ be such that*

$$\|y^{\tau_i} - \overline{y}\| \leqslant r$$

for $i = 1, 2$. Let $0 \leqslant \tau_1 < \tau_2 \leqslant T$ be fixed such that $\tau_2 - \tau_1 \geqslant 2T_0$, and let $u_T \in \mathfrak{U}_{\mathrm{ad}}$ be a global minimizer to J_{τ_1,τ_2} defined in (5.5.2), with y_T denoting the associated solution to (5.5.3) with $y_T(\tau_2) = y^{\tau_2}$.

Then, there exists a constant $C = C(f, T_0, \overline{y}, r) > 0$ independent of $T, \tau_1, \tau_2 > 0$ such that

$$\|u_T\|^2_{L^2(\tau_1,\tau_2;\mathbb{R}^m)} + \|y_T - \overline{y}\|^2_{L^2(\tau_1,\tau_2;\mathbb{R}^d)} + \|y_T(t) - \overline{y}\|^2 \leqslant C\left(\|y^{\tau_1} - \overline{y}\|^2 + \|y^{\tau_2} - \overline{y}\|^2\right)$$

holds for all $t \in [\tau_1, \tau_2]$. Moreover, the map $r \longmapsto C(f, T_0, \overline{y}, r)$ is non-decreasing as a function from $\mathbb{R}_+$ to $\mathbb{R}_+$.

The key idea of the proof of Lemma 5.5.2 lies in the construction of an auxiliary suboptimal quasi-turnpike control (steering the corresponding trajectory from y^{τ_1} to y^{τ_2} in time $\tau_2 - \tau_1$, whilst remaining at $\overline{y}$ over an interval of length $\tau_2 - \tau_1 - 2T_0$; see the figure just below) in view of estimating each individual addend of $J_{\tau_1,\tau_2}(u_T)$, which is the minimal value of the functional J_{τ_1,τ_2}. This construction will yield the desired result.

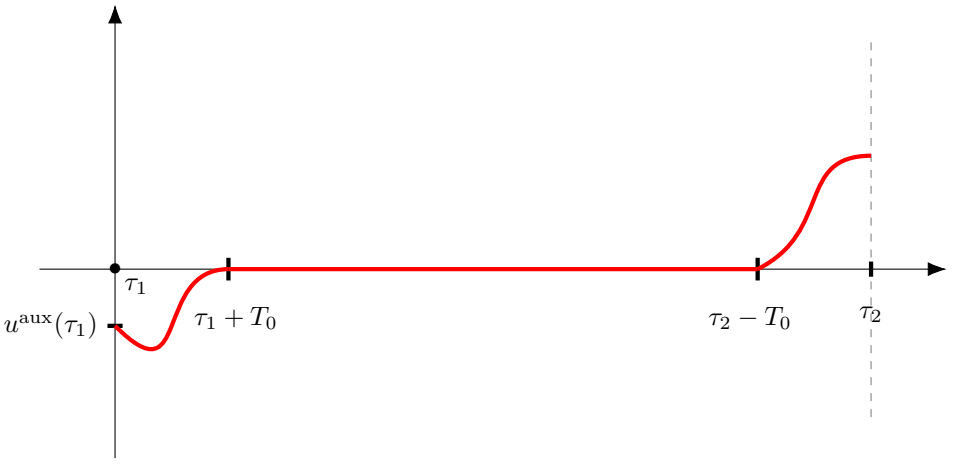
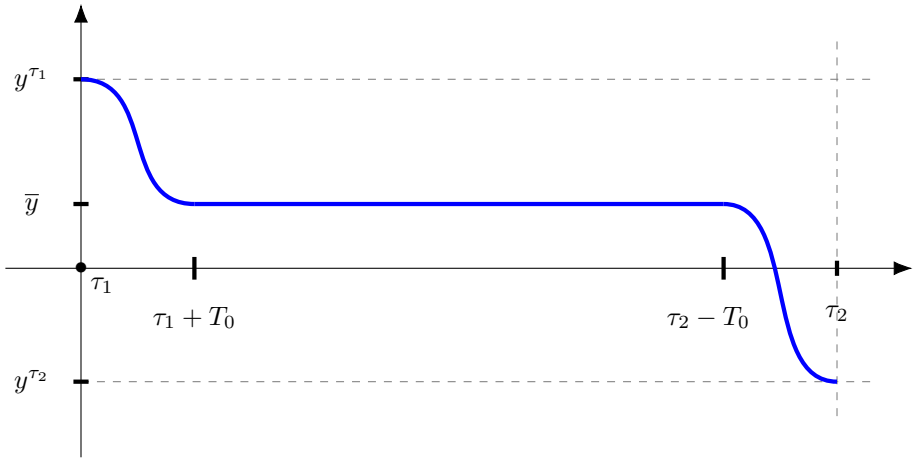


Figure 5.4: **Proof of Lemma 5.5.2.** The first two terms appearing in the estimate implied by Lemma 5.5.2 also appear in the functional $J_{\tau_1,\tau_2}(u_T)$. We thus construct an admissible quasi-turnpike control u^{aux} (red), for which the corresponding state y^{aux} (blue) coincides with $\overline{y}$ over $(\tau_1 + T_0, \tau_2 - T_0)$. In this way, as $J_{\tau_1,\tau_2}(u_T) \leqslant J_{\tau_1,\tau_2}(u^{\mathrm{aux}})$, and $J_{\tau_1,\,\tau_2}(u^{\mathrm{aux}})$ is independent of T, τ_1, τ_2, we can conclude. The estimate of the third term then follows from Lemma 5.4.1.



Proof of Lemma 5.5.2. Using the controllability assumption, we know the following.

- There exists a control $u^\dagger \in L^2(\tau_1, \tau_1 + T_0; \mathbb{R}^m)$ satisfying

$$\left\|u^\dagger\right\|^2_{L^2(\tau_1,\tau_1+T_0;\mathbb{R}^m)} \leqslant C(T_0) \left\|y^{\tau_1} - \overline{y}\right\|^2, \tag{5.5.4}$$

  for some $C(T_0) > 0$, and which is such that the corresponding solution $y^\dagger$ to

$$\begin{cases} \dot{y}^\dagger = f\left(y^\dagger, u^\dagger\right) & \text{in } (\tau_1, \tau_1 + T_0) \\ y^\dagger(\tau_1) = y^{\tau_1} \end{cases} \tag{5.5.5}$$

  satisfies $y^\dagger(\tau_1 + T_0) = \overline{y}$. By integrating (5.5.5), and using the Lipschitz character

of $f_0, \dots, f_m$, Grönwall's inequality, Cauchy-Schwarz and (5.5.4), we see that

$$\begin{aligned}\left\|y^\dagger(t)\right\| &\leqslant C_0\left(\|y^{\tau_1}\| + \left\|u^\dagger\right\|_{L^2(\tau_1,\tau_1+T_0;\mathbb{R}^m)} + 1\right)\exp\left(C_0\left\|u^\dagger\right\|_{L^2(\tau_1,\tau_1+T_0;\mathbb{R}^m)}\right)\\ &\leqslant C_1\Big(\|y^{\tau_1}\| + \|y^{\tau_1} - \overline{y}\| + 1\Big)\exp\Big(C_1\|y^{\tau_1} - \overline{y}\|\Big)\\ &\leqslant C_1\Big(\|y^{\tau_1}\| + r + 1\Big)\exp\Big(C_1 r\Big)\\ &\leqslant C_2\Big(\|\overline{y}\| + r + 1\Big)\exp\Big(C_2 r\Big)\end{aligned} \tag{5.5.6}$$

for some $C_0 = C_0(f, T_0) > 0$, $C_1 = C_1(f, T_0) > 0$, $C_2 = C_2(f, T_0) > 0$, and for every $t \in (\tau_1, \tau_1 + T_0)$. Then, by integrating (5.5.5) once again, and using $f_0(\overline{y}) = 0$, Cauchy-Schwarz and (5.5.6), we moreover see that

$$\begin{aligned}\left\|y^\dagger(t) - \overline{y}\right\| &\leqslant \|y^{\tau_1} - \overline{y}\| + \int_{\tau_1}^t \sum_{j=1}^m \left|u_j^\dagger(s)\right| \left\|f_j(y^\dagger)\right\| \,\mathrm{d}s + \int_{\tau_1}^t \left\|f(y^\dagger) - f(\overline{y})\right\| \,\mathrm{d}s\\ &\leqslant \|y^{\tau_1} - \overline{y}\| + C_3\left\|u^\dagger\right\|_{L^2(\tau_1,\tau_1+T_0;\mathbb{R}^m)} + C(f)\int_{\tau_1}^t \left\|y^\dagger(s) - \overline{y}\right\| \,\mathrm{d}s\end{aligned} \tag{5.5.7}$$

for some $C_3(f, T_0, r, \overline{y}) > 0$, with $C(f) > 0$ being the Lipschitz constant of the vector fields f_j. Finally, applying Grönwall's inequality to (5.5.7) and using (5.5.4), we deduce that

$$\left\|y^\dagger(t) - \overline{y}\right\| \leqslant C_4 \exp\left(C(f)T_0\right)\|y^{\tau_1} - \overline{y}\| \tag{5.5.8}$$

for some $C_4(f, T_0, \overline{y}, r) > 0$ independent of $T, \tau_1, \tau_2 > 0$, and for every $t \in (\tau_1, \tau_1 + T_0)$. Note that in view of (5.5.6), both C_3 and C_4 are non-decreasing with respect to the parameter $r > 0$.

- There exists a control $u^\ddagger \in L^2(\tau_1, \tau_1 + T_0; \mathbb{R}^m)$ satisfying

$$\left\|u^\ddagger\right\|^2_{L^2(\tau_1,\tau_1+T_0;\mathbb{R}^m)} \leqslant C(T_0)\left\|\overline{y} - y^{\tau_2}\right\|^2, \tag{5.5.9}$$

and which is such that the corresponding solution $y^\ddagger$ to

$$\begin{cases}\dot{y}^\ddagger = f\left(y^\ddagger, u^\ddagger\right) & \text{in } (\tau_1, \tau_1 + T_0)\\ y^\ddagger(\tau_1) = \overline{y}\end{cases} \tag{5.5.10}$$

satisfies $y^\ddagger(\tau_1 + T_0) = y^{\tau_2}$. By integrating (5.5.10), and using the Lipschitz character of $f_0, \dots, f_m$, Grönwall's inequality, Cauchy-Schwarz and (5.5.9), we see that

$$\begin{aligned}\left\|y^\ddagger(t)\right\| &\leqslant C_5\left(\|\overline{y}\| + \left\|u^\ddagger\right\|_{L^2(\tau_1,\tau_1+T_0;\mathbb{R}^m)} + 1\right)\exp\left(C_5\left\|u^\ddagger\right\|_{L^2(\tau_1,\tau_1+T_0;\mathbb{R}^m)}\right)\\ &\leqslant C_6\Big(\|\overline{y}\| + \|\overline{y} - y^{\tau_2}\| + 1\Big)\exp\Big(C_6\|\overline{y} - y^{\tau_2}\|\Big)\\ &\leqslant C_6\Big(\|\overline{y}\| + r + 1\Big)\exp\Big(C_6 r\Big)\end{aligned} \tag{5.5.11}$$

for some $C_5(f) > 0$ and $C_6(f, T_0) > 0$, and for every $t \in (\tau_1, \tau_1 + T_0)$. Then, by integrating (5.5.10) once again, and using $f_0(\overline{y}) = 0$, Cauchy-Schwarz and (5.5.11), we moreover see that

$$\begin{aligned}\left\|y^\ddagger(t) - \overline{y}\right\| &\leqslant \int_{\tau_1}^t \sum_{j=1}^m \left|u_j^\ddagger(s)\right| \left\|f_j(y^\ddagger)\right\| \,\mathrm{d}s + \int_{\tau_1}^t \left\|f\left(y^\ddagger\right) - f(\overline{y})\right\| \,\mathrm{d}s\\ &\leqslant C_7\left\|u^\ddagger\right\|_{L^2(\tau_1,\tau_1+T_0;\mathbb{R}^m)} + C(f)\int_{\tau_1}^t \left\|y^\ddagger(s) - \overline{y}\right\| \,\mathrm{d}s\end{aligned} \tag{5.5.12}$$

for some $C_7(f, T_0, r, \overline{y}) > 0$, with $C(f) > 0$ being the Lipschitz constant of the vector fields f_j. Finally, applying Grönwall's inequality to (5.5.12) and using (5.5.9), we deduce that

$$\left\|y^\ddagger(t) - \overline{y}\right\| \leqslant C_8 \exp\left(C(f)T_0\right) \left\|y^{\tau_2} - \overline{y}\right\| \tag{5.5.13}$$

for some $C_8(f, T_0, \overline{y}, r) > 0$ independent of $T, \tau_1, \tau_2 > 0$, and for every $t \in (\tau_1, \tau_1 + T_0)$. Note that in view of (5.5.6), both C_7 and C_8 are non-decreasing with respect to the parameter $r > 0$.

Now set

$$u^{\mathrm{aux}}(t) := \begin{cases} u^\dagger(t) & \text{in } (\tau_1, \tau_1 + T_0) \\ 0 & \text{in } (\tau_1 + T_0, \tau_2 - T_0) \\ u^\ddagger\left(t - (\tau_2 - \tau_1 - T_0)\right) & \text{in } (\tau_2 - T_0, \tau_2), \end{cases}$$

and let y^{aux} be the corresponding solution to (5.5.3). By construction, we have

$$y^{\mathrm{aux}}(t) = y^\dagger(t) \qquad \text{in } [\tau_1, \tau_1 + T_0],$$

and thus

$$y^{\mathrm{aux}}(t) = \overline{y} \qquad \text{in } [\tau_1 + T_0, \tau_2 - T_0], \tag{5.5.14}$$

whereas we also have $y^{\mathrm{aux}}(\tau_2) = y^{\tau_2}$, whence $u^{\mathrm{aux}} \in \mathfrak{U}_{\mathrm{ad}}$.

We now evaluate $J_{\tau_1,\,\tau_2}$ at u^{aux}, which by virtue of a simple change of variable as well as (5.5.14), (5.5.4), (5.5.8), (5.5.9) and (5.5.13), leads us to

$$\begin{aligned} J_{\tau_1,\tau_2}(u^{\mathrm{aux}}) &= \left\|u^\dagger\right\|_{L^2(\tau_1,\tau_1+T_0;\mathbb{R}^m)} + \left\|u^\ddagger\right\|_{L^2(\tau_1,\tau_1+T_0;\mathbb{R}^m)} \\ &\quad + \int_{\tau_1}^{\tau_1+T_0} \left\|y^\dagger(t) - \overline{y}\right\|^2 \mathrm{d}t + \int_{\tau_1}^{\tau_1+T_0} \left\|y^\ddagger(t) - \overline{y}\right\|^2 \mathrm{d}t \\ &\leqslant C_9\Big(\left\|\overline{y} - y^{\tau_1}\right\|^2 + \left\|\overline{y} - y^{\tau_2}\right\|^2 \Big) \end{aligned} \tag{5.5.15}$$

where $C_9 = C_9(f, \overline{y}, T_0, r) > 0$ is independent of $T, \tau_1, \tau_2 > 0$, and is non-decreasing with respect to r. Hence $u_T \in \mathfrak{U}_{\mathrm{ad}}$ is uniformly bounded with respect to $T, \tau_1, \tau_2 > 0$, as in view of (5.5.15) we have

$$\begin{aligned} \left\|y_T - \overline{y}\right\|^2_{L^2(\tau_1,\tau_2;\mathbb{R}^d)} + \left\|u_T\right\|^2_{L^2(\tau_1,\tau_2;\mathbb{R}^m)} &\leqslant J_{\tau_1,\tau_2}\left(u_T\right) \leqslant J_{\tau_1,\tau_2}\left(u^{\mathrm{aux}}\right) \\ &\leqslant C_9\Big(\left\|\overline{y} - y^{\tau_1}\right\|^2 + \left\|\overline{y} - y^{\tau_2}\right\|^2 \Big). \end{aligned}$$

An application of Lemma 5.4.1 combined with the uniform boundedness of $\left\|u_T\right\|_{L^2(\tau_1,\tau_2;\mathbb{R}^m)}$ with respect to $T, \tau_2, \tau_1 > 0$ suffices to conclude. □

Before proceeding with the proof of Theorem 5.1, we will need the following key lemma.

Lemma 5.5.3. *Let X be a Banach space, $T > 0$ and $f \in C^0([0,T]; X)$. For any $\tau \leqslant \frac{T}{2}$, there exist $t_1 \in [0, \tau)$ and $t_2 \in (T - \tau, T]$ such that*

$$\left\|f(t_i)\right\|_X \leqslant \frac{\left\|f\right\|_{L^2(0,T;X)}}{\sqrt{\tau}} \qquad \textit{for } i = 1, 2.$$

Proof of Lemma 5.5.3. Denote

$$\eta(\tau) := \frac{\left\|f\right\|_{L^2(0,T;X)}}{\sqrt{\tau}}.$$

We argue by contradiction. Assume that either

$$\left\|f(t)\right\|_X > \eta(\tau) \qquad \text{for all } t \in [0, \tau)$$

or

$$\|f(t)\|_X > \eta(\tau) \qquad \text{for all } t \in (T-\tau, T].$$

hold. Then we have

$$\int_0^T \|f(t)\|_X^2 \,\mathrm{d}t \geqslant \int_0^\tau \|f(t)\|_X^2 \,\mathrm{d}t + \int_{T-\tau}^T \|f(t)\|_X^2 \,\mathrm{d}t > \tau\eta(\tau)^2.$$

Hence

$$\eta(\tau)^2 < \frac{1}{\tau}\int_0^T \|f(t)\|_X^2 \,\mathrm{d}t = \eta(\tau)^2,$$

which yields a contradiction. This concludes the proof. □

5.5.2 Proof of Theorem 5.1

We are now in a position to prove our first main result.

Proof of Theorem 5.1. We begin by noting that (5.2.9) follows immediately from Lemma 5.5.1. We thus concentrate on proving (5.2.8) – we split the proof in two parts.

Before proceeding, let us first note that by Lemma 5.5.1, there exists some positive constant $C_1(f, T_0, \overline{y}, y^0) > 0$ such that whenever $T \geqslant 2T_0$,

$$\|y_T(t) - \overline{y}\| \leqslant C_1 \qquad \text{for all } \ t \in [0, T]. \tag{5.5.16}$$

Let $r > 0$ be the radius provided by Definition 6.4.1. By Lemma 5.5.2, we know that there exists a constant $C_2(f, T_0, \overline{y}, r) > 0$ such that for any $\tau_1, \tau_2 \in [0, T]$ such that $\tau_2 - \tau_1 \geqslant 2T_0$ and

$$\|y_T(\tau_i) - \overline{y}\| \leqslant r$$

for $i = 1, 2$, the estimate

$$\|y_T(t) - \overline{y}\| \leqslant C_2\Big(\|y_T(\tau_2) - \overline{y}\| + \|y_T(\tau_1) - \overline{y}\|\Big) \qquad \text{for all } \ t \in [\tau_1, \tau_2]$$

holds. Now let

$$\tau > 16C_2^4 + \frac{C_1^2}{r^2} + \frac{4C_1^2C_2^2}{r^2} + T_0 \tag{5.5.17}$$

and let

$$T \geqslant 2\tau + 2T_0$$

be fixed. The choice of the "buffer" time τ will become clear in what follows (in fact, it will also be seen that $T^* := \frac{\tau+T_0}{2}$ in the statement of the theorem).

Part 1: We note that for $t \in \big[0, \tau + T_0\big]$ and $t \in \big[T - (\tau + T_0), T\big]$, the desired estimate (5.2.8) can be obtained without too much difficulty, as the length of both time intervals is independent of T. Indeed, by (5.5.16), for any $\mu > 0$ we have

$$\begin{aligned} \|y_T(t) - \overline{y}\| \ &\leqslant C_1 = C_1 e^{\mu t}\ e^{-\mu t} \\ &\leqslant C_1 e^{\mu(\tau+T_0)}\ \Big(e^{-\mu t}\ + e^{-\mu(T-t)}\Big) \end{aligned} \tag{5.5.18}$$

for $t \in \big[0, \tau + T_0\big]$, and

$$\begin{aligned} \|y_T(t) - \overline{y}\| \ &\leqslant C_1 = C_1 e^{\mu(T-t)}\ e^{-\mu(T-t)} \\ &\leqslant C_1 e^{\mu(\tau+T_0)}\ \Big(e^{-\mu t}\ + e^{-\mu(T-t)}\Big) \end{aligned} \tag{5.5.19}$$

for $t \ \in \big[T - (\tau + T_0), T\big]$.

Thus, it only remains to be seen what happens when $t \ \in \big[\tau + T_0, T - (\tau + T_0)\big]$. We will address this case by means of a bootstrap argument in Part 2 just below.

Part 2: We now aim to show (5.2.8) for $t \ \in \big[\tau + T_0, T - (\tau + T_0)\big]$. To this end, we proceed in three steps.

Step 1). **Preparation.** Since $\tau \leqslant \frac{T}{2}$, by Lemma 5.5.3 and Lemma 5.5.1, there exist a couple of time instances $\tau_1 \in [0, \tau)$ and $\tau_2 \in (T-\tau, T]$ such that

$$\|y_T(\tau_i) - \overline{y}\| \leqslant \frac{\|y_T - \overline{y}\|_{L^2(0,T;\mathbb{R}^d)}}{\sqrt{\tau}} \leqslant \frac{C_1}{\sqrt{\tau}}. \tag{5.5.20}$$

Note that, by virtue of the choice of τ in (5.5.17), we have that $\frac{C_1}{\sqrt{\tau}} \leqslant r$ and thus

$$\|y_T(\tau_i) - \overline{y}\| \leqslant r \tag{5.5.21}$$

also holds. We shall now restrict our analysis onto $[\tau_1, \tau_2]$, and extrapolate onto the subset $\big[\tau, T-\tau\big]$. First note that $u_T|_{[\tau_1,\tau_2]}$ is a global minimizer[2] of J_{τ_1,τ_2} defined in (5.5.2) with fixed endpoints $y^{\tau_1} = y_T(\tau_1)$ and $y^{\tau_2} = y_T(\tau_2)$, and thus clearly $y_T|_{[\tau_1,\tau_2]}$ solves (5.5.3). As

$$\tau_2 - \tau_1 \geqslant T - 2\tau \geqslant 2T_0,$$

in view of (5.5.21), we may use Lemma 5.5.2 with the effect of deducing that

$$\|y_T(t) - \overline{y}\| \leqslant C_2\Big(\|y_T(\tau_1) - \overline{y}\| + \|y_T(\tau_2) - \overline{y}\|\Big) \tag{5.5.22}$$

holds for all $t \in [\tau_1, \tau_2]$. Setting

$$\kappa := \max\left\{1, \frac{C_1}{C_2}\right\},$$

and applying (5.5.20) to inequality (5.5.22), we deduce that

$$\|y_T(t) - \overline{y}\| \leqslant \kappa \frac{2C_2^2}{\sqrt{\tau}} \tag{5.5.23}$$

holds for all $t \in [\tau_1,\ \tau_2]$. As $\tau_1 \leqslant \tau$ and $T - \tau \leqslant \tau_2$, estimate (5.5.23) clearly holds for all $t \in \big[\tau, T-\tau\big]$.

Step 2). **Bootstrap.** Inequality (5.5.23) motivates performing a bootstrap – we will show that for any $n \in \mathbb{N}$ satisfying

$$n \leqslant \frac{1}{\tau}\left(\frac{T}{2} - T_0\right),$$

one has

$$\|y_T(t) - \overline{y}\| \leqslant \frac{\kappa}{2}\left(\frac{4C_2^2}{\sqrt{\tau}}\right)^n \qquad \text{for } t \in \big[n\tau, T - n\tau\big]. \tag{5.5.24}$$

The choice of n is done as to guarantee that $T - 2n\tau \geqslant 2T_0$ in view of a repeated application of Lemma 5.5.2. Note that (5.5.22), combined with the choice of τ in (5.5.17), also implies that

$$\|y_T(t) - \overline{y}\| \leqslant r \tag{5.5.25}$$

for all $t \in [\tau, T-\tau]$.

To prove (5.5.24), we proceed by induction. The case $n = 1$ clearly holds by (5.5.23). Thus, assume that (5.5.24) holds – we aim to show that (5.5.24) holds at step $n+1$. To this end, let

$$n+1 \leqslant \frac{1}{\tau}\left(\frac{T}{2} - T_0\right).$$

[2] This can be shown by contradiction.

This clearly implies that

$$\tau \leqslant \frac{T-2n\tau}{2}. \tag{5.5.26}$$

As in Step 1, since $T-2n\tau \geqslant 2T_0$, it can be seen that $u_T|_{[n\tau,T-n\tau]}$ is a global minimizer of $J_{n\tau,T-n\tau}$ defined in (5.5.2). Taking these facts into account, and noting that (5.5.25) holds[3], we can apply Lemma 5.5.3 on $[n\tau, T-n\tau]$ (note (5.5.26)), and Lemma 5.5.2 with $\tau_1 = n\tau$ and $\tau_2 = T-n\tau$, to deduce that there exist a couple of times $t_1 \in \big[n\tau,(n+1)\tau\big)$ and $t_2 \in \big(T-(n+1)\tau, T-n\tau\big]$ such that

$$\begin{aligned}\|y_T(t_i)-\overline{y}\| &\leqslant \frac{\|y_T-\overline{y}\|_{L^2(n\tau,T-n\tau;\mathbb{R}^d)}}{\sqrt{\tau}}\\ &\leqslant \frac{C_2}{\sqrt{\tau}}\Big(\|y_T(n\tau)-\overline{y}\|+\|y_T(T-n\tau)-\overline{y}\|\Big)\end{aligned}$$

We now use the induction hypothesis (5.5.24) to obtain

$$\|y_T(t_i)-\overline{y}\| \leqslant \kappa\frac{2C_2}{\sqrt{\tau}}\left(\frac{4C_2^2}{\sqrt{\tau}}\right)^n \tag{5.5.27}$$

Now since

$$t_2-t_1 \geqslant T-2(n+1)\tau \geqslant 2T_0,$$

and since $u_T|_{[t_1,t_2}$ is a global minimizer of J_{t_1,t_2} defined in (5.5.2), combining Lemma 5.5.2[4] and (5.5.27) we are led to deduce that

$$\begin{aligned}\|y_T(t)-\overline{y}\| &\leqslant C_2\Big(\|y_T(t_1)-\overline{y}\|+\|y_T(t_2)-\overline{y}\|\Big)\\ &\leqslant \frac{\kappa}{2}\frac{4C_2^2}{\sqrt{\tau}}\left(\frac{4C_2^2}{\sqrt{\tau}}\right)^n\end{aligned} \tag{5.5.28}$$

for $t \in [t_1,t_2]$. Since $t_1 < (n+1)\tau$ and $T-(n+1)\tau < t_2$, estimate (5.5.28) clearly also holds for $t \in \big[(n+1)\tau, T-(n+1)\tau\big]$. Identity (5.5.24) is thus proven.

Step 3). **Conclusion.** We now look to use (5.5.24) as to conclude the proof. Suppose that $t \in \big[\tau+T_0, T-(\tau+T_0)\big]$. We set

$$n(t) := \min\left\{\left\lfloor\frac{t}{\tau+T_0}\right\rfloor, \left\lfloor\frac{T-t}{\tau+T_0}\right\rfloor\right\},$$

where $\lfloor z\rfloor$ denotes the integer part of $z \in \mathbb{R}$. Clearly $n(t) \geqslant 1$ and

$$n(t)\tau \leqslant t \leqslant T-n(t)\tau.$$

Moreover, since $z \mapsto \frac{z-2T_0}{z}$ is non-decreasing,

$$\begin{aligned}n(t) \leqslant \frac{T}{2(\tau+T_0)} &= \frac{T}{2\tau}\frac{2(\tau+T_0)-2T_0}{2(\tau+T_0)}\\ &\leqslant \frac{T}{2\tau}\frac{T-2T_0}{T}\\ &= \frac{1}{\tau}\left(\frac{T}{2}-T_0\right).\end{aligned}$$

We may then apply (5.5.24) to obtain

$$\|y_T(t)-\overline{y}\| \leqslant \frac{\kappa}{2}\left(\frac{4C_2^2}{\sqrt{\tau}}\right)^{n(t)}. \tag{5.5.29}$$

[3] Note that $n\tau \geqslant \tau$ and $T-n\tau \leqslant T-\tau$, so (5.5.25) also holds for $t \in [n\tau, T-n\tau]$, hence Lemma 5.5.2 is applicable.

[4] May be applied once again since (5.5.25) holds for $t = t_1 \geqslant \tau$ and $t = t_2 \leqslant T-\tau$.

As $\tau > 16C_2^2$, we see that $\frac{4C_2^2}{\sqrt{\tau}} < 1$. Moreover, since either $n(t) \geqslant \left\lfloor \frac{t}{\tau+T_0} \right\rfloor - 1$ or $n(t) \geqslant \left\lfloor \frac{T-t}{\tau+T_0} \right\rfloor - 1$ holds, we may rewrite (5.5.29) to obtain

$$\begin{aligned} \|y_T(t) - \overline{y}\| &\leqslant \frac{\kappa}{2} \exp\left(-n(t)\log\left(\frac{\sqrt{\tau}}{4C_2^2}\right)\right) \\ &\leqslant \frac{\kappa}{2}\frac{\sqrt{\tau}}{4C_2^2}\left(\exp\left(-\frac{\log\left(\frac{\sqrt{\tau}}{4C_2^2}\right)}{\tau+T_0}t\right) + \exp\left(-\frac{\log\left(\frac{\sqrt{\tau}}{4C_2^2}\right)}{\tau+T_0}(T-t)\right)\right). \end{aligned} \tag{5.5.30}$$

Whence, (5.2.8) holds for all $t \in \big[\tau+T_0, T-(\tau+T_0)\big]$, with

$$C := \frac{\kappa}{2}\frac{\sqrt{\tau}}{4C_2^2}$$

and

$$\mu := \frac{\log\left(\frac{\sqrt{\tau}}{4C_2^2}\right)}{\tau+T_0} > 0. \tag{5.5.31}$$

By virtue of (5.5.18), (5.5.19) and (5.5.30), we deduce that (5.2.8) holds for all $t \in [0,T]$, with $T^* := \tau + T_0$,

$$C := \max\left\{C_1, \frac{\kappa}{2}\frac{\sqrt{\tau}}{4C_2^2}\right\}$$

and μ as in (5.5.31). This concludes the proof. □

5.5.3 Proof of Corollary 5.2.4

Let us now provide a proof to Corollary 5.2.4.

Proof of Corollary 5.2.4. By Theorem 5.1, with $\phi \equiv 0$, there exist constants $C_1 > 0$ and $\mu_1 > 0$ such that

$$\|y_T(t) - \overline{y}\| \leqslant C_1\Big(e^{-\mu_1 t} + e^{-\mu_1(T-t)}\Big)$$

holds for all $t \in [0,T]$. We now distinguish two cases.

- If $t \in \big[0, \frac{T}{2}\big]$, we also have

$$\|y_T(t) - \overline{y}\| \leqslant C_1\left(e^{-\mu_1 t} + e^{-\mu_1(T-t)}\right) \leqslant 2C_1 e^{-\mu_1\frac{t}{2}}. \tag{5.5.32}$$

  The desired estimates thus holds in this case.

- We now consider the case $t \in \big[\frac{T}{2}, T\big]$. First set

$$u^{\mathrm{aux}}(t) := \begin{cases} u_T(t) & \text{for } t \in \left[0, \dfrac{T}{2}\right] \\ 0 & \text{for } t \in \left[\dfrac{T}{2}, T\right]. \end{cases}$$

  The state y^{aux}, solution to (5.2.1) with $y^{\mathrm{aux}}(0) = y^0$ associated to u^{aux} is precisely

$$y^{\mathrm{aux}}(t) = \begin{cases} y_T(t) & \text{for } t \in \left[0, \dfrac{T}{2}\right] \\ y_T\left(\dfrac{T}{2}\right) & \text{for } t \in \left[\dfrac{T}{2}, T\right]. \end{cases}$$

Using the inequality $J_T(u_T) \leqslant J_T(u^{\mathrm{aux}})$ along with (5.5.32) gives

$$\begin{aligned}\int_{\frac{T}{2}}^{T} \|u_T(t)\|^2 \,\mathrm{d}t + \int_{\frac{T}{2}}^{T} \|y_T(t)-\overline{y}\|^2 \,\mathrm{d}t &\leqslant \int_{\frac{T}{2}}^{T} \left\|y_T\left(\frac{T}{2}\right)-\overline{y}\right\|^2 \mathrm{d}t \\ &\leqslant \int_{\frac{T}{2}}^{T} 4C_1^2 \exp(-\mu_1 T)\,\mathrm{d}t \\ &\leqslant 2C_1^2 T \exp(-\mu_1 T) \\ &\leqslant C_2 \exp\left(-\mu_1 \frac{T}{4}\right), \end{aligned} \tag{5.5.33}$$

for some $C_2 > 0$ independent of $T > 0$.

Now by Lemma 5.4.1, for any $t \in \left[\frac{T}{2}, T\right]$,

$$\|y_T(t)-\overline{y}\| \leqslant C_3\left(\left\|y_T\left(\frac{T}{2}\right)-\overline{y}\right\| + \|u_T\|_{L^2\left(\frac{T}{2},T;\mathbb{R}^m\right)} + \|y_T-\overline{y}\|_{L^2\left(\frac{T}{2},T;\mathbb{R}^d\right)}\right)$$

for some $C_3 > 0$ of the form

$$C_3 \sim \exp\left(\|u_T\|_{L^2\left(\frac{T}{2},T;\mathbb{R}^m\right)}\right).$$

In view of (5.2.9), $\|u_T\|_{L^2\left(\frac{T}{2},T;\mathbb{R}^m\right)}$ is bounded uniformly with respect to T, and thus $C_3 > 0$ is independent of T. Combining the above estimate with (5.5.33) leads us to

$$\|y_T(t)-\overline{y}\| \leqslant C_4 \exp\left(-\mu_1 \frac{T}{4}\right) \leqslant C_4 \exp\left(-\mu_1 \frac{t}{4}\right), \tag{5.5.34}$$

for some $C_4 > 0$ independent of $T > 0$ and for all $t \in \left[\frac{T}{2}, T\right]$.

Combining (5.5.32) and (5.5.34), we see that for $C_5 := \max\{2C_1, C_4\}$ and $\mu := \frac{\mu_1}{4}$, the stabilization estimate

$$\|y_T(t)-\overline{y}\| \leqslant C_5 e^{-\mu t},$$

for all $t \in [0,T]$. This concludes the proof. □

5.5.4 Proof of Corollary 5.2.5

We finish this section with the proof of Corollary 5.2.5, which stipulates an exponential decay of optimal controls in the context of driftless control-affine systems, namely (5.2.1) with a nonlinearity of the form

$$f(y,u) = \sum_{j=1}^{m} u_j f_j(y) \qquad \text{for} \quad (y,u) \in \mathbb{R}^d \times \mathbb{R}^m. \tag{5.5.35}$$

We recall that, here, $f_1, \ldots, f_m \in \mathrm{Lip}(\mathbb{R}^d; \mathbb{R}^d)$.

We begin with the following simple result.

Lemma 5.5.4. *Let $T_0 > 0$, $y^0 \in \mathbb{R}^d$ and $u_{T_0} \in L^2(0,T_0;\mathbb{R}^m)$ be given, and let $y_{T_0} \in C^0([0,T_0];\mathbb{R}^d)$ be the unique solution to*

$$\begin{cases} \dot{y}_{T_0} = f(y_{T_0}, u_{T_0}) & \text{in} \quad (0,T_0) \\ y_{T_0}(0) = y^0 \end{cases} \tag{5.5.36}$$

with f as in (5.5.35). Let $T > 0$, and define

$$u_T(t) := \frac{T_0}{T} u_{T_0}\left(t\frac{T_0}{T}\right) \qquad \text{for} \quad t \in [0,T],$$

and

$$y_T(t) := y_{T_0}\left(t\frac{T_0}{T}\right) \qquad \text{for } \ t \in [0,T].$$

Then $y_T \in C^0([0,T];\mathbb{R}^d)$ *is the unique solution to* (5.2.1) *with* $y_T(0) = y^0$ *associated to* u_T.

This sort of time-scaling in the context of driftless control affine systems is commonly used in control theoretical contexts – a canonical example is the proof of the Chow-Rashevskii controllability theorem, see [69, Chapter 3, Section 3.3]. We provide the short proof for completeness.

Proof of Lemma 7.2.1. Using the fact that y_{T_0} is the solution to (6.7.3) and the change of variable $\tau = s\frac{T}{T_0}$, we see that

$$\begin{aligned} y_T(t) := y_{T_0}\left(t\frac{T_0}{T}\right) &= y^0 + \int_0^{t\frac{T_0}{T}} f(y_{T_0}(s), u_{T_0}(s))\,\mathrm{d}s \\ &= y^0 + \int_0^t \frac{T_0}{T} f\left(y_{T_0}\left(\tau\frac{T_0}{T}\right), u_{T_0}\left(\tau\frac{T_0}{T}\right)\right)\,\mathrm{d}\tau \\ &= y^0 + \int_0^t f\left(y_T(\tau), u_T(\tau)\right)\,\mathrm{d}\tau. \end{aligned}$$

It follows that y_T solves (5.2.1) with $y_T(0) = y^0$, and we conclude by uniqueness. □

Proof of Corollary 5.2.5. As $t \longmapsto \|u_T(t)\|^2$ is in $L^1(0,T)$, by the Lebesgue differentiation theorem, we have

$$\|u_T(t)\|^2 = \lim_{h \searrow 0^+} \frac{1}{h} \int_t^{t+h} \|u_T(s)\|^2\,\mathrm{d}s$$

for almost every $t \in (0,T)$. Hence, we will aim at estimating the integral on the right hand side by constructing an appropriate auxiliary suboptimal auxiliary control, and conclude by passing to the limit as $h \to 0^+$. We will split the proof in two parts, namely separate the proof of (5.2.11) (i.e. $\phi \not\equiv 0$) and (5.2.12) (i.e. $\phi \equiv 0$); they mainly differ in the construction of the suboptimal auxiliary control required very last estimate (5.5.45) before concluding.

Part 1: Proof of (5.2.11)**.** Fix $t \in [0,T)$ and $0 < h \ll 1$ so that $t + 2h^2 + 2h \in [0,T]$. Let us set

$$u^{\mathrm{aux}}(s) := \begin{cases} u_T(s) & \text{for } \ s \in [0,t] \\ \dfrac{1}{2} u_T\left(t + \dfrac{s-t}{2}\right) & \text{for } \ s \in \left(t, t+2h^2\right] \\ \dfrac{h+2}{2} u_T\left(\left(\dfrac{h+2}{2}\right)s - \dfrac{h+2}{2}(t+2h^2) + t + h^2\right) & \text{for } t \in \left(t+2h^2, t+2h^2+2h\right] \\ u_T(s) & \text{for } s \in \left(t+2h^2+2h, T\right]. \end{cases}$$

The specific choice of u^{aux} will become clear in what follows – the factor h in the third line will be essential in the subsequent estimates. By Lemma 7.2.1, the state y^{aux}, solution to (5.2.1) associated to u^{aux} is precisely

$$y^{\mathrm{aux}}(s) = \begin{cases} y_T(s) & \text{for } \ s \in [0,t] \\ y_T\left(t + \dfrac{s-t}{2}\right) & \text{for } \ s \in \left(t, t+2h^2\right] \\ y_T\left(\left(\dfrac{h+2}{2}\right)s - \dfrac{h+2}{2}(t+2h^2) + t + h^2\right) & \text{for } \ t \in \left(t+2h^2, t+2h^2+2h\right] \\ y_T(s) & \text{for } \ t \in \left(t+2h^2+2h, T\right]. \end{cases}$$

Note in particular that $y^{\mathrm{aux}}(T) = y_T(T)$, whence $\phi(y^{\mathrm{aux}}(T)) = \phi(y_T(T))$. Our goal is then to rewrite the simple inequality $J_T(u_T) \leqslant J_T(u^{\mathrm{aux}})$ as to estimate the infinitesimal average of $\|u_T(t)\|^2$ taken about t by the infinitesimal average of $\|y_T(t)-\overline{y}\|^2$ taken about t, for which we have an exponential estimate by Theorem 5.1. We proceed as follows.

- On one hand, using the change of variable $\tau \coloneqq t + \frac{s-t}{2}$ we see that

$$\begin{aligned}\int_t^{t+2h^2} \|u^{\mathrm{aux}}(s)\|^2 \,\mathrm{d}s &= \int_t^{t+2h^2} \left\| \frac{1}{2} u_T\left(t + \frac{s-t}{2}\right)\right\|^2 \mathrm{d}s \\ &= \frac{1}{2}\int_t^{t+h^2} \|u_T(\tau)\|^2 \,\mathrm{d}\tau. \end{aligned} \tag{5.5.37}$$

  On another hand, via $\tau \coloneqq \left(\frac{h+2}{2}\right) s - \frac{h+2}{2}(t+2h^2) + t + h^2$ we see that

$$\begin{aligned}&\int_{t+2h^2}^{t+2h^2+2h} \|u^{\mathrm{aux}}(s)\|^2 \,\mathrm{d}s \\ &= \int_{t+2h^2}^{t+2h^2+2h} \left\| \frac{h+2}{2} u_T\left(\left(\frac{h+2}{2}\right) s - \frac{h+2}{2}(t+2h^2) + t + h^2\right)\right\|^2 \mathrm{d}s \\ &= \frac{h+2}{2}\int_{t+h^2}^{t+h^2+2h} \|u_T(\tau)\|^2 \,\mathrm{d}\tau. \end{aligned} \tag{5.5.38}$$

  Combining (5.5.37) and (5.5.38) and since $0 < h \ll 1$, it follows that

$$\begin{aligned}\int_0^T \|u^{\mathrm{aux}}(s)\|^2 \,\mathrm{d}s &= \int_0^t \|u^{\mathrm{aux}}(s)\|^2 \,\mathrm{d}s + \int_t^{t+2h^2} \|u^{\mathrm{aux}}(s)\|^2 \,\mathrm{d}s \\ &\quad + \int_{t+2h^2}^{t+2h^2+2h} \|u^{\mathrm{aux}}(s)\|^2 \,\mathrm{d}s + \int_{t+2h^2+2h}^{T} \|u^{\mathrm{aux}}(s)\|^2 \,\mathrm{d}s \\ &= \int_0^t \|u_T(s)\|^2 \,\mathrm{d}s + \frac{1}{2}\int_t^{t+h^2} \|u_T(\tau)\|^2 \,\mathrm{d}\tau \\ &\quad + \frac{h+2}{2}\int_{t+h^2}^{T+h^2+2h} \|u_T(\tau)\|^2 \,\mathrm{d}\tau + \int_{t+2h^2+2h}^{T} \|u_T(\tau)\|^2 \,\mathrm{d}\tau \\ &\leqslant \int_0^T \|u_T(s)\|^2 \,\mathrm{d}s - \frac{1}{2}\int_t^{t+h} \|u_T(s)\|^2 \,\mathrm{d}s \\ &\quad + \frac{h}{2}\int_{t+h^2}^{t+h^2+2h} \|u_T(\tau)\|^2 \,\mathrm{d}\tau. \end{aligned} \tag{5.5.39}$$

- We now focus on rewriting the state tracking term. On one hand, by means of $\tau \coloneqq t + \frac{s-t}{2}$ we see that

$$\begin{aligned}\int_t^{t+2h^2} \|y^{\mathrm{aux}}(s) - y\|^2 \,\mathrm{d}s &= \int_t^{t+2h^2} \left\| y_T\left(t + \frac{s-t}{2}\right) - \overline{y}\right\|^2 \mathrm{d}s \\ &= 2\int_t^{t+h^2} \|y_T(\tau) - \overline{y}\|^2 \,\mathrm{d}\tau. \end{aligned} \tag{5.5.40}$$

  On another hand, via $\tau \coloneqq \left(\frac{h+2}{2}\right) s - \frac{h+2}{2}(t+2h^2) + t + h^2$ we see that

$$\begin{aligned}&\int_{t+2h^2}^{t+2h^2+2h} \|y^{\mathrm{aux}}(s) - \overline{y}\|^2 \,\mathrm{d}s \\ &= \int_{t+2h^2}^{t+2h^2+2h} \left\| y_T\left(\left(\frac{h+2}{2}\right) s - \frac{h+2}{2}(t+2h^2) + t + h^2\right) - \overline{y}\right\|^2 \mathrm{d}s \\ &= \frac{2}{h+2}\int_{t+h^2}^{t+h^2+2h} \|y_T(\tau) - \overline{y}\|^2 \,\mathrm{d}\tau. \end{aligned} \tag{5.5.41}$$

Combining (5.5.40) and (5.5.41) and since $0 < h \ll 1$, we obtain

$$\begin{aligned}\int_0^T \|y^{\mathrm{aux}}(s)-\overline{y}\|^2\,\mathrm{d}s &= \int_0^t \|y^{\mathrm{aux}}(s)-\overline{y}\|^2\,\mathrm{d}s + \int_t^{t+2h^2} \|y^{\mathrm{aux}}(s)-\overline{y}\|^2\,\mathrm{d}s\\ &\quad + \int_{t+2h^2}^{t+2h^2+2h} \|y^{\mathrm{aux}}(s)-\overline{y}\|^2\,\mathrm{d}s + \int_{t+2h^2+2h}^{T} \|y^{\mathrm{aux}}(s)-\overline{y}\|^2\,\mathrm{d}s\\ &= \int_0^t \|y_T(s)-\overline{y}\|^2\,\mathrm{d}s + 2\int_t^{t+h^2} \|y_T(\tau)-\overline{y}\|^2\,\mathrm{d}\tau\\ &\quad + \frac{2}{h+2}\int_{t+h^2}^{t+h^2+2h^2} \|y_T(\tau)-\overline{y}\|^2\,\mathrm{d}\tau + \int_{t+2h^2+2h}^{T} \|y_T(\tau)-\overline{y}\|^2\,\mathrm{d}\tau\\ &\leqslant \int_0^T \|y_T(s)-\overline{y}\|^2\,\mathrm{d}s + \int_t^{t+h} \|y_T(s)-\overline{y}\|^2\,\mathrm{d}s. \quad (5.5.42)\end{aligned}$$

We may now proceed with the main argument. Using the optimality of u_T, and applying (5.5.39) and (5.5.42), we see that

$$\begin{aligned}J_T(u_T) \leqslant J_T(u^{\mathrm{aux}}) &= \phi(y^{\mathrm{aux}}(T)) + \int_0^T \|y^{\mathrm{aux}}(s)-\overline{y}\|^2\,\mathrm{d}s + \int_0^T \|u^{\mathrm{aux}}(s)\|^2\,\mathrm{d}s\\ &= \phi(y_T(T)) + \int_0^T \|y_T(s)-\overline{y}\|^2 + \int_t^{t+h} \|y_T(s)-\overline{y}\|^2\,\mathrm{d}s\\ &\quad + \int_0^T \|u_T(s)\|^2\,\mathrm{d}s - \frac{1}{2}\int_t^{t+h} \|u_T(s)\|^2\,\mathrm{d}s\\ &\quad + \frac{h}{2}\int_{t+h^2}^{t+h^2+2h} \|u_T(s)\|^2\,\mathrm{d}s. \quad (5.5.43)\end{aligned}$$

From (5.5.43), one clearly sees that

$$\frac{1}{2}\int_t^{t+h} \|u_T(s)\|^2\,\mathrm{d}s \leqslant \int_t^{t+h} \|y_T(s)-\overline{y}\|^2\,\mathrm{d}s + \frac{h}{2}\int_{t+h^2}^{t+h^2+2h} \|u_T(s)\|^2\,\mathrm{d}s. \quad (5.5.44)$$

We combine estimate (5.5.44) with (5.2.8) to deduce that

$$\begin{aligned}\frac{1}{h}\int_t^{t+h} \|u_T(s)\|^2\,\mathrm{d}s &\lesssim \frac{1}{h}\int_t^{t+h} \|y_T(s)-\overline{y}\|^2\,\mathrm{d}s + \int_{t+h^2}^{t+h^2+2h} \|u_T(s)\|^2\,\mathrm{d}s\\ &\leqslant \frac{C}{h}\int_t^{t+h} \left(e^{-\mu s}+e^{-\mu(T-s)}\right)^2\,\mathrm{d}s + \int_{t+h^2}^{t+h^2+2h} \|u_T(s)\|^2\,\mathrm{d}s\\ &\leqslant \frac{C}{h}\int_t^{t+h} \left(e^{-\mu t}+e^{-\mu(T-t-h)}\right)^2\,\mathrm{d}s + \int_{t+h^2}^{t+h^2+2h} \|u_T(s)\|^2\,\mathrm{d}s\\ &= C\left(e^{-\mu t}+e^{-\mu(T-t-h)}\right)^2 + \int_{t+h^2}^{t+h^2+2h} \|u_T(s)\|^2\,\mathrm{d}s \quad (5.5.45)\end{aligned}$$

for some $C > 0$ independent of T. Thus, by using the Lebesgue differentiation theorem and the Lebesgue dominated convergence theorem (applied to the integrable function $s \mapsto \|u_T(s)\|^2 \mathbf{1}_{(t+h^2,t+h^2+2h)}(s)$) in (5.5.45), we deduce that

$$\|u_T(t)\| = \lim_{h\searrow 0^+}\left(\frac{1}{h}\int_t^{t+h} \|u_T(s)\|^2\,\mathrm{d}s\right)^{1/2} \leqslant C\left(e^{-\mu t}+e^{-\mu(T-t)}\right),$$

as desired.

Part 2: Proof of (5.2.12). This part is somewhat simpler due to the fact that $\phi \equiv 0$. Hence the suboptimal control has a much simpler structure as we do not require a factor

of h to deal with a remainder term as in (5.5.43). From there on, the arguments are identical to those above.

Fix any $t \in [0,T)$ and $0 < h \ll 1$, so that $t+2h \in [0,T]$ and set

$$u^{\mathrm{aux}}(s) := \begin{cases} u_T(s) & \text{for } \ s \in [0,t] \\ \dfrac{1}{2} u_T\left(t + \dfrac{s-t}{2}\right) & \text{for } \ s \in (t, t+2h] \\ u_T(s-h) & \text{for } s \in (t+2h, T]. \end{cases}$$

By Lemma 7.2.1, the state y^{aux}, solution to (5.2.1) associated to u^{aux} is precisely

$$y^{\mathrm{aux}}(s) = \begin{cases} y_T(s) & \text{for } \ s \in [0,t] \\ y_T\left(t + \dfrac{s-t}{2}\right) & \text{for } \ s \in (t, t+2h] \\ y_T(s-h) & \text{for } \ t \in (t+2h, T]. \end{cases}$$

Arguing by means of simple changes of variable just as we did for obtaining (5.5.39) and (5.5.42), and using the suboptimality of u^{aux}, we can readily see that

$$\begin{aligned} J_T(u_T) \leqslant J_T(u^{\mathrm{aux}}) &= \int_0^T \|u^{\mathrm{aux}}(s)\|^2 \,\mathrm{d}s + \int_0^T \|y^{\mathrm{aux}}(s) - \overline{y}\|^2 \,\mathrm{d}s \\ &= \int_0^{T-h} \|u_T(s)\|^2 \,\mathrm{d}s - \frac{1}{2}\int_t^{t+h} \|u_T(s)\|^2 \,\mathrm{d}s \\ &\quad + \int_0^{T-h} \|y_T(s) - \overline{y}\|^2 + \int_t^{t+h} \|y_T(s) - \overline{y}\|^2 \,\mathrm{d}s \\ &\leqslant \int_0^{T} \|u_T(s)\|^2 \,\mathrm{d}s - \frac{1}{2}\int_t^{t+h} \|u_T(s)\|^2 \,\mathrm{d}s \\ &\quad + \int_0^{T} \|y_T(s) - \overline{y}\|^2 + \int_t^{t+h} \|y_T(s) - \overline{y}\|^2 \,\mathrm{d}s. \end{aligned} \tag{5.5.46}$$

From (5.5.46), one clearly sees that

$$\frac{1}{2}\int_t^{t+h} \|u_T(s)\|^2 \,\mathrm{d}s \leqslant \int_t^{t+h} \|y_T(s) - \overline{y}\|^2 \,\mathrm{d}s. \tag{5.5.47}$$

We combine estimate (5.5.47) with (5.2.12) to deduce that

$$\begin{aligned} \frac{1}{h}\int_t^{t+h} \|u_T(s)\|^2 \,\mathrm{d}s \lesssim \frac{1}{h}\int_t^{t+h} \|y_T(s) - \overline{y}\|^2 \,\mathrm{d}s &\leqslant \frac{C}{h}\int_t^{t+h} e^{-2\mu s} \,\mathrm{d}s \\ &\leqslant \frac{C}{h}\int_t^{t+h} e^{-2\mu t} \,\mathrm{d}s \\ &= Ce^{-2\mu t}. \end{aligned} \tag{5.5.48}$$

Thus by the Lebesgue differentiation theorem, using (5.5.48) we deduce that

$$\|u_T(t)\| = \lim_{h \searrow 0^+} \left(\frac{1}{h}\int_t^{t+h} \|u_T(s)\|^2 \,\mathrm{d}s\right)^{1/2} \leqslant Ce^{-\mu t},$$

as desired. This concludes the proof. □

5.6 Proof of Theorem 5.2

In this section, we provide details of the proof of Theorem 5.2. The proof of Corollary 5.3.2 follows by repeating the proof of Corollary 5.2.4 in the appropriate functional setting, so we omit it.

Proof of Theorem 5.2. Once (5.3.1) is written as a first order evolution equation set in $X := H_0^1(\Omega) \times L^2(\Omega)$ (see the proof of Lemma 5.4.4 for this setup), the only noticeable difference in the proof of Theorem 5.2 with respect to the proof of Theorem 5.1 are the specific quasi-turnpike lemmas one applies in the preparation (Lemma 5.6.1 in Part 1 & Step 1 of Part 2) and bootstrap (Lemma 5.6.3 in Step 2). So one simply repeats the proof of Theorem 5.1 whilst applying Lemma 5.6.1, Lemma 5.6.3 and Lemma 5.5.3 with X as above. Whence, the proof follows from these two lemmas, stated and proven just below. □

Lemma 5.6.1. *Let $\mathbf{y}^0 = (y_1^0, y_2^0) \in H_0^1(\Omega) \times L^2(\Omega)$ be given. Let $T > 0$ be fixed, and let $u_T \in L^2((0,T) \times \omega)$ be a global minimizer to J_T defined in (5.3.2), with y_T denoting the associated solution to (5.3.1). Then, there exists a constant $C = C(f, \phi, \omega, \Omega, \overline{y}, \mathbf{y}^0) > 0$ independent of $T > 0$ such that*

$$J_T(u_T) + \|y_T(t) - \overline{y}\|^2_{H_0^1(\Omega)} + \|\partial_t y_T(t)\|^2_{L^2(\Omega)} \leqslant C$$

holds for all $t \in [0,T]$.

Proof of Lemma 5.6.1. The proof follows the lines of that of Lemma 5.5.1, simply adapted to the PDE setting. Fix $T_0 \geqslant T_{\min}$ where $T_{\min} = T_{\min}(\omega, \Omega) > 0$ is the minimal controllability time for the semilinear wave equation.

We begin by considering the case $T > T_0$. By controllability, we know that exists some control $u^\dagger \in L^2((0,T_0) \times \omega)$ such that the corresponding solution $y^\dagger$ to

$$\begin{cases} \partial_t^2 y^\dagger - \Delta y^\dagger + f(y^\dagger) = u^\dagger \mathbf{1}_\omega & \text{in } (0,T_0) \times \Omega \\ y^\dagger = 0 & \text{on } (0,T_0) \times \partial\Omega \\ (y^\dagger, \partial_t y^\dagger)|_{t=0} = \mathbf{y}^0 & \text{in } \Omega. \end{cases}$$

satisfies $y^\dagger(T_0) = \overline{y}$ and $\partial_t y^\dagger(T_0) = 0$ a.e. in Ω. Now set

$$u^{\text{aux}}(t) := \begin{cases} u^\dagger(t) & \text{in } (0,T_0) \\ 0 & \text{in } (T_0,T) \end{cases}$$

and let y^{aux} be the corresponding solution to (5.3.1). Clearly

$$y^{\text{aux}}(t) = \overline{y} \quad \text{and} \quad \partial_t y^{\text{aux}}(t) = 0 \qquad \text{for } t \in [T_0,T] \text{ a.e. in } \Omega.$$

Combining this fact with $J_T(u_T) \leqslant J_T(u^{\text{aux}})$, we see that

$$J_T(u_T) \leqslant \phi(\overline{y}) + \left\|y^\dagger - \overline{y}\right\|^2_{L^2(0,T_0;H_0^1(\Omega))} + \left\|\partial_t y^\dagger\right\|^2_{L^2((0,T)\times\Omega)} + \left\|u^\dagger\right\|^2_{L^2((0,T_0)\times\omega)}.$$

As the right hand side in the above inequality is clearly independent of T, we conclude by applying Lemma 5.4.4.

Now suppose that $T \leqslant T_0$. We use the optimality inequality $J_T(u_T) \leqslant J_T(u_{T_0+1})$ to obtain

$$\begin{aligned} J_T(u_T) \leqslant \phi\left(y_{T_0+1}(T)\right) + & \left\|y_{T_0+1} - \overline{y}\right\|^2_{L^2(0,T;H_0^1(\Omega))} \\ & + \left\|\partial_t y_{T_0+1}\right\|^2_{L^2(0,T;L^2(\Omega))} + \left\|u_{T_0+1}\right\|^2_{L^2((0,T)\times\omega)}. \end{aligned}$$

By the previous case addressed just above, the trajectory $y_{T_0+1} \in C^0([0,T_0+1];L^2(\Omega))$ is bounded uniformly with respect to T. Hence, using the fact that $\phi \in C^0(L^2(\Omega);\mathbb{R}_+)$ and $T \leqslant T_0$, we deduce that

$$J_T(u_T) \leqslant C \tag{5.6.1}$$

for some $C > 0$ independent of T. Combining (5.6.1) with Lemma 5.4.4 allows us to conclude. □

We note that since $f \in \mathrm{Lip}(\mathbb{R})$, following the spirit of our finite-dimensional arguments, since (5.3.1) is a Lipschitz perturbation of an exactly controllable linear system, the following claim holds.

Claim 5.6.2 (Cost estimate)**.** *Let $T_0 > T_{\min}$, where $T_{\min} = T_{\min}(\Omega,\omega) > 0$ is the minimal controllability time for (5.3.1). There exists $r > 0$ and $C = C(T_0,\omega,f) > 0$ such that*

$$\inf_{\substack{u \\ \text{such that} \\ (y,\partial_t y)|_{t=0}=\mathbf{y}^0 \\ \text{and} \\ (y,\partial_t y)|_{t=T_0}=(\overline{y},0)}} \|u\|^2_{L^2((0,T_0)\times\omega)} \leqslant C\left(\left\|y_1^0-\overline{y}\right\|^2_{H_0^1(\Omega)} + \left\|y_2^0\right\|^2_{L^2(\Omega)}\right),$$

and

$$\inf_{\substack{u \\ \text{such that} \\ (y,\partial_t y)|_{t=0}=(\overline{y},0) \\ \text{and} \\ (y,\partial_t y)|_{t=T_0}=\mathbf{y}^1}} \|u\|^2_{L^2((0,T_0)\times\omega)} \leqslant C\left(\left\|y_1^1-\overline{y}\right\|^2_{H_0^1(\Omega)} + \left\|y_2^1\right\|^2_{L^2(\Omega)}\right),$$

hold for any $\mathbf{y}^0 = \left(y_1^0, y_2^0\right)$ and $\mathbf{y}^1 = \left(y_1^1, y_2^1\right)$ such that

$$\mathbf{y}^0, \mathbf{y}^1 \in \left\{ \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \in H_0^1(\Omega)\times L^2(\Omega) \,:\, \left\| \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} \overline{y} \\ 0 \end{bmatrix} \right\|_{H_0^1(\Omega)\times L^2(\Omega)} \leqslant r \right\},$$

where y solves (5.3.1) and $\overline{y} \in H_0^1(\Omega)$ is fixed as in (5.3.3).

As in the finite-dimensional case, the second quasi-turnpike result is one for an auxiliary control problem with fixed endpoints. For $0 \leqslant \tau_1 < \tau_2 \leqslant T$ and given $\mathbf{y}^{\tau_1}, \mathbf{y}^{\tau_2} \in H_0^1(\Omega)\times L^2(\Omega)$, this auxiliary problem consists in minimizing the nonnegative functional

$$J_{\tau_1,\tau_2}(u) := \int_{\tau_1}^{\tau_2} \|y(t)-\overline{y}\|^2_{H_0^1(\Omega)}\,\mathrm{d}t + \int_{\tau_1}^{\tau_2} \|\partial_t y(t)\|^2_{L^2(\Omega)} + \int_{\tau_1}^{\tau_2} \|u(t)\|^2_{L^2(\omega)}\,\mathrm{d}t \tag{5.6.2}$$

over all $u \in \mathfrak{U}_{\mathrm{ad}}$, where $y \in C^0([\tau_1,\tau_2];H_0^1(\Omega)) \cap C^1([\tau_1,\tau_2];L^2(\Omega))$ denotes the unique solution to

$$\begin{cases} \partial_t^2 y - \Delta y + f(y) = u\mathbf{1}_\omega & \text{in } (\tau_1,\tau_2)\times\Omega \\ y = 0 & \text{on } (\tau_1,\tau_2)\times\partial\Omega \\ (y,\partial_t y)|_{t=\tau_1} = \mathbf{y}^{\tau_1} & \text{in } \Omega. \end{cases} \tag{5.6.3}$$

and where

$$\mathfrak{U}_{\mathrm{ad}} := \left\{u \in L^2((\tau_1,\tau_2)\times\omega) \colon (y,\ \partial_t y)|_{t=\tau_2} = \mathbf{y}^{\tau_2}\right\}.$$

We recall that $f \in \mathrm{Lip}(\mathbb{R})$.

We now state and prove the wave equation analog of Lemma 5.5.2, which we recall, is the cornerstone of the bootstrap argument in our turnpike proof.

Lemma 5.6.3. *Let $T_0 > 0$ and $r > 0$ be provided by Claim 5.6.2, and let $\mathbf{y}^{\tau_1}, \mathbf{y}^{\tau_2}$ be such that*

$$\mathbf{y}^{\tau_i} \in \left\{ \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \in H_0^1(\Omega)\times L^2(\Omega) : \left\| \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} \overline{y} \\ 0 \end{bmatrix} \right\|_{H_0^1(\Omega)\times L^2(\Omega)} \leqslant r \right\}$$

for $i = 1, 2$. Let $T > 0$ and $0 \leqslant \tau_1 < \tau_2 \leqslant T$ be fixed such that $\tau_2 - \tau_1 \geqslant 2T_0$, and let $u_T \in \mathfrak{U}_{\mathrm{ad}}$ be a global minimizer to J_{τ_1,τ_2} defined in (5.6.2), with y_T denoting the associated solution to (5.6.3). Then, there exists a constant $C = C(f, T_0, \Omega, \omega) > 0$ independent of $T, \tau_1, \tau_2 > 0$ such that

$$\begin{aligned} J_{\tau_1,\tau_2}(u_T) + \|y_T(t) - \overline{y}\|^2_{H^1_0(\Omega)} + \|\partial_t y_T(t)\|^2_{L^2(\Omega)} \\ \leqslant C\left(\|y_1^{\tau_1} - \overline{y}\|^2_{H^1_0(\Omega)} + \|y_2^{\tau_1}\|^2_{L^2(\Omega)} + \|y_1^{\tau_2} - \overline{y}\|^2_{L^2(\Omega)} + \|y_2^{\tau_2}\|^2_{L^2(\Omega)}\right) \end{aligned}$$

holds for all $t \in [\tau_1, \tau_2]$.

Proof of Lemma 5.6.3. The proof follows the lines of that of Lemma 5.5.2, with some slight technical differences. We provide details for the sake of completeness. For notational purposes, it will be significantly simpler to operate in the canonical first order system framework presented in the proof of Lemma 5.4.4. For the same reason, we will also drop the subscripts of T.

We set $X := H^1_0(\Omega) \times L^2(\Omega)$, and we denote

$$\mathbf{y} := \begin{bmatrix} y \\ \partial_t y \end{bmatrix}, \qquad \overline{\mathbf{y}} := \begin{bmatrix} \overline{y} \\ 0 \end{bmatrix}.$$

We also recall the definition of the skew-adjoint operator

$$A := \begin{bmatrix} 0 & \mathrm{Id} \\ \Delta & 0 \end{bmatrix}, \qquad D(A) = D(\Delta) \times H^1_0(\Omega),$$

where $D(\Delta) = H^2(\Omega) \cap H^1_0(\Omega)$. Then the desired estimate simply writes as

$$J_{\tau_1,\tau_2}(u) + \|\mathbf{y}(t) - \overline{\mathbf{y}}\|^2_X \leqslant C\left(\|\mathbf{y}^{\tau_1} - \overline{\mathbf{y}}\|^2_X + \|\mathbf{y}^{\tau_2} - \overline{\mathbf{y}}\|^2_X\right)$$

for all $t \in [\tau_1, \tau_2]$. We proceed similarly as in the proof of Lemma 5.5.2. Using Claim 5.6.2, we know the following.

- There exists a control $u^\dagger \in L^2((\tau_1, \tau_1 + T_0) \times \omega)$ satisfying

$$\left\|u^\dagger\right\|^2_{L^2((\tau_1,\tau_1+T_0)\times\omega)} \leqslant C_0 \left\|\mathbf{y}^{\tau_1} - \overline{\mathbf{y}}\right\|^2_X, \tag{5.6.4}$$

  for some $C_0 = C_0(T_0, \omega, f) > 0$, and such that the corresponding solution $\mathbf{y}^\dagger = \begin{bmatrix} y^\dagger \\ \partial_t y^\dagger \end{bmatrix}$ to

$$\begin{cases} \partial_t \mathbf{y}^\dagger - A\mathbf{y}^\dagger + \begin{bmatrix} 0 \\ f(y^\dagger) \end{bmatrix} = \begin{bmatrix} 0 \\ u^\dagger \mathbf{1}_\omega \end{bmatrix} & \text{in } (\tau_1, \tau_1 + T_0) \\ \mathbf{y}^\dagger|_{t=\tau_1} = \mathbf{y}^{\tau_1} \end{cases}$$

  satisfies $\mathbf{y}^\dagger(\tau_1 + T_0) = \overline{\mathbf{y}}$ in X. By writing the Duhamel formula for $\mathbf{y}^\dagger - \overline{\mathbf{y}}$, and using the conservative character of $\{e^{tA}\}_{t>0}$ in X, Cauchy-Schwarz, the Lipschitz character of f and the Poincaré inequality, we see that

$$\begin{aligned} \left\|\mathbf{y}^\dagger(t) - \overline{\mathbf{y}}\right\|_X &\leqslant \left\|e^{tA}(\mathbf{y}^{\tau_1} - \overline{\mathbf{y}})\right\|_X + \int_{\tau_1}^t \left\|e^{(t-s)A}\begin{bmatrix} 0 \\ u^\dagger(s)\mathbf{1}_\omega \end{bmatrix}\right\|_X \mathrm{d}s \\ &\quad + \int_{\tau_1}^t \left\|e^{(t-s)A}\begin{bmatrix} 0 \\ (f(y^\dagger) - f(\overline{y})) \end{bmatrix}\right\|_X \mathrm{d}s \\ &\leqslant \left\|\mathbf{y}^{\tau_1} - \overline{\mathbf{y}}\right\|_X + \sqrt{T_0}\left\|u^\dagger\right\|_{L^2((\tau_1,\tau_1+T_0)\times\omega)} \\ &\quad + C(f,\Omega)\int_{\tau_1}^t \left\|\mathbf{y}^\dagger(s) - \overline{\mathbf{y}}\right\|_X \mathrm{d}s, \end{aligned} \tag{5.6.5}$$

with $C(f,\ \Omega) > 0$ depending solely on the Poincaré constant and the Lipschitz constant of f. Applying Grönwall's inequality to (5.6.5) and using (5.6.4), we deduce that

$$\left\|\mathbf{y}^\dagger(t) - \overline{\mathbf{y}}\right\|_X \leqslant C_1 \exp\left(C(f,\Omega)T_0\right) \left\|\mathbf{y}^{\tau_1} - \overline{\mathbf{y}}\right\|_X \tag{5.6.6}$$

holds for some $C_1(f, T_0, \omega) > 0$ independent of $T, \tau_1, \tau_2 > 0$, and for every $t \in (\tau_1, \tau_1 + T_0)$.

- There exists a control $u^\ddagger \in L^2((\tau_1, \tau_1 + T_0) \times \omega)$ satisfying

$$\left\|u^\ddagger\right\|^2_{L^2((\tau_1,\tau_1+T_0)\times\omega)} \leqslant C_0 \left\|\overline{\mathbf{y}} - \mathbf{y}^{\tau_2}\right\|^2_X, \tag{5.6.7}$$

and which is such that the corresponding solution $\mathbf{y}^\ddagger = \begin{bmatrix} y^\ddagger \\ \partial_t y^\ddagger \end{bmatrix}$ to

$$\begin{cases} \partial_t \mathbf{y}^\ddagger - A\mathbf{y}^\ddagger + \begin{bmatrix} 0 \\ f(y^\ddagger) \end{bmatrix} = \begin{bmatrix} 0 \\ u^\ddagger \mathbf{1}_\omega \end{bmatrix} & \text{in } (\tau_1, \tau_1 + T_0) \\ \mathbf{y}^\ddagger|_{t=\tau_1} = \overline{\mathbf{y}} \end{cases}$$

satisfies $\mathbf{y}^\ddagger(\tau_1 + T_0) = \mathbf{y}^{\tau_2}$ in X. Arguing just as above, we see that

$$\begin{aligned} \left\|\mathbf{y}^\ddagger(t) - \overline{\mathbf{y}}\right\|_X &\leqslant \int_{\tau_1}^t \left\| e^{(t-s)A} \begin{bmatrix} 0 \\ u^\ddagger(s)\mathbf{1}_\omega \end{bmatrix} \right\|_X \mathrm{d}s + \int_{\tau_1}^t \left\| e^{(t-s)A} \begin{bmatrix} 0 \\ \left(f(y^\ddagger) - f(\overline{y})\right) \end{bmatrix} \right\|_X \mathrm{d}s \\ &\leqslant \sqrt{T_0} \left\|u^\ddagger\right\|_{L^2((\tau_1,\tau_1+T_0)\times\omega)} + C(f,\Omega) \int_{\tau_1}^t \left\|\mathbf{y}^\ddagger(s) - \overline{\mathbf{y}}\right\|_X \mathrm{d}s, \end{aligned} \tag{5.6.8}$$

with $C(f,\ \Omega) > 0$ depending solely on the Poincaré constant and the Lipschitz constant of f. Applying Grönwall's inequality to (5.6.8) and using (5.6.7), we deduce that

$$\left\|\mathbf{y}^\ddagger(t) - \overline{\mathbf{y}}\right\|_X \leqslant C_2 \exp\left(C(f,\Omega)T_0\right) \left\|\mathbf{y}^{\tau_2} - \overline{\mathbf{y}}\right\|_X \tag{5.6.9}$$

holds for some $C_2(f, T_0, \omega) > 0$ independent of $T, \tau_1, \tau_2 > 0$, and for every $t \in (\tau_1, \tau_1 + T_0)$.

Now set

$$u^{\mathrm{aux}}(t) := \begin{cases} u^\dagger(t) & \text{in } (\tau_1, \tau_1 + T_0) \\ 0 & \text{in } (\tau_1 + T_0, \tau_2 - T_0) \\ u^\ddagger\left(t - (\tau_2 - \tau_1 - T_0)\right) & \text{in } (\tau_2 - T_0, \tau_2), \end{cases}$$

and let $\mathbf{y}^{\mathrm{aux}} = \begin{bmatrix} y^{\mathrm{aux}} \\ \partial_t y^{\mathrm{aux}} \end{bmatrix}$ be the corresponding solution to (5.6.3). By construction, we have

$$\mathbf{y}^{\mathrm{aux}}(t) = \mathbf{y}^\dagger(t) \qquad \text{in } [\tau_1, \tau_1 + T_0],$$

and thus

$$\mathbf{y}^{\mathrm{aux}}(t) = \overline{\mathbf{y}} \qquad \text{in } [\tau_1 + T_0, \tau_2 - T_0], \tag{5.6.10}$$

whereas we also have $\mathbf{y}^{\mathrm{aux}}(\tau_2) = \mathbf{y}^{\tau_2}$, whence $u^{\mathrm{aux}} \in \mathfrak{U}_{\mathrm{ad}}$.

We now evaluate $J_{\tau_1,\ \tau_2}$ at u^{aux}, which by virtue of a simple change of variable as well as (5.6.10), (5.6.4), (5.6.6), (5.6.7) and (5.6.9), leads us to

$$\begin{aligned} J_{\tau_1,\tau_2}(u^{\mathrm{aux}}) &= \left\|u^\dagger\right\|_{L^2((\tau_1,\tau_1+T_0)\times\omega)} + \left\|u^\ddagger\right\|_{L^2((\tau_1,\tau_1+T_0)\times\omega)} \\ &\quad + \int_{\tau_1}^{\tau_1+T_0} \left\|\mathbf{y}^\dagger(t) - \overline{\mathbf{y}}\right\|^2_X \mathrm{d}t + \int_{\tau_1}^{\tau_1+T_0} \left\|\mathbf{y}^\ddagger(t) - \overline{\mathbf{y}}\right\|^2_X \mathrm{d}t \\ &\leqslant C_3 \Big(\left\|\overline{\mathbf{y}} - \mathbf{y}^{\tau_1}\right\|^2_X + \left\|\overline{\mathbf{y}} - \mathbf{y}^{\tau_2}\right\|^2_X \Big) \end{aligned} \tag{5.6.11}$$

where $C_3(f, T_0, \Omega, \omega) > 0$ is independent of $T, \tau_1, \tau_2 > 0$. By virtue of the optimality of u and (5.6.11), we have

$$J_{\tau_1,\tau_2}(u) \leqslant J_{\tau_1,\tau_2}(u^{\mathrm{aux}}) \leqslant C_3\Big(\|\overline{\mathbf{y}} - \mathbf{y}^{\tau_1}\|_X^2 + \|\overline{\mathbf{y}} - \mathbf{y}^{\tau_2}\|_X^2 \Big).$$

An application of Lemma 5.4.4 suffices to conclude. □

5.7 Proof of Theorem 5.3

For the semilinear heat equation, we can adapt the proof strategy of Theorem 5.1 to directly prove the stabilization result stipulated by Theorem 5.3. We provide details of the proof, as it is not an immediate application of that of Theorem 5.1.

We recall that since $f \in \mathrm{Lip}(\mathbb{R})$, as presented in [219, Lemma 8.3] (and the references therein), given any $T > 0$, $y^0 \in L^2(\Omega)$ and $\overline{y} \in H_0^1(\Omega)$ solution to (5.3.3), there exists a control $u \in L^2((0,T)\times\omega)$ such that the unique solution y to (5.3.5) satisfies $y(T) = \overline{y}$, and

$$\|u\|_{L^2(\Omega)} \leqslant C(T, \omega, f) \left\|y^0 - \overline{y}\right\|_{L^2(\Omega)} \tag{5.7.1}$$

for some $C(T, \omega, f) > 0$ (the dependence on f is through the Lipschitz constant which is an upper bound for the potential appearing in the associated linear problem). Indeed, we may consider $z := y - \overline{y}$, and the control u steering z to 0 in time T is the same as that steering y to $\overline{y}$ in time T. But then, $\|u\|_{L^2(\Omega)} \leqslant C(T, \omega, f)\|z(0)\|_{L^2(\Omega)}$ from the linear system and a fixed-point argument.

Suppose $y^{\tau_1} \in L^2(\Omega)$ is given. Let $T > 0$ and $0 \leqslant \tau_1 < T$ be fixed. Consider

$$J_{\tau_1,T}(u) := \int_{\tau_1}^T \|y(t) - \overline{y}\|_{L^2(\Omega)}^2 \,\mathrm{d}t + \int_{\tau_1}^T \|u(t)\|_{L^2(\omega)}^2 \,\mathrm{d}t, \tag{5.7.2}$$

where y solves

$$\begin{cases} \partial_t y - \Delta y + f(y) = u\mathbf{1}_\omega & \text{in } (\tau_1, T)\times\Omega \\ y = 0 & \text{on } (\tau_1, T)\times\partial\Omega \\ y|_{t=\tau_1} = y^{\tau_1} & \text{in } \Omega. \end{cases} \tag{5.7.3}$$

We will only need the following lemma, which is similar to Lemma 5.6.3. In fact, the blueprint of the proof below is contained therein.

Lemma 5.7.1. *Suppose $y^{\tau_1} \in L^2(\Omega)$ is given. Let $T > 0$ and τ_1 be given such that $T > \tau_1$. Let $u_T \in L^2((\tau_1, T)\times\omega)$ be any global minimizer to $J_{\tau_1,T}$ defined in (5.7.2), with y_T denoting the corresponding solution to (5.7.3). Then, there exists a constant $C = C(f, \overline{y}, \omega) > 0$ independent of $T, \tau_1 > 0$ and y^{τ_1} such that*

$$J_{\tau_1,T}(u_T) + \|y_T(t) - \overline{y}\|_{L^2(\Omega)}^2 \leqslant C\,\|y^{\tau_1} - \overline{y}\|_{L^2(\Omega)}^2$$

holds for all $t \in [\tau_1, T]$.

Proof of Lemma 5.7.1. Fix an arbitrary $T_0 > 0$.

Let us first suppose that $T \geqslant \tau_1 + T_0$. By controllability to the steady state $\overline{y}$ (see the discussion around (5.7.1)), we know that exists a control $u^\dagger \in L^2((\tau_1, \tau_1 + T_0)\times\omega)$ satisfying

$$\left\|u^\dagger\right\|_{L^2((\tau_1,\tau_1+T_0)\times\omega)} \leqslant C_1\,\|y^{\tau_1} - \overline{y}\|_{L^2(\Omega)} \tag{5.7.4}$$

for some $C_1 = C_1(T_0, \omega, f) > 0$ and such that the corresponding solution $y^\dagger$ to

$$\begin{cases} \partial_t y^\dagger - \Delta y^\dagger + f(y^\dagger) = u^\dagger\mathbf{1}_\omega & \text{in } (\tau_1, \tau_1 + T_0)\times\Omega \\ y^\dagger = 0 & \text{on } (\tau_1, \tau_1 + T_0)\times\partial\Omega \\ y^\dagger|_{t=0} = y^0 & \text{in } \Omega. \end{cases}$$

satisfies $y^\dagger(\tau_1+T_0)=\overline{y}$ a.e. in Ω. Arguing as in the proof of Lemma 5.4.3, we see that

$$\begin{aligned}\left\|y^\dagger(t)-\overline{y}\right\|_{L^2(\Omega)} \leqslant \|y^{\tau_1}-\overline{y}\|_{L^2(\Omega)} + \sqrt{T_0}\left\|u^\dagger\right\|_{L^2((\tau_1,\tau_1+T_0)\times\omega)} \\ + C(f)\int_{\tau_1}^t \left\|y^\dagger(s)-\overline{y}\right\|_{L^2(\Omega)}\,\mathrm{d}s \end{aligned} \tag{5.7.5}$$

for $t\in(\tau_1,\tau_1+T_0)$, with $C(f)>0$ being the Lipschitz constant of f. Applying Grönwall's inequality to (5.7.5) and using (5.7.4), we deduce that

$$\left\|y^\dagger(t)-\overline{y}\right\|_{L^2(\Omega)} \leqslant C_2\exp\left(C(f)T_0\right)\|y^{\tau_1}-\overline{y}\|_{L^2(\Omega)} \tag{5.7.6}$$

for some $C_2(f,T_0,\omega)>0$ independent of $T,\tau_1,\tau_2>0$, and for every $t\in(\tau_1,\tau_1+T_0)$.

Now set

$$u^{\mathrm{aux}}(t):=\begin{cases} u^\dagger(t) & \text{in } (\tau_1,\tau_1+T_0)\\ 0 & \text{in } (\tau_1+T_0,T)\end{cases}$$

and let y^{aux} be the corresponding solution to (5.3.5). Clearly $y^{\mathrm{aux}}(t)=\overline{y}$ for $t\in[\tau_1+T_0,T]$, a.e. in Ω. Hence, using $J_{\tau_1,T}(u_T)\leqslant J_{\tau_1,T}(u^{\mathrm{aux}})$, (5.7.6) and (5.7.4), we see that

$$\begin{aligned} J_{\tau_1,T}(u_T) &\leqslant \left\|y^\dagger-\overline{y}\right\|^2_{L^2((\tau_1,\tau_1+T_0)\times\Omega)} + \left\|u^\dagger\right\|^2_{L^2((\tau_1,\tau_1+T_0)\times\omega)}\\ &\leqslant C_3\|y^{\tau_1}-\overline{y}\|^2_{L^2(\Omega)}\end{aligned}$$

for some $C_3(f,T_0,\omega)>0$ independent of $T,\tau_1>0$. Applying Lemma 5.4.3 suffices to conclude.

Now suppose that $\tau_1<T<T_0+\tau_1$. We may then use the optimality inequality $J_{\tau_1,T}(u_T)\leqslant J_{\tau_1,T}(u_{T_0+\tau_1})$, and since by the previous step, we know that

$$J_{\tau_1,T}(u_{T_0+\tau_1})\leqslant J_{\tau_1,T_0+\tau_1}(u_{T_0+\tau_1})\leqslant C_3\ \|y^{\tau_1}-\overline{y}\|^2_{L^2(\Omega)}$$

where $C_3=C_3(f,T_0,\omega)>0$ is independent of $T,\tau_1>0$, we deduce

$$J_{\tau_1,T}(u_T)\ \leqslant C_3\|y^{\tau_1}-\overline{y}\|^2_{L^2(\Omega)}\,. \tag{5.7.7}$$

We may conclude by combining (5.7.7) with Lemma 5.4.3. □

Proof of Theorem 5.3. The proof is of the same spirit as that of Theorem 5.1, the only difference being the fact that we only need to bootstrap "forward" in time due to the lack of final cost, which renders the proof significantly less technical. The control estimate follows from Lemma 5.7.1. We thus concentrate solely on estimating the state.

Fix

$$\tau>C_1^4$$

where $C_1=C_1(f,\overline{y},\omega)>0$ is the (square root of the) constant appearing in Lemma 5.7.1, and let $T_0>\tau$ be arbitrary and fixed[5]. Similarly to the proof of Theorem 5.1, we will have $T^*:=\tau+T_0$ in the statement. Let

$$T\geqslant\tau+T_0$$

be fixed.

First note that for $t\in[0,\tau+T_0]$, just as in Part 1 of the proof of Theorem 5.1, the desired estimate can easily be obtained for such t since the length of the time interval is independent of T. Hence, we will solely concentrate on the case $t\in[\tau+T_0,T]$. To this end, we will mimic the steps done in the proof of Theorem 5.1.

[5]Note that this choice is independent of the one done in the proof of Lemma 5.7.1.

Step 1). **Preparation.** Since $2\tau < \tau + T_0 < T$ and thus $\tau \leqslant \frac{T}{2}$, by Lemma 5.5.3 there exists a $\tau_1 \in [0, \tau)$ such that

$$\|y_T(\tau_1) - \overline{y}\|_{L^2(\Omega)} \leqslant \frac{\|y_T - \overline{y}\|_{L^2((0,T)\times\Omega)}}{\sqrt{\tau}} \leqslant \frac{C_1}{\sqrt{\tau}} \left\|y^0 - \overline{y}\right\|_{L^2(\Omega)}. \tag{5.7.8}$$

Now the control $u_T|_{[\tau_1,T]}$ minimizes $J_{\tau_1,T}$ with initial data $y^{\tau_1} = y_T(\tau_1)$ for (5.7.3), to which clearly the solution is $y_T|_{[\tau_1,T]}$. So by Lemma 5.7.1 and (6.7.3),

$$\|y_T(t) - \overline{y}\|_{L^2(\Omega)} \leqslant C_1 \|y_T(\tau_1) - \overline{y}\|_{L^2(\Omega)} \leqslant \frac{C_1^2}{\sqrt{\tau}} \left\|y^0 - \overline{y}\right\|_{L^2(\Omega)} \tag{5.7.9}$$

holds for all $t \in [\tau_1, T]$. Since $\tau_1 < \tau$, (5.7.9) also holds for all $t \in [\tau, T]$.

Step 2). **Bootstrap.** We bootstrap (5.7.9) and prove that for any $n \in \mathbb{N}$ satisfying

$$n \leqslant \frac{T}{2\tau},$$

the estimate

$$\|y_T(t) - \overline{y}\|_{L^2(\Omega)} \leqslant \left(\frac{C_1^2}{\sqrt{\tau}}\right)^n \left\|y^0 - \overline{y}\right\|_{L^2(\Omega)} \tag{5.7.10}$$

holds for all $t \in [n\tau, T]$. We proceed by induction. The case $n = 1$ holds by (5.7.9). Thus assume that (5.7.10) holds at some stage $n \in \mathbb{N}$ and suppose that

$$n + 1 \leqslant \frac{T}{2\tau}.$$

This clearly implies that

$$\tau \leqslant \frac{T - 2n\tau}{2}. \tag{5.7.11}$$

Now the control $u_T|_{[n\tau,T]}$ is a global minimizer of $J_{n\tau,T}$. We can thus apply Lemma 5.7.1 with $\tau_1 = n\tau$, and Lemma 5.5.3 (note (5.7.11)) on $[n\tau, T - n\tau]$, to deduce that there exists $t_1 \in [n\tau, (n+1)\tau)$ such that

$$\|y_T(t_1) - \overline{y}\|_{L^2(\Omega)} \leqslant \frac{\|y_T - \overline{y}\|_{L^2((n\tau,T)\times\Omega)}}{\sqrt{\tau}} \leqslant \frac{C_1}{\sqrt{\tau}} \|y_T(n\tau) - \overline{y}\|_{L^2(\Omega)}.$$

So now we apply the induction hypothesis (5.7.10) to deduce

$$\|y_T(t_1) - \overline{y}\|_{L^2(\Omega)} \leqslant \frac{C_1}{\sqrt{\tau}} \left(\frac{C_1^2}{\sqrt{\tau}}\right)^n \left\|y^0 - \overline{y}\right\|_{L^2(\Omega)}. \tag{5.7.12}$$

Since $u_T|_{[t_1,T]}$ is a global minimizer of $J_{t_1,T}$, we can apply Lemma 5.7.1 and use (5.7.12) to deduce that

$$\|y_T(t) - \overline{y}\|_{L^2(\Omega)} \leqslant C_1 \|y_T(t_1) - \overline{y}\|_{L^2(\Omega)} \leqslant \frac{C_1^2}{\sqrt{\tau}} \left(\frac{C_1^2}{\sqrt{\tau}}\right)^n \left\|y^0 - \overline{y}\right\|_{L^2(\Omega)} \tag{5.7.13}$$

holds for all $t \in [t_1, T]$. Clearly, as $t_1 < (n+1)\tau$, (5.7.13) also holds for all $t \in [(n+1)\tau, T]$. This concludes the induction proof, and so (5.7.10) does indeed hold.

Step 3). **Conclusion.** We now use (5.7.10) to conclude the proof. Suppose $t \in [\tau + T_0, T]$ is arbitrary and fixed. Set $n(t) := \left\lfloor \frac{t}{\tau + T_0} \right\rfloor$. Clearly $n(t) \geqslant 1$, $t \geqslant n(t)\tau$ and $n(t) \leqslant \frac{T}{2\tau}$ due to the choice of T_0. We may then apply (5.7.10) to find that

$$\|y_T(t) - \overline{y}\|_{L^2(\Omega)} \leqslant \left(\frac{C_1^2}{\sqrt{\tau}}\right)^{n(t)} \left\|y^0 - \overline{y}\right\|_{L^2(\Omega)} \tag{5.7.14}$$

Now since $\tau > C_1^4$ and $n(t) \geqslant \frac{t}{\tau+T_0} - 1$, we can see from (5.7.14) that

$$\begin{aligned}\|y_T(t) - \overline{y}\|_{L^2(\Omega)} &\leqslant \exp\left(-n(t)\log\left(\frac{\sqrt{\tau}}{C_1^2}\right)\right)\|y^0 - \overline{y}\|_{L^2(\Omega)} \\ &\leqslant \frac{\sqrt{\tau}}{C_1^2}\exp\left(-\frac{\log\left(\frac{\sqrt{\tau}}{C_1^2}\right)}{\tau+T_0}t\right)\|y^0 - \overline{y}\|_{L^2(\Omega)}\end{aligned}$$

The desired estimate thus holds for all $t \in [\tau + T_0, T]$, with

$$\mu := \frac{\log\left(\frac{\sqrt{\tau}}{C_1^2}\right)}{\tau+T_0} > 0$$

and

$$C := \frac{\sqrt{\tau}}{C_1^2}\|y^0 - \overline{y}\|_{L^2(\Omega)}.$$

This concludes the proof. □

5.8 Concluding remarks

We have presented a new methodology for proving the turnpike property for nonlinear optimal control problems set in large time horizons, under the assumption that the running target is a steady control-state pair, and that the system is controllable with a local estimate on the cost. These assumptions allow us to bypass necessary optimality conditions and a study of the adjoint system, and rather relies on calculus of variations–based arguments.

More precisely, we have concluded that

(1). The exponential turnpike property holds for optimal state trajectories of optimal control problems for nonlinear finite and infinite-dimensional dynamics, whenever the cost functional is coercive with respect to the distance of the state to the target steady state. The nonlinearity may be assumed to be only globally Lipschitz continuous (and thus possibly nonsmooth). The result holds without any smallness assumptions on the initial data.

(2). The last exponential arc (near $t = T$) can be removed whenever the optimal control problem is considered without a final time cost, and thus entails an exponential stabilization estimate for the optimal state trajectory.

Outlook. Let us conclude with a list of select problems related to our study.

- **Necessity of assuming that $\overline{y}$ is a steady state.** The assumption that the running target $\overline{y}$ in (5.2.3) is a steady state of the dynamics allows us to easily obtain quasi-turnpike strategies allowing us to obtain the key estimates in Lemma 5.5.1 and Lemma 5.5.2 (resp. Lemma 5.6.1, Lemma 5.6.3, Lemma 5.7.1 in the PDE setting). The case of controlled steady states $\overline{y}$ associated to a presecribed control $\overline{u}$ can readily be addressed by penalizing $u - \overline{u}$ over $[0, T]$ instead of solely u as noted in Remark 5.2.1. But we were unable to see if this is a necessary assumption in the nonlinear context in the absence of smallness conditions on the target, and whether the controlled steady state case can be covered by solely penalizing u. These questions merit in-depth investigation.

- **Weakening Definition 6.4.1.** An important hypothesis we made throughout is Definition 6.4.1, which required that, at least for data y^0, y^1 in the vicinity of the free steady state $\overline{y}$, the minimal L^2–norm control steering the system from y^0 to $\overline{y}$ may be estimated by $\|y^0 - \overline{y}\|$, and similarly for that from $\overline{y}$ to y^1. This is a hallmark of linear control systems, which is also expected for nonlinear systems for which controllability results are obtained by linearization or perturbation methods and a fixed-point argument. But in the general context of control-affine systems, such an assumption may appear restrictive, eventhough it is local. It is thus of interest to see how the results and methodology can be pertained whilst weakening Definition 6.4.1.

  In fact, more generally, it would be of interest to investigate whether the methodology presented herein can still be applied by only assuming approximate controllability with an adequate estimate on the control cost.

- **Turnpike with state or control constraints.** A problem which has not been extensively covered in the literature is the turnpike property with positivity (or box) constraints on either the state or the control. Slightly weaker integral turnpike results under such constraints have been obtained in [203] by means of *quantitative inequalities*. Such a study would complement the already existent nonlinear *controllability under constraints theory* – a topic covered in several recent works, see e.g. [173, 204, 219, 236] and the references therein.

- **More general control systems.** We have considered homogeneous Dirichlet boundary conditions in (5.3.1) and (5.3.5) merely to avoid additional technical details. The proofs of Theorem 5.2 (resp. Theorem 5.3) only require that the underlying dynamics are exactly controllable (resp. controllable to a steady state), thus, the same results hold with Neumann boundary conditions or boundary controls. Similarly, variable coefficients and lower order terms may be considered, as long as these coefficients are time-independent, as we are using a Duhamel formula along with a semigroup representation of the solution.

  In fact, we have chosen the wave and heat equation for the sake of presentation, but the respective results could possibly be extended to a more general scenario of exactly controllable semilinear systems with similar assumptions, e.g. dispersive equations (Schrödinger, Korteweg-de Vries), coupled systems, and so on. The necessity of a Duhamel formula may however be an impediment to the extension of our results to the context of quasilinear systems such as the porous medium equation (see [115] and the references therein).

- **Bilinear control systems.** It would also be of interest to establish the turnpike property for bilinear control systems. This would be the somewhat true analog of the control-affine systems presented herein, and under suitable assumptions on the nonlinearity, one could expect that our methodology applies to such cases as well. We have not addressed such systems for the simplicity of presentation and due to the controllability assumptions we make, as the controllability theory for bilinear problems is not complete (albeit, see [23, 45, 87, 202] for recent developments). Notwithstanding, our results should be applicable to a system of the form (see [23])

$$\begin{cases} \partial_t y - \partial_x^2 y = u(t) f(y) & \text{in } (0,T)\times(0,\pi) \\ \partial_x y(t,0) = \partial_x y(t,\pi) = 0 & \text{in } (0,T) \\ y|_{t=0} = y^0 & \text{in } (0,\pi) \end{cases}$$

  where u is a scalar control and f is an appropriate nonlinearity (see [23] for sufficient conditions for ensuring controllability, and globally Lipschitz for applying our methodology).

- **More general nonlinearities.** Finally, it would be of interest to investigate problems where our methodology does not immediately apply, such as the paradigmatic example of the cubic heat equation. This problem consists in seeing whether one may prove Theorem 5.3 (with the estimate on u_T changed by an estimate of $u_T - \overline{u}$) for minimizers u_T of

$$J_T(u) := \int_0^T \|y(t) - \overline{y}\|^2 \, \mathrm{d}t + \int_0^T \|u - \overline{u}\|^2 \, \mathrm{d}t$$

where y_T is the unique solution to

$$\begin{cases} \partial_t y - \Delta y + y^3 = u \mathbf{1}_\omega & \text{in } (0,T) \times \Omega \\ y = 0 & \text{on } (0,T) \times \partial\Omega \\ y|_{t=0} = y^0 & \text{in } \Omega, \end{cases} \tag{5.8.1}$$

and $\overline{y} \in H_0^1(\Omega)$ is a controlled steady state associated to some $\overline{u} \in L^2(\omega)$ (the case $\overline{u} \equiv 0$ is somewhat trivial due to the inherent stabilization to $\overline{y} \equiv 0$). Let us elaborate on a possible technical impediment in the direct application of our strategy. Clearly, for Theorem 5.3 to hold in this case, it would suffice to prove Lemma 5.7.1 for $f(s) = s^3$ (while replacing the estimate of u_T by an estimate of $u_T - \overline{u}$). To this end, first of all, for any $u \in L^2((0,T) \times \omega)$, using the variational formulation and standard arguments including Cauchy-Schwarz, Young with ϵ and Poincaré inequalities, one can find

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega |y(t,x)|^2 \, \mathrm{d}x \leqslant \epsilon \int_\omega \|u(t,x)\|^2 \, \mathrm{d}x$$

for a.e. $t \in [0,T]$, where $\epsilon > \frac{C(\Omega)}{4}$, whereas y solves (5.8.1), and thus

$$\|y\|_{C^0([0,T];L^2(\Omega))} \leqslant C_1(\Omega) \left(\|u\|_{L^2((0,T)\times\omega)} + \left\|y^0\right\|_{L^2(\Omega)} \right). \tag{5.8.2}$$

Following the proof of Lemma 5.4.3 for $f(s) = s^3$ and using (5.8.2), we may find

$$\|y(t) - \overline{y}\|_{L^2(\Omega)} \leqslant C \left(\left\|y^0 - \overline{y}\right\|_{L^2(\Omega)} + \|u - \overline{u}\|_{L^2((0,T)\times\omega)} + \|y - \overline{y}\|_{L^2((0,T)\times\Omega)} \right),$$

where now

$$C \sim \exp\left(\|u\|_{L^2((0,T)\times\omega)} \right). \tag{5.8.3}$$

It is precisely at this point where the issue appears, since simply by using the form of the functional, we are not in a position to prove that $\|u\|_{L^2((0,T)\times\omega)}$ is uniformly bounded with respect to T, but rather only $\|u - \overline{u}\|_{L^2((0,T)\times\omega)}$. Should this be possible, then one can expect our methodology to apply to the cubic heat equation as well, but as things stand, turnpike without smallness conditions in this case remains open.

Further examples worth analyzing include the heat equation with a convective nonlinearity $f(y, \nabla y)$, even in one space dimension (e.g. the Burgers equation); along these lines we refer to [275] for a local turnpike result for the 2d Navier-Stokes system. Similar questions can be asked for the semilinear wave equation, where the nonlinearity is sometimes only assumed to be superlinear (see [168] for a subcritical optimal control study) – our methodology a priori applies if the nonlinearity is either truncated by some cut-off, or if one manages to prove uniform estimates of $\|y_T\|_{L^\infty((0,T)\times\Omega)}$ with respect to T. Further nonlinear problems which could be investigated include hyperbolic systems (see [125] for a related study) or free boundary problems (see [116] for a control perspective).

Part III

Interplay of deep learning and control

Chapter 6

Large-time asymptotics in deep learning

Abstract. We study the behavior of supervised learning problems for neural ODEs when the final time horizon T is increased, a fact that may be interpreted as increasing the depth in the associated residual neural network (ResNet) setting.

For the classical L^2 (or Sobolev)–regularized empirical risk minimization problem, under homogeneity assumptions on the neural ODE flow, we prove that when T goes to infinity, the training error decays to zero with an (almost) polynomial, $\mathcal{O}\left(\frac{1}{T}\right)$–rate. In the context of regression tasks, the optimal parameters are also shown converge to minimal L^2 (resp. Sobolev)–norm parameters which interpolate the dataset. Moreover, motivated by the fact that the L^2–regularization context, a natural scaling between the time horizon T and the regularization parameter λ appears, using similar arguments, we obtain the same convergence results when λ goes to zero and the horizon is fixed. These results thus allow us to stipulate generalization properties in the overparametrized regime – now seen from the large depth and neural ODE perspective –, and are aligned with results on regularization path convergence (i.e. λ to zero) and implicit regularization of gradient descent for linear models or two-layer perceptrons.

To enhance the polynomial decay rates of the training error, we propose an augmented learning problem by adding an artificial regularization term of the state trajectory over the entire time horizon. In the context of training error for squared ℓ^2–loss, we obtain an exponential, $\mathcal{O}\left(e^{-\mu t}\right)$–rate of decay for the training error and for the optimal parameters in any time $t \in [0, T]$ – an improved estimate for the depth required to reach optimal training accuracy. This result is a particular manifestation of the so-called *turnpike property*, well-known in economics and optimal control theory.

The aforementioned asymptotic regimes are also discussed in the context of continuous space-time neural networks taking the form of nonlinear integro-differential equations, which provide a framework for addressing ResNets with variable widths.

Keywords. Deep learning, ResNets, neural ODEs, regularization path, optimal control, generalization, exponential decay, turnpike theory.

AMS Subject Classification. 49J15; 49M15; 49J20; 49K20; 93C20; 49N05.

This Chapter is taken from [95]:

> *Large-time behavior in deep supervised learning.*
> C. Esteve, B. Geshkovski, D. Pighin and E. Zuazua, 2020.
> https://arxiv.org/abs/2008.02491

Chapter Contents

6.1 Introduction

Modern supervised learning addresses the problem of predicting from data, which roughly consists in approximating an unknown function $f(\cdot)$ from N known but possibly noisy samples $\{\vec{x}_i, \vec{y}_i = f(\vec{x}_i)\}_{i=1}^N$. Depending on the nature of the labels $\vec{y}_i$, one distinguishes two types of supervised learning tasks, namely that of *classification* (labels take values in a finite set of m classes, e.g. $\{1, \dots, m\}$) and *regression* (labels take continuous values in $\mathbb{R}^m$). In many applications, the dimension d of each sample $\vec{x}_i$ may be big compared to the number/dimension m of the labels – in image classification for instance, a sample of the ImageNet dataset [167], which has $m = 1000$ classes, is an element of $\mathbb{R}^{65536}$.

A plethora of methods for finding $f(\cdot)$ efficiently with theoretical and empirical guarantees have been developed and investigated in the machine learning literature in recent decades. Prominent examples, to name a few, include linear parametric methods (e.g. linear or logistic regression), kernel-based methods (e.g. support vector machines), tree-based methods (e.g. decision trees) and so on. We refer to [119] for a comprehensive presentation of these topics.

Deep neural networks are parametrized computational architectures which propagate each individual sample of the input data $\{\vec{x}_i\}_{i=1}^N \in \mathbb{R}^{d\times N}$ across a sequence of linear parametric operators and simple nonlinearities. The so-called *residual neural networks*

(ResNets, [140]) may, in the simplest case, be cast as schemes of the mould

$$\begin{cases} \mathbf{x}_i^{k+1} = \mathbf{x}_i^k + \sigma(w^k \mathbf{x}_i^k + b^k) & \text{for } k \in \{0, \dots, N_{\text{layers}} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i \in \mathbb{R}^d \end{cases} \tag{6.1.1}$$

for all $i \in [N]$, where we set $[N] := \{1, \dots, N\}$. The unknowns are the states $\mathbf{x}_i^k \in \mathbb{R}^d$ for any $i \in [N]$, σ is an explicit scalar, Lipschitz continuous nonlinear function defined componentwise in (7.1.1), $\{w^k, b^k\}_{k=0}^{N_{\text{layers}}-1}$ are optimizable parameters (controls) with $w^k \in \mathbb{R}^{d \times d}$ – called weights, and $b^k \in \mathbb{R}^d$ – called biases, and $N_{\text{layers}} \geqslant 1$ designates the number of layers referred to as the depth. The training process consists in finding optimal parameters steering all of the network outputs $\mathbf{x}_i^{N_{\text{layers}}}$ as close as possible to the corresponding labels $\vec{y}_i$ by solving

$$\min_{\{w^k, b^k\}_{k=0}^{N_{\text{layers}}-1}} \frac{1}{N} \sum_{i=1}^{N} \text{loss}\left(P \mathbf{x}_i^{N_{\text{layers}}}, \vec{y}_i\right),$$

whilst guaranteeing reliable performance on unseen data (ensuring *generalization*). Here $\text{loss}(\cdot, \cdot)$ is a given continuous and nonnegative function which may change depending on the task in hand – for instance $\text{loss}(x, y) := \|x - y\|_{\ell^p}^p$ for $p = 1, 2$ is commonly used for regression tasks, while $\text{loss}(x, y) = \log(1 + \exp(-yx))$ may be used for binary classification, namely when $\vec{y}_i \in \{-1, +1\}$ (we refer to (6.2.8) for more general settings). On the other hand, $P : \mathbb{R}^d \to \mathbb{R}^m$ is an affine map which in practice is either part of the optimizable parameters or may be chosen at random. In our work, we shall assume that P is given and specified on a case-by-case basis.

Due to the inherent dynamical systems nature of ResNets, several recent works have aimed at studying an associated continuous-time formulation in some detail, a trend started with the works [89, 129]. This perspective is motivated by the simple observation that for any $i \in [N]$ and for $T > 0$, (7.1.1) is roughly the forward Euler scheme for the neural ordinary differential equation (neural ODE)

$$\begin{cases} \dot{\mathbf{x}}_i(t) = \sigma(w(t)\mathbf{x}_i(t) + b(t)) & \text{for } t \in (0, T) \\ \mathbf{x}_i(0) = \vec{x}_i \in \mathbb{R}^d. \end{cases} \tag{6.1.2}$$

The continuous-time formulation has been used to great effect for improving computational training performance – for instance, by using adaptive ODE solvers [61, 86] or indirect training algorithms based on the Pontryagin Maximum Principle [181, 25] –, and also for modeling purposes, including irregular time series modeling [234], and generative modeling through normalizing flows [120, 60]. It should be noted that the origins of continuous-time supervised learning go back to the 1980s – the neural network model proposed in [141] is a differential equation, whereas in [179] back-propagation is connected to the adjoint method arising in optimal control. Related works include studies on identification of the weights from data [6, 7] and controllability of continuous-time recurrent networks [247, 248].

The role of the final time horizon $T > 0$ however, which plays a key role in the control of dynamical systems, has not been discussed in the context of supervised learning problems via models such as (7.1.2). As each time-step of a discretization to (7.1.2) represents a different layer of the derived neural network (e.g. (7.1.1)), the time horizon $T > 0$ in (7.1.2) may serve as an indicator of the number of layers N_{layers} in the discrete-time context (7.1.1). Thus, a good a priori knowledge of the dynamics of the learning problem over longer time horizons is needed. Such an understanding would lead to potential rules for choosing the number of layers, and enlighten the possible generalization properties when the number of layers is large. In this work, we aim to bridge this gap by proposing several insights and an analysis of the role of the time horizon T.

6.1.1 Our contributions.

We shall focus this presentation on the neural ODE (7.1.2), but our results also hold for other systems, as seen in subsequent discussions.

1. We first consider the classical supervised learning problem, namely that of regularized empirical risk minimization[1]:

$$\inf_{\substack{[w,b]\in H^k(0,T;\mathbb{R}^{d_u}) \\ \mathbf{x}_i \text{ solves (7.1.2)}}} \underbrace{\frac{1}{N}\sum_{i=1}^{N} \mathrm{loss}\big(P\mathbf{x}_i(T),\, \vec{y}_i\big)}_{\text{training error}:=\mathcal{E}(\mathbf{x}(T))} + \underbrace{\lambda \Big\| [w,b] \Big\|^2_{H^k(0,T;\mathbb{R}^{d_u})}}_{\text{regularization}} \tag{6.1.3}$$

with $k = 0, 1$.

- When $\mathrm{loss}(Px, y) = 0$ at $Px = y$ (typical in the context of regression tasks where $\mathrm{loss}(Px, y) = \|Px - y\|^2_{\ell^2}$ and P is affine) and for σ 1–homogeneous, we show (see Theorem 6.1) that the training error $\mathcal{E}(\mathbf{x}_T(T))$ of the vector $\mathbf{x}_T = [\mathbf{x}_{T,1}, \dots, \mathbf{x}_{T,N}]^\top$ of solutions to (7.1.2) corresponding to any solution $[w_T, b_T]$ to the minimization problem (6.1.3), decays to 0 as $\mathcal{O}\left(\frac{1}{T}\right)$, whilst the optimal parameters $[w_T, b_T]$ converge, on a suitable time-scale, to a solution $[w^*, b^*]$ of

$$\inf_{\substack{[w,b]\in H^k(0,1;\mathbb{R}^{d_u}) \\ \mathbf{x}_i \text{ solves (7.1.2) in}[0,1] \\ \text{and} \\ P\mathbf{x}_i(1) = \vec{y}_i}} \Big\| [w,b] \Big\|^2_{H^k(0,1;\mathbb{R}^{d_u})}$$

  when $T \longrightarrow +\infty$.

- For classification tasks, where, for example, $\vec{y}_i \in \{-1, +1\}$ and considering $\mathrm{loss}(Px, y) = \log\left(1 + e^{-yPx}\right)$ (all results hold for multi-label tasks where $\vec{y}_i \in [m]$ for $m \geqslant 2$ via cross-entopy loss), in Theorem 6.3, we show that the training error $\mathcal{E}(\mathbf{x}_T(T))$ decays to 0 as $T \longrightarrow +\infty$ like

$$\log\left(1 + e^{-\gamma e^{T^\alpha}}\right) + \mathcal{O}\left(T^{2\alpha-1}\right)$$

  for all $\alpha \in \left(0, \frac{1}{2}\right)$, under the assumption that the *margin*

$$\gamma := \min_{i\in[N]} \vec{y}_i P\, \widehat{\mathbf{x}}_i(1)$$

  is positive for some neural ODE trajectory $\widehat{\mathbf{x}}(t)$ defined for $t \in [0, 1]$.

Let us put the above results into context. For neural ODEs for which L^2–regularization suffices, we remark that $T \longrightarrow +\infty$ is equivalent to $\lambda \searrow 0$. The latter is the convergence of the regularization path, studied in the literature (see Section 6.1.2) for linear models and multi-layer perceptrons (but not for more compound models such as the ones considered here), where the asymptotic limits can be shown to satisfy desirable generalization properties.

Using similar arguments as when $T \longrightarrow +\infty$, we obtain the same conclusions when $\lambda \searrow 0$ and T is fixed (Theorem 6.2 and Theorem 6.4). Consequently, Theorem 6.1 and Theorem 6.3 also stipulate generalization properties – namely, optimizing with $T \gg 1$, which may be interpreted as a larger depth for ResNets, has the practically desirable effect of making the training error close to zero, but by means of almost optimal parameters.

[1] Here $H^k(0, T; \mathbb{R}^{d_u})$ denotes the standard Sobolev space of square integrable functions from $(0, T)$ to $\mathbb{R}^{d_u}$ with k square integrable weak derivatives. We make precise the necessity of considering Sobolev regularization, namely $k = 1$, for compactness purposes in the context of (7.1.2) in Remark 6.2.2; we use the convention $H^0 := L^2$.

2. Parallel to (6.1.3), to enhance the convergence rate of the training error $\mathcal{E}(\mathbf{x}_T(T))$ to 0 as $T \longrightarrow +\infty$, when $\mathrm{loss}(x, y) = \|x-y\|_{\ell^2}^2$ we consider an augmented empirical risk minimization problem which consists in solving

$$\inf_{\substack{[w,b]\in H^k(0,T;\mathbb{R}^{d_u}) \\ \mathbf{x}_i \text{ solves } (7.1.2)}} \mathcal{E}(\mathbf{x}(T)) + \int_0^T \|\mathbf{x}(t) - \overline{\mathbf{x}}\|^2 \, \mathrm{d}t + \lambda \Big\| [w, b] \Big\|^2_{H^k(0,T;\mathbb{R}^{d_u})}, \tag{6.1.4}$$

where $P : \mathbb{R}^d \to \mathbb{R}^m$ appearing in $\mathcal{E}$ is Lipschitz (possibly nonlinear in classification tasks) and surjective, and $\overline{\mathbf{x}}_i \in P^{-1}(\{\vec{y}_i\})$ for all $i \in [N]$ are arbitrary and given.

Under a particular simultaneous controllability assumption, but without any regularity assumptions on the activation function σ or smallness assumptions on the dataset, we show (see Theorem 6.5) that optimal parameters $[w_T(t), b_T(t)]$ and the training error $\mathcal{E}(\mathbf{x}_T(t))$ of the corresponding vector $\mathbf{x}_T(t)$ of solutions to (7.1.1) decay to 0 as $\mathcal{O}(e^{-\mu t})$ for any $t \in [0, T]$ and some $\mu > 0$.

This result is in line with Theorem 6.1 and Theorem 6.3, but with a significantly improved rate of convergence, and thus a better estimate of the time horizon needed to be ε–close to the interpolation or separation regime for any given $\varepsilon > 0$. Due to the exponentially small global minimizers, numerical experiments show that the learned flow is simple, stipulating possible generalization properties. Theorem 6.5 is a manifestation of the so-called *turnpike property*, well-known in optimal control theory ([261]).

Problem (6.1.4) is motivated by the more computationally scalable training problem

$$\inf_{\substack{[w,b]\in H^k(0,T;\mathbb{R}^{d_u}) \\ \mathbf{x}_i \text{ solves } (7.1.2)}} \int_0^T \mathcal{E}(\mathbf{x}(t)) \, \mathrm{d}t + \lambda \Big\| [w, b] \Big\|^2_{H^k(0,T;\mathbb{R}^{d_u})}, \tag{6.1.5}$$

where the $\mathrm{loss}(\cdot, \cdot)$ appearing in $\mathcal{E}$ is continuous and nonnegative, but otherwise arbitrary (thus, possibly non-coercive). Whilst left without proof, numerical experiments stipulate that a similar decay for the training error and optimal parameters, and, combined with Theorem 6.5, motivate the usage of in practice (see Section 6.4.1).

To concur on the specific proof strategy, in Theorem 6.6, we show using a constructive method that the simultaneous controllability assumption needed for Theorem 6.5 is satisfied for a subclass of neural ODEs with C^1–regular activation functions σ, under smallness conditions on the data.

3. To address variable width architectures motivated by multi-layer perceptrons and convolutional neural networks, in Section 6.5 we study a continuous space-time neural network formulation taking the form of a scalar, integro-differential equation:

$$\begin{cases} \partial_t \mathbf{x}_i(t, x) = \sigma\left(\displaystyle\int_\Omega w(t, x, \xi) \mathbf{x}_i(t, \xi) \, \mathrm{d}\xi + b(t, x) \right) & \text{for } (t, x) \in (0, T) \times \Omega \\ \mathbf{x}_i(0, x) = \mathbf{x}_i^{\mathrm{in}}(x) & \text{for } x \in \Omega \end{cases} \tag{6.1.6}$$

for any $i \in [N]$. Here $\Omega \subset \mathbb{R}^{d_\Omega}$ is a bounded domain, $d_\Omega \geqslant 1$ is chosen based on the nature[2] of the inputs $\{\vec{x}_i\}_{i=1}^N$, whereas $\mathbf{x}_i^{\mathrm{in}} \in C^0(\overline{\Omega})$ interpolates $\vec{x}_i$ for any i. By means of some simple discretization arguments, we demonstrate that (6.1.6) is general in the sense that by taking initial data as a linear combination of Dirac masses, one recovers neural ODEs such as (7.1.2), while by imposing a specific structure on the weight $w(t, x, \xi)$, it allows for deducing various forms of convolutional neural networks as well.

[2]For instance, $d_\Omega = 3$ if $\vec{x}_i \in \mathbb{R}^{d_1 \times d_2 \times d_{\mathrm{ch}}}$ in the context of image data, and $d_\Omega = 1$ for vectorized data.

In Theorem 6.7 (resp. Theorem 6.8), we show that some of our finite-dimensional conclusions from Theorem 6.1 (resp. Theorem 6.5) transfer, under similar assumptions, to the continuous space-time networks such as (6.1.6).

6.1.2 Related work.

Our results are related to several questions studied in existing literature.

Universal approximation. On a first note, the asymptotic results presented herein may (heuristically) be interpreted as approximation results in the sense of the universal approximation theory. These are density results for neural networks, and in the simplest cases can be interpreted in terms of the elementary building blocks of measure theory such as the density of simple functions in Lebesgue spaces. The first result in this direction is the seminal work [76], which indicates that shallow neural networks with increasing width, i.e. a superposition of sufficiently many dilated and translated sigmoids, may approximate any continuous function on compact sets. We also refer to [142, 220] for an extension to multi-layer neural networks. Our results are somewhat dual to [76] – therein, to increase the approximation accuracy, the width is allowed to grow, whilst we fix the width and allow the depth to increase. We do note however that we prove approximation properties for the trained parameters, and for a fixed dataset, unlike what is commonly done in universal approximation theorem, where the parameters are not known explicitly. We refer to the thesis [211] for results and a comprehensive review of universal approximation results for ResNets, and to the recent works [182] and [237], for universal approximation results for neural ODE and for illuminating observations on the latter's working mechanisms.

Regularization path limit: $\lambda \searrow 0$. The regularization path limit $\lambda \searrow 0$ has also been addressed in some machine learning literature. This was initiated in [233, 232], where the authors study linear logistic regression, and show convergence to the max-margin as $\lambda \searrow 0$, under the assumption of linearly separable data. The max-margin, support vector machine solution, ([249]) is a special example among all solutions that fit the training data – another example includes minimal ℓ^2–norm solution for linear regression –; both these solutions can be shown to ensure generalization by virtue of explicit generalization error estimates [18, 151]. This insight stipulates a likely generalization capacity of our asymptotic limits as $T \longrightarrow +\infty$ or $\lambda \searrow 0$.

The results of [233, 232] have subsequently been extended in [269] (and some of the references therein) to multi-layer perceptrons with ReLU activations, where the intrinsic homogeneity of the network is used. The extended results further explain why optimizing the ℓ^2–regularized loss typically used in practice can lead to parameters with a large margin and good generalization. They further remark that the maximum possible margin is non-decreasing in the width of the architecture, so their generalization bounds improve as the size of the network grows. Thus, even if the dataset is already separable, it could still be useful to further over-parameterize to achieve better generalization. This is in line with common writing in the machine learning literature, highlighted in [276], that statistical models operating in the overparametrization regime – for instance, neural networks with significantly more trainable parameters than the number N of training data –, perform well experimentally as they fit the entire training dataset, namely the training error $\mathcal{E}$ is zero, but do so without overfitting.

6.1.3 Outline

The remainder of the paper is organized as follows.

- In Section 6.2, we give a brief but comprehensive presentation on the neural ODE perspective of deep supervised learning.

- In Section 6.3, we present our main results in the context of regularized empirical risk minimization (Theorem 6.1 and Theorem 6.3).

- In Section 6.4, we present our main result for the augmented empirical risk minimization problem, namely exponential decay of the training error with exponentially small parameters (Theorem 6.5). We also present the local simultaneous controllability result (Theorem 6.6).

- Finally in Section 6.5, we present the continuous analog of residual neural networks with variable widths, depict some possible approaches for passing from the continuous to the discrete case, and present extensions of Theorem 6.1 and Theorem 6.5 in this context.

- The proofs of all results may be found in Section 6.7.

6.1.4 Notation and assumptions

We denote $\dot{\mathbf{x}}(t) := \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}(t)$. For $a \in \mathbb{R}^n$, we denote by $a^\top$ its transpose. We use the notation $\mathbf{x}_T$ and u_T to display the dependence of these variables on the time horizon T. We designate by $\|a\|$ the standard euclidean norm when a is a vector, and the Frobenius norm when a is a matrix/tensor. We denote by $\mathrm{Lip}(\mathbb{R})$ the set of functions $f : \mathbb{R} \longrightarrow \mathbb{R}$ which are globally Lipschitz continuous, and by $L^2(0,T;\mathbb{R}^n)$ (resp. $H^1(0,T;\mathbb{R}^n)$) the Lebesgue (resp. Sobolev) space consisting of all functions $f : (0,T) \longrightarrow \mathbb{R}^n$ which are square integrable (resp. square integrable and with a square integrable weak derivative) – recall that $H^1(0,T;\mathbb{R}^n)$ is endowed with the norm $\|f\|^2_{H^1(0,T;\mathbb{R}^n)} := \|f\|^2_{L^2(0,T;\mathbb{R}^n)} + \|\dot{f}\|^2_{L^2(0,T;\mathbb{R}^n)}$.

Throughout the remainder of this work, we will work under the following couple of assumptions, which are universal in the context of machine learning.

Assumption 6.1.1. *We henceforth assume that we are given a training dataset*

$$\{\vec{x}_i, \vec{y}_i\}_{i=1}^N \subset \mathbb{R}^{d\times N} \times \mathbb{R}^{m\times N},$$

with $\vec{x}_i \neq \vec{x}_j$ for $i \neq j$. unless otherwise stated, any initial datum $\mathbf{x}^0 \in \mathbb{R}^{d\times N}$ for the systems under consideration will take the form $\mathbf{x}^0 = [\vec{x}_1, \dots, \vec{x}_N]$.

The following assumption is satisfied by most of the commonly used activation functions, including sigmoids such as $\sigma(x) = \tanh(x)$, and rectifiers: $\sigma(x) = \max\{\alpha x, x\}$ for $\alpha \in [0,1)$.

Assumption 6.1.2. *unless otherwise stated, we fix an activation function σ satisfying*

$$\sigma \in \mathrm{Lip}(\mathbb{R}) \quad \textit{and} \quad \sigma(0) = 0.$$

6.2 Roadmap to learning via neural ODEs

6.2.1 Feed-forward neural networks

The canonical example of a feed-forward neural network is the multi-layer perceptron (MLP), which generally takes the form

$$\begin{cases} \mathbf{x}_i^{k+1} = \sigma(w^k \mathbf{x}^k + b^k) & \text{for } k \in \{0, \dots, N_{\text{layers}} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i \in \mathbb{R}^d \end{cases} \tag{6.2.1}$$

for $i \in [N]$. The integer $N_{\text{layers}} \geqslant 1$ is the depth of the neural network (6.2.1), and each k is a layer. For any i, the vector $\mathbf{x}_i^k \in \mathbb{R}^{d_k}$ designates the state at the layer k, d_k

is referred to as the width of the layer k, while $w^k \in \mathbb{R}^{d_{k+1}\times d_k}$ and $b^k \in \mathbb{R}^{d_k}$ are the optimizable weight and bias parameters of the network (6.2.1). Finally, $\sigma \in \mathrm{Lip}(\mathbb{R})$ is a fixed nonlinear activation function – by abuse of notation, we define the vector-valued analog of σ component-wise, namely, $\sigma : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is defined by

$$\sigma(\mathbf{x})_j := \sigma(\mathbf{x}_j) \qquad \text{for } j \in [d].$$

Common choices include sigmoids such as $\sigma(x) = \tanh(x)$ or $\sigma(x) = \frac{1}{1+e^{-x}}$, and rectifiers: $\sigma(x) = \max\{x, ax\}$ for a fixed $0 \leqslant a < 1$. In practice, the activation σ is generally selected using cross-validation. It can readily be seen that the formulation (6.2.1) coincides with the more conventional formulation of neural networks as compositional structures of parametric affine operators and nonlinearities, as namely

$$\mathbf{x}_i^{N_{\text{layers}}} = (\sigma \circ \Lambda^k \circ \ldots \circ \sigma \circ \Lambda^0)(\vec{x}_i),$$

with $\Lambda^k \vec{x} := w^k \vec{x} + b^k$ for $k \in \{0, \ldots, N_{\text{layers}}\}$.

Note that the iterative nature of the MLP (6.2.1) stimulates permuting the order of the parametric affine maps and the nonlinearity σ, to the effect of considering the equivalent, but somewhat simpler system

$$\begin{cases} \mathbf{x}_i^{k+1} = w^k \sigma(\mathbf{x}^k) + b^k & \text{for } k \in \{0, \ldots, N_{\text{layers}} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i \in \mathbb{R}^d. \end{cases} \tag{6.2.2}$$

We will henceforth concentrate on residual neural networks (ResNets). Contrary to the multi-layer perceptrons (6.2.1) – (6.2.2), when considering ResNets one typically needs to assume that the width d_k is fixed over every layer k, namely $d_k = d$ for every k. We refer to Section 6.5 for variable width ResNets. In the fixed width context, a residual neural network generally takes the form

$$\begin{cases} \mathbf{x}_i^{k+1} = \mathbf{x}_i^k + g(u^k, \mathbf{x}_i^k) & \text{for } k \in \{0, \ldots, N_{\text{layers}} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i \in \mathbb{R}^d \end{cases} \tag{6.2.3}$$

for $i \in [N]$, where $\mathbf{x}_i^k \in \mathbb{R}^d$ for any i, k, $u^k := [w^k, b^k] \in \mathbb{R}^{d\times d+d}$ and g is as in (6.2.1) or (6.2.2). As explained in [193], other classes of networks (including specific subclasses of CNNs) can be fit into the residual network framework.

6.2.2 Neural ODEs

It is readily seen that (6.2.3) corresponds, modulo a scaling factor $\Delta t = \frac{T}{N_{\text{layers}}}$, to the forward Euler discretization of

$$\begin{cases} \dot{\mathbf{x}}_i(t) = g(u(t), \mathbf{x}_i(t)) & \text{in } (0, T) \\ \mathbf{x}_i(0) = \vec{x}_i \in \mathbb{R}^d, \end{cases} \tag{6.2.4}$$

for $i \in [N]$. Here $T > 0$ is a fixed time horizon, and $u(t) := [w(t), b(t)] \in \mathbb{R}^{d\times d+d}$. As per what precedes, the nonlinearity g in (6.2.4) may take the form

$$g(u(t), \mathbf{x}_i(t)) := \sigma(w(t)\mathbf{x}_i(t) + b(t)) \tag{6.2.5}$$

or

$$g(u(t), \mathbf{x}_i(t)) = w(t)\sigma(\mathbf{x}_i(t)) + b(t). \tag{6.2.6}$$

for $i \in [N]$. We will address both cases in our analytical study, and emphasize the stark differences between the two. The above parametrizations are not the lone considered in practice. In fact, one may consider, for instance, combinations of (6.2.5) and (6.2.6) which allow intermediate exploration (bottlenecks) in higher dimensions:

$$g(u(t), \mathbf{x}_i(t)) := w_2(t)\sigma(w_1(t)\mathbf{x}_i(t) + b_1(t)) + b_2(t) \tag{6.2.7}$$

where now $w_1(t) \in \mathbb{R}^{d_{\text{hid}}\times d}$, $w_2(t) \in \mathbb{R}^{d\times d_{\text{hid}}}$, $b_1(t) \in \mathbb{R}^{d_{\text{hid}}}$ and $b_2(t) \in \mathbb{R}^d$. In fact, (6.2.3) with g as in (6.2.7) is much like the original ResNet first presented in [140].

6.2.3 Training

For an input sample $\vec{x}_i \in \mathbb{R}^d$, the prediction of the neural ODE (6.2.4) is a flattening of the form $P\mathbf{x}_i(T) \in \mathbb{R}^m$ for some $P \in C^0(\mathbb{R}^d; \mathbb{R}^m)$. In practice,

$$\begin{aligned} Px &:= \mathrm{softmax}(p_1 x + p_2) \qquad && \text{(classification)}, \\ Px &:= p_1 x + p_2 && \text{(regression)}, \end{aligned} \tag{6.2.8}$$

where $p_1 \in \mathbb{R}^{m\times d}$ and $p_2 \in \mathbb{R}^m$ are optimizable parameters, and sofmax is defined by

$$\mathrm{softmax}(z)_j = \frac{e^{z_j}}{\sum_{\ell=1}^m e^{z_\ell}}$$

for $z \in \mathbb{R}^m$ and $j \in [m]$. In the context of binary classification, namely $m = 1$ with $\vec{y}_i = \pm 1$, one may also use $Px := \tanh(p_1 x + p_2)$. As mentioned in the introduction, we shall assume that the parameters p_1, p_2 defining P are given (but arbitrary and possibly picked at random unless specified) for technical reasons.

In supervised learning, one seeks to tune the parameters $[w, b]$ so that $P\mathbf{x}_i(T)$ most closely resembles $\vec{y}_i$ for $i \in [N]$. To this end, the Tikhonov-regularized, empirical risk minimization problem

$$\inf_{\substack{[w,b] \\ \mathbf{x}_i \text{ solves (6.2.4)}}} \frac{1}{N}\sum_{i=1}^N \mathrm{loss}\big(P\mathbf{x}_i(T), \vec{y}_i\big) + \lambda \Big\| [w, b] \Big\|^2_{H^k(0,T;\mathbb{R}^{d_u})} \tag{6.2.9}$$

where $\lambda > 0$ is fixed, and $\mathbf{x}_i$ solves (6.2.4) with g as in (6.2.5) or (6.2.6) (although, more general cases such as (6.2.7) can also be considered). Here $\mathrm{loss}(\cdot,\cdot) : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+$ is a given continuous function, which in our work we will choose on a case-by-case basis. Note that (6.2.9) is the empirical and regularized approximation of the expected risk minimization problem:

$$\inf_{\substack{[w,b] \\ \mathbf{x}_i \text{ solves (6.2.4)}}} \mathbb{E}\Big[\mathrm{loss}\left(P\mathbf{x}_\cdot(T), \cdot\right)\Big],$$

where $\mathbb{E}[f(\cdot,\cdot)] := \int_{\mathbb{R}^d\times\mathbb{R}^m} f(x,y)\,\mathrm{d}\rho(x,y)$, with $\mathbf{x}_{\vec{x}}$ denoting the solution to (6.2.4) with initial datum $\vec{x}$. Here $\rho : \mathbb{R}^d \times \mathbb{R}^m \to [0,1]$ is an unknown joint probability distribution, from which one samples the training dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$. We shall solely focus on the empirical problem in this work.

By virtue of the classical direct method in the calculus of variations, we may readily prove the existence of minimizer for a class of the learning problems we consider herein.

Proposition 6.2.1 (Existence of minimizers)**.** *Let $T > 0$, $\lambda > 0$, and let $\mathrm{loss} \in C^0(\mathbb{R}^m \times \mathbb{R}^m; \mathbb{R}_+)$ and $P \in C^0(\mathbb{R}^d; \mathbb{R}^m)$ be given. The minimization problems (6.2.9) and (6.1.4) admit a global solution $[w, b] \in H^k(0, T; \mathbb{R}^{d_u})$, with $k = 0$ for (6.3.3) or $k = 1$ for (6.3.2) – (6.2.7).*

Remark 6.2.2 (Sobolev regularization)**.** *We stress the possible need for considering a Sobolev H^1–regularization in the case of (6.3.2), as otherwise, we may not a priori guarantee the existence of a global minimizer. Indeed, an issue arises due to the specific nonlinear form of the neural ODE (6.3.2), which could be an impediment for passing to the limit in the equation using only weak convergences (consider, for instance, the sequence $\{\sin(nx)\}_{n=1}^{+\infty}$ and $\sigma(x) = \max\{x, 0\}$). This issue is specific to the continuous-time setting, as in the discrete-time thus finite dimensional optimization setting, weak and strong convergences coincide.*

6.3 Empirical risk minimization

Throughout the paper, we will focus on neural ODEs given by (6.2.4) with g as in (6.2.5) or (6.2.6). The results can thence be extrapolated to the case when g is parametrized by (6.2.7) whenever w_1 and b_1 (resp. w_2, b_2) are time-independent – we comment on further extensions on a case-by-case basis. As it will be rather convenient to work with the full stacked state trajectory $\mathbf{x}(t) = [\mathbf{x}_1(t), \dots, \mathbf{x}_N(t)]$, we introduce some further notation. We shall henceforth denote

$$d_u := d \times (d+1), \qquad d_x := d \times N.$$

Moreover, given $w \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$, we shall write

$$\mathbf{w} := \begin{bmatrix} w & & \\ & \ddots & \\ & & w \end{bmatrix} \in \mathbb{R}^{d_x \times d_x}, \qquad \mathbf{b} := \begin{bmatrix} b \\ \vdots \\ b \end{bmatrix} \in \mathbb{R}^{d_x}. \tag{6.3.1}$$

In view of the above discussion and noting (6.3.1), we will consider stacked neural ODEs in $\mathbb{R}^{d_x}$ such as

$$\begin{cases} \dot{\mathbf{x}}(t) = \sigma(\mathbf{w}(t)\mathbf{x}(t) + \mathbf{b}(t)) & \text{for } t \in (0,T) \\ \mathbf{x}(0) = \mathbf{x}^0 \in \mathbb{R}^{d_x}, \end{cases} \tag{6.3.2}$$

and

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{w}(t)\sigma(\mathbf{x}(t)) + \mathbf{b}(t) & \text{for } t \in (0,T) \\ \mathbf{x}(0) = \mathbf{x}^0 \in \mathbb{R}^{d_x}. \end{cases} \tag{6.3.3}$$

In this section, we consider the problem of regularized empirical risk minimization. For simplicity of notation, we henceforth denote the training error (empirical risk) by

$$\mathcal{E}(\mathbf{x}) := \frac{1}{N}\sum_{i=1}^{N} \mathrm{loss}\big(P\mathbf{x}_i, \vec{y}_i\big), \tag{6.3.4}$$

for $\mathbf{x} \in \mathbb{R}^{d_x}$, where $P \in C^\infty(\mathbb{R}^d; \mathbb{R}^m)$ and $\mathrm{loss}(\cdot,\cdot) \in C^0(\mathbb{R}^m \times \mathbb{R}^m; \mathbb{R}_+)$ are given – both will change with respect to the task in question (regression, classification), as discussed in (6.2.8).

For fixed $\lambda > 0$, we will study the behavior when $T \gg 1$ of global minimizers $[w_T, b_T]$ to the functional

$$J_{\lambda,T}(w,b) = \mathcal{E}(\mathbf{x}(T)) + \lambda \Big\| [w,b] \Big\|^2_{H^k(0,T;\mathbb{R}^{d_u})} \tag{6.3.5}$$

where $\mathbf{x} \in C^0([0,T]; \mathbb{R}^{d_x})$ is the unique solution to either (6.3.3) ($k=0$) or (6.3.2) ($k=1$) corresponding to the parameters $[w,b] \in H^k(0,T;\mathbb{R}^{d_u})$, noting (6.3.1).

6.3.1 Regression

We begin by considering the case wherein P and $\mathrm{loss}(\cdot,\cdot)$ are chosen in (6.3.4) so that $\mathrm{loss}(x,x) = 0$. This is for instance the case when loss is a distance inferred by a norm (e.g. $\mathrm{loss}(x,y) = \|x-y\|^p_{\ell^p}$, $p = 1, 2$), and P is an affine map. Such modeling assumptions are typically made in the context of regression tasks, wherein when minimizing the training error, one looks to interpolate the training data by means of the projected neural ODE flow. Our asymptotic convergence result will entail a property of this form, reflected by the following definition.

Definition 6.3.1 (Interpolation)**.** *Let $P : \mathbb{R}^d \to \mathbb{R}^m$ be any non-zero affine map. We say that (6.3.3) (resp. (6.3.2)) interpolates the dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ in some time $T > 0$ if there exists a time $T > 0$ and some parameters $[w,b] \in L^2(0,T;\mathbb{R}^{d_u})$ (resp. in $H^1(0,T;\mathbb{R}^{d_u})$) such that the unique solution $\mathbf{x}$ to (6.3.3) (resp. (6.3.2)), noting (6.3.1), satisfies*

$$P\mathbf{x}_i(T) = \vec{y}_i \qquad \text{for all} \quad i \in [N].$$

Let us first note that by means of an elementary time-scaling, if (6.3.3) or (6.3.2) interpolates the dataset in some time $T > 0$, it interpolates it in any time, in particular, in time 1. We will make use of this observation to simplify the subsequent presentation and analysis by assuming interpolation in time 1 without loss of generality.

Clearly, in view of the definition of $\mathcal{E}$ in (6.3.4) with loss and P as above, if interpolation holds, then the minimum of $\mathcal{E}$ (equal to 0) is attained. We may state our main result in this context.

Theorem 6.1. *Let $\lambda > 0$ be fixed. Suppose that $P : \mathbb{R}^d \to \mathbb{R}^m$ is any non-zero affine map, and suppose that $\mathrm{loss} \in C^0(\mathbb{R}^m \times \mathbb{R}^m; \mathbb{R}_+)$ in (6.3.4) is such that $\mathrm{loss}(x, x) = 0$. Assume that (6.3.3) (resp. (6.3.2) with σ 1–homogeneous) interpolates the dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ in time 1 in the sense of Definition 6.3.1. For any $T \geqslant 1$ let $[w_T, b_T] \in L^2(0, T; \mathbb{R}^{d_u})$ (resp. in $H^1(0, T; \mathbb{R}^{d_u})$) be any pair of global minimizers to $J_{\lambda,T}$ defined in (6.3.5), and let $\mathbf{x}_T$ be the unique associated solution to (6.3.3) (resp. (6.3.2)), noting (6.3.1). The following properties then hold.*

1. *There exists a constant $C = C(\mathbf{x}^0, \vec{y}, \lambda) > 0$ independent of T such that*
$$\mathcal{E}(\mathbf{x}_T(T)) \leqslant \frac{C}{T}.$$

2. *There exists a sequence $\{T_n\}_{n=1}^{+\infty}$, with $T_n > 0$ and $T_n \xrightarrow[n\longrightarrow+\infty]{} +\infty$, and some $\mathbf{x}_\circ \in \mathbb{R}^{d_x}$ with $\mathcal{E}(\mathbf{x}_\circ) = 0$ such that, along a subsequence,*
$$\mathbf{x}_{T_n}(T_n) \xrightarrow[n\longrightarrow+\infty]{} \mathbf{x}_\circ. \tag{6.3.6}$$

3. *For any $n \geqslant 1$, set*
$$\begin{aligned} w_n(t) &:= T_n\, w_{T_n}(t\, T_n) && \text{for } t \in [0,1], \\ b_n(t) &:= T_n\, b_{T_n}(t\, T_n) && \text{for } t \in [0,1]. \end{aligned}$$
Then along a subsequence,
$$\Big\| [w_n, b_n] - [w^*, b^*] \Big\|_{H^k(0,1;\mathbb{R}^{d_u})} \xrightarrow[n\longrightarrow+\infty]{} 0,$$
where $[w^*, b^*] \in H^k(0, 1; \mathbb{R}^{d_u})$ is some solution to the minimization problem
$$\inf_{\substack{[w,b]\in H^k(0,1;\mathbb{R}^{d_u}) \\ \mathbf{x} \text{ solves (6.3.2) (resp. (6.3.3)) in } [0,1] \\ \text{and} \\ P\mathbf{x}_i(1) = \vec{y}_i \quad \forall i}} \Big\| [w, b] \Big\|^2_{H^k(0,1;\mathbb{R}^{d_u})}.$$

Idea of proof. The proof of Theorem 6.1 may be found in Section 6.7.1. Let us motivate the main underlying idea.

Under the above assumptions, both (6.3.2) and (6.3.3) will be 1–homogeneous with respect to the parameters $[w(t), b(t)]$. Namely, both (6.3.2) and (6.3.3) (noting (6.3.1)) can be written as
$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), w(t), b(t)) & \text{in } (0, T) \\ \mathbf{x}(0) = \mathbf{x}^0, \end{cases} \tag{6.3.7}$$
where $\mathbf{f}(\mathbf{x}, \alpha w, \alpha b) = \alpha \mathbf{f}(\mathbf{x}, w, b)$ for $\alpha > 0$. Whilst in the case of (6.3.3) this homogeneity property holds for any activation function σ, we require σ to be 1–homogeneous for neural networks such as (6.3.2). This includes rectifiers, but excludes sigmoids.

A simple computation (see Lemma 7.2.1) then leads to noting that, given some parameters $u^1 := [w^1, b^1]$ and the solution $\mathbf{x}^1$ to
$$\begin{cases} \dot{\mathbf{x}}^1(t) = \mathbf{f}(\mathbf{x}^1(t), w^1(t), b^1(t)) & \text{in } (0, 1) \\ \mathbf{x}^1(0) = \mathbf{x}^0, \end{cases} \tag{6.3.8}$$

then $u_T(t) := \frac{1}{T} u^1(\frac{t}{T})$ for $t \in [0,T]$ is such that $\mathbf{x}_T(t) := \mathbf{x}^1(\frac{t}{T})$ solves (6.3.7). Under the interpolation assumption, we may find $u^1 \in H^k(0,T_0;\mathbb{R}^{d_u})$ such that the corresponding solution $\mathbf{x}^1$ satisfies $\mathcal{E}(\mathbf{x}^1(1)) = 0$, and then use the above scaling and the optimality of u_T to deduce

$$J_{\lambda,T}(u_T) \leqslant \mathcal{E}\left(\mathbf{x}^1(1)\right) + \frac{\lambda}{T} \left\|u^1\right\|^2_{H^k(0,1;\mathbb{R}^{d_u})} = \frac{\lambda}{T} \left\|u^1\right\|^2_{H^k(0,1;\mathbb{R}^{d_u})}$$

for $T \geqslant 1$. This will imply the decay estimate of $\mathcal{E}$, and combined with some more technical compactness arguments, will yield the remaining convergence results as well.

On another hand, considering the case of (6.3.3) and thus $k = 0$, we see that

$$\begin{aligned}
&\inf_{\substack{u_T = [w_T, b_T] \in L^2(0,T;\mathbb{R}^{d_u}) \\ \mathbf{x}_T \text{ solves (6.3.7)}}} \mathcal{E}(\mathbf{x}_T(T)) + \lambda \int_0^T \|u_T(t)\|^2 \, \mathrm{d}t \\
&= \inf_{\substack{u_T = [w_T, b_T] \in L^2(0,T;\mathbb{R}^{d_u}) \\ \mathbf{x}_T \text{solves (6.3.7)}}} \mathcal{E}(\mathbf{x}_T(T)) + \frac{\lambda}{T} \int_0^1 \|T u_T(sT)\|^2 \, \mathrm{d}s \\
&= \inf_{\substack{u^1 = [w^1, b^1] \in L^2(0,1;\mathbb{R}^{d_u}) \\ \mathbf{x}^1 \text{solves (6.3.8)}}} \mathcal{E}(\mathbf{x}^1(1)) + \frac{\lambda}{T} \int_0^1 \left\|u^1(s)\right\|^2 \, \mathrm{d}s.
\end{aligned}$$

This computation indicates that one may consider the behavior when $T \longrightarrow +\infty$ for fixed $\lambda > 0$ and that when $\lambda \searrow 0$ for fixed $T > 0$ in the same fashion. Although this scaling is specific to the L^2–regularization setting, it motivates completing Theorem 6.1 with the following result.

Theorem 6.2. *Under the assumptions of Theorem 6.1, fix $T > 0$, and for any $\lambda > 0$, let $[w_\lambda, b_\lambda] \in L^2(0,T;\mathbb{R}^{d_u})$ (resp. $H^1(0,T;\mathbb{R}^{d_u})$) be any pair of global minimizers to $J_{\lambda,T}$ defined in (6.3.5), and let $\mathbf{x}_\lambda$ be the unique associated solution to (6.3.3) (resp. (6.3.2)), noting (6.3.1). The following properties then hold.*

1. *There exists a constant $C = C(\mathbf{x}^0, \vec{y}, T) > 0$ independent of $\lambda > 0$ such that*

$$\mathcal{E}(\mathbf{x}_\lambda(T)) \leqslant C\lambda.$$

2. *There exists a sequence $\{\lambda_n\}_{n=1}^{+\infty}$, with $\lambda_n > 0$ and $\lambda_n \xrightarrow[n\to+\infty]{} 0$, and some $\mathbf{x}_\circ \in \mathbb{R}^{d_x}$ with $\mathcal{E}(\mathbf{x}_\circ) = 0$ such that, along a subsequence*

$$\mathbf{x}_{\lambda_n}(T) \xrightarrow[n\longrightarrow+\infty]{} \mathbf{x}_\circ.$$

3. *Moreover, along a subsequence,*

$$\left\| [w_{\lambda_n}, b_{\lambda_n}] - [w^*, b^*] \right\|_{H^k(0,T;\mathbb{R}^{d_u})} \xrightarrow[n\longrightarrow+\infty]{} 0,$$

where $[w^*, b^*]^\top \in H^k(0,T;\mathbb{R}^{d_u})$ is some solution to the minimization problem

$$\inf_{\substack{[w,b] \in H^k(0,T;\mathbb{R}^{d_u}) \\ \mathbf{x} \text{ solves (6.3.2) (resp. (6.3.3))} \\ \text{and} \\ P\mathbf{x}_i(T) = \vec{y}_i \quad \forall i}} \left\|[w,b]\right\|^2_{H^k(0,T;\mathbb{R}^{d_u})}.$$

As an intermezzo before proceeding with the classification tasks and an in-depth discussion, let us note that by means of an elementary Grönwall argument, we first show the following illustrative result, which stipulates a lower bound for the cost of the weights w in terms of the way the dataset is "spread out".

Proposition 6.3.2. *Let $P \in C^0(\mathbb{R}^d;\mathbb{R}^m)$ be surjective, and let $T>0$. Assume that for some parameters $[w,b]$, the solution $\mathbf{x}=[\mathbf{x}_1,\dots,\mathbf{x}_N]$ to either (6.3.3) or (6.3.2) satisfies*

$$P\mathbf{x}_i(T)=\vec{y}_i \qquad \text{for all } i\in[N].$$

Then

$$\|w\|_{L^1(0,T;\mathbb{R}^{d_u})} \geqslant C(\sigma) \max_{\substack{(i,j)\in[N]^2\\ i\neq j}} \inf_{\substack{\mathbf{x}_i^1\in P^{-1}(\{\vec{y}_i\})\\ \mathbf{x}_j^1\in P^{-1}(\{\vec{y}_j\})}} \log\left(\frac{\left\|\mathbf{x}_i^1-\mathbf{x}_j^1\right\|}{\left\|\mathbf{x}_i^0-\mathbf{x}_j^0\right\|}\right), \tag{6.3.9}$$

where $C(\sigma)>0$ is the Lipschitz constant of σ.

By virtue of Cauchy-Schwarz, (6.3.9) clearly implies

$$\|w\|_{L^2(0,T;\mathbb{R}^{d_u})} \geqslant \frac{C(\sigma)}{\sqrt{T}} \max_{\substack{(i,j)\in[N]^2\\ i\neq j}} \inf_{\substack{\mathbf{x}_i^1\in P^{-1}(\{\vec{y}_i\})\\ \mathbf{x}_j^1\in P^{-1}(\{\vec{y}_j\})}} \log\left(\frac{\left\|\mathbf{x}_i^1-\mathbf{x}_j^1\right\|}{\left\|\mathbf{x}_i^0-\mathbf{x}_j^0\right\|}\right).$$

6.3.2 Classification

We now consider the standard setting of classification tasks, namely wherein the labels $\vec{y}_i$ take values in a set of $m\geqslant 2$ classes – unless otherwise stated, we henceforth consider $\vec{y}_i\in[m]$ for all $i\in[N]$. We will focus on the cross-entropy loss in (6.3.4), which we recall, reads

$$\mathrm{loss}(P\mathbf{x}_i(T),\vec{y}_i) := -\log\left(\frac{e^{P\mathbf{x}_i(T)_{\vec{y}_i}}}{\sum_{j=1}^m e^{P\mathbf{x}_i(T)_j}}\right), \tag{6.3.10}$$

where $P:\mathbb{R}^d\longrightarrow\mathbb{R}^m$ is made precise later on. An important feature of the cross-entropy loss is the fact that it is not coercive with respect to the first variable – namely, as $P\mathbf{x}_i(T)_{\vec{y}_i}$ goes to infinity, the loss goes to zero. This is very much in line with intuition regarding the classification task, as one looks to separate the features with respect to their individual class in the label space $\mathbb{R}^m$.

Remark 6.3.3 (Binary classification)**.** *We note that for two classes, one could consider $\vec{y}_i\in\{-1,+1\}$ and train either with the logistic loss in (6.3.4)*

$$\mathrm{loss}(P\mathbf{x}_i(T),\vec{y}_i) := \log\left(1+e^{-\vec{y}_i P\mathbf{x}_i(T)}\right),$$

where $P:\mathbb{R}^d\to\mathbb{R}$ is an affine map, or even with the squared ℓ^2–loss

$$\mathrm{loss}(P\mathbf{x}_i(T),\vec{y}_i) := \|P\mathbf{x}_i(T)-\vec{y}_i\|^2$$

where $Px=\tanh(p_1\cdot x+p_2)$ with $p_1\in\mathbb{R}^d$ and $p_2\in\mathbb{R}$.

The problem consisting of classifying a given dataset is closely tied to the following rather intuitive notion of separability, which we will require in the subsequent results.

Definition 6.3.4 (Separability)**.** *Let $P:\mathbb{R}^d\longrightarrow\mathbb{R}^m$ be any non-zero affine map. We say that (6.3.3) (resp. (6.3.2)) separates the dataset $\{\vec{x}_i,\vec{y}_i\}_{i=1}^N$ with respect to P if there exists a time $T>0$ and some parameters $[w,b]\in L^2(0,T;\mathbb{R}^{d_u})$ (resp. in $H^1(0,T;\mathbb{R}^{d_u})$) such that the unique solution $\mathbf{x}$ to (6.3.3) (resp. (6.3.2)) satisfies*

$$P\mathbf{x}_i(T)_{\vec{y}_i} - \max_{\substack{j\in[N]\\ j\neq\vec{y}_i}} P\mathbf{x}_i(T)_j > 0 \qquad \text{for all } i\in[N].$$

In other words, a parametrized neural ODE flow separates the given dataset if the corresponding *margin* $\gamma_{[w,b]}$, defined as

$$\gamma_{[w,b]} := \min_{i\in[N]} \left(P\mathbf{x}_i(T)_{\vec{y}_i} - \max_{\substack{j\in[N]\\ j\neq\vec{y}_i}} P\mathbf{x}_i(T)_j \right) \tag{6.3.11}$$

is positive. We may now state our main result in the classification context, which entails a quantitative rate of decay as $T \longrightarrow +\infty$ of the training error with cross-entropy loss for ReLU activated neural ODEs.

Theorem 6.3. *Let $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ be a given dataset with $\vec{x}_i \in \mathbb{R}^d$ and $\vec{y}_i \in [m]$. Let $\lambda > 0$ be fixed, and let $\mathfrak{Q} : \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^d$ be a non-zero affine map such that $\mathfrak{Q}\vec{x}_i \geqslant 0$ for $i \in [N]$. Set*

$$\mathbf{x}_i^0 := \mathfrak{Q}\vec{x}_i \qquad \text{for } \ i \in [N],$$

and let $P \in \mathbb{R}^{m\times d}$ be any non-zero matrix such that (6.3.2) with $\sigma(x) = \max\{x, 0\}$ separates the dataset $\{\mathbf{x}_i^0, \vec{y}_i\}_{i=1}^N$ with respect to P in some time $T_0 > 0$ as per Definition 6.3.4, and let γ denote the associated margin as defined in (6.3.11). For any $T \geqslant T_0$, let $[w_T, b_T] \in H^1(0, T; \mathbb{R}^{d_u})$ be any pair of global minimizers to $J_{\lambda,T}$ defined in (6.3.5)–(7.1.3), and let $\mathbf{x}_T$ be the associated unique solution to (6.3.2) with $\sigma(x) = \max\{x, 0\}$. Then, there exists a constant $C = C(\mathbf{x}^0, \vec{y}, \lambda) > 0$ independent of $T > 0$ such that

$$\mathcal{E}(\mathbf{x}_T(T)) \leqslant \log\left(1 + (m-1)e^{-\gamma\, e^{\frac{T^\alpha}{2}}}\right) + CT^{2\alpha-1} \tag{6.3.12}$$

holds for any $\alpha \in \left(0, \frac{1}{2}\right)$.

We note that the above theorem is very specific to neural ODEs of the form (6.3.2) with ReLU activations, and the specific form of the cross-entropy loss, from which the first term in the estimate (6.3.12) is derived. This is due to the proof strategy, which relies on using the positivity of the right hand side to, in some sense, obtain a linear equation for the projected output features for some auxiliary parameters constructed within the proof, and thus have an explicit solution for these parameters of the form $\sim e^t$. This stimulates the appearance of the second exponential within the log in (6.3.12).

Much like what we observed in the regression setting, we can expect to link the limit as T goes to infinity with the convergence of the regularization path, namely the limit as $\lambda \searrow 0$. This is depictd in the following theorem.

Theorem 6.4. *Under the assumptions of Theorem 6.3, fix $T \geqslant T_0$ and for any $\lambda > 0$ let $[w_\lambda, b_\lambda] \in H^1(0, T; \mathbb{R}^{d_u})$ be any pair of global minimizers to $J_{\lambda,T}$ defined in (6.3.5)–(7.1.3), and let $\mathbf{x}_\lambda$ be the associated unique solution to (6.3.2) with $\sigma(x) = \max\{x, 0\}$. Then, there exists a constant $C = C(\mathbf{x}^0, \vec{y}, T) > 0$ independent of $\lambda > 0$ such that for any $\alpha \in \left(0, \frac{1}{2}\right)$,*

$$\mathcal{E}(\mathbf{x}_\lambda(T)) \leqslant \log\left(1 + (m-1)e^{-\gamma\, e^{\frac{\lambda^{-\alpha}}{2}}}\right) + C\lambda^{-2\alpha+1}.$$

6.3.3 Discussion

When training without explicit regularization (i.e. $\lambda = 0$), a common approach in the literature is to resort to algorithm-dependent generalization analysis, where the end results are surprisingly similar to the limit $\lambda \searrow 0$. In this case, the implicit bias of gradient descent ([249, 128]) indicates that in the overparametrized regime, after training a neural network (or other statistical model, such as linear regression) with gradient-based methods until zero training error, without requiring any explicit parameter regularization, among the many classifiers which overfit on the training dataset, the algorithm

selects the one which performs best on the test dataset (e.g. minimal ℓ^2–norm solution or max-margin solution). Recent works have shown that gradient descent can allow overparametrized multi-layer networks to attain arbitrarily low training error on fairly general datasets ([83, 8, 9]), and find minimum-norm/maximum-margin solutions that fit the data in the settings of logistic regression, deep linear networks, and symmetric matrix factorization ([128, 249, 149]). In [62, 63] overparametrization is approached from the point of view of the width of the neural network, unlike our depth-inspired perspective. The authors consider a 2-layer shallow perceptron with ReLU activation, and exhibit the Wasserstein gradient flow formulation of the descent scheme yielding controls, and they consequently prove that these controls approach global minimizers of the cost functional when the width increases, with the global minimizer being characterized as a max-margin classifier in a certain non-Hilbertian space of functions.

Remark 6.3.5 (Extensions)**.** *Let us comment on the assumptions of the asymptotic results given in what precedes.*

- *The issue that appears in the results presented above when considering neural ODEs of the form*

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{w}^1(t)\sigma\left(\mathbf{w}^2(t)\mathbf{x}(t) + \mathbf{b}^2(t)\right) + \mathbf{b}^1(t) & \text{in } (0,T) \\ \mathbf{x}(0) = \mathbf{x}^0, \end{cases} \tag{6.3.13}$$

  where $\mathbf{w}^1(t) \in \mathbb{R}^{d_x \times (d_{\text{hid}} \times N)}$ and $\mathbf{w}^2 \in \mathbb{R}^{(d_{\text{hid}} \times N) \times d_x}$ is the lack of homogeneity (and thus scaling) with respect to the parameters. Consequently, one cannot see that the squared L^2 (or Sobolev) norm of the parameters scales like $\frac{1}{T}$ as simply as before, a property which is the cornerstone of our proofs.

- *We note that the output layer parameters given by the affine map P are fixed, but in general arbitrary and may be picked at random (e.g., from a normal distribution), in most of the preceding results. This is due to the fact if we were to optimize P as well, we would have to ensure that the optimal P is bounded with respect to the limiting hyper-parameter (T or λ). This in turn could perhaps be ensured if we were to regularize the output layer as well, but would, in turn, be an impediment to the scaling arguments we use in all proofs since now the parameter regularization norm would not scale polynomially with T.*

Remark 6.3.6 (*Deep limits*)**.** *In [257] (see also [16]), the authors show, via Γ-convergence arguments, that the optimal control parameters in the discrete-time context converge to those of the continuous-time context when the time-step converges to 0. The latter is interpreted as an infinite layer limit when the final time horizon T in the continuous-time context is fixed (equal to 1). Our result is of different nature. Rather than aim to prove that the discrete-time controls converge to the continuous-time ones, we exhibit the continuous-time neural ODE representation, for which the final time horizon clearly commands the number of layers for the associated time-discretization when the time-step is fixed, and aim to characterize the possible phenomena which arise whenever this time horizon increases.*

6.4 Augmented empirical risk minimization

We are now interested seeing whether one can obtain better quantitative estimates for the decay of the training error $\mathcal{E}$ to 0 with respect to the time horizon ($\sim$ number of layers) $T > 0$ – namely, improve the $\mathcal{O}\left(\frac{1}{T}\right)$–rate of convergence of the training error to 0 manifested in Theorem 6.1 and Theorem 6.3.

We will henceforth solely concentrate on the ℓ^2–loss; in other words,

$$\mathcal{E}(\mathbf{x}) := \frac{1}{N}\sum_{i=1}^{N} \|P\mathbf{x}_i - \vec{y}_i\|^2 \tag{6.4.1}$$

for $\mathbf{x} \in \mathbb{R}^{d_x}$, where $P \in \mathrm{Lip}(\mathbb{R}^d; \mathbb{R}^m)$ is any given surjective and non-zero map, which, in the context of regression, is simply a non-zero affine map, while in the context of binary classification, may be an affine map composed with a sigmoid nonlinearity.

To obtain stronger quantitative (and in fact, exponential) estimates, we will introduce a slightly different learning problem, inspired from results in optimal control theory. For fixed $\lambda > 0$, we will study the behavior when $T \gg 1$ of global minimizers to the functional

$$J_T(w, b) := \mathcal{E}(\mathbf{x}(T)) + \int_0^T \|\mathbf{x}(t) - \overline{\mathbf{x}}\|^2 \, \mathrm{d}t + \lambda \Big\| [w, b] \Big\|^2_{H^k(0,T;\mathbb{R}^{d_u})}, \tag{6.4.2}$$

with $\mathcal{E}$ as in (6.4.1), and where $\overline{\mathbf{x}}_i \in P^{-1}(\{\vec{y}_i\})$ for all $i \in [N]$ are given. Once again, $k = 0$ for (6.3.3) and $k = 1$ for (6.3.2) and $\mathbf{x} \in C^0([0,T]; \mathbb{R}^{d_x})$ is the unique solution to (6.3.3) or (6.3.2) corresponding to the parameters $[w, b] \in H^k(0, T; \mathbb{R}^{d_u})$, noting (6.3.1).

We note that, contrary to the case where we minimizing the training error at the final time T, here, the same scaling does not appear which allows us to deduce an equivalence with $\lambda \longrightarrow 0$. Hence, we will solely be interested in the behavior when $T \gg 1$.

We will require the following controllability definition, which is rather expected in the context of the result that follows.

Definition 6.4.1 (Simultaneous controllability with linear cost)**.** *We say that* (6.3.3) *(resp.* (6.3.2)*) is simultaneously controllable with linear cost in time $T > 0$ if for any $\mathbf{x}^0 \in \mathbb{R}^{d_x}$ and $\mathbf{x}^1 \in \mathbb{R}^{d_x}$, there exists a time $T > 0$ and parameters $[w, b] \in H^k(0, T; \mathbb{R}^{d_u})$ such that the corresponding unique solution $\mathbf{x}$ to* (6.3.3) *(resp.* (6.3.2)*) satisfies $\mathbf{x}(T) = \mathbf{x}^1$ and*

$$\Big\| [w, b] \Big\|_{H^k(0,T;\mathbb{R}^{d_u})} \leqslant \mathfrak{C}(T) \left\| \mathbf{x}^0 - \mathbf{x}^1 \right\| \tag{6.4.3}$$

holds for some $\mathfrak{C}(T) > 0$.

We refer to (6.4.3), as this is an estimate typically encountered in linear systems, and also, since the cost is linearly proportional to the distance of the initial data $\mathbf{x}^0$ to the target $\mathbf{x}^1$. We refer to Theorem 6.6 for further analysis regarding Definition 6.4.1.

We are in a position to state our main result in the context of the augmented supervised learning problem consisting of minimizing (6.4.2).

Theorem 6.5 (Exponential decay)**.** *Fix $\lambda > 0$, let $P \in \mathrm{Lip}(\mathbb{R}^d; \mathbb{R}^m)$ be any given non-zero and surjective map and let $\overline{\mathbf{x}} \in \mathbb{R}^{d_x}$ with $\overline{\mathbf{x}}_i \in P^{-1}(\{\vec{y}_i\})$ be arbitrary but fixed. Suppose that system* (6.3.3) *(resp.* (6.3.2) *with σ 1–homogeneous) is controllable in some time $T_0 > 0$ in the sense of Definition 6.4.1. Then, there exists $T^* > 0$ and positive constants $C_1, C_2, \mu > 0$ depending on $\lambda, \vec{x}_i, \vec{y}_i, N$ such that for any $T \geqslant T^*$, any parameters $[w_T, b_T] \in H^k(0, T; \mathbb{R}^{d_u})$ minimizing* (6.4.2)*, where $k = 0$ in the case of* (6.3.3)*, and $k = 1$ in the case of* (6.3.2) *and the corresponding unique solution $\mathbf{x}_T$ to* (6.3.2) *(resp.* (6.3.3)*) satisfy*

$$\left\| w_T(t) \right\| + \left\| b_T(t) \right\| \leqslant C_1 e^{-\mu t}$$

for a.e. $t \in [0, T]$ and

$$\mathcal{E}(\mathbf{x}_T(t)) + \|\mathbf{x}_T(t) - \overline{\mathbf{x}}\| \leqslant C_2 e^{-\mu t}$$

for all $t \in [0, T]$.

Theorem 6.5 is a specific manifestation of the so-called *turnpike property*, a paradigm dating back to the works of von Neumann [267], and works in economics by Samuelson et al. [82]. A local turnpike theory, combining the Pontryagin Maximum Principle, linearization arguments and precise estimates on Riccati equations, and covering a wide variety of nonlinear optimal control problems is developed in [261] – with extensions to Lipschitz nonlinearities and avoiding smallness conditions found in [96]. Theorem 6.5 can be proven by a small adaptation of the proofs presented in [96].

In Figure 6.1 – Figure 6.2, we depict[3] a manifestation of the exponential decay estimates insinuated by Theorem 6.5 on a toy binary classification task ($\vec{y}_i \in \{-1, +1\}$) with $N = 2400$ training samples and 600 test samples, where $P(\cdot) = \tanh(p_1 \, \cdot + p_2)$ with p_1, p_2 picked at random from a normal distribution (whilst ensuring that P is surjective). To discretize the full continuous-time optimization problem, we use direct shooting, which is a *first discretize then optimize* approach. We consider the neural ODE (6.3.3) with $\sigma(x) = \tanh(x)$ (we use the ResNet (6.2.3)), with $T = 15$ (and thus 15 layers) and $\lambda = 10^{-2}$. Finally, we discretize the integrals using an elementary trapezoidal quadrature. We note that the learned flow has a distinctly simple variation in Figure 6.2, and, albeit on a toy task, we observe satisfactory generalization properties in Figure 6.3.

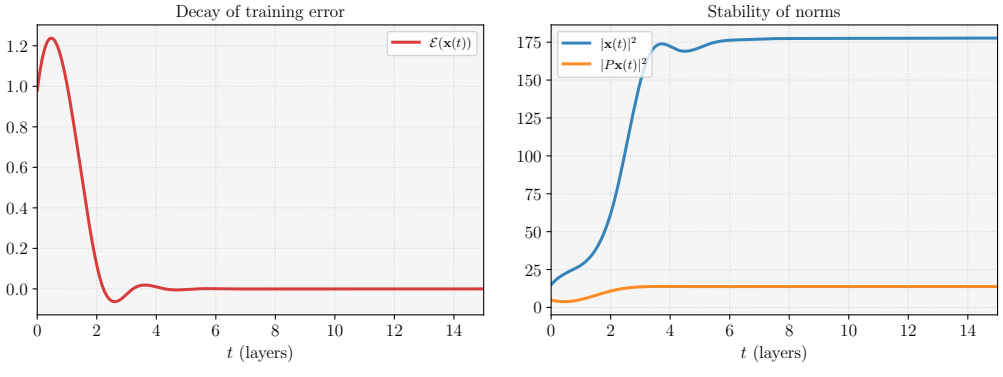


Figure 6.1: We visualize a manifestation of the exponential decay and stabilization results of Theorem 6.5 for the state $\mathbf{x}_T(t)$ (*right*) and the training error $\mathcal{E}(\mathbf{x}_T(t))$ (*left*) over $t \in [0, T]$. We observe that, after a finite time, the training error and trajectory remain at a steady configuration, so further times could be discarded from training.

The convergence rate entailed by Theorem 6.5 is not only noticeably stronger than that of Theorem 6.1 and Theorem 6.3, but the exponential estimate holds in any time $t \in [0, T]$ (i.e., at every layer when viewed from the discrete-time perspective) and not only for the output features. In fact, note that Theorem 6.5 is slightly different in nature to Theorem 6.1. This is because the integral term introduces a stronger time-scale in the

[3]Software experiments were done using `PyTorch` [214] (and may be found at `https://github.com/borjanG/dynamical.systems`), using the Adam optimizer [159] with learning rate equal to 10^{-3} and `TorchDiffEq` library [61]. Experiments were conducted on a personal MacBook Pro laptop (2.4 GHz Quad-Core Intel Core i5, 16GB RAM, Intel Iris Plus Graphics 1536 MB)

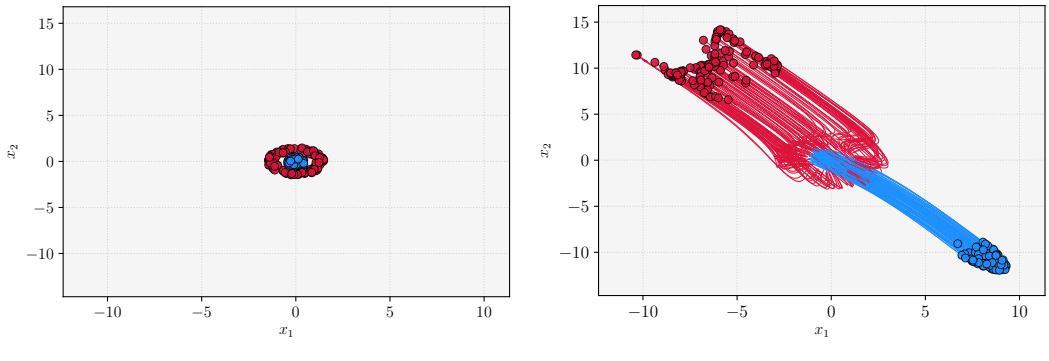


Figure 6.2: The training dataset (*left*) and the evolution of the trained neural ODE trajectories $\mathbf{x}_{T,i}(t)$ (*right*) in the phase plane – the learned flow is simple and varies little due to the exponentially small parameters, which ought to stipulate satisfactory generalization properties.

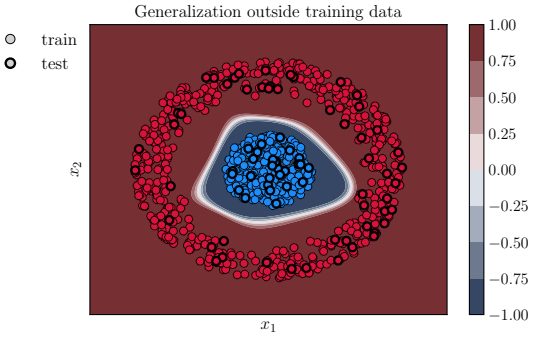


Figure 6.3: Trained classifier plot on $[-2,2]^2$ and evaluated on the test set; the simplicity of the learned flow indeed ensures satisfactory generalization as the shape of the dataset is learned adequately, and the test set is correctly classified.

behavior of the optimization problem as $T \longrightarrow +\infty$. To see this, consider the neural ODE (6.3.3) (whence $k = 0$) for simplicity, and as in (6.3.7),

$$\begin{aligned}
&\inf_{\substack{u_T \in L^2(0,T;\mathbb{R}^{d_u}) \\ \mathbf{x}_T \text{ solves (6.3.7)}}} \mathcal{E}(\mathbf{x}_T(T)) + \int_0^T \|\mathbf{x}_T(t) - \overline{\mathbf{x}}\|^2 \,\mathrm{d}t + \lambda \int_0^T \|u_T(t)\|^2 \,\mathrm{d}t \\
&= \inf_{\substack{u_T \in L^2(0,T;\mathbb{R}^{d_u}) \\ \mathbf{x}_T \text{ solves (6.3.7)}}} \mathcal{E}(\mathbf{x}_T(T)) + T \int_0^1 \left\|\mathbf{x}_T\left(\frac{s}{T}\right) - \overline{\mathbf{x}}\right\|^2 \,\mathrm{d}s + \frac{\lambda}{T} \int_0^1 \|T u_T(sT)\|^2 \,\mathrm{d}s \\
&= \inf_{\substack{u^1 \in L^2(0,1;\mathbb{R}^{d_u}) \\ \mathbf{x}^1 \text{ solves (6.3.8)}}} \mathcal{E}\left(\mathbf{x}^1(1)\right) + T \int_0^1 \|\mathbf{x}^1(s) - \overline{\mathbf{x}}\|^2 \,\mathrm{d}s + \frac{\lambda}{T} \int_0^1 \left\|u^1(s)\right\|^2 \,\mathrm{d}s. \qquad (6.4.4)
\end{aligned}$$

We see that, unlike Theorem 6.1, the integral term in (6.4.4) carries significance when $T \gg 1$, somewhat motivating the appearance of the exponential decay.

6.4.1 Motivating problem

Due to the specific nature of the proof of Theorem 6.5, which strongly relies on the fact that we may estimate the entire state $\mathbf{x}(t)$ via Grönwall arguments, we have restricted our study to an integral tracking term consisting of the squared $L^2(0,T;\mathbb{R}^{d_x})$–norm, albeit the final cost $\mathcal{E}(\mathbf{x}_T(T))$ allows us to study both classification and regression tasks. However, having to look for targets $\overline{\mathbf{x}}$ in the preimage of the labels $\vec{y}_i$ by P for any general task may not scale well computationally.

To alleviate this, at least numerically, we observe that the stabilization phenomenon for the output features (and also for the trajectories, although perhaps not with the same rate) persists when the term $\|\mathbf{x}(t) - \overline{\mathbf{x}}\|^2$ is replaced by the training error $\mathcal{E}(\mathbf{x}(t))$ with a general and possibly non-coercive loss, for instance, the cross-entropy loss on a multi-label classification tasks as seen in Figure 6.14 & Figure 6.17. In fact, we stipulate this stabilization phenomenon (be it exponential or not) to possibly hold for global minimizers of functionals of the form

$$J_T(w,b) := \int_0^T \mathcal{E}(\mathbf{x}(t)) \,\mathrm{d}t + \lambda \left\|[w,b]\right\|^2_{H^k(0,T;\mathbb{R}^{d_u})}, \qquad (6.4.5)$$

with $\mathcal{E}$ as in (6.3.4) and loss being continuous and nonnegative. We perform several numerical experiments to justify this claim.

Example 6.4.2 (Concentric spheres)**.** *Let us first depict the universality of the stabilization/turnpike property described by the estimates in Theorem 6.5 for the functional (6.4.5) on the concentric spheres dataset as above. We consider the same neural ODE (6.3.3), and this time, for variety, we consider ReLU activations (the same conclusions hold for* tanh*). We consider squared ℓ^2–loss in the training error $\mathcal{E}$ in (6.4.5), with the*

output layer having the form $Px = \tanh(p_1 x + p_2)$*, with* p_1, p_2 *both being part of the trainable parameters. We visualize the output of the experiments in Figure 6.4 – Figure 6.6 below.*

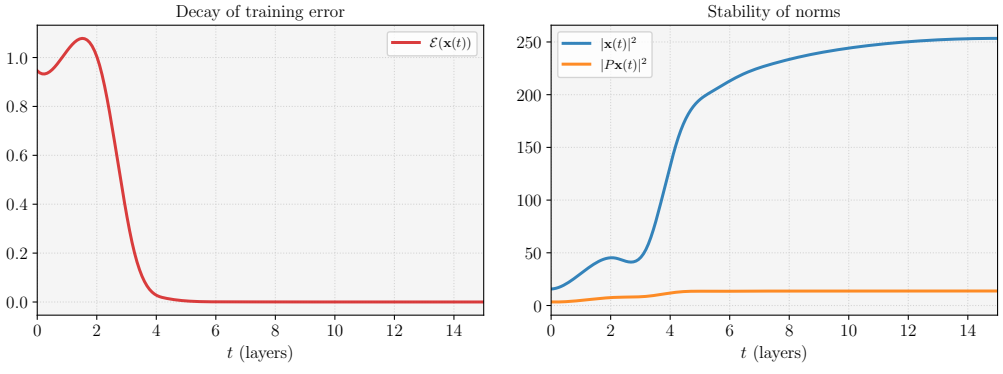


Figure 6.4: **Example 6.4.2:** The decay of the training error (*left*) and stabilization of optimal state trajectory (*right*) as stipulated by Theorem 6.5.

Example 6.4.3 (XOR)**.** *We now consider a binary classification task where a dimension augmentation of the input training data (as motivated by [86]) is beneficial – this is the case with the quintessensial XOR dataset (Figure 6.8) consisting of* $N = 3200$ *training samples and* 800 *test samples. We consider the same setup as in Example 6.4.2, and depict the output in Figure 6.7 – Figure 6.8, when the input data is immediately given by the inputs of the training dataset, and Figure 6.9 when we append a* 0 *to each input of the training dataset, and consider the dynamics evolving in* $\mathbb{R}^3$*. Whilst the trained neural ODE flow does separate both datasets in the phase space, a noticeable improvement in generalization capacity is observed in the augmented flow (Figure 6.10).*

Example 6.4.4 (Three labels)**.** *We now consider a toy multi-label classification task, with three labels, namely* $\vec{y}_i \in \{1, 2, 3\}$*, each label corresponding to a different color, consisting of* $N = 3200$ *training samples and* 800 *test samples. We consider the cross-entropy loss (7.1.3) in the training error in (6.4.5), and we only consider* L^2*–regularization of the parameters (instead of* H^1*). To further depict the universality of the stability phenomenon, we consider the neural ODE (6.2.4) – (6.2.7), where* $d_{\text{hid}} = 5$ *and* $\sigma = \tanh$*. The output layer is parametrized by* $Px = p_1 x + p_2$*, where* p_1, p_2 *are part of the trainable variables. We depict the results of the experiments in Figure 6.14 – Figure 6.16.*

Example 6.4.5 (MNIST)**.** *We finish this presentation by showing that the stabilization phenomenon may also be observed on more complex datasets such as MNIST [178].*

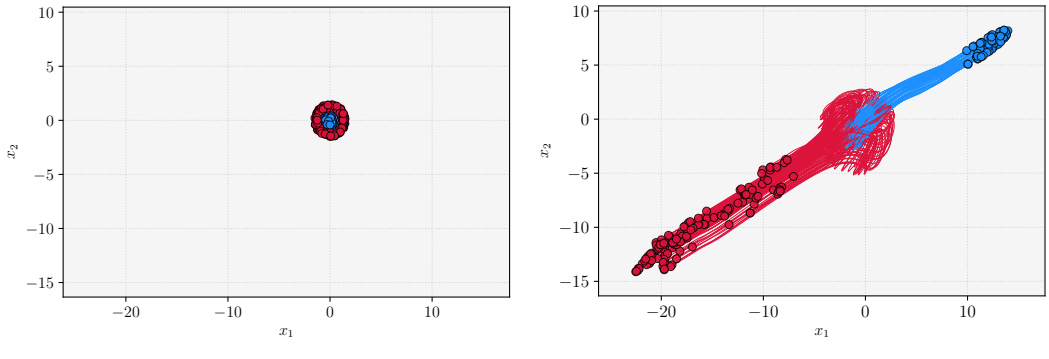


Figure 6.5: **Example 6.4.2:** The training dataset (*left*) and the evolution of the trained neural ODE trajectories $\mathbf{x}_{T,i}(t)$ (*right*) in the phase plane.

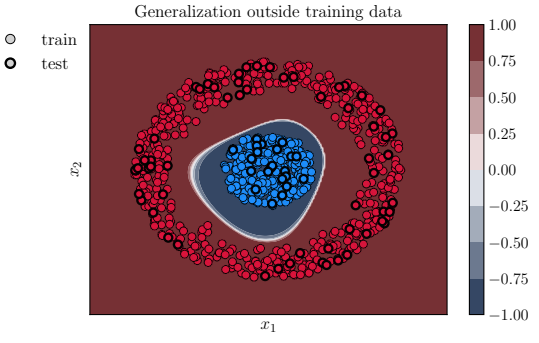


Figure 6.6: **Example 6.4.2:** Plot of the trained classifier on $[-2,2]^2$ and its evaluation on the test dataset; the learned flow ensures satisfactory generalization as the shape of the dataset is captures adequately.

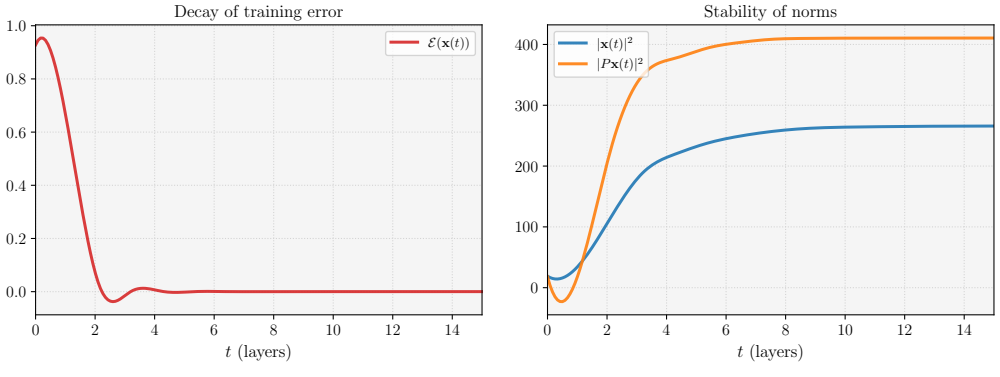


Figure 6.7: **Example 6.4.3:** The decay of the training error (*left*) and stabilization of optimal state trajectory (*right*) as stipulated by Theorem 6.5.

MNIST is a dataset consisting of handwritten digits from 0 *to* 9*, with a training set of 60000 samples, and a test set of 10000 samples. Each input sample* $\vec{x}_i$ *is a grayscale,* 28×28 *image of a handwritten digit, and thus an element of* $\mathbb{R}^{784}$*; the dataset has* 10 *labels:* $\vec{y}_i \in \{0, \dots, 9\}$*. We consider a similar setup as in Example 6.4.5 – the neural ODE is parametrized as* (6.2.4) – (6.2.7)*, where* $d_{\text{hid}} = 16$ *and* $\sigma = \tanh$*, and we consider cross-entropy loss in the training error in* (6.4.5) *and we only consider* L^2*–regularization of the parameters (instead of* H^1*), with* $T = 20$*, and the output layer is parametrized by* $Px = p_1 x + p_2$*, where* p_1, p_2 *are part of the trainable variables. We show the results of the experiments in Figure 6.17 – Figure 6.18.*

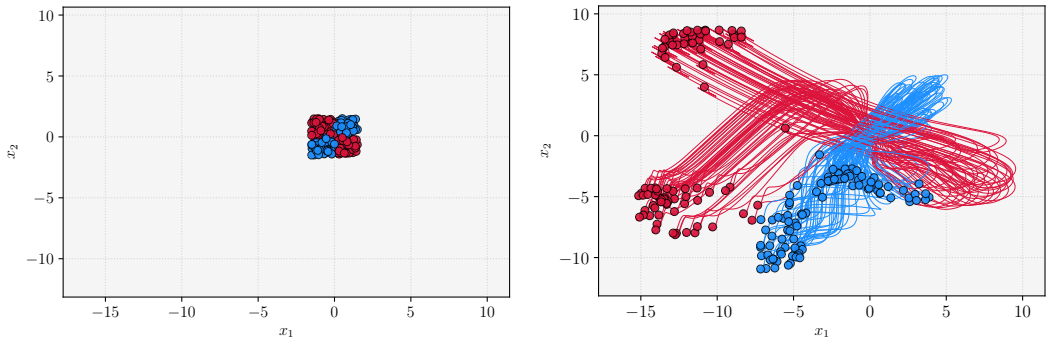


Figure 6.8: **Example 6.4.3:** The training dataset (*left*) and the evolution of the trained neural ODE trajectories $\mathbf{x}_{T,i}(t)$ (*right*) in the phase plane in $\mathbb{R}^2$.

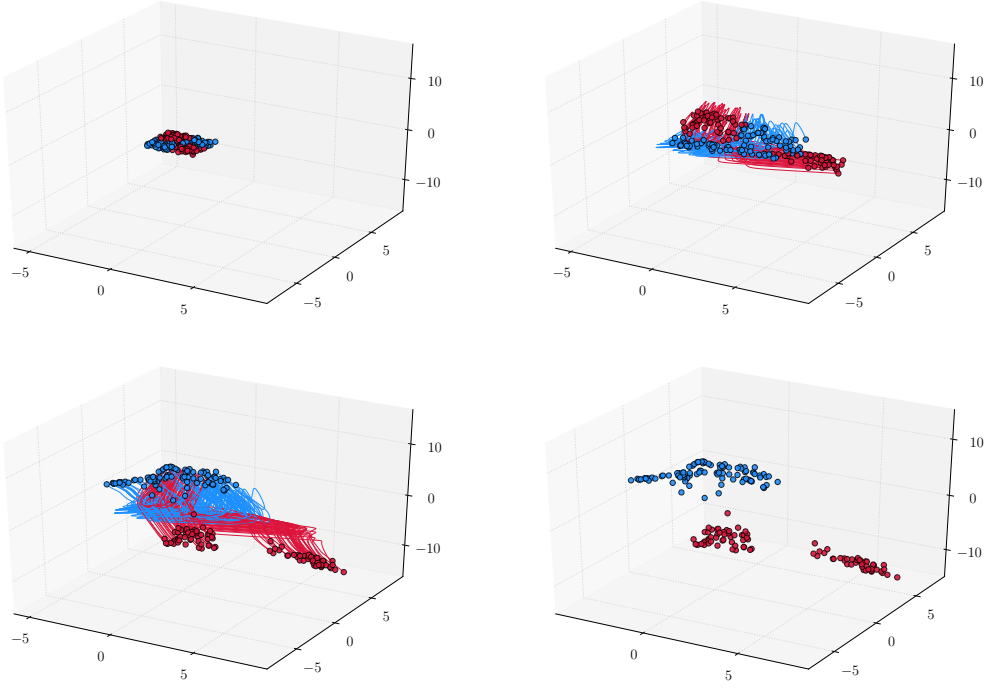


Figure 6.9: **Example 6.4.3:** The training dataset (*top left*) and the evolution of the trained neural ODE trajectories $\mathbf{x}_{T,i}(t)$ for $t \leqslant 5$ (*top right*) and $t \leqslant T$ (*bottom left*) in the phase space in $\mathbb{R}^3$, with the separated features $\mathbf{x}_{T,i}(T)$ (*bottom right*).

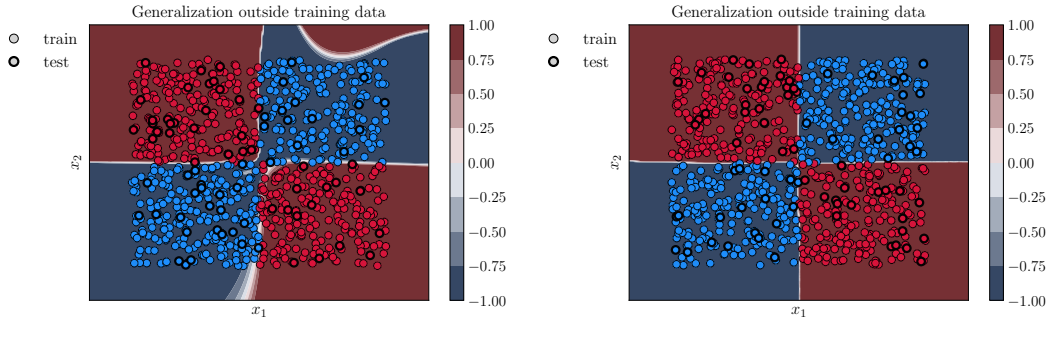


Figure 6.10: **Example 6.4.3:** Plot of the trained classifier via the neural ODE flow evolving in $\mathbb{R}^2$ (*left*) and in $\mathbb{R}^3$ (*right*) on $[-2.5, 2.5]^2$ and its evaluation on the test dataset. We see that the learned flow captures the shape of the dataset adequately in both cases, but with a slightly more satisfactory accuracy when the input data is augmented.

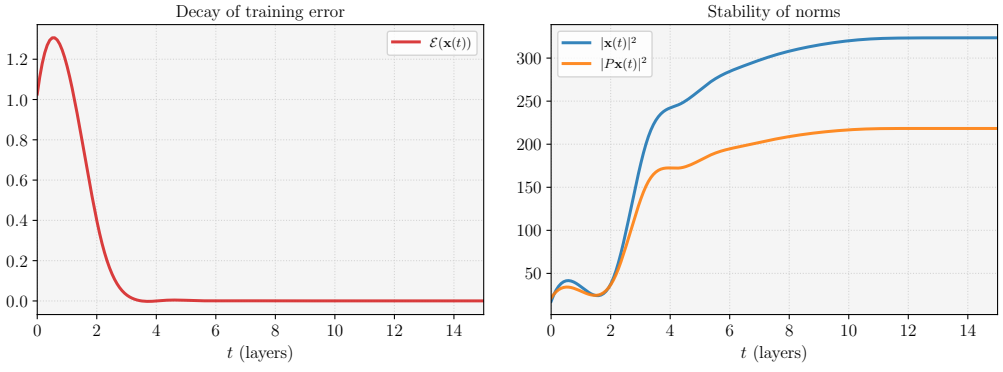


Figure 6.11: **Example 6.4.4:** The decay of the training error (*left*) and stabilization of optimal state trajectory (*right*) as stipulated by Theorem 6.5.

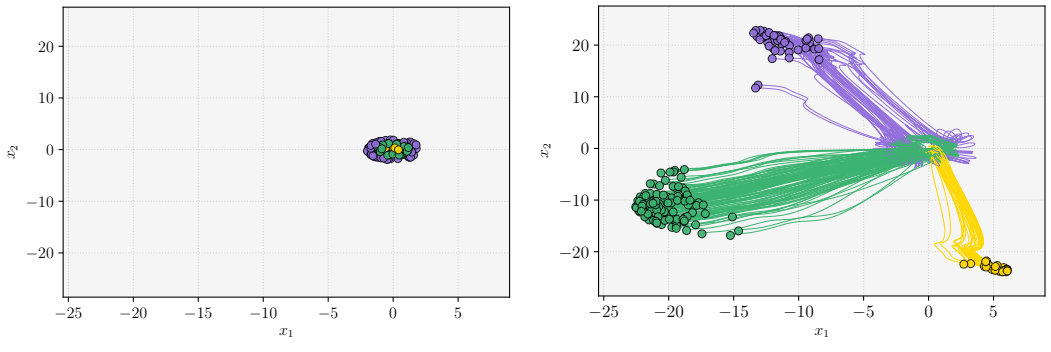


Figure 6.12: **Example 6.4.4:** The training dataset (*left*) and the evolution of the trained neural ODE trajectories $\mathbf{x}_{T,i}(t)$ (*right*) in the phase plane.

6.4.2 On Definition 6.4.1

The majority of our results stated in the preceding sections stipulate whether and how the neural network prediction approaches the zero training error regime ($\mathcal{E} = 0$ with $\mathcal{E}$ given in (6.3.4)) when the number of layers increases. It is thus of interest to also illuminate the properties of the parameters which allow the neural network prediction to reach precisely a minimizer of the training error $\mathcal{E}$.

To complete this section, we state the following controllability result, which namely contains an estimate on the control with respect to the distance of the target and the initial datum, which somewhat enhances the validity of the controllability assumption we make in Theorem 6.1. While such an estimate is standard in the linear systems setting, it is not provided by sufficient controllability conditions for nonlinear systems such as the Chow-Rashevski theorem [69, Chapter 3, Section 3.3].

Theorem 6.6. *Let $T > 0$ and assume that $N \leqslant d$. Let $\mathbf{x}^1 \in \mathbb{R}^{d_x}$ be given, and assume that the activation function $\sigma \in C^1(\mathbb{R}) \cap \mathrm{Lip}(\mathbb{R})$ is such that*

$$\left\{\sigma\left(\mathbf{x}_1^1\right), \ldots, \sigma\left(\mathbf{x}_i^1\right), \ldots, \sigma\left(\mathbf{x}_N^1\right)\right\}$$

is a system of linearly independent vectors in $\mathbb{R}^d$. Then, there exist universal constants $r > 0$ and $\mathfrak{C} > 0$ such that for any datum $\mathbf{x}^0 \in \mathbb{R}^{d_x}$ satisfying $\|\mathbf{x}^0 - \mathbf{x}^1\| \leqslant r$, there exists a weight matrix $w \in L^\infty(0, T; \mathbb{R}^{d \times d})$ such that the unique solution $\mathbf{x}$ to

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{w}(t)\sigma(\mathbf{x}(t)) & \text{in } (0, T) \\ \mathbf{x}(0) = \mathbf{x}^0, \end{cases}$$

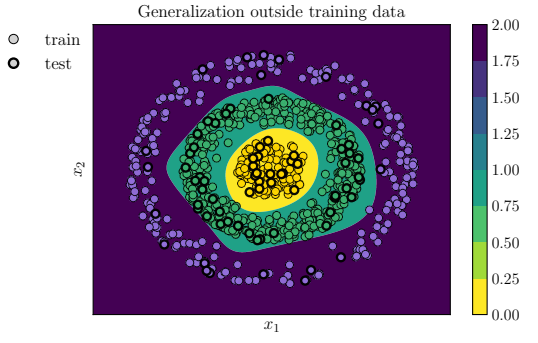


Figure 6.13: **Example 6.4.4:** Plot of the trained classifier on $[-2.5, 2.5]^2$ and its evaluation on the test dataset; the learned flow ensures satisfactory generalization as the shape of the dataset is captures adequately.

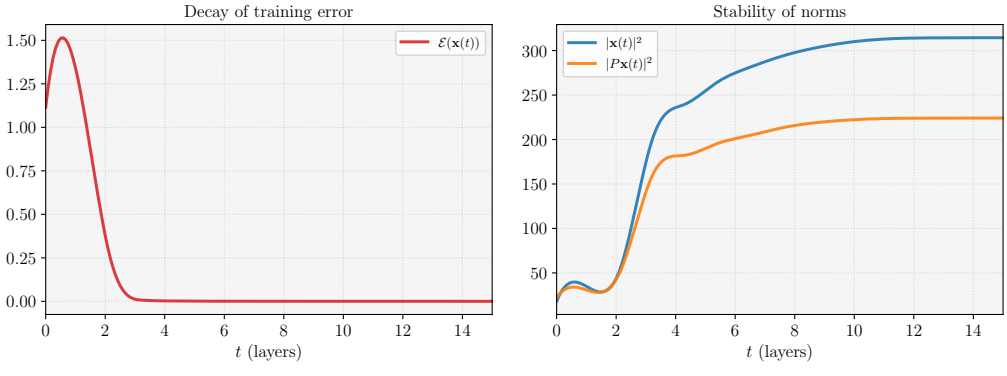


Figure 6.14: **Example 6.4.4:** The decay of the training error (*left*) and stabilization of optimal state trajectory (*right*) as stipulated by Theorem 6.5.

satisfies

$$\mathbf{x}(T) = \mathbf{x}^1,$$

and the following estimate holds

$$\|u\|_{L^\infty(0,T;\mathbb{R}^{d_u})} \leqslant \frac{\mathfrak{C}}{T} \left\|\mathbf{x}^0 - \mathbf{x}^1\right\|.$$

Remark 6.4.6. *The following observations are in order.*

- *For simplicity of presentation, we have not exhibited the bias parameter, namely the additive time-dependent control b. One can readlily check that, in the presence of this additional control, the assumption $N \leqslant d$ can be relaxed to $N \leqslant d + 1$.*

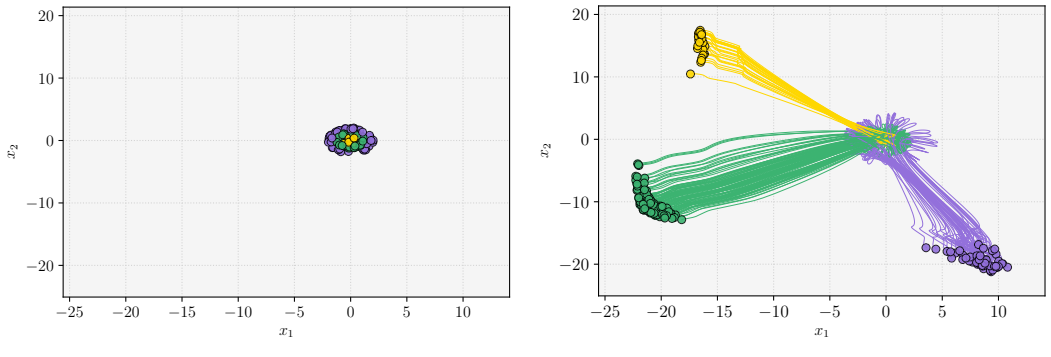


Figure 6.15: **Example 6.4.4:** The training dataset (*left*) and the evolution of the trained neural ODE trajectories $\mathbf{x}_{T,i}(t)$ (*right*) in the phase plane.

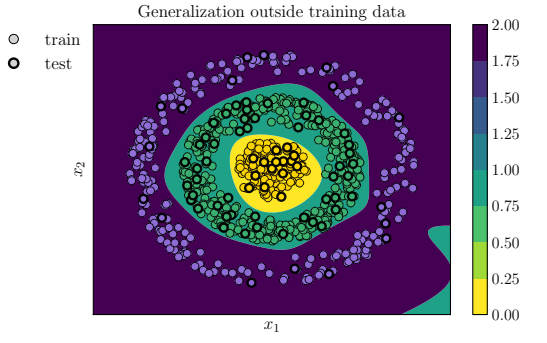


Figure 6.16: **Example 6.4.4:** Plot of the trained classifier on $[-2.5, 2.5]^2$ and its evaluation on the test dataset; the learned flow ensures satisfactory generalization as the shape of the dataset is captures adequately.

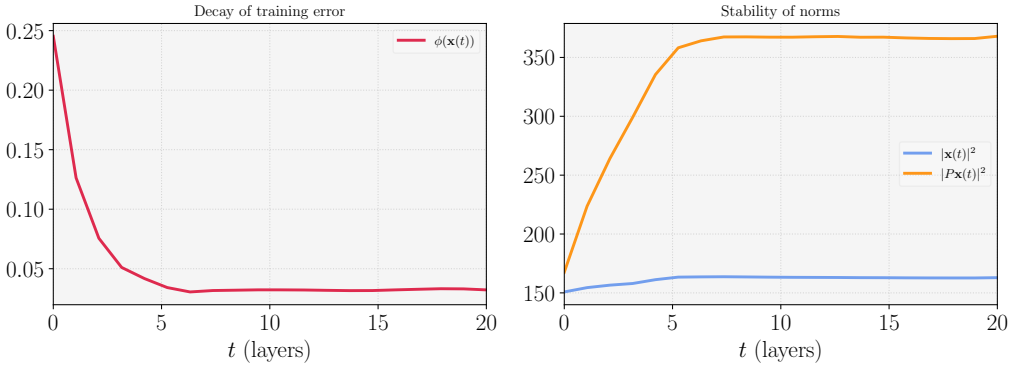


Figure 6.17: **Example 6.4.5:** The decay of the training error (*left*) and stabilization of optimal state trajectory (*right*) as stipulated by Theorem 6.5.

- *One could perhaps adapt the argument in the proof of Theorem 6.6 (given just below) to obtain a global result, assuming the existence of a continuous arc γ linking $\mathbf{x}^0$ and $\mathbf{x}^1$, such that*

$$\left\{\sigma\left(\gamma_1(s)\right), \dots, \sigma\left(\gamma_i(s)\right), \dots, \sigma\left(\gamma_N(s)\right)\right\}$$

  is a system of linearly independent vectors in $\mathbb{R}^d$ for any $s \in [0,1]$. Problems arise however whenever this condition is not satisfied. In any case, in view of the uniqueness results for ODEs and Proposition 6.3.2, we have to assume that $\mathbf{x}_i^0 \neq \mathbf{x}_j^0$ and $\mathbf{x}_i^1 \neq \mathbf{x}_j^1$, for $i \neq j$.

- *The case $N > d+1$ may be treated by linearizing around a non-steady trajectory. Note that in [69, Section 3.1, Theorem 3.6], the controllability of the linearized problem around a general trajectory suffices.*

In the discrete-time context of neural networks such as (6.2.1) or (6.2.3), the property analog to Definition 6.3.1 is also well explored in the literature, and is commonly called *finite sample expressivity* [276]. An additional interest is that of estimating the number of parameters – referred to as *the memorization capacity* – needed to manifest this property. For instance, in [276], the authors use an MLP with ReLU activations with two layers and $2N + d$ parameters to interpolate any labeling of size N in d dimensions. Their network inevitably has large width, but a network of depth $N_{\text{layers}} \geqslant 2$ can be conceived, in which each individual layer has only $\mathcal{O}\left(NN_{\text{layers}}^{-1}\right)$ parameters. For additional results, we refer the reader to [192, 274, 209].

In the ODE context, the property of finite sample expressivity finds its analog in the *complete* or *simultaneous controllability*, wherein one requires only 1 pair of controls to steer N trajectories of the same system to N prescribed targets – this is the property

Figure 6.18: **Example 6.4.5:** We depict the evolution of three individual samples $\mathbf{x}_i(t) \in \mathbb{R}^{784}$ at times $t \in \{0, 2, 8, 15, 19\}$ We see that each trajectory stabilizes to some steady configuration after time $t \geqslant 8$; curiously enough, the neural ODE tends to compress the input digit samples ahead of classifying them via the softmax applied to $P\mathbf{x}_i(t) \in \mathbb{R}^{10}$.

we show in Theorem 6.6. There have been some works on such controllability results of neural ODEs, mostly relying on geometrical techniques such as Lie brackets techniques (see [69, Chapter 3, Section 3.3]), under specific constraints on the activations function (see e.g. [75, 253]). We refer to [237] for further results in this direction.

6.5 Continuous space-time neural networks

We now come back to the scheme (6.2.3) defining a ResNet with $N_{\text{layers}} \geqslant 2$ layers. Whilst such networks are widely used in practice and provide reliable results, in the discrete-time context, they do not take into account variations of the dimensions of the weights and states over layers. Such variations may arise when considering *convolutional* and/or *pooling* layers, which are ubiquitous in tasks in computer vision. In such tasks, it is moreover of interest to view the data itself as being continuum objects.

To be more specific, we note that in the simplest nonlinear context, a residual network with variable dimensions analog to (6.2.3) takes the form (see [140])

$$\begin{cases} \mathbf{x}_i^{k+1} = P^k \mathbf{x}_i^k + \sigma(w^k \mathbf{x}_i^k + b^k) & \text{for } k \in \{0, \dots, N_{\text{layers}} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i. \end{cases} \tag{6.5.1}$$

Here, contrary to (6.2.3), we have $w^k \in \mathbb{R}^{d_{k+1} \times d_k}$ and $b^k \in \mathbb{R}^{d_{k+1}}$, and thus $\mathbf{x}^k \in \mathbb{R}^{d_k}$ for $k \in \{0, \dots, N_{\text{layers}}\}$, where $\{d_k\}_{k=0}^{N_{\text{layers}}}$ are given positive integers, called widths of the layers k. One imposes $d_0 = d$, and $P^k \in \mathbb{R}^{d_{k+1} \times d_k}$ is a projection/embedding operator which serves to match dimensions. Much like in the fixed width case, we may also write the variable-width ResNet when g is parametrized as in (6.2.6) or otherwise.

The continuous space-time network. It is not immediately obvious how one can see (6.5.1) as a numerical scheme for some continuous-time dynamical system in the flavor of (6.2.4). Nevertheless, this can be achieved by viewing the changing dimension over time-steps as an additional (spatial) variable, thus yielding an integro-differential equation in the continuum.

To be more precise, for any $i \in [N]$ we consider the scalar integro-differential equation

$$\begin{cases} \partial_t \mathbf{x}_i(t,x) = \sigma\left(\int_\Omega w(t,x,\xi)\mathbf{x}_i(t,\xi)\,\mathrm{d}\xi + b(t,x)\right) & \text{for } (t,x) \in (0,T)\times\Omega \\ \mathbf{x}_i(0,x) = \mathbf{x}_i^{\mathrm{in}}(x) & \text{for } x \in \Omega. \end{cases} \tag{6.5.2}$$

Here $\Omega \subset \mathbb{R}^{d_\Omega}$ is a bounded domain, where $d_\Omega \geqslant 1$. We emphasize that $\mathbf{x}_i(t,x) \in \mathbb{R}$ for $(t,x) \in (0,T)\times\Omega$, and similarly, $w(t,x,\xi) \in \mathbb{R}$ and $b(t,x) \in \mathbb{R}$ for $(x,\xi) \in \Omega\times\Omega$. The initial datum $\mathbf{x}_i^{\mathrm{in}} \in C^0(\overline{\Omega})$ is such that there exist $\{x_j\}_{j=1}^d \subset \Omega$ such that $\mathbf{x}_i^{\mathrm{in}}(x_j) = (\vec{x}_i)_j$. Such a datum can always be found (e.g. by interpolation). The continuum model (6.5.2) is proposed in [188] where well-posedness is established, and is also suggested in [89] albeit in a slightly different context. We distinguish two typical cases for choosing the shape of Ω as well as d_Ω.

- **Variable-width ResNets.** If in the discretized level, we seek to simply obtain a variable-width residual network such as (6.5.1) (or even the standard ResNet analog (6.2.3)), it suffices to consider $\Omega = (0,1)$, thus $d_\Omega = 1$. We give more detail on possible possible discretizations in Section 6.5.2 and Remark 6.5.1.

- **Convolutional Neural Networks.** The situation is slightly more delicate in the case of CNNs, which are typically used in computer vision. We provide a proposal covering the continuous-time analog of CNNs with partial generality.

  Assume that the dataset $\{\vec{x}_i\}_{i=1}^N$ consists of N images: $\vec{x}_i \in \mathbb{R}^{d_1\times d_2\times d_{\mathrm{ch}}}$ for any i; here d_1 (resp. d_2) denote the number of horizontal (resp. vertical) pixels in the image $\vec{x}_i$, whereas d_{ch} denotes the number of channels, i.e. the color format (e.g. $d_{\mathrm{ch}} = 1$ for grayscale, $d_{\mathrm{ch}} = 3$ for RGB). In this case, we consider $\Omega := \Omega_{\mathrm{img}}\times(0,1)$, where $\Omega_{\mathrm{img}} \subset \mathbb{R}^2$ is a rectangle. Thus $d_\Omega = 3$. Moreover, we assume that the weights w in (6.5.2) are compactly supported and of a specific *convolutional* form (as indicated in most works, this is more so a *cross-correlation* form), namely, for any i, the equation takes the form

  $$\partial_t \mathbf{x}_i(t,x,\zeta) = \sigma\left(\int_0^1 \int_{\Omega_{\mathrm{img}}} w(t,x+\xi,\omega)\mathbf{x}_i(t,\xi,\omega)\,\mathrm{d}\xi\,\mathrm{d}\omega + b(t,x,\zeta)\right)$$

  for $(t,x,\zeta) \in (0,T)\times\Omega_{\mathrm{img}}\times(0,1)$. We note that the variable $x \in \Omega_{\mathrm{img}}$ denotes a pixel, whereas $\zeta \in (0,1)$ is a continuous variable indicating, when discretized, the number of extracted features (namely the number of filters). The bias parameter b can be omitted in this case, if desired.

  One possible way to discretize the above continuous-time model and obtain a CNN-ResNet as in [140] is to follow the arguments in Section 6.5.2, where one would use a time-dependent grid for discretizing with respect to the variable $\zeta \in (0,1)$ as well, as the number of filters commonly varies over layers in CNNs. By discretizing Ω_{img} with a "shrinking" or "expanding" time-dependent rectangular grid, some effects of padding or pooling (but not max-pooling a priori) may also be considered. However, a full CNN–applicable theory is out of the scope of this work.

The mathematical theory of structural properties of CNNs is well-established – for instance, [197, 34, 198] provide, via a concept of Lipschitz stability to the action of diffeomorphisms, a characterization of of invariance and stabilitiy properties of input images, shown by using the so-called scattering transform, based on microlocal analysis techniques. They in particular define explicitly the weight kernels w by means of specific wavelets motivated by the fact that CNNs are specifically designed to exploit the prior properties of image data, and thus no optimization is involved. This differs significantly from the commonly used CNNs however, which adapt filters to training data by optimization.

Remark 6.5.1. *Observe that the continuous space-time model (6.5.2) (resp. (6.5.3)) is more general and englobes (6.2.4) – (6.2.5) (resp. (6.2.4) – (6.2.6)), where only the time variable is considered to be continuous. Indeed, fix d different points $\{x_1, \dots, x_d\} \in \Omega$, and let δ_{x_j} denote the Dirac mass centered at x_j. For any $i \in [N]$, we consider the initial datum*

$$\mathbf{x}_i^{\text{in}}(x) := \sum_{j=1}^{d} (\vec{x}_i)_j \, \delta_{x_j}(x) \qquad \text{for } x \in \Omega.$$

We write the weight w as

$$w(t,x,\zeta) := \sum_{j=1}^{d}\sum_{\ell=1}^{d} w_{j,\ell}(t)\delta_{x_j}(x)\delta_{x_\ell}(\zeta) \qquad \text{for } (t,x,\zeta) \in (0,T)\times\Omega\times\Omega,$$

yielding the matrix $[w_{j,\ell}(t)]_{1\leqslant j,\ell\leqslant d}$ of weights at time t, whereas the bias $b(t,x)$ is written as

$$b(t,x) := \sum_{j=1}^{d} b_j(t)\delta_{x_j}(x) \qquad \text{for } (t,x) \in (0,T)\times\Omega,$$

yielding the vector $[b_j(t)]_{1\leqslant j\leqslant d}$ of biases at time t. As $\mathbf{x}_i^{\text{in}}$, w and b are all linear combinations of Dirac masses, by plugging them in (6.5.2), we rewrite the integrals as sums, and setting, for any $i \in [N]$,

$$(\mathbf{x}_i)_j(t) := \int_\Omega \mathbf{x}_i(t,x)\,\mathrm{d}\delta_{x_j}(x)$$

for $j \in [d]$, we see that $(\mathbf{x}_i)_j$ solves

$$\begin{cases} (\dot{\mathbf{x}}_i)_j(t) = \sigma\left(\displaystyle\sum_{\ell=1}^{d} w_{j,\ell}(t)\,(\mathbf{x}_i)_\ell(t) + b_j(t)\right) & \text{for } t \in (0,T) \\ (\mathbf{x}_i)_j(0) = (\vec{x}_i)_j. \end{cases}$$

This is just the j–th equation of the (6.2.4) – (6.2.5) for $i \in [N]$.

Remark 6.5.2. *Correspondingly for $i \in [N]$ we may consider*

$$\begin{cases} \partial_t \mathbf{x}_i(t,x) = \displaystyle\int_0^1 w(t,x,\xi)\sigma(\mathbf{x}_i(t,\xi))\,\mathrm{d}\xi + b(t,x) & \text{in } (0,T)\times\Omega \\ \mathbf{x}_i(0,x) = \mathbf{x}_i^{\text{in}}(x) & \text{in } \Omega. \end{cases} \tag{6.5.3}$$

All of the above discussions also apply for this system.

6.5.1 The supervised learning problem

Given a training dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ with $\vec{x}_i \in \mathbb{R}^d$ and $\vec{y}_i \in \mathbb{R}^m$ for any i, and a time horizon $T > 0$, just as in the finite dimensional context, we begin by writing the equation satisfied by the stacked vector of states $\mathbf{x} := [\mathbf{x}_1, \dots, \mathbf{x}_N]$ corresponding to the stacked vector of data $\mathbf{x}^{\text{in}} := [\mathbf{x}_1^{\text{in}}, \dots, \mathbf{x}_N^{\text{in}}]$, where each $\mathbf{x}_i$ is the solution to either (6.5.2) or (6.5.3) corresponding to the datum $\mathbf{x}_i^{\text{in}}$, and control parameters $[w, b]$ which are the same for all i. The stacked continuous space-time neural networks we consider are thus either

$$\begin{cases} \partial_t \mathbf{x}(t,x) = \sigma\left(\displaystyle\int_\Omega \mathbf{w}(t,x,\xi)\mathbf{x}(t,\xi)\,\mathrm{d}\xi + \mathbf{b}(t,x)\right) & \text{in } (0,T)\times\Omega \\ \mathbf{x}(0,x) = \mathbf{x}^{\text{in}}(x) & \text{in } \Omega \end{cases} \tag{6.5.4}$$

or

$$\begin{cases} \partial_t \mathbf{x}(t,x) = \displaystyle\int_\Omega \mathbf{w}(t,x,\xi)\sigma(\mathbf{x}(t,\xi))\,\mathrm{d}\xi + \mathbf{b}(t,x) & \text{in } (0,T)\times\Omega \\ \mathbf{x}(0,x) = \mathbf{x}^{\text{in}}(x) & \text{in } \Omega. \end{cases} \tag{6.5.5}$$

Just as in the finite-dimensional case, the key point is to note how the controls $[w(t,x,\xi), b(t,x)]$ for $(t,x,\xi) \in (0,T) \times \Omega \times \Omega$ enter the systems:

$$\mathbf{w}(t,x,\xi) := \begin{bmatrix} w(t,x,\xi) & & \\ & \ddots & \\ & & w(t,x,\xi) \end{bmatrix} \in \mathbb{R}^{N\times N}, \qquad \mathbf{b}(t,x) := \begin{bmatrix} b(t,x) \\ \vdots \\ b(t,x) \end{bmatrix} \in \mathbb{R}^N. \tag{6.5.6}$$

Empirical risk minimization

As before, we first consider the consider the regularized empirical risk minimization problem

$$\inf_{\substack{[w,b]\in H^k(0,T;\mathfrak{U}) \\ \text{subject to}(6.5.4)\ (\text{resp.}(6.5.5))}} \mathcal{E}(\mathbf{x}(T)) + \lambda \Big\| [w,b] \Big\|^2_{H^k(0,T;\mathfrak{U})}, \tag{6.5.7}$$

where $\alpha > 0$ is fixed, $k = 0$ for (6.5.5) and $k = 1$ for (6.5.4),

$$\mathfrak{U} := L^2(\Omega \times \Omega) \times L^2(\Omega),$$

and we define the training error as (we concentrate[4] on L^2–loss):

$$\mathcal{E}(\mathbf{x}(T)) := \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i(T) - g(\vec{y}_i)\|^2_{L^2(\Omega)}, \tag{6.5.8}$$

where $g : \mathcal{Y} \longrightarrow L^2(\Omega)$ is arbitrary, but fixed. We note that, due to the fact that we are working with functions as inputs and labels, we do not require an output layer P which matches dimensions as in the finite-dimensional case. The optimization problem (6.5.7) admits a solution – the argument follows the same lines as the proof of Proposition 6.2.1.

In view of the rather universal nature of the proof to Theorem 6.1 and Theorem 6.5 in the finite-dimensional case, one may in fact roughly repeat the exact same proofs at most points, replacing throughout the finite dimensional euclidean spaces $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_u}$, by $L^2(\Omega)^N$ and $\mathfrak{U}$ respectively. Whence, we state the infinite-dimensional (partial) analog to Theorem 6.1.

Theorem 6.7. *Let $\lambda > 0$ be fixed, let $\mathbf{x}^{\text{in}} \in \big(C^0(\overline{\Omega})\big)^N$ be such that $\mathbf{x}^{\text{in}}_i(x_j) = (\vec{x}_i)_j$, and assume that (6.5.4) (resp. (6.5.5) with σ positively homogeneous of degree 1) interpolates the set $\{\mathbf{x}^{\text{in}}_i, y^{\text{out}}_i\}_{i=1}^N$ in some time $T_0 > 0$. For any $T > 0$, let $\mathbf{x}_T \in C^0\big([0,T]; L^2(\Omega)^N\big)$ be the unique solution to (6.5.4) (resp. (6.5.5) with σ positively homogeneous of degree 1), associated a global minimizer $u_T := [w_T, b_T] \in H^k(0,T;\mathfrak{U})$ of the functional in (6.5.7), where $k = 0$ in the case of (6.5.5) and $k = 1$ in the case of (6.5.4). The following properties then hold.*

1. *There exists a constant $C = C(\mathbf{x}^{\text{in}}, \vec{y}, \lambda) > 0$ independent of T such that*

$$\mathcal{E}(\mathbf{x}_T(T)) \leqslant \frac{C}{T}.$$

2. *There exists a sequence $\{T_n\}_{n=1}^{+\infty}$, with $T_n > 0$ and $T_n \xrightarrow[n\longrightarrow+\infty]{} +\infty$, and some $\mathbf{x}_\circ \in L^2(\Omega)^N$ with $\mathcal{E}(\mathbf{x}_\circ) = 0$ such that, along a subsequence,*

$$\mathcal{E}\left(\mathbf{x}_{T_n}(T_n)\right) \longrightarrow 0$$

[4] We do this choice for simplicity of presentation of the continuous space-time model and since we only present the results analog to the ℓ^2–loss in the neural ODE setting. One can define the functional version of classification losses such as cross-entropy by rather working with probability measures instead of L^2 data, or, sticking to binary classification tasks, apply a sigmoid nonlinearity to $\mathbf{x}_i(T)$. We leave these cases to the interested reader.

and

$$\mathbf{x}_{T_n}(T_n) \rightharpoonup \mathbf{x}^* \qquad \text{weakly in } L^2(\Omega)^N$$

as $n \longrightarrow +\infty$.

For the sake of completeness, we give a sketch of the proof – by indicating the only changes with respect to that of Theorem 6.1.

Proof of Theorem 6.7. We note that the infinite-dimensional analog of Lemma 7.2.1 may easily be shown to hold, and one may readily repeat precisely the same arguments as in the proof of Theorem 6.1, replacing $\mathbb{R}^{d_u}$ and $\mathbb{R}^{d_x}$ by $\mathfrak{U}$ and $L^2(\Omega)$ respectively throughout. The only difference occurs in regarding the arguments on strong L^2–convergence of the sequence of controls in the case $k = 1$ – in the infinite dimensional case, we may exhibit the Aubin-Lions compactness lemma instead of Rellich-Kondrachov to conclude. □

Augmented empirical risk minimization

We similarly consider the augmented supervised learning problem

$$\inf_{\substack{[w,b]\in H^k(0,T;\mathfrak{U}) \\ \text{subject to (6.5.4) (resp.(6.5.5))}}} \int_0^T \mathcal{E}(\mathbf{x}(t))\,\mathrm{d}t + \lambda \Big\| [w,b] \Big\|^2_{H^k(0,T;\mathfrak{U})}, \tag{6.5.9}$$

where $\mathcal{E}$ is as in (6.5.8) (note that, since there is no output layer P, and we consider L^2–loss, we may consider integrating the training error straight away). As expected, the analog exponential decay result holds for (6.5.9).

Theorem 6.8. *Fix $\lambda > 0$ and let $\mathbf{x}^{\mathrm{in}} \in C^0(\overline{\Omega})^N$. Assume that (6.5.5) (resp. (6.5.4)) is controllable in some time $T_0 > 0$ with linear cost.Then, there exists $T^* > 0$ and positive constants $C_1, C_2, \mu > 0$ depending on $\lambda, \mathbf{x}^{\mathrm{in}}, \vec{y}, N$, such that for any $T \geqslant T^*$, any parameters $[w_T, b_T] \in H^k(0,T;\mathfrak{U})$ solving the minimization problem (6.5.9), where $k = 0$ in the case of (6.5.5) and $k = 1$ in the case of (6.5.4), and the corresponding unique solution $\mathbf{x}_T \in C^0\big([0,T]; L^2(\Omega)^N\big)$ to (6.5.5) (resp. (6.5.4)), satisfy*

$$\|w_T(t)\|_{L^2(\Omega\times\Omega)} + \|b_T(t)\|_{L^2(\Omega)} \leqslant C_1 e^{-\mu t}$$

for a.e. $t \in [0,T]$ and

$$\mathcal{E}(\mathbf{x}_T(t)) \leqslant C_2 e^{-\mu t}$$

for all $t \in [0,T]$.

The proof is omitted and left to the reader, as it follows precisely the same arguments as that of Theorem 6.5.

6.5.2 From continuous to discrete

The passage from (6.5.2) to a discrete-time scheme such as (6.5.1) is not immediately obvious, and to our knowledge has not been presented in the literature. To proceed, it is important to observe the inherent link between the layer k and the width d_k in (6.5.1). This motivates discretizing (6.5.2) in the spatial variable $x \in (0,1)$ by using a *time-dependent grid*, which has a different number of nodes d_k at each time-step. We give more detail on this in what follows.

Let us demonstrate that (6.5.2) which reads[5] (we omit the dependence on i for clarity)

$$\begin{cases} \partial_t \mathbf{x}(t,x) = \sigma\left(\displaystyle\int_0^1 w(t,x,\xi)\mathbf{x}(t,\xi)\,\mathrm{d}\xi + b(t,x)\right) & \text{in } (0,T)\times(0,1) \\ \mathbf{x}(0,x) = \mathbf{x}^{\mathrm{in}}(x) & \text{in } (0,1), \end{cases}$$

[5]The choice of the spatial interval $[0,1]$ is completely arbitrary – one may of course consider any bounded interval of $\mathbb{R}$.

where $\mathbf{x}^{\text{in}}$ is such that $\mathbf{x}^{\text{in}}(x_j) = \vec{x}_{,j}$ for some $\{x_j\}_{j=1}^{d} \subset [0,1]$, can be discretized to read exactly as

$$\begin{cases} \mathbf{x}^{k+1} = P^k \mathbf{x}^k + \sigma\left(w^k \mathbf{x}^k + b^k\right) & \text{for } k \in \{0, \dots, N_{\text{layers}} - 1\} \\ \mathbf{x}^0 = \vec{x}. \end{cases} \tag{6.5.10}$$

Here $\mathbf{x}^k \in \mathbb{R}^{d_k}$, $w^k \in \mathbb{R}^{d_{k+1} \times d_k}$ and $b^k \in \mathbb{R}^{d_{k+1}}$, with $d_0 := d$ and $\{d_k\}_{k=1}^{N_{\text{layers}}}$ given positive integers, and $P^k \in \mathbb{R}^{d_{k+1} \times d_k}$.

The derivation below is purely for illustrative purposes – an adaptive solver ought to perform better than an adaptation of an Euler scheme as (6.5.10). Moreover, the subsequent arguments will of course also apply to (6.5.3).

Let

$$\left\{t^0, \dots, t^{N_{\text{layers}}}\right\}, \qquad \text{with } t^0 := 0 \text{ and } t^{N_{\text{layers}}} := T,$$

be a given, non-decreasing sequence of time-steps. For simplicity of presentation, let us assume that the time-steps are uniform, namely $t^k = k\Delta t$ with $\Delta t = \frac{T}{N_{\text{layers}}}$, but more general time-adaptive sequences can be considered. For any $k \in \{0, \dots, N_{\text{layers}}\}$, let us assume that we are given a grid

$$\left\{x_j\left(t^k\right)\right\}_{j=1}^{d_k} \subset [0,1]$$

which is ordered and uniformly distributed. For simplicity of presentation, in our discussion we will assume that $x_1(t^k) = 0$ and $x_{d_k}(t^k) = 1$ for any k. However by means of an time-step-dependent dilation, this restriction may be removed. We note that, not only there might be no overlap of grid nodes over different time-steps, but moreover, the number of grid nodes changes at each time-step k.

We will seek for an appropriate discretization of

$$\partial_t \mathbf{x}(t^{k+1}, x_j(t^{k+1})) = \sigma\left(\int_0^1 w\left(t^{k+1}, x_j(t^{k+1}), \xi\right) \mathbf{x}(t^k, \xi)\, \mathrm{d}\xi + b\left(t^{k+1}, x_j(t^{k+1})\right)\right) \tag{6.5.11}$$

for $k \in \{0, \dots, N_{\text{layers}} - 1\}$ and $j \in \{1, \dots, d_{k+1}\}$. Hence, in view of the preceding discussion, some kind of interpolation may needed to justify a backward Euler discretization of the time derivative $\partial_t \mathbf{x}$ appearing in (6.5.11) at the grid nodes.

For any given $k \in \{0, \dots, N_{\text{layers}} - 1\}$ and $j \in \{1, \dots, d_k\}$, we shall henceforth denote

$$x_j^k := x_j(t^k), \qquad \mathbf{x}_j^k := \mathbf{x}(t^k, x_j^k).$$

Following through the above discussion, the main issue in writing down a forward difference discretization to $\partial_t \mathbf{x}(t^{k+1}, x_j(t^{k+1}))$ appears whenever for a given k one has $d_k \neq d_{k+1}$, as it is a priori not possible to make sense of the expression $\mathbf{x}(t^{k+1}, x_j(t^{k+1})) - \mathbf{x}(t^k, x_j(t^k))$ for $j \neq 1$. Indeed, all $\iota \in \{2, \dots, d_k\}$ are such that $x_\iota(t^k) \notin \{x_j(t^{k+1})\}_{j=1}^{d_{k+1}}$, due to the uniformity of the grid.

Let us give an elementary argument for addressing this issue. Given k and given any $j \in \{1, \dots, d_{k+1}\}$, there clearly exists $\iota \in \{2, \dots, d_k\}$ such that $x_j^{k+1} \in [x_{\iota-1}^k, x_\iota^k]$. For such indices, we may thus define the linear interpolant

$$\widehat{\mathbf{x}}_j^k := \mathbf{x}_\iota^k + \frac{\mathbf{x}_\iota^k - \mathbf{x}_{\iota-1}^k}{x_\iota^k - x_{\iota-1}^k}\left(x_j^{k+1} - x_\iota^k\right). \tag{6.5.12}$$

This is nothing but an approximation of the first order Taylor expansion of $\mathbf{x}(t^{k+1}, x_j(t^{k+1}))$ with respect to the second variable. Using this interpolant, we may consider the simple forward difference

$$\partial_t \mathbf{x}(t^{k+1}, x_j(t^{k+1})) \approx \frac{\mathbf{x}_j^{k+1} - \widehat{\mathbf{x}}_j^k}{\Delta t} \tag{6.5.13}$$

for any $k \in \{0, \dots, N_{\text{layers}} - 1\}$ and any $j \in \{1, \dots, d_{k+1}\}$. We may now use any Newton-Cotes formula to discretize the integral term in (6.5.11): for $j \in \{1, \dots, d_{k+1}\}$, we write

$$\int_0^1 w\left(t^{k+1}, x_j(t^{k+1}), \xi\right) \mathbf{x}(t^k, \xi)\, \mathrm{d}\xi \approx \sum_{\iota=1}^{d_k} \alpha_\iota w\left(t^{k+1}, x_j(t^{k+1}), x_\iota(t^k)\right) \mathbf{x}(t^k, x_\iota(t^k)). \tag{6.5.14}$$

Here, $\alpha_\iota > 0$ are the corresponding weights of the chosen Newton-Cotes formula.

Let us now define

$$\mathbf{x}^k := \begin{bmatrix} \mathbf{x}(t^k, x_1(t^k)) \\ \vdots \\ \mathbf{x}(t^k, x_{d_k}(t^k)) \end{bmatrix} \in \mathbb{R}^{d_k}, \qquad b^k := \begin{bmatrix} b(t^{k+1}, x_1(t^{k+1})) \\ \vdots \\ b(t^{k+1}, x_{d_{k+1}}(t^{k+1})) \end{bmatrix} \in \mathbb{R}^{d_{k+1}}$$

and

$$w^k := \left[\alpha_\iota w(t^{k+1}, x_j(t^{k+1}), x_\iota(t^k))\right]_{1 \leqslant j \leqslant d_{k+1} 1 \leqslant \iota \leqslant d_k} \in \mathbb{R}^{d_{k+1} \times d_k}.$$

The above definitions, as well as (6.5.13) and (6.5.14) applied to (6.5.11), lead us to (6.5.10), where Δt has been "omitted" as a factor of the nonlinearity. In view of (6.5.12), the operator $P^k \in \mathbb{R}^{d_{k+1}} \times \mathbb{R}^{d_k}$ takes the explicit

$$P^k = \sum_{j=1}^{d_{k+1}} \left(\left\{ 1 + \frac{x_j^{k+1} - x_{\iota(j)}^k}{x_{\iota(j)}^k - x_{\iota(j)-1}^k} \right\} \overline{e}_j e_{\iota(j)}^\top - \frac{x_j^{k+1} - x_{\iota(j)}^k}{x_{\iota(j)}^k - x_{\iota(j)-1}^k} \overline{e}_j e_{\iota(j)-1}^\top \right),$$

where $\iota(j) \in \{2, \dots, d_k\}$ is such that $x_j^{k+1} \in [x_{\iota(j)-1}^k, x_{\iota(j)}^k]$, while $\{\overline{e}_j\}_{j=1}^{d_{k+1}}$ and $\{e_j\}_{j=1}^{d_k}$ denote the canonical bases of $\mathbb{R}^{d_{k+1}}$ and $\mathbb{R}^{d_k}$ respectively. We notice that the matrix P^k only has 2 non-zero elements at every row $j \in \{1, \dots, d_{k+1}\}$. This concludes our derivation.

Remark 6.5.3 (Generating moving grids)**.** *Whilst we have assumed a very simple given time-dependent grid, one may certainly generate more sophisticated moving grids – we refer to [35] for a comprehensive overview on the existing methods, which have found extensive use in the discretization of partial differential equations manifesting shock waves and/or free boundaries.*

6.6 Concluding remarks

In this work, we have addressed the behavior when the time horizon goes to infinity of general but widely used learning problems for neural ODEs.

- In the classical empirical risk minimization problem with a Tikhonov parameter regularization, we concluded via Theorem 6.1 – Theorem 6.3 that when T is large enough, the obtained optimal/trained parameters for neural ODEs are such that the corresponding trajectories reach zero training error with a quantitative rate (thus, stipulate an approximation property of the trained model with respect to T), whilst doing so with the least oscillations possible. In the associated discrete-time, residual neural network setting, this result indicates that adding more layers before training would guarantee the optimal trajectories approach the zero training error regime, but do so without overfitting. In more practical terms, to ensure that the global minimizer is near zero training error, while training, one could systematically decrease the time horizon T whilst keeping the regularization parameter $\lambda > 0$ fixed.

- To obtain better quantitative estimates on the time horizon (and thus, number of layers) required to be ε–close to the zero training error regime, for a given tolerance $\varepsilon > 0$, we introduced a minimization problem wherein we added a tracking term which regularizes the state trajectories over the entire time horizon. In Theorem 6.5, we show that the training error and the optimal parameters are in $\mathcal{O}\left(e^{-\mu t}\right)$

for all $t \in [0, T]$. This result, along with numerical experiments, demonstrates a strong approximation rate of the trained neural ODE flow (which ought to be compared with universal approximation results, in which, a key caveat is that there is no scalable method to compute the theoretically guaranteed parameters), with parameters which are exponentially small, and thus stipulate that the flow would tend to oscillate little. Moreover, the exponential decay estimate also ensures that T need not be chosen too large to render the training error small.

6.6.1 Outlook

We present a list of questions and topics which would be complementary to our work.

- **Generalization bounds.** To complement our analytical study on the long time horizon/large layer regime, it would be of interest to provide generalization error bounds for the limiting, least L^2–norm parameters in the interpolation regime obtained in Theorem 6.1, via, for instance, commonly used metrics such as the VC dimension [263] or Rademacher complexity [18].
- **Exponential decay for (6.4.5) and non ℓ^2–losses.** We provided a proof of the exponential decay of the training error and optimal parameters in the context of ℓ^2–loss, and without regularizing the output $P\mathbf{x}_i(t)$ but rather the features $\mathbf{x}_i(t)$ over all time/layer $t \in [0, T]$. We stipulate that, whenever P is Lipschitz (and possibly real analytic) and such that the training error attains its minimum (e.g. when P is a matrix, or a matrix composed with a smoothly truncated sigmoid), the exponential decay result could hold by making use of a Łojasiewicz inequality argument. This is a prospective work. On the other hand, addressing analytically the (exponential) decay stipulated by the numerical experiments presented herein for non ℓ^2–losses such as cross-entropy remains an open problem.
- **Unsupervised learning.** As discussed in the introduction, the neural ODE representation of deep supervised learning has seen fruitful applications in the context of generative modeling via normalizing flows, a popular topic in the context of unsupervised learning. In unsupervised learning, one does not dispose of input-label samples, but rather only data which is unlabeled, and aims to generate a learned representation much like supervised learning. It would be of interest, in view of the existing applications, to investigate the potential use of the results presented in this work to the context of unsupervised learning.

6.7 Appendix: Proofs

6.7.1 Proof of Theorem 6.1

We note that both (6.3.2) and (6.3.3) can be written in the compact form

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{f}(w(t), b(t), \mathbf{x}(t)) & \text{in } (0, T) \\ \mathbf{x}(0) = \mathbf{x}^0 \in \mathbb{R}^{d_x}, \end{cases} \tag{6.7.1}$$

with

$$\mathbf{f}(0, 0, \mathbf{x}) = 0, \qquad \mathbf{f}(\alpha w, \alpha b, \mathbf{x}) = \alpha \mathbf{f}(w, b, \mathbf{x}) \quad \text{for } \alpha > 0. \tag{6.7.2}$$

We will refer to $u := [w, b]$ as *the control* of the ODE system, in accordance with control theory vocabulary. We begin with following short but key lemma.

Lemma 6.7.1. *Let $T_0 > 0$ and $[w_{T_0}, b_{T_0}] \in H^k(0, T_0; \mathbb{R}^{d_u})$ be given, and let $\mathbf{x}_{T_0}$ be the unique solution to*

$$\begin{cases} \dot{\mathbf{x}}_{T_0}(t) = \mathbf{f}(w_{T_0}(t), b_{T_0}(t), \mathbf{x}_{T_0}(t)) & \text{in } (0, T_0) \\ \mathbf{x}_{T_0}(0) = \mathbf{x}^0 \in \mathbb{R}^{d_x}, \end{cases} \tag{6.7.3}$$

(i.e. (6.7.1) on $(0,T_0)$) with $\mathbf{f}$ *as in either (6.3.3) or (6.3.2), thus satisfying (6.7.2). Let* $T>0$*, and define*

$$w_T(t) := \frac{T_0}{T} w_{T_0}\left(t\frac{T_0}{T}\right), \qquad b_T(t) := \frac{T_0}{T} b_{T_0}\left(t\frac{T_0}{T}\right) \qquad \text{for } t\in[0,T], \tag{6.7.4}$$

and

$$\mathbf{x}_T(t) := \mathbf{x}_{T_0}\left(t\frac{T_0}{T}\right) \qquad \text{for } t\in[0,T]. \tag{6.7.5}$$

Then $\mathbf{x}_T$ *is the unique solution to (6.7.1) (with the same* $\mathbf{f}$ *as in (6.7.3)) associated to* $[w_T, b_T]$*.*

We omit the proof, which follows by writing the integral formulation of $\mathbf{x}_T(t)$ and a change of variable in the intervening integral. This sort of time-scaling in the context of *driftless control affine* systems is commonly used in control theoretical contexts – a canonical example is the proof of the Chow-Rashevskii controllability theorem, see [69, Chapter 3, Section 3.3].

The following corollary is an immediate consequence.

Corollary 6.7.2. *Let* $\mathbf{x}^0\in\mathbb{R}^{d_x}$ *and* $\mathbf{x}^1\in\mathbb{R}^{d_x}$*. If (6.3.3) (resp. (6.3.2) with* σ *1–homogeneous) is controllable in some time* $T_0>0$*, then (6.3.3) (resp. (6.3.2)) is controllable in any time* $T>0$*.*

Proof of Corollary 6.7.2. Let $[w_{T_0}, b_{T_0}]\in H^k(0,T_0;\mathbb{R}^{d_u})$, with $k=0$ for (6.3.3) and $k=1$ for (6.3.2) be such that the corresponding solution $\mathbf{x}_{T_0}$ to (6.7.3) satisfies $\mathbf{x}_{T_0}(T_0)=\mathbf{x}^1$. Let $T>0$ and consider $[w_T, b_T]$ defined in (7.2.1). The corresponding solution $\mathbf{x}_T$ to (6.7.1) is thus given by (7.2.1) – we clearly observe that $\mathbf{x}_T(T)=\mathbf{x}_{T_0}\left(T\frac{T_0}{T}\right)=\mathbf{x}_{T_0}(T_0)=\mathbf{x}^1$. This concludes the proof. □

We are now in a position to prove the main result.

Proof of Theorem 6.1. We will henceforth, for notational convenience, extensively make use of the notation $u:=[w,b]$. We will focus on the neural ODE (6.3.3) and hence $k=0$. The case (6.3.2) and $k=1$ follows exactly the same arguments, and we will comment on the key differences at the end of the proof.

Part 1. We begin by showing

$$\mathcal{E}\left(\mathbf{x}_T(T)\right) \lesssim T^{-1} \tag{6.7.6}$$

uniformly in T. By the interpolation assumption, there exists some $u^1\in L^2(0,1;\mathbb{R}^{d_u})$ such that the associated solution $\mathbf{x}^1$ to (6.3.3) on $[0,1]$ satisfies $\mathcal{E}(\mathbf{x}^1(1))=0$. Using the optimality of u_T and the scaling relations from Lemma 7.2.1, we obtain

$$\begin{aligned} J_{\lambda,T}\left(u_T\right) &= \mathcal{E}\left(\mathbf{x}_T(T)\right) + \lambda\left\|u_T\right\|^2_{L^2(0,T;\mathbb{R}^{d_u})} \\ &\leqslant \mathcal{E}(\mathbf{x}^1(1)) + \frac{\lambda}{T}\left\|u^1\right\|^2_{L^2(0,1;\mathbb{R}^{d_u})} \end{aligned}$$

for all $T>0$. Since $\mathcal{E}(\mathbf{x}^1(1))=0$ by the interpolation assumption, the above inequality implies

$$0\leqslant \mathcal{E}\left(\mathbf{x}_T(T)\right) \leqslant \frac{\lambda}{T}\left\|u^1\right\|^2_{L^2(0,1;\mathbb{R}^{d_u})} \tag{6.7.7}$$

for all $T>0$. Estimate (6.7.7) clearly implies (6.7.6).

Part 2. We now look to prove (6.3.6). To this end, we will look to show that $\{\mathbf{x}_T(T)\}_{T>0}$ is a bounded subset of $\mathbb{R}^{d_x}$. This will allow us to extract a converging sequence, whose limit will be shown to lie in $\{\mathcal{E}=0\}$.

For any $T > 0$, set

$$u^{\mathrm{aux}}(t) := \frac{1}{T} u^1\left(\frac{t}{T}\right) \qquad \text{for } t \in [0, T].$$

We argue similarly as in Part 1. Making use of Lemma 7.2.1 once again, and since $\mathcal{E}(\mathbf{x}^1(1)) = 0$, we see that

$$\begin{aligned} J_{\lambda,T}\left(u^{\mathrm{aux}}\right) &= \mathcal{E}\left(\mathbf{x}^1(1)\right) + \frac{\lambda}{T} \left\|u^1\right\|^2_{L^2(0,1;\mathbb{R}^{d_u})} \\ &= \frac{\lambda}{T} \left\|u^1\right\|^2_{L^2(0,1;\mathbb{R}^{d_u})}. \end{aligned} \tag{6.7.8}$$

Using the optimality of u_T, one sees that

$$J_{\lambda,T}\left(u^{\mathrm{aux}}\right) \geqslant J_{\lambda,T}\left(u_T\right) \geqslant \lambda \left\|u_T\right\|^2_{L^2(0,T;\mathbb{R}^{d_u})}. \tag{6.7.9}$$

Combining (6.7.9) and (6.7.8), we deduce that

$$\left\|u_T\right\|^2_{L^2(0,T;\mathbb{R}^{d_u})} \leqslant \frac{1}{T} \left\|u^1\right\|^2_{L^2(0,1;\mathbb{R}^{d_u})} \tag{6.7.10}$$

for any $T > 0$. Now by integrating (6.3.3), and using the fact that σ is globally Lipschitz continuous with constant $C(\sigma) > 0$ and satisfies $\sigma(0) = 0$, for any $t \in [0, T]$ we have

$$\left\|\mathbf{x}_T(t) - \mathbf{x}^0\right\| \leqslant NC(\sigma) \int_0^t \left\|w_T(s)\right\| \left\|\mathbf{x}_T(s)\right\| \,\mathrm{d}s + N \left\|b_T\right\|_{L^1(0,T;\mathbb{R}^d)}.$$

By using the Grönwall inequality, we obtain

$$\left\|\mathbf{x}_T(T) - \mathbf{x}^0\right\| \leqslant N \left\|b_T\right\|_{L^1(0,T;\mathbb{R}^d)} \exp\left(NC(\sigma) \int_0^T \left\|w_T(s)\right\| \,\mathrm{d}s\right),$$

whereas by Cauchy-Schwarz, it follows that

$$\left\|\mathbf{x}_T(T) - \mathbf{x}^0\right\| \leqslant \sqrt{T} N \left\|b_T\right\|_{L^2(0,T;\mathbb{R}^d)} \exp\left(\sqrt{T} NC(\sigma) \left\|w_T\right\|_{L^2(0,T;\mathbb{R}^{d\times d})}\right).$$

At this point, employing (6.7.10), we deduce

$$\left\|\mathbf{x}_T(T) - \mathbf{x}^0\right\| \leqslant N \left\|u^1\right\|_{L^2(0,1;\mathbb{R}^{d_u})} \exp\left(NC(\sigma) \left\|u^1\right\|_{L^2(0,1;\mathbb{R}^{d_u})}\right).$$

Since u^1 is independent of T, we conclude that the set $\{\mathbf{x}_T(T)\}_{T>0}$ is bounded. Whence, there exists a sequence $\{T_n\}_{n=1}^{+\infty}$ with $T_n > 0$ and $T_n \longrightarrow +\infty$ as $n \longrightarrow +\infty$ and some $\mathbf{x}_\circ \in \mathbb{R}^{d_x}$ such that

$$\mathbf{x}_{T_n}(T_n) \longrightarrow \mathbf{x}_\circ \qquad \text{as } n \longrightarrow +\infty. \tag{6.7.11}$$

Since $\mathcal{E}\left(\mathbf{x}_{T_n}(T_n)\right) \longrightarrow 0$ as $n \longrightarrow +\infty$ by (6.7.6), by continuity of $\mathcal{E}$, we have $\mathcal{E}(\mathbf{x}_\circ) = 0$. This concludes the proof of (6.3.6).

Part 3. We now address the third statement of the theorem. To this end, we will first show that the sequence $\{u_n\}_{n=1}^{+\infty}$ defined in the statement is bounded in $L^2(0,1;\mathbb{R}^{d_u})$.

Let $u^\dagger \in L^2(0,1;\mathbb{R}^{d_u})$ be any solution to

$$\inf_{\substack{u \in L^2(0,1;\mathbb{R}^{d_u}) \\ \mathbf{x} \text{ solves (6.3.3)} \\ \text{and} \\ \mathcal{E}(\mathbf{x}(1))=0}} \int_0^1 \|u(t)\|^2 \,\mathrm{d}t. \tag{6.7.12}$$

Denote by $\mathbf{x}^\dagger$ the corresponding solution to (6.3.3) on $[0,1]$. We claim that

$$\left\|u_n\right\|_{L^2(0,1;\mathbb{R}^{d_u})} \leqslant \left\|u^\dagger\right\|_{L^2(0,1;\mathbb{R}^{d_u})}, \qquad \text{for all } n \geqslant 1. \tag{6.7.13}$$

We prove this claim by contradiction. Indeed, assume that we had

$$\left\|u^\dagger\right\|_{L^2(0,1;\mathbb{R}^{d_u})} < \|u_n\|_{L^2(0,1;\mathbb{R}^{d_u})} \qquad \text{for some } n \geqslant 1.$$

We consider

$$u_n^\dagger(t) := \frac{1}{T_n} u^\dagger\left(\frac{t}{T_n}\right) \qquad \text{for } t \in [0, T_n],$$

whose corresponding state $\mathbf{x}_n^\dagger$, solution to (6.3.3) on $[0, T_n]$, satisfies $\mathbf{x}_n^\dagger(T_n) = \mathbf{x}^\dagger(1)$ by Lemma 7.2.1. On another hand, by assumption we have $\mathcal{E}(\mathbf{x}^\dagger(1)) = 0$. It then follows that

$$\begin{aligned} J_{\lambda,T_n}\left(u_n^\dagger\right) &= \frac{\lambda}{T_n}\left\|u^\dagger\right\|^2_{L^2(0,1;\mathbb{R}^{d_u})} \\ &< \mathcal{E}\left(\mathbf{x}_{T_n}(T_n)\right) + \frac{\lambda}{T_n}\|u_n\|^2_{L^2(0,1;\mathbb{R}^{d_u})} = J_{T_n}\left(u_{T_n}\right), \end{aligned}$$

which contradicts the fact that u_{T_n} minimizes J_{T_n}. Hence, (6.7.13) holds, and $\{u_n\}_{n=1}^{+\infty}$ is bounded in $L^2(0,1;\mathbb{R}^{d_u})$. Consequently, by the Banach-Alaoglu theorem, there exists $u^* = [w^*, b^*] \in L^2(0,1;\mathbb{R}^{d_u})$ such that

$$u_n \rightharpoonup u^* \qquad \text{weakly in } L^2(0,1;\mathbb{R}^{d_u}),$$

along some subsequence as $n \longrightarrow +\infty$. Moreover, using the properties of equation (6.3.3) (see the arguments in the proof of Proposition 6.2.1), we deduce that the trajectory $\mathbf{x}_n$ associated to u_n satisfies

$$\mathbf{x}_n \longrightarrow \mathbf{x}^* \qquad \text{strongly in } C^0([0,1];\mathbb{R}^{d_x}) \tag{6.7.14}$$

as $n \longrightarrow +\infty$, where $\mathbf{x}^*$ is the solution to (6.3.3) on $[0,1]$, associated to u^*. On another hand, note that by Lemma 7.2.1, $\mathbf{x}_{T_n}(t) = \mathbf{x}_n(\frac{t}{T_n})$ for $t \in [0, T_n]$, whence $\mathbf{x}_{T_n}(T_n) = \mathbf{x}_n(1)$ and thus, combining (6.7.14) and (6.7.11), we see that $\mathbf{x}^*(1) = \mathbf{x}_\circ$. Consequently, u^* is a control such that $\mathcal{E}(\mathbf{x}^*(1)) = \mathcal{E}(\mathbf{x}_\circ) = 0$, thus satisfying the constraint in (6.7.12). In view of this, we may also use (6.7.13) and the weak lower semicontinuity of the L^2–norm to write

$$\begin{aligned} \left\|u^\dagger\right\|_{L^2(0,1;\mathbb{R}^{d_u})} \leqslant \|u^*\|_{L^2(0,1;\mathbb{R}^{d_u})} &\leqslant \liminf_{n\longrightarrow+\infty} \|u_n\|_{L^2(0,1;\mathbb{R}^{d_u})} \\ &\leqslant \lim_{n\longrightarrow+\infty} \|u_n\|_{L^2(0,1;\mathbb{R}^{d_u})} \\ &\leqslant \limsup_{n\longrightarrow+\infty} \|u_n\|_{L^2(0,1;\mathbb{R}^{d_u})} \\ &\leqslant \left\|u^\dagger\right\|_{L^2(0,1;\mathbb{R}^{d_u})}, \end{aligned} \tag{6.7.15}$$

clearly implying that

$$\lim_{n\longrightarrow+\infty} \|u_n\|_{L^2(0,1;\mathbb{R}^{d_u})} = \|u^*\|_{L^2(0,1;\mathbb{R}^{d_u})}.$$

Hence, as weak convergence and convergence of the norms in L^2 implies strong convergence in L^2, we deduce that

$$u_n \longrightarrow u^* \qquad \text{strongly in } L^2(0,1;\mathbb{R}^{d_u})$$

along soe subsequence as $n \longrightarrow +\infty$. Moreover, from (6.7.15) we deduce that, since $u^\dagger$ is a solution to (6.7.12) and since u^* satisfies the constraints therein, u^* is a solution to (6.7.12) as well, which concludes the proof for (6.3.3) and $k = 0$.

In the case (6.3.2) and $k = 1$, one may clearly repeat the above reasoning, replacing $L^2(0,T;\mathbb{R}^{d_u})$ by $H^1(0,T;\mathbb{R}^{d_u})$ throughout, with some key additions.

In Part 1, we first note that instead of (6.7.8), one has

$$J_{\lambda,T}(u^{\mathrm{aux}}) = \mathcal{E}(\mathbf{x}^1(1)) + \frac{\lambda}{T}\left\|u^1\right\|^2_{L^2(0,1;\mathbb{R}^{d_u})} + \frac{\lambda}{T^3}\left\|\dot{u}^1\right\|^2_{L^2(0,1;\mathbb{R}^{d_u})}$$
$$= \frac{\lambda}{T}\left\|u_{T_0}\right\|^2_{L^2(0,1;\mathbb{R}^{d_u})} + \frac{\lambda}{T^3}\left\|\dot{u}^1\right\|^2_{L^2(0,1;\mathbb{R}^{d_u})}.$$

This is not an impediment to (6.7.9), which remains true, and one can clearly deduce that $\{\mathbf{x}_T(T)\}_{T>0}$ is bounded as well. Similarly, (6.7.7) holds with a bound of the form

$$0 \leqslant \mathcal{E}(\mathbf{x}_T(T)) \leqslant \frac{\lambda}{T}\left\|u^1\right\|^2_{L^2(0,1;\mathbb{R}^{d_u})} + \frac{\lambda}{T^3}\left\|\dot{u}^1\right\|^2_{L^2(0,1;\mathbb{R}^{d_u})}.$$

Whence the remainder of parts 1 and 2 hold in this context as well.

In Part 3, we emphasize the sole key difference between (6.3.3) and (6.3.2) – the weak L^2–convergence of $\{u_n\}_{n=1}^{+\infty}$ is a priori not sufficient to entail the strong convergence in (6.7.14) in the case of (6.3.2). However, by the Rellich-Kondrachov compactness theorem, the weak H^1–convergence of $\{u_n\}_{n=1}^{+\infty}$ implies a strong L^2–convergence along a subsequence, which would yield (6.7.14) by arguing just as in the proof of Proposition 6.2.1.

This concludes the proof. □

6.7.2 Proof of Theorem 6.2

The proof closely follows the lines of that just above. Let us consider $k = 1$, since the case $k = 0$ is equivalent to Theorem 6.2. We present minimal details for completeness.

Proof of Theorem 6.2. We again make use of the notation $u := [w, b]$. We first show

$$\mathcal{E}(\mathbf{x}_\lambda(T)) \lesssim \lambda \tag{6.7.16}$$

uniformly in $\lambda > 0$ – we argue as in the proof of Theorem 6.1 just above, exhibiting, by the interpolation assumption, parameters $u^1 \in L^2(0,1;\mathbb{R}^{d_u})$ such that $\mathcal{E}(\mathbf{x}^1(1)) = 0$. We may obtain an estimate like (6.7.7) and conclude. Now, the same arguments as in Part 2 of the proof of Theorem 6.1 may be used to deduce that $\{\mathbf{x}_\lambda(T)\}_{\lambda>0}$ is a bounded subset of $\mathbb{R}^d$, and hence there exists a sequence $\{\lambda_n\}_{n=1}^{+\infty}$ of positive numbers with $\lambda_n \searrow 0$ as $n \longrightarrow +\infty$ and some $\mathbf{x}_\circ \in \mathbb{R}^{d_x}$ such that

$$\mathbf{x}_{\lambda_n}(T) \xrightarrow[n\longrightarrow+\infty]{} \mathbf{x}_\circ.$$

Using (6.7.16) we deduce that $\mathcal{E}(\mathbf{x}_\circ) = 0$. Finally, the proof of the last fact is identical to that done for Theorem 6.1, so we omit it. □

6.7.3 Proof of Proposition 6.3.2

The proof of Proposition 6.3.2 is a straightforward Grönwall argument. We sketch it for completeness.

Proof of Proposition 6.3.2. For simplicity of presentation but without any loss of generality, we will henceforth concentrate on system (6.3.3). For any $t \in [0,T]$, $i \in [N]$ and $j \in [N]$, we have

$$\mathbf{x}_i(t) - \mathbf{x}_j(t) = \mathbf{x}_i^0 - \mathbf{x}_j^0 + \int_0^t w(\tau)\Big(\sigma(\mathbf{x}_i(\tau)) - \sigma(\mathbf{x}_j(\tau))\Big)\,\mathrm{d}\tau.$$

Using the Lipschitz character of σ, we get

$$\|\mathbf{x}_i(t) - \mathbf{x}_j(t)\| \leqslant \left\|\mathbf{x}_i^0 - \mathbf{x}_j^0\right\| + \int_0^t \|w(\tau)\|\,\|\sigma(\mathbf{x}_i(\tau)) - \sigma(\mathbf{x}_j(\tau))\|\,\mathrm{d}\tau$$
$$\leqslant \left\|\mathbf{x}_i^0 - \mathbf{x}_j^0\right\| + C(\sigma)\int_0^t \|w(\tau)\|\,\|\mathbf{x}_i(\tau) - \mathbf{x}_j(\tau)\|\,\mathrm{d}\tau.$$

We apply the Grönwall inequality with the effect of

$$\|\mathbf{x}_i(t) - \mathbf{x}_j(t)\| \leqslant \exp\left(C(\sigma)\int_0^t \|w(\tau)\|\,\mathrm{d}\tau\right)\left\|\mathbf{x}_i^0 - \mathbf{x}_j^0\right\|.$$

We evaluate the above expression at final time $t = T$ to obtain

$$\left\|\mathbf{x}_i^1 - \mathbf{x}_j^1\right\| \leqslant \exp\left(C(\sigma)\int_0^T \|w(\tau)\|\,\mathrm{d}\tau\right)\left\|\mathbf{x}_i^0 - \mathbf{x}_j^0\right\|,$$

for some $\mathbf{x}_i^1 \in P^{-1}\left(\{\vec{y}_i\}\right)$ and $\mathbf{x}_j^1 \in P^{-1}\left(\{\vec{y}_j\}\right)$, whence

$$\exp\left(C(\sigma)\int_0^T \|w(\tau)\|\,\mathrm{d}\tau\right) \geqslant \frac{\left\|\mathbf{x}_i^1 - \mathbf{x}_j^1\right\|}{\left\|\mathbf{x}_i^0 - \mathbf{x}_j^0\right\|}.$$

Taking the log on both sides we obtain (6.3.9). □

6.7.4 Proof of Theorem 6.3

We now provide a proof of our main result in the context of classification tasks.

Proof of Theorem 6.3. Let $\left[\widehat{w}, \widehat{b}\right] \in H^k(0, T_0; \mathbb{R}^{d_u})$ be a pair of parameters which separates the dataset $\left\{\mathbf{x}_i^0, \vec{y}_i\right\}_{i=1}^N$ with respect to P in time $T_0 > 0$, i.e. such that the solution $\widehat{\mathbf{x}} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_N]$ to (6.3.2) with initial condition $\mathbf{x}^0 = [\mathbf{x}_1^0, \ldots, \mathbf{x}_N^0]$ corresponding to $\left[\widehat{w}, \widehat{b}\right]$ satisfies

$$\min_{i\in[N]}\left\{P\,\widehat{\mathbf{x}}_i(T_0)_{\vec{y}_i} - \max_{\substack{j\in[N]\\ j\neq\vec{y}_i}} P\,\widehat{\mathbf{x}}_i(T_0)_j\right\} =: \gamma > 0. \tag{6.7.17}$$

Now fix an arbitrary $\alpha \in \left(0, \frac{1}{2}\right)$, and, for any $T > 0$, define

$$\left[w^\dagger(t), b^\dagger(t)\right] := \begin{cases} \dfrac{2T_0}{T}\left[\widehat{w}\left(t\dfrac{2T_0}{T}\right), \widehat{b}\left(t\dfrac{2T_0}{T}\right)\right] & \text{for } t \in \left[0, \dfrac{T}{2}\right] \\ T^{\alpha-1}\left[\mathrm{Id}_d, 0_d\right] & \text{for } t \in \left(\dfrac{T}{2}, T\right], \end{cases}$$

where Id_d is the identity matrix in $\mathbb{R}^{d\times d}$ and 0_d is the zero vector in $\mathbb{R}^d$. By virtue of the scaling in Lemma 7.2.1, for $t \in \left[\frac{T}{2}, T\right]$, the trajectories $\mathbf{x}^\dagger = \left[\mathbf{x}_1^\dagger, \ldots, \mathbf{x}_N^\dagger\right]$ associated to $\left[w^\dagger, b^\dagger\right]$ are given by the solution to

$$\begin{cases} \dot{\mathbf{x}}_i^\dagger(t) = \sigma\left(T^{\alpha-1}\mathbf{x}_i^\dagger(t)\right) & \text{for } t \in \left[\dfrac{T}{2}, T\right] \\ \mathbf{x}_i^\dagger\left(\dfrac{T}{2}\right) = \widehat{\mathbf{x}}_i\left(T_0\right). \end{cases} \tag{6.7.18}$$

Moreover, since $\sigma(x) = \max\{x, 0\}$, and thus σ being nonnegative, the right hand side in (6.3.2) is nonnegative. Using the assumption that the initial conditions are of the form $\mathbf{x}_i^0 = \mathfrak{Q}\vec{x}_i \geqslant 0$, it follows that $\widehat{\mathbf{x}}_i(T_0) \geqslant 0$ for all $i \in [N]$. We can therefore drop σ from (6.7.19) and deduce that $P\mathbf{x}_i^\dagger(t)$ solves

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}P\mathbf{x}_i^\dagger(t) = T^{\alpha-1}P\mathbf{x}_i^\dagger(t) & \text{for } t \in \left[\dfrac{T}{2}, T\right] \\ P\mathbf{x}_i^\dagger\left(\dfrac{T}{2}\right) = P\widehat{\mathbf{x}}_i\left(T_0\right). \end{cases} \tag{6.7.19}$$

Hence, we have

$$P\mathbf{x}_i^\dagger(t) = P\widehat{\mathbf{x}}_i(T_0)e^{T^{\alpha-1}(t-T/2)}, \qquad \text{for all } t \in \left[\frac{T}{2}, T\right].$$

Now, using the definition of the cross-entropy loss and the margin γ in (6.7.17), we compute, for any $i \in [N]$,

$$\begin{aligned}
\operatorname{loss}\left(\mathbf{x}_i^\dagger(T), \vec{y}_i\right) &= -\log\left(\frac{e^{P\widehat{\mathbf{x}}_i(T_0)_{\vec{y}_i} e^{\frac{T^\alpha}{2}}}}{\sum_{j=1}^m e^{P\widehat{\mathbf{x}}_i(T_0)_j e^{\frac{T^\alpha}{2}}}}\right) \\
&= \log\left(1 + \sum_{j\neq\vec{y}_i} e^{\left(P\widehat{\mathbf{x}}_i(T_0)_j e^{\frac{T^\alpha}{2}}\right) - \left(P\widehat{\mathbf{x}}_i(T_0)_{\vec{y}_i} e^{\frac{T^\alpha}{2}}\right)}\right) \\
&\leqslant \log\left(1 + (m-1)\exp\left(-\gamma\,\exp\left(\frac{T^\alpha}{2}\right)\right)\right).
\end{aligned}$$

Then, we can estimate

$$\mathcal{E}\left(\mathbf{x}^\dagger(T)\right) \leqslant \log\left(1 + (m-1)\exp\left(-\gamma\,\exp\left(\frac{T^\alpha}{2}\right)\right)\right). \tag{6.7.20}$$

On the other hand, using the definition of $\left[w^\dagger, b^\dagger\right]$, we deduce

$$\begin{aligned}
\left\|\left[w^\dagger, b^\dagger\right]\right\|^2_{H^1(0,T;\mathbb{R}^{d_u})} &= \left\|\left[w^\dagger, b^\dagger\right]\right\|^2_{H^1\left(0,\frac{T}{2};\mathbb{R}^{d_u}\right)} + \left\|\left[w^\dagger, b^\dagger\right]\right\|^2_{H^1\left(\frac{T}{2},T\right)} \\
&\leqslant \frac{C_1}{T} + C_2\, T^{2(\alpha-1)}T,
\end{aligned}$$

for some constants $C_1, C_2 > 0$ depending only on λ, T_0 and $\left[\widehat{w}, \widehat{b}\right]$. From this estimate, together with (6.7.20), we obtain, for $T > T_0$,

$$J_{\lambda,T}\left(w^\dagger, b^\dagger\right) \leqslant \log\left(1 + (m-1)\exp\left(-\gamma\,\exp\left(\frac{T^\alpha}{2}\right)\right)\right) + CT^{2\alpha-1},$$

for some constant $C > 0$ depending on λ, T_0, $\left[\widehat{w}, \widehat{b}\right]$, but independent of T. Using the above estimate, we may conclude from the optimality of $[w_T, b_T]$, as

$$\begin{aligned}
\mathcal{E}\left(\mathbf{x}_T(T)\right) \leqslant J_{\lambda,T}\left(w_T, b_T\right) &\leqslant J_{\lambda,T}\left(w^\dagger, b^\dagger\right) \\
&\leqslant \log\left(1 + (m-1)\exp\left(-\gamma\,\exp\left(\frac{T^\alpha}{2}\right)\right)\right) + CT^{2\alpha-1}.
\end{aligned}$$

□

6.7.5 Proof of Theorem 6.6

The following short functional analysis lemma will be of use in the proof of Theorem 6.6. We omit the proof, which follows by using the open mapping theorem (see e.g. [33, Theorem 2.6, pp. 35]).

Lemma 6.7.3. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two real Hilbert spaces. Let*

$$\Lambda : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$$

be a linear, bounded and surjective operator. Then

$$\begin{aligned}
\Gamma : \mathcal{H}_2 &\longrightarrow \mathcal{H}_1 \\
y &\longmapsto \operatorname*{arg\,min}_{x\in\Lambda^{-1}(\{y\})} \|x\|^2_{\mathcal{H}_1}
\end{aligned}$$

is linear and bounded.

Proof of Theorem 6.6. Inspired by the techniques in [72] and the so-called *staircase* method introduced in [219] (see also [236]), we define the continuous arc

$$\begin{aligned}\gamma : [0,1] &\longrightarrow \mathbb{R}^{d_x} \\ s &\longmapsto (1-s)\mathbf{x}^0 + s\mathbf{x}^1.\end{aligned}$$

By assumption,

$$\left\{\sigma\left(\mathbf{x}_1^1\right), \dots, \sigma\left(\mathbf{x}_i^1\right), \dots, \sigma\left(\mathbf{x}_N^1\right)\right\}$$

is a linearly independent system of vectors in $\mathbb{R}^d$ for any $s \in [0,1]$. Thus, by using the continuity of γ, there exists an $\eta > 0$, such that whenever $\left\|\mathbf{x}^1 - \mathbf{x}^0\right\| \leqslant \eta$,

$$\left\{\sigma\left(\gamma_1(s)\right), \dots, \sigma\left(\gamma_i(s)\right), \dots, \sigma\left(\gamma_N(s)\right)\right\} \tag{6.7.21}$$

is also a system of linearly independent vectors in $\mathbb{R}^d$ for any $s \in [0,1]$. Following the framework of Lemma 6.7.3, for any $s \in [0,1]$, define

$$\begin{aligned}\Lambda_s : \mathbb{R}^{d\times d} &\longrightarrow \mathbb{R}^{d_x} \\ w &\longmapsto w\sigma\left(\gamma\left(s\right)\right).\end{aligned}$$

By the linear independence of the system of vectors (6.7.21), Λ_s is surjective for any $s \in [0,1]$. Hence, using Lemma 6.7.3, we see that

$$\begin{aligned}\Gamma_s : \mathbb{R}^{d_x} &\longrightarrow \mathbb{R}^{d\times d} \\ y &\longmapsto \underset{w\in\Lambda_s^{-1}(\{y\})}{\arg\min} \|w\|,\end{aligned}$$

is a linear and bounded operator for any $s \in [0,1]$, and, since (6.7.21) is independent and the arc γ is continuous, $\{\Gamma_s\}_{s\in[0,1]}$ is uniformly bounded in operator norm:

$$\|\Gamma_s\|_{\mathcal{L}\left(\mathbb{R}^{d_x};\mathbb{R}^{d\times d}\right)} \leqslant C \tag{6.7.22}$$

for some $C > 0$ independent of $T > 0$. Now, for $t \in [0,T]$, set

$$w(t) \coloneqq \Gamma_{s_t}\left(\frac{\mathbf{x}^1 - \mathbf{x}^0}{T}\right), \tag{6.7.23}$$

with $s_t \coloneqq \frac{t}{T}$. Note that for any $t \in [0,T]$, the vector $w(t) \in \mathbb{R}^{d\times d}$ solves the linear system of equations

$$w(t)\sigma\left(\mathbf{x}_i(t)\right) = \dot{\mathbf{x}}_i(t) \quad \text{for } i \in [N],$$

where

$$\mathbf{x}(t) \coloneqq \gamma\left(\frac{t}{T}\right) = \left(1 - \frac{t}{T}\right)\mathbf{x}^0 + \frac{t}{T}\mathbf{x}^1.$$

Hence, $\mathbf{x}(t)$ solves

$$\begin{cases}\dot{\mathbf{x}}_i(t) = \mathbf{w}(t)\sigma(\mathbf{x}_i(t)) & \text{for } t \in (0,T) \\ \mathbf{x}_i(0) = \gamma(0) = \mathbf{x}_i^0 \\ \mathbf{x}_i\left(T\right) = \gamma(1) = \mathbf{x}_i^1,\end{cases}$$

for any $i \in [N]$. This thus demonstrates the existence of a control w steering the stacked dynamics from $\mathbf{x}^0$ to $\mathbf{x}^1$ in time T.

Let us conclude by showing that w satisfies the stated estimate. By the definition of w in (6.7.23) as well as (6.7.22), for any $t \in [0,T]$ we have

$$\|w(t)\| = \left\|\Gamma_t\left(\frac{\mathbf{x}^1 - \mathbf{x}^0}{T}\right)\right\| \leqslant \frac{C}{T}\left\|\mathbf{x}^1 - \mathbf{x}^0\right\|,$$

as desired. □

6.7.6 Proof of Proposition 6.2.1

For the sake of completeness, and usage of the arguments in some of the other proofs, we sketch a proof of the existence of minimizers via the classical direct method of the calculus of variations.

Proof of Proposition 6.2.1. We shall concentrate solely on the case $k = 0$, as modulo an application of the Rellich-Kondrachov compactness theorem, the arguments are exactly the same in the case $k = 1$. We fix $\lambda = 1$ for simplicity.

Let $\{[w_n, b_n]\}_{n=1}^{+\infty} \subset L^2(0, T; \mathbb{R}^{d_u})$ be a minimizing sequence, namely a sequence satisfying

$$\lim_{n\to+\infty} J_T(w_n, b_n) = \inf_{[w,b]\in L^2(0,T;\mathbb{R}^{d_u})} J_T(w, b).$$

For any $n \geqslant 1$, denote by $\mathbf{x}_n \in C^0([0, T]; \mathbb{R}^{d_x})$ the unique solution to (6.3.3) – (6.3.1) associated to $[w_n, b_n]$ and the initial datum $\mathbf{x}^0$. Note that

$$J_T(w, b) \geqslant \int_0^T \Big\| [w(t), b(t)] \Big\|^2 \, \mathrm{d}t,$$

whence J_T is coercive in the sense that $J_T(u) \to +\infty$ when $\|u\|_{L^2} \to +\infty$. Since J_T is coercive, it follows that $\{[w_n, b_n]\}_{n=1}^{+\infty}$ is bounded in $L^2(0, T; \mathbb{R}^{d_u})$. Therefore, there exists a pair $\left[w^\dagger, b^\dagger\right] \in L^2(0, T; \mathbb{R}^{d_u})$ such that

$$\begin{aligned} w_n &\rightharpoonup w^\dagger \qquad \text{weakly in } L^2(0, T; \mathbb{R}^{d\times d}) \\ b_n &\rightharpoonup b^\dagger \qquad \text{weakly in } L^2(0, T; \mathbb{R}^{d}) \end{aligned}$$

along a subsequence as $n \to +\infty$. Of course, the same convergences thence hold for $\mathbf{w}_n := \mathrm{diag}_N(w_n)$ to $\mathbf{w}^\dagger := \mathrm{diag}_N(w^\dagger)$, as well as $\mathbf{b}_n := [b_n, \dots, b_n]$ to $\mathbf{b}^\dagger := [b^\dagger, \dots, b^\dagger]$. Let $\mathbf{x}^\dagger \in C^0([0, T]; \mathbb{R}^{d_x})$ be the unique solution to (6.3.3) associated to $[w^\dagger, b^\dagger]$ and the initial datum $\mathbf{x}^0$. Let us prove that

$$\mathbf{x}_n \longrightarrow \mathbf{x}^\dagger \qquad \text{strongly in } C^0([0, T]; \mathbb{R}^{d_x}) \tag{6.7.24}$$

along the aforementioned subsequence as $n \to +\infty$. Take an arbitrary $t \in [0, T]$. Note that

$$\begin{aligned} \mathbf{x}_n(t) - \mathbf{x}^\dagger(t) &= \int_0^t \Big[\mathbf{w}_n(\tau)\sigma(\mathbf{x}_n(\tau)) + \mathbf{b}_n(\tau)\Big] \, \mathrm{d}\tau - \int_0^t \left[\mathbf{w}^\dagger(\tau)\sigma\left(\mathbf{x}^\dagger(\tau)\right) + \mathbf{b}^\dagger(\tau)\right] \, \mathrm{d}\tau \\ &= \int_0^t \Big[\mathbf{w}_n(\tau)\sigma(\mathbf{x}_n(\tau)) - \mathbf{w}_n(\tau)\sigma\left(\mathbf{x}^\dagger(\tau)\right)\Big] \, \mathrm{d}\tau \\ &\quad + \int_0^t \Big[\mathbf{w}_n(\tau)\sigma\left(\mathbf{x}^\dagger(\tau)\right) - \mathbf{w}^\dagger(\tau)\sigma\left(\mathbf{x}^\dagger(\tau)\right)\Big] \, \mathrm{d}\tau \\ &\quad + \int_0^t \Big[\mathbf{b}_n(\tau) - \mathbf{b}^\dagger(\tau)\Big] \, \mathrm{d}\tau. \end{aligned}$$

Hence, using the fact that σ is globally Lipschitz with constant $c(\sigma) > 0$,

$$\begin{aligned} \left\|\mathbf{x}_n(t) - \mathbf{x}^\dagger(t)\right\| &\leqslant \int_0^t \|\mathbf{w}_n(\tau)\| \left\|\sigma\Big(\mathbf{x}_n(\tau)\Big) - \sigma\left(\mathbf{x}^\dagger(\tau)\right)\right\| \, \mathrm{d}\tau \\ &\quad + \left\| \int_0^t \sigma\left(\mathbf{x}^\dagger(\tau)\right) \left[\mathbf{w}_n(\tau) - \mathbf{w}^\dagger(\tau)\right] \, \mathrm{d}\tau \right\| \\ &\quad + \left\| \int_0^t \Big[\mathbf{b}_n(\tau) - \mathbf{b}^\dagger(\tau)\Big] \, \mathrm{d}\tau \right\| \\ &\leqslant c(\sigma) \int_0^t \|\mathbf{w}_n(\tau)\| \left\|\mathbf{x}_n(\tau) - \mathbf{x}^\dagger(\tau)\right\| \, \mathrm{d}\tau + c_n, \end{aligned}$$

with

$$c_n \coloneqq \left\| \int_0^t \sigma\left(\mathbf{x}^\dagger(\tau)\right)\left[\mathbf{w}_n(\tau) - \mathbf{w}^\dagger(\tau)\right] \mathrm{d}\tau \right\| + \left\| \int_0^t \left[\mathbf{b}_n(\tau) - \mathbf{b}^\dagger(\tau)\right] \mathrm{d}\tau \right\|.$$

Using Grönwall's inequality, Cauchy-Schwarz, and the boundedness of the L^2–norm of $\{\mathbf{w}_n\}_{n=1}^{+\infty}$ by some constant $M > 0$ independent of t, we thence obtain

$$\begin{aligned}
\left\|\mathbf{x}_n(t) - \mathbf{x}^\dagger(t)\right\| &\leqslant c_n \exp\left(c(\sigma)\int_0^t \left\|\mathbf{w}_n(\tau)\right\| \mathrm{d}\tau\right) \\
&\leqslant c_n \exp\left(c(\sigma)\sqrt{T}\left\|\mathbf{w}_n\right\|_{L^2(0,T;\mathbb{R}^{d\times d\times N})}\right) \\
&\leqslant c_n \exp\left(c(\sigma)\sqrt{T}M\right).
\end{aligned}$$

As $c_n \longrightarrow 0$ along any subsequence as $n \to +\infty$ by virtue of the weak convergences of $\{\mathbf{w}_n\}_{n=1}^{+\infty}$ to $\mathbf{w}^\dagger$ and $\{\mathbf{b}_n\}_{n=1}^{+\infty}$ to $\mathbf{b}^\dagger$, we deduce (6.7.24).

Now using the weak lower semicontinuity of the squared $L^2(0,T;\mathbb{R}^{d_u})$–norm, the continuity of $\mathcal{E}$, (6.7.24) and – if there is an integral tracking term of the state – the Lebesgue dominated convergence theorem, we deduce

$$\begin{aligned}
\inf_{[w,b]\in L^2(0,T;\mathbb{R}^{d_u})} J_T(w,b) &= \lim_{n\to+\infty} J_T(w_n,b_n) \\
&\geqslant \liminf_{n\to+\infty} J_T(w_n,b_n) \\
&\geqslant J_T(w^\dagger,b^\dagger).
\end{aligned}$$

Whence $\left[w^\dagger, b^\dagger\right]$ is a minimizer. This concludes the proof. □

Chapter 7

Sparse approximation in learning via Neural ODEs

Abstract. We consider the continuous-time, neural ODE perspective of deep supervised learning, and study the role of the final time horizon T, which may be interpreted as the depth of the associated residual neural network (ResNet). We focus on a cost consisting of an integral of the empirical risk over the time horizon and L^1–parameter regularization, and under homogeneity assumptions on the dynamics (typical for ReLU activations), we prove that any global minimizer is sparse, in the sense there exists a positive stopping time T^* beyond which the optimal parameters vanish. Moreover, under appropriate interpolation assumptions of the model, we may provide quantitative estimates on the stopping time T^*, and on the training error of the neural ODE trajectories at the stopping time. The latter stipulates a quantitative approximation property of neural ODE flows with sparse parameters. In practical terms, when extrapolated to the ResNet context, a shorter time-horizon in the optimal control problem can be interpreted as considering a shallower ResNet, which may lower the computational cost of training.

Keywords. Deep Learning; Neural ODEs; Supervised Learning; Sparsity; Optimal control; Nonlinear systems.
AMS Subject Classification. 49J15; 49M15; 49J20; 49K20; 93C20; 49N05.

Chapter 7 is taken from [272]:

> *Sparse approximation in learning via neural ODEs.*
> C. Esteve Yagüe and B. Geshkovski, 2021.
> https://arxiv.org/abs/2102.13566

Chapter Contents

7.1 Introduction

Sparsity is a highly desirable property in many machine learning and optimization tasks due to the inherent reduction of computational complexity. When induced by ℓ^1–regularization for instance, it has been used extensively for simplifying a machine learning task by selecting a strict subset of the available features to be used in an automatized manner. An illustrative example is the well-known Lasso (least absolute shrinkage and selection operator, [239, 258]), which consists in minimizing a least squares cost function and an ℓ^1–penalty for an affine parametric model, and enforces a subset of the trainable parameters to become zero. As a consequence, the associated features may safely be removed.

Following this line of reasoning, in this work, we study supervised learning problems viewed from a continuous-time, neural ODE perspective, and we demonstrate the appearance of sparsity patterns for L^1–regularized minimization problems.

We recall that supervised learning addresses the problem of predicting from data, which consists in approximating an unknown function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ from N known and possibly noisy samples $\{\vec{x}_i, \vec{y}_i = f(\vec{x}_i)\}_{i=1}^N$. Depending on the nature of the space of labels $\mathcal{Y}$, one distinguishes two types of supervised learning tasks, namely that of *classification* (labels take values in a finite set of m classes, e.g. $\mathcal{Y} = \{1, \dots, m\}$) and *regression* (labels take continuous values in $\mathcal{Y} \subset \mathbb{R}^m$). Heuristically, supervised learning consists in constructing a map

$$f_{\text{approx}} : \mathcal{X} \longrightarrow \mathcal{P}(\mathcal{Y}),$$

which, desirably, is such that for any $x \in \mathcal{X}$ and for any Borel measurable $A \subset \mathcal{Y}$, $f_{\text{approx}}(x)(A) \simeq 1$ whenever $f(x) \in A$, and $f_{\text{approx}}(x)(A) \simeq 0$ whenever $f(x) \notin A$; here, $\mathcal{P}(\mathcal{Y})$ denotes the space of probability measures on $\mathcal{Y}$. In other words, one looks for a map f_{approx} which approximates the map $x \longmapsto \delta_{f(x)}$ where δ_z stands for the Dirac measure centered at z. The map f_{approx} is generally chosen from a class of parametric functions. As one only has N samples of f, the parameters are tuned in order to fit f_{approx} to these data by minimizing a specific loss functional.

Deep neural networks constitute a popular method for constructing f_{approx} – they are parametrized computational architectures which propagate each individual sample of the input data $\{\vec{x}_i\}_{i=1}^N \in \mathbb{R}^{d \times N}$ across a sequence of affine parametric operators and simple nonlinearities. The so-called *residual neural networks* (ResNets, [140]) may, in the simplest case, be cast as schemes of the mould

$$\begin{cases} \mathbf{x}_i^{k+1} = \mathbf{x}_i^k + \sigma\left(w^k \mathbf{x}_i^k + b^k\right) & \text{for } k \in \{0, \dots, N_{\text{layers}} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i \in \mathbb{R}^d \end{cases} \tag{7.1.1}$$

for all $i \in \{1, \dots, N\}$. The unknown states are $\mathbf{x}_i^k \in \mathbb{R}^d$ for any $i \in \{1, \dots, N\}$, σ is an explicit scalar, Lipschitz continuous nonlinear function defined componentwise in (7.1.1), $\left\{w^k, b^k\right\}_{k=0}^{N_{\text{layers}}-1}$ are optimizable parameters (controls) with $w^k \in \mathbb{R}^{d \times d}$ – called weights, and $b^k \in \mathbb{R}^d$ – called biases, and $N_{\text{layers}} \geqslant 1$ designates the number of layers referred to as the depth.

Due to the inherent dynamical systems nature of ResNets, several recent works have aimed at studying an associated continuous-time formulation in some detail, a trend started with the works [89, 129]. This perspective is motivated by the simple observation that for any $i \in \{1, \dots, N\}$ and for $T > 0$, (7.1.1) is roughly the forward Euler scheme for the neural ordinary differential equation (neural ODE)

$$\begin{cases} \dot{\mathbf{x}}_i(t) = \sigma(w(t)\mathbf{x}_i(t) + b(t)) & \text{for } t \in (0, T) \\ \mathbf{x}_i(0) = \vec{x}_i \in \mathbb{R}^d. \end{cases} \tag{7.1.2}$$

We shall focus our interest on parametrizing f_{approx} by the flows of neural ODEs such as (7.1.2). This may be done by setting $f_{\text{approx}} : x \longmapsto \mu(\mathbf{x}(T))$, where $\mathbf{x}(T)$ solves

(7.1.2) with $\mathbf{x}(0) = x$, and $\mu : \mathbb{R}^d \longrightarrow \mathcal{P}(\mathcal{Y})$ is chosen appropriately. In practice, the time-dependent parameters $[w, b]$ are found by solving the regularized empirical risk minimization problem

$$\min_{[w,b]} \underbrace{\frac{1}{N}\sum_{i=1}^{N} \operatorname{loss}\big(P\mathbf{x}_i(T), \vec{y}_i\big)}_{:=\mathcal{E}(\mathbf{x}(T))} + \Big\|[w,b]\Big\|^p_{L^p(0,T)},$$

where $p \in \{1, 2\}$, $P : \mathbb{R}^d \longrightarrow \mathbb{R}^m$ is assumed to be a given[1] affine map, and $\operatorname{loss}(\cdot,\cdot) : \mathbb{R}^m \times \mathcal{Y} \longrightarrow \mathbb{R}_+$ is such that $\mathbf{x} \longmapsto \operatorname{loss}(\mathbf{x}, y)$ is continuous for all $y \in \mathcal{Y}$, $\operatorname{loss}(\mathbf{x}, y) \neq 0$ whenever $\mu(\mathbf{x}) \neq \delta_y$, and $\operatorname{loss}(\mathbf{x}, y) \longrightarrow 0$ when $\mu_{\mathbf{x}} \longrightarrow \delta_y$ in an appropriate sense of measures (e.g. for the Wasserstein distance). Common examples of loss functions include the cross-entropy loss for classification tasks

$$\operatorname{loss}\big(P\mathbf{x}, \vec{y}\big) := -\log\left(\frac{e^{(P\mathbf{x})_{\vec{y}}}}{\sum_{j=1}^m e^{(P\mathbf{x})_j}}\right), \tag{7.1.3}$$

where $P\mathbf{x} \in \mathbb{R}^m$ and $\vec{y} \in \{1, \dots, m\}$, in which case, $\mu := \operatorname{softmax} \circ P$, or the mean squared loss for regression tasks

$$\operatorname{loss}\big(P\mathbf{x}, \vec{y}\big) := \big\|P\mathbf{x} - \vec{y}\,\big\|^2_{\ell^2}$$

where now $\vec{y} \in \mathcal{Y} \subset \mathbb{R}^m$, in which case, $\mu := P$.

As each time-step of a discretization to (7.1.2) represents a different layer of the derived neural network (7.1.1), the time horizon $T > 0$ in (7.1.2) may serve as an indicator of the number of layers N_{layers} in the discrete-time context. Thus, a good a priori knowledge of the dynamics of the learning problem over longer time horizons is desirable in view of discovering approximation and generalization properties of the trained neural ODE flow. This perspective has been taken in [95] for L^2–regularized supervised learning problem. Herein, we complete this study with new results and insights for L^1–regularized learning problems.

7.1.1 Setup

We assume we are given a training dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ where $\vec{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ and $\vec{y}_i \in \mathcal{Y}$. We henceforth set $d_x := d \times N$, and consider stacked neural ODEs of the form

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), u(t)) & \text{for } t \in (0, T) \\ \mathbf{x}(0) = \mathbf{x}^0 \in \mathbb{R}^{d_x}, \end{cases} \tag{7.1.4}$$

where $T > 0$ and $\mathbf{x}^0 = [\vec{x}_1, \dots, \vec{x}_N] \in \mathbb{R}^{d_x}$. The nonlinearity $\mathbf{f} : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \longrightarrow \mathbb{R}^{d_x}$ may take the form

$$\mathbf{f}(\mathbf{x}, u) = \sigma\left(\begin{bmatrix} w & & \\ & \ddots & \\ & & w \end{bmatrix}\mathbf{x} + \begin{bmatrix} b \\ \vdots \\ b \end{bmatrix}\right) \tag{7.1.5}$$

for $\mathbf{x} \in \mathbb{R}^{d_x}$ and $u = [w, b] \in \mathbb{R}^{d_u}$ with $d_u := d^2 + d$, and $\sigma \in \operatorname{Lip}(\mathbb{R})$ is defined componentwise so that each component of $\mathbf{f}$ coincides with the canonical neural ODE given in (7.1.2). Permutations may also be considered, e.g.

$$\mathbf{f}(\mathbf{x}, u) = \begin{bmatrix} w & & \\ & \ddots & \\ & & w \end{bmatrix}\sigma(\mathbf{x}) + \begin{bmatrix} b \\ \vdots \\ b \end{bmatrix}. \tag{7.1.6}$$

[1] In practice, P is either part of the optimizable parameters, or may be chosen at random. Whilst we fix P for technical purposes, numerical experiments indicate that the results presented in what follows persist when P is optimized as well.

The key assumption we make in what follows is that $\mathbf{f}$ is 1–homogeneous with respect to the parameters u, i.e.

$$\mathbf{f}(\mathbf{x}, \alpha u) = \alpha \mathbf{f}(\mathbf{x}, u) \qquad \text{for all } (\mathbf{x}, u) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \text{ and } \alpha > 0. \tag{7.1.7}$$

This is clearly the case for $\mathbf{f}$ parametrized as in (7.1.6), whilst for (7.1.5), we shall moreover assume that σ is 1–homogeneous – a canonical example of such an activation function is the ReLU $\sigma(x) = \max\{x, 0\}$.

Remark 7.1.1. *Since $\sigma \in \mathrm{Lip}(\mathbb{R})$, for any $\mathbf{x}^0 \in \mathbb{R}^{d_x}$ and $u \in L^1(0, T; \mathbb{R}^{d_u})$, (7.1.4) with $\mathbf{f}$ as above admits a unique solution $\mathbf{x} \in C^0([0, T]; \mathbb{R}^{d_x})$. This can be shown by combining a fixed point and Grönwall argument to the integral formulation of (7.1.4).*

The supervised learning problem we address in this work consists in minimizing, for $T > 0$, a functional of the form

$$J_T(u) := \int_0^T \mathcal{E}(\mathbf{x}(t))\, \mathrm{d}t + \int_0^T \|u(t)\|_1\, \mathrm{d}t, \tag{7.1.8}$$

over $u = [w, b] \in \mathfrak{U}_{\mathrm{ad},T}$, where $\mathcal{E}$ denotes the empirical risk defined by

$$\mathcal{E}(\mathbf{x}(t)) := \frac{1}{N} \sum_{i=1}^{N} \mathrm{loss}(P\mathbf{x}_i(t), \vec{y}_i). \tag{7.1.9}$$

Here $\mathbf{x} \in C^0([0, T]; \mathbb{R}^{d_x})$ solves (7.1.4), $P : \mathbb{R}^d \longrightarrow \mathbb{R}^m$ is a given affine map, and

$$\mathfrak{U}_{\mathrm{ad},T} := \Big\{ u \in L^1(0, T; \mathbb{R}^{d_u}) \colon \ \|u(t)\|_1 \leqslant M \ \text{ a.e. in } (0, T) \Big\}$$

for a fixed thresholding constant $M > 0$[2]. Finally, $\mathrm{loss}(\cdot, \cdot) : \mathbb{R}^m \times \mathcal{Y} \longrightarrow \mathbb{R}_+$ is assumed to satisfy

$$\mathrm{loss}(\cdot, y) \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}^m; \mathbb{R}_+) \qquad \text{and} \qquad \inf_{\mathbf{x} \in \mathbb{R}^m} \mathrm{loss}(\mathbf{x}, y) = 0, \qquad \text{for any } y \in \mathcal{Y}. \tag{7.1.10}$$

This is the case for most losses considered in practice, including the ones defined in the introduction that precedes. We shall make use of the entry-wise ℓ^1–norm $\|\cdot\|_1$ on $\mathbb{R}^{d_u}$, defined as $\|u\|_1 := \sum_{k=1}^{d_u} |u_k|$ for $u = [u_1, \ldots, u_{d_u}] \in \mathbb{R}^{d_u}$. We emphasize that our results would clearly hold for different norms on $\mathbb{R}^{d_u}$ (e.g. the euclidean norm or max norm) by the equivalence of norms.

7.1.2 Main result

We will be interested in studying the behavior of global minimizers to (7.1.8) and the corresponding solutions to (7.1.4). Due to the fact that the empirical risk $\mathcal{E}(\mathbf{x}(t))$ is regularized over the entire time interval $[0, T]$, one expects that any minimizer u_T steers the trajectories – as fast as possible – to a configuration for which $\mathcal{E}(\mathbf{x}_T(t))$ is small, and then remain in that configuration by using parameters of small amplitude, or eventually, no parameters at all.

Throughout the paper, we will assume that the neural ODE can interpolate the dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$, either in finite or in infinite time, namely, we shall suppose that there exist parameters such that its corresponding trajectory makes the training error $\mathcal{E}$ defined in (7.1.9) vanish, either in finite or in infinite time.

Definition 7.1.2 (Interpolation)**.** *Let $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ be a given dataset with $\vec{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ and $\vec{y}_i \in \mathcal{Y}$.*

[2]The L^1–regularization in (7.1.8) enforces the use of sparse parameters concentrated near $t = 0$. We include an L^∞–constraint in the definition of $\mathfrak{U}_{\mathrm{ad},T}$ in order to prevent degeneracy.

1. *We say that* (7.1.4) interpolates *the dataset* $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ *in time* $T_0 > 0$ *if there exists* $u \in L^\infty(0, T_0; \mathbb{R}^{d_u})$ *such that the corresponding solution* $\mathbf{x} \in C^0([0, T_0]; \mathbb{R}^{d_x})$ *to* (7.1.4) *satisfies*
$$\mathcal{E}(\mathbf{x}(T_0)) = 0.$$

2. *We say that* (7.1.4) asymptotically interpolates *the dataset* $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ *if there exist* $T_0 > 0$*, a function* $h \in C^\infty([T_0, \infty); \mathbb{R}_+)$ *satisfying*
$$\dot{h}(t) < 0 \qquad \text{for } t \geqslant T_0 \qquad \text{and} \qquad \lim_{t \longrightarrow \infty} h(t) = 0,$$
and $u \in L^\infty(\mathbb{R}_+; \mathbb{R}^{d_u})$ *such that the corresponding solution* $\mathbf{x} \in C^0([0, \infty); \mathbb{R}^{d_x})$ *to* (7.1.4) *satisfies*
$$\mathcal{E}(\mathbf{x}(t)) \leqslant h(t) \qquad \text{for } t \geqslant T_0.$$

We consider asymptotic interpolation due to the occurrence of non-coercive losses which do not attain their minimum, exemplified in the context of classification tasks with losses such as the cross-entropy defined in (7.1.3). In fact, in Proposition 7.4.2 below, we prove that, under suitable assumptions, the asymptotic interpolation property for the cross-entropy loss holds with
$$h(t) = \log\left(1 + (m-1)e^{-\gamma e^t}\right),$$
where $\gamma > 0$ is the *margin* defined by (we set $[m] := \{1, \dots, m\}$)
$$\gamma := \min_{i \in [N]} \left\{ P\mathbf{x}_i(T_0)_{\vec{y}_i} - \max_{\substack{j \in [m] \\ j \neq \vec{y}_i}} P\mathbf{x}_i(T_0)_j \right\}.$$

On another hand, (finite-time) interpolation can be shown to hold, for instance, for $\mathbf{f}$ as in (7.1.6) with loss attaining its minimum 0 and P surjective – see [75, 95, 237] for results in this direction.

We are in position to state our main result, which ensures that any minimizer u_T of J_T is sparse in the sense that $u_T \equiv 0$ on (T^*, T) for some $T^* \in (0, T]$. Moreover, under interpolation assumptions, we may provide estimates on the stopping time T^* and the training error $\mathcal{E}(\mathbf{x}_T(T^*))$.

Theorem 7.1. *Let* $T > 0$ *and* $M > 0$ *be fixed, and let* $u_T \in \mathfrak{U}_{\mathrm{ad},T}$ *be any (should it exist*[3]*) global minimizer to* J_T *defined in* (7.1.8)*, with* $\mathcal{E}$ *as in* (7.1.9)*,* loss *satisfying* (7.1.10) *and* $\mathbf{f}$ *satisfying* (7.1.7)*. Let* $\mathbf{x}_T \in C^0([0, T]; \mathbb{R}^{d_x})$ *denote the corresponding solution to* (7.1.4)*. Then, there exists a time* $T^* \in (0, T]$ *such that*
$$\begin{aligned} &\|u_T(t)\|_1 = M && \text{for a.e. } t \in (0, T^*), \\ &\|u_T(t)\|_1 = 0 && \text{for a.e. } t \in (T^*, T) \end{aligned} \tag{7.1.11}$$
and
$$\mathcal{E}(\mathbf{x}_T(T^*)) \leqslant \mathcal{E}(\mathbf{x}_T(t)) \qquad \text{for } t \in [0, T]. \tag{7.1.12}$$
Moreover,

1. *If system* (7.1.4) *interpolates the dataset in some time* $T_0 > 0$ *as per Definition 7.1.2, then there exists a time* $T(M) > 0$ *and a constant* $\mathfrak{C}(M) > 0$*, both independent of* T*, such that*
$$T^* \leqslant T(M) \qquad \text{and} \qquad \mathcal{E}(\mathbf{x}_T(T^*)) \leqslant \frac{\mathfrak{C}(M)}{T}.$$

[3] One can show that a minimizer exists when $\mathbf{f}$ is as in (7.1.6) by means of the direct method in the calculus of variations. However, for $\mathbf{f}$ as in (7.1.5), ensuring compactness does not appear straightforward.

2. *If system (7.1.4) asymptotically interpolates the dataset as per Definition 7.1.2, there exists a constant* $\mathfrak{C}(M) > 0$ *independent of* T *such that*

$$T^* \leqslant \frac{\mathfrak{C}(M)}{M} h^{-1}\left(\frac{1}{T}\right) + \frac{1}{M}$$

and

$$\mathcal{E}(\mathbf{x}_T(T^*)) \leqslant \frac{\mathfrak{C}(M)}{T} h^{-1}\left(\frac{1}{T}\right) + \frac{1}{T},$$

where h^{-1} *denotes the inverse function of* h.

Remark 7.1.3. *A couple of pertinent remarks are in order.*

- *Observe that when* $h(t)$ *in Definition 7.1.2 is such that* $h(t) \sim o\left(T^{-1}\right)$ *as* $t \sim \infty$, *Theorem 7.1–*(2) *implies that*

$$T^* = o(T) \qquad \text{and} \qquad \mathcal{E}(\mathbf{x}_T(T^*)) = o(1).$$

- *On another hand we also observe that, having the stopping time* T^*, *or at least an upper bound of it, allows one to reduce the supervised learning problem to an equivalent one over a shorter time-horizon but with a final cost, namely minimizing a functional of the form*

$$J_T^*(u) := \int_0^{T^*} \mathcal{E}(\mathbf{x}(t))\,\mathrm{d}t + \int_0^{T^*} \|u(t)\|_1\,\mathrm{d}t + (T - T^*)\mathcal{E}(\mathbf{x}(T^*)).$$

When extrapolated to the discrete-time, ResNet context, a shorter time-horizon in the optimal control problem can be interpreted as considering a shallower ResNet, which naturally lowers the computational cost of the training process.

Remark 7.1.4 (Dimension reduction)**.** *A related concept to sparsity is that of* coordinate-wise sparsity *– sometimes referred to as* switching *–, which is described by the property*

$$u_j(t)u_k(t) = 0 \qquad \text{for } j,k \in [d_u],\ j \neq k \qquad \text{and} \qquad \text{for a.e. } t \in (0,T).$$

In other words, this entails that at most one component of $u(t)$ *is non-zero at time* t. *We refer the reader to the work of Zuazua [282] for a comprehensive overview of switching in the context of linear systems (both finite and infinite dimensional). In [153], the authors study the occurrence of coordinate-wise sparsity for infinite-time horizon optimal control problems for nonlinear ODE systems, and stipulate that such a property occurs when one considers a parameter regularization term of the form*

$$\int_0^T \left(\|u(t)\|_1 + 2 \sum_{\substack{j,k \in [d_u] \\ j \neq k}} |u_j(t)u_k(t)|^{1/2} \right) \mathrm{d}t = \int_0^T \left(\sum_{j=1}^{d_u} |u_j(t)|^{1/2} \right)^2 \mathrm{d}t.$$

We refer to [11, 145, 152] for further related works on optimal control problems with L^1*–regularization terms, but which do not apply to our setup due to underlying assumptions on linearity or inifinite-time horizons.*

The interest of this discussion stems from the possibility of interpreting coordinate-wise sparsity and switching as allowing the flow to alternate dimensions over different time instances in the discrete, ResNet context, which could allow for a variable width interpretation of the neural ODE models. The role of ℓ^1*–regularization in signal compression and dimension reduction by inducing sparsity is well explored (see, for instance, the seminal works [80, 81, 43, 44]). Since our methodology for the proof of Theorem 1.10 essentially relies on the homogeneity of the neural ODE with respect to the parameters and the invariance of the* $L^1(0,T;\mathbb{R}^{d_u})$*–norm with respect to the induced scaling, it could be plausible*

to stipulate the occurrence of coordinate-wise sparsity in our finite-time horizon context by adapting our arguments presented to the parameter regularization just above, which is also invariant by the induced scaling. We however leave the proof for a forthcoming work.

In Figure 7.1 – Figure 7.4 below, we visualize[4] the conclusions of Theorem 7.1 on a toy binary classification task ($\mathcal{Y} := \{-1, 1\}$) for the neural ODE (7.1.4) – (7.1.5) (we use the scheme given in (7.1.1)), with $\sigma \equiv \tanh$, and using cross-entropy loss. We set $T = 15$, $M = 2.75$, and work with the training dataset displayed in Figure 7.3 consisting of $N = 3000$ samples. We use an elementary trapezoidal rule for discretizing the intervening integrals.

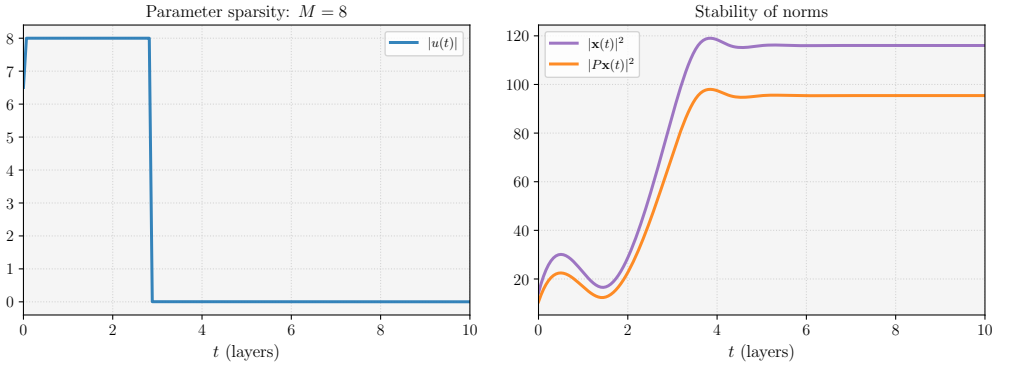


Figure 7.1: We depict a manifestation of the first part of Theorem 7.1 for a binary classification task in the setup presented just above. *Left:* the sparsity of the optimal parameters $u_T = [w_T, b_T]$ over time/layer with $M = 8$; *Right:* The norms of the associated state trajectory and projected output (see Figure 7.3). One notes a phase transition at the stopping time $T^* \sim 3$.

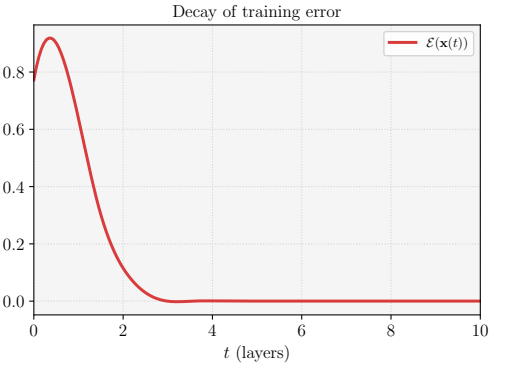


Figure 7.2: We also depict a manifestation the second part of Theorem 7.1, which stipulates a bound of the training error at the stopping time $T^* \sim 3$ – we in fact see that the training error stabilizes beyond the stopping time.

7.1.3 Related work

We give a brief overview on some related literature.

Sparse approximation via neural networks. There is a plethora of works in the literature on approximation theory regarding the *universal approximation* properties of multi-layer perceptrons. In [27] for instance, the authors derive lower bounds on the connectivity and the memory requirements of multi-layer perceptrons guaranteeing uniform

[4] Software experiments were done using `PyTorch` [214] (and may be found at `https://github.com/borjanG/dynamical.systems`), using the Adam optimizer [159] and `TorchDiffEq` library [61]. Experiments were conducted on a personal MacBook Pro laptop (2.4 GHz Quad-Core Intel Core i5, 16GB RAM, Intel Iris Plus Graphics 1536 MB).

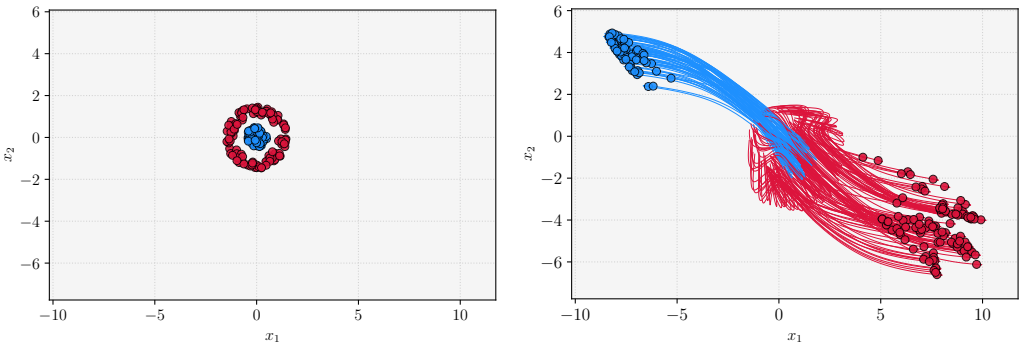


Figure 7.3: We depict the evolution of the state trajectories of the neural ODE, in the setting of Figure 7.1. *Left:* Initial configuration of training data; *Right:* Evolution and the final configuration $\mathbf{x}_i(T)$ of the trajectories, for $i \in [N]$.

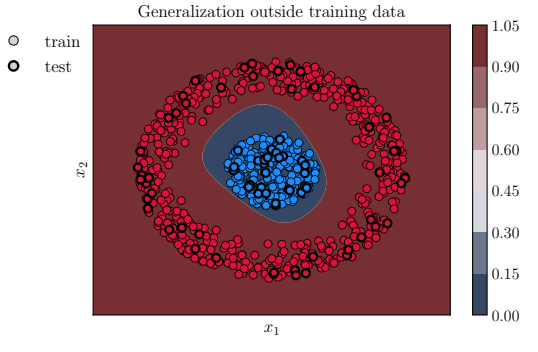


Figure 7.4: We see that the sparsely trained neural ODE flow captures the shape of the training dataset, and accurately classifies the test data.

approximation rates for arbitrary function classes in $L^2(\mathbb{R}^d)$. A key caveat, as for all universal approximation results, is that there is no guarantee that the training algorithm will find the constructed parameters exhibiting the approximation property. On another hand, our result, albeit specific to the dataset one considers, is guaranteed for the global minimizer, which may be found by training.

Neural ODEs. The continuous-time neural ODE formulation of deep neural networks has been used to great effect for improving computational training performance – for instance, by using adaptive ODE solvers [61, 86] or indirect training algorithms based on the Pontryagin Maximum Principle [181, 25] –, and also for modeling purposes, including irregular time series modeling [234], and generative modeling through normalizing flows [120, 60]. It should be noted that the origins of continuous-time supervised learning go back to the 1980s – the neural network model proposed in [141] is a differential equation, whereas in [179] back-propagation is connected to the adjoint method arising in optimal control. Related works include studies on identification of the weights from data [6, 7] and controllability of continuous-time recurrent networks [247, 248].

Long-time optimal control. The behavior displayed in Theorem 7.1 is somewhat reminiscent of the well-known *turnpike property* in optimal control and economics: over long time horizons, the optimal pair $(u_T(t), \mathbf{x}_T(t))$ should be "near" the optimal steady pair (u_s, x_s), namely a solution to the problem

$$\inf_{(u_s,x_s)\in\mathbb{R}^{d_u}\times\mathbb{R}^{d_x}} \mathcal{E}(\mathbf{x}_s) + \|u\|_1 \qquad \text{subject to } \mathbf{f}(\mathbf{x}_s, u) = 0.$$

Note that, due to the 1-homogeneity of $\mathbf{f}$, we have $\mathbf{f}(\mathbf{x}, 0) = 0$ for all $\mathbf{x} \in \mathbb{R}^d$. Hence, at least when $\mathcal{E}$ attains its minimum (e.g. regression tasks), we can deduce that $(0, \mathbf{x}^\dagger)$, where $\mathbf{x}^\dagger \in \mathbb{R}^{d_x}$ is any zero of $\mathcal{E}$, designates an optimal stationary solution. In [95, 96], the authors prove that, when an L^2–regularization term for the control is considered in

the functional J_T, an exponential decay estimate for the training error and the optimal parameters can be obtained at any time $t \in [0, T]$. Thus, at least heuristically, whereas in Theorem 7.1 one sees a sharp phase transition for the optimal parameters at time T^*, in the L^2–regularized results, this transition is somehow diffused and compensated by an exponential decay.

We also refer the reader to the recent work [126] for a study of linear problems with L^1–penalties.

7.1.4 Outline

The remainder of the paper is organized as follows. In Section 7.2, we provide some necessary backbone results which mainly rely on the homogeneity of the neural ODE with respect to the parameters. We tackle the proof of Theorem 7.1 in Section 7.3. We provide a proof of the asymptotic interpolation property for ReLU activated neural ODEs for the cross-entropy loss in Section 7.4.

7.2 Preliminary lemmas

A key point in our forthcoming arguments is the possibility of scaling a trajectory of a neural ODE set in a time-interval $[0, T_0]$ into a trajectory of the same neural ODE but set on a time-interval $[0, T_1]$.

Lemma 7.2.1. *Let $T_0 > 0$, $x_0 \in \mathbb{R}^d$, $u_{T_0} \in L^1(0, T_0; \mathbb{R}^{d_u})$, and let $\mathbf{x}_{T_0} \in C^0([0, T_0]; \mathbb{R}^{d_x})$ be the unique solution to (7.1.4) on $[0, T_0]$ associated to u_{T_0}. Let $T > 0$, and define*

$$u_T(t) := \frac{T_0}{T} u_{T_0}\left(t\frac{T_0}{T}\right) \qquad \text{for } t \in [0, T],$$

and

$$\mathbf{x}_T(t) := \mathbf{x}_{T_0}\left(t\frac{T_0}{T}\right) \qquad \text{for } t \in [0, T].$$

Then $\mathbf{x}_T \in C^0([0, T]; \mathbb{R}^{d_x})$ is the unique solution to (7.1.4) associated to u_T.

Such time-scaling arguments are standard in the context of driftless control affine systems, and is used, for instance, in the proof of the Chow-Rashevskii theorem, see [69, Chapter 3, Section 3.3]. We omit the proof, which is straightforward.

We now state and prove a result which ensures that when minimizing a functional of the form J_T defined in (7.1.8) over $\mathfrak{U}_{\mathrm{ad},T}$, one only needs to take into account sparse parameters saturating the constraint $\|u(t)\|_1 \leqslant M$ until the function $t \longmapsto \mathcal{E}(\mathbf{x}(t))$ reaches its minimum over the interval $[0, T]$.

Let us begin by making precise what we mean by sparse parameters.

Definition 7.2.2 (Sparse parameters)**.** *Let $M > 0$ and $0 < T^* \leqslant T$ be fixed. We say that the parameters $u = [w, b] \in \mathfrak{U}_{\mathrm{ad},T}$ are* sparse *in $(0, T^*)$ if*

$$\|u(t)\|_1 = M \qquad \text{for a.e. } t \in (0, T^*), \tag{7.2.1}$$
$$\|u(t)\|_1 = 0 \qquad \text{for a.e. } t \in (T^*, T). \tag{7.2.2}$$

For any $T^* > 0$, we denote by $\mathfrak{U}_{\mathrm{sp},T^*}$ the subset of $\mathfrak{U}_{\mathrm{ad},T}$ consisting of parameters which are sparse in $(0, T^*)$.

Proposition 7.2.3. *Let $\mathrm{loss}(\cdot,\cdot) : \mathbb{R}^m \times \mathcal{Y} \longrightarrow \mathbb{R}_+$ satisfy (7.1.10), and let $T > 0$ and $M > 0$ be fixed. Let $u_T = [w_T, b_T] \in \mathfrak{U}_{\mathrm{ad},T}$ be a global minimizer of J_T defined in (7.1.8),*

and let $\mathbf{x}_T$ *be the corresponding unique solution to* (7.1.4)*. Then* $u_T \in \mathfrak{U}_{\mathrm{sp},T^*}$*, where* T^* *is defined*[5] *as*

$$T^* := \min\left\{t \in [0,T]\colon \mathcal{E}(\mathbf{x}_T(t)) = \min_{s\in[0,T]} \mathcal{E}(\mathbf{x}_T(s))\right\}.$$

The core of the proof lies in the following lemma, which ensures that if an admissible pair of parameters does not saturate the L^∞–constraint before some time T^*, then it can always be improved by means of the scaling property from Lemma 7.2.1.

Lemma 7.2.4. *Let* $u_T \in \mathfrak{U}_{\mathrm{ad},T}$ *and* $T^* > 0$ *be as in Proposition 7.2.3. Assume that, for some* $\theta \in (0,1)$*, there exists a finite collection of disjoint non-empty intervals* $\{(a_i,b_i)\}_{i=1}^{\mathfrak{I}}$ *with* $(a_i,b_i) \subset (0,T^*)$ *for which*

$$\|u_T(t)\|_1 \leqslant (1-\theta)M \qquad \text{for a.e. } t \in \bigcup_{i=1}^{\mathfrak{I}}(a_i,b_i), \tag{7.2.3}$$

and

$$\mathcal{E}(\mathbf{x}_T(t)) - \mathcal{E}(\mathbf{x}_T(T^*)) \geqslant \theta \qquad \text{for all } t \in \bigcup_{i=1}^{\mathfrak{I}}(a_i,b_i) \tag{7.2.4}$$

hold. Then, there exist parameters $\overline{u} \in \mathfrak{U}_{\mathrm{ad},T}$ *satisfying*

$$\overline{u}(t) = 0 \qquad \text{for a.e. } t \in (T^*-\tau, T), \tag{7.2.5}$$

and

$$J_T(\overline{u}) \leqslant J_T(u_T) - \theta\tau,$$

where

$$\tau := \theta \sum_{i=1}^{\mathfrak{I}}(b_i - a_i).$$

We may now provide a proof to Proposition 7.2.3.

Proof of Proposition 7.2.3. We argue by contradiction. Suppose that $u_T \in \mathfrak{U}_{\mathrm{ad},T}$ is a global minimizer of J_T such that $u_T \notin \mathfrak{U}_{\mathrm{sp},T^*}$, where $T^* > 0$ is defined as in the statement. Hence, either condition (7.2.1) or condition (7.2.2) do not hold.

Case 1: (7.2.2) does not hold. Let us suppose that condition (7.2.2) does not hold. Consider $\overline{u} \in \mathfrak{U}_{\mathrm{ad},T}$ defined as

$$\overline{u}(t) = \begin{cases} u_T(t) & \text{for } t \in [0,T^*] \\ 0 & \text{for } t \in (T^*,T]. \end{cases}$$

By the 1-homogeneity of $\mathbf{f}$ with respect to u, we have $\mathbf{f}(\cdot,0) \equiv 0$, and so

$$\overline{\mathbf{x}}(t) = \overline{\mathbf{x}}(T^*) = \mathbf{x}_T(T^*), \qquad \text{for } t \in [T^*,T].$$

In view of the definition of T^*, the above identity implies that

$$\int_0^T \mathcal{E}(\overline{\mathbf{x}}(t))\,\mathrm{d}t \leqslant \int_0^T \mathcal{E}(\mathbf{x}_T(t))\,\mathrm{d}t.$$

In addition, the fact that (7.2.2) does not hold implies that

$$\begin{aligned}\int_0^T \|\overline{u}(t)\|_1\,\mathrm{d}t &= \int_0^{T^*} \|u_T(t)\|_1\,\mathrm{d}t \\ &< \int_0^{T^*} \|u_T(t)\|_1\,\mathrm{d}t + \int_{T^*}^T \|u_T(t)\|_1\,\mathrm{d}t = \int_0^T \|u_T(t)\|_1\,\mathrm{d}t.\end{aligned}$$

[5] Note that the min defining T^* is clearly well defined, as the set in question is bounded, and also closed as the preimage of the singleton $\left\{\min_{s\in[0,T]} \mathcal{E}(\mathbf{x}_T(s))\right\}$ under the continuous map $t \longmapsto \mathcal{E}(\mathbf{x}(t))$.

We thus deduce that $J_T(\overline{u}) < J_T(u_T)$, which contradicts the optimality of u_T.

Case 2: (7.2.1) does not hold. If (7.2.1) is not fulfilled, then there must exist $\theta \in (0,1)$ such that the set

$$\mathbf{A}_\theta := \left\{ t \in (0,T^*) : \ \|u_T(t)\|_1 \leqslant (1-\theta)M \right\}$$

has positive Lebesgue measure, namely $\mu(\mathbf{A}_\theta) > 0$ where $\mu(\cdot)$ henceforth denotes the Lebesgue measure. Now, note that for a fixed $\delta \in (0, \mu(\mathbf{A}_\theta))$, using elementary set theory we have

$$\mathbf{A}_\theta \cap (0, T^* - \delta) = \mathbf{A}_\theta \setminus \Big((0,T^*) \setminus [T^* - \delta, T^*) \Big) = \mathbf{A}_\theta \setminus [T^* - \delta, T^*),$$

whence the set

$$\mathbf{B}_\theta := \mathbf{A}_\theta \cap (0, T^* - \delta)$$

also has positive Lebesgue measure: $\mu(\mathbf{B}_\theta) > 0$. By classical results in Lebesgue measure theory (see [273, Thm. 3.25]), for all $\varepsilon > 0$ there exists a finite collection of disjoint nonempty intervals $\{(a_i, b_i)\}_{i=1}^{n(\varepsilon)}$, with $(a_i, b_i) \subset (0, T^* - \delta)$, such that the set

$$\mathbf{O}_\varepsilon := \bigcup_{i=1}^{n} (a_i, b_i)$$

satisfies

$$\mu\left(\mathbf{O}_\varepsilon \setminus \mathbf{B}_\theta\right) < \varepsilon \quad \text{and} \quad \mu\left(\mathbf{B}_\theta \setminus \mathbf{O}_\varepsilon\right) < \varepsilon. \tag{7.2.6}$$

This implies in particular that

$$\mu\left(\mathbf{O}_\varepsilon\right) > \mu(\mathbf{B}_\theta) - \varepsilon. \tag{7.2.7}$$

Now let $\varepsilon \in (0, \mu(\mathbf{B}_\theta))$ be arbitrary and to be fixed later, and let $\{(a_i, b_i)\}_{i=1}^{n(\varepsilon)}$ be the corresponding collection of disjoint intervals satisfying (7.2.6), with $\mathbf{O}_\varepsilon$ denoting the union of these intervals as defined above. Set

$$u^\varepsilon(t) := \begin{cases} u_T(t) & \text{for } t \in (0,T) \setminus (\mathbf{O}_\varepsilon \setminus \mathbf{B}_\theta) \\ 0 & \text{for } t \in \mathbf{O}_\varepsilon \setminus \mathbf{B}_\theta. \end{cases}$$

Since $u_T \in \mathfrak{U}_{\mathrm{ad},T}$, it may be seen that

$$\|u^\varepsilon(t)\|_1 \leqslant M \qquad \text{for a.e. } t \in (0,T).$$

Now let $\mathbf{x}^\varepsilon$ denote the solution to (7.1.4) associated to u^ε. By virtue of the specific form of $\mathbf{f}$, the Lipschitz continuity of σ, and the Grönwall inequality, we may readily deduce that there exists a constant $C_1 = C_1(T, M, N) > 0$ independent of ε such that

$$\left\|\mathbf{x}^\varepsilon(t) - \mathbf{x}_T(t)\right\|_1 \leqslant C_1 \int_0^T \left\|u^\varepsilon(s) - u_T(s)\right\|_1 \mathrm{d}s \qquad \text{for } t \in [0,T]. \tag{7.2.8}$$

On the other hand, by using (7.2.6), we also deduce that

$$\int_0^T \left\|u^\varepsilon(s) - u_T(s)\right\|_1 \mathrm{d}s \leqslant M\mu\left(\mathbf{O}_\varepsilon \setminus \mathbf{B}_\theta\right) < M\varepsilon. \tag{7.2.9}$$

Combining (7.2.8) and (7.2.9) leads us to

$$\left\|\mathbf{x}^\varepsilon(t) - \mathbf{x}_T(t)\right\|_1 < C_1 M \varepsilon \qquad \text{for } t \in [0,T].$$

Now clearly, since $\mathbf{x}_T \in C^0([0,T]; \mathbb{R}^{d_x})$, the stacked trajectory $\mathbf{x}_T(t)$ remains in a compact set of $\mathbb{R}^{d_x}$ for all $t \in [0,T]$. Hence, by the locally Lipschitz character of $\mathrm{loss}(\cdot, \vec{y})$, the estimate

$$\left| \mathcal{E}\left(\mathbf{x}^\varepsilon(t)\right) - \mathcal{E}\left(\mathbf{x}_T(t)\right) \right| \leqslant C_2 M \varepsilon, \tag{7.2.10}$$

holds for some $C_2 = C_2(T, M, N, \text{loss}) > 0$ independent of ε, and for all $t \in [0, T]$. On the other hand, using only the definition of T^*, one sees that there exists some $\gamma > 0$ such that

$$\mathcal{E}(\mathbf{x}_T(t)) - \gamma \geqslant \mathcal{E}(\mathbf{x}_T(T^*)) \qquad \text{for all } \ t \in [0, T^* - \delta]. \tag{7.2.11}$$

Estimate (7.2.10) combined with (7.2.11) yields

$$\mathcal{E}\big(\mathbf{x}^\varepsilon(t)\big) - \mathcal{E}\big(\mathbf{x}_T(T^*)\big) \geqslant \gamma - C_2\, M\varepsilon \qquad \text{for all } t \in [0, T^* - \delta].$$

Since $\varepsilon \in (0, \mu(\mathbf{B}_\theta))$ is arbitrary, we choose ε small enough to ensure that $\gamma - C_2 M\varepsilon > 0$. Setting

$$\theta^* := \min\{\theta, \gamma - C_2 M\varepsilon\},$$

we observe that, by definition, u^ε satisfies

$$\|u^\varepsilon(t)\|_1 \leqslant (1 - \theta^*)\, M, \qquad \text{for a.e. } t \in \mathbf{O}_\varepsilon,$$

and moreover,

$$\mathcal{E}\big(\mathbf{x}^\varepsilon(t)\big) - \mathcal{E}\big(\mathbf{x}_T(T^*)\big) \geqslant \theta^* \qquad \text{for all } t \in \mathbf{O}_\varepsilon$$

holds. We may thus apply Lemma 7.2.4, which ensures the existence of parameters $\overline{u}^\varepsilon$ for which

$$J_T(\overline{u}^\varepsilon) \leqslant J_T(u^\varepsilon) - (\theta^*)^2 \mu(\mathbf{O}_\varepsilon) \tag{7.2.12}$$

holds. As a consequence of (7.2.9) and (7.2.10), we have

$$J_T(u^\varepsilon) \leqslant J_T(u_T) + (1 + C_2 T)\, M\varepsilon,$$

which, when combined with (7.2.12) and (7.2.7), yields

$$J_T(\overline{u}^\varepsilon) < J_T(u_T) + (1 + C_2 T) M\varepsilon - (\theta^*)^2(\mu(\mathbf{B}_\theta) - \varepsilon).$$

Looking at the above inequality, we may note that, by choosing $\varepsilon > 0$ sufficiently small (for instance $\varepsilon \leqslant \frac{(\theta^*)^2 \mu(\mathbf{B}_\theta)}{(1 + C_2 T) M}$ suffices), we may ensure that

$$J_T(\overline{u}_{\varepsilon,n}) < J_T(u_T),$$

which contradicts the optimality of u_T. This concludes the proof. □

We conclude this section with a proof of Lemma 7.2.4.

Proof of Lemma 7.2.4. We will argue by induction over the number of intervals $\mathfrak{I} \geqslant 1$, constructing appropriately the parameters $\overline{u}$ explicitly in each step via affine transformations of u_T – the desired estimates will follow by using the time-scaling invariance of the L^1–norm of the parameters.

Step 1). **Initialization.** Let us first assume that $\mathfrak{I} = 1$. Consider

$$\overline{u}(t) := \begin{cases} u_T(t) & \text{for } \ t \in (0, a_1) \\ \dfrac{b_1 - a_1}{c_1 - a_1} u_T\left((t - a_1)\dfrac{b_1 - a_1}{c_1 - a_1} + a_1\right) & \text{for } \ t \in [a_1,\, c_1) \\ u_T\,(t + b_1 - c_1) & \text{for } t \in [c_1,\, T^* - (b_1 - c_1)), \\ 0 & \text{for } t \in [T^* - (b_1 - c_1),\, T), \end{cases}$$

where $c_1 \in (a_1, b_1)$ is chosen so that

$$\frac{b_1 - a_1}{c_1 - a_1}(1 - \theta) = 1,$$

which is equivalent to

$$b_1 - c_1 = \theta(b_1 - a_1) =: \tau.$$

Observe that as a consequence of (7.2.3), the parameters $\overline{u}(t)$ satisfy the constraint $\|\overline{u}(t)\|_1 \leqslant M$ for a.e. $t \in (0,T)$. In addition, by virtue of the choice of c_1, and the definition of τ, $\overline{u}(t)$ also satisfies (7.2.5). Now, making use of the scaling provided by Lemma 7.2.1, and the fact that $\mathbf{f}(\cdot,0) \equiv 0$, one can check that the state trajectory $\overline{\mathbf{x}}(t)$ associated to $\overline{u}(t)$ is exactly given by

$$\overline{\mathbf{x}}(t) = \begin{cases} \mathbf{x}_T(t) & \text{for } \ t \in [0, a_1) \\ \mathbf{x}_T\left((t-a_1)\dfrac{b_1-a_1}{c_1-a_1} + a_1\right) & \text{for } \ t \in [a_1,\, c_1) \\ \mathbf{x}_T\left(t + b_1 - c_1\right) & \text{for } \ t \in [c_1,\, T^* - (b_1 - c_1)), \\ \mathbf{x}_T(T^*) & \text{for } \ t \in [T^* - (b_1 - c_1),\, T]. \end{cases}$$

Moreover, observe that since $\tau := b_1 - c_1$,

$$\mathcal{E}(\overline{\mathbf{x}}(t)) = \mathcal{E}(\mathbf{x}_T(T^*)) \qquad \text{for } t \in [T^* - \tau, T]. \tag{7.2.13}$$

Let us now evaluate the functional J_T along $\overline{u}$. We start by computing the L^1–norm of $\overline{u}$:

$$\begin{aligned} \|\overline{u}\|_{L^1(0,T;\mathbb{R}^{d_u})} &= \int_0^{a_1} \|u_T(t)\|_1 \,\mathrm{d}t + \int_{c_1}^{T^*-(b_1-c_1)} \|u_T(t+b_1-c_1)\|_1 \,\mathrm{d}t \\ &\quad + \frac{b_1-a_1}{c_1-a_1}\int_{a_1}^{c_1} \left\| u_T\left((t-t_1)\frac{b_1-a_1}{c_1-a_1}+a_1\right)\right\|_1 \mathrm{d}t \\ &\leqslant \|u_T\|_{L^1(0,T;\mathbb{R}^{d_u})}\,, \end{aligned} \tag{7.2.14}$$

where we made use of the elementary changes of variables

$$t \longmapsto t + (b_1 - c_1) \quad \text{ and } \quad t \longmapsto (t-a_1)\frac{b_1-a_1}{c_1-a_1} + a_1$$

for the second and third integral respectively. On the other hand, by virtue of (7.2.13), the assumption (7.2.4), and the fact that

$$\mathcal{E}(\mathbf{x}_T(T^*)) = \min_{s\in[0,T]} \mathcal{E}(\mathbf{x}_T(s))$$

via definition of T^*, the same chain of change of variables as above can be used to estimate the integral of the training error in as follows:

$$\begin{aligned} \int_0^T \Big(\mathcal{E}(\overline{\mathbf{x}}(t)) - \mathcal{E}\big(\mathbf{x}_T(T^*)\big)\Big)\,\mathrm{d}t &= \int_0^{a_1} \Big(\mathcal{E}(\mathbf{x}_T(t)) - \mathcal{E}\big(\mathbf{x}_T(T^*)\big)\Big)\,\mathrm{d}t \\ &\quad + \underbrace{\frac{c_1-a_1}{b_1-a_1}}_{1-\theta}\int_{a_1}^{b_1} \Big(\mathcal{E}(\mathbf{x}_T(s)) - \mathcal{E}(\mathbf{x}_T(T^*))\Big)\,\mathrm{d}s \\ &\quad + \int_{b_1}^{T^*} \Big(\mathcal{E}(\mathbf{x}_T(s)) - \mathcal{E}(\mathbf{x}_T(T^*))\Big)\,\mathrm{d}s \\ &\leqslant \int_0^T \Big(\mathcal{E}(\mathbf{x}_T(t)) - \mathcal{E}(\mathbf{x}_T(T^*))\Big)\,\mathrm{d}t - \theta^2(b_1-a_1). \end{aligned}$$

By combining the above inequality with (7.2.14), it follows that

$$J_T\left(\overline{u}\right) \leqslant J_T\left(u_T\right) - \theta^2(b_1 - a_1).$$

The statement of the Lemma thus holds for $\mathfrak{I} = 1$.

Step 2). **Heredity.** Let us suppose that, for some $n \geqslant 1$, the statement of the lemma holds whenever $\mathfrak{I} = n$, and let u_T satisfy (7.2.3) and (7.2.4) with $\mathfrak{I} = n+1$. Assume without loss of generality that $a_1 > a_i$ for all $i \in \{2, \dots, \mathfrak{I}\}$. Using precisely the same argument as in Step 1, we can construct a pair of parameters $\overline{u}_1$ satisfying

$$\overline{u}_1(t) = 0 \qquad \text{for a.e. } t \in (T^* - \tau_1, T)$$

with $\tau_1 = \theta(b_1 - a_1)$, and

$$J_T(\overline{u}_1) \leqslant J_T(u_T) - \theta^2(b_1 - a_1),$$

and which is such that $\overline{u}_1(t) = u_T(t)$ for all $t \in (0, t_1)$. Now observe that, since $a_1 > a_i$ for all $i \geqslant 2$, and in view of (7.2.13), it follows that $\overline{u}_1$ satisfies (7.2.3) and (7.2.4) with $\mathfrak{I} - 1 = n$ number of intervals and with $T_1^* = T^* - \tau_1$ instead of T^*. By the induction hypothesis, we conclude that there exists an admissible pair of parameters $\overline{u} \in \mathfrak{U}_{\mathrm{ad},T}$ such that

$$\overline{u}(t) = 0 \qquad \text{for a.e. } t \in (T_1^* - \tau, T)$$

with $\tau = \theta \sum_{i=2}^{\mathfrak{I}} (b_i - a_i)$, and

$$\begin{aligned} J_T(\overline{u}) &\leqslant J_T(\overline{u}_1) - \theta^2 \sum_{i=2}^{\mathfrak{I}} (b_i - a_i) \\ &\leqslant J_T(u_T) - \theta^2 \sum_{i=1}^{\mathfrak{I}} (b_i - a_i). \end{aligned}$$

This concludes the proof. □

7.3 Proof of Theorem 7.1

We are now in a position to complete the proof to Theorem 7.1.

Proof of Theorem 7.1. Properties (7.1.11) and (7.1.12) for the minimizers of J_T follow directly from Proposition 7.2.3. Let us give the proof of the statements *(1)* and *(2)* in Theorem 7.1.

Proof of *(1)*. If the interpolation property holds, then there exist $T_0 > 0$ and a pair of parameters $u_{T_0} = [w_{T_0}, b_{T_0}] \in L^\infty(0, T_0; \mathbb{R}^{d_u})$ such that the associated trajectory $\mathbf{x}_{T_0} \in C^0([0, T_0]; \mathbb{R}^{d_x})$ of (7.1.4) satisfies $\mathcal{E}(\mathbf{x}_{T_0}(T_0)) = 0$. Set

$$T_1 := \frac{T_0 \, \|u_{T_0}\|_{L^\infty(0,T_0;\mathbb{R}^{d_u})}}{M},$$

and consider the pair $u_{T_1} = [w_{T_1}, b_{T_1}]$ defined by

$$u_{T_1}(t) := \frac{M}{\|u_{T_0}\|_{L^\infty(0,T_0;\mathbb{R}^{d_u})}} u_{T_0}\left(t \frac{T_0}{T_1}\right) \qquad \text{for } t \in (0, T_1).$$

Observe that $u_{T_1} \in \mathfrak{U}_{\mathrm{ad},T_1}$. Furthermore, in view of Lemma 7.2.1, the associated solution $\mathbf{x}_{T_1}$ to (7.1.4), is given by $\mathbf{x}_{T_1}(t) = \mathbf{x}_{T_0}\left(t \frac{T_0}{T_1}\right)$, and hence,

$$\mathcal{E}(\mathbf{x}_{T_1}(T_1)) = 0.$$

Now for any $T > 0$, we define $\overline{u} \in \mathfrak{U}_{\mathrm{ad},T}$ by

$$\overline{u}(t) = \begin{cases} u_{T_1}(t) & \text{for } t \in (0, T) \cap (0, T_1) \\ 0 & \text{for } t \in (0, T) \setminus (0, T_1). \end{cases}$$

Then, it follows that

$$\begin{aligned} J_T(\overline{u}) &\leqslant \int_0^{T_1} \mathcal{E}(\mathbf{x}_{T_1}(t))\,\mathrm{d}t + M\,T_1 \\ &= \frac{\|u_{T_0}\|_{L^\infty(0,T_0;\mathbb{R}^{d_u})}}{M} \int_0^{T_0} \mathcal{E}(\mathbf{x}_{T_0}(t))\,\mathrm{d}t + \|u_{T_0}\|_{L^\infty(0,T_0;\mathbb{R}^{d_u})}\,T_0 \\ &= \frac{C_1(u_{T_0})}{M} + C_2(u_{T_0}), \end{aligned} \tag{7.3.1}$$

where $C_1(u_{T_0}), C_2(u_{T_0}) > 0$ are independent of T and M. In view of (7.1.11), any minimizer u_T of J_T satisfies $u_T \in \mathfrak{U}_{\mathrm{sp},T^*}$ for some $T^* \in (0,T]$. Hence, using (7.3.1), we obtain

$$\begin{aligned} J_T(u_T) = M\,T^* + \int_0^T \mathcal{E}(\mathbf{x}_T(t))\,\mathrm{d}t &\leqslant J_T(\overline{u}) \\ &\leqslant \frac{C_1(u_{T_0})}{M} + C_2(u_{T_0}). \end{aligned}$$

From the above estimates, we deduce that

$$T^* \leqslant \frac{C_1(u_{T_0})}{M^2} + \frac{C_2(u_{T_0})}{M} := T(M).$$

Moreover, using (7.1.12), we also deduce that

$$\begin{aligned} T\mathcal{E}(\mathbf{x}_T(T^*)) \leqslant J_T(u_T) &\leqslant J_T(\overline{u}) \\ &\leqslant \frac{C_1(u_{T_0})}{M} + C_2(u_{T_0}) =: \mathfrak{C}(M), \end{aligned}$$

which implies *(1)*, as desired.

Proof of *(2)*. If the asymptotic interpolation property holds, then there exist $T_0 > 0$, a function h as in Definition 7.1.2, and a pair of parameters $u^\dagger = \left[w^\dagger, b^\dagger\right] \in L^\infty(\mathbb{R}_+;\mathbb{R}^{d_u})$ such that the corresponding solution $\mathbf{x}^\dagger$ to (7.1.4) satisfies

$$\mathcal{E}(\mathbf{x}^\dagger(t)) \leqslant h(t) \qquad \text{for all } t > T_0. \tag{7.3.2}$$

Combining this knowledge with the continuity of the map $t \longmapsto \mathcal{E}(\mathbf{x}^\dagger(t))$, we can readily deduce that there exists a constant $C_0 > 0$ depending only on $T_0 > 0$ such that

$$\mathcal{E}(\mathbf{x}^\dagger(t)) \leqslant C_0 \qquad \text{for } t > 0.$$

On another hand, we know by (7.1.11) that there exists $T^* > 0$ such that $u_T \in \mathfrak{U}_{\mathrm{sp},T^*}$. Whilst we cannot give an upper bound for T^* which is uniform in T, we will prove that there exists a constant $\mathfrak{C}(M) > 0$, independent of T, such that

$$T^* \leqslant \frac{\mathfrak{C}(M)}{M} h^{-1}\left(\frac{1}{T}\right) + \frac{1}{M} \quad \text{and} \quad \mathcal{E}(\mathbf{x}_T(T^*)) \leqslant \frac{\mathfrak{C}(M)}{T} h^{-1}\left(\frac{1}{T}\right) + \frac{1}{T} \tag{7.3.3}$$

hold for any $T > \frac{1}{h(T_0)}$. Observe that, by virtue of Definition 7.1.2, the function h is a bijection from $(0,\infty)$ to $(0,h(T_0))$, and so $h^{-1}\left(\frac{1}{T}\right)$ is clearly well defined for all $T > \frac{1}{h(T_0)}$.

Let us henceforth denote

$$\mathfrak{m} := \frac{M}{\left\|u^\dagger\right\|_{L^\infty(\mathbb{R}_+;\mathbb{R}^{d_u})}},$$

and we define the auxiliary parameters $u^\ddagger \in L^\infty(\mathbb{R}_+;\mathbb{R}^{d_u})$ by

$$u^\ddagger(t) = \mathfrak{m} u^\dagger(\mathfrak{m} t) \qquad \text{for } t \in \mathbb{R}_+.$$

For any $T_1 > 0$, we also define

$$u_{T_1}(t) = \begin{cases} u^\ddagger(t) & \text{for } t \in (0, T_1] \\ 0 & \text{for } t > T_1. \end{cases}$$

Observe that $u_{T_1} \in \mathfrak{U}_{\mathrm{ad},T}$ for any $T > 0$. By virtue of Lemma 7.2.1, the state trajectory associated to u_{T_1} is precisely

$$\mathbf{x}_{T_1}(t) = \begin{cases} \mathbf{x}^\dagger(\mathfrak{m}t) & \text{for } \ t \in (0, T_1) \\ \mathbf{x}^\dagger(\mathfrak{m}T_1) & \text{for } t \geqslant T_1. \end{cases}$$

Now, in view of the definition of u_{T_1}, for any $T > 0$, we have

$$\begin{aligned} J_T(u_{T_1}) &\leqslant \int_0^{T_1} \mathcal{E}\left(\mathbf{x}^\dagger(t\mathfrak{m})\right) \mathrm{d}t + (T - T_1)_+ \mathcal{E}\left(\mathbf{x}^\dagger(T_1\mathfrak{m})\right) + M\,T_1 \\ &\leqslant (C_0 + M)\,T_1 + T\,\mathcal{E}\left(\mathbf{x}^\dagger(T_1\mathfrak{m})\right), \end{aligned} \tag{7.3.4}$$

were $(T - T_1)_+ := \max\{0, T - T_1\}$. We use this term in order to cover the case when $T < T_1$.

Let us now prove (7.3.3) using (7.3.4). For any $T > \frac{1}{h(T_0)}$, we set

$$T_1 := \frac{1}{\mathfrak{m}} h^{-1}\left(\frac{1}{T}\right).$$

Combining the optimality of u_T with (7.3.4), and the fact that $u_T \in \mathfrak{U}_{\mathrm{sp},T^*}$ for some $T^* \in (0, T]$, we obtain

$$\begin{aligned} J_T(u_T) = M\,T^* + \int_0^T \mathcal{E}(\mathbf{x}_T(t))\,\mathrm{d}t &\leqslant J_T(u_{T_1}) \\ &\leqslant \mathfrak{C}(M)\,h^{-1}\left(\frac{1}{T}\right) + T\,\mathcal{E}\left(\mathbf{x}^\dagger\left(h^{-1}\left(\frac{1}{T}\right)\right)\right), \end{aligned}$$

for any $T > \frac{1}{h(T_0)}$, where the constant

$$\mathfrak{C}(M) := \frac{(C_0 + M)}{\mathfrak{m}}$$

is clearly independent of T. The last inequality, combined with (7.1.12), (7.3.2), and the fact that h^{-1} is non-increasing, allows us to deduce that

$$M\,T^* + T\,\mathcal{E}(\mathbf{x}_T(T^*)) \leqslant \mathfrak{C}(M)\,h^{-1}\left(\frac{1}{T}\right) + 1 \qquad \text{for all } T > \frac{1}{h(T_0)}.$$

The estimate just above implies (7.3.3), and the desired statement *(2)* then follows. This concludes the proof. □

7.4 An example of asymptotic interpolation

In this section, we present an example of a typical case arising in practical applications where the asymptotic interpolation property in Definition 7.1.2 is expected to hold. In such a scenario, one may use the statement *(2)* of Theorem 7.1 to give quantitative upper bounds for the stopping time T^*, as well as for the training error $\mathcal{E}(\mathbf{x}_T(T^*))$. We shall consider a classification problem for which we are given a training dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$. Here, each input $\vec{x}_i$ is a vector in $\mathcal{X} \subset \mathbb{R}^d$, and the labels $\vec{y}_i$ are elements in $\mathcal{Y} := [m]$, and we henceforth make use of the notation $[m] := \{1, \dots, m\}$.

We focus on neural ODEs of the form (7.1.5), i.e. for any $i \in [N]$ we consider

$$\begin{cases} \dot{\mathbf{x}}_i(t) = \sigma(w(t)\mathbf{x}_i(t) + b(t)) & \text{for } \ t \in (0,T) \\ \mathbf{x}_i(0) = \mathfrak{Q}\,\vec{x}_i, \end{cases} \tag{7.4.1}$$

where $T > 0$ is arbitrarily chosen, the state $\mathbf{x}(t)$ evolves in $\mathbb{R}^{d+m}$ and $\sigma(x) = \max\{x, 0\}$. Moreover, $\mathfrak{Q}$ is the simple linear transformation

$$\mathfrak{Q}\,x = \begin{bmatrix} \mathrm{Id}_d \\ 0_{m,d} \end{bmatrix} x \qquad \text{for } x \in \mathbb{R}^d,$$

which canonically embeds $\mathbb{R}^d$ into $\mathbb{R}^{d+m}$ (as used in [86], for instance). Here and henceforth, $\mathrm{Id}_d \in \mathbb{R}^{d\times d}$ denotes the identity matrix, and $0_{m,d} \in \mathbb{R}^{m\times d}$ is a zero matrix. Finally, we define the linear map $P : \mathbb{R}^{d+m} \longrightarrow \mathbb{R}^m$ as

$$P\mathbf{x} = [0_{m,d}\ \mathrm{Id}_m]\ \mathbf{x} \qquad \text{for } \mathbf{x} \in \mathbb{R}^{d+m}, \tag{7.4.2}$$

and we focus on the cross-entropy loss for multi-class classification, defined in (7.1.3).

Note that the cross-entropy loss defined in (7.1.3) is strictly positive for any input in $\mathbb{R}^m \times \mathcal{Y}$ and thus, we cannot expect to find parameters $[w, b]$ for which the associated training error of the neural ODE trajectory $\mathcal{E}(\mathbf{x}(T))$ equals 0. However, we could render the latter small if we could find parameters $[w, b]$ such that, for any $i \in [N]$, we have

$$P\mathbf{x}_i(T)_{\vec{y}_i} \gg \max_{\substack{j\in[m] \\ j\neq\vec{y}_i}} P\mathbf{x}_i(T_0)_j.$$

This motivates the following discussion. We shall prove that the flow of the neural ODE (7.4.1) just above asymptotically interpolates the dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ in the sense of Definition 7.1.2, whenever there exist parameters $[w, b]$ which *accurately classify* the points in the data set in finite time, by which we mean the following.

Definition 7.4.1. *Let $T_0 > 0$. We say that the parameters $[w_0, b_0] \in L^\infty(0, T_0; \mathbb{R}^{d_u})$* accurately classify *the dataset $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ if, for any $i \in [N]$, the solution $\mathbf{x}_i$ to (7.4.1) on $[0, T_0]$ corresponding to $[w_0, b_0]$, satisfies*

$$P\mathbf{x}_i(T_0)_{\vec{y}_i} > \max_{\substack{j\in[m] \\ j\neq\vec{y}_i}} P\mathbf{x}_i(T_0)_j.$$

The following result then holds for the cross-entropy loss and the ReLU activated neural ODE flows in (7.4.1).

Proposition 7.4.2. *Assume that there exists a time $T_0 > 0$ and parameters $[w_0, b_0] \in L^\infty(0, T_0; \mathbb{R}^{d_u})$ which correctly classify the dataset $\{\mathfrak{Q}\,\vec{x}_i, \vec{y}_i\}_{i=1}^N$ by means of (7.4.1) in time T_0 in the sense of Definition 7.4.1. Then, the asymptotic interpolation property as per Definition 7.1.2 holds with*

$$h(t) = \log\left(1 + (m-1)e^{-\gamma e^{t-T_0}}\right),$$

where $\gamma > 0$ is the margin *defined by*

$$\gamma := \min_{i\in[N]} \left\{ P\mathbf{x}_i(T_0)_{\vec{y}_i} - \max_{\substack{j\in[m] \\ j\neq\vec{y}_i}} P\mathbf{x}_i(T_0)_j \right\}.$$

Before proceeding with the proof, we may apply Theorem 7.1 to obtain the desired quantitative estimates for the training error and the stopping time T^*. Assuming that the dataset may be classified accurately in time $T_0 > 0$ by the ReLU activated neural

ODE (7.4.1) with parameters of margin γ, we obtain the following estimate for the stopping time T^*:

$$T^* \leqslant \frac{\mathfrak{C}(M)}{M}\left(T_0 + \log\left(\frac{1}{\gamma}\log\left(\frac{m-1}{e^{\frac{1}{T}}-1}\right)\right)\right) + \frac{1}{M},$$

whenever $T > \log\left(1+(m-1)e^{-\gamma}\right)$. For the training error on the other hand, we have the estimate:

$$\mathcal{E}\left(\mathbf{x}_T(T^*)\right) \leqslant \frac{\mathfrak{C}(M)}{T}\left(T_0 + \log\left(\frac{1}{\gamma}\log\left(\frac{m-1}{e^{\frac{1}{T}}-1}\right)\right)\right) + \frac{1}{T}.$$

We conclude this section with the proof of Proposition 7.4.2.

Proof of Proposition 7.4.2. Let us consider the pair of parameters

$$\left[w^\dagger(t), b^\dagger(t)\right] := \begin{cases} [w_0(t), b_0(t)] & \text{for } t \in [0, T_0] \\ \left[\begin{bmatrix} 0_d & 0_{d,m} \\ 0_{m,d} & \mathrm{Id}_m \end{bmatrix}, 0_{d+m}\right] & \text{for } t > T_0 \end{cases}$$

defined on $\mathbb{R}_+$. Let us note that, for any $i \in [N]$, the solution $\mathbf{x}_i^\dagger$ of (7.4.1) on $\mathbb{R}_+$ associated to this pair then also solves

$$\begin{cases} \dot{\mathbf{x}}_i^\dagger(t) = \sigma\left(\begin{bmatrix} 0_{d,d} & 0_{d,m} \\ 0_{m,d} & \mathrm{Id}_m \end{bmatrix} \mathbf{x}_i^\dagger(t)\right) & \text{for } t \in (T_0, \infty) \\ \mathbf{x}_i^\dagger(T_0) = \mathbf{x}_i(T_0), \end{cases}$$

where $\mathbf{x}_i(t)$ is the solution to (7.4.1) on $[0, T_0]$ associated to $[w_0, b_0]$ for $i \in [N]$. Since the weight in the system just above is a diagonal matrix, all components of the state vector $\mathbf{x}_i^\dagger(t)$ are mutually independent for all $t > T_0$. We now define $\mathbf{z}_i(t) := P\mathbf{x}_i^\dagger(t)$ for $t \geqslant 0$ and $i \in [N]$. By virtue of the definition of P in (7.4.2), and the fact that $\sigma(x) = \max\{x, 0\} \geqslant 0$, we see that $\mathbf{z}_i(t) \geqslant 0$ for all $t \geqslant 0$ and $i \in [N]$. In view of all this, it may be seen that $\mathbf{z}_i(t)$ also satisfies

$$\begin{cases} \dot{\mathbf{z}}_i(t) = \mathbf{z}_i(t) & \text{for } t \in (T_0, \infty) \\ \mathbf{z}_i(T_0) = \mathbf{x}_i(T_0), \end{cases}$$

for $i \in [N]$, and thus

$$P\mathbf{x}_i^\dagger(t) = \mathbf{z}_i(t) = \mathbf{z}_i(T_0)e^{t-T_0} \qquad \text{for } t > T_0.$$

We now evaluate the cross-entropy loss along $\left(P\mathbf{x}_i^\dagger(t), \vec{y}_i\right)$ for $t > T_0$. By the definition of the margin γ, we have

$$\mathbf{z}_i(T_0)_j \leqslant \mathbf{z}_i(T_0)_{\vec{y}_i} - \gamma \qquad \text{for all } i \in [N] \quad \text{and} \quad j \in [m],\ j \neq \vec{y}_i,$$

which allows us to compute

$$\begin{aligned} \operatorname{loss}\left(P\mathbf{x}_i^\dagger(t), \vec{y}_i\right) &= -\log\left(\frac{e^{\mathbf{z}_i(t)_{\vec{y}_i}}}{\sum_{j=1}^m e^{\mathbf{z}_i(t)_j}}\right) \\ &= -\log\left(\frac{e^{(\mathbf{z}_i(T_0)e^{t-T_0})_{\vec{y}_i}}}{\sum_{j=1}^m e^{(\mathbf{z}_i(T_0)e^{t-T_0})_j}}\right) \\ &= -\log\left(\frac{e^{\mathbf{z}_j(T_0)_{\vec{y}_i}e^{t-T_0}}}{e^{\mathbf{z}_i(T_0)_{\vec{y}_i}e^{t-T_0}} + \sum_{j\neq\vec{y}_i} e^{\mathbf{z}_i(T_0)_j e^{t-T_0}}}\right) \\ &\leqslant -\log\left(\frac{e^{\mathbf{z}_i(T_0)_{\vec{y}_i}e^{t-T_0}}}{e^{\mathbf{z}_i(T_0)_{\vec{y}_i}e^{t-T_0}} + (m-1)e^{(\mathbf{z}_i(T_0)_{\vec{y}_i}-\gamma)e^{t-T_0}}}\right) \\ &= \log\left(1 + (m-1)e^{-\gamma e^{t-T_0}}\right). \end{aligned}$$

As the above inequality holds for any $i \in [N]$, we conclude that, for all $t > T_0$, we have

$$\mathcal{E}\left(\mathbf{x}_j^\dagger(t)\right) \leqslant \frac{1}{N}\sum_{i=1}^{N} \text{loss}\left(P\mathbf{x}_j^\dagger(t), y_j\right) \leqslant \log\left(1 + (m-1)e^{-\gamma e^{t-T_0}}\right).$$

Defining $h(t) := \log\left(1 + (m-1)e^{-\gamma e^{t-T_0}}\right)$, it is readily seen that h satisfies the properties required in Definition 7.1.2. This concludes the proof. □

7.5 Concluding remarks

In this paper, we have presented a manifestation of sparsity and approximation properties for L^1–regularized supervised learning problems for neural ODEs. Our main result ensures that any global minimizer u_T is sparse, in the sense that $u_T \equiv 0$ on (T^*, T) for some $T^* \in (0, T]$. Moreover, under appropriate interpolation assumptions, we may provide estimates on the stopping time T^* and the training error $\mathcal{E}(\mathbf{x}_T(T^*))$. When extrapolated to the discrete-time, ResNet context, a shorter time-horizon in the optimal control problem can be interpreted as considering a shallower ResNet, which naturally lowers the computational cost of the training process.

Conclusion

In this thesis, we presented various contributions on questions related to free boundary problems and the foundations of deep learning, through the lens of control theory.

Our first contribution regards the controllability properties of several free boundary problems. By combining a plethora of classical and modern methods, we have concluded that one may expect, in many cases, for the controllability properties to transfer from the PDE (e.g. heat, Burgers) to its free boundary analog in one-dimension. This is mainly due to the fact that the free boundary is space independent. Whilst not proven, we observe, in the context of the two-dimensional Stefan problem, that the euclidean dimension of the moving domain may play a role in the controllability properties (or lack thereof) of the linearized system. In fact, the boundary controllability of the classical Stefan problem (which is the zero-surface tension limit of the Gibbs-Thomson system in the uncontrolled setting) may not be derived from the controllability of the Gibbs-Thomson system in two space dimensions. This could in part be because of the fact that the free boundary is space dependent, and the linearized classical Stefan problem is uncoupled and of cascade form, with the free boundary condition manifesting itself as an infinite-dimensional constraint on the control for the PDE component.

Our second contribution regards the turnpike property in finite and infinite dimensional nonlinear optimal control. When the running target is chosen as a stationary solution of the free nonlinear dynamics, we prove the exponential turnpike property without any smallness conditions on the data or the target. Due to the specificity of our proof, we bypass the usage of the optimality system or linearization techniques, which in turn allows us to address finite-dimensional, control-affine systems with globally Lipschitz nonlinearities, commonly encountered in the context of deep learning, and semilinear PDEs with globally Lipschitz nonlinearities.

Our third contribution regards the role of the time horizon in deep supervised learning problems with L^2 (or Sobolev) parameter regularization. We obtain several quantitative approximation properties for the trained/optimal parameters and the associated neural ODE flow with respect to the final time horizon, with the specific rate depending on the loss function at hand. Moreover, due to an underlying homogeneity assumption of the dynamics, we deduce that the final horizon scales inversely to the regularization hyperparameter in the cost functional. This allows us to obtain an equivalence between the convergence of the final time horizon to infinity and the convergence of the regularization hyperparameter to zero. The latter, combined with the convergence of the optimal parameters to minimal norm parameters which interpolate the dataset in the regression setting, allows us to stipulate generalization properties – namely, optimizing with a large time horizon, which may be interpreted as a larger depth for ResNets, has the practically desirable effect of making the training error close to zero, but by means of almost optimal parameters. As elaborated in the open problems section, we expect a similar convergence property of the parameters in the classification setting.

To enhance the decay rates of the training error, we proposed an augmented learning problem by adding an artificial regularization term of the state trajectory over the entire time horizon. We obtained an exponential rate of decay for the training error and for the optimal parameters in any time– an improved estimate for the depth required to reach

optimal training accuracy. Moreover, since the trained parameters are exponentially small, this would entail that the flow would tend to oscillate little, and stipulate possible generalization properties. In particular, this result indicates that there is no need to consider too large final time horizons in such supervised learning problems.

All of our approximation results ought to be compared with universal approximation results, in which, a key caveat is that there is no scalable method to compute the theoretically guaranteed parameters.

Our final contribution regards the appearance of sparsity patterns for supervised learning problems for neural ODEs. In the context of the augmented learning problem with L^1–parameter regularization, under homogeneity assumptions on the dynamics (typical for ReLU activations), we showed that the trained parameters are sparse in the sense there exists a positive stopping time beyond which the optimal parameters vanish. In practical terms, when extrapolated to the ResNet context, a shorter time-horizon in the optimal control problem can be interpreted as considering a shallower ResNet, which lowers the computational cost of training. We may also provide quantitative estimates on the stopping time, and on the training error of the neural ODE trajectories at the stopping time. The latter stipulates a quantitative approximation property of neural ODE flows with sparse parameters, in line with what we had deduced for the L^2–regularized problem.

Conclusión

En esta tesis, presentamos diversas contribuciones sobre cuestiones relacionadas con problemas de frontera libre y la teoría fundamental del aprendizaje profundo desde el punto de vista de la teoría de control.

Nuestra primera contribución concierne las propiedades de controlabilidad de varios problemas de frontera libre. Combinando los métodos clásicos con otros más modernos, se puede esperar, en muchos casos, que las propiedades de controlabilidad se transfieran desde la EDP (por ejemplo, la ecuación del calor o de Burgers) a su análogo de frontera libre en una dimensión. Esto se debe principalmente al hecho de que la frontera libre es independiente de la variable espacial. Aunque no está probado, observamos, en el contexto del problema de Stefan en en dos dimensiones espaciales, que la dimensión euclidiana del dominio puede ser relevante en las propiedades de controlabilidad (o falta de ellas) del sistema linealizado. De hecho, la controlabilidad desde el borde del problema clásico de Stefan (que es el límite del sistema Gibbs-Thomson sin control cuando la tensión superficial tiende a cero) no puede se derivarse de la controlabilidad del sistema Gibbs-Thomson en dos dimensiones espaciales. Esto podría deberse, en parte, al hecho de que el frontera libre depende del espacio, y el problema clásico de Stefan linealizado está desacoplado teniendo forma de cascada, con el frontera libre manifestándose como una restricción de dimensión infinita en el control de la componente EDP.

Nuestra segunda contribución concierne la propiedad de Turnpike en problemas de control óptimo no lineal en dimensión finita e infinita. Cuando el estado objetivo es una solución estacionaria de la dinámica libre, probamos la propiedad de Turnpike exponencial sin imponer hipótesis de pequeñez sobre los datos iniciales o el estado objetivo. Dada la estrategia que proponemos para la demostración, conseguimos evitar el uso del sistema de optimalidad o técnicas de linealización, y de este modo, nuestras técnicas nos permiten abordar sistemas de control afín en dimensión finita con no linealidades globalmente Lipschitz, que se encuentran comúnmente en el contexto del aprendizaje profundo, y EDP semilineales con no linealidades globalmente Lipschitz.

En nuestra tercera contribución, estudiamos el papel del horizonte temporal en problemas de aprendizaje supervisado profundo con regularización L^2 (o Sobolev) de los parámetros. Obtenemos varias estimaciones cuantitativas de aproximación para los parámetros entrenados / óptimos y para el flujo de la EDO neuronal asociada con respecto al horizonte temporal, con una tasa de decaimiento específica dependiendo de la función de coste en cuestión. Además, bajo hipótesis de homogeneidad de la dinámica subyacente, deducimos que el horizonte temporal escala inversamente al hiperparámetro de regularización en el funcional de coste. Esto nos permite obtener una equivalencia entre el límite cuando el horizonte temporal va al infinito y el límite cuando el hiperparámetro de regularización va a cero. Por otro lado, se podrían estipular propiedades de generalización, es decir, optimizar con un horizonte temporal grande, que podría interpretarse como una ResNet de profundidad mayor, tiene el efecto deseable, a nivel práctico, de hacer que el error de entrenamiento sea cercano a cero sin sobreajustar. Como se explicó en la sección de problemas abiertos, esperamos que los parámetros tengan una propiedad de convergencia similar en los problemas de clasificación.

Para mejorar las tasas de decaimiento del error de entrenamiento, proponemos tam-

bién un problema de aprendizaje aumentado, agregando un término artificial de regularización de la trayectoria del estado en todo el horizonte temporal. En este escenario, obtenemos una tasa de decaimiento exponencial para el error de entrenamiento y para los parámetros óptimos en todo el intervalo de tiempo. Este resultado permite deducir una estimación mejorada de la profundidad óptima requerida para alcanzar una precisión de entrenamiento prefijada. Además, dado que los parámetros entrenados son exponencialmente pequeños, el flujo tiende a oscilar poco, y por tanto se podrían estipular posibles propiedades de generalización. En particular, este resultado indica que no hay necesidad de considerar horizontes temporales demasiado grandes en tales problemas de aprendizaje supervisado.

Todos nuestros resultados de aproximación deben compararse con los resultados de aproximación universal existentes, en los cuales, es importante señalar que no existe un método escalable para obtener los parámetros garantizados teóricamente.

En nuestra última contribución, estudiamos la aparición de patrones de tipo sparse para problemas de aprendizaje supervisado mediante EDOs neuronales. En el contexto del problema de aprendizaje aumentado con regularización L^1 de los parámetros, bajo supuestos de homogeneidad en la dinámica (típico de las activaciones de tipo ReLU), mostramos que los parámetros entrenados son sparse en el sentido de que existe un tiempo de parada positivo, más allá del cual, los parámetros óptimos son nulos. En términos prácticos, cuando se extrapola al contexto ResNet, un horizonte temporal más corto en el problema de control óptimo se puede interpretar como una ResNet menos profunda, lo que reduce el coste computacional del entrenamiento. También proporcionamos estimaciones cuantitativas para el tiempo de parada y para el error de entrenamiento de las trayectorias de las EDOs neuronales. Este resultado estipula una propiedad de aproximación cuantitativa de los flujos asociados a EDOs neuronales con parámetros sparse, en línea con lo que habíamos deducido para el mismo problema de aprendizaje con regularización L^2.

Bibliography

[1] Alabau-Boussoira, F., Cannarsa, P., and Fragnelli, G. Carleman estimates for degenerate parabolic operators with applications to null controllability. *J. Evol. Equ. 6* (2006), 161–204.

[2] Alazard, T. Stabilization of the water-wave equations with surface tension. *Ann. PDE 3*, 2 (2017), 17.

[3] Alazard, T. Boundary observability of gravity water waves. *Ann. Inst. H. Poincaré Anal. Non Linéaire 35*, 3 (2018), 751–779.

[4] Alazard, T. Stabilization of gravity water waves. *J. Math. Pur. Appl. 114* (2018), 51–84.

[5] Alazard, T., Baldi, P., and Han-Kwan, D. Control of water waves. *J. Eur. Math. Soc. 20*, 3 (2018), 657–745.

[6] Albertini, F., and Sontag, E. D. For neural networks, function determines form. *Neural Networks 6*, 7 (1993), 975–990.

[7] Albertini, F., Sontag, E. D., and Maillot, V. Uniqueness of weights for neural networks. *Artificial Neural Networks for Speech and Vision* (1993), 115–125.

[8] Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in Neural Information Processing Systems* (2019), pp. 6158–6169.

[9] Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning* (2019), PMLR, pp. 242–252.

[10] Almgren, F., and Wang, L. Mathematical existence of crystal growth with Gibbs-Thomson curvature effects. *J. Geom. Anal. 10*, 1 (2000), 1.

[11] Alt, W., and Schneider, C. Linear-quadratic control problems with l1-control cost. *Optimal Control Appl. Methods 36*, 4 (2015), 512–534.

[12] Amann, H. *Linear and quasilinear parabolic problems*, vol. 1. Springer, 1995.

[13] Ammar-Khodja, F., Micu, S., and Münch, A. Controllability of a string submitted to unilateral constraint. In *Ann. Inst. H. Poincaré Anal. Non Linéaire* (2010), vol. 27, pp. 1097–1119.

[14] Angenent, S. Large time asymptotics for the porous medium equation. In *Nonlinear diffusion equations and their equilibrium states*. Math. Sci. Res. Inst. Publ., Springer, New York, 1988, pp. 21–34.

[15] Athanasopoulos, I., Caffarelli, L., and Milakis, E. Parabolic obstacle problems. quasi-convexity and regularity. *arXiv preprint arXiv:1601.01516* (2016).

[16] Avelin, B., and Nyström, K. Neural ODEs as the deep limit of ResNets with constant weights. *Anal. Appl.* (2020), 1–41.

[17] Badra, M., and Takahashi, T. Feedback stabilization of a simplified 1d fluid–particle system. *Ann. Inst. H. Poincaré Anal. Non Linéaire 31*, 2 (2014), 369–389.

[18] Bartlett, P. L., and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res. 3*, Nov (2002), 463–482.

[19] Beale, J. The initial value problem for the Navier-Stokes equations with a free surface. *Comm. Pure Appl. Math. 34* (1981), 359–392.

[20] Beale, J. Large-time regularity of viscous surface waves. *Arch. Ration. Mech. Anal. 84* (1984), 307–352.

[21] Beauchard, K., Cannarsa, P., and Guglielmi, R. Null controllability of Grushin-type operators in dimension two. *J. Eur. Math. Soc. 16*, 1 (2014), 67–101.

[22] Beauchard, K., and Marbach, F. Quadratic obstructions to small-time local controllability for scalar-input systems. *J. Differ. Equ. 264*, 5 (2018), 3704–3774.

[23] Beauchard, K., and Marbach, F. Unexpected quadratic behaviors for the small-time local null controllability of scalar-input parabolic equations. *J. Math. Pur. Appl. 136* (2020), 22–91.

[24] Beauchard, K., and Zuazua, E. Some controllability results for the 2d Kolmogorov equation. *Ann. Inst. H. Poincaré Anal. Non Linéaire 26*, 5 (2009), 1793–1815.

[25] Benning, M., Celledoni, E., Ehrhardt, M. J., Owren, B., and Schönlieb, C.-B. Deep learning as optimal control problems: Models and numerical methods. *J. Comput. Dyn. 6*, 2 (2019), 171.

[26] Bensoussan, A., Da Prato, G., Delfour, M. C., and Mitter, S. K. *Representation and control of infinite dimensional systems*, 2 ed. Systems & Control : Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA, 2007.

[27] Bölcskei, H., Grohs, P., Kutyniok, G., and Petersen, P. Optimal approximation with sparsely connected deep neural networks. *SIAM J. Math. Data Sci.*, 1 (2019), 8–45.

[28] Bonn, D., Eggers, J., Indekeu, J., and Meunier, J. Wetting and spreading. *Rev. Mod. Phys. 81*, 739 (2009).

[29] Bonneton, P., Lannes, D., Martins, K., and Michallet, H. A nonlinear weakly dispersive method for recovering the elevation of irrotational surface waves from pressure measurements. *Coastal Engineering 138* (2018), 1–8.

[30] Boulakia, M., and Guerrero, S. Local null controllability of a fluid-solid interaction problem in dimension 3. *J. Eur. Math. Soc 15* (2013), 825–856.

[31] Boulakia, M., and Osses, A. Local null controllability of a two-dimensional fluid-structure interaction problem. *ESAIM Control Optim. Calc. Var. 14*, 1 (2008), 1–42.

[32] Brézis, H. Problèmes unilatéraux. *J. Math. Pur. Appl. 51* (1972), 1–168.

[33] Brezis, H. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

[34] Bruna, J., and Mallat, S. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence 35*, 8 (2013), 1872–1886.

[35] Budd, C. J., Huang, W., and Russell, R. D. Adaptivity with moving grids. *Acta Numer. 18* (2009), 111–241.

[36] Caboussat, A. Numerical simulation of two-phase free surface flows. *Arch. Comput. Methods Eng. 12*, 2 (2005), 165–224.

[37] Caboussat, A., and Rappaz, J. Analysis of a one-dimensional free boundary flow problem. *Numer. Math. 101* (2005), 67–86.

[38] Caffarelli, L. The regularity of free boundaries in higher dimensions. *Acta Math. 139* (1977), 155–184.

[39] Caffarelli, L. Some aspects of the one-phase Stefan problem. *Indiana Univ. Math. J. 27* (1978), 73–77.

[40] Caffarelli, L., and Evans, L. Continuity of the temperature in the two-phase Stefan problem. *Arch. Ration. Mech. Anal. 81* (1983), 199–220.

[41] Caffarelli, L., Petrosyan, A., and Shahgholian, H. Regularity of a free boundary in parabolic potential theory. *J. Amer. Math. Soc. 17*, 4 (2004), 827–869.

[42] Caffarelli, L. A. The obstacle problem revisited. *J. Fourier Anal. Appl. 4*, 4 (1998), 383–402.

[43] Candes, E. J., Romberg, J. K., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math. 59*, 8 (2006), 1207–1223.

[44] Candes, E. J., and Tao, T. Decoding by linear programming. *IEEE Trans. Inform. Theory 51*, 12 (2005), 4203–4215.

[45] Cannarsa, P., Floridia, G., and Khapalov, A. Y. Multiplicative controllability for semilinear reaction–diffusion equations with finitely many changes of sign. *J. Math. Pur. Appl. 108*, 4 (2017), 425–458.

[46] Cannarsa, P., Fragnelli, G., and Rocchetti, D. Null controllability of degenerate parabolic operators with drift. *Netw. Heterog. Media 2*, 4 (2007), 695–715.

[47] Cannarsa, P., Fragnelli, G., and Rocchetti, D. Controllability results for a class of one-dimensional degenerate parabolic problems in nondivergence form. *J. Evol. Equ. 8*, 4 (2008), 583–616.

[48] Cannarsa, P., Martinez, P., and Vancostenoble, J. Persistent regional null controllability for a class of degenerate parabolic equations. *Comm. Pure Appl. Anal. 3*, 4 (2004), 607–635.

[49] Cannarsa, P., Martinez, P., and Vancostenoble, J. Null controllability of degenerate heat equations. *Adv. Differ. Equ. 10*, 2 (2005), 153–190.

[50] Cannarsa, P., Martinez, P., and Vancostenoble, J. Carleman estimates for a class of degenerate parabolic operators. *SIAM J. Control. Optim. 47*, 1 (2008), 1–19.

[51] Cannarsa, P., Martinez, P., and Vancostenoble, J. *Global Carleman estimates for degenerate parabolic operators with applications*, vol. 239 of *Mem. Amer. Math. Soc.* 2016.

[52] Castro, A., Córdoba, D., Fefferman, C., Gancedo, F., and Gómez-Serrano, J. Splash singularities for the free boundary Navier-Stokes equations. *Ann. PDE 5*, 1 (2019), 12.

[53] Castro, Á., Córdoba, D., Fefferman, C., Gancedo, F., and López-Fernández, M. Rayleigh-Taylor breakdown for the Muskat problem with applications to water waves. *Ann. of Math.* (2012), 909–948.

[54] Castro, A., Córdoba, D., Fefferman, C. L., Gancedo, F., and Gómez-Serrano, J. Splash singularity for water waves. *Proceedings of the National Academy of Sciences 109*, 3 (2012), 733–738.

[55] Cazenave, T. *An introduction to semilinear elliptic equations*, vol. 164. Editora do Instituto de Matemática, Universidade Federal do Rio de Janeiro, 2006.

[56] Chalmers, B. *Principles of solidification.* Krieger, Huntington, N.Y., 1977.

[57] Chaves-Silva, F. W., Rosier, L., and Zuazua, E. Null controllability of a system of viscoelasticity with a moving control. *J. Math. Pur. Appl. 101*, 2 (2014), 198–222.

[58] Chayes, L., del Mar González, M., Gualdani, M. P., and Kim, I. Global existence and uniqueness of solutions to a model of price formation. *SIAM J. Math. Anal. 41*, 5 (2009), 2107–2135.

[59] Chen, G.-Q., and Feldman, M. Global solutions of shock reflection by large-angle wedges for potential flow. *Ann. of Math.* (2010), 1067–1182.

[60] Chen, R. T., Behrmann, J., Duvenaud, D. K., and Jacobsen, J.-H. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems* (2019), pp. 9916–9926.

[61] Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems* (2018), pp. 6571–6583.

[62] Chizat, L., and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems* (2018), pp. 3036–3046.

[63] Chizat, L., and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Proceedings of Thirty Third Conference on Learning Theory, PMLR* (2020), vol. 125, pp. 1305–1338.

[64] Chowdhury, S., Maity, D., Ramaswamy, M., and Raymond, J.-P. Local stabilization of the compressible Navier–Stokes system, around null velocity, in one dimension. *J. Differ. Equ. 259*, 1 (2015), 371–407.

[65] Cindea, N., Micu, S., Roventa, I., and Tucsnak, M. Particle supported control of a fluid-particle system. *J. Math. Pures Appl. 104*, 2 (2014), 311–353.

[66] Clapeyron, B., and Lamé, G. Mémoire sur l'équilibre intérieur des corps solides homogènes. *J. Reine Angew. Math. 1831*, 7 (1831), 145–169.

[67] Colombo, M., Spolaor, L., and Velichkov, B. On the asymptotic behavior of the solutions to parabolic variational inequalities. *J. Reine Angew. Math. 1*, ahead-of-print (2020).

[68] Colombo, R. M. Hyperbolic phase transitions in traffic flow. *SIAM J. Appl. Math. 63*, 2 (2003), 708–721.

[69] Coron, J.-M. *Control and nonlinearity.* No. 136. American Mathematical Soc., 2007.

[70] Coron, J.-M., Diaz, J. I., Drici, A., and Mignazzini, T. Global null controllability of the 1-dimensional nonlinear slow diffusion equation. *Chin. Ann. Math. 34* (2013), 333–344.

[71] Coron, J.-M., and Nguyen, H.-M. Null controllability and finite time stabilization for the heat equations with variable coefficients in space in one dimension via backstepping approach. *Arch. Ration. Mech. Anal. 225*, 3 (2017), 993–1023.

[72] Coron, J.-M., and Trélat, E. Global steady-state controllability of one-dimensional semilinear heat equations. *SIAM J. Control Optim. 43*, 2 (2004), 549–569.

[73] Cortes, C., and Vapnik, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.

[74] Crosetto, P., Reymond, P., Deparis, S., Kontaxakis, D., Stergiopulos, N., and Quarteroni, A. Fluid–structure interaction simulation of aortic blood flow. *Comput. & Fluids 43*, 1 (2011), 46–57.

[75] Cuchiero, C., Larsson, M., and Teichmann, J. Deep neural networks, generic universal interpolation, and controlled ODEs. *SIAM J. Math. Data Sci. 2*, 3 (2020), 901–919.

[76] Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* (1989), 303–314.

[77] Demarque, R., and Fernández-Cara, E. Local null controllability of one-phase Stefan problems in 2d star-shaped domains. *J. Evol. Equ. 18*, 1 (2018), 245–261.

[78] Denzler, J., and McCann, R. J. Fast diffusion to self-similarity: complete spectrum, long-time asymptotics, and numerology. *Arch. Ration. Mech. Anal. 175*, 3 (2005), 301–342.

[79] Dodet, G., Mélet, A., Ardhuin, F., Bertin, X., Idier, D., and Almar, R. The contribution of wind-generated waves to coastal sea-level changes. *Surveys in Geophysics 40*, 6 (2019), 1563–1601.

[80] Donoho, D. L. Compressed sensing. *IEEE Trans. Inform. Theory 52*, 4 (2006), 1289–1306.

[81] Donoho, D. L., and Elad, M. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proc. Natl. Acad. Sci. USA 100*, 5 (2003), 2197–2202.

[82] Dorfman, R., Samuelson, P., and Solow, R. *Linear Programming and Economic Analysis*. Dover Books on Advanced Mathematics. Dover Publications, 1958.

[83] Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning* (2019), PMLR, pp. 1675–1685.

[84] Dunbar, W. B., Petit, N., Rouchon, P., and Martin, P. Boundary control of a nonlinear Stefan problem. In *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)* (2003), vol. 2, IEEE, pp. 1309–1314.

[85] Dunbar, W. B., Petit, N., Rouchon, P., and Martin, P. Motion planning for a nonlinear Stefan problem. *ESAIM Control Optim. Calc. Var. 9* (2003), 275–296.

[86] Dupont, E., Doucet, A., and Teh, Y. W. Augmented Neural ODEs. In *Advances in Neural Information Processing Systems* (2019), pp. 3134–3144.

[87] Duprez, M., and Lissy, P. Bilinear local controllability to the trajectories of the Fokker-Planck equation with a localized control. *arXiv preprint arXiv:1909.02831* (2019).

[88] Duyckaerts, T., Zhang, X., and Zuazua, E. On the optimality of the observability inequalities for parabolic and hyperbolic systems with potentials. In *Ann. Inst. H. Poincaré Anal. Non Linéaire* (2008), vol. 25, Elsevier, pp. 1–41.

[89] E, W. A proposal on machine learning via dynamical systems. *Commun. Math. Stat. 5*, 1 (2017), 1–11.

[90] El-Hachem, M., McCue, S. W., Jin, W., Du, Y., and Simpson, M. J. Revisiting the Fisher–Kolmogorov–Petrovsky–Piskunov equation to interpret the spreading–extinction dichotomy. *Proceedings of the Royal Society A 475*, 2229 (2019), 20190378.

[91] Ervedoza, S. Control issues and linear projection constraints on the control and on the controlled trajectory. *arXiv preprint arXiv:1909.01649* (2019).

[92] Ervedoza, S., and Savel, M. Local boundary controllability to trajectories for the 1d compressible Navier Stokes equations. *ESAIM Control Optim. Calc. Var. 24*, 1 (2018), 211–235.

[93] Ervedoza, S., and Zuazua, E. Sharp observability estimates for heat equations. *Arch. Ration. Mech. Anal. 202*, 3 (2011), 975–1017.

[94] Escher, J., Prüss, J., and Simonett, G. Analytic solutions for a Stefan problem with Gibbs-Thomson correction. *J. Reine Angew. Math.* (2003), 1–52.

[95] Esteve, C., Geshkovski, B., Pighin, D., and Zuazua, E. Large-time asymptotics in deep learning. *arXiv preprint arXiv:2008.02491* (2020).

[96] Esteve, C., Geshkovski, B., Pighin, D., and Zuazua, E. Turnpike in Lipschitz-nonlinear optimal control. *arXiv preprint arXiv:2011.11091* (2020).

[97] Fattorini, H., and Russell, D. Exact controllability theorems for linear parabolic equations in one space dimension. *Arch. Ration. Mech. Anal. 43* (1971), 272–292.

[98] Feireisl, E. On the motion of rigid bodies in a viscous compressible fluid. *Arch. Ration. Mech. Anal. 167*, 4 (2003), 281.

[99] Fernández-Cara, E., and de Sousa, I. T. Local null controllability of a free-boundary problem for the semilinear 1d heat equation. *Bull. Braz. Math. Soc. 48* (2017), 303–315.

[100] Fernández-Cara, E., and Doubova, A. Some control results for simplified one-dimensional models of fluid-solid interaction. *Math. Models Methods Appl. Sci 15*, 5 (2005), 783–824.

[101] Fernández-Cara, E., Guerrero, S., Imanuvilov, O., and Puel, J.-P. Local exact controllability of the Navier-Stokes system. *J. Math. Pures Appl. 83*, 12 (2004), 1501–1542.

[102] Fernández Cara, E., Hernández, F., and Limaco Ferrel, J. Local null controllability of a 1d Stefan problem. *Bull. Braz. Math. Soc.* (2018), 1–25.

[103] Fernández-Cara, E., Limaco Ferrel, J., and Dias Bezerra de Menezes, S. On the controllability of a free-boundary problem for the 1d heat equation. *Syst. Cont. Lett. 87* (2016).

[104] Fernández-Cara, E., and Zuazua, E. Null and approximate controllability for weakly blowing up semilinear heat equations. In *Ann. Inst. H. Poincaré Anal. Non Linéaire* (2000), vol. 17, pp. 583–616.

[105] Figalli, A. Regularity of interfaces in phase transitions via obstacle problems. In *Proceedings of the International Congress of Mathematicians 2018 (ICM 2018)* (2019), vol. 1, World Scientific, pp. 225–247.

[106] Formaggia, L., Quarteroni, A., and Veneziani, A. *Cardiovascular Mathematics: Modeling and simulation of the circulatory system*, vol. 1. Springer Science & Business Media, 2010.

[107] Fragnelli, G. Null controllability of degenerate parabolic equations in non divergence form via Carleman estimates. *Discrete Contin. Dyn. Syst. Ser. S 6*, 3 (2013), 687–701.

[108] Friedman, A. The Stefan problem in several space variables. *Trans. Amer. Math. Soc. 133* (1968), 51–87.

[109] Friedman, A. *Variational Principles and Free-Boundary Problems.* Wiley-Interscience, New York, 1982.

[110] Friedman, A., and Kinderlehrer, D. A one phase Stefan problem. *Indiana Univ. Math. J. 24* (1975), 1005–1035.

[111] Friedman, A., and Reitich, F. The Stefan problem with small surface tension. *Trans. Amer. Math. Soc. 328* (1991), 465–515.

[112] Fu, X., Yong, J., and Zhang, X. Exact controllability for multidimensional semilinear hyperbolic equations. *SIAM J. Control Optim. 46*, 5 (2007), 1578–1614.

[113] Fursikov, A., and Imanuvilov, O. Y. Controllability of evolution equations. vol. 34 of *Lecture Notes Series*. Soul National University Research Institute of Mathematics Global Analysis Research Center, Seoul, 1996.

[114] Gérard-Varet, D., and Hillairet, M. Regularity issues in the problem of fluid structure interaction. *Arch. Ration. Mech. Anal. 195*, 2 (2010), 375–407.

[115] Geshkovski, B. Null-controllability of perturbed porous medium gas flow. *ESAIM Control Optim. Calc. Var. 26* (2020), 85.

[116] Geshkovski, B., and Zuazua, E. Controllability of one-dimensional viscous free boundary flows. *Hal preprint* (2019).

[117] Giacomelli, L., Knüpfer, H., and Otto, F. Smooth zero-contact-angle solutions to a thin-film equation around the steady state. *J. Differ. Equ. 6* (2008), 1454–1506.

[118] Gnann, M. V. Well-posedness and self-similar asymptotics for a thin-film equation. *Siam J. Math. Anal. 47* (2015), 2868–2902.

[119] Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning.* 2016.

[120] Grathwohl, W., Chen, R. T., Bettencourt, J., Sutskever, I., and Duvenaud, D. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367* (2018).

[121] Gravner, J., and Quastel, J. Internal DLA and the Stefan problem. *Ann. Probab.* (2000), 1528–1562.

[122] Grüne, L., Schaller, M., and Schiela, A. Exponential sensitivity and turnpike analysis for linear quadratic optimal control of general evolution equations. *J. Differ. Equ. 268*, 12 (2020), 7311–7341.

[123] Guerrero, S., and Imanuvilov, O. Y. Remarks on non controllability of the heat equation with memory. *ESAIM Control Optim. Calc. Var. 19*, 1 (2013), 288–300.

[124] Gueye, M. Exact boundary controllability of 1-D parabolic and hyperbolic degenerate equations. *SIAM J. Control. Optim. 52*, 4 (2014), 2037–2054.

[125] Gugat, M., and Hante, F. M. On the turnpike phenomenon for optimal boundary control problems with hyperbolic systems. *SIAM J. Control Optim. 57*, 1 (2019), 264–289.

[126] Gugat, M., Schuster, M., and Zuazua, E. The finite-time turnpike phenomenon for optimal control problems: Stabilization by non-smooth tracking terms. *arXiv preprint arXiv:2006.07051* (2020).

[127] Gugat, M., Trélat, E., and Zuazua, E. Optimal Neumann control for the 1d wave equation: Finite horizon, infinite horizon, boundary tracking terms and the turnpike property. *Systems Control Lett. 90* (2016), 61–70.

[128] Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems* (2018), pp. 9461–9471.

[129] Haber, E., and Ruthotto, L. Stable architectures for deep neural networks. *Inverse Problems 34*, 1 (2017), 014004.

[130] Hadžić, M., and Shkoller, S. Well-posedness for the classical Stefan problem and the zero surface tension limit. *Arch. Ration. Mech. Anal. 223*, 1 (2017), 213–264.

[131] Hadžić, M. Orthogonality conditions and asymptotic stability in the Stefan problem with surface tension. *Arch. Ration. Mech. Anal. 203*, 3 (2012), 719–745.

[132] Hadzić, M., and Guo, Y. Stability in the Stefan problem with surface tension (i). *Comm. Part. Diff. Eq. 35* (2010), 201–244.

[133] Hadzic, M., and Raphaël, P. On melting and freezing for the 2d radial Stefan problem. *J. Eur. Math. Soc. 21*, 11 (2019), 3259–3341.

[134] Hadžić, M., and Shkoller, S. Global stability and decay for the classical Stefan problem. *Comm. Pure Appl. Math. 68*, 5 (2015), 689–757.

[135] Hansen, S., and Tucsnak, M. Some new applications of Russell's principle to infinite dimensional vibrating systems. *Annual Reviews in Control 44* (2017), 184–198.

[136] Hanzawa, E. Classical solutions of the Stefan problem. *Tôhoku Math. J. 2* (1981), 297–335.

[137] Hardy, G. Note on a theorem of Hilbert. *Math. Z. 6* (1920), 314–317.

[138] Hartmann, A., Kellay, K., and Tucsnak, M. From the reachable space of the heat equation to hilbert spaces of holomorphic functions. *J. Eur. Math. Soc. 22*, 10 (2020), 3417–3440.

[139] Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. The entire regularization path for the support vector machine. *J. Mach. Learn. Res. 5*, Oct (2004), 1391–1415.

[140] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[141] Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences 79*, 8 (1982), 2554–2558.

[142] Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural networks* (1989), 359–366.

[143] Imanuvilov, O., and Takahashi, T. Exact controllability of a fluid-rigid body system. *J. Math. Pur. Appl. 87* (2007), 408–437.

[144] Ingham, A. E. Some trigonometrical inequalities with applications to the theory of series. *Math. Z. 41*, 1 (1936), 367–379.

[145] Ito, K., and Kunisch, K. A variational approach to sparsity optimization based on lagrange multiplier theory. *Inverse Problems 30*, 1 (2013), 015001.

[146] Ivanov, S., and Pandolfi, L. Heat equation with memory: Lack of controllability to rest. *J. Math. Anal. Appl. 355*, 1 (2009), 1–11.

[147] J. Prüss, G. S. Stability of equilibria for the Stefan problem with surface tension. *SIAM J. Math. Anal. 40* (2008), 675–698.

[148] J. Prüss, J. Saal, G. S. Existence of analytic solutions for the classical Stefan problem. *Math. Ann. 338*, 3 (2007), 703–755.

[149] Ji, Z., and Telgarsky, M. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300* (2018).

[150] Joly, R., and Laurent, C. A note on the semiglobal controllability of the semilinear wave equation. *SIAM J. Control Optim. 52*, 1 (2014), 439–450.

[151] Kakade, S. M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems* (2009), pp. 793–800.

[152] Kalise, D., Kunisch, K., and Rao, Z. Infinite horizon sparse optimal control. *J. Optim. Theory Appl. 172*, 2 (2017), 481–517.

[153] Kalise, D., Kunisch, K., and Rao, Z. Sparse and switching infinite horizon optimal controls with mixed-norm penalizations. *ESAIM Control Optim. Calc. Var. 26* (2020), 61.

[154] Kamenomostskaja, S. On Stefan's problem. *Math. Sbornik 53* (1965), 485–514.

[155] Kienzler, C. Flat fronts and stability for the porous medium equation. *Comm. Part. Diff. Eq. 41*, 12 (2016), 1793–1838.

[156] Kim, T., Adalsteinsson, D., and Lin, M. C. Modeling ice dynamics as a thin-film Stefan problem. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation* (2006), pp. 167–176.

[157] Kinderlehrer, D., and Nirenberg, L. Regularity in free boundary problems. *Ann. Scuola Norm. Sup. Pisa (4) 4* (1977), 373–391.

[158] Kinderlehrer, D., and Nirenberg, L. The smoothness of the free boundary in the one phase Stefan problem. *Comm. Pure Appl. Math. 31* (1978), 257–282.

[159] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[160] Koch, H. *Non-Euclidean singular integrals and the porous medium equation.* Habilitation, University of Heidelberg, 1999.

[161] Koga, S., Diagne, M., and Krstic, M. Output feedback control of the one-phase Stefan problem. In *2016 IEEE 55th Conference on Decision and Control (CDC)* (2016), IEEE, pp. 526–531.

[162] Koga, S., Diagne, M., and Krstic, M. Control and state estimation of the one-phase Stefan problem via backstepping design. *IEEE Transactions on Automatic Control 64*, 2 (2018), 510–525.

[163] Koga, S., Diagne, M., Tang, S., and Krstic, M. Backstepping control of the one-phase Stefan problem. In *2016 American Control Conference (ACC)* (2016), IEEE, pp. 2548–2553.

[164] Koga, S., and Krstic, M. Single-boundary control of the two-phase Stefan system. *Systems Control Lett. 135* (2020), 104573.

[165] Kohn, J., and Nirenberg, L. Degenerate elliptic-parabolic equations of second order. *Comm. Pure Appl. Math. 20*, 4 (1967), 797–872.

[166] Koike, K. Long-time behavior of a point mass in a one-dimensional viscous compressible fluid and pointwise estimates of solutions. *J. Differ. Eq. 271* (2021), 356–413.

[167] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (2012), pp. 1097–1105.

[168] Kunisch, K., and Meinlschmidt, H. Optimal control of an energy-critical semilinear wave equation in 3d with spatially integrated control constraints. *J. Math. Pur. Appl.* (2020).

[169] Ladyzhenskaja, O. A., Solonnikov, V., and Uraltseva, N. *Linear and quasilinear equations of parabolic type*, vol. 23 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, R.I., 1968.

[170] Lannes, D. Well-posedness of the water-waves equations. *J. Amer. Math. Soc. 18*, 3 (2005), 605–654.

[171] Lannes, D. *The water waves problem: mathematical analysis and asymptotics*, vol. 188. American Mathematical Soc., 2013.

[172] Lasry, J.-M., and Lions, P.-L. Mean field games. *Jpn. J. Math 2*, 1 (2007), 229–260.

[173] Le Balc'h, K. Local controllability of reaction-diffusion systems around nonnegative stationary states. *ESAIM Control Optim. Calc. Var. 26* (2020), 55.

[174] Lebeau, G., and Robbiano, L. Contrôle exact de léquation de la chaleur. *Comm. Part. Diff. Eq. 20*, 1-2 (1995), 335–356.

[175] Lebeau, G., and Zuazua, E. Null-controllability of a system of linear thermoelasticity. *Arch. Ration. Mech. Anal. 141*, 4 (1998), 297–329.

[176] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature 521*, 7553 (2015), 436–444.

[177] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural computation 1*, 4 (1989), 541–551.

[178] LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist 2* (2010).

[179] LeCun, Y., Touresky, D., Hinton, G., and Sejnowski, T. A theoretical framework for back-propagation. In *Proceedings of the 1988 connectionist models summer school* (1988), vol. 1, CMU, Pittsburgh, Pa: Morgan Kaufmann, pp. 21–28.

[180] Lequeurre, J. Null controllability of a fluid-structure system. *SIAM J. Control Optim. 51*, 3 (2013), 1841–1872.

[181] Li, Q., Chen, L., Tai, C., and E, W. Maximum principle based algorithms for deep learning. *J. Mach. Learn. Res. 18*, 1 (2017), 5998–6026.

[182] Li, Q., Lin, T., and Shen, Z. Deep learning via dynamical systems: An approximation perspective. *arXiv preprint arXiv:1912.10382* (2019).

[183] Lin, H., and Jegelka, S. Resnet with one-neuron hidden layers is a universal approximator. In *Advances in Neural Information Processing Systems* (2018), pp. 6169–6178.

[184] Lions, J.-L. *Contrôllabilité exaacte, perturbations et stabilisation de systèmes distribués, Tome 1 - contrôllabilité exacte*. Masson, 1988.

[185] Lions, J.-L. Exact controllability, stabilization and perturbations for distributed systems. *SIAM Rev. 30*, 1 (1988), 1–68.

[186] Lions, J. L., and Magenes, E. *Non-homogeneous boundary value problems and applications*, vol. 2. Springer Science & Business Media, 2012.

[187] Lions, P.-L. On the existence of positive solutions of semilinear elliptic equations. *SIAM Review 24*, 4 (1982), 441–467.

[188] Liu, H., and Markowich, P. Selection dynamics for deep neural networks. *J. Differ. Eq. 269*, 12 (2020), 11540–11574.

[189] Liu, X. Null controllability of a class of Newtonian filtration equations. *J. Math. Anal. Appl. 342*, 1096–1106.

[190] Liu, X., and Gao, H. Controllability of a class of Newtonian filtration equations with control and state constraints. *SIAM J. Control Optim. 46*, 6 (2007), 2256–2279.

[191] Liu, Y., Takahashi, T., and Tucsnak, M. Single input controllability of a simplified fluid-structure interaction model. *ESAIM Control Optim. Calc. Var. 19*, 1 (2013), 20–42.

[192] Livni, R., Shalev-Shwartz, S., and Shamir, O. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems* (2014), pp. 855–863.

[193] Lu, Y., Zhong, A., Li, Q., and Dong, B. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *International Conference on Machine Learning* (2018), pp. 3276–3285.

[194] Luckhaus, S. Solutions for the two-dimensional Stefan problem with the Gibbs-Thomson law for melting temperature. *European J. Appl. Math.* (1990), 101–111.

[195] Maity, D., and Tucsnak, M. A maximal regularity approach to the analysis of some particulate flows. In *Particles in flows*, Adv. Math. Fluid Mech. Birkhäuser/Springer, 2017, pp. 1–75.

[196] Maity, D., Tucsnak, M., and Zuazua, E. Controllability and positivity constraints in population dynamics with age structuring and diffusion. *J. Math. Pures Appl.* (2018).

[197] Mallat, S. Group invariant scattering. *Comm. Pure Appl. Math. 65*, 10 (2012), 1331–1398.

[198] Mallat, S. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374*, 2065 (2016), 20150203.

[199] Maronnier, V., Picasso, M., and Rappaz, J. Numerical simulation of free surface flows. *J. Comput. Phys. 155*, 2 (1999), 439–455.

[200] Martin, P., Rosier, L., and Rouchon, P. Null controllability of the heat equation using flatness. *Automatica 50*, 12 (2014), 3067–3076.

[201] Matano, H. *Asymptotic behavior of the free boundaries arising in one phase Stefan problems in multi-dimensional spaces*, vol. 5 of *Lecture Notes in Num. Appl. Anal.* Kinokuniya, Tokyo, 1982.

[202] Mazari, I., and Ruiz-Balet, D. A fragmentation phenomenon for a non-energetic optimal control problem: optimisation of the total population size in logistic diffusive models. *arXiv preprint arXiv:2005.08515* (2020).

[203] Mazari, I., and Ruiz-Balet, D. Quantitative stability for eigenvalues of Schrödinger operator & application to the turnpike property for a bilinear optimal control problem. *Hal preprint* (2020).

[204] Mazari, I., Ruiz-Balet, D., and Zuazua, E. Constrained control of gene-flow bistable reaction-diffusion equations: spatially heterogeneous models and infection-dependent models. *Hal preprint* (2020).

[205] McCann, R. J., and Seis, C. The spectrum of a family of fourth-order nonlinear diffusions near the global attractor. *Comm. Part. Diff. Eq. 40* (2015), 191–218.

[206] Meirmanov, A. On the classical solution of the multidimensional Stefan problem for quasilinear parabolic equations. *Math. Sb. 112* (1980), 170–192.

[207] Mignot, F. Contrôle dans les inéquations variationelles elliptiques. *J. Funct. Anal. 22*, 2 (1976), 130–185.

[208] Miller, L. The control transmutation method and the cost of fast controls. *SIAM J. Control Optim. 45*, 2 (2006), 762–772.

[209] Montanari, A., and Zhong, Y. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *arXiv preprint arXiv:2007.12826* (2020).

[210] Moyano, I. Flatness for a strongly degenerate 1-D parabolic equation. *Math. Control Signals Systems 28*, 4 (2016).

[211] Müller, J. Universal flow approximation with deep residual networks. *arXiv preprint arXiv:1910.09599* (2019).

[212] Muskat, M. *The Flow of Homogeneous Fluids Through Porous Media.* McGraw-Hill, New York, 1937.

[213] Nakoulima, O. Contrôlabilité à zéro avec contraintes sur le contrôle. *CR Math. 339*, 6 (2004), 405–410.

[214] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in PyTorch.

[215] Petrosyan, A., and Shahgholian, H. Parabolic obstacle problems applied to finance. *Recent developments in nonlinear partial differential equations 439* (2007), 117–133.

[216] Phan, D., and Rodrigues, S. Stabilization to trajectories for parabolic equations. *Math. Control Signals Systems 30*, 2 (2018), 11.

[217] Pighin, D. *Long time control with applications.* PhD thesis, 2020.

[218] Pighin, D. The turnpike property in semilinear control. *arXiv:2004.03269* (2020).

[219] Pighin, D., and Zuazua, E. Controllability under positivity constraints of semilinear heat equations. *Math. Control Relat. Fields 8*, 3&4 (2018), 935.

[220] Pinkus, A. Approximation theory of the mlp model in neural networks. *Acta Numer. 8*, 1 (1999), 143–195.

[221] Porretta, A., and Zuazua, E. Long time versus steady state optimal control. *SIAM J. Control Optim. 51*, 6 (2013), 4242–4273.

[222] Porretta, A., and Zuazua, E. Remarks on long time versus steady state optimal control. In *Mathematical Paradigms of Climate Science*. Springer, 2016, pp. 67–89.

[223] Prüss, J., and SIMONETT, G. On the two-phase Navier–Stokes equations with surface tension. *Interfaces Free Bound. 12* (2010), 311–345.

[224] Prüss, J., and Simonett, G. *Moving interfaces and quasilinear parabolic evolution equations*, vol. 105. Springer, 2016.

[225] Prüss, J., Simonett, G., and Zacher, R. Qualitative behavior of solutions for thermodynamically consistent Stefan problems with surface tension. *Arch. Ration. Mech. Anal. 207*, 2 (2013), 611–667.

[226] Radkevitch, E. The Gibbs-Thompson correction and conditions for the existence of a classical solution of the modified Stefan problem. *Soviet Math. Dokl. 43* (1991), 274–278.

[227] Radkevitch, E. Conditions for the existence of a classical solution of a modified Stefan problem (the Gibbs-Thomson law). *Mat. Sb. (translation in Russian Acad. Sci. Sb. Math. 75 (1993), 221–246) 183* (1992), 77–101.

[228] Ramaswamy, M., Roy, A., and Takahashi, T. Remark on the global null controllability for a viscous Burgers-particle system with particle supported control. *Applied Mathematics Letters* (2020), 106483.

[229] Raymond, J.-P., and Vanninathan, M. Null controllability in a fluid-solid structure model. *J. Differ. Equ. 248*, 7 (2010), 1826–1865.

[230] Röger, M. Solutions for the Stefan problem with Gibbs-Thomson law by a local minimisation. *Interfaces Free Bound. 6* (2004), 105–133.

[231] Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review 65*, 6 (1958), 386.

[232] Rosset, S., Zhu, J., and Hastie, T. Margin maximizing loss functions. *Advances in Neural Information Processing Systems 16* (2003), 1237–1244.

[233] Rosset, S., Zhu, J., and Hastie, T. Boosting as a regularized path to a maximum margin classifier. *J. Mach. Learn. Res. 5*, Aug (2004), 941–973.

[234] Rubanova, Y., Chen, R. T., and Duvenaud, D. K. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems* (2019), pp. 5320–5330.

[235] Rubinstein, L. *The Stefan problem*, vol. 8. American Mathematical Soc., 2000.

[236] Ruiz-Balet, D., and Zuazua, E. Control under constraints for multi-dimensional reaction-diffusion monostable and bistable equations. *J. Math. Pur. Appl. 143* (2020), 345–375.

[237] Ruiz-Balet, D., and Zuazua, E. Neural ODE control for classification, approximation and transport. *In preparation* (2021).

[238] San Martín, J., Starovoitov, V., and Tucsnak, M. Global weak solutions for the two-dimensional motion of several rigid bodies in an incompressible viscous fluid. *Arch. Ration. Mech. Anal. 161*, 2 (2002), 113–147.

[239] Santosa, F., and Symes, W. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Statist. Comput. 7* (1986), 1307–1330.

[240] Schoof, C. Marine ice-sheet dynamics. part 1. the case of rapid sliding. *Journal of Fluid Mechanics 573* (2007), 27.

[241] Schoof, C. Marine ice sheet stability. *Journal of Fluid Mechanics 698* (2012), 62–72.

[242] Seidman, T. I., Avdonin, S. A., and Ivanov, S. A. The 'window problem' for series of complex exponentials. *J. Fourier Anal. Appl. 6* (2000), 233–254.

[243] Seis, C. Long-time asymptotics for the porous medium equation: The spectrum of the linearized operator. *J. Differ. Equ. 256*, 3 (2014), 1191–1223.

[244] Seis, C. Invariant manifolds for the porous medium equation. *arXiv preprint arXiv:1505.06657* (2015).

[245] Seis, C. The thin-film equation close to self-similarity. *Analysis & PDE 11*, 5 (2018), 1303–1342.

[246] Serfaty, S., and Serra, J. Quantitative stability of the free boundary in the obstacle problem. *Analysis & PDE 11*, 7 (2018), 1803–1839.

[247] Sontag, E., and Sussmann, H. Complete controllability of continuous-time recurrent neural networks. *Systems Control Lett. 30*, 4 (1997), 177–183.

[248] Sontag, E. D., and Qiao, Y. Further results on controllability of recurrent neural networks. *Systems Control Lett. 36*, 2 (1999), 121–129.

[249] Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res. 19*, 1 (2018), 2822–2878.

[250] Stefan, J. Versuche über die scheinbare adhäsion. *Annalen der Physik 230*, 2 (1875), 316–318.

[251] Su, P., Tucsnak, M., and Weiss, G. Stabilizability properties of a linearized water waves system. *Systems Control Lett. 139* (2020), 104672.

[252] Szegö, G. *Orthogonal Polynomials*, vol. 23. Colloquium Publications, 1939.

[253] Tabuada, P., and Gharesifard, B. Universal approximation power of deep neural networks via nonlinear control theory. *arXiv preprint arXiv:2007.06007* (2020).

[254] Tanner, L. The spreading of silicone oil drops on horizontal surfaces. *Journal of Physics D: Applied Physics 12*, 9 (1979).

[255] Tenenbaum, G., and Tucsnak, M. New blow-up rates for fast controls of Schrödinger and heat equations. *J. Differ. Equ. 243* (2007), 70–100.

[256] Tenenbaum, G., and Tucsnak, M. On the null-controllability of diffusion equations. *ESAIM Control Optim. Calc. Var. 17*, 4 (2013), 1088–1100.

[257] Thorpe, M., and van Gennip, Y. Deep limits of residual neural networks. *arXiv preprint arXiv:1810.11741* (2018).

[258] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*, 1 (1996), 267–288.

[259] Trélat, E. Linear turnpike theorem. *arXiv preprint arXiv:2010.13605* (2020).

[260] Trélat, E., Zhang, C., and Zuazua, E. Steady-state and periodic exponential turnpike property for optimal control problems in Hilbert spaces. *SIAM J. Control Optim. 56*, 2 (2018), 1222–1252.

[261] Trélat, E., and Zuazua, E. The turnpike property in finite-dimensional nonlinear optimal control. *J. Differ. Equ. 258*, 1 (2015), 81–114.

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

[263] Vapnik, V. *The nature of statistical learning theory.* Springer Science & Business Media, 2013.

[264] Vázquez, J., and Zuazua, E. Large time behavior for a simplified 1d model of fluid-solid interaction. *Commun. Part. Diff. Eq. 28*, 9-10 (2003), 1705–1738.

[265] Vázquez, J., and Zuazua, E. Lack of collision in a simplified 1d model of fluid-solid interaction. *Math. Models Methods Appl. Sci 16*, 5 (2006), 637–678.

[266] Visintin, A. Supercooling and superheating effects in phase transitions. *IMA J. Appl. Math.* (1985), 233–256.

[267] von Neumann, J. A model of general economic equilibrium. *Review of Economic Studies 13*, 1 (1945), 1–9.

[268] Vázquez, J. L. *The Porous Medium Equation: Mathematical theory and applications.* Oxford Publishing Group, 2007.

[269] Wei, C., Lee, J. D., Liu, Q., and Ma, T. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems* (2019), pp. 9712–9724.

[270] Wohlmuth, B. Variationally consistent discretization schemes and numerical algorithms for contact problems. *Acta Numer. 20* (2011), 569–734.

[271] Wolpert, D. H., and Macready, W. G. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation 1*, 1 (1997), 67–82.

[272] Yagüe, C. E., and Geshkovski, B. Sparse approximation in learning via neural odes. *arXiv preprint arXiv:2102.13566* (2021).

[273] Yeh, J. *Real analysis: theory of measure and integration second edition.* World Scientific Publishing Company, 2006.

[274] Yun, C., Sra, S., and Jadbabaie, A. Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity. In *Advances in Neural Information Processing Systems* (2019), pp. 15558–15569.

[275] Zamorano, S. Turnpike property for two-dimensional Navier–Stokes equations. *J. Math. Fluid Mech. 20*, 3 (2018), 869–888.

[276] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530* (2016).

[277] Zhang, X., and Zuazua, E. Exact controllability of the semi-linear wave equation. *Unsolved Problems in Mathematical Systems and Control Theory* (2004), 173.

[278] Zhu, H. Control of three dimensional water waves. *Arch. Ration. Mech. Anal.* (2020), 1–74.

[279] Ziane, T. A., Ouzzane, H., and Zair, O. A Carleman estimate for the two dimensional heat equation with mixed boundary conditions. *Comptes Rendus Mathematique 351*, 3-4 (2013), 97–100.

[280] Zuazua, E. Exact controllability for semilinear wave equations in one space dimension. In *Ann. Inst. H. Poincaré Anal. Non Linéaire* (1993), vol. 10, Elsevier, pp. 109–129.

[281] Zuazua, E. Controllability and observability of partial differential equations: some results and open problems. In *Handbook of differential equations: evolutionary equations*, vol. 3. Elsevier, 2007, pp. 527–621.

[282] Zuazua, E. Switching control. *J. Eur. Math. Soc. 13*, 1 (2010), 85–117.

[283] Zuazua, E. Large time control and turnpike properties for wave equations. *Annual Reviews in Control 44* (2017), 199–210.

[284] Zuazua, E. Asymptotic behavior of scalar convection-diffusion equations. *arXiv preprint arXiv:2003.11834* (2020).